RNA GENE DISCOVERY THROUGH THE CLASSIFICATION OF RNA SECONDARY STRUCTURE ELEMENTS

by

Nicholas Erho
B. Sc., Computing Science, Trinity Western University, 2007

A Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Applied Science
in the School
of
Computing Science

© Nicholas Erho 2010
SIMON FRASER UNIVERSITY
Fall 2010

All rights reserved. This work may not be
reproduced in whole or in part, by photocopy
or other means, without the permission of the author.

APPROVAL

Name:	Nicholas Erho
Degree:	Master of Applied Science
Title of Thesis:	RNA Gene Discovery Through the Classification of RNA Secondary Structure Elements

Examining Committee: Dr. Robert D. Cameron
Chair

Dr. Kay C. Wiese, *Senior Supervisor*
Associate Professor, Computing Science
Simon Fraser University

Dr. Ke Wang, *Supervisor*
Professor, Computing Science
Simon Fraser University

Dr. Mohamed Hefeeda, *SFU Examiner*
Associate Professor, Computing Science
Simon Fraser University

Date Approved: November 29, 2010

SFU SIMON FRASER UNIVERSITY
LIBRARY

Declaration of Partial Copyright Licence

The author, whose copyright is declared on the title page of this work, has granted to Simon Fraser University the right to lend this thesis, project or extended essay to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its users.

The author has further granted permission to Simon Fraser University to keep or make a digital copy for use in its circulating collection (currently available to the public at the "Institutional Repository" link of the SFU Library website <www.lib.sfu.ca> at: <http://ir.lib.sfu.ca/handle/1892/112>) and, without changing the content, to translate the thesis/project or extended essays, if technically possible, to any medium or format for the purpose of preservation of the digital work.

The author has further agreed that permission for multiple copying of this work for scholarly purposes may be granted by either the author or the Dean of Graduate Studies.

It is understood that copying or publication of this work for financial gain shall not be allowed without the author's written permission.

Permission for public performance, or limited permission for private scholarly use, of any multimedia materials forming part of this work, may have been granted by the author. This information may be found on the separately catalogued multimedia material and in the signed Partial Copyright Licence.

While licensing SFU to permit the above uses, the author retains copyright in the thesis, project or extended essays, including the right to change the work for subsequent purposes, including editing and publishing the work in whole or in part, and licensing other parties, as the author may desire.

The original Partial Copyright Licence attesting to these terms, and signed by this author, may be found in the original bound copy of this work, retained in the Simon Fraser University Archive.

Simon Fraser University Library
Burnaby, BC, Canada

Abstract

A multi-layered exploratory look at the use of secondary structure elements in ribonucleic acid (RNA) gene finding is performed. Individual structural element metrics are analyzed for their ability to act as structural RNA gene signals. Additionally, each structural element is analyzed for its ability to detect structural RNA gene sequences by training and testing classifiers which utilize the structural element's metrics to classify candidate RNA sequences. Finally, groups of structural elements are examined, by voting the prediction results of the individual structural element classifiers together to determine if a candidate sequence is a structural RNA gene. The tests reveal that the *external loop*, *structure*, *stemloop*, *hairpin loop*, and *tail* structural elements produce significant signals for structural RNA genes. Many groups of structural elements were found to have potential but particularly the *stemloop* and *hairpin loop* structural element combination stood out for its practicality and strong results.

To my family, for being family.

“Medicine makes people ill, mathematics makes them sad, and theology makes them sinful.”

— Luther, Martin (1483-1546)

Acknowledgments

I would first like to acknowledge my senior supervisor, Dr. Kay Wiese, for introducing me to the problem of RNA gene finding and providing the freedom for me to explore different facets of computing science. Kay provided not only computer hardware I needed to complete this research but also the financial support to attend CIBCB 2010 Conference in Montreal, Canada. I am grateful for the encouragement and guidance he provided.

A special thanks to Dr. Ke Wang for being a supervisor on this research project and introducing me to the concepts of data mining. Likewise, I would like to thank Dr. Mohamed Hefeeda for serving as my external examiner and Dr. Robert Cameron for serving as the chair for my defense.

Denny Chen and Andrew Hendriks were two of my colleges in Dr. Wiese's lab. Denny would helpfully engage in technical discussions with me, often providing insights and new concepts to try. While Andrew is an excellent teacher explaining many research related concepts to me, I found his analysis of contemporary society to be most enlightening and thought provoking. I was privileged that Kirt Noël, a previous student of Dr. Wiese, took the time to meet with me to discuss his thesis research and provide me with several suggestions for directions my research could take.

My sincere thanks to my parents and friends for their encouragement and support. Particularly the support of Darlene Erho, Linda Jia, and Rajon Bhuiyan who proof read my papers and thesis. I would not have finished my thesis without the encouragement of these people.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Structural ribonucleic acid (RNA) gene finding is essentially a search problem—finding a needle in a haystack. The needle is represented by the structural RNA gene (SRNAG), and the genome, excluding the SRNAG regions, is represented by the hay. Unfortunately, when SRNAG finding is viewed at the nucleotide sequence level, the needles and hay appear similar because the sequences of SRNAGs are not highly conserved across evolution [44, 50]. So perhaps a better analogy for the problem of SRNAG finding would be finding a needle produced by one manufacturer in a stack of needles produced by another manufacturer. This problem is further complicated by the fact that there are many SRNAGs within a genome changing the goal of the search problem to the more general task of finding as many of these genes as possible while keeping the number of false positives to a minimum. Likewise, there are many different families of SRNAGs, often with very different properties. Even within each family the properties of these SRNAGs can vary significantly and yet genes from all these RNA families need to be located. In the analogy, this is represented by a blurring of the target needles' specifications so that there is a general idea of what a target needle might be like, yet no specific characteristics to search for. In a similar way, the properties of the background genome are a product of evolutionary forces which often leads to significant variation in properties. These concepts transform the haystack analogy further, so that there are a set of manufacturers whose needles need to be located in a pile of needles produced by other needle manufacturers while all the needles have some variance from their specified properties.

At the onset this problem looks daunting, as by simply observing the needles there is no accurate way to distinguish between the two sets of manufacturers. Fortunately, in the analogy, individual manufacturers produce needles in distinctive ways, applying different forces and processes to mold and form them. These differences in the manufacturing process can help determine properties of needles which can be exploited to distinguish a needle produced by one manufacturer from that of another. In the same way evolutionary forces driving the formation and preservation of SRNAGs differs from the evolutionary pressure on the genome's non-SRNAG regions. Hopefully, these differences can be used to exploit underlying properties of SRNAGs which are useful in SRNAG finding.

SRNAGs utilize their molecular shape to provide functionality in the cell. If the shape of these molecules is altered, the SRNAG may not be able to carry out its role in the cell which can lead to the loss of a cellular function or even cell death. For this reason the shape of SRNAGs is highly conserved across evolution. Since non-gene genome sections have no direct relationship with cellular functionality, there is little evolutionary pressure placed on these regions of the genome and hence the RNA structures produced by non-SRNAG regions will tend to have different properties than regions holding functioning SRNAGs. Discovery and exploitation of the metric distribution differences caused by evolutionary pressure in SRNAG and non-SRNAG regions of the genome is a key component of SRNAG finding.

Solving the RNA gene finding problem is key for the advancement of several biological research areas, as RNA gene finding is aimed to discover and exploit sequence features unique to RNA genes using computational techniques. The computational techniques may provide a statistically significant means to distinguish the RNA genes from the protein coding and non-coding sections of a genome [39]. With the advancement of genome sequencing technology came a wealth of data for biologists to analyze. Part of that analysis process involves annotating the different regions of the genomes. A statistically significant method of distinguishing SRNAGs from the rest of the genome is a feasible way to help annotate these genomes, as it can indicate areas of the genome which are likely to be SRNAGs, allowing researchers to focus their time verifying the most probable candidate genes. Likewise, there is a feeling within the biology research community that there are potentially a large number of undiscovered RNA genes, making a high accuracy RNA gene finder immensely valuable to the scientific community, just as protein gene finders have already been [50, 9]. It is possible to discover structural RNAs through biochemical processes, but these methods are slow, costly, and often ineffective [50]. Furthermore, since SRNAGs are some of the building blocks in the cell, accurate SRNAG annotation gives biology researchers more cellular components to work with, opening the door to experimental monitoring of expression levels, functional assay by deletion or mutagenesis, structural and functional analysis, and identification of interaction partners [9, 39].

There are several goals of this thesis, the first is to propose and test a novel SRNAG finding method which exploits a collection of structural information, derived from folding a given RNA sequence, as a distinguishing signal for SRNAGs. This SRNAG finding method will not only act a proof of concept showing the ability to use specific structural elements for SRNAG finding, but will allow allow the exploration of which structural elements are useful for SRNAG finding, which is the second goal of this thesis. The third goal involves figuring out which properties of the structural elements allow them to produce strong SRNAG finding signals. In addition to which specific structural elements can be used to produce strong SRNAG finding signals, the fourth goal of this research is to determine which groups of structural elements work well together to produce a strong SRNAG finding signal. Achieving this forth goal is done using a configurable voting system to combine the results of several structural element models to build a multi-structural element SRNAG prediction system.

In order to achieve these objectives this thesis is broken down into several chapters. The *Introduction* chapter defines the problem of SRNAG gene finding, provides motivation for solving it, and briefly covers some of the key biological concepts used in this thesis. Chapter 2 discusses previous methods for SRNAG finding and compares them to the SRNAG finding approach taken in this thesis. This novel approach to SRNAG finding is developed in Chapter 3, where each step in the classification process is discussed in detail. The *Methods* chapter describes the two experiments used to test the SRNAG finder. One experiment deals with testing the classification system under favorable conditions, while the other tests the classifier in harsh conditions. Chapter 5 presents the results and analysis of the experiments described in Chapter 4. Finally the last chapter gives a conclusion and discusses the future direction of this research. To provide additional background information, Appendix A gives some basic statistical definitions, Appendix B introduces the theory behind support vector machines from a high level, geometric perspective, Appendix C provides information on nucleotide shuffling, and Appendix D supplies an explanation as to why protein finding methods are not suitable for RNA gene finding. The last appendix, Appendix E, tabulates most of the results discussed in this thesis.

1.1 Different Types of RNA Gene Finding

There are many different methods available for RNA gene finding. The choice of technique depends on the data available in the gene finding problem and the range of different types of genes attempting to be discovered. Typically, RNA gene finding methods are split into two categories: homology based and *ab-initio* prediction methods. Although there are many different homology based approaches to RNA gene finding, typically they exploit similarities in evolutionary related sequences to create signals for RNA genes [39]. For example, although RNA gene sequences are less conserved across evolution than protein genes, because RNA gene structures are conserved the RNA gene sequences do display certain conservation characteristics [39, 7]. If the pairing of certain nucleotides is critical to a structural RNA gene's function, then there will be a tendency for these nucleotides to undergo correlated mutations. Homology gene finding methods can exploit theses correlated mutations to produce a RNA gene signal. When *ab-initio* RNA gene finding methods are used, no sequence annotation or external sequences are required other than the genome sequence to be searched. This restriction makes *ab-initio* RNA gene finding much more difficult than homology based approaches and is "more or less an unsolved problem" [39]. The SRNAG finding method presented in this thesis attempts to solve the *ab-initio* SRNAG gene finding problem, hence the rest of this document does not deal with homology RNA gene finding techniques.

1.2 RNA and DNA

While RNA molecules are the focus of this thesis, they are best described alongside the better known deoxyribonucleic acid (DNA), as they are closely related molecules. RNA, like DNA, are nucleic acids, macromolecules composed of polymers of monomeric nucleotides [7]. Nucleotides are comprised of three elements: a nitrogenous base, a five-carbon sugar, and a phosphate group [7, 44]. Both DNA and RNA link the 5-carbon sugar of one nucleotide to the phosphate of the next nucleotide in an alternating pattern through a shared oxygen atom to form the backbone of the nucleic acid chain [7]. These bonds to the shared oxygen atom attach the phosphate group to the carbon at the 3' end of the sugar and to the carbon at the 5' end of the sugar, giving the sequence directionality [44, 7]. By convention the 5' terminus refers to the start of the sequence and the 3' terminus refers to the end of the sequence [44, 7]. In DNA this backbone uses pentose sugars, while in RNA ribose sugars are used [7]. Furthermore, although DNA and RNA share the nitrogenous bases Adenine (A), Cytosine (C), and Guanine (G), they have one base that differs [7]; DNA uses the base Thymine (T), while RNA uses the base Uracil (U) [7]. Certain nitrogenous bases can bond to each other through base pairing, allowing DNA and RNA strands to have complex yet well defined interactions [44]. In general, these interactions follow Watson-Crick base pairing rules where GC, AT, and AU can bond; however less frequently non-Watson-Crick base pairing can happen where GU bonds form [7, 44]. These base pairing interactions are due to intermolecular hydrogen bonds, where three hydrogen bonds link GC, while only two hydrogen bonds link AT and AU [7, 44]. The fact that GC base pairs have three hydrogen bonds while AT and AU base pairs only have two hydrogen bonds helps make the pairs more stable [44, 73]. Likewise, G and C base stacking energies contribute more to the stability of the molecules than T, A, and U bases [73]. The increased stability of GC pairs over the other pairing configurations result in DNA and RNA sequence tending to have higher GC content, as higher concentration of GC content allows for stabler molecules[1] [52, 73]. Finally, another difference between DNA and RNA is that DNA is composed of two chains of deoxyribonucleotides bound to each other through base pairing interactions, while RNA is typically a single chain of ribonucleotides, although this single strand has the ability to fold back on itself through base pairing interactions to form double stranded sections. It is this folding through base pairing which enables structural RNAs to form specific molecular shapes, allowing them to be functional molecules within the cell.

1.3 Genomes and Transcription

Both DNA and RNA use patterns of nucleotide bases to encode genetic information. Genetic information is a term describing the set of instructions needed by the cell to build cellular components such as proteins and functional RNAs. Although DNA is typically viewed as the long-term storage molecule of the cell, it is known that in some retroviruses, RNA takes on this role, making up the virus's genome [44]. Typically, genomes are comprised of several chromosomes, which in turn are comprised of DNA molecules [44, 7]. For the purposes of this thesis, a genome is considered simply as a single sequence given by the

[1] As GC content increase AT or AU content necessarily decrease.

alphabet A,T,C,G, containing some regions which are transcribed[2] into RNA. The regions of the genome which are not transcribed are known as non-coding (NC) regions [44]. It is important to note that even after a RNA strand is transcribed it still may undergo various post processing steps. These extra steps are especially common in eukaryotes[3] and include the processing events of end-modification, splicing, cutting, and chemical modification [7]. Because these events modify the RNA transcript, the sequence of the final RNA molecule may be different than the genomic sequence it was copied from, which means that the genomic information that is being used to fold or find structural RNA genes may at times not accurately reflect the final structural RNA molecule. Post transcription modification (PTM) is an important consideration in the task of SRNAG finding as typically the genome sequence is searched without consideration for PTM, yet it is possible that PTM changes the sequence of the resulting RNA enough that the gene finder does not detect the genome sequence. For a SRNAG finder to consider PTMs it would add considerable complexity to an already challenging problem, so this thesis does not deal with this added complexity.

1.4 Classes of RNAs

Although the product of transcription is always an RNA sequence[4], there are many classes of RNAs. The breakdown of RNA categories is shown in Figure 1.1. If the RNA transcribed is later translated into a protein it is termed messenger RNA (mRNA) and is part of the class of coding RNAs, as its primary purpose is to transport information from the genome [7]. Non-coding (ncRNA), also known as functional RNAs (fRNA), are not translated but conduct their biological role as RNA sequences [39]. Within this class of ncRNAs there is a subclass of structural RNAs which are RNA transcripts that gain their cellular role by folding into tertiary structures [7, 39, 44]. These tertiary structures can catalyze chemical reactions or play other biological roles in the cell in a way analogous to protein activity [44]. Within the structural RNA class there is transfer RNA (tRNA), ribosomal RNA (rRNA), small nuclear RNA (snRNA) [44], small nucleolar RNA (snoRNA), ribonuclease P (RNase P), transfer-messenger-RNA (TmRNA), and signal recognition particle RNA (SRP RNA) [7].

[2] Transcription is the process by which a complementary RNA sequence is created using a DNA template [44, 7].

[3] Eukaryotes are cells which contain a nucleus to hold their genetic material, as apposed to prokaryotes where the genetic material is suspended in the cytoplasm [7]

[4] RNA molecules are typically less than a few thousand nucleotides in length [7].

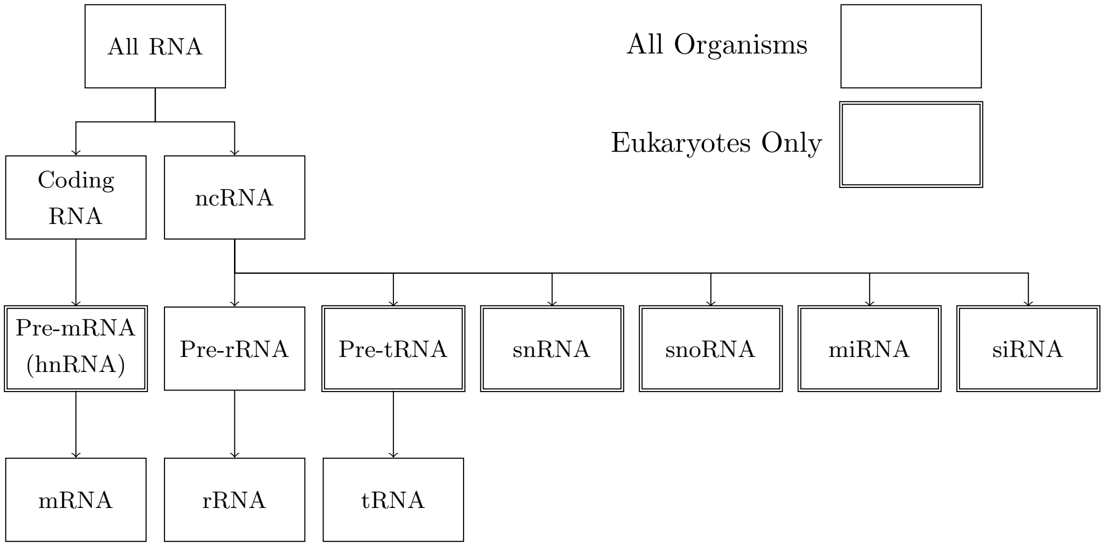

Figure 1.1: RNA Hierarchy. Repoduction of figure in [7].

tRNA, the first type of RNA gene discovered [39], are small molecules, around 80 – 120 nucleotides in length, characterized by their cloverleaf shape. The unique shape of the tRNA molecule provides an active site to bind specific amino acids and an anticodon region, allowing it to interface with a RNA sequence at specific locations so that it can perform its role of transporting the next amino acids to a growing polypeptide chain in the process of translation. rRNA also plays a role in translation, where it acts as the catalytic component of ribosomes. Ribosomes are responsible for binding to mRNA and carrying out protein synthesis. There are several different types of rRNA, including 5S rRNA, 5.8S rRNA, 16S rRNA, 18S rRNA, 23S rRNA, and 28S rRNA. Likewise, snRNA are small RNA molecules involved in RNA splicing, transcription factors, RNA polymerases II, and telomere regulation processes. Within the class of snRNA exists snoRNA, which play a role in RNA biogenesis and facilitate chemical modification of rRNA, tRNA, and other snRNAs. All snRNAs are located in the nucleus of eukaryotic cells. RNase P is a ribozyme responsible for cleaving RNA in the cell, and TmRNA with a three fold purpose rescues stalled ribosomes, tags incomplete polypeptides, and promotes aberrant mRNA for degradation. Finally, SRP RNA is a protein-RNA complex that detects and moves specific proteins across the endoplasmic reticulum (ER).

It is important to note that there are other functional RNAs that may not exploit their structure to accomplish their cellular role, but instead solely make use of their primary sequence [44]. Examples of functional RNAs include microRNA (miRNA), which play a role in gene expression regulation and small interfering RNA (siRNA), which play a role in the RNA interference pathway.

Although SRNAG finders have been described which are able to locate only a single type of RNA gene, many are general SRNAG finders and are able to locate a number of different RNA gene types. There are advantages and disadvantages with both types of RNA gene finders. A general finder only has to be run once to annotate many different RNA gene types, while many different specific SRNAG finders would have to be run to fully annotate the SRNAGs in a genome, likely a time consuming task. However, specific SRNAG finders, often of a simpler design, will run faster than a general SRNAG finder. Likewise due to their focused approach, a specific SRNAG finder may be able to locate the specific RNA genes with higher accuracy, while due to its more flexible construction a general SRNAG finder may be able to locate previously undiscovered gene classes, as it will tend to consider trends among most SRNAGs instead of considering only specific features of a certain type of RNA molecule. This thesis focuses on a general approach to SRNAG finding, where seven classes of structural RNAs are considered: tRNA, 5S rRNA, 16S rRNA, 23S rRNA, RNase P, SRP RNA, and TmRNA. These structural RNAs cover the variety of structural RNAs while still having large collections of known examples to use for training and testing purposes.

1.5 Sequence, Structure, Function, and Thermodynamic Models

A fundamental theme in modern biology is the idea that the function of a molecule is derived from its structure which in turn is derived from the molecular sequence [7, 44]. SRNAGs are no exception to this theme, where the structure of the RNA molecule is directly dictated by the sequence of its gene's nucleotide bases [7, 44]. This sequence-structure relationship which determines how the RNA molecule folds is driven largely by thermodynamics. Hence the RNA structure prediction problem is often approximated as a thermodynamics problem in which the RNA sequence represents a thermodynamic system under certain temperature and pressure constraints [77]. Within this thermodynamic system, the base pairing configuration which results in a RNA structure with the highest energy loss[5] is deemed a

[5] In the terms of Gibbs free energy, this is the lowest overall energy stored by the RNA molecule system.

highly stable structure and in turn a highly probable configuration for the RNA molecule to naturally reside. Dynamic programming algorithms, like those found in the popular RNA secondary structure prediction software MFOLD [77] and RNAfold in the Vienna RNA Package, can numerically find this minimum free energy configuration in a running time of $O(n^3)$, where n is the length of the sequence folded and the energy model is assumed to be a constant time calculation. It should be made clear that these thermodynamic RNA secondary structure prediction models are only approximate and the real secondary structure of the RNA sequence may vary, often substantially, from the predicted structure [44]. This relationship between the sequence, structure, and function is what the SRNAG finder will exploit in this thesis. Since the structure of structural RNA molecules provides a function that is related to their structure, the structure-function relationship can be exploited by searching for specific structure properties as the structures of SRNAGs will tend to be preserved across evolution. As already mentioned, while the structure of SRNAGs is preserved across evolution, the sequence is not; however, the sequence-structure relationship can be utilized by modeling regions of the genome as thermodynamic systems and folding them to obtain their secondary structure. In this way, the genome nucleotide sequences can be evaluated by determining their structure through folding and then measuring properties of the structure for the classification engine to utilize.

1.6 RNA Structural Elements

As mentioned in the previous section the SRNAG finder presented in this thesis will collect data from the secondary structures of folded RNA sequences. This section defines RNA secondary structure and builds a vocabulary of components which comprise the secondary structures.

The primary structure of RNA is given simply by its sequence, while the secondary structure of RNA deals with planar base pairing interactions between antiparallel segments of the sequence [70]. Tertiary RNA structure deals with interactions between either two helices, two unpaired regions, or one unpaired region and a double stranded helix [70]. These tertiary interactions often break the planar nature of secondary structure forcing a three-dimensional representation of the molecule. Interactions between multiple folded RNA molecules or protein molecules and RNA molecules is known as quaternary structure. This thesis focuses on RNA structures no higher than secondary structure as secondary structure prediction has efficient, polynomial running time algorithms, while finding stable tertiary structures is known to be NP-hard [76], making it impractical for RNA gene finding.

For the purposes of this thesis several definitions of RNA structural elements are needed. Pictorial representations of the structural elements can be found in Fig. 1.3 and Fig. 1.4, and a summary of the descriptions can be found in Table 1.1. The most basic of all structural elements is the *unpaired* element, which is simply any region of unpaired bases. Complementary to the *unpaired* element is the *stack* which is defined as any uninterrupted sequence of double stranded RNA.

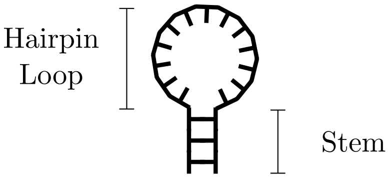

Figure 1.2: Basic Stemloop.

If a stacking region closes an *unpaired* region, this forms a basic *stemloop* (see Figure 1.2). The loop component of the *stemloop* is called the *hairpin loop*, while the *stack* component is known as the *stem*. It is not required for the *stem* to be completely paired, as it may contain some *unpaired* elements. A *stem*, with unpaired nucleotides on both sides of the double stranded region are *internal loops*; however, if unpaired nucleotides only interrupt on one side of the *stem*, the structural element is a *bulge*. For the purposes of this thesis when referring to a collection of both *bulges* and *internal loops*, the term *loop* will be used.

A *multiloop* is a structural element from which three or more *stems* protrude with the possibility of *unpaired* elements between the *stems* spacing them apart. These *unpaired* regions between the protruding *stems* are known as *joints*. In each RNA structure, there is always one special *multiloop*-like component called the *external loop*. This structure contains both ends of the nucleotide sequence and contains one or more *stems* branching off from it. If the RNA strand ends are unpaired, these *unpaired* regions are defined as *tails*. In this thesis when referring to a collection of *unpaired* regions in a *multiloop* or *external loop*, that is, *tails* and *joints*, the term *joint-tail* is used and when referring to both *multiloops* and *external loops*, the term *junction* is used, because these structures typically are a junction for multiple branches protruding from them.

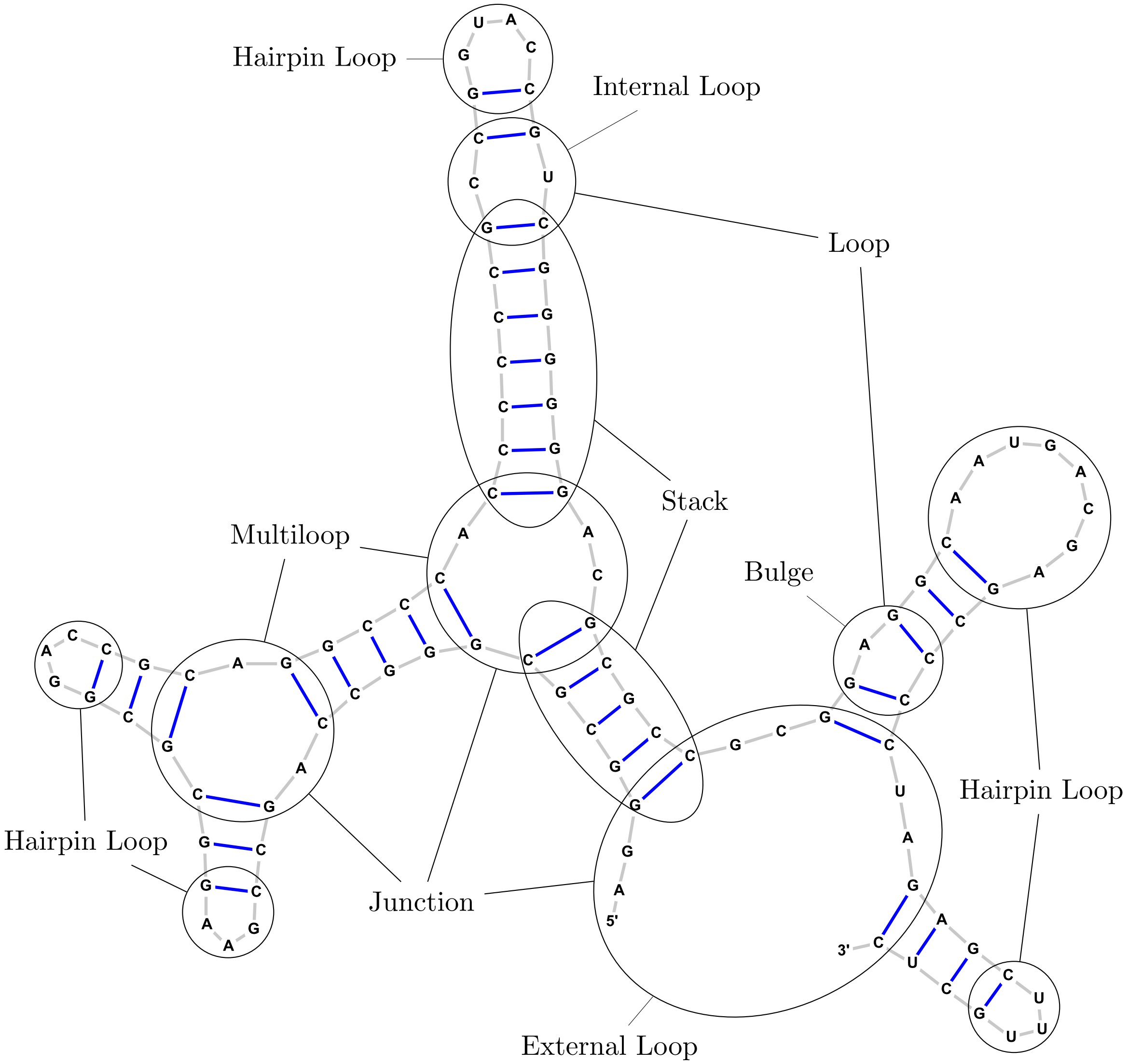

Figure 1.3: RNA Structural Elements A. This figure depicts a RNA secondary structure with labeled structural elements.

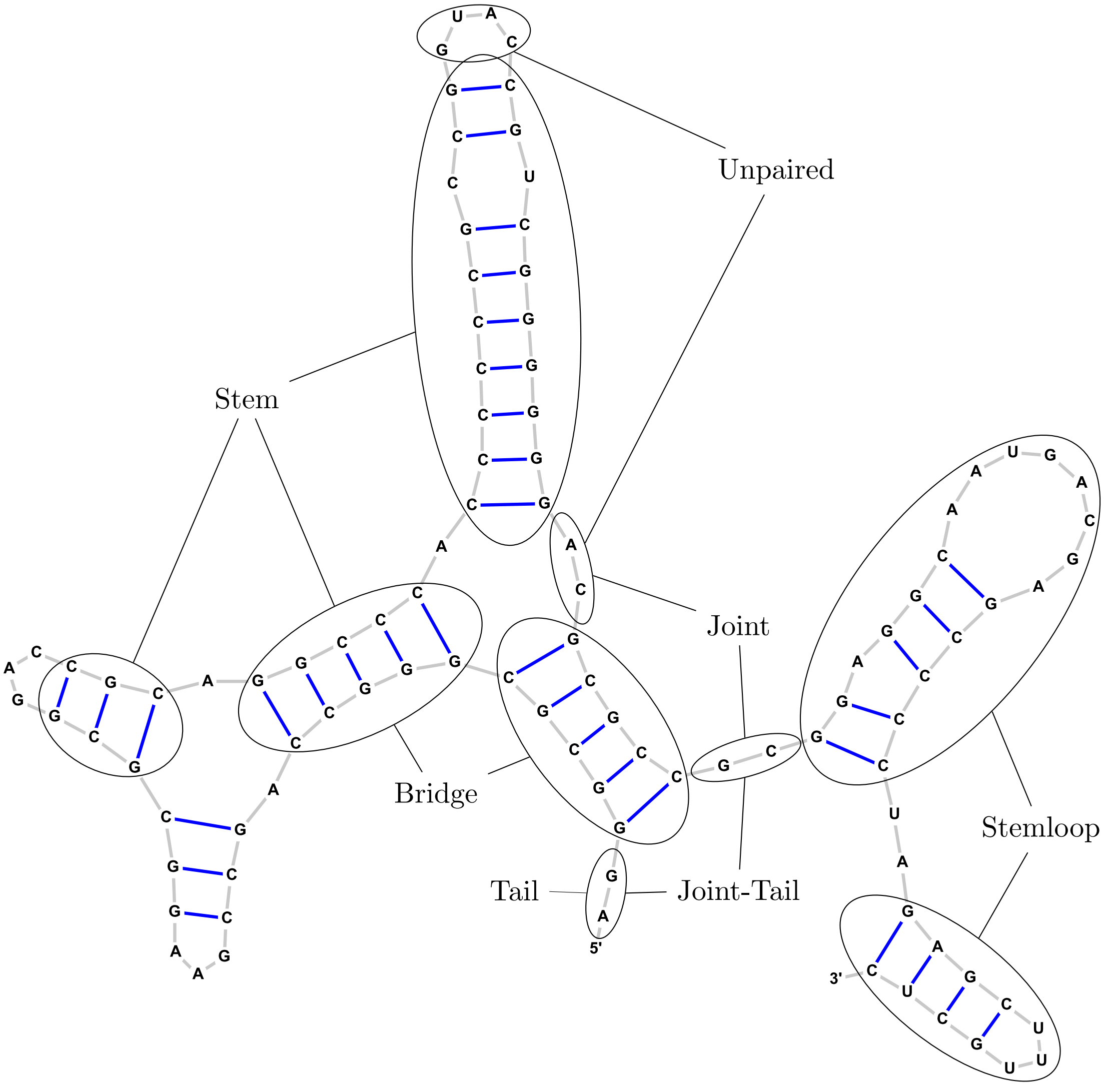

Figure 1.4: RNA Structural Elements B. This figure depicts a RNA secondary structure with labeled structural elements.

Structure	**Description**
Stack	Any uninterrupted sequence of double stranded RNA.
Unpaired	Any unpaired region.
Bulge	Any unpaired region interrupting a stack on only one side of the double stranded region.
Internal Loop	Any unpaired region interrupting a stack on both sides of the double stranded region.
Loop	Either an internal loop or a bulge.
Stem	Any double stranded region composed of at least one stack and any number of internal loops and bulges.
Stemloop	Any sequence region which has folded back on itself to produce a stem and a hairpin loop.
Hairpin Loop	Any loop created directly due to a sequence folding back on itself.
Bridge	Any stem which joins two junction elements.
Tail	Any unpaired region at the ends of the nucleotide sequence.
External loop	The only structural element containing the ends of the nucleotide sequence, with one or more stems protruding from it.
Multiloop	Any region other than an external loop in which three or more stems protrude.
Junction	Either an external loop or a multiloop.
Joint	Any unpaired region in a junction between protruding stems.
Joint-Tail	Either a tail or a joint.
Structure	Entire structure, composed of all sub-components.

Table 1.1: Structural Element Descriptions. Each of the structural elements investigated in this thesis are listed along with their description.

Larger RNA gene structures often have several *junctions* connected by a *stem*-like component. This structural element which connects *junction* elements is identified as a *bridge*. Finally the whole structure, made up of all the different structural components is likewise classed a structural element and is known as the *structure*. The relationship between the structural elements is shown in Figure 1.5.

1.7 Chapter Review

This chapter introduces the problem of SRNAG finding and provides the motivation behind solving it. Additionally, the goals and direction of the thesis are presented. The basic

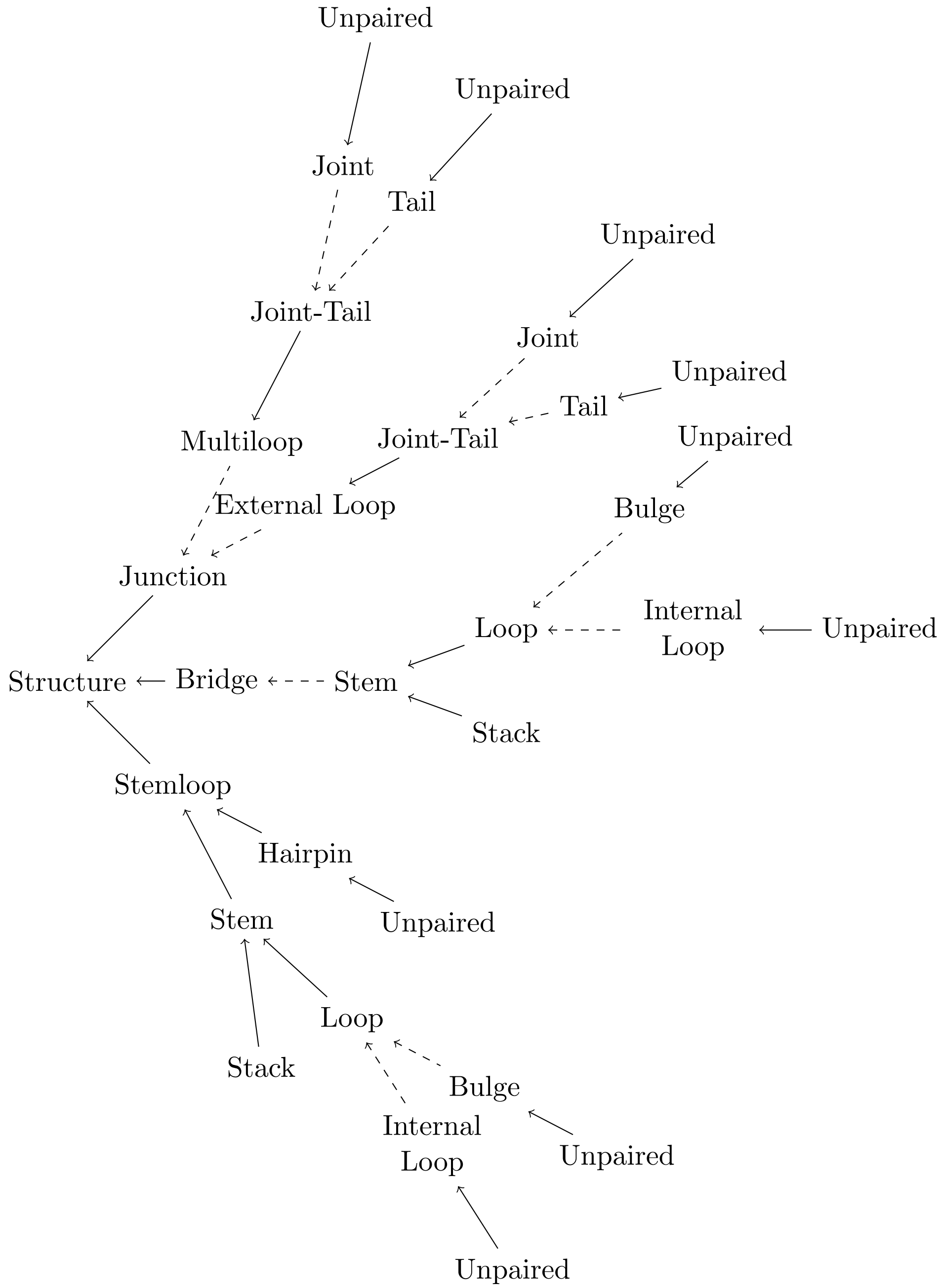

Figure 1.5: Structural Element Hierarchy. Solid lines indicate a has-a relationship and dashed lines show a is-a relationship between the structural elements.

biochemistry of RNA and DNA and the roles of these molecules in the cell is discussed and RNA secondary structure elements are defined. The next chapter deals with the historical approaches to the SRNAG finding problem and discusses the novel method presented in this thesis.

Chapter 2

Background

As evident by the many classes of RNA gene prediction discussed in the previous chapter, a large set of diverse researchers have tackled the problem of RNA gene finding. This chapter outlines the history of *ab-initio* RNA gene finding from its conception to the current state of the art methods. All the RNA gene finding methods presented rely on some unique property, such as free energy, base composition, motifs, or structural patterns, of SRNAG regions which can be used as a signal to distinguish SRNAG regions from NC and coding regions of the genome. These historical gene finding methods are then contrasted with the method introduced in this thesis.

2.1 History of ab-initio RNA gene finders

Staden pioneered RNA gene finding by developing the first RNA gene finder in 1980 [58]. He was motivated by the realization that as the large number of sequencing projects underway came to completion the need to search these sequences for tRNA gene regions quickly and accurately would grow to a point where biologists could no longer effectively approach the problem by visually scanning the DNA sequence [58]. Staden utilized the knowledge that tRNA secondary structures conform to the shape of a cloverleaf pattern and the fact that their sequences typically contain certain bases at specific locations in their structure. This conservation of structure was exploited by developing a computer program which would search a DNA sequence for these common features and display the prospective tRNA genes to the users in both one and two dimensional forms [58, 21]. The authors reported that the method was successfully applied to locate tRNAs in mitochondria DNA, noting that mitochondria tRNAs differ in several aspects to previously studied tRNAs although still retaining the key structural elements needed by the program [58, 21, 4].

While Staden exploited a conserved structure pattern to search for tRNAs, Maizel's group, in 1988, attempted to show that a generic RNA gene finder could be built based on using secondary structure free energy as a statistically significant metric for distinguishing SRNAG regions from NC regions [12, 31, 29]. The group demonstrated this claim by using a Monte Carlo method to show that observed minimum free energy of a gene region is lower than expected by chance and showed the statistical significance of the Monte Carlo simulation with a z-score calculation (see Appendix A.1.1), where the difference between the observed minimum free energy of a region and the average minimum free energy of a population of the shuffled sequences, with the same length and nucleotide composition, but different nucleotide order, is divided by the standard deviation of that shuffled sequence population's minimum free energy [31]. In order to calculate the free energy along the genome, Maizel's group use a fixed length window which was slid along the genome sequence to select successive genome regions for folding (see Figure 2.1), using a modified vectorized version of MFOLD[1] [31]. Once the authors had calculated the z-score from these sequence segments and the observed free energy, the z-scores were plotted allowing researchers to visually detect SRNAG regions where the z-score dipped significantly. Using this vectorized version of the folding algorithm the authors were able to reduce the complexity of the algorithm to $O(n^3)$, where n is the number of nucleotides in the genome. It is important to understand that n could be potentially very large and even with the vectorized version of the folding algorithm running on a Cray X-MP 24 supercomputer, the proposed method could not handle a window size of greater than 200 nucleotide bases or a genomic sequence greater than 1000 bases in a practical time [12, 44].

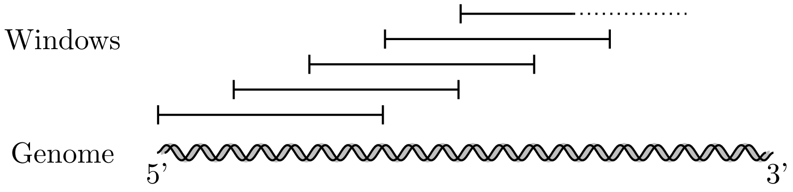

Figure 2.1: Sliding Window Scheme. This figure is based on Figure 2.1 in [44].

In a subsequent paper Maizel's group describes how they tackle the running time issue

[1] The dynamic programming algorithm proposed by Zucker.

by reducing the processing requirements of the Monte Carlo simulation by using a regression model to predict the mean and standard deviation of the minimum free energy of a population of random sequences with a given mononucleotide content. In short, they first used a regression model to determine the relationship between length and mononucleotide content value by computing the free energy mean and standard deviation for different length sequences for three groups of random sequences each consisting of the same mononucleotide content[2] [12]. Then they used this energy-length relationship to create a regression model to predict how the free energy will change as GC content increases given a length by using a least-squares fit on the free energy mean and standard deviation computed from increasing GC content populations. The resulting relationship allows for a length and nucleotide content to be used to predict the average and standard deviation of the free energy of a sample population with those characteristics [12]. With this equation in hand, the previous SRNAG finding method described by these researchers can calculate an estimation for the mean and standard deviation of a window's free energy, instead of having to compute it [12]. This new method was reported to be able to recognize SRNAG regions as effectively as the first method but it took only 70 seconds to calculate what the previous method did in 150 hours [12, 44]. Although the new method improved the computational efficiency greatly, it still did not address the $O(n^3)$ running time due to the folding process [12, 44].

These two studies claimed to show that the observed free energy values for a region were at least three standard deviation units less than the average free energy values for the shuffled sequences and the predicted RNA genes were "consistent with the empirical result from analysis using ribonuclease and cobra venom nuclease digestions" and hence free energy is a useful metric in SRNAG finding [31, 12]. However, the methods presented in these papers were met with some criticism because even with the reduced complexity of the second algorithm the computational steps required makes Mazial's group's method impractical to use with longer segment lengths and because the longest subsequences tend to be the best thermodynamically it is difficult to identify "meaningful high-scoring subsequence unless one repeats the search with multiple different fixed-length window sizes" [50].

Rivas and Eddy disputed the usefulness of free energy as a SRNAG finding metric basing their dispute on three test methods for RNA gene prediction [50]. The first method was a reimplementation of the method proposed by Maizel's group using a thermodynamic

[2] Group 1: A: 42%, C: 6%, G: 20%, U: 32% Group 2: A: 18%, C: 18%, G: 50%, U: 14% Group 3: A: 6%, C: 42%, G: 30%, U: 22%

model to produce a signal, the second method used a stochastic (probabilistic) context free grammar (SCFG) to produce a signal, and the third method simply used base composition as a signal.

The implementation of the SCFG closely mimicked the rules used in the MFOLD thermodynamic model except the thermodynamic scores were replaced with probabilities [50]. The SCFG used a log-odds scoring system which causes the score of a subsequence to get worse as the subsequence becomes longer, in effect acting as a "local alignment" and allowing high scoring subsequences to be identified under a maximum target length, w [50]. This property of the SCFG method contrasts the thermodynamic energy method where, as previously discussed, the longest subsequence tended to be the best scoring one and hence in order to have a fair comparison between several different length subsequences the algorithm must be run several times with different window sizes [50]. Since the SCFG is a trained model, it permits the inclusion of statistical biases which are not represented in the current RNA thermodynamic models [50]. The time complexity of this SCFG algorithm is $O(Lw^2)$ to process a genome of length L with a maximum sized subsequence w being evaluated [50]. The authors tested the SCFG method on several genomes of different average GC content and found that there was "a strong correlation between the strength of a tRNA hit and the difference in GC base composition between the tRNA and the background base composition of the given organism" [50]. This evidence that the difference in GC content is the real signal that their SCFG was using led the authors to develop another SCFG that did not use structural information but only base composition to generate a RNA gene signal. This new SCFG produced results "remarkably similar" to the results obtained using the structural information as well [50]. In order to support their observation that the signal produced by structural information is greatly overshadowed by the signal produced by base composition, two experiments were devised.

First they tested whether the two structural algorithms (thermodynamic and structural SCFG) could distinguish a real RNA gene from a shuffled one. The results of this experiment showed that the base composition model retained its scoring shape after the shuffle, providing evidence that the base composition remained intact over the shuffling; however, the structural algorithms also retained their hits showing that the scoring of these structural models is mainly due to base composition [50]. It should also be noted that as the shuffled region was extended past the gene region "the shuffled sequence scores tend to smear out" as the base composition is changed [50]. The next test was to determine whether the secondary structure "contributes a significant component to the score" [50]. In this experiment they embedded a real RNA gene into a random sequence of the same base composition and attempted to locate the gene using the different gene finders [50]. The results showed that in most examples the embedded RNA genes were not found which indicates that the secondary structure signal is not strong enough to distinguish a real RNA from background signals of the same base composition [50]. It should be noted that after coming to the conclusion that MFE was not a statistically significant signal for RNA gene finding, Rivas' group shifted their focus away from ab-initio RNA gene finding and concentrated their efforts on homology RNA gene finding [52].

Workman and Krogh pointed out that since Zuker's algorithm for folding RNA and calculating minimum free energies relies on the stabilizing terms of stacked bases, only random RNA with the same dinucleotide frequency can be used to draw any valid conclusions when comparing MFEs [71]. Clote *et al.* used this fact to refute Rivas' group's claim that secondary structure is not a significant signal for RNA gene finding citing, that Rivas' group had performed mononucleotide shuffling in their experiments, making them invalid [14]. Clote demonstrated the significance of secondary structure in RNA finding with a new method which made use of RNAfold in the Vienna RNA Package to find the minimum free energy of a sequence and used a dinucleotide shuffler for generating the random population of sequences needed in calculating the average and standard deviation free energy values [14]. In addition, Clote modified the z-score formula in a novel way, creating what he called an "asymptotic z-score", having the property of providing "an asymptotic limit for the mean and standard deviation of minimum free energy per nucleotide for random RNA". In another improvement, Clote created a free energy asymptotic mean and standard deviation lookup table for a "complete set of dinucleotide frequencies"[3], which provided a significant speed-up when processing large genomes [14]. Even with the use of this lookup table the running time of their algorithm is $O(NL^2)$, where N is the size of the window and L is the length of the genome [14]. The paper reports that these asymptotic z-scores produced a "higher signal to noise ratio" than classical z-scores, but notes that the researchers are unclear of what causes this difference [14]. They tested the algorithm on several RNAs[4],

[3] Up to two decimal places [14].

[4] tRNAs, type III hammerhead ribozymes, SECIS sequences, srpRNAs, snRNAs, U1 small nucleotide particle, U2 small nucleotide particle, mRNA

showing that SRNAGs have "significantly lower folding energy than random RNA of the same dinucleotide frequency" [14].

Carter et al. took a more inclusive approach to RNA gene finding using base composition, transition frequency, and sequence motifs, in addition to MFE as signals for their RNA gene finder, RNAGENiE [9]. Basing their hypothesis on the idea that the evolutionary forces that create diversity in NC genome regions will also create a characteristically different distribution set for features of SRNAG regions, the researchers proposed searching for fRNAs based on a variety of possible characteristic signals [9]. From a high level standpoint the authors constructed their fRNA gene finder by using machine learning techniques to extract signal data from known fRNA gene regions and NC regions and applied the "learned rules" to predict the location of novel fRNAs in unannotated regions of the genome [9].

The training dataset was constructed by selecting genome regions of known fRNAs and NC regions, as the authors felt that using randomized RNA sequences for the negative training set would not represent the distribution of features in real genomes with enough accuracy [9]. Any duplicate rRNA and tRNA sequences were removed from the training set to control bias[5] and each sequence left in the training set was partitioned into windows of 80 nucleotides in length[6] overlapping by 40 nucleotides[7] [9]. After some initial tests this training set was altered because the database of NC regions was around 80 times larger than that of the RNA genes. This is a problem when training a neural network (NN), where the size of the positive and negative training sets should be similar, so the large NC training set was partitioned into several smaller sets having a size similar to that of the RNA gene training set [9]. This partitioning resulted in five datasets, which allowed five different NNs and SVMs to be trained independently, enabling cross checking between the various models for agreement of a decision, helping reduce the number of false positives [9].

The researchers trained both a NN and a SVM model on the feature set collected from the training data [9]. The back-propagation, feed forward type NNs had a single input for each feature in the dataset, a single hidden layer, and two output nodes (see Figure 2.2). One output node to indicate whether the given window's features are from a RNA gene region and the other to indicate if the features are predicted to be from a NC region [9].

[5] This left a training set consisting of 8400 fRNA nucleotides and 675322 NC nucleotides [9].

[6] This size was picked as it roughly corresponds to the size of tRNAs [9].

[7] Resulting in 7705 ncRNA windows and 188 unique fRNA windows consisting of 38.3% 23S rRNAs, 26.6% miscellaneous small RNAs, 20.2% 16S rRNAs, 13.3% tRNAs, and 1.6% 5S rRNAs [9].

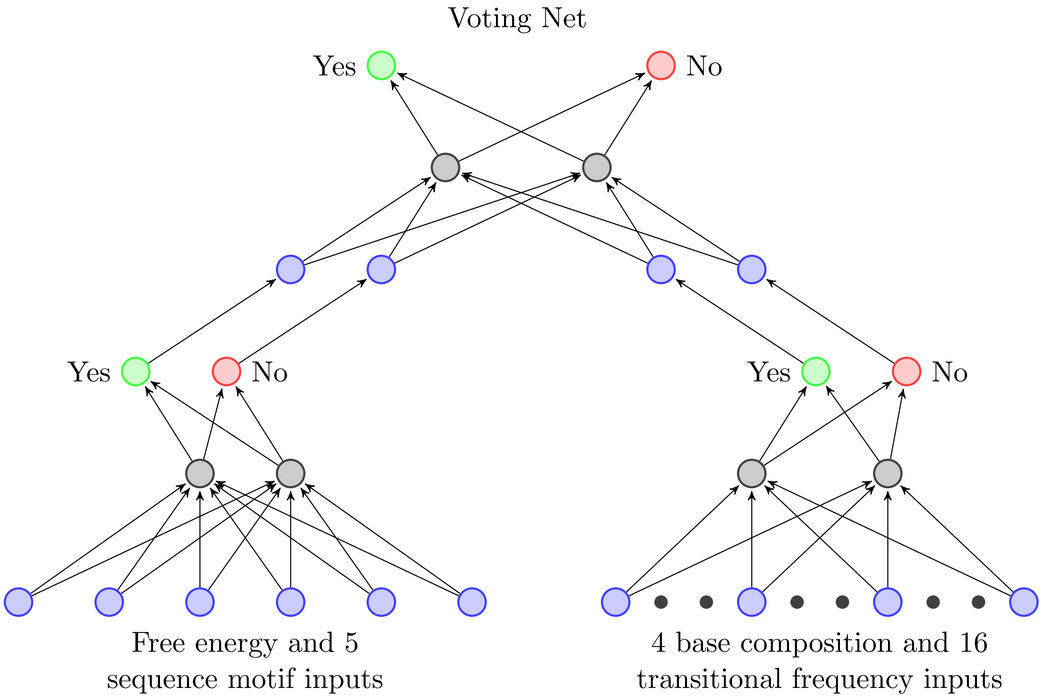

Figure 2.2: RNAGENiE's Neural Network. This figure is based on Figure 1 in [9].

Input for the SVM was simply the feature set data [9]. These input features included mononucleotide base composition[8], dinucleotide base composition[9], free energy of the folded window sequence, and the motifs: UNCG, GNRA, CUYG, AAR, CUAG[10] [9]. Although the computational complexity of the method is not reported in the paper, the major component of this complexity is most likely due to the free energy calculations for each window.

Using a jackknife testing procedure[11] the composition inputs on the neural network produced a prediction accuracy of greater than 85% and the motif inputs alone produced an accuracy of over 81% [9]. When all 20 inputs were tested by the jackknife procedure, the model resulted in a prediction accuracy over 92% [9]. These tests showed that not only the method was working, but also that both the motif inputs and the composition inputs were contributing to the prediction score [9]. Even with both groups of inputs contributing, the researchers found a correlation between the successfulness of their gene finder and the average GC content of the genome; however, as seen in a number of outliers, this correlation did not always hold true [9]. For the various bacteria genomes tested, the reported accuracy ranged between greater than 88% to just under 100% [9]. This accuracy is exceptional, especially considering the researchers tested some genomes with a ratio between structural genes GC content and genome background GC content of around 1:1, showing once again that RNAGENiE was not simply using base composition for RNA gene finding. The tests with the SVM models produced results which were similar but "somewhat less accurate" than the neural network results [9]. In addition to the success reported in the paper it has been shown that some of the hits that were taken to be false positives have been identified as RNA genes in unrelated laboratory experiments [44].

In Kirt Noël's thesis *Examining Stem-Loops as a Sequence Signal for Identifying Structural RNA Genes*, a RNA gene finding algorithm, named Wave, is developed which intended to rely on the statistical significance of various quantifiable properties of *stemloops* located in the genomic code to distinguish SRNAGs from the NC and protein coding regions [45, 44, 46]. The system used an ad hoc algorithm for scanning a genome and detecting sequence regions that could potentially form SRNAGs *stemloops* [44]. Since Noël noted that

[8] %A, %T, %G, and %C

[9] %AA, %AT, %AG, %AC, %TA, etc.

[10] N, any (A, T, G, or C); R, purine (G or A); Y, pyrimidine (C or U)

[11] Removing one training example, retraining the model, and testing the new model on the previously removed example, then repeating this process for each example in the training set, and averaging the prediction results [9].

stemloops are the most important structural element in SRNAGs, analogous to the protein secondary structure α-helices and β-sheets, he hypothesized that *stemloops* formed by the genomic sequence will tend to be longer and more frequent in the SRNAG coding regions [44]. Based on this hypothesis, Noël developed a SRNAG finding method using the size, shape, and relative location of the detected *stemloops* to compose a signal for SRNAGs [44].

Although the *stemloop* finding algorithm described in [44] has the potential to find all possible theoretical *stemloops* in a genome sequence, allowing for *bulges* and *internal loops*, Noël used it to report only the largest possible non-overlapping *stemloops* in the genome which did not violate any of the constraint parameters[12] [44]. When this algorithm was tested on the E. coli ssu rRNA gene, the algorithm only found 19 out of the 32 *stemloops* in the secondary structure computed using dynamic programming and of these 19 only 9 *stemloops* were in the correct location [44]. This poor accuracy is most likely caused by the *stemloop* finder both allowing for *internal loops* and *bulges* and finding only the largest *stemloops*. Unlike the previous methods discussed, this method did not make use of free energy as a metric and hence even with the *stemloop* finding step, the algorithm only has a computational complexity of $O(n^2)$ in the worst case scenario [44]. Noël reported that in practice the Wave algorithm had a running time of $O(n)$ determined through stopwatch runtime evaluation [44].

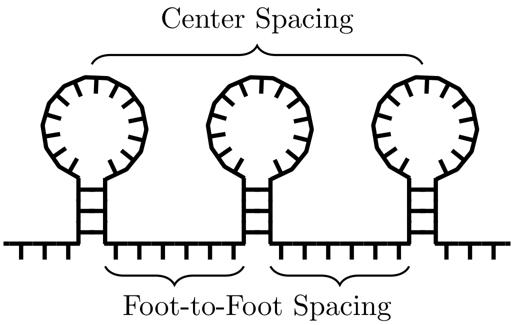

Figure 2.3: Center Spacing and Foot-to-Foot Spacing. This figure is based on Figure 4.2 and Figure 4.3 in [44].

Once found, the *stemloop* was processed to extract the metrics. These metrics included

[12] *Hairpin loop* size, minimum *hairpin loop* closure, maximum *bulge* size, maximum *internal loop* size, minimum *bulge* or *internal loop* closure, minimum GC base pair content, maximum GU base pair content, overall minimum number of base pairs [44]

span[13], center point spacing (CS)[14], foot to foot spacing (FS)[15], and number of base pairs (bps)[16] [44]. In addition to these atomic metrics Noël combined, through scaling, the CS and bps metrics and FS, and bps metrics to create two composite metrics [44].

These six metrics were then tested for their ability to distinguish tRNA and rRNA from coding and NC regions of genomes of differing background GC content. Noël found that the FS, span, and bps metric values at some GC content levels were not useful in distinguishing the SRNAG parts of the genome from the NC and coding parts. However, the CS metric showed some success by remaining distinguishable across the GC content spectrum, yet when the GC content between the genome and RNA genes is roughly equal, even this signal becomes very weak [44]. Likewise, the combined metrics showed little effect in amplifying the signals of their components [44]. Overall, the *stemloop* metrics showed success predicting RNA genes in genomes of high average AT content, but when applied to genomes of high GC content many false positives were reported [44]. These results indicate that the success of the gene finder relied heavily, although indirectly, on base composition for its RNA gene finding signal.

2.2 A Novel Method

Although many of these methods for *ab-initio* SRNAG finding produced good results in genomes of high AT content, most of them resulted in a high number of false positives in genomes of high GC content. A high number of false positives makes such methods impractical as research tools because with such uncertainty regarding any prediction, researchers can not trust the tool for their analysis and it is inefficient and costly for biologists to experimentally validate a large number of prediction claims when many of them are not actually RNA genes. Likewise all the ab initio SRNAG finders described in the previous section, except for Noël's method using stemloops and Rivas and Eddy's method using only base composition, have running times of at least $O(Lw^2)$ where L is the length of the genome and w is the window size.

This thesis sets the ground work in order to address both of these issues. The method

[13] The number of nucleotides which comprise the *stemloop*.

[14] The average distance from top of the loop on one *stemloop* to its neighbours (see Figure 2.3).

[15] The average distance from the bottom of the base of a *stemloop* to each of its neighbours (see Figure 2.3).

[16] The number of bond pairs found in the *stemloop* structure.

presented in this thesis explores many new metrics hopefully some of which can act as new signals for SRNAG regions and since this method focuses on the extraction of these metrics from specific structural elements the running time of the ab initio RNA gene finding is hoped to be addressed by removing the need to fold candidate sequences. So while this novel method is aimed at predicting SRNAG regions with high specificity, it is also an exploratory tool geared to lay the foundations for future SRNAG finders.

Although substantially different, Noël's thesis work inspired many aspects of this thesis. Noël's concept of using structural element analysis for signal metrics is expanded. Only *stemloops* were analyzed in Noël's thesis and this was done using Wave to locate the RNA genes. As noted previously, although Wave had a low running time, the algorithm did not have a high accuracy in locating *stemloops* in the annotated genomes tested, which may have affected the ability for the *stemloops* to act as a RNA gene signal. In this thesis, *stemloops* will be evaluated again for their ability to act as a signal for RNA gene discovery; however unlike Noël's work, a number of other structural elements (*structure*, *hairpins*, *bulges*, *multiloops*, *stems*, *stacks*, *external loops*, *internal loops*, *tails*, *joints*, *unpaired*, *junctions*, *joint-tails*, *bridges*, and *loops*) will be tested as well. To rectify the problem of Wave not accurately being able to find *stemloops* and possibly influencing the results of Noël's experiments, a scheme of folding consecutive windows will be used to guarantee that the metrics analyzed reflect the theoretical RNA gene secondary structure as closely as currently possible. It should be clear that although a folding window scheme is used in this thesis, the purpose is only to be able to easily and accurately locate the secondary structure elements for analysis. Once the secondary structure elements' ability as a gene finding signal has been verified, algorithms can be created, similar to Wave, which can attempt to locate the structural element in the genome without the need for folding a window. Moving in this direction, where individual elements are located instead of windows folded, is beneficial for two main reasons. First, sliding window schemes are unable to adapt to RNA genes of different sizes, while structural element analysis is not limited to selecting specific sized windows, allowing for a high diversity of genes and their lengths to be discovered. Second, as previously discussed, doing metric analysis from a folded sliding window has at best a running time of $O(L * w^2)$, where w is the window size, and L is the length of the window, while it is possible for individual structural elements to be detected with enough accuracy in a running time lower than this, as demonstrated in the $O(n^2)$ running time of Wave. Although beneficial, the development of Wave like algorithms for locating secondary structure elements in genomes is not pursued in this thesis due to time and efficiency[17] constraints.

While Noël's work inspired the use of structural elements for signal data, the classification method used by Carter's group in RNAGENiE inspired the use of machine learning for metric analysis. Carter's use of NNs and SVMs allowed for the effective use of multiple metrics in the classification of RNA genes; likewise, this thesis uses SVMs to analyze multiple metrics. While RNAGENiE used motif metrics, composition metrics, and free energy for classification, the method presented in this thesis does not utilize any motif metrics, but does make use of composition and free energy along with many other metrics not used by RNAGENiE. The primary reason motif data was not used is simply that the focus of this project is on the use of structural elements for RNA gene finding, but future versions of the SRNAG finder could use them as an additional metric in conjunction to the metrics chosen for this thesis. Carter found the NNs performed better than SVMs for RNA gene finding, however, SVMs are used exclusively in this project for classification for three reasons. First, SVMs always find a global optimal model due to the geometric, numerical methods used, while NNs trained with back-propagation usually only converge to a local optimal solution [8, 53]. Second, the classification feature set in this project is large and unlike NNs where in order to control model complexity the feature set size needs to be small, SVMs automatically select their model size through the selection of support vectors [53]. Third, the development of SVMs has been based on sound theory with experimentation to verify their effectiveness, while NN development followed a more "heuristic" path from experimentation and testing [63].

Likewise, since the work of Chen, Le, and Clote demonstrate the importance of MFE, dinucleotide composition, and mononucleotide composition to RNA gene finding, such metrics are included in the feature set of analyzed properties of the purposed method.

In brief, this thesis presents a method which trains a classification system based on SVMs for the detection of SRNAGs. A database composed positive SRNAG sequence examples and negative dinucleotide shuffled RNA sequence examples is used to train a sophisticated voting network of SVMs, which attempts to classify metrics derived from many different RNA secondary structure elements. These secondary structure elements are collected by first folding the RNA sequence and then parsing out the individual structures and substructures. Then many properties of these structural elements are measured and recorded. This metric

[17] There is no need to develop algorithms to find some of the tested structural elements that do not play a significant role SRNAG detection.

data is passed to the support vector machines to train a classification engine. Once the classifiers in the voting network are trained, this classification system is applied to classify the metric data collected though a similar process as the training process. A much more detailed explanation of the methodology used is provided in Chapter 3 and Chapter 4.

2.3 Chapter Review

This chapter described several important ab-initio RNA gene finding methods, detailing the drawbacks and benefits of each method. Through [31, 12, 50, 14], it has been seen that in order to be a useful and statistically sound RNA detection signal, the free energy z-scores must be calculated against populations generated with dinucleotide shuffling. Particularly [50, 9, 44] noted the importance of base composition in RNA gene regions and the need to test RNA gene finding methods either against a simple base composition signal or in genomes where the difference between the background GC content and the GC content of the RNA genes is low. Out of all the RNA gene finding methods discussed RNAGENiE appeared to be the most successful method, predicting many classes of RNA genes with high accuracy and a relatively low false positive rate. Finally, a brief description of the RNA gene discovery method introduced in this thesis was given, contrasting it to methods described in this chapter.

Chapter 3

Building a SRNAG Classifier

At the heart of the RNA gene discovery method proposed in this thesis is the extraction of the RNA structural features to be used as a signal to distinguish RNA gene regions. There are three main steps involved in this extraction process: folding the RNA sequence into its secondary structure, parsing out the individual structural elements, and extracting the structural features by measuring specific properties of the RNA structural element. To utilize the structural element metrics in SRNAG finding, a model is built, which will attempt to label the metrics as coming from a SRNAG or not. These models will be both built and tested using the extracted metrics, the process of which is explained in detail in Chapter 4. This chapter deals with the three steps of the metric extraction process and explains the steps needed to use SVMs to classify RNA gene sequences.

3.1 Folding

In order to extract the structural features from RNA, the RNA sequences need to be folded into their MFE secondary structures. As stated before, secondary structures are used in this project because efficient algorithms exist for determining these structures using dynamic programming, allowing for the many sequence windows of a genome to be folded in a reasonable amount of time. The RNAfold software from the Vienna RNA Package[1] is utilized to conduct this folding process.

Treating the folding process as a black box, the system works as shown in Figure 3.1.

[1] `http://www.tbi.univie.ac.at/RNA/`

First the RNA sequences are piped into the RNAfold software in *Multiple Fasta* format, with each sequence header having a unique ID for tracking it through the rest of the gene finding system. RNAfold uses its dynamic programming algorithm to fold the RNA sequence into its secondary structure, which is then written to a file in *Multiple Fasta Dot-Bracket* format. This file format contains all the sequence header information and the same nucleotide sequence that the input Fasta file had, and also includes the structure of the sequence expressed in dot bracket form. Dot bracket notation was introduced by Hofacker et al. and uses matching parenthesis to indicate bonds and dots to denote free bases [26].

Known native SRNAG structures could be used for the classifier training process instead of folded RNA sequences. Native structures are easily accessible and could be handled in much the same way as the folded structures. Likewise, they would allow for more realistic secondary structure metric data to be collected. However, native structures can only be used in the collection of training data and folding would still need to be used for the testing data collection process. This difference in the method for data collection could lead to different data distributions between the training and testing sets, causing a SVM trained on the training set to perform poorly on the tested set. So for consistency in the data collection process folding is used to generate the secondary structures in both the training and testing phases ensuring the distributions learned by the SVM will be exploitable by the test dataset.

3.2 Parsing

As discussed in the previous section, the folding process produces a structure in dot bracket notation aligned with the nucleotide sequence. In order to locate the structural elements in the structure, a recursive function is used with an execution path that mimics a RNA secondary structure tree graph (see Figure 3.2). Its most basic inception, seen in Algorithm 1, moves from the beginning of the structure sequence to the end, recursing each time an open parenthesis is found and backtracks from a current recursion when a closing parenthesis is found. In line 5 when a recursive call ends, it updates the current symbol pointer, i, to the next symbol location, allowing the algorithm to handle *external loop* and *multiloops*. Unpaired or '.' symbols are processed but ignored and the algorithm exits when the sequence's

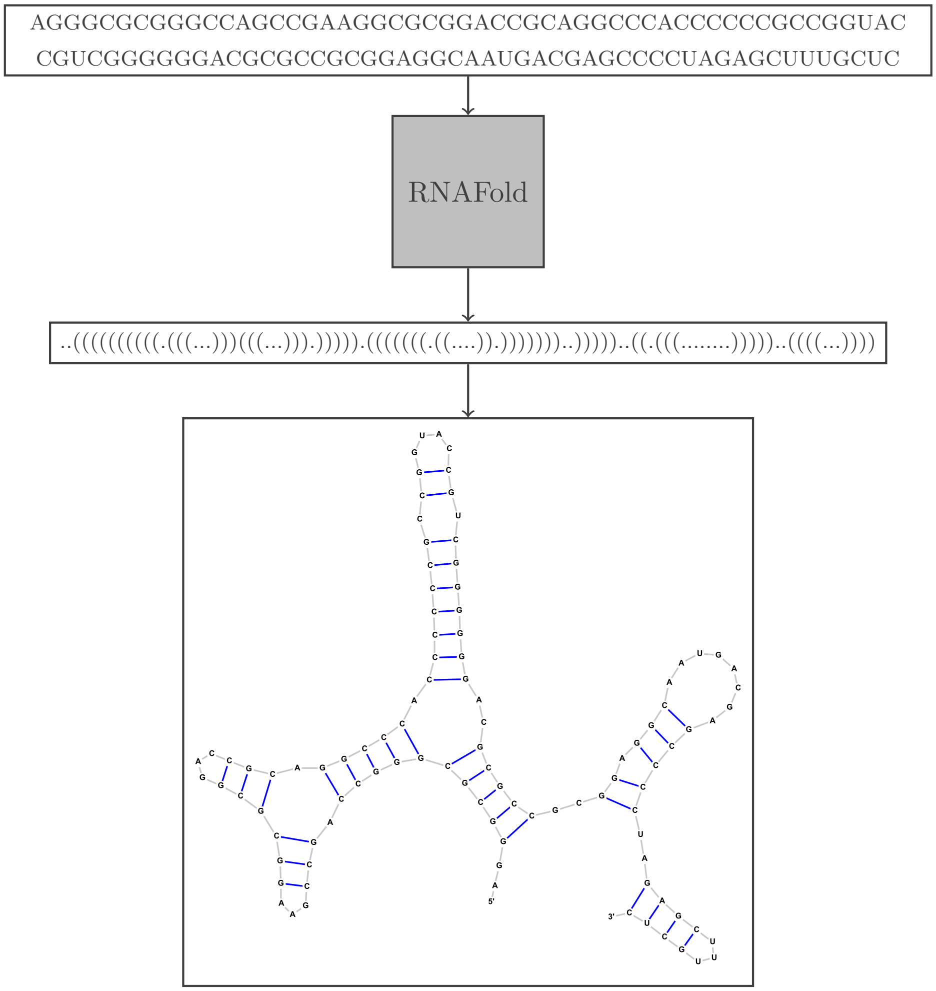

Figure 3.1: Folding Process. This figure shows a nucleotide sequence being fed into RNAfold which will fold the sequence producing a dot bracket representation of the nucleotide sequence's structure. The secondary structure is shown at the end of the process.

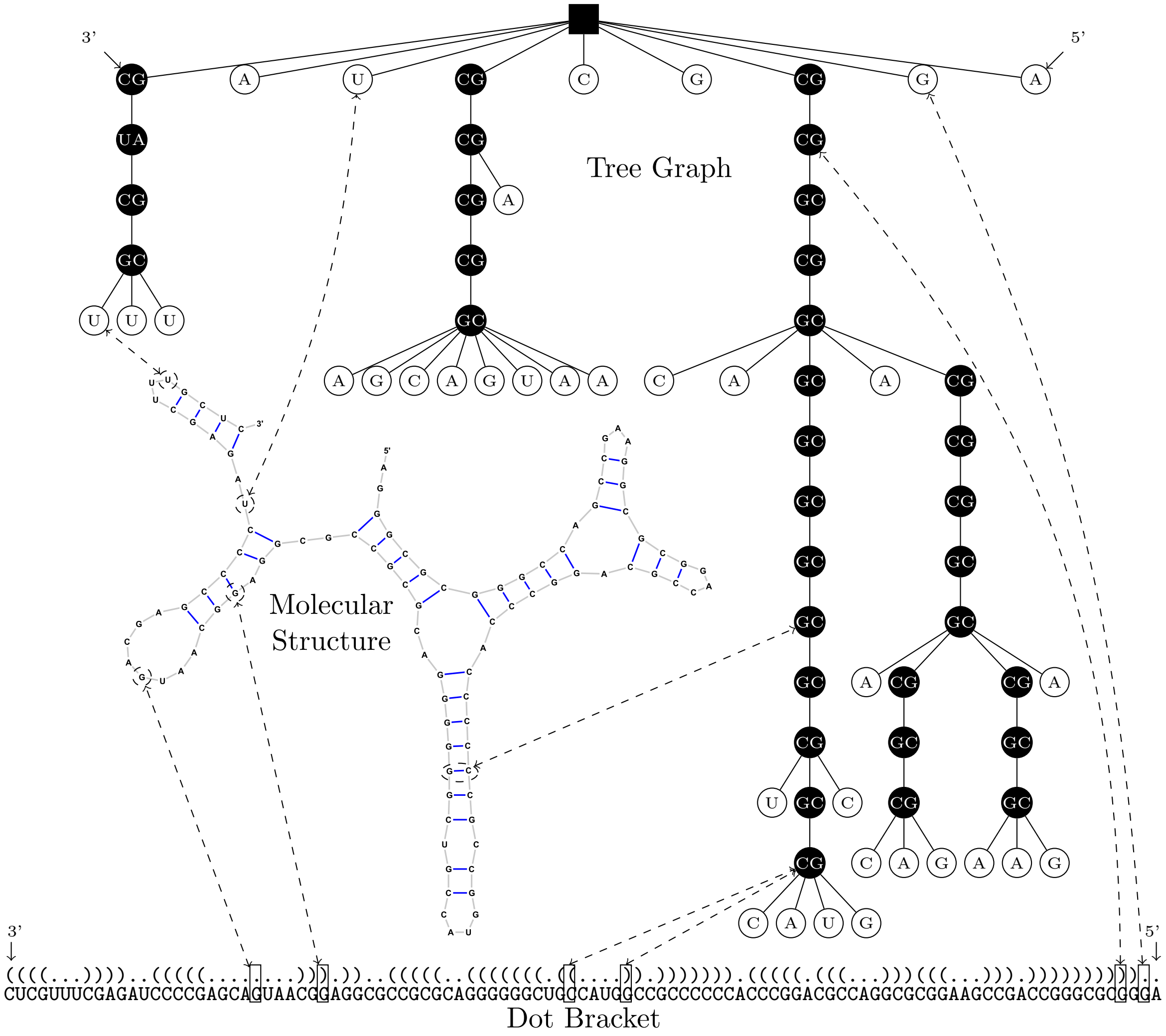

Figure 3.2: Secondary Structure Representation Relationship. Shows the relationship among the dot bracket notion, the molecular structure, and the tree graph representation of the RNA secondary structure. The dashed lines are for reference only, showing the mapping between certain points of one representation to another. It should be noted that the nucleotide and structure sequence are printed in opposite direction from convention, to ease the mapping of the different representations. The tree graph representation is similar to the tree in Figure 1 in [56].

null terminal is found.[2]

Algorithm 1 $ParseTree(i)$

```
Require: $structure$ to be a $null$ terminated string with matching parenthesis.
while $structure[i] \neq null$ do
  if $structure[i] =$ ‘.’ then
    $i = i + 1$
  else if $structure[i] =$ ‘(’ then
    $i = ParseTree(i + 1)$
  else if $structure[i] =$ ‘)’ then
    return $i + 1$
return $i$
```

Once a tree graph representation of the secondary structure is created, certain signatures, based on the order and number of paired and unpaired child nodes, allow each node to be labeled as a specific structural element. Figure 3.3 shows several of these signatures. A *hairpin loop* is always found at the end of a branch and is characterized by a node with only unpaired child nodes. *Stacks* on the other hand always have exactly one child, which is paired. The signature for *internal loops* and *bulges* is similar to the one for *stacks*, except that in addition to a lone paired child, *internal loops* have unpaired child nodes on both sides of the paired node, while *bulges* have unpaired child nodes on either side of the paired node, but not both sides. *Multiloops* are characterized by having more than one paired child node; between or to the left or right of these child nodes could exist any number of unpaired child nodes. Although an *external loop* typically has similar characteristics to those of a *multiloop*, its only requirement is that it contain the two ends of the RNA sequence and at least one paired node. If the *external loop* does not contain any paired child then the sequence given to the parser is completely unpaired. The *external loop* will aways be the root node of the tree. Algorithm 2 takes a list of child nodes and labels the parent node according to the child node signatures described. The function in Algorithm 3 also labels a node, but this function is only called on the root node, and hence outputs either an *unpaired* or *external-loop* label.

Algorithm 4 is Algorithm 1 with the labeling functions incorporated. Compared to Algorithm 1, the labeling algorithm collects information about the sibling nodes, particularly the location of unpaired and paired nodes and passes them to Algorithm 2, which labels the

[2] Assuming that every open parenthesis is matched with a closing one.

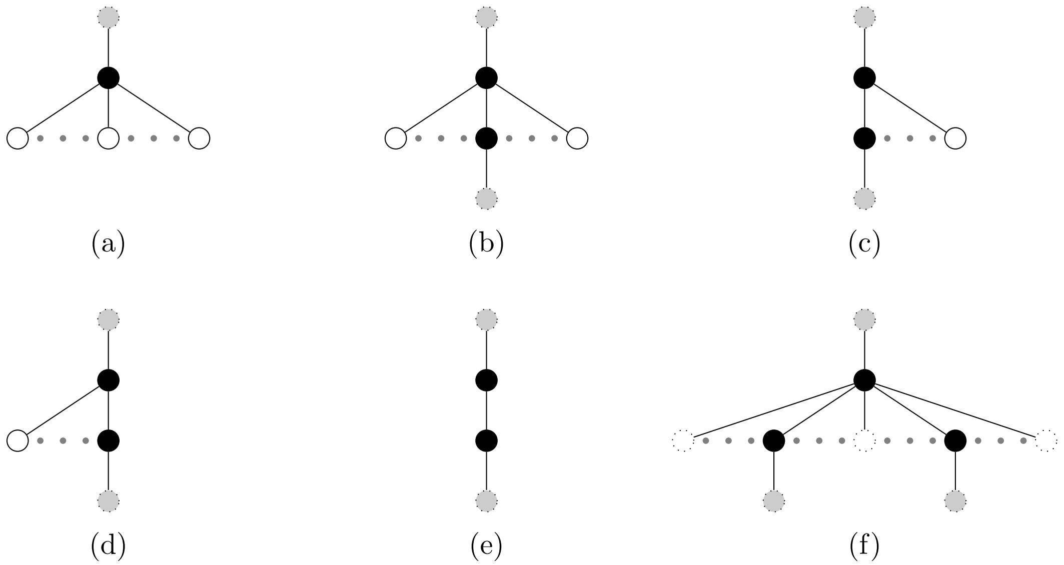

Figure 3.3: Basic Parse Signatures. The graph structure signatures shown are (a) *Hairpin*, (b) *Internal Loop*, (c) and (d) *Bulge*, (e) *Stack*, and (f) *Multiloop*. Gray nodes with a dotted outline indicate continuing tree nodes and white nodes with dotted lines indicate unpaired nodes which may or may not exist. The ellipses indicate that there could be more child nodes.

Algorithm 2 *BasicLabel*(*siblings*)

```
Require: siblings to be a list populated with unpaired and paired symbols.
if $siblings.count(paired) = 0$ then
    return hairpin
else if $siblings.count(paired) = 1$ then
    if $siblings.first = unpaired$ and $siblings.last = unpaired$ then
        return internal-loop
    else if $siblings.first = unpaired$ then
        return bulge
    else if $siblings.last = unpaired$ then
        return bulge
    else
        return stem
else
    return multiloop
```

Algorithm 3 *BasicLabelRoot*(*siblings*)

```
Require: $siblings$ to be a list populated with $unpaired$ and $paired$ symbols.
if $siblings.count(paired) = 0$ then
    return $unpaired$
else
    return $external\text{-}loop$
```

Algorithm 4 *ParseTreeBasicLabel*(*i*)

```
Require: $structure$ to be a $null$ terminated string.
$siblings \leftarrow \emptyset$
while $structure[i] \neq null$ do
    if $structure[i] =$ '.' then
        $siblings.append(unpaired)$
        $i = i + 1$
    else if $structure[i] =$ '(' then
        $siblings.append(paired)$
        $i = ParseTreeBasicLabel(i + 1)$
    else if $structure[i] =$ ')' then
        $BasicLabel(siblings)$
        return $i + 1$
$BasicLabelRoot(siblings)$
return $i$
```

node before moving back up the branch. Since for any input structure that has no unmatched parentheses Algorithm 4 will only ever hit line 12 and 13 after the *null* terminating symbol is found, Algorithm 3 will be run once for the root node. After running Algorithm 4 on a structure sequence, each node in the tree will be labeled with a localized label, which is a label pertaining only to the node and its children; however for the larger, multinode structures, like *stemloops*, an extension will need to be made.

Using the higher level labeling function found in Algorithm 5 in addition to the two basic labeling functions (Algorithms 2 and 3), multinode secondary structure elements can be located and labeled. In order to label larger sections of the tree it is necessary to know what precedes in the tree. For example, if the current node is known to be a *stack*, it is possible that the *stack* could belong to either a *bridge* or a *stemloop*, however the only way to determine this would be to check if the branch leads to a *multiloop* or *hairpin*. So in order to perform this higher level labeling, as the recursive procedure backtracks up the parse tree, it passes messages up the tree. For example, when a *hairpin loop* is reached at the bottom of a branch, that node is labeled as being part of a *stemloop*, and the *stemloop* message is passed to its parent, which again is labeled as a *stemloop*. This passing up the tree and labeling continues until a *multiloop* is reached, at which point the message is changed to *bridge*, and the parent node of the *multiloop* is labeled as such. This process continues until the whole structure is annotated.

Algorithm 5 $CompoundLabel(current, child)$

```
if $current = hairpin$ then
  return $stemloop$
else if $current \in \{internal\text{-}loop, bulge, stem\}$ then
  if $child = stemloop$ then
    return $stemloop$
  else
    return $bridge$
else
  return $none$
```

It should be noted that this section has not touched on locating every structural component as many of the components are easily determined from the structural elements already discussed. *Joints*, for example, are any group of unpaired child nodes in a *multiloop*, and *tails* are any group of unpaired child nodes descending from the root node on the outermost

Algorithm 6 $ParseTreeCompoundLabel(i)$

```
Require: $structure$ to be a $null$ terminated string.
$siblings \leftarrow \emptyset$
$childtype \leftarrow none$
while $structure[i] \neq null$ do
    if $structure[i] =$ '.' then
        $siblings.append(unpaired)$
        $i = i + 1$
    else if $structure[i] =$ '(' then
        $siblings.append(paired)$
        $(i, childtype) = ParseTreeBasicLabel(i + 1)$
    else if $structure[i] =$ ')' then
        $BasicLabel(siblings)$
        $CompoundLabel(BasicLabel(siblings), childtype)$
        return $i + 1, CompoundLabel(BasicLabel(siblings), childtype)$
$BasicLabelRoot(siblings)$
return $(i, none)$
```

sides of any paired child nodes.[3] Likewise, *stems* are directly related to *bridges*, and can be located in *stemloops* simply by breaking off the *hairpin loop* component. For *joints*, *tails*, *hairpin loops*, *bulges*, and *internal loops* it is trivial to parse the component into its *stack* and *unpaired* sub-components. It should be clear that there is no need to locate any aggregate structural elements, as the component metric data can be aggregated directly from each structural element in the aggregate. The data aggregation process will be discussed in more detail in Section 3.4. Finally it should be mentioned that the final parsing algorithm (Algorithm 6) looks at each dot bracket symbol once as it works its way through the structure sequence, so it is plainly evident that the parsing algorithm runs in $O(n)$ time, where n is the length of the structure sequence. Likewise the space complexity is also linear, as even if one was to build the tree in memory, each node would correspond to no more than half the symbols in the structure sequence since each opening parenthesis needs to be paired with a closing parenthesis to form a node.

[3] The unpaired nodes of the *external loop* between the paired children, are *joints*.

3.3 Feature Extraction

Once the structural elements have been labeled, each element is measured in a variety of ways to create a feature set, which is written to the disk in plain text table format. Algorithmically most of the features are simple to extract from the excised nucleotide and structural sequence that makes up the structural element. Only certain metrics are extracted from each structural element, as in many cases the measurement does not logically apply to an element of that type. Table 3.1 lists all the metrics and shows which structural elements they apply to. It is important to note that some of the features listed in Table 3.1 are actually a group of several related features. It should be made clear that there is no strong rationale behind including many of the features that are going to be explored, as it can often be hard to figure out which features are going to produce a strong RNA gene signal without trying them first. So a kitchen sink approach is taken, where as many different metrics are tested as possible within practical limits. The rest of this section goes through each feature listed, describing it in full and detailing the extraction method where applicable.

3.3.1 Size

The feature *size* can be extracted from any of the structural elements and is simply a measure of the number of nucleotides that make up the structure.

3.3.2 Linguistic Complexity

Linguistic complexity (LC) is defined as “the ratio of the number of [substrings] present in the string of interest to the maximum number of [substrings] for a string of the same length over the same alphabet” [60]. More simply, LC is a measure of the frequency of repeated segments in a string; the more repeats that are present in a sequence the lower the LC. LC applies to two metrics: (1) Nucleotide Sequence Linguistic Complexity (NLC) and (2) Structure Sequence Linguistic Complexity (SLC), where the first is the LC of the nucleotide sequence and the second is the LC of the dot bracket structure sequence. Since all structural elements have both a nucleotide sequence and a structure sequence, this metric is applied to every structural element. LC can be calculated in $O(n)$ time and $O(n)$ space, implemented using compressed suffix trees [60], so LC will not present a processing or memory bottleneck for this project.

Feature / **Structure**	Structure	Stemloop	Stem	Bridge	Stack	Loop	Internal Loop	Bulge	Hairpin Loop	Junction	Multiloop	External Loop	Joint-Tail	Joint	Tail	Unpaired
Size	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
NLC	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SLC	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mononucleotide Composition	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Dinucleotide Composition	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CS		✓			✓	✓	✓	✓	✓				✓	✓	✓	✓
FS		✓			✓	✓	✓	✓	✓				✓	✓	✓	✓
PP	✓	✓	✓	✓						✓	✓	✓				
Bond Composition	✓	✓	✓	✓	✓					✓	✓	✓			✓	
Sides Ratio							✓									
Gibbs Free Energy	✓															
Average Sides Ratio		✓	✓	✓												
Average Size		✓	✓	✓						✓	✓	✓				
Sub-element Composition	✓	✓	✓	✓						✓	✓	✓				
Number Sub-element Composition	✓															
Avg. CS		✓	✓	✓						✓	✓	✓				
Avg. FS		✓	✓	✓						✓	✓	✓				
Avg. Type NLC		✓	✓	✓						✓	✓	✓				
Avg. Type SLC		✓	✓	✓						✓	✓	✓				
Avg. Mononucleotide Composition		✓	✓	✓						✓	✓	✓				
Avg. Dinucleotide Composition		✓	✓	✓						✓	✓	✓				
Avg. Bond Composition		✓	✓	✓						✓	✓	✓				

Table 3.1: Structural Element Features. This table shows which features are extracted from each structural element.

3.3.3 Nucleotide Composition

Composition, like LC, is really two groups of features: (1) mononucleotide composition and (2) dinucleotide composition. Appendix C defines mononucleotide and dinucleotide composition. Their extraction is as simple as counting the number of occurrences of each nucleotide or combinations of two nucleotides in the sequence. Since the composition values are reported as frequencies each composition count is divided by the sum of all the categories for the given group. This means mononucleotide composition has a total of four metrics (A%, C%, G%, U%), one for each nucleotide, and dinucleotide composition has a total of sixteen metrics (AA%, AC%, AG%, AU%, CA%, CC%, CG%, CU%, GA%, GC%, GG%, GU%, UA%, UC%, UG%, UU%), one for every combination of each nucleotide, totaling to twenty metrics in the composition group. Again, since every structural element has a nucleotide sequence the composition group of metrics is applied to every structural element.

3.3.4 Spacing

Spacing is the combination of Center Spacing (CS) and Foot-to-Foot Spacing (FS). Both these metrics have been described for the *stemloop* element in Chapter 2 and the same general principles apply to the other structural elements. When applied to a single stranded structural element[4] the properties are exactly the same as were described for *stemloops* in Chapter 2. To reiterate, CS is the average distance between the center of the given structure to the centers of the two adjacent structures of the same kind, while FS is the average distance between the edge of the structure to the nearest edge of the next structure moving outwards. For double stranded structural elements,[5] CS distance is the average distance between the center of the element and the centers of the two adjacent elements of the same type moving up or down the bound strands of the *stem*, while FS for double stranded structures is the average distance between the two closest edges of the elements of the same type above and below the element in question. This feature is extracted from every single stranded structural element, except for the *structure* element as there are no adjacent *structure* elements to compute the distance. Likewise, it extracted from every double stranded element which appears in a *stem*.

[4] *Joint-Tail*, *joint*, *tail*, *hairpin Loop*, *stemloop*, and *unpaired*.

[5] *Bulge*, *Internal Loop*, *Loop*, and *Stack*

3.3.5 Percent Paired

Percent Paired (PP) is extracted by dividing the number of '(' symbols in the structure sequence by the total length of the sequence to give the percentage of nucleotides in the structural element which are paired. This feature is only applied to structural elements which many contain any number of paired and unpaired nucleotides (*stem*, *stemloop*, *bridge*, *external loop*, *multiloop*, *junction*, *structure*).

3.3.6 Bond Composition

The frequency of GC, AU, and GU bonds appearing in the structural element is given by the bond composition metric group (GC% Bond, AU% Bond, GU% Bond). These three metrics are easily calculated from the nucleotide and structure sequences, as every time a bond is found in the structure sequence the two nucleotides which make up the bond are counted. These counts are totaled and the total number of each bond type is divided by the total number of bonds in the structural element. Each structural element that includes paired nucleotides will record these percentages, which means that only the *unpaired*, *tail*, *joint*, and *joint-tail* are excluded from extracting this metric.

3.3.7 Sides Ratio

Side Ratio (SR) only applies to the *internal loop* structural element and is a measure of the ratio of the length of one side of the *internal loop* versus the length of the other side. It should be noted that this metric is such that the smaller side is divided by the larger side and so it is independent of whether the larger side is on the left or right.

3.3.8 Gibbs Free Energy

Through the nucleotide sequence folding process that is needed to produce the structure sequences, the free energy of the structure is calculated. This metric, therefore, applies only to the *structure* element.

3.3.9 Average Side Ratio

Within any *stem* structure or a structure that contains a *stem*, multiple *internal loops* can exist. The metric Average Side Ratio, is simply a measure of the average Side Ratio (see

Section 3.3.7) of all the *internal loops* within a given *stem*. It should be noted that this metric is a special case of the aggregate metrics.

3.3.10 Sub-element Composition

Sub-element composition (% [TYPE]) is simply a measure of the percentage of nucleotides each substructure type contributes to a structural element. There is one value reported for each sub-structural element in the composite element. For example, the *stem* feature set would contain four metric values, one for each of the sub-elements: *stacks*, *internal loops*, *bulges*, and *loops*. Likewise, for *stemloops* with their additional *hairpin loop*, five metrics are used.

3.3.11 Number Sub-element Composition

Similar to Sub-element Composition, the Number Sub-element Composition (% Num. [Type]) is the percentage of the number of each sub-element type that occurs in the *structure*. Only the *structure* element uses this metric. Values are reported for every sub-element, meaning there is a value for each and every structural element other than the *structure*—a total of 15 metric values.

3.3.12 Aggregate Metrics

Several metrics are aggregate metrics, meaning that they attempt to provide a summary of several sub-elements' metric values with a single value. Each type of sub-element within a structural element is aggregated separately from the other types. Take a *stem* for example; a *stem* can contain any number of *stacks*, *internal loops*, *bulges*, and *loops*. For each of these sub-structure groups an aggregate value for the given metric will be calculated. Meaning that a *stem* will end up with four aggregate values: one for each structural element. Aggregate features only apply to structural elements which contain any number of a structural sub-element type. *Stemloops*, *stems*, and *bridges* produce aggregate values over the *stack*, *internal loop*, *bulge*, and *loop* structural elements, while *junctions*, *external Loops*, and *multiloops* produce their aggregates over *joint*, *tail*, and *joint-tail* structural elements. Originally, in addition to averaging the aggregates, the minimum and maximum values were also reported; however too much data was produced[6] making the processing time impractical with the inclusion of these two latter metrics, so they were dropped. Aggregate metrics include: average size, average CS, average FS, average NLC, average SLC, average mononucleotide composition, average dinucleotide composition, average bond composition, and average SR. All of these metrics are simply averages of the metrics to which their names refer for each individual sub-element type.

It is important to note that two types of aggregation occur. Value aggregation, previously described, is the process of proving a summary metric value for a group of sub-elements. Depicted in Section 3.4, data aggregation attempts to create a larger training and prediction dataset by grouping the data records of related structural elements into a single dataset.

3.4 Classification Engine

As mentioned in Chapters 1 and 2, SVMs are used exclusively in this project for classification. This section will detail the classification engine while Chapter 4 will detail the training and testing of the classification models. Specifically LIBSVM[7] [10] is the SVM software used in this thesis.

As discussed in the previous section, after the metrics are extracted from a structural element they are stored in a dataset file. The second type of aggregation happens in this stage; that is, aggregation of multiple groups of records. Section 3.3.12 detailed metric aggregation, where the metric of sub-elements are averaged. In contrast, the aggregation at this stage involves taking records from two related groups of elements and merging them into a new group. For example, it is easy to see that both *bulges* and *internal loops* are related, as they are simply an unbound region within a *stem*. The *loop* dataset is an aggregate of those two structural elements, meaning that the *loop* dataset contains both the *bulge* and *internal loop* feature records.[8] Data aggregation will allow for the detection of trends which might not appear when the specific structural elements are evaluated independently and if it happens that an aggregate structure is as useful as the individual structural elements,

[6] By including minimum and maximum aggregates between six and eight additional metric values are produced for each aggregate metric.

[7] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

[8] Only the metrics that appear in both the structural elements' data are in the combined dataset.

it will simplify the problem of locating structural elements for RNA gene finding as only one structural element needs to be located instead of two. Figure 1.5 graphically shows which structural elements are aggregates through the *is-a* relationships. To summarize the hierarchy, *junction* is a data aggregate of *multiloop* and *external loop* and *joint-tail* is an aggregate of *joint* and *tail*. Since a Branch is a specialization of a *stem*, the *stem* group can be considered an aggregate of the *stems* found in *stemloops* and those which make up *bridges*.

The data that is produced after the feature extraction process is raw metric data in table format. In order to use the data with the LIBSVM software, the data needs to be converted into the LIBSVM file format, where each record is composed of a class label, followed by a feature index[9] followed by the metric value, separated by a colon. The data is then normalized by scaling it between 0 and 1 using a tool provided with the LIBSVM software.[10]

With the datasets for each structural element separate, each of the structural elements can be tested individually for its own ability to classify prospective gene sequences. However, it is expected that by using more than one structural element to classify a sequence the classification accuracy will be improved. To facilitate this combining of the individual models, a voting system is used, where every model being used for the classification will cast a weighted vote for what it considers the correct class of the sequence.

In addition to combining different models, when it comes to classifying sequences, often there will be multiple structural elements in the same prospective secondary structure. For example, a SRNAG secondary structure might have five *stemloops* and two *bridges*. Each *stemloop* found will contribute to the classification process, as increasing the sample size should give a better overall prediction accuracy. The way all the elements of the same type are used is again through a system of voting. The voting system takes its inspiration from probability, however it needs to be made clear that the result of the voting scheme is not a true probability as it is known that the probability values used in the voting are not independent. In this voting scheme each of the individual structural elements vote on whether the elements have come from a SRNAG or a non-SRNAG sequence. The vote for

[9] The feature index allows for LIBSVM to handle sparse datasets, although this feature is not used in this thesis project.

[10] When applied to non-training data, it is possible that the scaling factor may scale the data below 0 or above 1, as the data value may fall outside of the range of the training dataset.

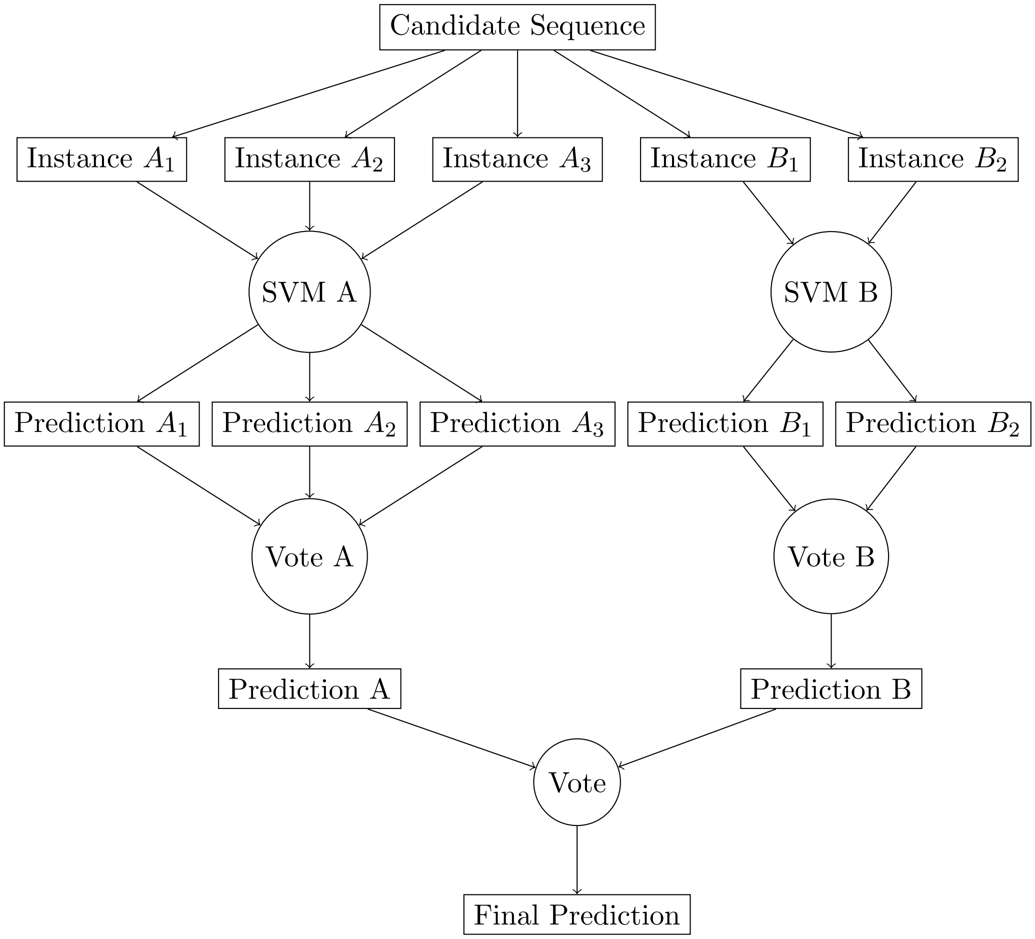

Figure 3.4: Voting Scheme. The voting scheme is broken down into two layers. On the first layer instances for each classifier type are processed by their respective classifier, which predicts a label for each instance and calculates a probability value that the label is correct. These probability predictions are voted together using Equation 3.1 to produce one prediction for each classifier. In the second layer both these predictions are voted together using Equation 3.1 to produce an overall prediction for the sequence which produced the secondary structure from which the instances were collected. In the diagram, A and B represent different structural elements, such as *stemloop* and *bridge*. Likewise, while only two structural element types are shown in the figure, more structural elements could be included in the voting system.

a given label is calculated by multiplying all the decision probabilities which agree with the label and one minus the decision probability, for all the decisions which do not agree with label. Equation 3.1 details this calculation mathematically, where S represents the set of all decision labels, $v(L = X)$ is the outcome of the vote for label X, and $c(L = X)$ is the certainty value produced by the SVM for label X of the element represented by L.

$$v(L = X) = \prod_{L \in S} c(L = X) * \prod_{L \in S} (1 - c(L \neq X)) \tag{3.1}$$

The label with the highest vote value becomes the predicted label for the structural elements of that type.

As already mentioned, once an element model has produced a vote these models will need to vote on the overall label of the sequence in question. This higher level voting is done again using Equation 3.1, where each structural element outputs the calculated vote and a label, and the votes for the same label are multiplied together while votes for a different label are subtracted from one and then multiplied. The label with the highest vote is the predicted class for the sequence being classified by the classification engine. Figure 3.4 illustrates this voting process.

The voting system for this SRNAG finder needed to be flexible to allow for different configurations of the classifier. The voting system previous described allows new configurations or groups of structural elements to be tested without the need for retraining of the voting system or reclassification of the individual structural elements. This lack of retraining and reclassification allows for many voting experiments to be attempted. Specifically, tests with all the combinations of two and three structural element groups would not have been possible if it was a complex and time consuming process to reconfigure the voting system. Likewise this voting system design proved to be quick to develop, as it is not much more complicated than accumulating the number of votes for each class.

3.5 Candidate Sequence Example

In order to flesh out the details of this SRNAG finding system further, an example candidate sequence will be tracked through the system. It will be assumed that the structural element SVMs are already trained and the goal is to label a sequence collected from a genome. The candidate sequence is shown below:

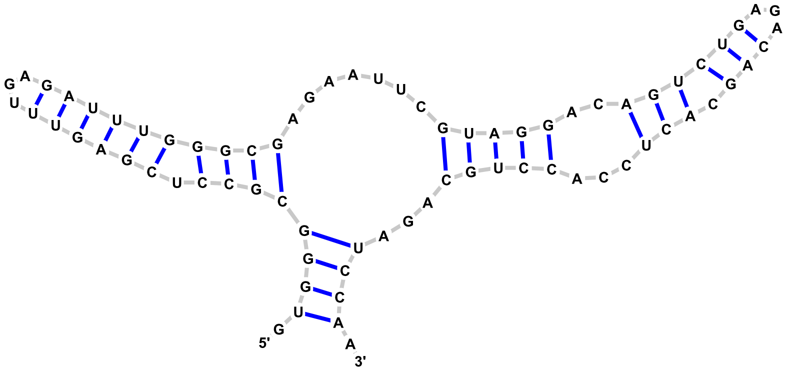

Figure 3.5: Example Secondary Structure. This figure shows the secondary structure of the example sequence.

```
GUGGGCGCCUCGAGUUUGAGAUUUGGGCGAGAAUUCGUAGGACAGUCUGAGACAGCACUCCACCUGCAGAUCCAA
```

The first step for the candidate sequence is to be folded into its secondary structure shown in Figure 3.5. This is done using *RNAfold* as described in Section 3.1 which produces the following secondary structure in dot-bracket format.

```
.((((((((.((((((...)))))))))).......(((((..((((((...))).)))...)))))...)))).
```

The dot-bracket sequence is used to parse the secondary structure into its structural elements, which are shown in Table 3.2. Metrics are then extracted from each of the structural elements in Table 3.2. For brevity only the *hairpin loop*, *internal loop*, and *joint* structural elements will be focused on from this point. The metric data for these structural elements is shown in Table 3.3.

Each of the data columns in Table 3.3 becomes a features set which is processed by its respective SVM. The two *hairpin loop* structural element data records will be processed on the *hairpin loop* SVM, the *internal loop* structural element will be processed with the *internal loop* SVM, and the two *joint* records will be processed with the *joint* SVM. Before any of these metric records are processed they are first scaled using scaling factors created by the SVM training set to normalize the data.

The scaled data records become the instances in Figure 3.6. Each of the structural element instances are passed to the trained SVM which will make a prediction based on the instance's metric values. A probability indicating the likelihood that the predicted class is the correct class is also returned by the SVM. Figure 3.6 shows that for the example

<table>
<tr><th>Element</th><th colspan="10">Structures</th></tr>
<tr><td>Structure</td><td colspan="10">GUGGGCGCCUCGAGUUUGAGAUUUGGGCGAGAAUUCGUAGGACAGUCUGAGACAGCACUCCACCUGCAGAUCCAA
.((((((((.((((((...)))))))))).......(((((..((((((...))).)))...)))))...)))).</td></tr>
<tr><td>External Loop</td><td colspan="10">GU AA
.( ).</td></tr>
<tr><td>Multiloop</td><td colspan="10">GC GAGAAUUCG CAGAU
)( ).......( )...(</td></tr>
<tr><td>Junction</td><td colspan="5">GU AA
.( ).</td><td colspan="5">GC GAGAAUUCG CAGAU
)( ).......( )...(</td></tr>
<tr><td>Bridge</td><td colspan="10">UGGG ACCU
(((( ))))</td></tr>
<tr><td>Stemloop</td><td colspan="5">CGCCUCGAGUUUGAGAUUUGGGCG
((((.((((((...))))))))))</td><td colspan="5">GUAGGACAGUCUGAGACAGCACUCCACCUGC
(((((..((((((...))).)))...)))))</td></tr>
<tr><td>Hairpin Loop</td><td colspan="5">UUGAG
(...)</td><td colspan="5">GAGAC
(...)</td></tr>
<tr><td>Stem</td><td colspan="3">UGGG ACCU
(((( ))))</td><td colspan="3">CGCCUCGAGUU GAUUUGGGCG
((((.(((((( ))))))))))</td><td colspan="4">GUAGGACAGUCUG CAGCACUCCACCUGC
(((((..(((((( ))).)))...)))))</td></tr>
<tr><td>Bulge</td><td colspan="5">CUC GG
(.( ))</td><td colspan="5">UC GCA
(( ).)</td></tr>
<tr><td>Internal Loop</td><td colspan="10">GACA UCCAC
(..( )...)</td></tr>
<tr><td>Loop</td><td colspan="3">CUC GG
(.( ))</td><td colspan="3">UC GCA
(( ).)</td><td colspan="4">GACA UCCAC
(..( )...)</td></tr>
<tr><td>Tail</td><td colspan="5">G
.</td><td colspan="5">A
.</td></tr>
<tr><td>Joint</td><td colspan="5">AGAAUUC
.......</td><td colspan="5">AGA
...</td></tr>
<tr><td>Joint-Tail</td><td colspan="2">G
.</td><td colspan="2">A
.</td><td colspan="3">AGAAUUC
.......</td><td colspan="3">AGA
...</td></tr>
<tr><td>Stack</td><td colspan="2">UGGG ACCU
(((( ))))</td><td colspan="2">CGCC GGCG
(((( ))))</td><td colspan="2">CGAGUU GAUUUG
(((((( ))))))</td><td colspan="2">GUAGG CCUGC
((((( )))))</td><td>AGU ACU
((( )))</td><td>CUG CAG
((( )))</td></tr>
<tr><td>Unpaired</td><td>G
.</td><td>U
.</td><td>UGA
...</td><td>AGAAUUC
.......</td><td>AC
..</td><td>AGA
...</td><td>C
.</td><td>CCA
...</td><td>AGA
...</td><td>A
.</td></tr>
</table>

Table 3.2: Parse Structural Elements. Shows the dot-bracket and nucleotide sequences for each of the structural elements in the candidate sequence's structure.

Metric	**Hairpin 1**	**Hairpin 2**	**Internal Loop 1**	**Joint 1**	**Joint 2**
Size	5	5	9	7	3
NLC	0.5000	1.0000	0.9000	1.0000	0.8333
SLC	0.5000	0.8333	0.5667	0.2917	0.5000
A%	0.3333	0.6667	0.4000	0.4286	0.6667
U%	0.3333	0.0000	0.0000	0.2857	0.0000
C%	0.0000	0.6667	0.6000	0.1429	0.0000
G%	0.0000	0.6667	0.0000	0.1429	0.3333
AA%	0.0000	0.0000	0.0000	0.1667	0.0000
AU%	0.0000	0.0000	0.0000	0.1667	0.0000
AC%	0.0000	0.0000	0.3333	0.0000	0.0000
AG%	0.0000	0.5000	0.0000	0.1667	0.5000
UA%	0.0000	0.0000	0.0000	0.0000	0.0000
UU%	0.0000	0.0000	0.0000	0.1667	0.0000
UC%	0.0000	0.0000	0.0000	0.1667	0.0000
UG%	0.5000	0.0000	0.0000	0.0000	0.0000
CA%	0.0000	0.0000	0.3333	0.0000	0.0000
CU%	0.0000	0.0000	0.0000	0.0000	0.0000
CC%	0.0000	0.0000	0.3333	0.0000	0.0000
CG%	0.0000	0.0000	0.0000	0.0000	0.0000
GA%	0.5000	0.5000	0.0000	0.1667	0.5000
GU%	0.0000	0.0000	0.0000	0.0000	0.0000
GC%	0.0000	0.0000	0.0000	0.0000	0.0000
GG%	0.0000	0.0000	0.0000	0.0000	0.0000
CS	33	33	N/A	36	36
FS	31	31	N/A	32	32
AU% Bond	0.0000	0.0000	0.5000	N/A	N/A
GC% Bond	1.0000	0.0000	0.5000	N/A	N/A
GU% Bond	0.0000	1.0000	0.0000	N/A	N/A
SR	N/A	N/A	0.6667	N/A	N/A

Table 3.3: Metric Data. Shows the structural element metric data for *hairpin loop*, *internal loop*, and *joint* structural elements from the candidate sequence's structure.

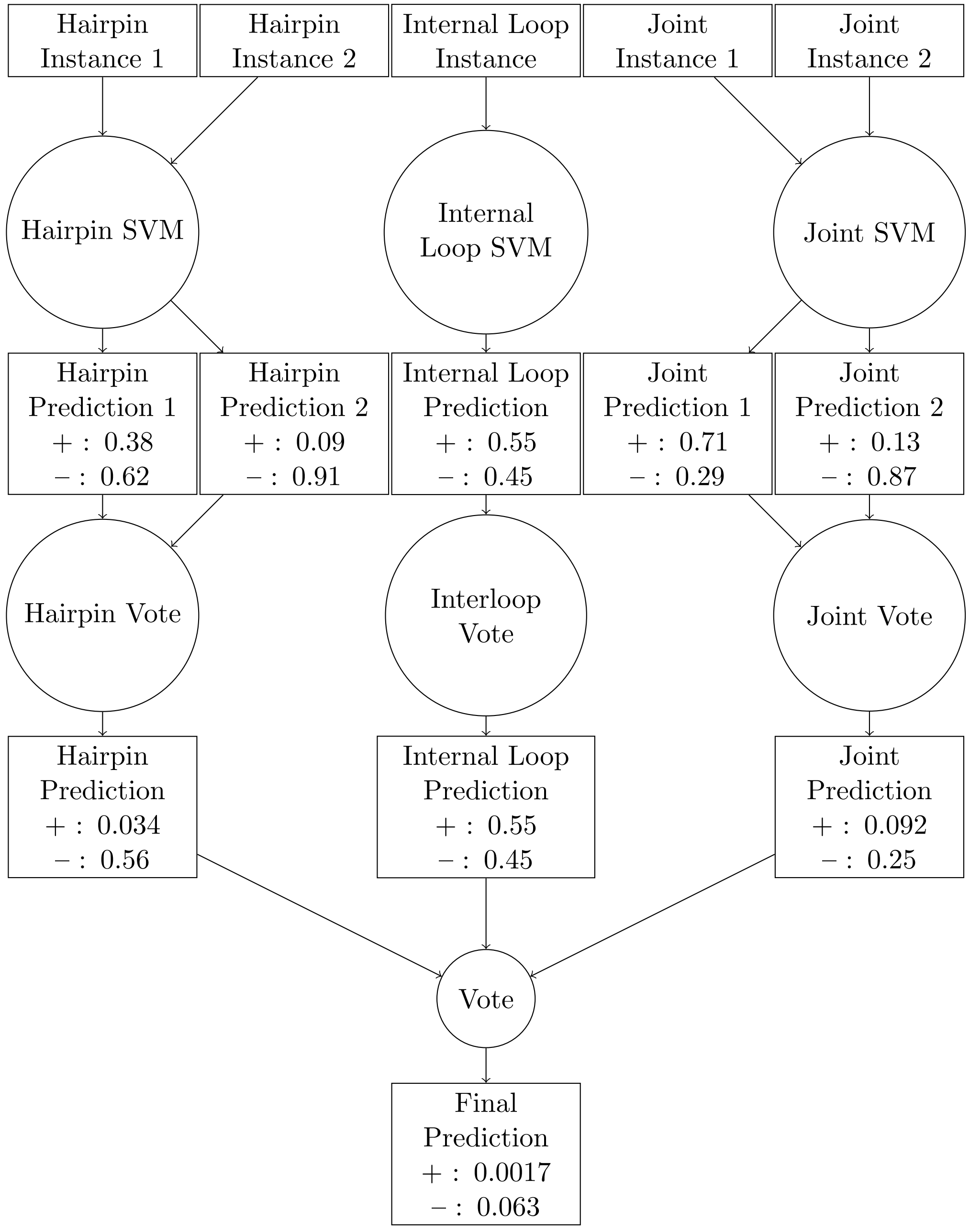
Hairpin Instance 1
Hairpin Instance 2
Internal Loop Instance
Joint Instance 1
Joint Instance 2
Hairpin SVM
Internal Loop SVM
Joint SVM
Hairpin Prediction 1 + : 0.38 − : 0.62
Hairpin Prediction 2 + : 0.09 − : 0.91
Internal Loop Prediction + : 0.55 − : 0.45
Joint Prediction 1 + : 0.71 − : 0.29
Joint Prediction 2 + : 0.13 − : 0.87
Hairpin Vote
Interloop Vote
Joint Vote
Hairpin Prediction + : 0.034 − : 0.56
Internal Loop Prediction + : 0.55 − : 0.45
Joint Prediction + : 0.092 − : 0.25
Vote
Final Prediction + : 0.0017 − : 0.063

Figure 3.6: Voting Example. Each of the columns of metric data in Table 3.3 are represented by the structural element instances in this diagram. The metric data is used by the SVMs to determine a class and the probability of that class being correct for the each of the instances. The diagram shows these probabilities along with the probability for the other class being correct as well. The two layers of the voting process is shown along with the calculated values.

sequence's structure the *hairpin loops* are both predicted to be from a non-SRNAG sequence with probabilities of 0.62 and 0.91. Since this is a binary classification the probability that the class other than the one predicted is correct is calculated by simply subtracting the SVM output probabilities from 1. In the case of the *hairpin loop* instances, the probabilities that they are from a SRNAG sequence are 0.38 $(1 - 0.62)$ and 0.09 $(1 - 0.91)$, respectively. The *internal loop* structural element parsed was predicted to come from a SRNAG sequence with a probability of 0.55. One of the *joint* structural elements was predicted to be from a SRNAG with a probability of 0.71 while the other *joint* element was predicted to not come from SRNAG with a probability of that prediction being correct of 0.87.

As described in Section 3.4, these predictions need to be combined together. The voting process treats each of the instance predictions as independent probabilistic events and attempts to use basic probability theory to combine them. Knowing that the *hairpin loop* structural elements receive a 0.62 and 0.91 chance of being from a non-SRNAG sequence, to calculate if both these elements come from a non-SRNAG sequence their probability values are multiplied together resulting in the value 0.56. To determine the probability value that both of these sequences came from a SRNAG their probabilities of being from a SRNAG are multiplied together $(0.38 \times 0.09 = 0.034)$. Since the probability that they are not from a SRNAG sequence is higher than the one that they are, the first level of the system would predict the sequence from which these two hairpin structures came from is not a SRNAG.

This same process happens to the *internal loop* and *joint* groups of predictions as well. In the case of the *internal loop* prediction since there are no other *internal loop* structures with which it can vote in the first layer of the system the probability values provided by the SVM are unaffected. The two *joint* structures with their different predicted labels also are combined through multiplication leading to a probability value of 0.092 (0.13×0.71) that they come from a SRNAG and a probability value of 0.25 (0.29×0.87) that they do not. This means the first layer of the voting system would predict that the *joint* structural elements came from a non-SRNAG sequence, while that the *internal loop* comes from a SRNAG.

To determine a prediction using all the structural elements, multiplication is once again used. All the probability values for a given label produced by the first layer of the voting system are multiplied together. This second layer predicts there is a 0.0017 $(0.034 \times 0.55 \times 0.092)$ probability that all the structural element instances came from a SRNAG, and a 0.063 $(0.56 \times 0.45 \times 0.25)$ probability that it did not. Since the probability of all structural element instances coming from a non-SRNAG sequence is higher than the probability that they did not, the system predicts that the candidate sequence is not a SRNAG. It should be made clear that although each of the instances is treated as an independent event they are not truly independent events and hence, while a relative comparison of the probability values is useful in determining which class the instances come from, the values themselves have little merit.

3.6 Chapter Review

This chapter introduces the major technological components of this thesis, allowing for a candidate nucleotide sequence to be folded, parsed, and for the structural element metrics to be extracted from it. Furthermore, the architecture of the classification engine is introduced detailing not only which SVM software is used, but how the SVM classifications are combined to give an overall prediction result of a candidate sequence. In regards to the classification of an unknown sequence the major computational complexity speed bump remains the $O(n^3)$ folding process. Finally a candidate sequence is walked through the system showing the results at each step as it is folded, parsed, has its features extracted, and is classified. The next chapter will explain the methodology used to train and test the RNA gene classifier proposed in this section.

Chapter 4

Methods

Presented in this chapter are the two major experiments used to validate the SRNAG finder and the hypotheses on which it relies. The first experiment deals with testing the SRNAG finder under perfect conditions where it must predict the label of whole SRNAG and non-SRNAG sequences, while the second experiment deals with SRNAG finding under unfavorable conditions by having the gene finder classify RNA sequence windows cut from SRNAG and non-SRNAG sequences. Each of these experiments is broken down further into two parts: one which deals with the structural elements and their metrics and one which deals with voting among structural elements. Specifically this chapter details the construction of the training and testing sets used by each experiment and the methodology used to train and test each of the classification models in the experiments.

As just mentioned, once the SRNAG finder proposed in Chapter 3 was built it is first tested using a dataset constructed from SRNAG segments and shuffled SRNAG segments, as seen in Figure 4.1 which shows the whole gene experiment. SRNAG segments and shuffled SRNAG segments provide the best possible scenario for this SRNAG finder, because the SRNAG finder uses RNA folding to generate the classification metrics. If the SRNAG segments are cut or concatenated with non-SRNAG sequence data the secondary structure produced through folding may not be closely related to the original SRNAG's secondary structure. Testing the SRNAG finder under perfect conditions checks whether the SRNAG software is working and that the key concepts utilized by the SRNAG are valid. Likewise, having the secondary structures reflect the natural shape of the molecules in nature provides a clean, accurate secondary structure element dataset for determining which structural elements and metrics produce strong SRNAG signals. Lastly, testing in favorable conditions

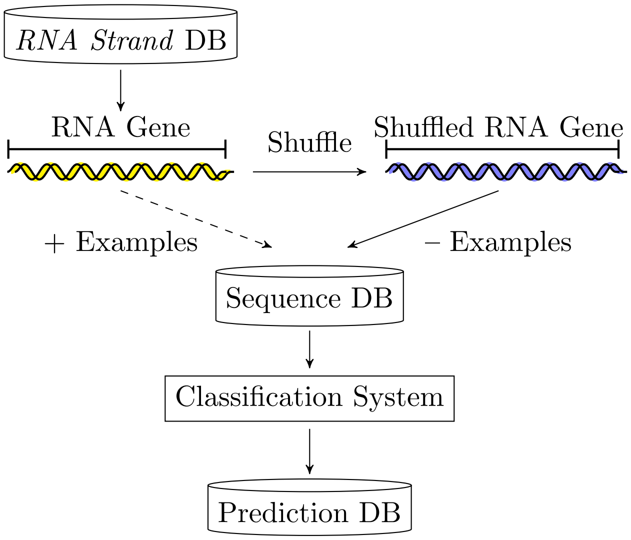

Figure 4.1: Whole RNA Gene Experiment. In the first experiment RNA genes are taken from *RNA Strand* and used as the positive sequence examples, while their dinucleotide shuffled counterparts are used as the negative sequence examples.

provides an upper bound for the SRNAG finder's performance.

Although testing under favorable conditions can provide insight into the method used for finding SRNAGs, the true colours of a SRNAG finder come to light when it is tested in adverse conditions. In this thesis the SRNAG finder is tested under conditions it would face if it was given a genome sequence with the same background dinucleotide composition as that of the SRNAGs' dinucleotide composition. Because the SRNAG finder proposed in this thesis relies on RNA folding, the genome segment will be cut into manageable sized sequences using a sliding window (described in Section 2.2). This second experiment is summarized in Figure 4.2. Since the genome being processed is unannotated the splitting of the genome most likely will not occur in such a way as to preserve whole SRNAGs, leading to candidate sequences which will contain only portions of SRNAGs or SRNAGs mixed with NC sequenced data. As stated in the last paragraph, this corruption of the candidate SRNAG segments can have a devastating effect on their secondary structures and hence the feature dataset will be very dirty. This sort of test gives a realistic lower bound for the performance of the classifier, verifying its performance in a harsh situation.

The next section outlines the building of the dataset for the first experiment, followed by a section on building the dataset for the second experiment. The training and testing

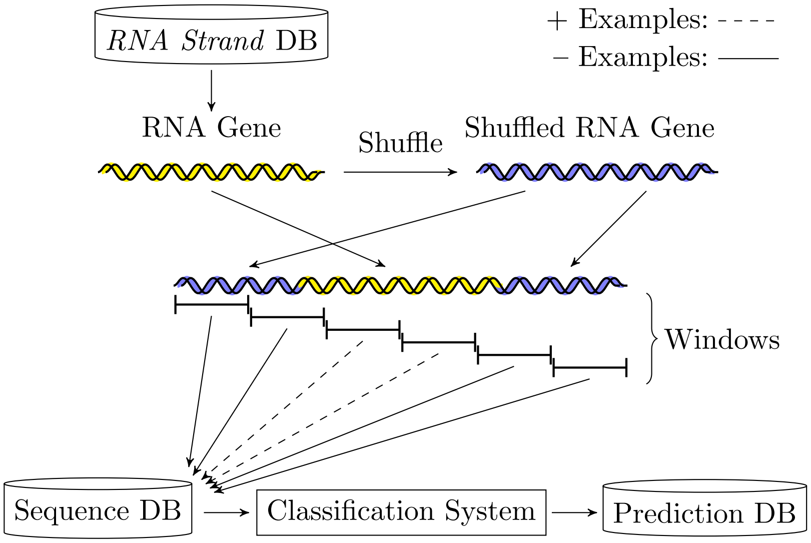

Figure 4.2: Sliding Window Experiment. The second experiment uses artificial mini-genomes created by embedding a whole RNA gene in its dinucleotide shuffled counterpart. A sliding window is run along these mini-genomes to segment it into candidate sequences for the SRNAG finder.

methodology sections follow.

4.1 Building a Dataset for Favorable Conditions

The dataset for the first experiment was built by acquiring 3386 gene sequences from *RNA Strand*[1]. Since these RNA gene sequences from *RNA Strand* are stored using the more general IUPAC nucleic acid codes, the characters representing more than one possible nucleotide are replaced with one of their representing nucleotides with an appropriate probability. The IUPAC coding scheme uses the alphabet $\Sigma = \{A, G, U, N, R, Y, K, M\}$ and the translations shown in Table 4.1. So, for example, in this step if a N appears in the nucleotide sequence for each A, U, G, or C there is a 25% probability that the nucleotide would replace it. Likewise, if there is a Y in the sequence there is be a 50% probability of it being replaced by a C and a 50% probability of being replaced by U. This conversion from IUPAC codes to standard RNA bases is necessary as the folding software which is used to predict the secondary structure of the RNA sequence does not handle these generalized bases and it

[1] www.rnasoft.ca/strand/

Name	**IUPAC Code**	**Nucleotide Representation**
Any	N	A, U, G, or C
Purine	R	G or A
Pyrimidine	Y	C or U
Keto	K	G or U
Amino	M	A or G

Table 4.1: IUPAC Nucleic Acid Codes

eases base composition analysis as probabilities do not need to be taken into account. Furthermore since the dataset size is large and these generalization codes are infrequent it is expected that the conversion process will not have a significant effect on the overall results.

These standard RNA sequences make up the positive example set, as they represent many different classes of real RNA genes. The number of genes representing each class is shown in Table 4.2 in the *Positive* examples row. The negative examples are generated from the gene sequences by dinucleotide shuffling the genes. Dinucleotide shuffling, as described in Appendix C, breaks up the gene by randomizing the nucleotide order while maintaining dinucleotide frequency.

Gene Class	16S rRNA	23S rRNA	5S rRNA	RNase P	SRP RNA	TmRNA	tRNA
Positive	723	205	161	470	394	726	707
Negative	723	205	161	470	394	726	707

Table 4.2: Positive and Negative Gene Dataset Examples. The positive examples are the ones downloaded from *RNA Strand*[3] while the negative examples are generated from the positive examples through dinucleotide shuffling, which is the reason there are an equal number of positive and negative examples for each class of genes.

Dinucleotide shuffling is chosen as the method for generating the negative dataset examples for a number of reasons. First, the other obvious source of negative examples would be the NC regions of a fully annotated genome; however, this alternative source may actually contain an undiscovered structural RNA gene which would pollute the negative example set. It is also possible that RNA shuffling by random chance could produce a sequence identical to a RNA gene, but this is highly unlikely. So in order to help preserve the purity of the example sets, nucleotide shuffling is used to generate the negative examples from the gene dataset. Second, because dinucleotide shuffling just shuffles the existing gene, while not affecting the dinucleotide base composition, it is guaranteed that the training and testing set is unbiased in regards to dinucleotide based composition, helping to ensure that trends caused by base composition are not learned by the prediction model as a distinguishing factor. Third, dinucleotide shuffling can be performed easily and quickly.

It should be made clear that the positive and negative examples are partitioned into the training set and testing set before the negative examples are generated in order to keep the dinucleotide content between the positive and negative sets balanced. This results in a training dataset with 80% of the example sequences and a testing dataset with 20% of the example sequences (see Table 4.3). The partitioning of the dataset is done through random selection for each gene class so that all the gene classes are represented fairly in both groups.

Gene Class	**Training**	**Testing**
16S rRNA	578	145
23S rRNA	164	41
5S rRNA	129	32
RNase P	376	94
SRP RNA	315	79
TmRNA	581	145
tRNA	586	141
Non-gene	2709	677
Total	5438	1354

Table 4.3: Training and Testing Sets. Shows the number of training and testing sequences for each class of RNA. The values for the number of training sequence should be roughly 80% of the total sequences in each category, leaving 20% of the whole dataset for the testing sequences.

4.2 Building a Dataset for Unfavorable Conditions

Originally, the RNA gene finder was going to utilize the models created from the training dataset described in the previous section, however this proved to be a poor decision, as those models are trained under favorable conditions which will lead to a very different set of feature distributions than when sliding window segments are used for candidate sequences. As stated in Chapter 1, since SVMs are trained classifiers, the testing dataset needs to come from the same distributions as the training dataset in order for SVMs to be effective. This means that each of the experiments have to have their own training and testing sets.

Since it is extremely time-consuming to fold windows along an entire genome, in order to train and test the gene finder, mini artificial genomes are constructed. Genes are selected from the gene database described in Section 4.1 and dinucleotide shuffled. The first half of the shuffled sequence is concatenated in front of the SRNAG used to produce it and the second half of the shuffled sequence is concatenated after it. The concatenation points are annotated. This methodology is similar to one of the experiments *Rivas et al.* used in [50] to show that their RNA gene finder was doing nothing more than using GC content as a signal for RNA gene finding. This technique for evaluating the usefulness of the RNA gene finder is chosen because it takes considerably less processing time to search the artificial mini genome for genes than an entire real genome. Likewise, it can be reasonably guaranteed that no RNA gene is present in the shuffled regions (see Section 4.1) and finding RNA genes within a region of the same GC content is the discerning test for RNA gene finders. Indicating that if the RNA gene finder is able to distinguish between RNA genes in RNA sequences of uniform dinucleotide content then it will most likely be as or more accurate in genomes where the GC content of the RNA genes is different than that of the surrounding genome.

A sliding window of length 80 nucleotides is used to generate the sequences for the training set. This specific size parameter is used because *Carter et al.* showed success using the same parameter value in their RNA gene finder [9]. Unlike in *Carter et al.* the window is shifted by a full 80 nucleotides each time. This fully disjoint shift allows data to be collected from all the nucleotides in the sequence without increasing the processing requirements.

Since each window of RNA will generate at least one record in the metric datasets, in order to keep the size of our dataset reasonable and consistent with the previous experiment, a quota scheme is used. Table 4.3 shows the number of genes used in the training set of the first experiment; these values make up the quotes of each category. So as windows are generated by moving sequentially along a strand of RNA, the quote value is decreased for the window's category until the quote gets to zero, at which point no more windows of that category are collected. This results in a new training set which has the exact same number of RNA segments of each type as the previous experiment. It should be noted, that non-gene windows have an individual category for each of the different gene types, so there is an equal number of windows generated from dinucleotide shuffled RNA of each gene type as from the original gene sequences of each type. Finally it should be made clear that a window is labeled with a gene label if more than 50% of it belongs to the gene segment, otherwise it is labeled as non-gene. The testing data set is created in the same way, with the quota values from the testing side of Table 4.3.

4.3 Training the Classification Models

With these sequence datasets in place, the metric dataset can be extracted from the training sequences as described in Chapter 3. This training data, will then be transformed and normalized according to the description given in Section 3.4 and then used to train the SVMs[4]—one for each structural element of the two experiments.

After the extraction of the feature set from the structural elements, there will be a bias in terms of the number of records for the SRNAG and non-SRNAG training examples. This bias is due to randomized RNA sequences being less likely to fold into RNA structures as structurally complex as known gene sequences, therefore the known gene sequences will tend to have more structural elements such as *stemloops* and *multiloops*. In order to re-balance the training sets, an even number of positive and negative examples are randomly selected from the metric dataset. When the dataset is balanced it is also reduced to a more manageable size. Since many of the structural elements will typically be numerous within, a structure the size of the element model training sets can become too large to process the SVM training in a reasonable amount of time. For example, the *unpaired* training set of the first experiment has over 400,000 examples and the *stemloop* set has over 56,000 (see Table 4.4). With such large datasets it is impossible to perform a single training, let alone a grid search in a practical time frame. For this reason as part of the dataset balancing, at most 20,000 records are taking from each element metric dataset to create a new training set. In the case where 20,000 data points are not available, the minimum value between the number of SRNAG and non-SRNAG data points is used. It is important to note that although the initial dataset has no base composition bias, it is possible that after the structure metric extraction a natural bias will appear. For example, it is possible that the *stemloops* of SRNAGs will end up with a higher GC content than the *stemloops* of non-SRNAGs. Biases such as these that appear naturally are exactly what the SRNAG finder will exploit to locate SRNAGs within a genome.

In order to calibrate the parameters of the SVM, a grid search is used, as described in Section B, which will try all the error penalty rates (C) and kernel parameter values (γ) within a specified search space. To perform the grid search, the grid search script provided with LIBSVM is used. The grid search uses cross validation to test the accuracy of the chosen parameters and allows any search space to be specified. Initially, a coarse grain

[4] A radial basis function is used for the kernel.

	Favorable Conditions		**Unfavorable Conditions**	
Structural Element	**Initial Size**	**Balanced Size**	**Initial Size**	**Balanced Size**
Bridge	40978	20000	10434	9986
Bulge	45061	20000	24145	20000
External Loop	5438	5438	5438	5438
Joint-Tail	126309	20000	62148	20000
Hairpin Loop	56855	20000	21051	20000
Internal Loop	89509	20000	30540	20000
Joint	120335	20000	57312	20000
Loop	134569	20000	54685	20000
Multiloop	40978	20000	15639	15482
Junction	45147	20000	18387	18274
Stack	217282	20000	938212	20000
Stem	97832	20000	31485	20000
Stemloop	56855	20000	21051	20000
Structure	5438	5438	5438	5438
Tail	5974	5974	6836	6648
Unpaired	407242	20000	168424	20000

Table 4.4: Training Set Resizing and Balancing. The Initial Size columns show the number of positive and negative training examples that were produced by the metric extraction process. The Balanced Size column shows the number of training examples after the training set has been resized and rebalanced. This is the final size of the training set for each structural element of each experiment. Note that the *Bridge*, *Multiloop*, *Junction*, and *Tail* rows have an initial size less than 20,000 yet a balanced set size smaller than the initial set size under unfavorable conditions. The reason for this reduction is due to there being less secondary structure element examples of one class than the other, so in order to create a set balanced in terms of class, the smaller class set size is matched with randomly selected examples from the larger class.

search space is used for configuration allowing regions which result in a good configuration to be identified and a fine grained search performed in the targeted area. The initial coarse grained search uses the default parameters in the LIBSVM grid search software, searching between -5 and 15 for the C parameter in steps of 2 and between 3 and -15 for the $log_2(\gamma)$ parameter in steps of -2. After the coarse grain evaluation is performed the data is plotted using a contour surface, such as the one seen in Figure 4.3, which allows one to easily spot the space for the fine grained search space.

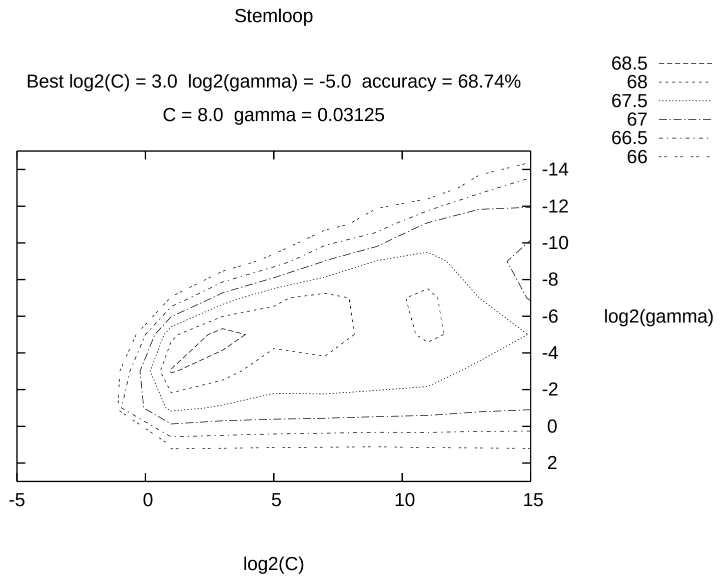

Figure 4.3: Grid Search Example Contour Map. This figure shows a contour map of the accuracy of a SVM tested using all the possible values of C between the -5 and 15 with a step of 2 and all the possible values of $log_2(\gamma)$ between 3 and -15 with steps of -2. This map of the coarse grain search enables easy visualization of the configuration parameter values which result in a strong classifier model, by grouping similar scoring regions into the same contour layer and labeling the layer by accuracy value.

After the fine grain parameter space evaluation is complete and plotted on a contour, the center of the area with the highest accuracy is used as the parameters for the SVM. With these parameters the SVM machine is retrained, with a parameter which allows the SVM to output a probability for its decision along with the label. This probability training is done after the grid search has been completed because it is more time-consuming for the

SVM training to build probability into the model and is largely unnecessary for the grid search. This final trained model is then saved to the disk along with the scaling parameters for later use in testing of the SRNAG finder.

It should be noted that there are other possible configuration techniques, such as local search, however grid search was chosen for two reasons. First, grid search exhaustively tries all the values specified by the step in the range, insuring a good spread of possible values are tested. Second, grid search can be easily parallelized as each configuration test is independent of all others. This allows for multiple computers to work on the same configuration problem at the same time, helping reduce the configuration time, which is significant because of the number of features and the large size of the training datasets.

4.4 Testing the SRNAG Finder

The structural element SVM models, trained in Section 4.3, are evaluated using the testing sequence datasets. Like the training sequence datasets, the test sequences are folded, parsed, the features extracted, and the data scaled; however, unlike the training set, the testing sets are not re-balanced for bias caused by the SRNAG sequences possibly having more structural elements than shuffled RNA sequences. Instead, all the testing records for each structural element are processed by the structural element's classification model, where the record label is hidden. These predictions are stored and analyzed in several different ways. The prediction records for each model are used to determine which structural elements produce strong SRNAG signals and which metrics in these structural element models contribute most to that signal. Likewise, the model predictions are voted together in several different ways allowing for groups of structural elements to be tested for their ability to classify SRNAGs. For each test, the hidden labels are compared to the overall prediction allowing true positives (TP), false positives (FP), true negatives (TN), false negatives (FN), accuracy (Acc.), precision (Prec.), recall, and F-measure (F-mea.) statistics to be calculated for the test. These statistics allow for a comparison and evaluation of the SRNAG finder and its component models, which will be presented in Chapter 5.

4.5 Chapter Review

The two experiments used to test the SRNAG finding method presented in this thesis were explained. Primarily, the creation of the datasets, the training of the models, and the testing of the models was outlined. The analysis of the results obtained by the tests is presented in the next chapter.

Chapter 5

Metric, Classifier, and Voting Analysis

While the last chapter explained the methodology of the two experiments conducted to test the classifiers, metrics, voting, and overall gene discovery engine, this chapter analyzes the results of those experiments. The overall results of the classifiers are discussed, that is the prediction results of all the models voting together to predict the sequence's class. Then the structural element models are ranked according to their classification ability and the structural elements which produce the strongest SRNAG finding signal are analyzed in depth. This analysis includes examining which metrics of the structural elements contribute most to its prediction power. Various voting schemes are also explored: individual structural element voting, double structural element voting, triple structural element voting, and progressive voting. Lastly, a method for reducing false positives is performed with the results discussed.

5.1 All-Inclusive Voting Analysis

To determine a benchmark for the performance of the SRNAG finder, the voting scheme was used to combine the prediction results of all of the SVM models in each experiment to create a classifier which used all the data available in the system for predicting candidate sequence labels. Each of the all-inclusive classifiers for each experiment was tested on their respective testing sets. The results are tabulated in Table 5.1 for the first experiment and Table 5.2

Gene Type	**TP**	**FP**	**TN**	**FN**	**Acc.**	**Prec.**	**Recall**	**F-mea.**
16S rRNA	0.4672	0.0146	0.4854	0.0328	0.9526	0.9697	0.9343	0.9517
23S rRNA	0.4583	0.0139	0.4861	0.0417	0.9444	0.9706	0.9167	0.9429
5S rRNA	0.4844	0.0000	0.5000	0.0156	0.9844	1.0000	0.9688	0.9841
RNase P	0.4890	0.0165	0.4835	0.0110	0.9725	0.9674	0.9780	0.9727
SRP RNA	0.4744	0.0513	0.4487	0.0256	0.9231	0.9024	0.9487	0.9250
TmRNA	0.3858	0.0079	0.4921	0.1142	0.8780	0.9800	0.7717	0.8634
tRNA	0.5000	0.0000	0.5000	0.0000	1.0000	1.0000	1.0000	1.0000
All	0.4537	0.0174	0.4826	0.0463	0.9363	0.9631	0.9073	0.9344

Table 5.1: All-Inclusive Voting Classification Statistics for Experiment 1. These are the prediction results from the experiment where whole genes and their shuffled counterparts make up the example set and every SVM classifier is used in the voting process.

for the second experiment. From the first experiment it can be seen that for the collection of test sequences, the combined classifier was able to achieve an accuracy of 93.63% and a F-measure value of 0.93. For specific RNA gene classes the classification engine was able to achieve F-measure values above 0.86, with a strong majority of the RNA gene classes' F-measure values above 0.92. These results show that given favorable conditions the SRNAG finder could achieve an accuracy of 93% for an assortment of RNA gene types with the metrics extracted in this research.

Unfortunately, when ideal conditions are not met, this strong ability to distinguish between SRNAG and non-SRNAG sequences degrades, as seen in the results for the second experiment shown in Table 5.2. Overall the second experiment's classifier showed some ability to distinguish between SRNAG and non-SRNAG sequence segments, achieving a classification F-measure of almost 0.7. Some specific gene classes proved to be easier for the classification engine in the second experiment to predict. The SRNAG finder's classification of 5S rRNA's managed to garner a F-measure value over 0.8, while RNase P received a F-measure over 0.78. These individual class results are not excellent, but show that even in unfavorable conditions the all-inclusive set of structural element models can distinguish between some groups of SRNAG sequences and their complementary randomized RNA sequences.

Gene Type	**TP**	**FP**	**TN**	**FN**	**Acc.**	**Prec.**	**Recall**	**F-mea.**
16SrRNA	0.4069	0.2828	0.2172	0.0931	0.6241	0.5900	0.8138	0.6841
23S rRNA	0.3659	0.3049	0.1951	0.1341	0.5610	0.5455	0.7317	0.6250
5S rRNA	0.4677	0.1935	0.3065	0.0323	0.7742	0.7073	0.9355	0.8056
RNase P	0.4681	0.2287	0.2713	0.0319	0.7394	0.6718	0.9362	0.7822
SRP RNA	0.3924	0.2532	0.2468	0.1076	0.6392	0.6078	0.7848	0.6851
TmRNA	0.4138	0.3345	0.1655	0.0862	0.5793	0.5530	0.8276	0.6630
tRNA	0.3750	0.2500	0.2500	0.1250	0.6250	0.6000	0.7500	0.6667
All	0.4174	0.2792	0.2208	0.0826	0.6382	0.5992	0.8349	0.6977

Table 5.2: All-Inclusive Voting Classification Statistics for Experiment 2. These are the prediction results from the experiment done with a sliding window where every SVM classifier is used in the voting process.

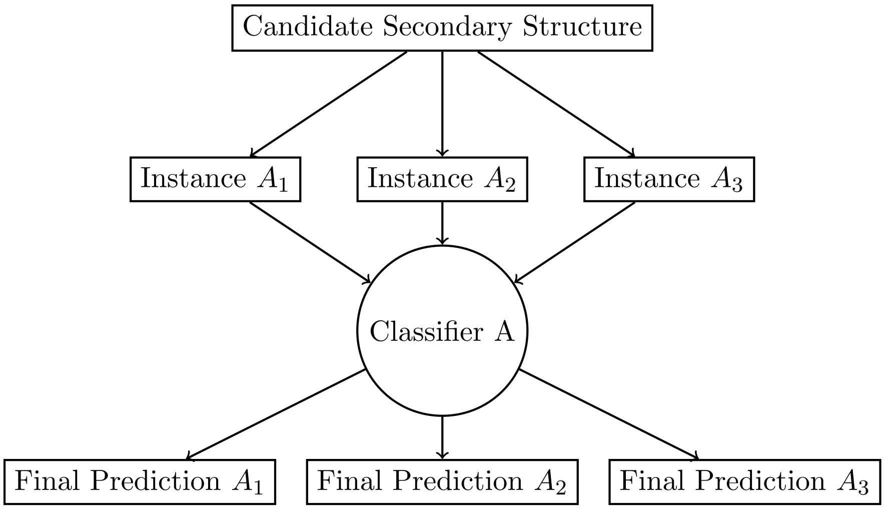

Figure 5.1: Structural Element Classifier Test. Each instance of a structural element is collected from a secondary structure and processed by the structural element's classifier to predict the label of the secondary structure's sequence. These predictions are compared to the real label and collected together to evaluate the performance of the classifier.

5.2 Individual Structural Element and Metric Analysis

Now that the benchmark has been set for the whole group of structural elements and metrics tested, a look at the prediction power of the individual structural element classifiers will help determine which of the structural elements and metrics are contributing most to the prediction power of these all-inclusive classifiers. This is done as shown in Figure 5.1, where each structural element of a given type is extracted from a test sequence and then classified by its structural element SVM model to generate performance statistics for the classifier. Table 5.3 lists each of the structural element models from the first experiment in order of descending F-measure value. The highest ranked structural element in Table 5.3 received a F-measure value of almost 0.84. When compared to the F-measure for all SRNAG classes in Table 5.1 it can be seen that through the use of voting an increase of almost 0.1 is achieved. This trend, showing the voting scheme's ability to increase the prediction power of the classifiers is also seen in the data from the second experiment. Table 5.4 contains the results for the individual structural element models from the second experiment. These results from the second experiment show that the highest ranking structural element model was the *structure* element's model with a F-measure of just under 0.62. This F-measure value is around 0.08 points lower than the all-inclusive classifier results from the second experiment in Table 5.2.

5.2.1 External Loop and Tail Analysis

The *external loop* of the first experiment ranked highest, with a F-measure of 0.84. This result is unexpected in conjunction with the poor performance of the *multiloop* and *junction* models which scored F-measure values of around 0.5 and prediction accuracies around 65%. Because the *external loop* is analogous to the *multiloop* and the *junction* is the *external loop*'s aggregate structural element, it is expected that these three structural elements would have similarly performing SVM models. The fact that only the *multiloop* and *junction* performed similarly is easily explained, as every structure will only ever have one *external loop* while possibly many *multiloops*, which causes the trends produced by the *external loop* to become diffused by the more numerous *multiloop* data. The unique performance of the *external loop* reveals that it is exploiting some special metric biases.

In a similar way, although the *tail* model in the first experiment was able to predict the correct label for the testing examples 71.31% of the time, its analogous structural element,

Structure	**TP**	**FP**	**TN**	**FN**	**Acc.**	**Prec.**	**Recall**	**F-mea.**
External Loop	0.4010	0.0541	0.4454	0.0995	0.8464	0.8811	0.8012	0.8392
Structure	0.3764	0.0319	0.4681	0.1236	0.8446	0.9220	0.7529	0.8289
Stemloop	0.3932	0.1665	0.3000	0.1404	0.6931	0.7025	0.7368	0.7193
Hairpin	0.3478	0.0956	0.3708	0.1858	0.7186	0.7844	0.6518	0.7120
Tail	0.3116	0.1187	0.4015	0.1686	0.7131	0.7241	0.6495	0.6848
Joint	0.3265	0.2785	0.2000	0.1950	0.5265	0.5397	0.6261	0.5797
Joint-Tail	0.2899	0.2405	0.2399	0.2297	0.5298	0.5466	0.5580	0.5522
Bridge	0.2048	0.0282	0.4542	0.3128	0.6590	0.8790	0.3956	0.5457
Stem	0.2288	0.0833	0.3898	0.2981	0.6187	0.7332	0.4343	0.5455
Stack	0.2665	0.2299	0.2750	0.2285	0.5416	0.5368	0.5384	0.5376
Multiloop	0.1774	0.0062	0.4762	0.3402	0.6535	0.9661	0.3427	0.5059
Junction	0.1783	0.0110	0.4730	0.3378	0.6513	0.9421	0.3455	0.5056
Unpaired	0.1917	0.1638	0.3455	0.2990	0.5372	0.5393	0.3907	0.4531
Internal Loop	0.1590	0.0793	0.4560	0.3057	0.6150	0.6673	0.3422	0.4524
Loop	0.1559	0.1093	0.4277	0.3071	0.5836	0.5878	0.3368	0.4282
Bulge	0.1425	0.1085	0.4319	0.3170	0.5744	0.5677	0.3101	0.4011

Table 5.3: Individual Structural Element Model Statistics for Experiment 1. This table shows the raw statistics for the prediction results achieved by the classification models from experiment 1. The structural elements are listed in order of descending F-measure.

the *joint*, achieved an unimpressive accuracy of only 52.65% even though *joints* should have a stronger bias as they are more numerous than *tails* and thus have a larger impact on the shape of the secondary structure. The contradictory performance of the *external loop* and *tail*'s analogous and aggregate models points to strong metric biases in these structural elements developed due to evolutionary pressure. In the second experiment the external loop and tail structural elements did not perform nearly as well, achieving F-measure values of 0.54 and 0.20 respectively. This drop in prediction power is due to the sliding window resulting in windows which cut off the gene tails or extend past the end of the gene and contain shuffled RNA at the ends of their sequences.

It is important to note that the *external loop* and *tail* metric signals in the first experiment, shown in Table 5.5 and Table 5.6 respectively, are naturally occurring biases. For example, the highest F-score ranked *external loop* metric of the first experiment is the average G composition in the *external loop*'s *tails*. Table 5.5 indicates that SRNAGs on average will have a value of 11.73% for this metric, which is much lower than the 25% value expected by random chance. This 25% probability of G bases in the *tail* elements is roughly

Structure	**TP**	**FP**	**TN**	**FN**	**Acc.**	**Prec.**	**Recall**	**F-mea.**
Structure	0.3024	0.1753	0.3247	0.1976	0.6271	0.6330	0.6048	0.6186
Stemloop	0.3120	0.2008	0.2878	0.1995	0.5997	0.6084	0.6100	0.6092
Hairpin	0.2380	0.0892	0.3994	0.2734	0.6374	0.7274	0.4654	0.5676
External Loop	0.2616	0.2161	0.2839	0.2384	0.5455	0.5476	0.5232	0.5351
Joint-Tail	0.2605	0.2198	0.2764	0.2432	0.5370	0.5424	0.5172	0.5295
Joint	0.2548	0.2233	0.2714	0.2505	0.5262	0.5330	0.5042	0.5182
Stack	0.2322	0.2090	0.2980	0.2609	0.5302	0.5263	0.4709	0.4970
Loop	0.1973	0.1795	0.3492	0.2740	0.5465	0.5236	0.4186	0.4652
Unpaired	0.1807	0.1646	0.3444	0.3102	0.5252	0.5233	0.3682	0.4322
Stem	0.1478	0.0781	0.4141	0.3600	0.5619	0.6541	0.2910	0.4028
Internal Loop	0.1421	0.0881	0.4320	0.3378	0.5741	0.6174	0.2961	0.4002
Bridge	0.1322	0.0554	0.4542	0.3582	0.5864	0.7045	0.2696	0.3899
Bulge	0.1165	0.1109	0.4362	0.3364	0.5527	0.5122	0.2571	0.3424
Multiloop	0.0896	0.0128	0.4968	0.4009	0.5864	0.8750	0.1826	0.3022
Junction	0.0608	0.0142	0.4887	0.4363	0.5495	0.8103	0.1222	0.2124
Tail	0.0613	0.0543	0.4437	0.4407	0.5050	0.5303	0.1221	0.1985

Table 5.4: Individual Structural Element Model Statistics for Experiment 2. This table shows the raw statistics for the prediction results achieved by the classification models from experiment 2. The structural elements are listed in order of descending F-measure.

observed by the randomly shuffled SRNAGs, where the difference in value is likely due to an uneven distribution of the four bases in the downloaded SRNAGs. The problem is that in the second experiment, as additional nongene nucleotides are included as part of tail, the G composition bias washes out; likewise, if the SRNAG segment length is cut short excluding nucleotides that were originally part of the SRNAG's secondary structure, the composition bias would be distorted. This washing out is exactly what is seen when the first experiment is compared with the second experiment. The F-score for the G composition of *tails* in *external loops* drops from 0.0718 in the first experiment, where it had the highest F-score, to 0.0001 in the second experiment, ranking it 68th out of the 104 metrics tested (see Table E.18). Comparing a SRNAG to shuffled RNA it can be seen from this experiment that SRNAGs tend to have *external loops* with a lower G composition in their tails, higher sequence repetition in their joint nucleotide sequences, smaller subtrees of elements branching off, and fewer joints than the shuffled RNA sequences. On the other hand *tail* structural elements have distinguishing biases based on nucleotide base composition and repetition in the nucleotide sequence. In essence, a SRNAG finder could attempt to exploit the properties of *external loops* and *tails*. However, constructing a SRNAG finder which uses segment folding to extract metrics to make use of the *external loop* and *tail* metrics would require selecting candidate SRNAG segments very precisely, as the greater the misalignment of the candidate segment and the embedded SRNAG segment, the larger the deterioration of the metric biases. More importantly for this research, it would be very hard to develop an ad hoc method for locating *tails* in a genome as the *tail* structural element is very small, often only consisting of a couple of nucleotides. Likewise, the sequence components which make up the structural element are distributed throughout the nucleotide sequence of the secondary structure making it also very challenging to detect using an ad hoc method. This criteria makes it challenging to construct an efficient SRNAG finder which can exploit the features of *external loops* and *tails*.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
Avg. Tail G%	0.0718	0.1173	0.1881	0.2332	0.2415
Avg. Joint NLC	0.0683	0.2816	0.4316	0.5140	0.4593
Avg. Joint CS	0.0682	3.7600	1.9277	4.0320	2.0743
Avg. Joint-Tail G%	0.0655	0.1306	0.1840	0.2315	0.2097
Joint%	0.0575	0.0898	0.1659	0.1792	0.2057
G%	0.0567	0.0930	0.1194	0.1485	0.1131
Avg. Joint FS	0.0561	1.1378	0.7635	1.0057	0.0805
Avg. Joint A%	0.0537	0.1192	0.2475	0.2523	0.3229
Size	0.0489	10.3385	7.0960	13.4796	7.1502
Avg. Joint Size	0.0488	1.0654	2.2723	2.3128	3.2940

Table 5.5: External Loop Metric Statistics for Experiment 1. Lists, in descending order of F-score, the top 10 metrics for the *external loop* structural element of the first experiment.

5.2.2 Structure Analysis

Being the largest and most inclusive structural element, it is no surprise that the *structure* structural element ranks high in Table 5.3 and 5.4. The *structure* structural element has a classification accuracy of 84.46% for the first experiment and 62.71% in the second experiment. This drop in prediction accuracy seen in the second experiment is expected as the *structure*, being a global structural element, would be affected by the changes in the secondary structure created from a SRNAG sequence which includes additional nucleotides

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
G%	0.0385	0.1401	0.2446	0.2480	0.2935
UU%	0.0300	0.1234	0.2741	0.0469	0.1463
NLC	0.0284	0.8745	0.1464	0.9189	0.1124
AC%	0.0190	0.0844	0.1788	0.0421	0.1251
GU%	0.0181	0.0231	0.1088	0.0637	0.1790
CC%	0.0142	0.0596	0.1389	0.0284	0.1142
FS	0.0133	124.3385	250.8350	72.2504	189.4729
CS	0.0130	127.2597	251.1728	75.5594	189.7670
U%	0.0112	0.2901	0.3409	0.2236	0.2706
AA%	0.0087	0.0692	0.1869	0.1071	0.2245

Table 5.6: Tail Metric Statistics for Experiment 1. Lists, in descending order of F-score, the top 10 metrics for the *tail* structural element of the first experiment.

or excludes part of the gene. As seen in Table 5.7, the highest ranked metric for the first experiment is the percentage of joint nucleotides within the *structure*, while the highest ranked metric in the second experiment for the *structure* structural element is MFE. As discussed in Chapter 2 it is expected that MFE would perform well for SRNAG finding, yet it did not show significance in the first experiment where it ranked very low receiving a F-score of 0.0016. However, on closer inspection the first experiment used training and testing sequences of many different lengths. Since MFE is directly related to the length of the RNA sequence [50], even though the training and testing datasets were balanced in terms of length, the large amount of variation in length in the positive sequences (and hence negative sequences) leads to high variation in the MFE metric as well. This variation results in a low F-score for the MFE metric because as the standard deviation of the positive MFE distribution and negative MFE distribution increases these distributions will tend to overlap significantly. This increased variation will cause low F-scores for MFE, even in-spite of the natural tendency for SRNAGs to have lower MFE values than random RNA. The lower average MFE for SRNAG's can be easily seen in Table 5.7. In the second experiment where the length of the RNA sequences was fixed, MFE was able to be a powerful gene finding signal as the MFE distributions had significantly less variation.

As already mentioned, the percentage of *joint* nucleotides is the highest ranked metric for the *structure* structural element of the first experiment. The averages of the two classes reveals that nongene RNA is slightly more likely to fold into a structure with a higher volume of joints. Since *joints* are unpaired regions it makes sense that the second highest ranked metric is percent paired, where the nongene secondary structures tend to be less paired than the SRNAG ones. These high scoring metrics are most likely a product of evolutionary pressures on SRNAG sequences causing them to develop well-defined pairing configurations which hold the RNA sequences in their functional shapes. Since no such evolutionary pressure is applied to the shuffled gene sequences there is a different probability distribution dictating which nucleotides will form a bond. Likewise, RNA genes tend to have fewer *tails* than non-SRNAG. Since the SRNAGs used came from *RNA Strand* where the nucleotide sequences are trimmed to encompass no more than the SRNAG, it makes sense that SRNAGs would have fewer *tails* in their folded structure as it would be more likely for a random segment of RNA than for a perfectly sized gene segment to fold into a structure with *tails*. *Stems* fully compose *bridges* while partially compose *stemloops* so it is expected that the density of these important and related structures in the secondary structure would be both highly ranked and grouped together with similar F-scores in Table 5.7, where each of these structures is more prevalent in SRNAGs then nongenes. The same is true about the important *hairpin loop* structural element, which are more frequent in SRNAG structures than non-SRNAG structures. Once again because secondary structure is dependent on the whole sequence it is no surprise that the metrics which were shown to be useful in the first experiment were not found to be useful signals in the second experiment where the segment data is dirty. The second experiment found MFE to be a useful metric and in addition to this metric it is seen from Table 5.8 that on average the SRNAG windows studied tend to be composed of fewer *bridge* nucleotides, have slightly lower nucleotide sequence linguistic complexity, and are composed more of *stemloops*.

Although in the first experiment no bias in base composition and *structure* size was present in the sequence dataset, it is noteworthy that in a real genome there would most likely be some bias in the composition which could be exploited to help classify genome windows. Because the *structure* structural element requires folding the whole sequence, it does not meet some of the requirements hoped to be achieved by the RNA structural elements in this thesis. Mainly the *structure* structural element fails to allow an improvement in algorithm speed, and it does not allow the use of smaller pieces of the secondary structure to be used to generate signals for SRNAG, allowing more naturally defined locations and sizes of SRNAGs to be annotated in the genome. However even with these drawbacks, the *structure* structural element is still important in terms of RNA gene finding, as new *structure*

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
Joint%	0.0921	0.2913	0.1826	0.3571	0.1783
PP	0.0513	0.6190	0.0463	0.5960	0.0548
% Num. Tail	0.0382	0.0276	0.0029	0.0288	0.0024
% Num. Stemloop	0.0228	0.0004	0.0016	0.0001	0.0001
% Num. Hairpin Loop	0.0228	0.0004	0.0016	0.0001	0.0001
% Num. Stem	0.0224	0.0008	0.0028	0.0002	0.0001
% Num. Bridge	0.0218	0.0003	0.0012	0.0001	0.0001
% Num. Multiloop	0.0218	0.0003	0.0012	0.0001	0.0001
Internal Loop%	0.0212	0.1463	0.0639	0.1650	0.0646
% Num. Stack	0.0189	0.0023	0.0090	0.0006	0.0004
% Num. Bulge	0.0168	0.0005	0.0018	0.0001	0.0001
⋮	⋮	⋮	⋮	⋮	⋮
MFE	0.0016	-297.3350	319.0345	-272.8374	296.5180

Table 5.7: Structure Metric Statistics for Experiment 1. Lists the top 10 metrics and MFE of the *structure* structural element for the first experiment and ranks them in descending order by F-score.

metrics which are found to be useful for SRNAG finding can help improve the accuracy of existing SRNAG finding methods by adding additional data into the prediction systems. Although the other metrics which ranked high in the second experiment may be of some use, MFE is the dominant metric when it comes to using a fixed length sliding window. The first experiment's metrics may be harder to make use of, since the data is based on folding sequences of nearly the exact gene segment versus shuffled RNA sequences. Currently there is no computationally feasible way to extract such clean data without already knowing where the genes lie within the genome. Nonetheless, the metrics in the first experiment can provide some possible hints at metrics, such as using higher pairing potential and higher density of *stems* for SRNAG signals–two signals which may not necessarily require folding to extract.

5.2.3 Stemloop and Hairpin Analysis

The first experiment produced classification results where the *stemloop* SVM model achieved a 0.72 F-measure value and the *hairpin* received a F-measure of 0.71. Table 5.9 and Table 5.10 show that the highest two F-score ranked metrics for these two secondary structure elements are CS and FS, where in both cases SRNAGs are shown to have more closely spaced

		Gene		Nongene	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
MFE	0.0190	-22.8759	7.7273	-20.8579	6.8840
Bridge%	0.0061	0.0671	0.1150	0.0861	0.1274
NLC	0.0049	0.9871	0.0046	0.9877	0.0039
Stemloop%	0.0042	0.7495	0.1677	0.7269	0.1818
Multiloop%	0.0029	0.0753	0.1102	0.0874	0.1142
SLC	0.0027	0.9303	0.0213	0.9326	0.0223
Internal Loop%	0.0014	0.1237	0.0895	0.1305	0.0882
Bulge%	0.0013	0.0245	0.0321	0.0269	0.0355
AG%	0.0010	0.0733	0.0259	0.0716	0.0271
AA%	0.0009	0.0809	0.0440	0.0783	0.0440

Table 5.8: Structure Metric Statistics for Experiment 2. Lists the top 10 metrics of the *structure* structural element for the second experiment and ranks them in descending order by F-score.

stemloops and *hairpins*. Since each *stemloop* has a *hairpin* at its tip, it is not surprising that the spacing of these structural elements is related. Within Table 5.10 which shows the top metrics for the *hairpin* structural element model of the first experiment, it is observed that the rest of the metrics have F-scores over a factor of ten lower than the CS and FS metrics, indicating that most of the prediction power for the *hairpin* model comes from CS and FS. On the other hand, Table 5.9 shows a more diverse picture of the metrics, with the SRNAG sequences shown to have larger *stacks* in their *stemloops* and more pairing. However, the fact that both the *stemloop* SVM and *hairpin* SVM achieved similar prediction power, the high F-scores of the CS and FS metrics in both models, and that CS and FS are related in both structures, provides evidence that even in the *stemloop* model the contribution of the CS and FS metrics to the prediction power significantly outweighs the other metrics.

These results are interesting for a number of key reasons. First, since these experiments were done in the absence of base composition bias, the *stemloop* and *hairpin* CS and FS metrics should have similar RNA gene signals in genomes of any background base composition. Second, the figures in Table 5.7 for *stemloop* CS support Noël's hypothesis that properties of *stemloops* in a genome can provide useful SRNAG finding signals. This support is more rigorous than Noël's work because the *stemloops* were located using folded MFE secondary structures instead of an ad hoc method. Third, the structure metric data tabulated in Table 5.7 also correlates these finding showing that the secondary structures of SRNAG sequences have a bias towards containing a higher concentration of *stemloops*. A higher concentration of stemloops means they must be closer together. Fourth, in addition to CS, FS is also shown to have some signals for RNA gene finding in the absence of base composition bias, but weaker signals than CS signals. Lastly, both CS and FS metrics produce strong signals even for *hairpins*, which might remove the need to locate the full *stemloop* and just search for the smaller *hairpin* structures.

Recently, a method for locating generic *hairpin loops* was presented by Jennifer Smith [57]. Using small co-variance models Smith achieved an extremely low false positive rate yet quick genome scans [57]. Such a method could easily exploit the *hairpin* spacing distribution with the *hairpins* it found to act as an additional RNA gene indicator. Furthermore, it may be possible to use *stemloop* and *hairpin* spacing in SRNAG finding methods which use a folded sliding window to extract MFE; however, this assertion is not supported by the second experiment, where the prediction power of *stemloop* and *hairpin* is poor, achieving F-measures of 0.61 and 0.57 respectively. Note that these two structural elements did rank high in Table 5.4, only the *structure* structural element has a higher F-measure.

In the second experiment, the CS and FS metrics are very low in terms of F-score showing that they produced very little SRNAG finding signal. Although there are many possible causes for these poor signals, a major factor is due to the window size used in the experiment. An 80 nucleotide window was used following in the footsteps of Carter [9]. Carter was using the window of 80 nucleotides to exploit the metrics of MFE, base composition, and motif presents; however given that the average *stemloop* and *hairpin* CS seen in Table 5.9 for SRNAG is over 51 nucleotides, it is possible that a window of 80 nucleotides is not large enough to capture enough sequence information for two or more full *stemloops*. For this reason it is possible that if a larger sequence window was used, the bias seen in these metrics would improve. On the other hand there is no guarantee that the problems caused by using a sliding window will not cause the folded SRNAG window segment's secondary structure to be deformed to the point that *stemloop* and *hairpin* CS and FS have no statistical distinguishing difference between the SRNAG segments and random RNA segments.

Although there is no guarantee that the folded SRNAG window will represent some portion of the real SRNAG's secondary structure, the metrics collected from higher F-score ranked metrics of the folded sliding windows still may be useful in revealing some bias in SRNAG sequences. Table 5.11 shows that SRNAG windows tend to fold into structures containing *stemloops* with larger *stack* sizes, with more of the nucleotides composing *stacks*, and a higher percentage of pairing nucleotides. Clearly these three properties are related, as stacks are the only paired structure within the *stemloop*. Furthermore, these biases are exactly what would be expected since evolution should drive SRNAGs toward higher pairing patterns. Unfortunately the hairpin element model showed little ability to distinguish RNA in the second experiment.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
CS	0.0234	51.2419	20.5950	58.1415	24.9893
FS	0.0225	24.3202	16.6562	30.0029	21.5590
Avg. Stack Size	0.0197	9.2210	3.5871	8.2691	3.1423
Stack%	0.0177	0.6399	0.1128	0.6062	0.1395
PP	0.0143	0.6490	0.0980	0.6244	0.1080
Avg. Stack CC%	0.0073	0.1177	0.1806	0.0889	0.1577
Avg. Hairpin Loop GG%	0.0065	0.0337	0.1076	0.0532	0.1357
Avg. Stack GG%	0.0065	0.1497	0.1923	0.1203	0.1748
Avg. Hairpin Loop AG%	0.0058	0.0547	0.1119	0.0730	0.1265
Avg. Stack C%	0.0053	0.3001	0.1111	0.2836	0.1155

Table 5.9: Stemloop Metric Statistics for Experiment 1. Lists the top 10 metrics of the *stemloop* structural element for the first experiment and ranks them in descending order by F-score.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
CS	0.0232	51.3026	20.7227	58.1448	25.0076
FS	0.0215	47.0594	20.6320	53.5970	24.9224
GG%	0.0065	0.0337	0.1078	0.0532	0.1357
AG%	0.0058	0.0550	0.1123	0.0730	0.1265
SLC	0.0051	0.3867	0.0799	0.3746	0.0865
Size	0.0051	5.2336	2.0237	5.5494	2.3551
AA%	0.0047	0.1887	0.2227	0.1587	0.2191
GC% Bond	0.0037	0.6890	0.4629	0.6320	0.4823
AU%	0.0032	0.0552	0.1079	0.0685	0.1213
AU% Bond	0.0027	0.2066	0.4049	0.2491	0.4325

Table 5.10: Hairpin Metric Statistics for Experiment 1. Lists the top 10 metrics of the *hairpin* structural element for the first experiment and ranks them in descending order by F-score.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
Avg. Stack Size	0.0114	8.5623	3.2362	7.9108	2.7886
Stack%	0.0082	0.6177	0.1193	0.5956	0.1258
PP	0.0079	0.6309	0.1003	0.6130	0.1022
Avg. Stack SLC	0.0048	0.4459	0.1060	0.4606	0.1059
GC% Bond	0.0045	0.5931	0.2174	0.5897	0.2222
Avg. Hairpin Loop SLC	0.0037	0.3801	0.0852	0.3697	0.0894
Avg. Hairpin Loop Size	0.0037	5.4184	2.2269	5.7052	2.6124
AU% Bond	0.0032	0.2831	0.1958	0.2852	0.1971
Avg. Hairpin Loop GG%	0.0026	0.0420	0.1186	0.0561	0.1399
Avg. Loop AA%	0.0024	0.0643	0.1599	0.0505	0.1274
⋮	⋮	⋮	⋮	⋮	⋮
CS	0.0004	29.3482	8.4823	29.3207	8.4559
⋮	⋮	⋮	⋮	⋮	⋮
FS	8e-06	5.7058	4.4837	5.7958	4.6108

Table 5.11: Stemloop Metric Statistics for Experiment 2. Lists the top 10 metrics along with CS and FS of the *stemloop* structural element for the second experiment and ranks them in descending order by F-score.

5.2.4 Poor Performing Structural Element Models

Observe in Table 5.3 and Table 5.4 that there are many structural elements shown to be ineffective for SRNAG finding using the metrics presented in this thesis. Particularly for the first experiment, the structural elements which did not result in models of high prediction power are unlikely to improve in further experiments as the first experiment represents a best case scenario for SRNAG finding. If the structural elements failed to produce a signal under these excellent conditions it means that there are no large enough differences between the features of the secondary structure of shuffled RNA and real SRNAGs for a classifier to distinguish between them.

It was expected that the *unpaired* structural element would not perform well as *unpaired* elements are small and contain little structural information; they were included in the structural element set simply for completion. On the other hand, it was surprising to see how poorly several other structural elements performed under these ideal conditions. The *stems* and *bridges* are relatively large, important structural elements yet did not achieve F-measures above 0.55; likewise, *multiloops* and *junctions* are critical to the global secondary structure yet achieved F-measures below 0.51. In the second experiment it is easy to see from Table 5.4 that none of the poor predicting structural elements from the first experiment were able to produce a strong RNA gene signal. The structural elements which performed well in the first experiment, yet failed to perform well in the second experiment have previously been discussed. In the next section, voting is used to try to improve the SRNAG finder's prediction accuracy by utilizing the predictions of several structural elements.

5.3 Composite Model Analysis

The first section in this chapter discussed the results of the two experiments when all the structural element models were used in the voting process and the previous section analyzed metrics involved in producing the prediction signals in those models. This section looks at the creation of a composite model where a limited number of structural element models are used in the voting process.

5.3.1 Individual Structural Element Voting

One of the goals of this thesis is to reduce the number of structural elements which need to be measured while maintaining a strong classification accuracy. The logical starting place to look is at voting within a structural element type (i.e. all the *stemloops* of a candidate sequence voting). Table 5.3 for the first experiment and Table 5.4 for the second experiment show the predictive power of each structural element type based on individual instances of structural element data. For all the structural elements except the *structure* and *external loop* elements there is a possibility that a single RNA sequence could fold into a structure with multiple elements of the same type. Table 5.12 and Table 5.13, for the first and second experiments respectively, show the prediction results for all the instances of a structural element type within a sequence voted together (see Figure 5.2). Clearly the *structure* and *external loop* elements will have the same values in the non-voting tables and the voting tables since each structure can have only one *structure* element and *external loop* element. As expected, the introduction of more information into the classification process increases the performance of the system. This is seen in the comparison of Table 5.3 with Table 5.12 and Table 5.4 with Table 5.13 where every structural element's prediction power in the first experiment increased due to voting (except the *structure* element and *external loop* element) and 11 of the structural elements in the second experiment increased in prediction power due to voting. In the first experiment some of these increases were as high as 0.27 and many of the structural elements which ranked low when considered individually, ranked much higher with the voting process. In the second experiment, the increases were not as large. The highest increase being 0.23 F-measure points for the *unpaired* structural element, but like the first experiment the increases often changed the ranking of the structural elements in terms of F-measure. It is interesting to note that structural elements which are numerous in secondary structures will tend to gain the most through this sort of voting, as more votes means an increase of information in the system, which should lead to higher prediction accuracy.

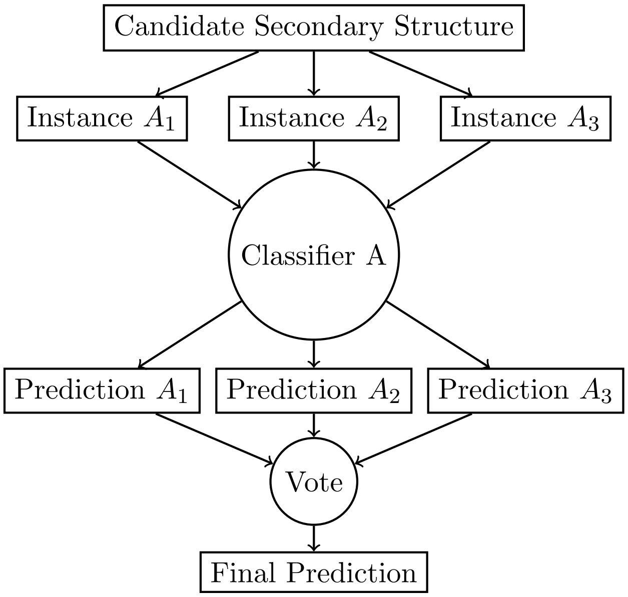

Figure 5.2: Individual Structural Element Voting. Every instance of a structural element type is collected from a secondary structure and processed by the structural element's classifier. The predictions of the classifier are voted together to predict the class of the candidate structure.

5.3.2 Structural Element Paired Voting

In the same way, Table 5.14 and Table 5.15 records the results of voting with combinations of two structural elements (see Figure 5.3), sorted in descending F-measure value. From the first table a number of trends are observed. First, as expected, most of the high ranking structural element combinations are composed of elements which also rank high when

Structures	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
External Loop	0.4006	0.0541	0.4459	0.0994	0.8465	0.8811	0.8012	0.8392
Hairpin	0.4015	0.0627	0.4373	0.0985	0.8388	0.8649	0.8031	0.8328
Structure	0.3764	0.0319	0.4681	0.1236	0.8446	0.9220	0.7529	0.8289
Stemloop	0.3996	0.0782	0.4218	0.1004	0.8214	0.8364	0.7992	0.8174
Stem	0.3755	0.1187	0.3813	0.1245	0.7568	0.7598	0.7510	0.7553
Bridge	0.3591	0.1081	0.3919	0.1409	0.7510	0.7686	0.7181	0.7425
Multiloop	0.3041	0.0174	0.4826	0.1959	0.7867	0.9459	0.6081	0.7403
Tail	0.3716	0.1525	0.3475	0.1284	0.7191	0.7090	0.7432	0.7257
Internal Loop	0.3514	0.1236	0.3764	0.1486	0.7278	0.7398	0.7027	0.7208
Loop	0.3649	0.1988	0.3012	0.1351	0.6660	0.6473	0.7297	0.6860
Stack	0.4710	0.4064	0.0936	0.0290	0.5647	0.5369	0.9421	0.6840
Unpaired	0.3649	0.2095	0.2905	0.1351	0.6554	0.6353	0.7297	0.6792
Joint-Tail	0.4662	0.4431	0.0569	0.0338	0.5232	0.5127	0.9324	0.6616
Junction	0.4662	0.4431	0.0569	0.0338	0.5232	0.5127	0.9324	0.6616
Joint	0.4624	0.4672	0.0328	0.0376	0.4952	0.4974	0.9247	0.6469
Bulge	0.3639	0.3282	0.1718	0.1361	0.5357	0.5258	0.7278	0.6105

Table 5.12: Individual Structural Element Voting Experiment 1. The voted classification results for the first experiment based on voting with each structural element individually. Each structural element is listed in descending F-measure value.

Structures	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
Stemloop	0.4091	0.3154	0.1846	0.0909	0.5937	0.5647	0.8182	0.6682
Hairpin	0.3265	0.1568	0.3432	0.1735	0.6698	0.6756	0.6531	0.6642
Stack	0.4796	0.4666	0.0334	0.0204	0.5130	0.5069	0.9592	0.6632
Unpaired	0.4861	0.4842	0.0158	0.0139	0.5019	0.5010	0.9722	0.6612
Joint-Tail	0.4212	0.3961	0.1039	0.0788	0.5250	0.5153	0.8423	0.6394
Structure	0.3024	0.1753	0.3247	0.1976	0.6271	0.6330	0.6048	0.6186
Loop	0.3469	0.3256	0.1744	0.1531	0.5213	0.5159	0.6939	0.5918
Joint	0.2913	0.2635	0.2365	0.2087	0.5278	0.5251	0.5826	0.5523
External Loop	0.2616	0.2161	0.2839	0.2384	0.5455	0.5476	0.5232	0.5351
Stem	0.2421	0.1642	0.3358	0.2579	0.5779	0.5959	0.4842	0.5343
Internal Loop	0.2106	0.1577	0.3423	0.2894	0.5529	0.5718	0.4212	0.4850
Bulge	0.1048	0.1020	0.3980	0.3952	0.5028	0.5067	0.2096	0.2966
Tail	0.0965	0.0826	0.4174	0.4035	0.5139	0.5389	0.1929	0.2842
Junction	0.0742	0.0204	0.4796	0.4258	0.5538	0.7843	0.1484	0.2496
Bridge	0.0547	0.0232	0.4768	0.4453	0.5315	0.7024	0.1095	0.1894
Multiloop	0.0371	0.0056	0.4944	0.4629	0.5315	0.8696	0.0742	0.1368

Table 5.13: Individual Structural Element Voting Experiment 2. The voted classification results for the second experiment based on voting with each structural element type individually. Each structural element is listed in descending F-measure value.

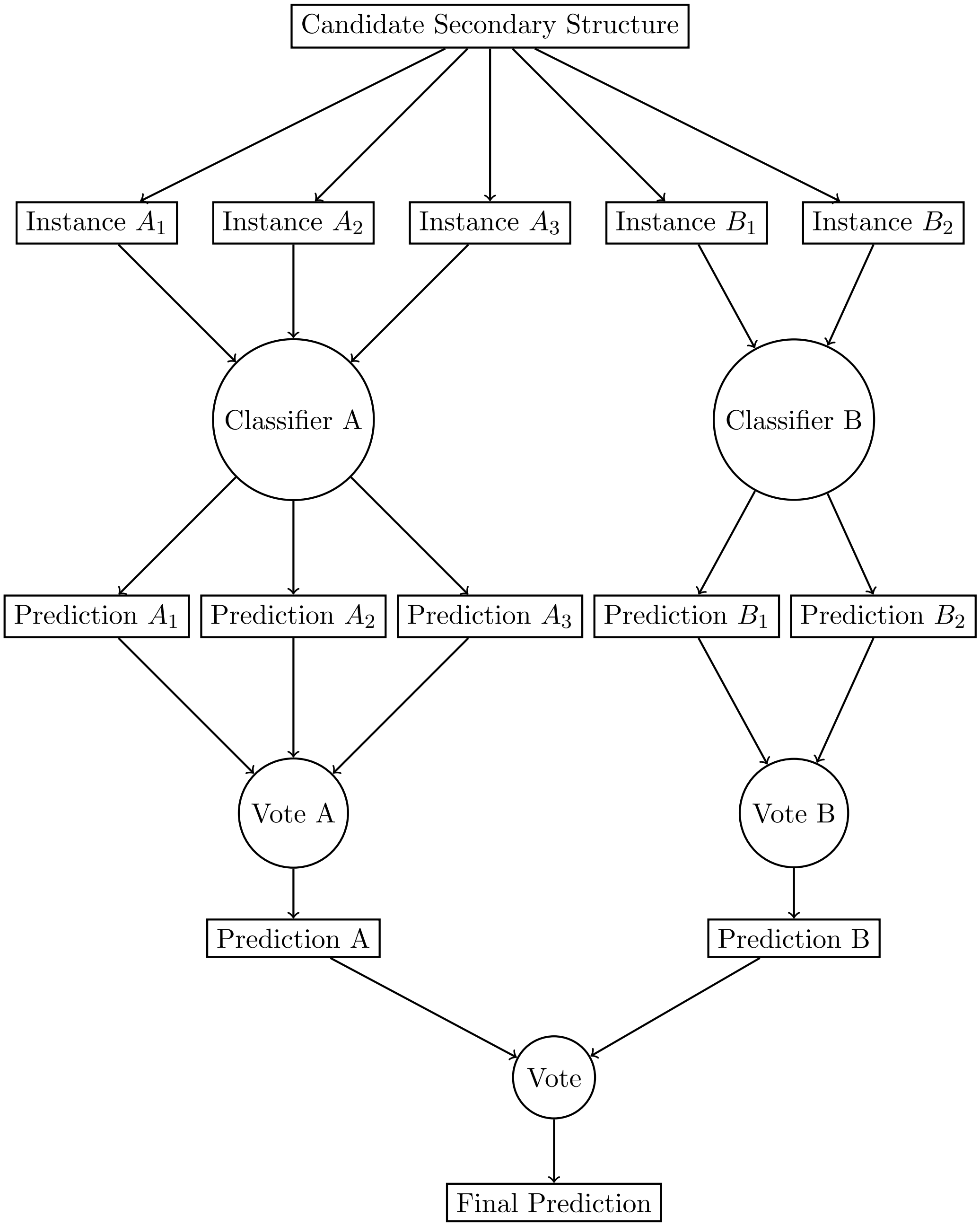

Figure 5.3: Structural Element Paired Voting. For every pair of structural element combinations, all the structural element instances from a candidate secondary structure are collected and process on their respective classifiers. The classifier prediction are first voted among themselves and then the results of that initial vote are voted together to produce a final prediction.

voted individually, such as the *external loop*, *hairpin*, and *structure*. However, the *junction* element, which is not a highly ranked individually voted structural element, also appears a number of times. Since the *junction* element has both *external loop* and *multiloop* element records, pairing it with a strong structural element probably helps compensate for the uncertainty seen in the *junction* class while still making use of the information provided by having both *external loop* data and *junction* data in the voting. Second, high ranking combinations which do not include the *structure* element tend to be made of structural elements that are disjoint. This observation coincides with the fact that disjoint structural elements will tend to add more unique information to the mix. In fact, the highest ranked structural element pair is based on data from the *hairpin* and *junction* structural elements, which are perfectly disjoint. It is worth pointing out that several of the structural element combinations do overlap. For example, the *structure* element appears in three of the five highest ranked structural element pairs, yet the *structure* structural element by definition will overlap any structural element paired with it. Considering the fact that the *structure* structural element ranks well when voted individually and contains mainly general metrics, therefore even though it will overlap its pair, each of the pairs still contributes some unique, more specified information to the prediction engine. Table 5.14 shows the first expectation is also not strictly adhered to. In the table it can be observed that the *junction* structural element is a member of the highest ranked element pair, yet the *junction* structural element ranks third from the bottom in the individual structural element voting. The reason for this unexpected observation is not obvious, as the *junction* structural element is a union of *external loop* and *multiloop* structural elements, which means it incorporates information from the high ranking *external loop* structural element, but also from the low ranking *multiloop* structural element. Furthermore, because the *junction* is an aggregation of records, each *junction* voting group will have one *external loop* and possibly many *multiloop* elements which should overpower the *junction* structural element with the weaker *multiloop* prediction data. Clearly, this is not the case as the junction element appears in the two highest structural element pairs in Table 5.14.

Overall, the first experiment's paired voting shows an improvement over individual voting as the highest ranked vote in the paired set is about 0.07 higher in terms of F-measure than the highest ranked individual structural element voting results. Pointedly, just because the highest ranked structural element pair, *hairpin* and *junction*, does involve a record aggregate structure, it could be seen as really three structural elements being part of the voting process. Nonetheless, even the highest ranked structural element pair which does not contain a record aggregate structure, the *external loop* and *hairpin* pair, still greatly improves on the best individual structural element voting. It has already been discussed that the *external loop* structural element is not an easy structural element to locate in an ad hoc fashion and that the *structure* structural element defeats the purpose of using structural elements for RNA gene finding. So the highest ranked structural element pair in Table 5.14 that is not based on either of these two structural elements is the *bridge* and *hairpin* pair, which ranked 11 receiving a F-measure of 0.8746. The *hairpin* and *multiloop* pair might be another viable combination, ranking slightly lower than the bridge hairpin combination with a F-measure of 0.8703.

The sliding window experiment matched the expected results more than the whole gene experiment. Table 5.15 shows that the highest ranked pairs in the second experiment are made up of structural elements which ranked well when voted individually, as seen in Table 5.13. Like the previous experiments' pair voting results, the top ranking pairs in Table 5.15 do not always match disjoint structural elements, but many of the highest ranked combinations are disjoint with the only exception the broadly based *structure* structural element pairing with the *hairpin* and *stemloop* elements. Since the second experiment deals with a sliding window, less care is needed in terms of which structural elements can be reasonably extracted; however, some structures have greater potential to be detected using ad hoc algorithms. The *structure* element for example requires $O(n^3)$ folding time to extract its metrics, but others like *hairpin*, *stemloop*, *stack*, and *loop* elements can probably be extracted with less processing power than required to fold the whole RNA window. For this reason structure pairs like the *hairpin* and *stack* pair with a F-measure of 0.6858 seem promising. Nonetheless it is important to note that even the highest ranked pair from the second experiment is only 0.0316 points of F-measure improvement when compared to the highest F-measure achieved with single element voting, while the *hairpin* and *stack* pair is only an improvement of 0.0176 F-measure points. These improvements are important, but less than expected considering a whole extra structural element is used in the voting process.

5.3.3 Structural Element Triad Voting

This process of adding another structural element into the voting process is continued with combinations of three structural elements. Table 5.16 shows these top 30 combinations

Elements	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
Hairpin & Junction	0.4604	0.0589	0.4411	0.0396	0.9015	0.8866	0.9208	0.9034
Junction & Structure	0.4585	0.0598	0.4402	0.0415	0.8986	0.8845	0.9170	0.9005
External Loop & Hairpin	0.4575	0.0647	0.4353	0.0425	0.8929	0.8762	0.9151	0.8952
Multiloop & Structure	0.4440	0.0512	0.4488	0.0560	0.8929	0.8967	0.8880	0.8923
External Loop & Structure	0.4672	0.0820	0.4180	0.0328	0.8851	0.8506	0.9344	0.8905
Bridge & Structure	0.4508	0.0724	0.4276	0.0492	0.8784	0.8616	0.9015	0.8811
Junction & Stemloop	0.4459	0.0695	0.4305	0.0541	0.8764	0.8652	0.8919	0.8783
External Loop & Multiloop	0.4459	0.0705	0.4295	0.0541	0.8755	0.8636	0.8919	0.8775
External Loop & Stemloop	0.4846	0.1226	0.3774	0.0154	0.8620	0.7981	0.9691	0.8753
Hairpin & Structure	0.4431	0.0695	0.4305	0.0569	0.8736	0.8644	0.8861	0.8751
Bridge & Hairpin	0.4344	0.0589	0.4411	0.0656	0.8755	0.8806	0.8687	0.8746
External Loop & Junction	0.4469	0.0801	0.4199	0.0531	0.8668	0.8480	0.8938	0.8703
Hairpin & Multiloop	0.4208	0.0463	0.4537	0.0792	0.8745	0.9008	0.8417	0.8703
Stemloop & Structure	0.4266	0.0550	0.4450	0.0734	0.8716	0.8858	0.8533	0.8692
External Loop & Internal Loop	0.4556	0.1014	0.3986	0.0444	0.8542	0.8180	0.9112	0.8621
Hairpin & Stemloop	0.4546	0.1004	0.3996	0.0454	0.8542	0.8191	0.9093	0.8618
Stem & Structure	0.4440	0.0869	0.4131	0.0560	0.8571	0.8364	0.8880	0.8614
Multiloop & Junction	0.4112	0.0454	0.4546	0.0888	0.8658	0.9006	0.8224	0.8597
Junction & Stem	0.4112	0.0463	0.4537	0.0888	0.8649	0.8987	0.8224	0.8589
Structure & Tail	0.4479	0.0965	0.4035	0.0521	0.8514	0.8227	0.8958	0.8577

Table 5.14: Experiment 1's Top 20 Structural Element Voting Pair Statistics. The 20 highest ranked structural element pairs by F-measure. This table is a truncation of Table E.97, which lists every combination of two structural elements voted together for the first experiment.

Elements	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
Hairpin & Structure	0.4119	0.2653	0.2347	0.0881	0.6466	0.6082	0.8237	0.6998
Stemloop & Structure	0.4573	0.3562	0.1438	0.0427	0.6011	0.5621	0.9147	0.6963
Hairpin & Stack	0.4202	0.3052	0.1948	0.0798	0.6150	0.5793	0.8404	0.6858
Stack & Stemloop	0.4917	0.4555	0.0445	0.0083	0.5362	0.5191	0.9833	0.6795
Joint-Tail & Hairpin	0.3952	0.2681	0.2319	0.1048	0.6271	0.5958	0.7904	0.6794
Hairpin & Stemloop	0.4267	0.3340	0.1660	0.0733	0.5928	0.5610	0.8534	0.6770
Loop & Structure	0.4276	0.3358	0.1642	0.0724	0.5918	0.5601	0.8553	0.6769
Internal Loop & Stemloop	0.4416	0.3636	0.1364	0.0584	0.5779	0.5484	0.8831	0.6766
Hairpin & Internal Loop	0.3766	0.2375	0.2625	0.1234	0.6391	0.6133	0.7532	0.6761
Hairpin & Unpaired	0.4276	0.3377	0.1623	0.0724	0.5900	0.5588	0.8553	0.6760
Stack & Structure	0.4731	0.4267	0.0733	0.0269	0.5464	0.5258	0.9462	0.6759
Stemloop & Unpaired	0.4583	0.3980	0.1020	0.0417	0.5603	0.5352	0.9165	0.6758
External Loop & Hairpin	0.4100	0.3043	0.1957	0.0900	0.6058	0.5740	0.8200	0.6753
Bulge & Stemloop	0.4212	0.3265	0.1735	0.0788	0.5946	0.5633	0.8423	0.6751
Loop & Stemloop	0.4712	0.4249	0.0751	0.0288	0.5464	0.5259	0.9425	0.6751
Hairpin & Loop	0.3599	0.2069	0.2931	0.1401	0.6531	0.6350	0.7199	0.6748
External Loop & Stemloop	0.4527	0.3942	0.1058	0.0473	0.5584	0.5345	0.9054	0.6722
Joint-Tail & Stemloop	0.4221	0.3340	0.1660	0.0779	0.5881	0.5583	0.8442	0.6721
Junction & Stemloop	0.4147	0.3219	0.1781	0.0853	0.5928	0.5630	0.8293	0.6707
Hairpin & Stem	0.3340	0.1623	0.3377	0.1660	0.6716	0.6729	0.6679	0.6704

Table 5.15: Experiment 2's Top 20 Structural Element Voting Pair Statistics. The 20 highest ranked structural element pairs by F-measure. This table is a truncation of Table E.98, which lists every combination of two structural elements voted together for the second experiment.

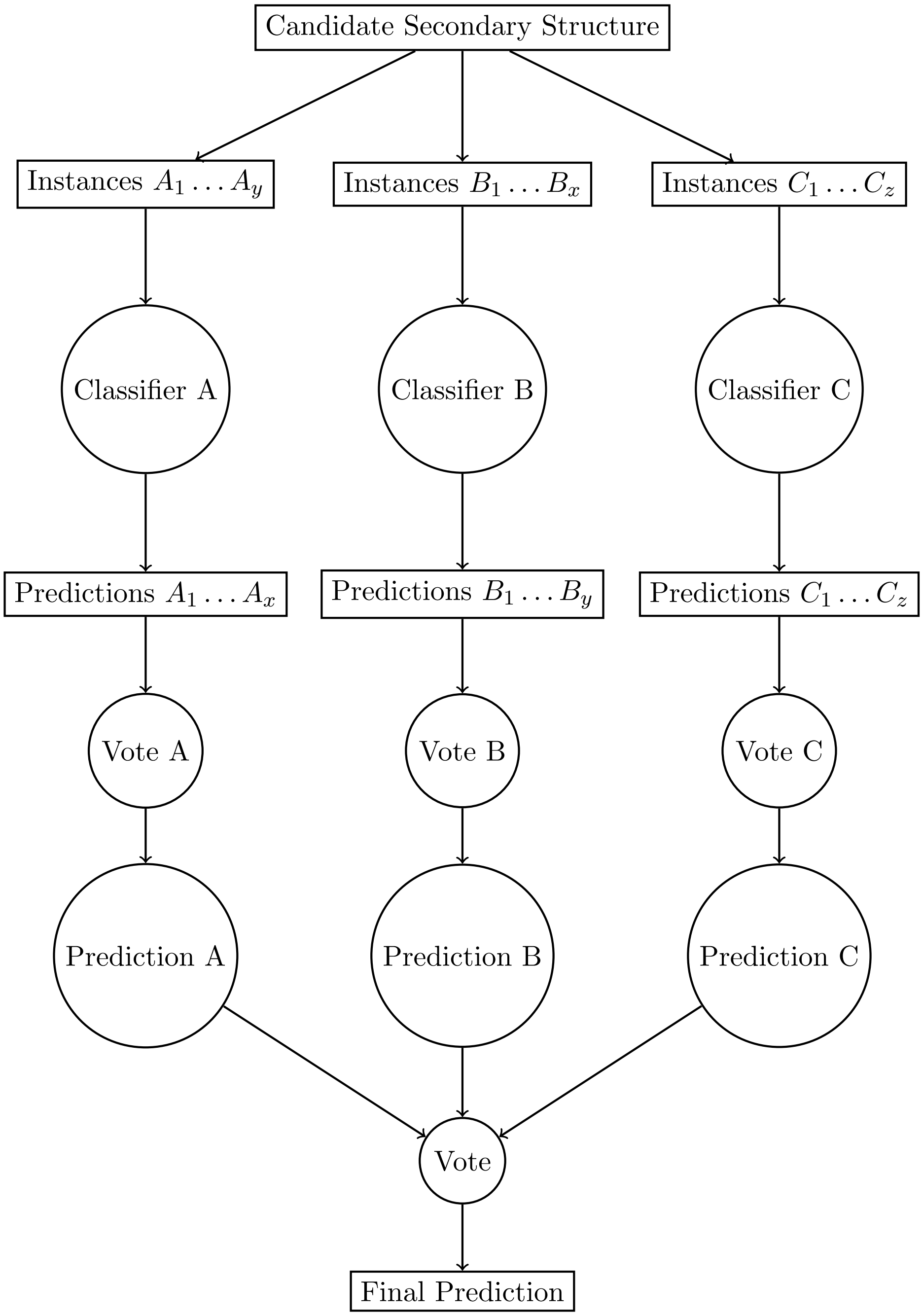
Candidate Secondary Structure
Instances $A_1 \dots A_y$
Instances $B_1 \dots B_x$
Instances $C_1 \dots C_z$
Classifier A
Classifier B
Classifier C
Predictions $A_1 \dots A_x$
Predictions $B_1 \dots B_y$
Predictions $C_1 \dots C_z$
Vote A
Vote B
Vote C
Prediction A
Prediction B
Prediction C
Vote
Final Prediction

Figure 5.4: Structural Element Triad Voting. For every triad of structural element combinations, all the structural element instances from a candidate secondary structure are collected and process on their respective classifiers. The classifier prediction are first voted among themselves and then the results of that initial vote are voted together to produce a final prediction.

of three structural elements for the first experiment, while Table 5.17 portrays the top 20 combinations for the second experiment. The first point which is worth noting about the highest ranked triads in Table 5.16 is that the improvement over the highest ranked pair in Table 5.14 is small, only 0.0064 F-measure. Since the highest ranking triad contains both the *external loop* and the *structure* structural elements it is not a realistic group of structural elements to extract by ad hoc means, as both the *external loop* and *structure* rely to some extent on the RNA segment matching the gene perfectly. Running down the structural element groups in Table 5.16, the highest ranked group which does not include the use of *external loops* or *structures* is the grouping of *bridge*, *hairpin loop*, and *multiloop* which ranked 30 with a F-measure of 0.8985. This feasible group with its extra structural element improved in F-measure by 0.0239 over the top two-element feasible group from the first experiment. Like the results from the first experiment, triad voting in the second experiment resulted in a very meager improvement as seen in Table 5.17. The *hairpin loop*, *stemloop*, and *structure* group only gained 0.0064 F-measure points. As expected, this highest ranked group is simply a union of the two highest ranked pairs in Table 5.17.

This process of adding another structural element and trying all the combinations could continue; however, the number of groups grows rapidly making the computational time impractical. The next section deals with a greedy approach to secondary structural element selection.

5.3.4 Progressive Addition Voting

Table 5.18 shows the classification results for all the classes of RNA genes as structural elements SVMs from the first experiment are progressively added into the set of structural element models used for voting. This table is a summary of the data presented in Appendix E Section E.3. In the table, the models are added in the order they rank in terms of F-measure. So the *external loop* model is added to the voting system which gives the largest increase in F-measure as it is the first model added. The F-measure difference of the system after the model is added is shown in Table 5.18. The *structure* SVM adds over 0.05 F-measure to the system, while the *stemloop* adds over 0.02 and the *hairpin loop* adds 0.01. Interestingly, even though the *tail* SVM achieved an F-measure value of 0.6848, it has a slight degrading effect on the classifier's prediction power. This effect could be simply due to the fact that the *external loop* and the *structure* structural elements already capture the information which allowed the *tail* SVM to be effective alone. With those key biases

Elements	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
External Loop & Junction & Structure	0.4575	0.0483	0.4517	0.0425	0.9093	0.9046	0.9151	0.9098
Hairpin & Junction & Structure	0.4595	0.0512	0.4488	0.0405	0.9083	0.8998	0.9189	0.9093
Hairpin & Internal Loop & Junction	0.4566	0.0483	0.4517	0.0434	0.9083	0.9044	0.9131	0.9087
Hairpin & Multiloop & Junction	0.4363	0.0241	0.4759	0.0637	0.9122	0.9476	0.8726	0.9085
External Loop & Hairpin & Structure	0.4546	0.0483	0.4517	0.0454	0.9064	0.9040	0.9093	0.9066
Hairpin & Junction & Stem	0.4440	0.0357	0.4643	0.0560	0.9083	0.9256	0.8880	0.9064
Multiloop & Junction & Structure	0.4672	0.0637	0.4363	0.0328	0.9035	0.8800	0.9344	0.9064
Hairpin & Junction & Unpaired	0.4527	0.0463	0.4537	0.0473	0.9064	0.9072	0.9054	0.9063
Bridge & Junction & Structure	0.4566	0.0512	0.4488	0.0434	0.9054	0.8992	0.9131	0.9061
Bridge & External Loop & Structure	0.4517	0.0454	0.4546	0.0483	0.9064	0.9087	0.9035	0.9061
Hairpin & Junction & Stemloop	0.4595	0.0550	0.4450	0.0405	0.9044	0.8931	0.9189	0.9058
Bridge & Hairpin & Junction	0.4431	0.0357	0.4643	0.0569	0.9073	0.9254	0.8861	0.9053
Junction & Structure & Tail	0.4508	0.0454	0.4546	0.0492	0.9054	0.9086	0.9015	0.9050
Junction & Stem & Structure	0.4643	0.0618	0.4382	0.0357	0.9025	0.8826	0.9286	0.9050
External Loop & Stemloop & Structure	0.4537	0.0492	0.4508	0.0463	0.9044	0.9021	0.9073	0.9047
Joint-Tail & Hairpin & Junction	0.4575	0.0541	0.4459	0.0425	0.9035	0.8943	0.9151	0.9046
External Loop & Hairpin & Stemloop	0.4604	0.0579	0.4421	0.0396	0.9025	0.8883	0.9208	0.9043
Bridge & Hairpin & Structure	0.4556	0.0521	0.4479	0.0444	0.9035	0.8973	0.9112	0.9042
Hairpin & Junction & Stack	0.4537	0.0512	0.4488	0.0463	0.9025	0.8987	0.9073	0.9030
Bridge & Multiloop & Structure	0.4527	0.0502	0.4498	0.0473	0.9025	0.9002	0.9054	0.9028
External Loop & Structure & Tail	0.4546	0.0531	0.4469	0.0454	0.9015	0.8954	0.9093	0.9023
Hairpin & Loop & Junction	0.4440	0.0405	0.4595	0.0560	0.9035	0.9163	0.8880	0.9020
External Loop & Stem & Structure	0.4537	0.0531	0.4469	0.0463	0.9006	0.8952	0.9073	0.9012
Internal Loop & Junction & Structure	0.4575	0.0579	0.4421	0.0425	0.8996	0.8876	0.9151	0.9011
Junction & Structure & Unpaired	0.4604	0.0627	0.4373	0.0396	0.8977	0.8801	0.9208	0.9000
External Loop & Hairpin & Multiloop	0.4633	0.0666	0.4334	0.0367	0.8967	0.8743	0.9266	0.8997
Hairpin & Multiloop & Structure	0.4440	0.0434	0.4566	0.0560	0.9006	0.9109	0.8880	0.8993
External Loop & Multiloop & Structure	0.4373	0.0357	0.4643	0.0627	0.9015	0.9245	0.8745	0.8988
External Loop & Hairpin & Internal Loop	0.4710	0.0772	0.4228	0.0290	0.8938	0.8592	0.9421	0.8987
Bridge & Hairpin & Multiloop	0.4402	0.0396	0.4604	0.0598	0.9006	0.9175	0.8803	0.8985

Table 5.16: Top 30 Structural Element Triple Voting Statistics For Experiment 1. Of the 560 possible combinations of structural elements, this table lists the top 30 triads which ranked highest by F-measure.

Elements	**TP**	**FP**	**TN**	**FN**	**Acc.**	**Prec.**	**Recall**	**F-mea.**
Hairpin & Stemloop & Structure	0.3980	0.2291	0.2709	0.1020	0.6688	0.6346	0.7959	0.7062
Hairpin & Stem & Structure	0.4091	0.2551	0.2449	0.0909	0.6540	0.6159	0.8182	0.7028
Hairpin & Multiloop & Structure	0.4156	0.2672	0.2328	0.0844	0.6484	0.6087	0.8312	0.7027
Hairpin & Junction & Structure	0.4165	0.2690	0.2310	0.0835	0.6475	0.6076	0.8330	0.7027
Hairpin & Loop & Structure	0.4054	0.2495	0.2505	0.0946	0.6558	0.6190	0.8108	0.7020
Hairpin & Structure & Unpaired	0.4119	0.2635	0.2365	0.0881	0.6484	0.6099	0.8237	0.7009
Bridge & Hairpin & Structure	0.4174	0.2746	0.2254	0.0826	0.6429	0.6032	0.8349	0.7004
Junction & Stemloop & Structure	0.4499	0.3349	0.1651	0.0501	0.6150	0.5733	0.8998	0.7004
Hairpin & Stack & Structure	0.4174	0.2755	0.2245	0.0826	0.6419	0.6024	0.8349	0.6998
Joint-Tail & Stemloop & Structure	0.4406	0.3191	0.1809	0.0594	0.6215	0.5800	0.8813	0.6996
Hairpin & Joint & Structure	0.4091	0.2616	0.2384	0.0909	0.6475	0.6100	0.8182	0.6989
Joint-Tail & Hairpin & Structure	0.4100	0.2635	0.2365	0.0900	0.6466	0.6088	0.8200	0.6988
Hairpin & Internal Loop & Structure	0.4174	0.2774	0.2226	0.0826	0.6401	0.6008	0.8349	0.6988
Bulge & Hairpin & Structure	0.4314	0.3033	0.1967	0.0686	0.6280	0.5871	0.8627	0.6987
Multiloop & Stemloop & Structure	0.4471	0.3330	0.1670	0.0529	0.6141	0.5731	0.8942	0.6986
Stem & Stemloop & Structure	0.4564	0.3506	0.1494	0.0436	0.6058	0.5655	0.9128	0.6984
Bulge & Stemloop & Structure	0.4638	0.3646	0.1354	0.0362	0.5993	0.5599	0.9276	0.6983
Hairpin & Structure & Tail	0.4156	0.2755	0.2245	0.0844	0.6401	0.6013	0.8312	0.6978
Stack & Stemloop & Structure	0.4527	0.3469	0.1531	0.0473	0.6058	0.5661	0.9054	0.6966
Stemloop & Structure & Unpaired	0.4555	0.3525	0.1475	0.0445	0.6030	0.5637	0.9109	0.6965

Table 5.17: Top 20 Structural Element Triple Voting Statistics For Experiment 2. Of the 560 possible combinations of structural elements, this table lists the top 20 triads which ranked highest by F-measure.

already contributed by the other structural element models, there is little left for the *tail* element to contribute but noise. Likewise, it was expected that because the *hairpin* model was added after the *stemloop* model, that it would have little to no positive effect on the voting results, yet it was able to produce a small improvement, which shows that it can help influence the decision making process. This positive influence could simply be due to more information being added to the system which was not available in the *external loop*, *structure*, or *stemloop* models or could be due to more weight being placed upon the *stemloop* and *hairpin* information. Although many of the remaining structural element models did have some positive effect on the prediction power of the voting model, the improvement is meager. The combination of these structural element SVMs improved the F-measure by less than 0.01. When voting with the exclusion of all the models which negatively affected the classification results, a small gain in F-measure of 0.011 is obtained, as seen in Table 5.19. The F-measure value, 0.9454, achieved by combining only the structural elements which produced a positive gain in F-measure in Table 5.18 is the highest of all the model groups considered.

In the same fashion the progressive voting is accomplished with the models from the second experiment. The summary of the results of this series of voting trials is shown in Table 5.20, while the raw results are listed in detail in Appendix E Section E.3. When only the first three highest ranked SVM models (*structure*, *stemloop*, and *hairpin*) are used, the highest F-measure seen in Table 5.20 is obtained. This value of 0.7062 is a little bit higher than using all the models combined. There are several models which negatively contribute to the results of the voted predictions: *external loop*, *joint-tail*, *joint*, *stack*, *bulge*, and *tail*. Considering that these models achieved low F-measure of around 0.5 or less, the fact that they negatively affect the voting is not surprising. Excluding the top three, the other structural element models which did positively contribute to the prediction accuracy had a combined impact on the F-measure of a little under 0.015. When the models which negatively affect the F-measure are removed from the set of models and the remaining models are used in the voting process, an increase in F-measure of 0.0169 was gained. These prediction statistics are presented in Table 5.21. When compared to the F-measure in Table 5.20 only an increase of 0.0084 is achieved. Such a small gain is hardly worth the cost of the 7 extra models needed to be processed. In conclusion, although a slight improvement can be gained by using additional structural elements only the *structure*, *stemloop*, and *hairpin* models are needed to achieve a F-measure of over 0.7 for the sliding window experiment.

Structures	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.	F-mea. Inc.
External Loop	0.4010	0.0541	0.4454	0.0995	0.8464	0.8811	0.8012	0.8392	0.8392
Structure	0.4672	0.0820	0.4180	0.0328	0.8851	0.8506	0.9344	0.8905	0.0513
Stemloop	0.4575	0.0463	0.4537	0.0425	0.9112	0.9080	0.9151	0.9115	0.0210
Hairpin Loop	0.4595	0.0328	0.4672	0.0405	0.9266	0.9333	0.9189	0.9261	0.0146
Tail	0.4566	0.0319	0.4681	0.0434	0.9247	0.9348	0.9131	0.9238	-0.0023
Joint	0.4508	0.0251	0.4749	0.0492	0.9257	0.9473	0.9015	0.9238	0.0000
Joint-Tail	0.4575	0.0319	0.4681	0.0425	0.9257	0.9349	0.9151	0.9249	0.0011
Bridge	0.4527	0.0280	0.4720	0.0473	0.9247	0.9418	0.9054	0.9232	-0.0017
Stem	0.4537	0.0270	0.4730	0.0463	0.9266	0.9438	0.9073	0.9252	0.0020
Stack	0.4546	0.0280	0.4720	0.0454	0.9266	0.9420	0.9093	0.9253	0.0001
Multiloop	0.4527	0.0241	0.4759	0.0473	0.9286	0.9494	0.9054	0.9269	0.0016
Junction	0.4604	0.0309	0.4691	0.0396	0.9295	0.9371	0.9208	0.9289	0.0020
Unpaired	0.4614	0.0309	0.4691	0.0386	0.9305	0.9373	0.9228	0.9300	0.0011
Internal Loop	0.4488	0.0145	0.4855	0.0512	0.9344	0.9688	0.8977	0.9319	0.0019
Loop	0.4517	0.0164	0.4836	0.0483	0.9353	0.9649	0.9035	0.9332	0.0013
Bulge	0.4537	0.0174	0.4826	0.0463	0.9363	0.9631	0.9073	0.9344	0.0012

Table 5.18: Progressive Voting for Experiment 1. Each structural element is added to the voting group one by one and the resulting prediction statistics are recorded in this table along with the difference in F-measure.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16SrRNA	0.4854	0.0182	0.4818	0.0146	0.9672	0.9638	0.9708	0.9673
23S rRNA	0.4722	0.0000	0.5000	0.0278	0.9722	1.0000	0.9444	0.9714
5S rRNA	0.5000	0.0156	0.4844	0.0000	0.9844	0.9697	1.0000	0.9846
RNase P	0.4945	0.0110	0.4890	0.0055	0.9835	0.9783	0.9890	0.9836
SRP RNA	0.4936	0.0513	0.4487	0.0064	0.9423	0.9059	0.9872	0.9448
TmRNA	0.4055	0.0276	0.4724	0.0945	0.8780	0.9364	0.8110	0.8692
tRNA	0.4706	0.0000	0.5000	0.0294	0.9706	1.0000	0.9412	0.9697
All	0.4681	0.0222	0.4778	0.0319	0.9459	0.9547	0.9363	0.9454

Table 5.19: Only Positive Contributing Structural Element Voting Statistics for Experiment 1. The voting statistics for the structural elements which produced a positive gain in F-measure in Table 5.18. This group of elements includes: Bulge, External Loop, Joint-Tail, Hairpin Loop, Internal Loop, Joint, Loop, Multiloop, Junction, Stack, Stem, Stemloop, Structure, and Unpaired.

5.4 Reducing False Positives

As it was already mentioned in Chapter 1, reducing false positives is crucial to RNA gene finding because it is very expensive for researchers to investigate predicted RNA genes which do not end up being real RNA genes. The SRNAG finder presented in this thesis can be adjusted to reduced the number of false positives by only labeling a candidate sequence a SRNAG if the vote for it being a SRNAG outweighs the vote against it being a SRNAG by a margin set with a parameter. As this margin becomes larger, the system will only be able to classify sequences as SRNAGs if they are highly probable to be so, classifying sequences which it is uncertain about as non-SRNAGs. Increasing this margin will have the effect of reducing false positives, but also increasing the number of false negatives.

Although this method of reducing false positives can be used for any set of structural elements, for brevity, only the data from Table 5.19 and Table 5.21 will be presented in this thesis as they achieved some of the strongest prediction results. Table 5.22 tabulates the classification results when a low false positive margin is used in the voting process for the first experiment. This table shows a false positive rate of just under 0.005 for the whole group of gene classes. Even with this low false positive rate, expanding the cutoff margin only drove the false negative rate up to 0.1139. Hence under favorable conditions the classifier is able to achieve a very low false positive rate, while still maintaining a reasonable accuracy of 88% by correctly labeling nearly all the non-SRNAG segments and just over 38% of the SRNAG segments.

Structures	**TP**	**FP**	TN	**FN**	**Acc.**	**Prec.**	**Recall**	**F-mea.**	**F-mea. Inc.**
Structure	0.3024	0.1753	0.3247	0.1976	0.6271	0.6330	0.6048	0.6186	0.6186
Stemloop	0.4434	0.3312	0.1688	0.0566	0.6122	0.5725	0.8868	0.6958	0.0772
Hairpin	0.3980	0.2291	0.2709	0.1020	0.6688	0.6346	0.7959	0.7062	0.0104
External Loop	0.4276	0.3052	0.1948	0.0724	0.6224	0.5835	0.8553	0.6938	-0.0124
Joint-Tail	0.4425	0.3340	0.1660	0.0575	0.6085	0.5699	0.8850	0.6933	-0.0005
Joint	0.4063	0.2746	0.2254	0.0937	0.6317	0.5967	0.8126	0.6881	-0.0052
Stack	0.3776	0.2217	0.2783	0.1224	0.6558	0.6300	0.7551	0.6869	-0.0012
Loop	0.3738	0.2078	0.2922	0.1262	0.6660	0.6427	0.7477	0.6913	0.0044
Unpaired	0.3961	0.2486	0.2514	0.1039	0.6475	0.6144	0.7922	0.6921	0.0008
Stem	0.4137	0.2783	0.2217	0.0863	0.6354	0.5979	0.8275	0.6942	0.0021
Internal Loop	0.3915	0.2338	0.2662	0.1085	0.6577	0.6261	0.7829	0.6958	0.0016
Bridge	0.3915	0.2319	0.2681	0.1085	0.6596	0.6280	0.7829	0.6969	0.0011
Bulge	0.4100	0.2681	0.2319	0.0900	0.6419	0.6047	0.8200	0.6961	-0.0008
Multiloop	0.4128	0.2690	0.2310	0.0872	0.6438	0.6054	0.8256	0.6986	0.0025
Junction	0.4054	0.2514	0.2486	0.0946	0.6540	0.6172	0.8108	0.7009	0.0023
Tail	0.4174	0.2792	0.2208	0.0826	0.6382	0.5992	0.8349	0.6977	-0.0032

Table 5.20: Progressive Voting for Experiment 2. Each structural element is added to the voting group, one by one, and the resulting prediction statistics are recorded in this table along with the difference in F-measure.

Gene Type	**TP**	**FP**	**TN**	**FN**	**Acc.**	**Prec.**	**Recall**	**F-mea.**
16S rRNA	0.4276	0.2448	0.2552	0.0724	0.6828	0.6359	0.8552	0.7294
23S rRNA	0.3902	0.2805	0.2195	0.1098	0.6098	0.5818	0.7805	0.6667
5S rRNA	0.4677	0.1774	0.3226	0.0323	0.7903	0.7250	0.9355	0.8169
RNase P	0.4574	0.2553	0.2447	0.0426	0.7021	0.6418	0.9149	0.7544
SRP RNA	0.3987	0.2278	0.2722	0.1013	0.6709	0.6364	0.7975	0.7079
TmRNA	0.4138	0.3172	0.1828	0.0862	0.5966	0.5660	0.8276	0.6723
tRNA	0.3750	0.2500	0.2500	0.1250	0.6250	0.6000	0.7500	0.6667
All	0.4239	0.2625	0.2375	0.0761	0.6614	0.6176	0.8479	0.7146

Table 5.21: Only Positive Contributing Structural Element Voting Statistics for Experiment 2. The voting statistics for the structural elements which produced a positive gain in F-measure in Table 5.20. This group of elements includes: Structure, Stemloop, Hairpin, Loop, Unpaired, Stem, Internal Loop, Bridge, Multiloop, and Junction.

Gene Type	**TP**	**FP**	**TN**	**FN**	**Acc.**	**Prec.**	**Recall**	**F-mea.**
16S rRNA	0.3869	0.0036	0.4964	0.1131	0.8832	0.9907	0.7737	0.8689
23S rRNA	0.4583	0.0000	0.5000	0.0417	0.9583	1.0000	0.9167	0.9565
5S rRNA	0.4375	0.0000	0.5000	0.0625	0.9375	1.0000	0.8750	0.9333
RNase P	0.4615	0.0055	0.4945	0.0385	0.9560	0.9882	0.9231	0.9545
SRP RNA	0.4295	0.0192	0.4808	0.0705	0.9103	0.9571	0.8590	0.9054
TmRNA	0.2677	0.0000	0.5000	0.2323	0.7677	1.0000	0.5354	0.6974
tRNA	0.4118	0.0000	0.5000	0.0882	0.9118	1.0000	0.8235	0.9032
All	0.3861	0.0048	0.4952	0.1139	0.8813	0.9877	0.7722	0.8667

Table 5.22: Low False Positive Prediction Statistics for Experiment 1. This table contains the same classification as Table 5.19 except an increased certainty margin is used to reduce the number of false positives to below half a percent.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.0793	0.0034	0.4966	0.4207	0.5759	0.9583	0.1586	0.2722
23S rRNA	0.0244	0.0000	0.5000	0.4756	0.5244	1.0000	0.0488	0.0930
5S rRNA	0.1774	0.0000	0.5000	0.3226	0.6774	1.0000	0.3548	0.5238
RNase P	0.1117	0.0106	0.4894	0.3883	0.6011	0.9130	0.2234	0.3590
SRP RNA	0.0190	0.0063	0.4937	0.4810	0.5127	0.7500	0.0380	0.0723
TmRNA	0.0655	0.0034	0.4966	0.4345	0.5621	0.9500	0.1310	0.2303
tRNA	0.1250	0.0000	0.5000	0.3750	0.6250	1.0000	0.2500	0.4000
All	0.0742	0.0046	0.4954	0.4258	0.5696	0.9412	0.1484	0.2564

Table 5.23: Low False Positive Prediction Statistics for Experiment 2. This table contains the same classification as Table 5.21 except an increased certainty margin is used to reduce the number of false positives to below half a percent.

Table 5.23 portrays a similar story for the second experiment where an increased cutoff margin was used. The false positive rate is just below 0.005, while the false negative rate has risen to 0.4258. This table shows that under harsh conditions where the genome background base composition is the same as the SRNAG composition, the SRNAG finder can classify the sequence windows just under 57% of the time, while maintaining a low false positive rate. Any SRNAGs segments found by this classifier would most likely be genuine SRNAGs. Nonetheless, with such a high false negative rate, the classifier would tend to miss most of the genes, only finding about 7%.

5.5 Comparison to Other SRNAG Finders

It is hard to compare these results to the *ab initio* SRNAG finding techniques discussed in Chapter 2 as many of the authors did not report statistics for the number of SRNAGs they were able to locate. An exception to this is in the paper describing RNAGENiE where Carter reported the classifier's prediction statistics. Carter's experiment methodology has a number of similarities to the methodology used to test the SRNAG finder described in this thesis. These similarities detailed in Chapter 2 include using an 80 nucleotide sliding window, several of the same metrics, and machine learning algorithms for classification. Because of these similarities and the fact that Carter reports the classification statistics, a comparison of the SRNAG finder presented in this thesis to RNAGENiE can be made.

Even though the highest prediction accuracy achieved by the SRNAG finder presented in this thesis was 95%, since this accuracy was achieved under favorable conditions it is unfair to use in the comparision. Hence, the highest sliding window accuracy achieved will be used in the comparison with RNAGENiE. With its sliding window dataset RNAGENiE was reported to have had an overall prediction accuracy of 92% [9], while the SRNAG finder presented in this thesis only scored a 67% accuracy on its own dataset. This 67% accuracy was achieved using the *structure*, *stemloop*, and *hairpin* elements together in a voting group. Even though RNAGENiE was reported to produce a higher prediction accuracy, the comparison is not perfectly fair as RNAGENiE utilized motif data and their negative dataset was created from genome NC regions. The negative dataset, used to test the SRNAG finder in this thesis, was created from shuffled SRNAG sequences so that no base composition bias could be exploited. Clearly this new SRNAG finder had a huge disadvantage as RNAGENiE's base composition inputs alone achieved a prediction accuracy greater than 85% [9]. Therefore, in light of what appear to be poor prediction results for the SRNAG finder in this thesis when it was processing the windows, it needs to be remembered that all the results in this thesis are in the absence of base composition bias. This absence of base composition bias means that any of the results discovered in this thesis should be applicable in any genome of any base composition background.

5.6 Chapter Review

This chapter presented the results for the whole SRNAG and sliding window experiments in parallel. The tests of the SRNAG finder under favorable conditions proved successful, revealing not only that the software was working correctly, but that the concepts utilized by SRNAG finder worked as expected. By analyzing each of the structural elements individually (without voting), it was seen that in the first experiment the *external loop*, *structure*, *stemloop*, *hairpin*, and *tail* structural elements were good candidates for producing strong SRNAG finding signals. Upon closer inspection it was found that of these five structural elements, the *external loop* and *tail* structural elements would be difficult to locate using ad hoc algorithms and the *structure* structural element would not allow the $O(n^3)$ algorithmic complexity to be addressed. This leaves the *stemloop* and *hairpin* elements as the most likely structural elements to be useful for SRNAG finding. Specifically it was found that the spacing of the *stemloops* and *hairpins* produced a strong signal for SRNAGs, confirming Noël's hypothesis in a rigorous way.

From the second experiment it was revealed that the sliding window had a negative effect on the structural element's ability to classify SRNAG and non-SRNAG sequences. The only two structural elements which stood out were the *structure* and *stemloop* elements, but even these structural elements did not achieve results as strong as the first experiment.

The results from the numerous voting experiments showed that the voting mechanism proposed in this thesis allowed several weaker structural element models to be combined into a single classifier with better prediction accuracy. Voting with a single structural element type proved to be quite successful in both experiments. As the size of the voting group increased, the gains produced by the additional structural element data shrank and the time required to test successively larger group sizes, drastically increased. As a result, a greedy structural element selection system was used to combat this growing time complexity which revealed that some of the structural elements caused a decrease in classifier performance. When only the structural element models which produced F-measure gains in the greedy selection test were combined, the highest F-measure value for the first experiment was achieved.

A mechanism for reducing false positives was introduced and tested. This mechanism used a parameter to increase the certainty required by the SRNAG finder to label a SRNAG. This mechanism was shown to work, reducing the false positive rate for both experiments to less than 0.005. Even with this reduction in false positives the first experiment's classification system was still able to locate many of the genes; however, this very low false positive rate caused the second experiment's classification system to mis-classify many SRNAGs as nongenes, drastically reducing the true positive rate to 7%.

Lastly, a comparison to RNAGENiE is made. Although RNAGENiE was able to beat the prediction accuracy of the SRNAG finder presented in this thesis, since the same dataset was not used by both RNA gene finders for this comparison, it reveals little about their true relative performance.

The next chapter draws some conclusions from the results presented in this chapter and outlines the direction this research will lead to in the future.

Chapter 6

Conclusion

This thesis has introduced and examined a new SRNAG finder, which uses structural elements to generate SRNAG detection signals. This method involved folding candidate sequences and extracting metrics from their secondary structure components. These metrics were then utilized by SVMs to classify whether the secondary structural elements had come from a SRNAG sequence or not. Since these individual structural element models are often weak predictors, voting was used to strengthen the prediction accuracy of the system.

Although this SRNAG finder was built and tested, its real purpose was to explore some of the underlying concepts which allowed it to function. Particularly it allowed each structural element's ability to produce a signal for SRNAGs to be investigated and the key properties of those secondary structure elements to be discovered. This investigation revealed that under favorable conditions the *external loop*, *structure*, *stemloop*, *hairpin*, and *tail* structural elements produced strong SRNAG signals. Unfortunately, several of these secondary structures are unsuitable for ad hoc detection which is required so that the $O(n^3)$ running time of contemporary *ab initio* SRNAG finders could be addressed. However, the *stemloop* and *hairpin* have been demonstrated to be detectable with some level of accuracy in a running time less than $O(n^3)$, making them prospective structural elements for use in SRNAG finding.

When artificial mini genomes were developed and a sliding window was used to select genome regions for evaluation, the sliding window would break up the hidden SRNAGs, destroying the structural elements' ability to classify the SRNAGs. However, the fixed window size allowed *structure*'s MFE metric to produce a SRNAG signal. MFE was already known to have some ability to produce a SRNAG signal, so this is no surprise, yet the MFE metric requires folding the sequence window and hence is not able to address the $O(n^3)$ run time. Although weak compared to the first experiment, the *stemloop* in the second experiment also produced a SRNAG signal. The *stemloop* metric which contributed most to this signal was the average stack size where it was found that SRNAGs tended to have larger stacks within their *stemloops*.

The experiment results revealed that the voting mechanism was able to significantly improve the results of the individual structural element SVM models. Keeping complexity at a minimum is important for any system, so using the least number of structural elements in the SRNAG finder helps control complexity. Beginning with using only data records produced for a sequence from a single structural element type it was easy to see the gains that could be accomplished through voting even without using more than a single structural element for classification. Next, every combination of two structural elements and then every combination of three was tested. As more structural elements were added to the SRNAG finding system, the prediction results rose making use of the additional information. However, even after adding only three structural elements to the system, the gains in prediction power from adding an additional structural element became insignificant. Later a greedy secondary structure model selection scheme was tried which added the voting power of each structural element iteratively. This greedy selection scheme revealed that some of the structural elements added, weaken the prediction results of the SRNAG finder and that only the first couple of structural elements contributed significant information into the system. Given the added complexity for diminishing returns, at most two to three structural elements should be used for SRNAG finding.

As already mentioned, when classifying between whole SRNAGs and random RNA, a group consisting of the *external loop*, *structure*, and *stemloop* structural elements excels, achieving a F-measure over 0.91. Since the *external loop* and *structure* elements can not be used to address the $O(n^3)$ run time, the favorable conditions experiments indicates that the *stemloop* and *hairpin* elements make a feasible, yet powerful structural element group, having a F-measure over 0.86. The sliding window experiment with its fixed length window allowed the *structure*, *stemloop*, and *hairpin* group to outrank the rest. Once again since this group contains the *structure* structural element which is relying on MFE for its prediction power, it can not address the $O(n^3)$ run time needed to calculate MFE. The next possible structural element which could be exchanged for the *structure* is the *loop*. The *loop* is a weak structural element model which does not contribute much to the prediction results.

Therefore, the sliding window experiment also revealed that a group consisting of *stemloops* and *hairpins* is a strong candidate group for SRNAG finding.

The last test performed in this thesis was to test the mechanism for reducing false positives. This mechanism utilized a parameter to increase the certainty needed by the SRNAG finder before it could label a candidate sequence as a SRNAG. Applied to the experiment which used the favorable conditions dataset, this mechanism drastically reduced the number of false positives while still maintaining to correctly classify a reasonable number of SRNAGs. The mechanism also worked in the second experiment, reducing the false positive rate to less than 0.005; however, without the support of a strong classification models the classifier tended to label nearly every candidate sequence as a non-SRNAG sequence. The results of the second experiment are not a failing of the mechanism, but reveal that a strong classification engine is needed to reduce false positives while still maintaining a high detection rate.

6.1 Future Work

This thesis laid the ground work to a more ambitious *ab-initio* SRNAG finding project. Designed to explore many different structural elements and many different metrics of those structural elements, this project saw which of them produce strong SRNAG signals. This project successfully achieved that goal, demonstrating the predictive potential of the *structure*, *stemloop*, and *hairpin* structural elements in the absence of base composition bias. The structural elements not included in this group have been shown to be useless or infeasible for the task of SRNAG finding and will not be explored further.

Although this research did not directly address the $O(n^3)$ running time of existing *ab initio* SRNAG finders, computational complexity required for structural element extraction was a requirement which was constantly referred to when choosing structural elements with SRNAG finding ability. In the end only the *stemloop* and *hairpin* elements produced SRNAG signals and are feasible to be detected without folding. Fortunately, Noël, already devised a *stemloop* finding algorithm, *Wave*, which runs in $O(n^2)$; however, as Noël demonstrated in his thesis *Wave* is not accurate enough to produce strong SRNAG finding signals from the *stemloops* it locates. *Wave* always finds the largest *stemloops* and allows for *internal loops* and *bulges* which most likely leads to this poor detection accuracy. If the accuracy of *Wave* could be improved, the results found in this thesis indicate it could possibly be used to produce SRNAG signals by exploiting *stemloops* spacing metrics in genomes of any background composition. Likewise, a simpler version of *Wave* which only detects hairpins could have similar potential.

The key metrics discovered by this thesis could be implemented in any SRNAG finder which makes use of MFE or requires folding the RNA sequence for metric extraction. For example, RNAGENiE could easily be modified to incorporate *stemloop* and *hairpin* spacing as an additional metric since it already uses MFE as a metric. In the same way, it would be interesting to add motif analysis into the SRNAG finder proposed in this thesis as they were able to produce a strong signal in RNAGENiE.

Working within the scope of this thesis project, it would be worth testing the SRNAG finder with a larger sliding window. It was already discussed that a sliding window of 80 nucleotides is not large enough to get an accurate measure of *stemloop* spacing; as on average 80 nucleotides is not large enough to contain two full *stemloops*. Likewise, rerunning the experiment using NC RNA sequences from real genomes, instead of shuffled SRNAGs for the negative example set would be interesting as it would demonstrate the SRNAG finders ability to classify SRNAGs in the presence of base composition. Finally, since Carter showed much success with NNs, it might be worth replacing the SVMs with NNs.

Appendix A

Statistics

This appendix gives a brief overview of the statistics utilized in this thesis. It is broken down into two main sections. In the first section the statistics which are used to compare structural element metrics are discussed, while the second section details the metrics which are used to compare the results from the classifier tests.

A.1 Metric Statistics

The two statistical techniques discussed in this thesis for comparing RNA secondary structure features are z-scores and f-scores. Z-scores, while not utilized by the SRNAG finding method presented in this thesis are a technique used by other SRNAG finders to produce SRNAG signals from the SRNAG features. On the other hand, F-scores are utilized by the methods in this thesis to rank the structural element metrics tested, making it easy to spot which metrics are likely to be effective in distinguishing between SRNAG and non-SRNAG sequences.

A.1.1 Z-scores

In order to justify the statistical significance of a metric value a z-score (standard score) can be calculated, which shows the number of units of standard deviations an observed value is away form the mean random score [31], allowing the comparison of observations from different normal distributions. The z-score formula follows,

$$z = \frac{x - \mu}{s}, \tag{A.1}$$

where z is the z-score, x is the raw observation to be standardized, μ is the mean observation for a population, and s is the standard deviation of observation for the population. It is important to note that the z-score is dependent on the population sample size from which μ and s are calculated, as the variation in the z-score tends to decrease as the sample size increases [31]. Scores with very negative or positive values tend to be significant as they are outside the probable region of the normal distribution [12].

A.1.2 F-score

F-score gives a measure of the discrimination of two sets of values [13]. In other words, F-score measures the amount of overlap between distributions of two datasets. This measure is particularly useful in classification as it provides a metric for indicating the strength of a feature in distinguishing between two classes of data. Higher F-score values indicate more disjoint feature data values and hence that the feature is useful for classification [13]. The following equation defines F-score:

$$f = \frac{\left(\bar{x}^{(+)} - \bar{x})^2 + (\bar{x}^{(-)} - \bar{x}\right)^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}\left(x_k^{(+)} - \bar{x}^{(+)}\right)^2 + \frac{1}{n_--1}\sum_{k=1}^{n_-}\left(x_k^{(-)} - \bar{x}^{(-)}\right)^2}, \tag{A.2}$$

where $\bar{x}$ is the average for the whole dataset and $\bar{x}^+$ and $\bar{x}^-$ are the averages of the different classes in the dataset; $x_k^{(+)}$ and $x_k^{(-)}$ are the k^{th} examples for the two classes in the dataset; and n_+ and n_- are the number of examples for each of the classes in the whole dataset [13].

One major disadvantage of relying on F-score as a measure of the discrimination of a metric is that F-score does not account for mutual information between features [13]. Figure A.1 demonstrates this problem, where a dataset with two features is plotted. The points of each class are easily partitionable when both features are considered together, yet this data produces low F-scores when each of the features are considered independently. So while F-score is a simple and generally effective means to measure the potential of a metric in classification, this measure is not definitive [13].

A.2 Classifier Statistics

There are four possible outcomes when a classifier predicts the class of a candidate instance: true positive (TP), false positive (FP), true negative (TN), and false negative (FN) [22]. In

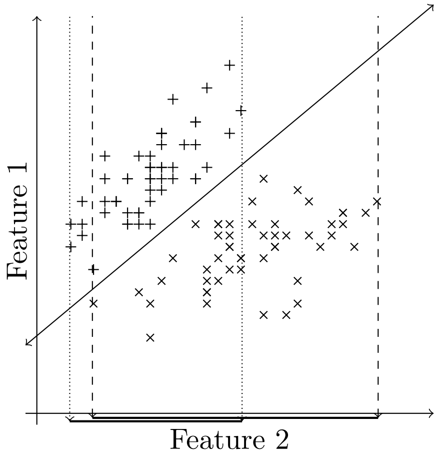

Figure A.1: Mutual Information Among Features. This figure shows a dataset with two classes which are completely disjoint when both features are considered together; however, when evaluated individually using a F-score the resulting value indicates low discrimination of the data classes. The dashed and doted lines reveal the range of Feature 2 values for the data point of each class and the solid line shows that the class data is completely disjoint. This figure is based on a figure in [13].

order to explain these four categories, assume a two class dataset exists with one class being positive and the other class being negative. If the actual class of the instance is positive and the classifier predicts positive then it falls in the TP category, while if the actual class is positive and the classifier predicts the instance to be a negative then it is a FN. Likewise, if the predicted class is negative and the actual class is negative then it is a TN, while if the real class is a negative and the classifier predicts the example to be a positive then its a FP. While these four categories of prediction results are informative on their own, they can reveal more information about the performance of the classifier by combining them to calculate accuracy, precision, recall, and f-measure.

A.2.1 Accuracy

Accuracy is simply a measure of the number of instances the classifier correctly predicted over the total number of instances attempted [22]. The following equation defines accuracy in terms of TP, FP, TN, and FN:

$$a = \frac{TP + TN}{TP + TN + FP + FN} \tag{A.3}$$

One disadvantage of using accuracy as a measure of classifier performance is that accuracy does not take into account the underlying distribution of positive and negative cases which occur in the testing dataset. This means that if you have a test dataset with 90% of the instances of one class and the rest of the other class, the classifier could easily achieve a 90% accuracy by classifying all the instances with the majority class label. Clearly, in a case where there is a strong bias in class label in the testing dataset, accuracy reveals little about the true performace of the classifier.

A.2.2 Precision

Precision is the accuracy of prediction given that a specific class has been predicted [22]. It is defined in terms of TP, FP, TN, and FN in the following equation:

$$p = \frac{TP}{TP + FP} \tag{A.4}$$

A.2.3 Recall

Recall is a measure of the sensitivity of the classifier, measuring the classifier's ability to select instances of a specific class from the dataset [22]. It is defined as follows:

$$r = \frac{TP}{TP + FN} \tag{A.5}$$

A.2.4 F-measure

F-measure is the weighted harmonic mean of precision and recall [22]. Balanced F-measure is defined as:

$$F = \frac{2 * p * r}{p + r}, \tag{A.6}$$

where p is precision, and r is recall. F-measure discourages a classification system which "sacrifices one measure for another too drastically" [22].

Appendix B

Support Vector Machines

The problem of RNA gene finding is reducible to a classification problem, where features extracted from a segment of a genomic sequence are used to predict a label. In the problem's simplest form these labels are binary (i.e. the segment is either from a RNA gene region or not). Although there are many useful techniques for data classification, this thesis primarily focuses on the use of Support Vector Machines (SVMs) for reasons discussed in Section 2.2.

A SVM is a computer algorithm which learns from a training set to solve a classification problem by maximizing a particular mathematical function in respect to the training data [43]. Although fully based on mathematical theory, SVMs have a geometric interpretation which makes them understandable without delving deep into mathematics. There are four basic concepts needed to understand the basics of SVMs: "the separating hyperplane, the maximum-margin hyperplane, the soft margin, and the kernel function" [43].

Imagine data from two features is extracted from a RNA gene and non-RNA gene sequence training set. This feature data is used to plot each sequence from the training set on a plane, where the symbol used for the points is determined by the label of the particular sequence the point represents. From Figure B.1 it is easy to see the two clusters formed by this data (one in the upper left corner and one in the lower right corner) and how they can be partitioned by a line. The equation which produces the partitioning line can be thought as a simple rule for classification: if an unlabeled sequence has feature data which falls on the RNA gene side or the non RNA gene, its label is predicted accordingly. This concept of finding an optimal cut through training set data can be applied to both higher and lower feature dimensions. In the case of one feature all that is needed is a point on a line which maximizes the separation of the labeled data points, as illustrated in Figure B.2. Likewise,

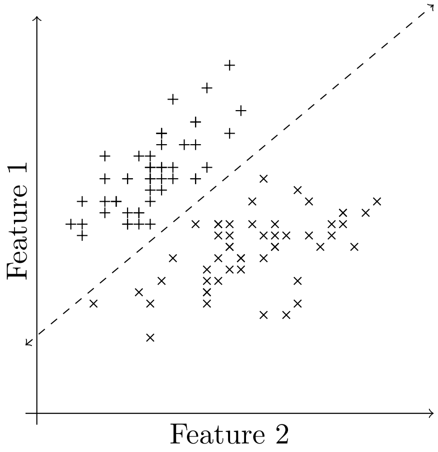

Figure B.1: Two Feature Separating Hyperplane Example.

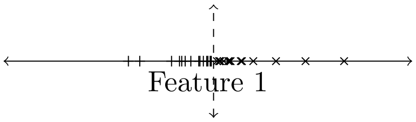

Figure B.2: One Feature Separating Hyperplane Example.

moving into three-dimensional space (three features) a cutting plane is used to partition the data instead of just a line [43]. This process is quite general and one can use a hyperplane to mathematically extrapolate this process to higher dimensional spaces [43, 8].

SVMs are not the only classifier to utilize this cutting plane concept [43]. There are often many different possible cuts that can be used [43]. SVMs have a unique method for selecting the cutting plane based on statistical learning theory, where they try to maximize the margin of the hyperplane, which is the distance from the hyperplane to the nearest expression vector (data point) [43]. By selecting a hyperplane that divides the data labels as evenly as possible the SVM is able to maximize its potential to correctly predict the class of unseen cases [43]. Because SVMs are a trained model and do not assume a normal distribution, it is important that the data classified by the SVM comes from the same distribution which is used to train the SVM [43].

As real world data is often dirty, the SVM algorithm deals with outliers that potentially negatively affect the hyperplane using a soft margin, that is, when the dataset contains a rogue data point, it is allowed to fall on the wrong side of the cutting plane [43]. This means that some of the data points will be on the opposite side of the cutting plane from the majority of their group [43]. Plainly, the SVM would not be a useful algorithm if it allowed too many of these misclassifications, hence user specified parameters can control roughly the number of misclassifications and how far across the line they are allowed to reside [43]. Such relaxation is in direct conflict with the size of the margin, so there is a trade off between violations and margin size, making the parameters tricky to configure [43].

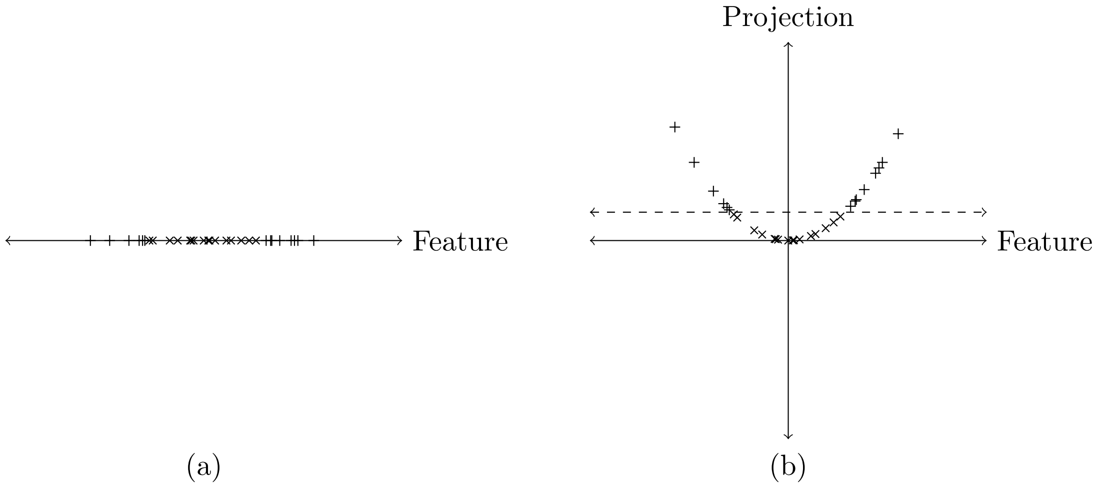

Figure B.3: Data Projection. This figure illustrates how a kernel method can be used to separate nonlinear divisible data using a projection into a higher dimension.

To explain the concept of a kernel function, consider another one dimensional classification problem, shown in graph (a) of Figure B.3. In this case, the data for one of the labels is clustered around the zero mark and the other label's data have large absolute values, creating a situation where no single point on the line can be used to partition the data into its respective groups (even applying a soft margin does not help). In cases like the one described, a kernel function can be introduced, which will add an additional dimension to the data. Figure B.3 (b) shows a projection of the dataset into a second dimension, where the values for the second dimension were computed by simply squaring the original values. In Figure B.3 (b) this projection separated the data allowing for a linear equation to cleanly partition the data. In basic terms, the kernel function is a mathematical trick that allows the SVM to preform a higher dimensional classification from a lower dimensional dataset [43]. It is important to note that for any given dataset with non-contradicting labels (two identical data points with different labels) there always exists a kernel function that will allow the data to be linearly separated; however, in order to achieve this, projecting the data into a very high-dimensional space may be needed [43, 8]. Although the data will always be separated perfectly in a very high-dimensional space, the hyperplane can be overly specific to the training set, overfitting the data [43]. The degree to which the SVM algorithm projects the data into higher dimensions is the second configurable parameter of SVMs. In this case one wants the kernel function to separate the data, but not to the point where it introduces irrelevant dimensions [43].

It often can be hard to predict a "good" configuration for the parameters of a SVM and often the only realistic method of configuration is through trial and error, where the trials are tested using cross-validation[1] [43]. A grid search can be used to automate this process of trial and error, by training and testing the SVM for each combination of values in a discrete space and then picking the best configuration of the configurations explored. Finally, it should be noted that as discussed so far, SVMs appear to be only able to do binary classification, however SVMs can be extended to achieve multiple classification by breaking a multiple classification into a series of binary classifications. For example, if a dataset contains data with the labels A, B, and C, the problem could be converted into a compound binary classification problem by first classifying between A and not A [43]. Then once the data points with the label A have been dealt with, the problem is reduced to a binary classification problem partition B and C labeled data points [43]. This discussion of SVMs has been mostly conceptual as this thesis deals with SVMs as a "black box." For a mathematical discussion of SVMs see [62] or [8].

[1] A method by which a dataset is divided into several equal parts and each partition is trained against the remaining partitions.

Appendix C

Nucleotide Shuffling

Often RNA gene finding experiments rely on nucleotide shuffling as either an integral part of the method or as a gene finding testing tool. Nucleotide shuffling transforms a RNA gene sequence into a random RNA sequence with the similar sequence composition properties as the RNA gene but without the same structural properties. This destruction of the structural properties of an RNA gene while preserving certain composition properties is useful when validating whether a gene finder is actually using structural information for RNA gene finding or is relying on sequence composition. Furthermore, nucleotide shuffling is helpful in generating populations of sequence with the same sequence composition when calculating z-scores.

Mononucleotide and dinucleotide shuffling are the two varieties of RNA sequence shuffling commonly used. When mononucleotide shuffling is performed on a sequence, it is a simple procedure of running through each nucleotide in the sequence and randomly choosing a nucleotide to swap it with from the part of the sequence not yet processed. Mononucleotide shuffling will preserve sequence composition, length, but just change the order of the nucleotides, as swapping ensures each nucleotide in the original sequence is also in the shuffled one. Dinucleotide shuffling allows a random sequence to be generated which preserves the base composition, length, and transitional frequency of the sequence. The transitional frequency of a sequence is the rate at which one nucleotide follows another one.

A method for generating these dinucleotide shuffled sequences is proposed and proved correct by Altschul and Erickson [1]. First a digraph is constructed from the original sequence where each node in the digraph represents a nucleotide and each edge in the digraph represents a transition from one nucleotide to another [1]. The sequence is processed one nucleotide at a time and for each transition a new edge is added to the graph [1]. As an example the sequence “AATAAGCCGAT” is used to compose the digraph in Figure C.1. Once the sequence has been completely processed, a random walk is formed in the graph which passes over every edge exactly once [1]. The sequence of nucleotides created by this path will be a dinucleotide shuffle of the original sequence [1]. Finally it should be noted that because, as already described, stacking energies contribute greatly to the stability of a RNA gene, using mononucleotide shuffling instead of dinucleotide shuffling can skew experiments that rely on it [14].

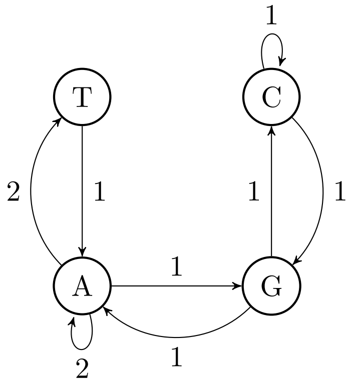

Figure C.1: Dinucleotide Shuffling Digraph. The digraph produced by the nucleotide transitions found in the sequence “AATAAGCCGAT”.

Appendix D

Protein Finding Methods

Unlike proteins which heavily rely on specific DNA sequences to produce their structures, RNA genes are less dependent upon their specific sequences. As mentioned elsewhere in this thesis, this is because the structure of RNA genes is derived directly from their sequences whereas proteins go through the process of translation, mapping their sequence in consecutive groups of three nucleotides, known as codons, to specific amino acids, producing the protein sequence [50]. The mapping produced by translation has a major effect on the resulting amino acid sequence. For example, consider the mRNA transcript "AAUGAU-UAUA", if the first nucleotide was used as the start of the translation reading frame the resulting amino acid sequence would be "Asparagine-Aspartic Acid-Tyrosine"[1]. However, if the second nucleotide was used as the start of the reading frame then the resulting protein sequence would be "Methionine-Isoleucine-Isoleucine". As it can be seen from this example, the reading frame for translation can have a devastating effect on protein gene sequence and hence structure. So if a protein gene sequence suffers a deletion or addition mutation where a nucleotide is removed or added respectfully, the resulting reading frame after that mutation will be altered, causing a major change in the protein sequence which ultimately will affect the protein's structure and function. Although point mutations, where a single nucleotide is changed from one base to another do not cause a change as significant as a deletion or addition mutation, they can still cause the amino acid sequence of the protein to be altered enough to disrupt the function of the protein. Since an organism with a mutation that causes a disruption of a protein's function will not be as fit as other organisms in the

[1] See [7] for amino acid encodings.

species without the mutation, the mutation will be selected against by evolution. Therefore it can be seen that there is evolutionary pressure to ensure protein sequence is conserved. This evolutionary pressure is not as strong in RNA genes, because the sequence of RNA genes is directly used to construct the RNA gene structure. If a small change occurs in the RNA gene from a mutation, it may have very little effect on the resulting structure of the RNA gene and hence very little evolutionary pressure on the sequence. Likewise, RNA gene sequences can undergo correlated mutations, which occur when two paired nucleotides both mutate to another set of pairing nucleotides. For example, the nucleotide pair A-T might be converted into C-G. This type of correlated mutation often has very little impact on the structure of the RNA gene and often there is little evolutionary pressure against it. Since protein gene sequences are conserved across evolution sequence alignment, open reading frame, hexamer frequency, and codon bias signals are very successful as protein gene finding tools, but are not effective for RNA gene finding [50].

Appendix E

Data

E.1 Structural Element Metric Statistics

E.1.1 Experiment 1

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
SLC	0.0099	0.5789	0.2063	0.6195	0.2021
Avg. Stack%	0.0095	0.8813	0.1423	0.8510	0.1683
Avg. Stack Size	0.0072	9.6812	3.6537	9.0829	3.4417
Avg. Internal Loop SLC	0.0061	0.3180	0.3802	0.3792	0.3946
Loop%	0.0058	0.1063	0.1190	0.1247	0.1234
PP	0.0058	0.8937	0.1190	0.8753	0.1234
Avg. Loop SLC	0.0054	0.4199	0.4072	0.4802	0.4067
Avg. Bulge FS	0.0053	4.9465	1.7936	4.8986	1.6883
Avg. Loop FS	0.0053	4.8693	1.7360	4.7771	1.5636
Avg. Bulge CS	0.0052	5.3056	1.9742	5.2469	1.9515
Avg. Internal Loop FS	0.0051	4.8135	1.7976	4.7178	1.6193
Avg. Loop CS	0.0050	5.5777	1.9785	5.4423	1.8477
Avg. Internal Loop CS	0.0049	5.6463	2.0522	5.5076	1.9145
Avg. Internal Loop NLC	0.0049	0.4041	0.4611	0.4692	0.4654
Size	0.0046	23.5521	19.5865	26.4218	22.5898

Avg. Loop NLC	0.0046	0.5096	0.4693	0.5728	0.4608
Avg. Bulge NLC	0.0045	0.2741	0.4360	0.3334	0.4610
Avg. Bulge SLC	0.0043	0.2432	0.4004	0.2972	0.4262
Internal Loop%	0.0041	0.0869	0.1145	0.1018	0.1186
Avg. Bulge GC% Bond	0.0036	0.1694	0.3225	0.2093	0.3458
Avg. Internal Loop GC% Bond	0.0034	0.2233	0.3268	0.2625	0.3393
Avg. Loop G%	0.0032	0.1335	0.2230	0.1601	0.2432
Avg. Internal Loop G%	0.0032	0.1328	0.2291	0.1597	0.2520
Avg. Stack FS	0.0031	2.8300	1.1443	2.8215	1.1204
Avg. Loop GC% Bond	0.0030	0.2947	0.3481	0.3335	0.3514
Avg. Loop C%	0.0030	0.0783	0.1691	0.0978	0.1823
Avg. Internal Loop C%	0.0029	0.0629	0.1393	0.0791	0.1543
Avg. Stack C%	0.0026	0.2962	0.1067	0.2854	0.1079
Avg. Stack SLC	0.0025	0.4180	0.1035	0.4283	0.1074
Avg. Bulge Size	0.0024	0.4996	1.0796	0.6086	1.2020
Avg. Stack CS	0.0024	6.2801	1.8011	6.1400	1.7406
Avg. Bulge G%	0.0023	0.0390	0.1580	0.0560	0.1909
Avg. Stack CC%	0.0023	0.1061	0.1727	0.0906	0.1560
Avg. Internal Loop AU% Bond	0.0022	0.1447	0.2548	0.1693	0.2641
Avg. Stack G%	0.0022	0.3516	0.0988	0.3423	0.1015
Avg. Bulge C%	0.0021	0.0417	0.1678	0.0581	0.1958
Avg. Internal Loop Size	0.0020	1.8342	2.4914	2.0613	2.5092
Avg. Loop AU% Bond	0.0020	0.1731	0.2576	0.1962	0.2621
NLC	0.0018	0.9175	0.0929	0.9249	0.0804
Bulge%	0.0018	0.0195	0.0401	0.0229	0.0423
Avg. Bulge AU%	0.0017	0.0838	0.2053	0.1014	0.2229
Avg. Internal Loop CC%	0.0017	0.0059	0.0479	0.0105	0.0624
Avg. Loop CC%	0.0016	0.0061	0.0474	0.0105	0.0605
Avg. Loop Size	0.0014	1.8318	2.2510	1.9965	2.1841
Avg. Internal Loop SR	0.0013	0.9175	0.1743	0.9042	0.1826
CC%	0.0013	0.0882	0.1063	0.0809	0.0984

Avg. Stack GG%	0.0013	0.1342	0.1850	0.1215	0.1721
Avg. Bulge U%	0.0012	0.0750	0.2283	0.0911	0.2456
Avg. Stack UG%	0.0011	0.0556	0.1269	0.0644	0.1307
GG%	0.0011	0.1221	0.1134	0.1147	0.1060
Avg. Internal Loop A%	0.0010	0.1619	0.2479	0.1784	0.2483
Avg. Internal Loop CG%	0.0010	0.0079	0.0505	0.0114	0.0590
Avg. Loop CG%	0.0010	0.0074	0.0473	0.0106	0.0537
GC%	0.0009	0.1066	0.1578	0.0977	0.1348
Avg. Stack GC%	0.0009	0.1307	0.2704	0.1149	0.2406
Avg. Stack U%	0.0009	0.2016	0.1058	0.2078	0.1054
C%	0.0009	0.2797	0.1051	0.2737	0.1036
Avg. Internal Loop U%	0.0008	0.0809	0.1804	0.0917	0.1875
Avg. Loop U%	0.0008	0.1150	0.2223	0.1275	0.2246
GC% Bond	0.0008	0.5905	0.2109	0.5793	0.2094
A%	0.0008	0.1742	0.1012	0.1796	0.0992
Avg. Stack NLC	0.0007	0.8812	0.1124	0.8745	0.1340
Avg. Stack UU%	0.0007	0.0395	0.1097	0.0449	0.1117
Avg. Loop GA%	0.0006	0.0385	0.1227	0.0324	0.1067
U%	0.0006	0.2031	0.1047	0.2082	0.1035
Avg. Stack A%	0.0006	0.1462	0.0965	0.1509	0.0949
UG%	0.0006	0.0634	0.0606	0.0665	0.0623
Avg. Internal Loop AG%	0.0006	0.0253	0.0965	0.0304	0.1060
Avg. Bulge CC%	0.0006	0.0019	0.0308	0.0038	0.0460
G%	0.0006	0.3430	0.0968	0.3385	0.0952
Avg. Internal Loop GA%	0.0005	0.0409	0.1327	0.0349	0.1174
UU%	0.0005	0.0448	0.0637	0.0474	0.0638
Avg. Bulge CG%	0.0005	0.0016	0.0268	0.0030	0.0381
Avg. Internal Loop AU%	0.0004	0.0146	0.0585	0.0171	0.0643
Avg. Internal Loop UG%	0.0004	0.0272	0.1307	0.0220	0.1051
Avg. Bulge CU%	0.0004	0.0031	0.0408	0.0050	0.0505
AU% Bond	0.0004	0.2963	0.1918	0.3038	0.1880

Avg. Loop UG%	0.0004	0.0251	0.1185	0.0206	0.0953
Avg. Internal Loop UC%	0.0004	0.0089	0.0573	0.0113	0.0609
Avg. Bulge GG%	0.0004	0.0018	0.0289	0.0031	0.0377
Avg. Internal Loop GU% Bond	0.0004	0.0705	0.1704	0.0771	0.1695
Avg. Stack AA%	0.0003	0.0193	0.0775	0.0221	0.0800
Avg. Stack CA%	0.0003	0.0348	0.1027	0.0385	0.1043
Avg. Bulge A%	0.0003	0.1302	0.3017	0.1409	0.3053
Avg. Stack GC%	0.0003	0.0551	0.1268	0.0592	0.1270
GA%	0.0003	0.0619	0.0642	0.0640	0.0654
AC%	0.0003	0.0448	0.0558	0.0469	0.0577
Avg. Internal Loop GC%	0.0003	0.0083	0.0582	0.0104	0.0614
GU% Bond	0.0003	0.1132	0.1115	0.1169	0.1119
Avg. Loop US%	0.0003	0.0096	0.0566	0.0115	0.0587
Avg. Internal Loop UU%	0.0003	0.0161	0.0868	0.0191	0.0907
Avg. Loop AU%	0.0003	0.0171	0.0654	0.0191	0.0702
Avg. Stack GU%	0.0003	0.0843	0.1526	0.0799	0.1436
Avg. Loop AG%	0.0002	0.0250	0.0906	0.0280	0.0961
Avg. Bulge GU%	0.0002	0.0035	0.0413	0.0047	0.0470
Avg. Loop GU% Bond	0.0002	0.0778	0.1665	0.0823	0.1626
Avg. Bulge AG%	0.0002	0.0082	0.0739	0.0063	0.0550
Avg. Stack UC%	0.0002	0.0544	0.1270	0.0576	0.1257
Avg. Loop GC%	0.0002	0.0081	0.0548	0.0096	0.0558
AA%	0.0002	0.0376	0.0542	0.0391	0.0544
Avg. Loop A%	0.0002	0.2187	0.2947	0.2264	0.2825
Avg. Bulge GU% Bond	0.0002	0.0326	0.1216	0.0354	0.1247
Avg. Loop CU%	0.0002	0.0100	0.0579	0.0116	0.0604
Avg. Bulge UG%	0.0002	0.0038	0.0435	0.0050	0.0506
Avg. Stack CG%	0.0002	0.0647	0.1931	0.0699	0.1893
Avg. Internal Loop GG%	0.0002	0.0196	0.0911	0.0217	0.0927
Avg. Bulge AU%	0.0002	0.0078	0.0611	0.0095	0.0689
Avg. Internal Loop UA%	0.0001	0.0188	0.0683	0.0204	0.0723

Avg. Loop GG%	0.0001	0.0173	0.0810	0.0191	0.0821
Avg. Bulge AA%	0.0001	0.0191	0.1098	0.0168	0.0985
CG%	0.0001	0.0661	0.0929	0.0685	0.0962
UA%	0.0001	0.0389	0.0617	0.0404	0.0632
Avg. Loop UA%	0.0001	0.0203	0.0720	0.0218	0.0757
CU%	0.0001	0.0579	0.0599	0.0565	0.0605
Avg. Loop UU%	0.0001	0.0181	0.0907	0.0199	0.0887
Avg. Internal Loop AA%	0.0001	0.0520	0.1413	0.0553	0.1445
Avg. Bulge CA%	0.0001	0.0051	0.0495	0.0062	0.0556
Avg. Stack AU%	0.0001	0.0293	0.1151	0.0314	0.1227
Avg. Stack UA%	0.0001	0.0316	0.1313	0.0338	0.1303
Avg. Loop GU%	0.0001	0.0101	0.0573	0.0111	0.0570
Avg. Loop CA%	0.0001	0.0149	0.0655	0.0162	0.0698
Avg. Bulge UA%	0.0001	0.0084	0.0699	0.0095	0.0688
CA%	0.0001	0.0410	0.0526	0.0418	0.0530
Avg. Bulge AC%	0.0001	0.0055	0.0510	0.0063	0.0550
Avg. Internal Loop GU%	0.0001	0.0097	0.0568	0.0105	0.0572
Avg. Internal Loop AC%	5e-05	0.0172	0.0767	0.0162	0.0689
Avg. Stack CU%	4e-05	0.0598	0.1316	0.0617	0.1281
Avg. Bulge GC%	4e-05	0.0020	0.0292	0.0023	0.0319
AU%	4e-05	0.0369	0.0603	0.0361	0.0585
Avg. Bulge UC%	4e-05	0.0039	0.0447	0.0044	0.0458
Avg. Loop AA%	4e-05	0.0555	0.1421	0.0541	0.1376
Avg. Bulge UU%	3e-05	0.0074	0.0659	0.0078	0.0661
Avg. Loop AC%	3e-05	0.0174	0.0747	0.0167	0.0691
Avg. Internal Loop CU%	2e-05	0.0104	0.0673	0.0111	0.0615
Avg. Internal Loop CA%	2e-05	0.0148	0.0685	0.0156	0.0702
Avg. Stack AC%	2e-05	0.0449	0.1179	0.0442	0.1120
GU%	2e-05	0.0716	0.0680	0.0723	0.0682
UC%	1e-05	0.0538	0.0621	0.0541	0.0640
Avg. Stack AG%	8e-06	0.0512	0.1230	0.0520	0.1189

Avg. Bulge GA%	2e-06	0.0081	0.0669	0.0077	0.0596
AG%	1e-06	0.0600	0.0581	0.0600	0.0597

Table E.1: Experiment 1 Stem Metric Statistics. Ranks each stem metric from the first experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
CS	0.0234	51.2419	20.5950	58.1415	24.9893
FS	0.0225	24.3202	16.6562	30.0029	21.5590
Avg. Stack Size	0.0197	9.2210	3.5871	8.2691	3.1423
Stack%	0.0177	0.6399	0.1128	0.6062	0.1395
PP	0.0143	0.6490	0.0980	0.6244	0.1080
Avg. Stack CC%	0.0073	0.1177	0.1806	0.0889	0.1577
Avg. Hairpin Loop GG%	0.0065	0.0337	0.1076	0.0532	0.1357
Avg. Stack GG%	0.0065	0.1497	0.1923	0.1203	0.1748
Avg. Hairpin Loop AG%	0.0058	0.0547	0.1119	0.0730	0.1265
Avg. Stack C%	0.0053	0.3001	0.1111	0.2836	0.1155
Avg. Hairpin Loop SLC	0.0051	0.3864	0.0800	0.3746	0.0865
Avg. Hairpin Loop Size	0.0051	5.2396	2.0148	5.5494	2.3551
Avg. Stack G%	0.0050	0.3564	0.1033	0.3415	0.1107
Avg. Hairpin Loop AA%	0.0047	0.1897	0.2236	0.1587	0.2191
Avg. Stack SLC	0.0044	0.4294	0.1078	0.4442	0.1161
Avg. Bulge FS	0.0044	4.9117	1.6795	4.7531	1.6186
Avg. Bulge NLC	0.0042	0.2541	0.4250	0.3112	0.4528
Avg. Bulge SLC	0.0042	0.2235	0.3877	0.2761	0.4172
Avg. Bulge CS	0.0041	5.3227	1.8730	5.1170	1.9184
Avg. Bulge GC% Bond	0.0038	0.1557	0.3127	0.1961	0.3394
GC% Bond	0.0037	0.6016	0.2184	0.5869	0.2211
Avg. Stack UG%	0.0036	0.0496	0.1210	0.0646	0.1336
Avg. Internal Loop SLC	0.0033	0.3159	0.3734	0.3594	0.3927
Avg. Hairpin Loop AU%	0.0032	0.0557	0.1084	0.0685	0.1213
Avg. Loop SLC	0.0032	0.4109	0.3999	0.4572	0.4091
Avg. Internal Loop GA%	0.0032	0.0479	0.1442	0.0328	0.1144
Avg. Loop FS	0.0029	4.8513	1.6959	4.6180	1.4910
Avg. Loop GA%	0.0028	0.0437	0.1315	0.0308	0.1051
AU% Bond	0.0027	0.2838	0.1975	0.2945	0.1975

Avg. Internal Loop FS	0.0026	4.8131	1.7501	4.5601	1.5454
Avg. Loop G%	0.0025	0.1311	0.2145	0.1538	0.2425
Bulge%	0.0025	0.0158	0.0348	0.0193	0.0374
Hairpin%	0.0024	0.2551	0.1407	0.2698	0.1587
Avg. Hairpin Loop GA%	0.0023	0.1196	0.1646	0.1041	0.1515
Avg. Loop CS	0.0023	5.6648	1.9669	5.2972	1.8067
Avg. Loop C%	0.0023	0.0775	0.1678	0.0945	0.1810
Avg. Internal Loop G%	0.0022	0.1303	0.2196	0.1519	0.2491
Avg. Internal Loop CS	0.0021	5.7498	2.0266	5.3610	1.8732
Avg. Bulge Size	0.0021	0.4800	1.1009	0.5798	1.2096
Loop%	0.0020	0.0959	0.1089	0.1058	0.1119
Avg. Bulge G%	0.0020	0.0371	0.1555	0.0523	0.1858
Avg. Bulge C%	0.0019	0.0392	0.1647	0.0552	0.1919
Avg. Internal Loop C%	0.0018	0.0638	0.1413	0.0763	0.1531
Avg. Loop NLC	0.0018	0.5076	0.4667	0.5471	0.4652
Avg. Internal Loop NLC	0.0016	0.4087	0.4599	0.4457	0.4646
AC%	0.0016	0.0513	0.0468	0.0551	0.0497
UG%	0.0016	0.0607	0.0484	0.0645	0.0507
Avg. Stack CA%	0.0016	0.0305	0.0967	0.0385	0.1064
Avg. Hairpin Loop CC%	0.0016	0.0271	0.0927	0.0354	0.1081
Avg. Loop AA%	0.0016	0.0638	0.1530	0.0522	0.1362
Avg. Hairpin Loop UU%	0.0015	0.0703	0.1627	0.0581	0.1417
Avg. Stack U%	0.0015	0.1961	0.1093	0.2045	0.1113
Avg. Internal Loop GC% Bond	0.0015	0.2268	0.3264	0.2522	0.3383
Avg. Bulge AU%	0.0013	0.0787	0.2003	0.0938	0.2166
Avg. Loop GC% Bond	0.0013	0.2955	0.3482	0.3210	0.3528
Avg. Bulge U%	0.0013	0.0685	0.2184	0.0841	0.2377
Avg. Stack A%	0.0013	0.1398	0.0993	0.1466	0.0997
Avg. Hairpin Loop AG%	0.0012	0.0519	0.1093	0.0595	0.1147
Avg. Stack GC%	0.0012	0.1340	0.2765	0.1154	0.2451
SLC	0.0012	0.8827	0.0457	0.8860	0.0495

Avg. Loop CC%	0.0012	0.0070	0.0520	0.0107	0.0621
Avg. Internal Loop CG%	0.0012	0.0074	0.0481	0.0111	0.0581
GC% Bond	0.0011	0.6016	0.2184	0.5869	0.2211
CG%	0.0011	0.0696	0.0627	0.0654	0.0638
Avg. Loop CG%	0.0011	0.0071	0.0460	0.0105	0.0536
Avg. Stack UU%	0.0010	0.0370	0.1074	0.0441	0.1121
Avg. Internal Loop CC%	0.0010	0.0070	0.0537	0.0104	0.0619
AU%	0.0010	0.0428	0.0508	0.0459	0.0516
Avg. Internal Loop AU% Bond	0.0010	0.1412	0.2491	0.1568	0.2570
CC%	0.0010	0.0654	0.0624	0.0617	0.0645
Avg. Internal Loop GC%	0.0009	0.0066	0.0473	0.0101	0.0615
Internal Loop%	0.0009	0.0800	0.1048	0.0866	0.1060
Avg. Internal Loop UG%	0.0009	0.0275	0.1328	0.0203	0.1009
Avg. Stack CG%	0.0009	0.0612	0.1877	0.0731	0.1973
Avg. Bulge CC%	0.0009	0.0016	0.0301	0.0039	0.0479
Avg. Hairpin Loop UC%	0.0009	0.0451	0.1031	0.0393	0.0976
CA%	0.0009	0.0456	0.0452	0.0484	0.0466
Avg. Stack NLC	0.0009	0.8717	0.1265	0.8630	0.1614
Avg. Hairpin Loop C%	0.0009	0.1574	0.1701	0.1685	0.1784
Size	0.0009	28.0147	19.2559	29.1799	21.4146
Avg. Loop UG%	0.0009	0.0256	0.1231	0.0192	0.0927
Avg. Loop AU% Bond	0.0008	0.1683	0.2530	0.1833	0.2577
Avg. Bulge AG%	0.0008	0.0104	0.0876	0.0063	0.0544
AU% Bond	0.0008	0.2838	0.1975	0.2945	0.1975
G%	0.0008	0.3129	0.0893	0.3080	0.0932
Avg. Hairpin Loop GC%	0.0007	0.0460	0.1085	0.0407	0.1001
GG%	0.0007	0.1008	0.0781	0.0965	0.0786
Avg. Hairpin Loop UG%	0.0007	0.0456	0.1027	0.0516	0.1107
Avg. Stack AA%	0.0006	0.0173	0.0737	0.0209	0.0786
C%	0.0006	0.2424	0.0848	0.2384	0.0874
U%	0.0006	0.2069	0.0994	0.2117	0.1004

Avg. Hairpin Loop CU%	0.0006	0.0350	0.0931	0.0397	0.0972
Avg. Loop GC%	0.0006	0.0068	0.0460	0.0093	0.0555
Avg. Hairpin Loop CA%	0.0006	0.0617	0.1183	0.0565	0.1126
Avg. Hairpin Loop A%	0.0005	0.3886	0.2249	0.3775	0.2246
GU% Bond	0.0005	0.1147	0.1155	0.1186	0.1197
A%	0.0005	0.2378	0.0962	0.2419	0.1005
GC%	0.0005	0.0770	0.0744	0.0739	0.0684
Avg. Hairpin Loop NLC	0.0005	0.8681	0.1072	0.8730	0.1098
Avg. Bulge GG%	0.0005	0.0018	0.0281	0.0033	0.0396
Avg. Bulge CG%	0.0004	0.0016	0.0278	0.0030	0.0384
Avg. Bulge AC%	0.0004	0.0042	0.0454	0.0062	0.0547
Avg. Loop U%	0.0004	0.1122	0.2170	0.1207	0.2213
Avg. Bulge AA%	0.0004	0.0204	0.1138	0.0162	0.0963
UA%	0.0004	0.0482	0.0536	0.0502	0.0539
GA%	0.0004	0.0735	0.0555	0.0757	0.0570
Avg. Internal Loop GG%	0.0004	0.0177	0.0825	0.0210	0.0910
Avg. Internal Loop AA%	0.0004	0.0585	0.1494	0.0529	0.1428
Avg. Loop A%	0.0003	0.2262	0.2995	0.2159	0.2803
Avg. Internal Loop AU%	0.0003	0.0141	0.0558	0.0162	0.0625
Avg. Stack GC%	0.0003	0.0525	0.1242	0.0567	0.1268
Avg. Bulge A%	0.0003	0.1213	0.2943	0.1318	0.2977
Avg. Bulge CU%	0.0003	0.0035	0.0453	0.0049	0.0500
Avg. Loop GG%	0.0003	0.0163	0.0750	0.0187	0.0811
Avg. Internal Loop AC%	0.0003	0.0183	0.0788	0.0159	0.0697
Avg. Stack FS	0.0003	2.9197	1.1824	2.8533	1.1478
GU% Bond	0.0002	0.1147	0.1155	0.1186	0.1197
Avg. Bulge GU%	0.0002	0.0033	0.0418	0.0047	0.0466
Avg. Bulge UG%	0.0002	0.0032	0.0427	0.0046	0.0483
AA%	0.0002	0.0780	0.0734	0.0757	0.0802
Avg. Internal Loop U%	0.0002	0.0816	0.1791	0.0864	0.1828
Avg. Internal Loop UU%	0.0002	0.0162	0.0875	0.0185	0.0893

CU%	0.0001	0.0535	0.0458	0.0524	0.0462
Avg. Internal Loop SR	0.0001	0.9127	0.1786	0.9088	0.1807
Avg. Hairpin Loop G%	0.0001	0.2318	0.1773	0.2364	0.1944
UU%	0.0001	0.0490	0.0615	0.0504	0.0626
Avg. Loop AU%	0.0001	0.0167	0.0636	0.0180	0.0676
Avg. Internal Loop AG%	0.0001	0.0266	0.1002	0.0288	0.1036
Avg. Internal Loop GU%	0.0001	0.0112	0.0601	0.0100	0.0550
Avg. Hairpin Loop CG%	0.0001	0.0384	0.0989	0.0365	0.0955
Avg. Stack AG%	0.0001	0.0495	0.1212	0.0522	0.1218
Avg. Hairpin Loop U%	0.0001	0.2222	0.2188	0.2177	0.2056
Avg. Internal Loop CA%	0.0001	0.0163	0.0705	0.0149	0.0684
AG%	0.0001	0.0723	0.0501	0.0716	0.0537
Avg. Bulge GA%	0.0001	0.0063	0.0585	0.0072	0.0557
Avg. Internal Loop UC%	0.0001	0.0097	0.0613	0.0110	0.0603
Avg. Internal Loop CU%	0.0001	0.0119	0.0762	0.0107	0.0605
Avg. Bulge CA%	0.0001	0.0054	0.0519	0.0061	0.0546
Avg. Bulge AU%	0.0001	0.0079	0.0625	0.0087	0.0651
Avg. Stack UA%	0.0001	0.0297	0.1300	0.0316	0.1284
Avg. Stack AC%	0.0001	0.0408	0.1102	0.0425	0.1123
Avg. Internal Loop GU% Bond	0.0001	0.0777	0.1792	0.0750	0.1688
Avg. Loop GU% Bond	0.0001	0.0832	0.1735	0.0806	0.1637
Avg. Internal Loop Size	0.0001	1.9395	2.6061	1.9745	2.5133
Avg. Bulge GU% Bond	5e-05	0.0319	0.1222	0.0335	0.1222
GU%	5e-05	0.0641	0.0492	0.0647	0.0502
Avg. Loop US%	5e-05	0.0102	0.0593	0.0111	0.0575
Avg. Stack UC%	4e-05	0.0533	0.1264	0.0546	0.1247
Avg. Loop GU%	4e-05	0.0115	0.0616	0.0108	0.0561
Avg. Bulge GC%	3e-05	0.0017	0.0256	0.0019	0.0272
Avg. Loop AG%	3e-05	0.0279	0.0980	0.0268	0.0946
Avg. Loop AC%	3e-05	0.0174	0.0737	0.0166	0.0700
Avg. Stack CU%	3e-05	0.0608	0.1328	0.0625	0.1319

Avg. Loop UA%	3e-05	0.0221	0.0736	0.0211	0.0757
Avg. Stack AU%	3e-05	0.0299	0.1152	0.0309	0.1244
Avg. Hairpin Loop GU%	2e-05	0.0518	0.1078	0.0507	0.1086
UC%	2e-05	0.0482	0.0458	0.0477	0.0454
Avg. Bulge UC%	2e-05	0.0039	0.0441	0.0042	0.0444
Avg. Bulge UU%	2e-05	0.0073	0.0663	0.0067	0.0602
Avg. Hairpin Loop UA%	1e-05	0.0738	0.1239	0.0743	0.1245
Avg. Loop Size	1e-05	1.9481	2.3928	1.9302	2.2133
Avg. Internal Loop UA%	9e-06	0.0200	0.0686	0.0194	0.0707
Avg. Bulge UA%	9e-06	0.0089	0.0711	0.0090	0.0674
NLC	6e-06	0.9618	0.0297	0.9616	0.0320
Avg. Loop UU%	3e-06	0.0186	0.0931	0.0190	0.0866
Avg. Loop CU%	3e-06	0.0115	0.0638	0.0112	0.0592
Avg. Loop CA%	3e-06	0.0161	0.0664	0.0158	0.0692
Avg. Stack GU%	2e-06	0.0788	0.1453	0.0794	0.1462
Avg. Internal Loop A%	1e-06	0.1700	0.2519	0.1693	0.2454
Avg. Stack CS	1e-06	6.2529	1.8212	5.9132	1.7104

Table E.2: Experiment 1 Stemloop Metric Statistics. Ranks each stemloop metric from the first experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		Nongene	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
Size	0.0153	24.7129	20.2033	30.1691	23.8731
Loop%	0.0135	0.0975	0.1103	0.1238	0.1160
PP	0.0135	0.9025	0.1103	0.8762	0.1160
Avg. Stack FS	0.0130	2.6973	1.0702	2.7844	1.0865
Avg. Stack CS	0.0123	6.3172	1.7607	6.4046	1.7383
SLC	0.0123	0.5561	0.2100	0.6027	0.2106
Internal Loop%	0.0122	0.0772	0.1043	0.1010	0.1116
Avg. Internal Loop NLC	0.0119	0.3996	0.4630	0.5008	0.4647
Avg. Stack%	0.0118	0.8949	0.1216	0.8680	0.1254
Avg. Internal Loop SLC	0.0110	0.3232	0.3904	0.4058	0.3957
Avg. Loop NLC	0.0106	0.5119	0.4730	0.6073	0.4524
Avg. Internal Loop Size	0.0102	1.6902	2.3105	2.1778	2.4989
Avg. Internal Loop CS	0.0101	5.5350	2.0799	5.6750	1.9474
Avg. Loop Size	0.0099	1.6684	2.0287	2.0856	2.1411
Avg. Loop CS	0.0097	5.4821	1.9902	5.6083	1.8800
Avg. Internal Loop FS	0.0094	4.8332	1.8502	4.8980	1.6819
Avg. Loop SLC	0.0092	0.4323	0.4169	0.5111	0.4015
Avg. Loop FS	0.0090	4.9058	1.7774	4.9591	1.6238
Avg. Internal Loop GC% Bond	0.0071	0.2201	0.3280	0.2763	0.3402
Avg. Loop GC% Bond	0.0065	0.2940	0.3477	0.3501	0.3488
Avg. Bulge CS	0.0064	5.2905	2.0600	5.3907	1.9777
Avg. Bulge FS	0.0062	4.9968	1.8924	5.0597	1.7483
Avg. Internal Loop A%	0.0062	0.1513	0.2416	0.1905	0.2517
Avg. Internal Loop SR	0.0050	0.9232	0.1692	0.8981	0.1851
Avg. Internal Loop C%	0.0048	0.0629	0.1378	0.0829	0.1558
Avg. Bulge NLC	0.0046	0.3007	0.4490	0.3633	0.4702
Avg. Internal Loop AU% Bond	0.0046	0.1498	0.2628	0.1861	0.2724
Avg. Internal Loop G%	0.0045	0.1361	0.2416	0.1702	0.2554
Avg. Bulge SLC	0.0043	0.2691	0.4147	0.3254	0.4363

Avg. Loop G%	0.0042	0.1371	0.2351	0.1686	0.2438
Avg. Loop C%	0.0042	0.0796	0.1697	0.1023	0.1841
Avg. Loop AU% Bond	0.0041	0.1795	0.2631	0.2134	0.2670
Avg. Internal Loop GU% Bond	0.0034	0.0609	0.1574	0.0800	0.1704
Avg. Loop A%	0.0034	0.2070	0.2858	0.2405	0.2848
Avg. Internal Loop CC%	0.0033	0.0046	0.0408	0.0106	0.0631
Avg. Bulge GC% Bond	0.0031	0.1883	0.3337	0.2270	0.3535
Avg. Internal Loop AA%	0.0029	0.0436	0.1308	0.0586	0.1466
Avg. Bulge Size	0.0028	0.5273	1.0497	0.6472	1.1906
Avg. Bulge G%	0.0027	0.0420	0.1623	0.0611	0.1975
Avg. Internal Loop U%	0.0025	0.0805	0.1836	0.0988	0.1935
Avg. Loop CC%	0.0025	0.0052	0.0425	0.0102	0.0582
Avg. Bulge C%	0.0022	0.0443	0.1683	0.0619	0.2010
Avg. Bulge AU%	0.0022	0.0913	0.2118	0.1118	0.2307
Avg. Loop GU% Bond	0.0022	0.0698	0.1553	0.0845	0.1611
Avg. Loop AG%	0.0021	0.0211	0.0816	0.0296	0.0981
Avg. Loop AA%	0.0021	0.0445	0.1265	0.0566	0.1396
Avg. Internal Loop AG%	0.0019	0.0236	0.0934	0.0326	0.1090
Avg. Stack GU%	0.0016	0.0929	0.1625	0.0805	0.1401
AA%	0.0015	0.0367	0.0523	0.0407	0.0520
Avg. Loop U%	0.0015	0.1196	0.2299	0.1366	0.2285
A%	0.0014	0.1785	0.0942	0.1854	0.0912
Avg. Internal Loop UC%	0.0014	0.0076	0.0505	0.0118	0.0618
Avg. Loop CU%	0.0013	0.0081	0.0500	0.0121	0.0620
Avg. Loop UA%	0.0012	0.0177	0.0692	0.0228	0.0758
Avg. Internal Loop UA%	0.0011	0.0171	0.0681	0.0217	0.0743
Avg. Stack GG%	0.0011	0.1117	0.1719	0.1230	0.1685
Avg. Internal Loop GU%	0.0011	0.0076	0.0522	0.0111	0.0600
Avg. Bulge U%	0.0010	0.0845	0.2415	0.1005	0.2555
Avg. Loop US%	0.0010	0.0085	0.0519	0.0120	0.0604
Bulge%	0.0010	0.0203	0.0386	0.0228	0.0398

NLC	0.0010	0.9244	0.0825	0.9294	0.0742
Avg. Stack SLC	0.0010	0.4013	0.0949	0.4070	0.0903
GC%	0.0010	0.1025	0.1349	0.0946	0.1139
Avg. Loop GU%	0.0010	0.0082	0.0512	0.0114	0.0582
C%	0.0009	0.2742	0.0979	0.2685	0.0946
Avg. Internal Loop GA%	0.0008	0.0315	0.1137	0.0379	0.1212
Avg. Loop CG%	0.0008	0.0079	0.0489	0.0107	0.0538
Avg. Internal Loop CG%	0.0008	0.0087	0.0545	0.0117	0.0601
Avg. Internal Loop CU%	0.0008	0.0085	0.0545	0.0116	0.0628
Avg. Stack NLC	0.0007	0.8945	0.0870	0.8900	0.0818
Avg. Bulge CU%	0.0007	0.0028	0.0380	0.0051	0.0511
Avg. Internal Loop AU%	0.0006	0.0151	0.0620	0.0183	0.0666
Avg. Stack Size	0.0006	10.3486	3.6571	10.1755	3.5227
Avg. Internal Loop CA%	0.0006	0.0130	0.0664	0.0165	0.0726
UU%	0.0005	0.0466	0.0618	0.0493	0.0606
Avg. Loop AU%	0.0005	0.0173	0.0680	0.0206	0.0734
Avg. Internal Loop UU%	0.0005	0.0159	0.0854	0.0200	0.0925
Avg. Stack GC%	0.0005	0.1264	0.2608	0.1142	0.2344
Avg. Stack UC%	0.0005	0.0558	0.1275	0.0616	0.1270
Avg. Loop UU%	0.0005	0.0170	0.0846	0.0213	0.0913
Avg. Loop CA%	0.0005	0.0132	0.0637	0.0166	0.0706
Avg. Bulge CG%	0.0005	0.0017	0.0256	0.0030	0.0378
UC%	0.0005	0.0573	0.0591	0.0547	0.0543
Avg. Bulge GU% Bond	0.0005	0.0326	0.1178	0.0379	0.1279
Avg. Stack AC%	0.0004	0.0515	0.1285	0.0464	0.1117
GU% Bond	0.0004	0.1106	0.1053	0.1148	0.1004
GU%	0.0004	0.0746	0.0649	0.0721	0.0571
AG%	0.0004	0.0597	0.0540	0.0618	0.0524
Avg. Bulge A%	0.0004	0.1414	0.3097	0.1532	0.3150
Avg. Bulge AU%	0.0003	0.0081	0.0609	0.0105	0.0735
UA%	0.0003	0.0402	0.0618	0.0422	0.0591

G%	0.0003	0.3358	0.0912	0.3327	0.0875
Avg. Stack UU%	0.0003	0.0419	0.1114	0.0459	0.1111
Avg. Loop GA%	0.0003	0.0313	0.1091	0.0346	0.1087
Avg. Stack GC%	0.0003	0.0581	0.1297	0.0624	0.1271
Avg. Bulge UU%	0.0003	0.0067	0.0619	0.0093	0.0732
GC% Bond	0.0003	0.5749	0.1981	0.5689	0.1921
Avg. Bulge CC%	0.0002	0.0024	0.0333	0.0036	0.0434
Avg. Bulge GG%	0.0002	0.0017	0.0281	0.0028	0.0350
Avg. Stack C%	0.0002	0.2907	0.0997	0.2878	0.0968
Avg. Stack U%	0.0002	0.2093	0.0997	0.2121	0.0968
Avg. Bulge AG%	0.0002	0.0049	0.0473	0.0064	0.0559
Avg. Bulge UA%	0.0002	0.0081	0.0704	0.0101	0.0705
Avg. Bulge GA%	0.0002	0.0107	0.0784	0.0085	0.0645
Avg. Stack AU%	0.0002	0.0285	0.1142	0.0321	0.1203
Avg. Bulge GU%	0.0002	0.0037	0.0407	0.0048	0.0475
Avg. Bulge CA%	0.0001	0.0048	0.0452	0.0063	0.0571
Avg. Stack CG%	0.0001	0.0700	0.2012	0.0656	0.1780
UG%	0.0001	0.0668	0.0577	0.0681	0.0545
Avg. Stack CA%	0.0001	0.0410	0.1107	0.0385	0.1013
U%	0.0001	0.2115	0.0975	0.2133	0.0951
Avg. Stack CC%	0.0001	0.0894	0.1596	0.0928	0.1536
GA%	0.0001	0.0640	0.0611	0.0651	0.0561
Avg. Bulge UG%	0.0001	0.0047	0.0440	0.0055	0.0535
Avg. Bulge AC%	0.0001	0.0075	0.0595	0.0063	0.0553
Avg. Stack AA%	0.0001	0.0222	0.0826	0.0237	0.0819
Avg. Bulge UC%	0.0001	0.0039	0.0448	0.0048	0.0477
Avg. Stack UA%	0.0001	0.0343	0.1325	0.0366	0.1326
Avg. Internal Loop UG%	0.0001	0.0262	0.1266	0.0243	0.1103
CU%	0.0001	0.0585	0.0555	0.0575	0.0524
Avg. Stack CU%	0.0001	0.0583	0.1294	0.0606	0.1229
Avg. Stack AG%	0.0001	0.0534	0.1251	0.0517	0.1149

Avg. Internal Loop AC%	0.0001	0.0157	0.0724	0.0166	0.0679
Avg. Loop UG%	0.0001	0.0239	0.1116	0.0224	0.0987
Avg. Stack G%	0.0001	0.3445	0.0915	0.3433	0.0878
Avg. Stack A%	0.0001	0.1555	0.0915	0.1567	0.0877
Avg. Bulge GC%	0.0001	0.0023	0.0320	0.0028	0.0373
CA%	4e-05	0.0418	0.0499	0.0426	0.0466
Avg. Loop GG%	4e-05	0.0185	0.0879	0.0196	0.0832
AU% Bond	3e-05	0.3144	0.1818	0.3163	0.1736
CG%	2e-05	0.0672	0.0900	0.0681	0.0822
Avg. Internal Loop GG%	2e-05	0.0218	0.1016	0.0228	0.0949
Avg. Loop AC%	2e-05	0.0176	0.0757	0.0169	0.0679
AC%	1e-05	0.0476	0.0525	0.0474	0.0497
GG%	8e-06	0.1156	0.1030	0.1151	0.0957
AU%	6e-06	0.0383	0.0595	0.0379	0.0550
Avg. Stack UG%	4e-06	0.0646	0.1352	0.0642	0.1266
Avg. Internal Loop GC%	3e-06	0.0105	0.0698	0.0106	0.0614
CC%	2e-06	0.0828	0.0960	0.0826	0.0899
Avg. Loop GC%	2e-06	0.0098	0.0632	0.0099	0.0563
Avg. Bulge AA%	1e-06	0.0177	0.1050	0.0176	0.1015

Table E.3: Experiment 1 Bridge Metric Statistics. Ranks each bridge metric from the first experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
Avg. Tail G%	0.0718	0.1173	0.1881	0.2332	0.2415
Avg. Joint NLC	0.0683	0.2816	0.4316	0.5140	0.4593
Avg. Joint CS	0.0682	3.7600	1.9277	4.0320	2.0743
Avg. Joint-Tail G%	0.0655	0.1306	0.1840	0.2315	0.2097
Joint%	0.0575	0.0898	0.1659	0.1792	0.2057
G%	0.0567	0.0930	0.1194	0.1485	0.1131
Avg. Joint FS	0.0561	1.1378	0.7635	1.0057	0.0805
Avg. Joint A%	0.0537	0.1192	0.2475	0.2523	0.3229
Size	0.0489	10.3385	7.0960	13.4796	7.1502
Avg. Joint Size	0.0488	1.0654	2.2723	2.3128	3.2940
Avg. Joint-Tail NLC	0.0417	0.8309	0.2226	0.9053	0.1292
Avg. Joint SLC	0.0411	0.1923	0.3268	0.3290	0.3487
AA%	0.0375	0.0743	0.1499	0.1415	0.1951
Avg. Tail AC%	0.0357	0.0881	0.1507	0.0395	0.1018
Avg. Tail UU%	0.0350	0.1253	0.2616	0.0475	0.1327
Avg. Joint-Tail UU%	0.0328	0.1103	0.2308	0.0454	0.1028
Avg. Tail CC%	0.0314	0.0689	0.1396	0.0271	0.0916
Avg. Joint-Tail AC%	0.0305	0.0884	0.1440	0.0459	0.0944
Avg. Joint-Tail AA%	0.0302	0.0630	0.1373	0.1164	0.1694
UU%	0.0279	0.1250	0.2525	0.0584	0.1241
Avg. Joint-Tail CC%	0.0271	0.0683	0.1371	0.0304	0.0886
GU%	0.0262	0.0271	0.0900	0.0649	0.1390
Avg. Tail GU%	0.0258	0.0211	0.0917	0.0618	0.1546
Avg. Joint-Tail GU%	0.0257	0.0220	0.0794	0.0550	0.1220
Avg. Joint AA%	0.0240	0.0323	0.1172	0.0822	0.1958
CC%	0.0223	0.0737	0.1397	0.0371	0.1027
Avg. Tail NLC	0.0220	0.8096	0.2477	0.8776	0.2093
AC%	0.0196	0.0926	0.1406	0.0573	0.1091
Avg. Joint-Tail CA%	0.0178	0.0711	0.1298	0.0413	0.0909

Avg. Joint-Tail AG%	0.0170	0.0289	0.0690	0.0513	0.0996
Avg. Tail AA%	0.0170	0.0593	0.1506	0.1038	0.1895
Avg. Tail CA%	0.0167	0.0725	0.1405	0.0402	0.1084
Avg. Joint C%	0.0166	0.0502	0.1533	0.0950	0.1926
GC%	0.0159	0.0183	0.0757	0.0426	0.1139
Avg. Joint-Tail GC%	0.0159	0.0152	0.0636	0.0367	0.1030
Avg. Joint-Tail Size	0.0156	3.7313	2.3429	4.3963	2.9611
Avg. Joint-Tail GA%	0.0147	0.0456	0.0973	0.0735	0.1307
Avg. Tail GA%	0.0138	0.0443	0.1074	0.0758	0.1558
Joint-Tail%	0.0129	0.6265	0.2026	0.6688	0.1691
Avg. Tail AG%	0.0128	0.0277	0.0754	0.0497	0.1154
Avg. Joint U%	0.0125	0.0643	0.1750	0.1063	0.2011
AG%	0.0124	0.0406	0.0848	0.0625	0.1097
Avg. Tail C%	0.0124	0.1940	0.2273	0.1467	0.1952
Avg. Joint G%	0.0120	0.0678	0.1843	0.1110	0.2093
CA%	0.0120	0.0805	0.1381	0.0531	0.1107
Avg. Joint AG%	0.0116	0.0113	0.0612	0.0309	0.1137
Avg. Tail U%	0.0113	0.2762	0.3142	0.2171	0.2334
Avg. Joint-Tail CS	0.0111	4.1960	2.0337	4.4271	1.6563
Avg. Joint GA%	0.0106	0.0165	0.0817	0.0383	0.1263
Avg. Tail GC%	0.0103	0.0149	0.0776	0.0339	0.1066
Avg. Joint-Tail U%	0.0099	0.2632	0.2957	0.2127	0.2003
A%	0.0087	0.2262	0.1440	0.2530	0.1412
Avg. Joint AU%	0.0084	0.0165	0.0799	0.0335	0.1051
Tail%	0.0084	0.5368	0.2499	0.4896	0.2665
Avg. Joint-Tail C%	0.0083	0.1953	0.2159	0.1595	0.1737
CG%	0.0072	0.0143	0.0642	0.0263	0.0769
Avg. Joint UA%	0.0069	0.0184	0.0753	0.0336	0.1051
UA%	0.0063	0.0518	0.1045	0.0693	0.1163
Avg. Tail Size	0.0063	3.7609	2.6397	4.2400	3.3848
NLC	0.0062	0.7580	0.1376	0.7783	0.1587

Avg. Joint AC%	0.0060	0.0175	0.0811	0.0323	0.1079
C%	0.0060	0.1334	0.1333	0.1148	0.1049
Avg. Tail GG%	0.0059	0.0207	0.0657	0.0346	0.1090
Avg. Joint CC%	0.0058	0.0064	0.0479	0.0168	0.0847
Avg. Joint GC%	0.0056	0.0067	0.0454	0.0187	0.1042
Avg. Joint-Tail GG%	0.0054	0.0223	0.0619	0.0342	0.0961
Avg. Joint-Tail UG%	0.0053	0.0198	0.0642	0.0304	0.0810
Avg. Tail UG%	0.0052	0.0188	0.0688	0.0306	0.0939
UG%	0.0051	0.0260	0.0802	0.0382	0.0906
Avg. Joint UC%	0.0050	0.0082	0.0583	0.0180	0.0788
Avg. Joint CA%	0.0050	0.0116	0.0570	0.0221	0.0887
Avg. Joint UU%	0.0050	0.0121	0.0675	0.0236	0.0936
U%	0.0048	0.1740	0.1789	0.1526	0.1199
GA%	0.0046	0.0651	0.1398	0.0842	0.1403
Avg. Joint-Tail CG%	0.0045	0.0117	0.0574	0.0198	0.0628
Avg. Tail FS	0.0042	1.3536	1.8006	1.1205	0.3185
Avg. Joint UG%	0.0039	0.0072	0.0470	0.0145	0.0680
Avg. Tail CS	0.0035	4.3571	2.4175	4.5407	1.8622
Avg. Tail CG%	0.0034	0.0108	0.0567	0.0181	0.0684
SLC	0.0032	0.7048	0.3902	0.6541	0.4770
Avg. Joint CG%	0.0029	0.0047	0.0464	0.0104	0.0580
Avg. Joint-Tail FS	0.0028	1.2912	1.4603	1.0847	0.2430
Avg. Joint CU%	0.0028	0.0084	0.0513	0.0144	0.0628
Avg. Joint-Tail A%	0.0021	0.3605	0.2634	0.3843	0.2414
GU% Bond	0.0020	0.0855	0.2406	0.1073	0.2553
Avg. Joint-Tail UA%	0.0016	0.0477	0.1033	0.0562	0.1094
Avg. Joint GG%	0.0016	0.0078	0.0540	0.0130	0.0758
GG%	0.0016	0.0332	0.0867	0.0412	0.1121
Avg. Joint-Tail SLC	0.0015	0.5353	0.2277	0.5527	0.2096
Avg. Tail SLC	0.0015	0.5163	0.2455	0.5360	0.2571
UC%	0.0010	0.0327	0.0966	0.0386	0.0874

Avg. Joint GU%	0.0010	0.0123	0.0746	0.0172	0.0783
GC% Bond	0.0007	0.6744	0.4181	0.6532	0.3897
Avg. Joint-Tail CU%	0.0006	0.0374	0.1056	0.0326	0.0806
AU%	0.0004	0.0679	0.1406	0.0734	0.1296
Avg. Joint-Tail UC%	0.0004	0.0273	0.0881	0.0307	0.0747
Avg. Tail CU%	0.0004	0.0377	0.1103	0.0334	0.0985
Avg. Tail A%	0.0003	0.3487	0.2821	0.3598	0.2825
CU%	0.0002	0.0441	0.1158	0.0408	0.0958
Avg. Tail UA%	0.0002	0.0492	0.1180	0.0525	0.1197
Avg. Tail AU%	0.0001	0.0627	0.1482	0.0592	0.1359
Avg. Joint-Tail AU%	3e-05	0.0610	0.1400	0.0595	0.1120
Avg. Tail UC%	3e-05	0.0276	0.0921	0.0285	0.0846
AU% Bond	1e-06	0.2402	0.3831	0.2396	0.3463

Table E.4: Experiment 1 External Loop Metric Statistics. Ranks each external loop metric from the first experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
Avg. Joint FS	0.0081	1.2161	1.5395	1.0177	0.2591
Avg. Joint-Tail FS	0.0081	1.2161	1.5395	1.0177	0.2591
Avg. Joint CC%	0.0023	0.0175	0.0562	0.0237	0.0702
Avg. Joint-Tail CC%	0.0023	0.0175	0.0562	0.0237	0.0702
CA%	0.0018	0.0618	0.1117	0.0529	0.0978
Joint%	0.0016	0.5883	0.1244	0.5984	0.1298
Joint-Tail%	0.0016	0.5883	0.1244	0.5984	0.1298
Avg. Joint C%	0.0014	0.1388	0.1458	0.1500	0.1540
Avg. Joint-Tail C%	0.0014	0.1388	0.1458	0.1500	0.1540
Avg. Joint Size	0.0014	3.8014	2.0079	3.9574	2.1261
Avg. Joint-Tail Size	0.0014	3.8014	2.0079	3.9574	2.1261
NLC	0.0013	0.8955	0.0787	0.8901	0.0831
Avg. Joint NLC	0.0013	0.9072	0.0766	0.9017	0.0820
Avg. Joint-Tail NLC	0.0013	0.9072	0.0766	0.9017	0.0820
Avg. Joint CA%	0.0013	0.0464	0.0915	0.0402	0.0808
Avg. Joint-Tail CA%	0.0013	0.0464	0.0915	0.0402	0.0808
C%	0.0013	0.0867	0.0759	0.0922	0.0795
Avg. Joint CS	0.0012	4.0294	1.9903	3.9119	1.3182
Avg. Joint-Tail CS	0.0012	4.0294	1.9903	3.9119	1.3182
CC%	0.0011	0.0241	0.0745	0.0294	0.0847
Avg. Joint GA%	0.0009	0.0822	0.1232	0.0751	0.1192
Avg. Joint-Tail GA%	0.0009	0.0822	0.1232	0.0751	0.1192
Avg. Joint GU%	0.0007	0.0342	0.0826	0.0300	0.0746
Avg. Joint-Tail GU%	0.0007	0.0342	0.0826	0.0300	0.0746
GU%	0.0007	0.0468	0.1043	0.0415	0.0978
GU% Bond	0.0007	0.1142	0.1787	0.1046	0.1746
SLC	0.0006	0.4759	0.1662	0.4679	0.1749
A%	0.0006	0.2704	0.1125	0.2759	0.1175
GC%	0.0004	0.0326	0.0894	0.0292	0.0807

Avg. Joint SLC	0.0004	0.5932	0.1588	0.5867	0.1653
Avg. Joint-Tail SLC	0.0004	0.5932	0.1588	0.5867	0.1653
GA%	0.0004	0.1003	0.1396	0.0949	0.1435
Avg. Joint UC%	0.0004	0.0217	0.0611	0.0241	0.0634
Avg. Joint-Tail UC%	0.0004	0.0217	0.0611	0.0241	0.0634
Avg. Joint AG%	0.0004	0.0564	0.1015	0.0604	0.1031
Avg. Joint-Tail AG%	0.0004	0.0564	0.1015	0.0604	0.1031
AU% Bond	0.0004	0.2615	0.2536	0.2714	0.2599
Avg. Joint UU%	0.0003	0.0380	0.0842	0.0348	0.0814
Avg. Joint-Tail UU%	0.0003	0.0380	0.0842	0.0348	0.0814
Avg. Joint GG%	0.0003	0.0245	0.0737	0.0274	0.0784
Avg. Joint-Tail GG%	0.0003	0.0245	0.0737	0.0274	0.0784
CU%	0.0003	0.0289	0.0730	0.0317	0.0758
Avg. Joint G%	0.0003	0.2172	0.1873	0.2108	0.1898
Avg. Joint-Tail G%	0.0003	0.2172	0.1873	0.2108	0.1898
UC%	0.0003	0.0300	0.0781	0.0328	0.0818
Avg. Joint CU%	0.0003	0.0208	0.0570	0.0229	0.0601
Avg. Joint-Tail CU%	0.0003	0.0208	0.0570	0.0229	0.0601
AG%	0.0003	0.0734	0.1202	0.0776	0.1235
Avg. Joint UA%	0.0003	0.0613	0.0951	0.0580	0.0901
Avg. Joint-Tail UA%	0.0003	0.0613	0.0951	0.0580	0.0901
Avg. Joint CG%	0.0003	0.0154	0.0494	0.0171	0.0522
Avg. Joint-Tail CG%	0.0003	0.0154	0.0494	0.0171	0.0522
Avg. Joint AC%	0.0002	0.0594	0.1055	0.0624	0.1068
Avg. Joint-Tail AC%	0.0002	0.0594	0.1055	0.0624	0.1068
UA%	0.0002	0.0824	0.1147	0.0789	0.1122
G%	0.0002	0.1188	0.0818	0.1168	0.0852
Avg. Joint U%	0.0002	0.1727	0.1678	0.1682	0.1691
Avg. Joint-Tail U%	0.0002	0.1727	0.1678	0.1682	0.1691
GG%	0.0001	0.0329	0.0936	0.0348	0.0998
AC%	0.0001	0.0744	0.1208	0.0765	0.1238

UU%	0.0001	0.0523	0.1067	0.0505	0.1104
CG%	0.0001	0.0219	0.0659	0.0233	0.0661
AA%	0.0001	0.2029	0.2054	0.2058	0.2119
U%	4e-05	0.1125	0.0889	0.1135	0.0927
Avg. Joint AA%	3e-05	0.1638	0.1779	0.1658	0.1792
Avg. Joint-Tail AA%	3e-05	0.1638	0.1779	0.1658	0.1792
Avg. Joint AU%	1e-05	0.0618	0.0984	0.0610	0.0958
Avg. Joint-Tail AU%	1e-05	0.0618	0.0984	0.0610	0.0958
UG%	1e-05	0.0362	0.0864	0.0354	0.0803
AU%	9e-06	0.0808	0.1175	0.0815	0.1169
Avg. Joint UG%	6e-06	0.0249	0.0623	0.0251	0.0623
Avg. Joint-Tail UG%	6e-06	0.0249	0.0623	0.0251	0.0623
Avg. Joint GC%	4e-06	0.0249	0.0669	0.0248	0.0710
Avg. Joint-Tail GC%	4e-06	0.0249	0.0669	0.0248	0.0710
Avg. Joint A%	1e-06	0.4712	0.2114	0.4707	0.2165
Avg. Joint-Tail A%	1e-06	0.4712	0.2114	0.4707	0.2165
Size	1e-06	17.4431	6.1393	17.4250	6.2932
GC% Bond	1e-06	0.6244	0.2797	0.6240	0.2834

Table E.5: Experiment 1 Multiloop Metric Statistics. Ranks each multiloop metric from the first experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
Avg. Joint-Tail FS	0.0075	1.2219	1.5157	1.0240	0.2583
Avg. Tail G%	0.0047	0.0107	0.0666	0.0222	0.1011
Avg. Joint FS	0.0038	1.2126	1.5011	1.0170	0.2524
Avg. Joint Size	0.0029	3.5529	2.1805	3.8010	2.3140
Avg. Joint CC%	0.0025	0.0166	0.0557	0.0230	0.0718
Joint-Tail%	0.0025	0.5917	0.1336	0.6051	0.1356
CA%	0.0024	0.0637	0.1146	0.0529	0.0991
Avg. Tail GU%	0.0024	0.0019	0.0280	0.0059	0.0510
Avg. Tail UU%	0.0024	0.0112	0.0857	0.0045	0.0432
Avg. Joint-Tail Size	0.0022	3.7943	2.0410	3.9991	2.2227
Avg. Tail AC%	0.0022	0.0080	0.0522	0.0038	0.0335
Avg. Joint-Tail CA%	0.0022	0.0487	0.0955	0.0403	0.0818
Avg. Tail CC%	0.0021	0.0063	0.0464	0.0026	0.0293
Avg. Joint C%	0.0021	0.1306	0.1481	0.1448	0.1589
Avg. Joint-Tail UU%	0.0020	0.0442	0.1077	0.0358	0.0838
Avg. Tail AA%	0.0016	0.0053	0.0478	0.0099	0.0659
Joint%	0.0016	0.5432	0.1922	0.5585	0.1854
Avg. Tail GA%	0.0013	0.0040	0.0349	0.0072	0.0529
Avg. Tail AG%	0.0012	0.0025	0.0241	0.0047	0.0385
Avg. Tail CA%	0.0010	0.0066	0.0471	0.0038	0.0354
Avg. Tail GC%	0.0010	0.0014	0.0241	0.0032	0.0344
Avg. Joint NLC	0.0010	0.8509	0.2331	0.8648	0.1977
A%	0.0010	0.2663	0.1160	0.2737	0.1201
UU%	0.0010	0.0585	0.1278	0.0512	0.1118
Avg. Joint CA%	0.0008	0.0432	0.0890	0.0385	0.0818
Avg. Joint-Tail AG%	0.0008	0.0538	0.0988	0.0595	0.1028
Avg. Joint AG%	0.0007	0.0522	0.0989	0.0576	0.1045
Avg. Tail GG%	0.0006	0.0019	0.0208	0.0033	0.0351
Avg. Joint UC%	0.0006	0.0206	0.0610	0.0235	0.0651

Avg. Joint-Tail GG%	0.0006	0.0243	0.0730	0.0280	0.0803
AG%	0.0006	0.0704	0.1171	0.0761	0.1223
Avg. Tail UG%	0.0006	0.0017	0.0214	0.0029	0.0303
Avg. Joint-Tail C%	0.0005	0.1437	0.1536	0.1509	0.1560
Avg. Joint-Tail U%	0.0005	0.1803	0.1848	0.1724	0.1728
Avg. Tail Size	0.0005	0.3404	1.3420	0.4033	1.6238
Avg. Joint GU%	0.0005	0.0322	0.0820	0.0287	0.0750
Avg. Joint A%	0.0005	0.4395	0.2372	0.4499	0.2375
Avg. Joint-Tail CG%	0.0005	0.0152	0.0503	0.0174	0.0533
Avg. Tail NLC	0.0005	0.0731	0.2436	0.0835	0.2655
Avg. Tail C%	0.0004	0.0175	0.0877	0.0140	0.0740
C%	0.0004	0.0910	0.0838	0.0944	0.0825
Avg. Joint CU%	0.0004	0.0196	0.0565	0.0221	0.0604
AA%	0.0004	0.1911	0.2038	0.1996	0.2112
Avg. Joint GA%	0.0004	0.0765	0.1222	0.0716	0.1204
UC%	0.0004	0.0302	0.0795	0.0334	0.0824
Avg. Joint-Tail UC%	0.0004	0.0223	0.0639	0.0247	0.0646
Avg. Joint GG%	0.0004	0.0230	0.0726	0.0260	0.0782
Avg. Joint-Tail CS	0.0004	4.0404	1.9785	3.9604	1.3620
Avg. Tail CG%	0.0004	0.0010	0.0172	0.0017	0.0218
Avg. Joint CG%	0.0004	0.0146	0.0493	0.0165	0.0528
Size	0.0004	16.7875	6.5333	17.0497	6.4837
Avg. Joint AC%	0.0004	0.0555	0.1039	0.0595	0.1073
Avg. Tail U%	0.0003	0.0247	0.1224	0.0207	0.0961
G%	0.0003	0.1167	0.0861	0.1199	0.0887
Avg. Joint-Tail AA%	0.0003	0.1549	0.1770	0.1611	0.1789
Avg. Joint-Tail GA%	0.0003	0.0792	0.1222	0.0749	0.1204
GU% Bond	0.0003	0.1110	0.1854	0.1049	0.1838
Avg. Tail CS	0.0003	4.3711	2.4512	4.5407	1.8622
Avg. Joint AA%	0.0003	0.1522	0.1775	0.1578	0.1825
AU% Bond	0.0003	0.2594	0.2678	0.2684	0.2695

Avg. Joint-Tail CC%	0.0002	0.0222	0.0693	0.0243	0.0722
CG%	0.0002	0.0217	0.0675	0.0236	0.0672
CU%	0.0002	0.0301	0.0781	0.0325	0.0780
NLC	0.0002	0.9281	0.1823	0.9243	0.1921
Avg. Tail SLC	0.0002	0.0465	0.1650	0.0510	0.1761
Avg. Joint-Tail SLC	0.0002	0.5880	0.1669	0.5835	0.1703
GG%	0.0002	0.0328	0.0926	0.0354	0.1011
Avg. Joint-Tail CU%	0.0002	0.0222	0.0629	0.0239	0.0624
Avg. Joint-Tail GC%	0.0002	0.0242	0.0669	0.0259	0.0747
Avg. Joint SLC	0.0001	0.5571	0.2140	0.5622	0.2049
GA%	0.0001	0.0975	0.1408	0.0939	0.1433
Avg. Joint UU%	0.0001	0.0354	0.0825	0.0337	0.0827
Avg. Tail A%	0.0001	0.0316	0.1311	0.0342	0.1369
Avg. Joint-Tail UA%	0.0001	0.0598	0.0959	0.0578	0.0921
Avg. Joint-Tail UG%	0.0001	0.0244	0.0624	0.0256	0.0643
CC%	0.0001	0.0288	0.0840	0.0301	0.0866
Avg. Joint UA%	0.0001	0.0572	0.0943	0.0557	0.0919
Avg. Joint-Tail NLC	0.0001	0.9005	0.1016	0.9020	0.0876
Avg. Joint G%	0.0001	0.2044	0.1921	0.2013	0.1939
Avg. Joint-Tail G%	0.0001	0.2101	0.1889	0.2127	0.1918
Avg. Tail UA%	0.0001	0.0044	0.0378	0.0050	0.0400
Avg. Joint UG%	0.0001	0.0232	0.0609	0.0241	0.0629
GU%	4e-05	0.0450	0.1032	0.0437	0.1027
UA%	4e-05	0.0791	0.1137	0.0780	0.1127
Avg. Joint GC%	3e-05	0.0234	0.0656	0.0242	0.0748
AC%	3e-05	0.0761	0.1227	0.0746	0.1226
Avg. Joint-Tail AC%	2e-05	0.0620	0.1096	0.0608	0.1058
GC%	2e-05	0.0314	0.0879	0.0305	0.0845
Tail%	2e-05	0.0485	0.1714	0.0466	0.1655
AU%	2e-05	0.0794	0.1199	0.0807	0.1182
Avg. Joint-Tail GU%	2e-05	0.0331	0.0822	0.0323	0.0806

Avg. Tail UC%	2e-05	0.0025	0.0285	0.0027	0.0274
Avg. Joint-Tail AU%	2e-05	0.0615	0.1030	0.0609	0.0974
GC% Bond	2e-05	0.6297	0.2949	0.6268	0.2953
Avg. Joint AU%	1e-05	0.0575	0.0979	0.0584	0.0970
UG%	1e-05	0.0352	0.0865	0.0357	0.0813
Avg. Tail FS	1e-05	1.3635	1.8457	1.1205	0.3185
U%	1e-05	0.1177	0.1020	0.1172	0.0963
Avg. Tail CU%	9e-06	0.0034	0.0346	0.0032	0.0319
Avg. Joint-Tail A%	7e-06	0.4612	0.2184	0.4625	0.2204
Avg. Joint CS	3e-06	4.0173	1.9700	3.9186	1.3719
Avg. Joint U%	1e-06	0.1624	0.1712	0.1623	0.1734
SLC	0e+00	0.4935	0.2016	0.4841	0.2245
Avg. Tail AU%	0e+00	0.0057	0.0476	0.0056	0.0453

Table E.6: Experiment 1 Junction Metric Statistics. Ranks each junction metric from the first experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		Nongene	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
Joint%	0.0921	0.2913	0.1826	0.3571	0.1783
PP	0.0513	0.6190	0.0463	0.5960	0.0548
% Num. Tail	0.0382	0.0286	0.0295	0.0288	0.0294
% Num. Stemloop	0.0228	0.0004	0.0016	0.0001	0.0001
% Num. Hairpin Loop	0.0228	0.0004	0.0016	0.0001	0.0001
% Num. Stem	0.0224	0.0008	0.0028	0.0002	0.0001
% Num. Bridge	0.0218	0.0003	0.0012	0.0001	0.0001
% Num. Multiloop	0.0218	0.0003	0.0012	0.0001	0.0001
Internal Loop%	0.0212	0.1463	0.0639	0.1650	0.0646
% Num. Stack	0.0189	0.0023	0.0090	0.0006	0.0004
% Num. Bulge	0.0168	0.0005	0.0018	0.0001	0.0001
% Num. Internal Loop	0.0152	0.0009	0.0038	0.0003	0.0002
Bulge%	0.0121	0.0331	0.0217	0.0381	0.0241
Multiloop%	0.0078	0.2000	0.0889	0.1852	0.0782
Stemloop%	0.0074	0.5701	0.1558	0.5439	0.1485
Stack%	0.0067	0.8480	0.1086	0.8279	0.1347
GU% Bond	0.0065	0.1232	0.0411	0.1299	0.0420
Bridge%	0.0043	0.2854	0.1291	0.3029	0.1377
Stem%	0.0043	0.7531	0.0761	0.7428	0.0814
NLC	0.0041	0.9937	0.0234	0.9959	0.0071
MFE	0.0016	-297.3350	319.0345	-272.8374	296.5180
% Num. Joint	0.0016	0.3012	0.0303	0.3036	0.0291
SLC	0.0012	0.9724	0.0297	0.9748	0.0392
Tail%	0.0010	0.0497	0.0878	0.0542	0.0904
Hairpin Loop%	0.0004	0.1023	0.0416	0.1039	0.0402
AU% Bond	0.0004	0.3102	0.1157	0.3057	0.1159
GC% Bond	0.0002	0.5666	0.1250	0.5630	0.1279
GU%	0e+00	0.0622	0.0148	0.0622	0.0148
AG%	0e+00	0.0718	0.0170	0.0718	0.0170

CG%	0e+00	0.0660	0.0271	0.0660	0.0271
GG%	0e+00	0.0933	0.0356	0.0933	0.0356
AU%	0e+00	0.0479	0.0242	0.0479	0.0242
A%	0e+00	0.2482	0.0520	0.2482	0.0520
UA%	0e+00	0.0512	0.0263	0.0512	0.0263
GA%	0e+00	0.0684	0.0169	0.0684	0.0169
GC%	0e+00	0.0719	0.0246	0.0719	0.0246
AA%	0e+00	0.0750	0.0308	0.0750	0.0308
UC%	0e+00	0.0500	0.0170	0.0500	0.0170
UG%	0e+00	0.0621	0.0178	0.0621	0.0178
AC%	0e+00	0.0533	0.0142	0.0533	0.0142
CU%	0e+00	0.0541	0.0159	0.0541	0.0159
C%	0e+00	0.2405	0.0496	0.2405	0.0496
U%	0e+00	0.2163	0.0563	0.2163	0.0563
CC%	0e+00	0.0660	0.0308	0.0660	0.0308
UU%	0e+00	0.0526	0.0325	0.0526	0.0325
G%	0e+00	0.2950	0.0549	0.2950	0.0549
CA%	0e+00	0.0541	0.0146	0.0541	0.0146
Size	0e+00	774.2023	794.3973	774.2023	794.3973

Table E.7: Experiment 1 Structure Metric Statistics. Ranks each structure metric from the first experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
FS	0.0035	2.8478	1.2532	2.8503	1.2897
Size	0.0029	9.2299	4.2774	8.7877	4.0410
SLC	0.0025	0.4351	0.1284	0.4479	0.1277
CS	0.0021	6.2301	1.9451	6.0954	1.9430
CC%	0.0011	0.0992	0.1994	0.0861	0.1887
GG%	0.0007	0.1294	0.2190	0.1183	0.2125
GC% Bond	0.0005	0.5772	0.2564	0.5651	0.2612
U%	0.0005	0.2114	0.1282	0.2175	0.1306
C%	0.0005	0.2886	0.1282	0.2825	0.1306
UG%	0.0005	0.0608	0.1661	0.0684	0.1740
UU%	0.0004	0.0420	0.1387	0.0475	0.1467
A%	0.0003	0.1523	0.1190	0.1566	0.1201
AU% Bond	0.0003	0.3046	0.2380	0.3132	0.2402
G%	0.0003	0.3477	0.1190	0.3434	0.1201
CG%	0.0003	0.0642	0.2323	0.0723	0.2435
GC%	0.0002	0.1198	0.3017	0.1108	0.2929
GA%	0.0002	0.0561	0.1578	0.0605	0.1630
AA%	0.0002	0.0208	0.0998	0.0234	0.1057
GU%	0.0001	0.0868	0.1894	0.0826	0.1859
GU% Bond	0.0001	0.1182	0.1528	0.1218	0.1577
CA%	0.0001	0.0372	0.1313	0.0400	0.1356
UC%	0.0001	0.0552	0.1567	0.0585	0.1607
CU%	2e-05	0.0627	0.1656	0.0644	0.1675
AG%	2e-05	0.0528	0.1537	0.0543	0.1555
UA%	2e-05	0.0337	0.1669	0.0351	0.1678
AU%	1e-05	0.0344	0.1631	0.0333	0.1610
NLC	0e+00	0.8879	0.1095	0.8878	0.1096
AC%	0e+00	0.0446	0.1425	0.0445	0.1424

Table E.8: Experiment 1 Stack Metric Statistics. Ranks each stack metric from the first experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		Nongene	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
FS	0.0165	1.1354	0.6542	1.0133	0.1865
GA%	0.0020	0.0821	0.2096	0.0644	0.1846
CS	0.0015	3.5832	1.8591	3.4386	1.8082
SLC	0.0015	0.6541	0.2693	0.6754	0.2778
AA%	0.0009	0.1319	0.2645	0.1168	0.2577
A%	0.0007	0.4151	0.3644	0.3962	0.3750
CC%	0.0005	0.0161	0.0916	0.0205	0.1090
G%	0.0005	0.2447	0.3207	0.2589	0.3442
CA%	0.0004	0.0382	0.1312	0.0329	0.1248
C%	0.0004	0.1454	0.2606	0.1561	0.2777
UU%	0.0004	0.0401	0.1525	0.0343	0.1399
UG%	0.0003	0.0393	0.1600	0.0341	0.1440
NLC	0.0003	0.9142	0.1241	0.9182	0.1256
Size	0.0001	3.0853	2.5950	3.0214	2.7226
AG%	0.0001	0.0487	0.1612	0.0525	0.1663
UA%	0.0001	0.0489	0.1449	0.0458	0.1440
GG%	0.0001	0.0297	0.1393	0.0326	0.1465
U%	0.0001	0.1947	0.2942	0.1888	0.2956
AC%	0.0001	0.0429	0.1442	0.0400	0.1395
CG%	0.0001	0.0179	0.0951	0.0196	0.1035
GU%	0.0001	0.0277	0.1124	0.0260	0.1135
AU%	3e-05	0.0443	0.1432	0.0427	0.1363
GC%	2e-05	0.0221	0.1057	0.0212	0.1087
CU%	2e-05	0.0211	0.1020	0.0218	0.1040
UC%	4e-06	0.0230	0.1063	0.0226	0.1085

Table E.9: Experiment 1 Unpaired Metric Statistics. Ranks each unpaired metric from the first experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		Nongene	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
AA%	0.0019	0.0664	0.2125	0.0490	0.1817
A%	0.0015	0.4395	0.4435	0.4069	0.4436
G%	0.0013	0.1394	0.3013	0.1637	0.3293
AG%	0.0011	0.0264	0.1368	0.0188	0.1029
CG%	0.0009	0.0055	0.0539	0.0095	0.0761
C%	0.0008	0.1512	0.3207	0.1679	0.3355
CC%	0.0007	0.0070	0.0672	0.0109	0.0867
NLC	0.0005	0.9583	0.1011	0.9629	0.0939
GU% Bond	0.0004	0.1138	0.3176	0.1083	0.3108
CU%	0.0004	0.0115	0.0851	0.0152	0.0983
UU%	0.0003	0.0275	0.1413	0.0226	0.1228
GG%	0.0003	0.0074	0.0659	0.0099	0.0805
AU%	0.0002	0.0316	0.1393	0.0272	0.1269
GC% Bond	0.0002	0.5765	0.4941	0.5874	0.4923
SLC	0.0002	0.8483	0.2320	0.8548	0.2345
GA%	0.0001	0.0260	0.1267	0.0237	0.1175
GC% Bond	0.0001	0.5765	0.4941	0.5874	0.4923
GU% Bond	0.0001	0.1138	0.3176	0.1083	0.3108
AC%	0.0001	0.0211	0.1146	0.0191	0.1064
U%	0.0001	0.2699	0.4003	0.2615	0.3954
Size	0.0001	1.7582	1.5168	1.7822	1.6916
GU%	5e-05	0.0128	0.0865	0.0139	0.0908
UG%	4e-05	0.0130	0.0877	0.0147	0.0951
CA%	3e-05	0.0169	0.0937	0.0179	0.1033
GC%	3e-05	0.0081	0.0703	0.0074	0.0673
UC%	2e-05	0.0144	0.0952	0.0135	0.0907
AU% Bond	2e-05	0.3096	0.4624	0.3043	0.4601
AU% Bond	2e-06	0.3096	0.4624	0.3043	0.4601
UA%	0e+00	0.0277	0.1312	0.0275	0.1276

Table E.10: Experiment 1 Bulge Metric Statistics. Ranks each bulge metric from the first experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
GA%	0.0038	0.0960	0.2268	0.0705	0.1935
SLC	0.0016	0.7232	0.2148	0.7420	0.2224
UG%	0.0015	0.0577	0.2090	0.0429	0.1704
A%	0.0011	0.3716	0.2932	0.3510	0.2961
CC%	0.0011	0.0142	0.0831	0.0201	0.1009
G%	0.0008	0.2989	0.3023	0.3162	0.3296
AC%	0.0007	0.0389	0.1258	0.0323	0.1147
AA%	0.0006	0.1213	0.2297	0.1086	0.2249
CG%	0.0005	0.0179	0.0868	0.0217	0.0952
Size	0.0005	4.2036	2.4361	4.0861	2.5457
C%	0.0005	0.1447	0.2107	0.1545	0.2215
GU% Bond	0.0004	0.1565	0.3633	0.1551	0.3620
AU% Bond	0.0003	0.3424	0.4745	0.3328	0.4712
AG%	0.0003	0.0558	0.1627	0.0607	0.1729
CA%	0.0002	0.0341	0.1176	0.0307	0.1143
U%	0.0002	0.1848	0.2711	0.1783	0.2705
UA%	0.0001	0.0432	0.1201	0.0403	0.1164
GC%	0.0001	0.0186	0.0965	0.0207	0.1041
GC% Bond	0.0001	0.5012	0.5000	0.5121	0.4999
SR	0.0001	0.8114	0.2616	0.8068	0.2701
AU% Bond	0.0001	0.3424	0.4745	0.3328	0.4712
NLC	0.0001	0.9196	0.1071	0.9219	0.1087
UU%	4e-05	0.0392	0.1548	0.0372	0.1470
UC%	4e-05	0.0212	0.1024	0.0227	0.1031
GU% Bond	1e-05	0.1565	0.3633	0.1551	0.3620
GG%	1e-05	0.0442	0.1531	0.0430	0.1527
GU%	9e-06	0.0220	0.0972	0.0214	0.0983
AU%	3e-06	0.0345	0.1048	0.0341	0.1063
GC% Bond	2e-06	0.5012	0.5000	0.5121	0.4999
CU%	1e-06	0.0221	0.1062	0.0218	0.1010

Table E.11: Experiment 1 Internal Loop Metric Statistics. Ranks each internal loop metric from the first experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
GA%	0.0023	0.0726	0.2017	0.0546	0.1724
FS	0.0018	4.8341	1.9004	4.7395	1.7864
CS	0.0015	5.5490	2.1696	5.4169	2.0935
A%	0.0012	0.3943	0.3522	0.3704	0.3536
CC%	0.0009	0.0118	0.0782	0.0172	0.0976
AA%	0.0009	0.1029	0.2256	0.0894	0.2141
SLC	0.0009	0.7650	0.2285	0.7798	0.2326
G%	0.0009	0.2455	0.3112	0.2650	0.3370
UG%	0.0008	0.0427	0.1791	0.0331	0.1489
CG%	0.0006	0.0138	0.0776	0.0176	0.0897
C%	0.0006	0.1469	0.2529	0.1591	0.2662
AC%	0.0005	0.0329	0.1225	0.0279	0.1123
GU% Bond	0.0004	0.1422	0.3493	0.1382	0.3451
Size	0.0002	3.3857	2.4598	3.3126	2.5404
NLC	0.0002	0.9326	0.1067	0.9354	0.1060
AU% Bond	0.0001	0.3314	0.4707	0.3247	0.4683
GC% Bond	0.0001	0.5264	0.4993	0.5371	0.4986
U%	0.0001	0.2133	0.3227	0.2056	0.3204
UU%	0.0001	0.0353	0.1505	0.0322	0.1390
CA%	0.0001	0.0283	0.1105	0.0261	0.1103
UA%	0.0001	0.0380	0.1241	0.0360	0.1207
AU% Bond	0.0001	0.3314	0.4707	0.3247	0.4683
AU%	5e-05	0.0335	0.1175	0.0317	0.1139
GC%	5e-05	0.0151	0.0887	0.0165	0.0947
GU% Bond	3e-05	0.1422	0.3493	0.1382	0.3451
CU%	3e-05	0.0186	0.0998	0.0195	0.1001
GC% Bond	1e-05	0.5264	0.4993	0.5371	0.4986
AG%	1e-05	0.0460	0.1552	0.0471	0.1550
UC%	8e-06	0.0189	0.1001	0.0192	0.0983
GG%	0e+00	0.0319	0.1318	0.0320	0.1343
GU%	0e+00	0.0190	0.0939	0.0190	0.0961

Table E.12: Experiment 1 Loop Metric Statistics. Ranks each loop metric from the first experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
CS	0.0232	51.3026	20.7227	58.1448	25.0076
FS	0.0215	47.0594	20.6320	53.5970	24.9224
GG%	0.0065	0.0337	0.1078	0.0532	0.1357
AG%	0.0058	0.0550	0.1123	0.0730	0.1265
SLC	0.0051	0.3867	0.0799	0.3746	0.0865
Size	0.0051	5.2336	2.0237	5.5494	2.3551
AA%	0.0047	0.1887	0.2227	0.1587	0.2191
GC% Bond	0.0037	0.6890	0.4629	0.6320	0.4823
AU%	0.0032	0.0552	0.1079	0.0685	0.1213
AU% Bond	0.0027	0.2066	0.4049	0.2491	0.4325
GA%	0.0023	0.1194	0.1648	0.1041	0.1515
CC%	0.0016	0.0274	0.0933	0.0354	0.1081
UU%	0.0015	0.0705	0.1635	0.0581	0.1417
AC%	0.0012	0.0519	0.1093	0.0595	0.1147
UC%	0.0009	0.0451	0.1035	0.0393	0.0976
C%	0.0009	0.1580	0.1705	0.1685	0.1784
GC%	0.0007	0.0466	0.1092	0.0407	0.1001
UG%	0.0007	0.0461	0.1033	0.0516	0.1107
CU%	0.0006	0.0353	0.0938	0.0397	0.0972
CA%	0.0006	0.0616	0.1183	0.0565	0.1126
A%	0.0005	0.3875	0.2250	0.3775	0.2246
GU% Bond	0.0005	0.1044	0.3057	0.1189	0.3237
NLC	0.0005	0.8680	0.1070	0.8730	0.1098
G%	0.0001	0.2321	0.1776	0.2364	0.1944
CG%	0.0001	0.0384	0.0993	0.0365	0.0955
U%	0.0001	0.2223	0.2195	0.2177	0.2056
GU%	2e-05	0.0515	0.1076	0.0507	0.1086
UA%	1e-05	0.0737	0.1240	0.0743	0.1245

Table E.13: Experiment 1 Hairpin Loop Metric Statistics. Ranks each Hairpin Loop metric from the first experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
G%	0.0385	0.1401	0.2446	0.2480	0.2935
UU%	0.0300	0.1234	0.2741	0.0469	0.1463
NLC	0.0284	0.8745	0.1464	0.9189	0.1124
AC%	0.0190	0.0844	0.1788	0.0421	0.1251
GU%	0.0181	0.0231	0.1088	0.0637	0.1790
CC%	0.0142	0.0596	0.1389	0.0284	0.1142
FS	0.0133	124.3385	250.8350	72.2504	189.4729
CS	0.0130	127.2597	251.1728	75.5594	189.7670
U%	0.0112	0.2901	0.3409	0.2236	0.2706
AA%	0.0087	0.0692	0.1869	0.1071	0.2245
CA%	0.0082	0.0664	0.1492	0.0413	0.1227
GC%	0.0077	0.0159	0.0939	0.0372	0.1372
C%	0.0053	0.1907	0.2507	0.1546	0.2353
AG%	0.0052	0.0338	0.1030	0.0519	0.1368
GA%	0.0039	0.0569	0.1603	0.0797	0.1926
CG%	0.0033	0.0106	0.0610	0.0187	0.0810
Size	0.0027	4.0368	3.1692	4.4015	3.9832
UG%	0.0020	0.0227	0.0886	0.0322	0.1134
GG%	0.0019	0.0262	0.0926	0.0365	0.1364
AU%	0.0003	0.0664	0.1735	0.0592	0.1556
CU%	0.0003	0.0394	0.1199	0.0347	0.1199
A%	0.0001	0.3791	0.3210	0.3738	0.3315
SLC	0.0001	0.5576	0.2631	0.5632	0.2838
UC%	4e-05	0.0310	0.1088	0.0299	0.1018
UA%	4e-06	0.0557	0.1519	0.0544	0.1389

Table E.14: Experiment 1 Tail Metric Statistics. Ranks each tail metric from the first experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
FS	0.0065	1.2256	1.7359	1.0192	0.3363
CC%	0.0009	0.0174	0.0928	0.0240	0.1141
Size	0.0008	3.7328	3.3283	3.9320	3.5174
C%	0.0008	0.1376	0.2314	0.1512	0.2446
NLC	0.0005	0.9079	0.1244	0.9021	0.1305
CS	0.0004	3.9841	2.6768	3.8850	2.1464
GA%	0.0004	0.0826	0.2027	0.0746	0.1888
SLC	0.0003	0.6001	0.2722	0.5895	0.2770
CA%	0.0003	0.0452	0.1438	0.0404	0.1325
G%	0.0003	0.2208	0.2948	0.2107	0.2923
UC%	0.0002	0.0214	0.0962	0.0243	0.1033
GU%	0.0002	0.0325	0.1234	0.0291	0.1164
AC%	0.0002	0.0577	0.1642	0.0626	0.1735
GG%	0.0001	0.0246	0.1174	0.0272	0.1258
AG%	0.0001	0.0576	0.1689	0.0604	0.1670
CU%	0.0001	0.0208	0.0942	0.0227	0.0955
CG%	0.0001	0.0155	0.0851	0.0170	0.0872
UU%	0.0001	0.0359	0.1269	0.0339	0.1215
U%	3e-05	0.1688	0.2497	0.1656	0.2458
UA%	3e-05	0.0595	0.1515	0.0576	0.1485
GC%	8e-06	0.0251	0.1170	0.0247	0.1179
AA%	7e-06	0.1640	0.2876	0.1661	0.2912
AU%	3e-06	0.0596	0.1610	0.0601	0.1564
A%	3e-06	0.4729	0.3438	0.4724	0.3465
UG%	0e+00	0.0246	0.1039	0.0245	0.0989

Table E.15: Experiment 1 Joint Metric Statistics. Ranks each joint metric from the first experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
FS	0.0064	1.2256	1.7255	1.0239	0.3305
Size	0.0009	3.7541	3.3255	3.9565	3.5443
C%	0.0005	0.1406	0.2327	0.1515	0.2442
CC%	0.0005	0.0196	0.0967	0.0242	0.1142
CA%	0.0004	0.0466	0.1448	0.0405	0.1321
UU%	0.0004	0.0395	0.1375	0.0346	0.1231
SLC	0.0003	0.5975	0.2720	0.5881	0.2773
GA%	0.0003	0.0811	0.2004	0.0748	0.1889
CS	0.0003	4.0126	2.6767	3.9314	2.1560
UC%	0.0002	0.0218	0.0966	0.0246	0.1032
NLC	0.0002	0.9063	0.1257	0.9029	0.1296
GG%	0.0002	0.0245	0.1163	0.0277	0.1264
AG%	0.0001	0.0565	0.1660	0.0600	0.1655
CG%	0.0001	0.0153	0.0841	0.0171	0.0869
U%	0.0001	0.1740	0.2556	0.1687	0.2475
CU%	0.0001	0.0217	0.0952	0.0234	0.0969
AC%	5e-05	0.0593	0.1658	0.0615	0.1713
G%	5e-05	0.2161	0.2925	0.2125	0.2925
UA%	3e-05	0.0592	0.1509	0.0575	0.1481
AA%	2e-05	0.1602	0.2849	0.1631	0.2884
GU%	2e-05	0.0318	0.1219	0.0309	0.1206
A%	9e-06	0.4693	0.3432	0.4673	0.3463
UG%	7e-06	0.0244	0.1029	0.0249	0.0996
GC%	2e-06	0.0250	0.1165	0.0253	0.1189
AU%	1e-06	0.0601	0.1617	0.0602	0.1565

Table E.16: Experiment 1 Joint-Tail Metric Statistics. Ranks each joint-tail metric from the first experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

E.1.2 Experiment 2

		Gene		Nongene	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
GU% Bond	0.0054	0.1128	0.1780	0.0875	0.1683
C%	0.0054	0.0939	0.0811	0.1056	0.0862
GU%	0.0050	0.0555	0.1204	0.0396	0.0925
Avg. Joint C%	0.0042	0.1524	0.1460	0.1702	0.1623
Avg. Joint-Tail C%	0.0042	0.1524	0.1460	0.1702	0.1623
UC%	0.0039	0.0336	0.0842	0.0425	0.0998
Avg. Joint CC%	0.0035	0.0230	0.0653	0.0313	0.0807
Avg. Joint-Tail CC%	0.0035	0.0230	0.0653	0.0313	0.0807
Avg. Joint UC%	0.0031	0.0230	0.0597	0.0291	0.0747
Avg. Joint-Tail UC%	0.0031	0.0230	0.0597	0.0291	0.0747
Avg. Joint G%	0.0030	0.2319	0.2054	0.2121	0.2002
Avg. Joint-Tail G%	0.0030	0.2319	0.2054	0.2121	0.2002
G%	0.0029	0.1259	0.0902	0.1166	0.0907
Joint%	0.0023	0.5790	0.1341	0.5904	0.1296
Joint-Tail%	0.0023	0.5790	0.1341	0.5904	0.1296
CC%	0.0021	0.0311	0.0804	0.0396	0.0968
Avg. Joint U%	0.0021	0.1607	0.1726	0.1776	0.1903
Avg. Joint-Tail U%	0.0021	0.1607	0.1726	0.1776	0.1903
U%	0.0020	0.1051	0.0922	0.1124	0.0936
UU%	0.0018	0.0444	0.1043	0.0512	0.1201
Avg. Joint SLC	0.0017	0.6080	0.1683	0.5965	0.1646
Avg. Joint-Tail SLC	0.0017	0.6080	0.1683	0.5965	0.1646
Avg. Joint GU%	0.0015	0.0385	0.0995	0.0314	0.0818
Avg. Joint-Tail GU%	0.0015	0.0385	0.0995	0.0314	0.0818
AG%	0.0014	0.0799	0.1213	0.0715	0.1092
Avg. Joint UU%	0.0013	0.0330	0.0817	0.0369	0.0880
Avg. Joint-Tail UU%	0.0013	0.0330	0.0817	0.0369	0.0880
SLC	0.0013	0.4919	0.1810	0.4805	0.1728

Avg. Joint NLC	0.0012	0.9099	0.0746	0.9056	0.0792
Avg. Joint-Tail NLC	0.0012	0.9099	0.0746	0.9056	0.0792
Avg. Joint A%	0.0011	0.4550	0.2169	0.4401	0.2259
Avg. Joint-Tail A%	0.0011	0.4550	0.2169	0.4401	0.2259
Avg. Joint UA%	0.0011	0.0555	0.0911	0.0508	0.0852
Avg. Joint-Tail UA%	0.0011	0.0555	0.0911	0.0508	0.0852
Avg. Joint CG%	0.0010	0.0168	0.0509	0.0203	0.0599
Avg. Joint-Tail CG%	0.0010	0.0168	0.0509	0.0203	0.0599
CU%	0.0010	0.0296	0.0720	0.0347	0.0880
Avg. Joint CS	0.0010	4.0928	1.4273	4.1756	1.3032
Avg. Joint-Tail CS	0.0010	4.0928	1.4273	4.1756	1.3032
Avg. Joint CU%	0.0009	0.0218	0.0589	0.0255	0.0716
Avg. Joint-Tail CU%	0.0009	0.0218	0.0589	0.0255	0.0716
UA%	0.0009	0.0735	0.1183	0.0686	0.1095
AA%	0.0008	0.2038	0.2438	0.1917	0.2183
Size	0.0007	16.1238	5.5751	16.3925	5.7299
Avg. Joint Size	0.0007	3.7075	2.0378	3.7981	2.0742
Avg. Joint-Tail Size	0.0007	3.7075	2.0378	3.7981	2.0742
Avg. Joint AA%	0.0007	0.1589	0.1834	0.1502	0.1769
Avg. Joint-Tail AA%	0.0007	0.1589	0.1834	0.1502	0.1769
GC% Bond	0.0006	0.6538	0.2907	0.6677	0.2813
NLC	0.0005	0.8961	0.0812	0.8931	0.0860
AU% Bond	0.0005	0.2334	0.2596	0.2448	0.2555
CA%	0.0005	0.0557	0.1168	0.0606	0.1040
Avg. Joint CA%	0.0005	0.0409	0.0862	0.0446	0.0861
Avg. Joint-Tail CA%	0.0005	0.0409	0.0862	0.0446	0.0861
Avg. Joint GC%	0.0005	0.0310	0.0835	0.0349	0.0858
Avg. Joint-Tail GC%	0.0005	0.0310	0.0835	0.0349	0.0858
Avg. Joint GG%	0.0004	0.0259	0.0766	0.0299	0.0839
Avg. Joint-Tail GG%	0.0004	0.0259	0.0766	0.0299	0.0839
Avg. Joint AU%	0.0003	0.0504	0.0928	0.0545	0.0913

Avg. Joint-Tail AU%	0.0003	0.0504	0.0928	0.0545	0.0913
Avg. Joint AG%	0.0003	0.0602	0.1062	0.0582	0.1034
Avg. Joint-Tail AG%	0.0003	0.0602	0.1062	0.0582	0.1034
GG%	0.0002	0.0395	0.1059	0.0374	0.1022
AU%	0.0001	0.0708	0.1330	0.0756	0.1231
Avg. Joint AC%	0.0001	0.0601	0.1072	0.0572	0.0999
Avg. Joint-Tail AC%	0.0001	0.0601	0.1072	0.0572	0.0999
UG%	0.0001	0.0320	0.0689	0.0336	0.0754
Avg. Joint UG%	0.0001	0.0234	0.0611	0.0246	0.0622
Avg. Joint-Tail UG%	0.0001	0.0234	0.0611	0.0246	0.0622
GC%	0.0001	0.0382	0.0977	0.0416	0.1047
AC%	0.0001	0.0731	0.1322	0.0720	0.1186
Avg. Joint FS	0.0001	1.0394	0.7753	1.0310	0.3636
Avg. Joint-Tail FS	0.0001	1.0394	0.7753	1.0310	0.3636
A%	0.0001	0.2541	0.1112	0.2558	0.1200
Avg. Joint GA%	2e-05	0.0679	0.1135	0.0649	0.1059
Avg. Joint-Tail GA%	2e-05	0.0679	0.1135	0.0649	0.1059
CG%	1e-05	0.0257	0.0834	0.0267	0.0712
GA%	1e-06	0.0871	0.1346	0.0868	0.1347

Table E.17: Experiment 2 Multiloop Metric Statistics. Ranks each multiloop metric from the second experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		Nongene	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
Joint%	0.0045	0.1616	0.2061	0.1346	0.1954
Avg. Joint NLC	0.0043	0.4525	0.4588	0.3927	0.4549
Avg. Joint CS	0.0041	4.6924	2.5529	4.5974	2.5241
Tail%	0.0041	0.5486	0.2662	0.5825	0.2651
CG%	0.0040	0.0264	0.0683	0.0360	0.0836
Avg. Joint FS	0.0031	1.0466	0.9819	1.0459	1.0900
Avg. Joint SLC	0.0028	0.2681	0.3254	0.2341	0.3196
Avg. Joint Size	0.0027	2.4026	3.7966	2.0175	3.5728
Avg. Joint-Tail CG%	0.0027	0.0229	0.0638	0.0303	0.0778
Avg. Joint A%	0.0023	0.2042	0.2927	0.1765	0.2854
Avg. Tail C%	0.0021	0.1719	0.2059	0.1914	0.2168
Size	0.0021	14.3156	7.6785	13.6317	7.2071
Avg. Joint AA%	0.0020	0.0763	0.1843	0.0603	0.1747
Avg. Joint UA%	0.0019	0.0350	0.1140	0.0259	0.0923
C%	0.0019	0.1272	0.1102	0.1369	0.1135
Avg. Joint G%	0.0019	0.1191	0.2299	0.0998	0.2182
Avg. Tail CC%	0.0016	0.0262	0.0868	0.0339	0.1044
Avg. Tail CG%	0.0014	0.0238	0.0826	0.0302	0.0854
Avg. Joint-Tail CC%	0.0014	0.0282	0.0812	0.0347	0.0931
Avg. Joint-Tail C%	0.0014	0.1761	0.1839	0.1899	0.1895
Avg. Joint C%	0.0011	0.0873	0.1900	0.0750	0.1780
Avg. Joint-Tail CA%	0.0011	0.0474	0.0952	0.0414	0.0850
CC%	0.0011	0.0334	0.0911	0.0396	0.1013
Avg. Tail UU%	0.0010	0.0552	0.1314	0.0473	0.1170
Avg. Joint AG%	0.0010	0.0325	0.1212	0.0254	0.1035
Avg. Joint CU%	0.0008	0.0167	0.0697	0.0128	0.0659
Avg. Tail UC%	0.0008	0.0309	0.0802	0.0360	0.0971
Avg. Tail U%	0.0008	0.2201	0.2232	0.2080	0.2136
Avg. Joint CA%	0.0008	0.0226	0.0823	0.0182	0.0780

AA%	0.0007	0.1361	0.1818	0.1269	0.1682
CA%	0.0006	0.0558	0.0990	0.0509	0.0954
GC% Bond	0.0006	0.6171	0.4162	0.6381	0.4208
Avg. Joint UC%	0.0006	0.0161	0.0780	0.0126	0.0647
Avg. Tail CA%	0.0006	0.0452	0.1064	0.0403	0.0969
GC%	0.0006	0.0374	0.0922	0.0420	0.1016
Avg. Joint-Tail UU%	0.0006	0.0516	0.1133	0.0465	0.1009
Avg. Joint U%	0.0005	0.0903	0.1759	0.0824	0.1804
Avg. Tail UG%	0.0005	0.0434	0.1169	0.0387	0.0954
UC%	0.0005	0.0386	0.0799	0.0424	0.0939
Joint-Tail%	0.0005	0.7102	0.1583	0.7171	0.1616
Avg. Joint GG%	0.0005	0.0218	0.1161	0.0172	0.0952
Avg. Joint-Tail U%	0.0005	0.2171	0.1969	0.2088	0.1927
Avg. Tail GU%	0.0004	0.0503	0.1291	0.0453	0.1125
Avg. Joint-Tail UG%	0.0004	0.0428	0.1044	0.0391	0.0838
CU%	0.0004	0.0478	0.0961	0.0440	0.0963
UG%	0.0004	0.0532	0.1160	0.0490	0.1000
UU%	0.0004	0.0632	0.1319	0.0584	0.1166
Avg. Joint-Tail AA%	0.0004	0.1136	0.1630	0.1076	0.1497
Avg. Joint-Tail GC%	0.0003	0.0340	0.0886	0.0375	0.0988
Avg. Joint UG%	0.0003	0.0166	0.0809	0.0138	0.0697
Avg. Joint AC%	0.0003	0.0305	0.1028	0.0269	0.1036
GG%	0.0003	0.0550	0.1363	0.0505	0.1227
AU% Bond	0.0003	0.2458	0.3650	0.2332	0.3669
NLC	0.0003	0.2414	0.3956	0.2657	0.4169
Avg. Joint-Tail UC%	0.0002	0.0331	0.0760	0.0357	0.0893
Avg. Joint-Tail A%	0.0002	0.3435	0.2271	0.3369	0.2222
GU% Bond	0.0002	0.1371	0.2983	0.1287	0.2951
Avg. Joint-Tail CU%	0.0002	0.0387	0.0845	0.0363	0.0875
Avg. Joint-Tail AU%	0.0002	0.0580	0.1134	0.0553	0.0903
Avg. Joint UU%	0.0001	0.0159	0.0746	0.0176	0.0785

Avg. Joint-Tail GU%	0.0001	0.0497	0.1129	0.0474	0.1029
AU%	0.0001	0.0687	0.1217	0.0664	0.0988
U%	0.0001	0.1647	0.1257	0.1622	0.1238
Avg. Tail AU%	0.0001	0.0551	0.1210	0.0529	0.1051
Avg. Tail A%	0.0001	0.3233	0.2578	0.3182	0.2506
G%	0.0001	0.1672	0.1223	0.1695	0.1229
Avg. Tail GA%	0.0001	0.0646	0.1405	0.0619	0.1333
Avg. Tail G%	0.0001	0.2549	0.2466	0.2502	0.2473
A%	0.0001	0.2511	0.1467	0.2485	0.1434
GA%	0.0001	0.0716	0.1130	0.0735	0.1237
Avg. Joint AU%	0.0001	0.0299	0.1071	0.0283	0.1072
Avg. Tail AG%	0.0001	0.0621	0.1278	0.0602	0.1233
AC%	0.0001	0.0635	0.1098	0.0619	0.1089
Avg. Joint-Tail SLC	0.0001	0.5171	0.2075	0.5141	0.2125
Avg. Tail AA%	0.0001	0.1021	0.1789	0.0996	0.1660
Avg. Tail SLC	4e-05	0.5103	0.2438	0.5071	0.2465
Avg. Joint-Tail CS	3e-05	5.0212	2.2131	5.0046	2.0481
SLC	2e-05	0.5865	0.4335	0.5915	0.4530
Avg. Joint-Tail AG%	2e-05	0.0626	0.1115	0.0616	0.1115
Avg. Tail GC%	2e-05	0.0353	0.1090	0.0362	0.1059
Avg. Joint CG%	2e-05	0.0097	0.0634	0.0092	0.0548
Avg. Joint-Tail UA%	1e-05	0.0561	0.0925	0.0554	0.0927
Avg. Tail CS	1e-05	5.0123	2.3627	5.0090	2.1911
Avg. Joint-Tail GG%	1e-05	0.0443	0.1089	0.0451	0.1147
UA%	1e-05	0.0684	0.1015	0.0677	0.1018
Avg. Tail NLC	9e-06	0.8887	0.1817	0.8876	0.1853
Avg. Joint GA%	7e-06	0.0316	0.1066	0.0309	0.1225
Avg. Joint CC%	7e-06	0.0148	0.0796	0.0153	0.0862
Avg. Joint-Tail Size	6e-06	5.1945	3.8196	5.1764	3.4450
Avg. Joint GC%	6e-06	0.0137	0.0734	0.0133	0.0734
AG%	5e-06	0.0712	0.1153	0.0706	0.1185

Avg. Tail UA%	4e-06	0.0534	0.1122	0.0529	0.1025
Avg. Tail AC%	3e-06	0.0479	0.1058	0.0483	0.1082
GU%	3e-06	0.0566	0.1140	0.0570	0.1169
Avg. Joint-Tail FS	3e-06	1.1970	0.8362	1.2027	0.7024
Avg. Joint GU%	2e-06	0.0179	0.0798	0.0181	0.0885
Avg. Tail FS	2e-06	1.2502	0.8029	1.2535	0.7011
Avg. Joint-Tail GA%	2e-06	0.0648	0.1185	0.0651	0.1222
Avg. Joint-Tail AC%	1e-06	0.0520	0.0970	0.0518	0.0981
Avg. Tail CU%	1e-06	0.0376	0.0909	0.0374	0.1002
Avg. Joint-Tail G%	1e-06	0.2564	0.2223	0.2560	0.2280
Avg. Joint-Tail NLC	0e+00	0.9079	0.1115	0.9077	0.1162
Avg. Tail Size	0e+00	5.0042	4.1801	4.9995	3.7536
Avg. Tail GG%	0e+00	0.0436	0.1276	0.0437	0.1234

Table E.18: Experiment 2 External Loop Metric Statistics. Ranks each external loop metric from the second experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
Avg. Joint-Tail CG%	0.0021	0.0212	0.0606	0.0276	0.0740
C%	0.0020	0.1182	0.1041	0.1275	0.1068
CG%	0.0020	0.0262	0.0727	0.0332	0.0804
Avg. Joint-Tail C%	0.0018	0.1696	0.1747	0.1851	0.1826
Avg. Joint-Tail CC%	0.0018	0.0268	0.0772	0.0331	0.0879
Avg. Joint G%	0.0015	0.1498	0.2290	0.1331	0.2190
Avg. Joint UA%	0.0014	0.0406	0.1086	0.0327	0.0898
CC%	0.0013	0.0328	0.0883	0.0388	0.0981
Avg. Joint NLC	0.0012	0.5769	0.4429	0.5436	0.4495
Avg. Tail U%	0.0012	0.1603	0.2142	0.1453	0.2023
Avg. Joint CS	0.0012	4.4363	2.1655	4.3931	2.0400
UC%	0.0012	0.0373	0.0811	0.0437	0.0986
Avg. Tail UU%	0.0011	0.0402	0.1148	0.0327	0.1004
Avg. Joint AA%	0.0010	0.0987	0.1877	0.0872	0.1801
Avg. Joint FS	0.0009	1.0435	0.8992	1.0394	0.8278
Avg. Joint Size	0.0009	2.7574	3.4581	2.5494	3.3170
Avg. Joint A%	0.0008	0.2724	0.2960	0.2548	0.2957
Avg. Joint SLC	0.0008	0.3605	0.3281	0.3408	0.3273
Avg. Tail CC%	0.0008	0.0191	0.0750	0.0232	0.0872
Avg. Tail NLC	0.0008	0.6471	0.4248	0.6230	0.4351
Avg. Tail CG%	0.0007	0.0173	0.0712	0.0214	0.0737
GU% Bond	0.0007	0.1305	0.2711	0.1160	0.2636
Avg. Tail CA%	0.0007	0.0329	0.0930	0.0287	0.0843
GC% Bond	0.0007	0.6270	0.3865	0.6483	0.3847
Avg. Tail SLC	0.0007	0.3715	0.3079	0.3565	0.3110
Avg. Tail CS	0.0006	5.0123	2.3627	4.9957	2.1968
Avg. Tail FS	0.0006	1.2502	0.8029	1.2546	0.7111
Avg. Joint-Tail UC%	0.0006	0.0304	0.0720	0.0345	0.0867
Avg. Tail UG%	0.0006	0.0316	0.1016	0.0272	0.0827

Size	0.0006	14.8073	7.2120	14.4150	6.9131
Avg. Joint AG%	0.0006	0.0400	0.1180	0.0350	0.1045
Avg. Tail A%	0.0006	0.2354	0.2628	0.2235	0.2559
Avg. Tail GU%	0.0005	0.0366	0.1124	0.0325	0.0979
Avg. Tail C%	0.0005	0.1251	0.1916	0.1345	0.2018
AA%	0.0005	0.1546	0.2028	0.1463	0.1875
Avg. Joint UU%	0.0005	0.0205	0.0770	0.0240	0.0840
Avg. Tail G%	0.0004	0.1856	0.2390	0.1755	0.2365
Avg. Joint CC%	0.0004	0.0171	0.0761	0.0197	0.0837
GC%	0.0004	0.0376	0.0937	0.0414	0.1022
Avg. Joint-Tail GU%	0.0004	0.0466	0.1096	0.0430	0.0976
GU%	0.0004	0.0563	0.1158	0.0527	0.1117
Avg. Joint-Tail GC%	0.0003	0.0332	0.0873	0.0368	0.0960
GG%	0.0003	0.0508	0.1289	0.0469	0.1174
Avg. Joint-Tail CA%	0.0003	0.0457	0.0929	0.0425	0.0855
Avg. Tail UC%	0.0003	0.0225	0.0698	0.0255	0.0836
Avg. Joint-Tail AA%	0.0003	0.1259	0.1700	0.1200	0.1596
Avg. Joint-Tail G%	0.0003	0.2497	0.2180	0.2429	0.2206
Avg. Tail Size	0.0003	3.6434	4.2047	3.4932	3.8891
Avg. Joint-Tail UG%	0.0003	0.0376	0.0950	0.0349	0.0788
Joint-Tail%	0.0002	0.6745	0.1629	0.6794	0.1622
Avg. Tail AU%	0.0002	0.0401	0.1061	0.0367	0.0912
UG%	0.0002	0.0474	0.1057	0.0444	0.0942
Avg. Tail GA%	0.0002	0.0470	0.1233	0.0434	0.1160
Avg. Tail AA%	0.0002	0.0743	0.1593	0.0693	0.1458
Avg. Tail AG%	0.0002	0.0452	0.1125	0.0425	0.1078
Tail%	0.0002	0.3994	0.3334	0.4086	0.3468
Avg. Joint-Tail A%	0.0002	0.3738	0.2298	0.3684	0.2292
Avg. Joint-Tail UA%	0.0001	0.0559	0.0921	0.0535	0.0900
AG%	0.0001	0.0735	0.1170	0.0714	0.1172
Avg. Joint AC%	0.0001	0.0386	0.1048	0.0357	0.1046

Avg. Joint GG%	0.0001	0.0229	0.1068	0.0210	0.0921
UA%	0.0001	0.0698	0.1063	0.0675	0.1037
Avg. Joint-Tail UU%	0.0001	0.0465	0.1060	0.0440	0.0990
Avg. Joint CU%	0.0001	0.0181	0.0670	0.0164	0.0675
G%	0.0001	0.1559	0.1160	0.1539	0.1169
CA%	0.0001	0.0558	0.1042	0.0540	0.0986
Avg. Joint UG%	0.0001	0.0184	0.0761	0.0170	0.0661
Avg. Joint-Tail SLC	0.0001	0.5418	0.2017	0.5390	0.2025
Avg. Tail UA%	0.0001	0.0389	0.0987	0.0367	0.0873
Avg. Joint CA%	0.0001	0.0276	0.0838	0.0256	0.0795
CU%	0.0001	0.0428	0.0905	0.0410	0.0935
NLC	0.0001	0.1419	0.3712	0.1513	0.3946
Avg. Joint GC%	0.0001	0.0184	0.0767	0.0199	0.0784
AU% Bond	0.0001	0.2425	0.3396	0.2357	0.3365
Avg. Joint-Tail AG%	0.0001	0.0619	0.1101	0.0606	0.1095
Avg. Joint CG%	0.0001	0.0116	0.0603	0.0128	0.0576
Avg. Joint GU%	0.0001	0.0235	0.0861	0.0217	0.0844
Avg. Joint-Tail NLC	0.0001	0.9084	0.1028	0.9073	0.1049
Joint%	0.0001	0.2751	0.2652	0.2708	0.2750
GA%	5e-05	0.0758	0.1194	0.0774	0.1277
Avg. Tail CU%	5e-05	0.0273	0.0794	0.0263	0.0864
AC%	4e-05	0.0661	0.1163	0.0646	0.1117
Avg. Joint-Tail U%	4e-05	0.2017	0.1922	0.1981	0.1920
Avg. Joint-Tail AU%	3e-05	0.0560	0.1082	0.0544	0.0898
Avg. Joint-Tail CU%	3e-05	0.0341	0.0787	0.0332	0.0839
Avg. Tail AC%	3e-05	0.0349	0.0928	0.0341	0.0929
Avg. Tail GG%	3e-05	0.0317	0.1106	0.0309	0.1061
Avg. Joint-Tail GG%	2e-05	0.0393	0.1014	0.0407	0.1070
A%	2e-05	0.2519	0.1380	0.2510	0.1372
Avg. Joint-Tail CS	2e-05	4.7674	2.0703	4.7478	1.9022
UU%	1e-05	0.0581	0.1252	0.0566	0.1186

U%	1e-05	0.1485	0.1205	0.1470	0.1181
Avg. Joint U%	1e-05	0.1095	0.1777	0.1092	0.1871
Avg. Joint-Tail FS	1e-05	1.1539	0.8229	1.1519	0.6319
Avg. Joint-Tail AC%	9e-06	0.0542	0.0999	0.0536	0.0999
Avg. Joint-Tail Size	8e-06	4.7902	3.4910	4.7640	3.1755
AU%	8e-06	0.0692	0.1248	0.0683	0.1052
Avg. Joint C%	6e-06	0.1050	0.1814	0.1039	0.1794
Avg. Joint AU%	4e-06	0.0354	0.1038	0.0356	0.1018
Avg. Tail GC%	3e-06	0.0257	0.0943	0.0253	0.0912
Avg. Joint GA%	1e-06	0.0414	0.1097	0.0416	0.1206
SLC	1e-06	0.5592	0.3808	0.5597	0.3956
Avg. Joint UC%	0e+00	0.0179	0.0736	0.0180	0.0700
Avg. Joint-Tail GA%	0e+00	0.0657	0.1172	0.0657	0.1197

Table E.19: Experiment 2 Junction Metric Statistics. Ranks each junction metric from the second experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		Nongene	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
Avg. Stack Size	0.0068	8.5136	3.2173	8.0098	2.8230
SLC	0.0039	0.6144	0.1961	0.6387	0.1903
Avg. Stack SLC	0.0029	0.4478	0.1053	0.4590	0.1045
Avg. Internal Loop SLC	0.0026	0.3343	0.3800	0.3742	0.3923
Avg. Loop AA%	0.0022	0.0618	0.1586	0.0482	0.1265
Avg. Stack%	0.0019	0.8520	0.1685	0.8367	0.1774
Avg. Internal Loop NLC	0.0019	0.4263	0.4602	0.4668	0.4655
Avg. Loop SLC	0.0019	0.4353	0.4014	0.4715	0.4068
Avg. Loop FS	0.0017	4.4838	1.4281	4.3760	1.3850
Avg. Internal Loop GC% Bond	0.0017	0.2472	0.3392	0.2754	0.3543
Avg. Internal Loop FS	0.0015	4.4285	1.4784	4.3330	1.4281
Avg. Internal Loop C%	0.0015	0.0767	0.1560	0.0893	0.1663
Avg. Loop CS	0.0014	5.2972	1.7400	5.1101	1.7783
Avg. Internal Loop G%	0.0014	0.1368	0.2331	0.1551	0.2557
Avg. Loop NLC	0.0013	0.5315	0.4631	0.5659	0.4627
Avg. Internal Loop CS	0.0013	5.3657	1.8073	5.1873	1.8255
AC%	0.0012	0.0396	0.0560	0.0437	0.0616
Avg. Loop G%	0.0011	0.1396	0.2286	0.1560	0.2471
Avg. Loop GC% Bond	0.0011	0.3182	0.3566	0.3423	0.3626
Avg. Bulge FS	0.0011	4.5969	1.5219	4.4718	1.4712
Avg. Stack AC%	0.0010	0.0357	0.1020	0.0422	0.1123
Loop%	0.0010	0.1277	0.1356	0.1363	0.1352
PP	0.0010	0.8723	0.1356	0.8637	0.1352
GA%	0.0010	0.0557	0.0635	0.0598	0.0677
Avg. Bulge AC%	0.0010	0.0082	0.0659	0.0046	0.0442
Avg. Internal Loop AA%	0.0010	0.0581	0.1531	0.0490	0.1315
Avg. Bulge CS	0.0009	5.0559	1.8212	4.8566	1.7902
Avg. Bulge AG%	0.0009	0.0090	0.0675	0.0054	0.0509
Avg. Loop C%	0.0009	0.0978	0.1905	0.1099	0.1977

Internal Loop%	0.0008	0.1056	0.1314	0.1127	0.1307
Avg. Bulge SLC	0.0008	0.2524	0.4038	0.2761	0.4192
Avg. Bulge GC% Bond	0.0007	0.1786	0.3292	0.1971	0.3405
AU%	0.0007	0.0374	0.0651	0.0339	0.0596
Avg. Internal Loop U%	0.0007	0.0880	0.1874	0.0981	0.1946
Avg. Stack GC%	0.0007	0.1430	0.2782	0.1294	0.2574
Avg. Bulge AA%	0.0007	0.0192	0.1127	0.0141	0.0906
Avg. Stack CC%	0.0006	0.1018	0.1710	0.0931	0.1611
Avg. Bulge NLC	0.0006	0.2862	0.4407	0.3100	0.4536
Avg. Loop AG%	0.0006	0.0302	0.1001	0.0254	0.0922
Avg. Stack GC%	0.0006	0.0496	0.1182	0.0556	0.1256
Avg. Stack GG%	0.0006	0.1369	0.1879	0.1279	0.1774
GG%	0.0006	0.1261	0.1216	0.1202	0.1155
Avg. Internal Loop CG%	0.0006	0.0091	0.0531	0.0118	0.0615
Avg. Loop AU%	0.0006	0.0161	0.0608	0.0188	0.0648
Avg. Stack FS	0.0005	2.9543	1.1949	2.9248	1.2346
GU%	0.0005	0.0686	0.0688	0.0718	0.0727
Avg. Stack C%	0.0005	0.2975	0.1129	0.2925	0.1140
AA%	0.0004	0.0390	0.0581	0.0365	0.0554
Avg. Internal Loop CC%	0.0004	0.0110	0.0669	0.0140	0.0729
Avg. Internal Loop AU%	0.0004	0.0155	0.0607	0.0178	0.0650
Avg. Stack UU%	0.0004	0.0375	0.1070	0.0416	0.1091
Avg. Stack UA%	0.0004	0.0275	0.1186	0.0326	0.1313
Avg. Bulge AU%	0.0004	0.0063	0.0556	0.0085	0.0656
Avg. Internal Loop AU% Bond	0.0004	0.1451	0.2511	0.1549	0.2547
Avg. Stack G%	0.0004	0.3568	0.1040	0.3529	0.1057
CG%	0.0003	0.0672	0.1004	0.0710	0.1088
CA%	0.0003	0.0427	0.0549	0.0409	0.0555
Avg. Internal Loop CU%	0.0003	0.0122	0.0714	0.0148	0.0726
Avg. Loop CG%	0.0003	0.0089	0.0498	0.0107	0.0543
AG%	0.0003	0.0599	0.0631	0.0578	0.0638

CC%	0.0003	0.0883	0.1129	0.0845	0.1070
Avg. Stack CG%	0.0002	0.0799	0.2136	0.0730	0.2017
Avg. Internal Loop SR	0.0002	0.9092	0.1831	0.9035	0.1856
Avg. Loop US%	0.0002	0.0123	0.0595	0.0145	0.0647
Avg. Bulge UU%	0.0002	0.0091	0.0759	0.0069	0.0631
Avg. Stack GU%	0.0002	0.0754	0.1417	0.0794	0.1467
U%	0.0002	0.1999	0.1101	0.2031	0.1094
Avg. Internal Loop UC%	0.0002	0.0132	0.0680	0.0153	0.0699
Avg. Loop AU% Bond	0.0002	0.1724	0.2558	0.1808	0.2543
Avg. Loop CC%	0.0002	0.0114	0.0716	0.0137	0.0699
Avg. Bulge G%	0.0002	0.0437	0.1675	0.0492	0.1807
Bulge%	0.0002	0.0222	0.0448	0.0236	0.0460
Avg. Stack U%	0.0002	0.1985	0.1111	0.2015	0.1115
Avg. Stack NLC	0.0002	0.8738	0.1226	0.8702	0.1319
UC%	0.0002	0.0517	0.0638	0.0536	0.0665
Avg. Bulge CU%	0.0002	0.0032	0.0376	0.0043	0.0450
Avg. Loop UU%	0.0002	0.0229	0.1031	0.0204	0.0884
NLC	0.0002	0.9153	0.0894	0.9177	0.0854
Avg. Internal Loop CA%	0.0002	0.0150	0.0616	0.0169	0.0757
Avg. Internal Loop AC%	0.0002	0.0143	0.0601	0.0157	0.0658
Avg. Stack CS	0.0002	5.9590	1.7035	5.7663	1.7511
Avg. Loop U%	0.0002	0.1269	0.2352	0.1334	0.2306
Avg. Bulge GC%	0.0001	0.0029	0.0420	0.0021	0.0309
Avg. Stack AA%	0.0001	0.0179	0.0739	0.0196	0.0771
Avg. Stack CU%	0.0001	0.0565	0.1278	0.0596	0.1277
GC%	0.0001	0.1131	0.1595	0.1098	0.1558
Avg. Internal Loop Size	0.0001	2.0349	2.6620	2.0921	2.5816
Avg. Loop CU%	0.0001	0.0124	0.0680	0.0139	0.0655
Avg. Bulge C%	0.0001	0.0583	0.2001	0.0632	0.2111
GU% Bond	0.0001	0.1212	0.1212	0.1241	0.1224
UG%	0.0001	0.0627	0.0638	0.0643	0.0647

Avg. Bulge GU% Bond	0.0001	0.0353	0.1278	0.0330	0.1233
Avg. Stack A%	0.0001	0.1392	0.0997	0.1411	0.0995
Avg. Loop GU% Bond	0.0001	0.0838	0.1695	0.0807	0.1666
Avg. Bulge A%	0.0001	0.1130	0.2773	0.1185	0.2861
Size	0.0001	21.9827	17.0357	22.3084	17.3015
Avg. Bulge Size	0.0001	0.5475	1.1872	0.5715	1.2148
Avg. Loop A%	0.0001	0.2101	0.2846	0.2045	0.2703
GC% Bond	0.0001	0.5983	0.2216	0.5945	0.2214
Avg. Bulge AU%	0.0001	0.0862	0.2134	0.0907	0.2124
Avg. Internal Loop AG%	0.0001	0.0287	0.0982	0.0272	0.0981
Avg. Internal Loop UA%	0.0001	0.0205	0.0707	0.0215	0.0776
Avg. Bulge UC%	0.0001	0.0034	0.0395	0.0045	0.0482
Avg. Stack AU%	0.0001	0.0326	0.1296	0.0303	0.1214
Avg. Stack AG%	0.0001	0.0458	0.1156	0.0479	0.1159
Avg. Loop AC%	0.0001	0.0163	0.0667	0.0151	0.0607
Avg. Loop UA%	0.0001	0.0216	0.0746	0.0227	0.0812
Avg. Bulge U%	0.0001	0.0850	0.2448	0.0899	0.2474
Avg. Loop CA%	0.0001	0.0164	0.0691	0.0176	0.0744
Avg. Stack UC%	0.0001	0.0545	0.1254	0.0568	0.1263
G%	0.0001	0.3477	0.1005	0.3463	0.1005
Avg. Loop GC%	3e-05	0.0105	0.0550	0.0099	0.0578
UU%	3e-05	0.0459	0.0659	0.0464	0.0655
Avg. Internal Loop UG%	3e-05	0.0214	0.1078	0.0200	0.0965
UA%	2e-05	0.0380	0.0649	0.0372	0.0635
Avg. Loop UG%	2e-05	0.0206	0.1003	0.0193	0.0913
C%	2e-05	0.2827	0.1104	0.2819	0.1095
A%	2e-05	0.1697	0.1056	0.1687	0.1030
Avg. Internal Loop GU% Bond	1e-05	0.0745	0.1697	0.0757	0.1727
Avg. Internal Loop A%	9e-06	0.1653	0.2466	0.1636	0.2358
Avg. Internal Loop GG%	9e-06	0.0211	0.0935	0.0219	0.0969
AU% Bond	9e-06	0.2805	0.1982	0.2814	0.1962

Avg. Loop Size	8e-06	2.0213	2.3578	2.0314	2.2864
Avg. Bulge GU%	8e-06	0.0047	0.0498	0.0050	0.0511
Avg. Loop GA%	8e-06	0.0301	0.1024	0.0292	0.1040
Avg. Bulge GA%	7e-06	0.0062	0.0556	0.0064	0.0545
Avg. Internal Loop GC%	7e-06	0.0110	0.0566	0.0106	0.0631
Avg. Bulge GG%	6e-06	0.0034	0.0438	0.0036	0.0470
Avg. Stack CA%	5e-06	0.0360	0.1007	0.0367	0.1024
Avg. Bulge CG%	4e-06	0.0019	0.0279	0.0020	0.0288
Avg. Internal Loop GA%	3e-06	0.0322	0.1124	0.0316	0.1161
Avg. Internal Loop GU%	3e-06	0.0113	0.0626	0.0115	0.0640
Avg. Bulge CC%	3e-06	0.0037	0.0520	0.0038	0.0460
Avg. Loop GU%	3e-06	0.0124	0.0666	0.0123	0.0640
Avg. Stack UG%	2e-06	0.0615	0.1318	0.0623	0.1293
Avg. Bulge CA%	2e-06	0.0063	0.0596	0.0065	0.0576
Avg. Loop GG%	2e-06	0.0195	0.0846	0.0198	0.0884
Avg. Bulge UA%	1e-06	0.0080	0.0638	0.0082	0.0663
Avg. Internal Loop UU%	1e-06	0.0207	0.1007	0.0209	0.0970
Avg. Bulge UG%	0e+00	0.0049	0.0530	0.0048	0.0494
CU%	0e+00	0.0570	0.0646	0.0570	0.0658

Table E.20: Experiment 2 Stem Metric Statistics. Ranks each stem metric from the second experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		Nongene	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
Size	0.0092	14.3775	9.3802	16.1137	9.3887
Avg. Stack CS	0.0091	5.6057	1.6503	5.5696	1.8326
Avg. Loop SLC	0.0089	0.3183	0.4027	0.3921	0.4200
Avg. Loop NLC	0.0081	0.3827	0.4623	0.4616	0.4719
Avg. Stack FS	0.0080	2.8512	1.2250	2.7968	1.2773
Avg. Internal Loop C%	0.0072	0.0427	0.1273	0.0665	0.1625
Avg. Internal Loop SLC	0.0071	0.2333	0.3642	0.2911	0.3872
Avg. Internal Loop NLC	0.0068	0.2880	0.4310	0.3546	0.4527
Avg. Loop U%	0.0058	0.0839	0.2095	0.1176	0.2411
Avg. Internal Loop GG%	0.0056	0.0070	0.0471	0.0200	0.1014
Avg. Bulge SLC	0.0055	0.1404	0.3266	0.1998	0.3818
Avg. Internal Loop GC% Bond	0.0050	0.1767	0.3208	0.2194	0.3471
Avg. Internal Loop UA%	0.0049	0.0082	0.0490	0.0181	0.0815
Avg. Loop FS	0.0048	3.9899	1.2496	4.0013	1.5297
Avg. Bulge FS	0.0046	4.2747	1.2334	4.0478	1.3206
Avg. Bulge NLC	0.0046	0.1597	0.3584	0.2186	0.4061
Avg. Internal Loop U%	0.0044	0.0527	0.1578	0.0747	0.1883
Avg. Loop GG%	0.0041	0.0088	0.0593	0.0206	0.1044
Avg. Bulge CS	0.0041	4.6162	1.4493	4.2674	1.5777
Avg. Loop GC% Bond	0.0039	0.2460	0.3638	0.2863	0.3652
Avg. Internal Loop FS	0.0038	4.0298	1.4078	3.9641	1.5609
Avg. Bulge AU%	0.0037	0.0008	0.0121	0.0054	0.0510
Avg. Loop CS	0.0037	4.6924	1.4797	4.4975	1.8638
PP	0.0035	0.9134	0.1252	0.9002	0.1208
Loop%	0.0035	0.0866	0.1252	0.0998	0.1208
Avg. Loop AG%	0.0034	0.0313	0.1149	0.0182	0.0778
Avg. Internal Loop Size	0.0031	1.2567	2.1994	1.4795	2.2616
Avg. Loop GU% Bond	0.0031	0.0494	0.1461	0.0706	0.1787
Avg. Internal Loop CS	0.0029	4.8386	1.5824	4.5547	1.9021

Avg. Internal Loop AC%	0.0029	0.0050	0.0390	0.0102	0.0552
Avg. Bulge A%	0.0028	0.0556	0.2066	0.0852	0.2575
Avg. Bulge AC%	0.0028	0.0050	0.0458	0.0014	0.0253
UG%	0.0028	0.0561	0.0689	0.0647	0.0672
Avg. Loop Size	0.0028	1.3727	2.1338	1.5632	2.1143
Avg. Bulge GC% Bond	0.0027	0.1086	0.2859	0.1432	0.3061
Avg. Loop A%	0.0027	0.1378	0.2559	0.1651	0.2700
Avg. Bulge U%	0.0027	0.0476	0.1939	0.0721	0.2373
Avg. Internal Loop AG%	0.0026	0.0287	0.1091	0.0189	0.0813
Avg. Loop AU% Bond	0.0026	0.1159	0.2428	0.1380	0.2455
Avg. Internal Loop GU% Bond	0.0025	0.0405	0.1362	0.0584	0.1714
Avg. Bulge AU%	0.0024	0.0436	0.1648	0.0625	0.1813
NLC	0.0023	0.8988	0.0935	0.9073	0.0867
Avg. Loop UU%	0.0023	0.0121	0.0857	0.0227	0.1165
Internal Loop%	0.0022	0.0720	0.1220	0.0814	0.1169
Avg. Stack%	0.0022	0.9007	0.1477	0.8887	0.1394
AC%	0.0021	0.0378	0.0633	0.0430	0.0602
Avg. Internal Loop UU%	0.0021	0.0098	0.0777	0.0201	0.1203
SLC	0.0021	0.5614	0.1836	0.5762	0.1935
Avg. Internal Loop CC%	0.0020	0.0074	0.0650	0.0148	0.0924
Avg. Loop GA%	0.0019	0.0141	0.0746	0.0227	0.0995
Avg. Internal Loop A%	0.0018	0.1043	0.2197	0.1224	0.2244
Avg. Bulge UU%	0.0018	0.0033	0.0452	0.0085	0.0743
CC%	0.0018	0.0976	0.1256	0.0868	0.1084
GU% Bond	0.0018	0.1083	0.1282	0.1210	0.1296
Avg. Loop C%	0.0017	0.0696	0.1899	0.0845	0.1968
Avg. Internal Loop GA%	0.0016	0.0144	0.0750	0.0224	0.0985
GG%	0.0016	0.1354	0.1304	0.1244	0.1186
Avg. Stack Size	0.0016	8.2340	3.0376	8.4837	2.9390
Avg. Internal Loop CA%	0.0015	0.0057	0.0355	0.0092	0.0628
UC%	0.0015	0.0490	0.0693	0.0528	0.0642

Bulge%	0.0014	0.0146	0.0404	0.0184	0.0450
GA%	0.0014	0.0530	0.0689	0.0568	0.0663
Avg. Stack AU%	0.0011	0.0278	0.1249	0.0194	0.0975
Avg. Bulge UC%	0.0011	0.0015	0.0227	0.0005	0.0076
Avg. Loop CC%	0.0011	0.0094	0.0766	0.0154	0.0933
Avg. Internal Loop AU% Bond	0.0011	0.0953	0.2279	0.1071	0.2325
Avg. Loop CU%	0.0011	0.0049	0.0411	0.0073	0.0508
Avg. Bulge CU%	0.0011	0.0004	0.0092	0.0019	0.0265
CG%	0.0011	0.0642	0.1084	0.0703	0.1041
Avg. Internal Loop AA%	0.0011	0.0431	0.1458	0.0334	0.1140
Avg. Bulge Size	0.0010	0.3079	1.0028	0.3831	1.0312
Avg. Internal Loop CU%	0.0010	0.0050	0.0421	0.0074	0.0570
Avg. Stack UC%	0.0010	0.0558	0.1429	0.0647	0.1441
Avg. Stack SLC	0.0009	0.4575	0.1005	0.4510	0.0979
U%	0.0009	0.1879	0.1204	0.1953	0.1077
Avg. Stack UG%	0.0009	0.0458	0.1255	0.0564	0.1350
Avg. Bulge UA%	0.0008	0.0069	0.0736	0.0040	0.0443
Avg. Loop AA%	0.0008	0.0449	0.1488	0.0355	0.1202
CA%	0.0008	0.0348	0.0591	0.0390	0.0583
Avg. Bulge GC%	0.0008	0.0022	0.0289	0.0006	0.0128
GC%	0.0008	0.1486	0.2060	0.1404	0.1869
Avg. Bulge G%	0.0008	0.0262	0.1302	0.0326	0.1555
Avg. Bulge GU% Bond	0.0007	0.0152	0.0917	0.0218	0.1044
Avg. Internal Loop G%	0.0005	0.1128	0.2418	0.1214	0.2486
AG%	0.0005	0.0579	0.0675	0.0553	0.0620
Avg. Internal Loop SR	0.0005	0.9441	0.1521	0.9351	0.1695
Avg. Loop AU%	0.0005	0.0117	0.0734	0.0153	0.0699
Avg. Bulge CA%	0.0005	0.0064	0.0567	0.0041	0.0549
Avg. Loop G%	0.0005	0.1200	0.2418	0.1278	0.2449
Avg. Stack CA%	0.0005	0.0271	0.0975	0.0331	0.1067
AA%	0.0004	0.0340	0.0638	0.0309	0.0550

Avg. Internal Loop UC%	0.0004	0.0068	0.0594	0.0084	0.0575
Avg. Loop UA%	0.0004	0.0141	0.0865	0.0178	0.0750
Avg. Stack AC%	0.0003	0.0420	0.1235	0.0471	0.1304
Avg. Stack UA%	0.0003	0.0261	0.1305	0.0306	0.1339
GC% Bond	0.0003	0.6245	0.2425	0.6155	0.2155
UU%	0.0003	0.0421	0.0705	0.0442	0.0678
G%	0.0003	0.3617	0.1052	0.3590	0.0978
Avg. Stack AG%	0.0002	0.0496	0.1360	0.0475	0.1222
Avg. Stack NLC	0.0002	0.8775	0.0980	0.8803	0.0950
Avg. Stack CC%	0.0002	0.1025	0.1822	0.0954	0.1693
Avg. Stack GC%	0.0002	0.1714	0.3273	0.1621	0.3090
Avg. Internal Loop GC%	0.0002	0.0068	0.0575	0.0077	0.0708
C%	0.0002	0.2967	0.1212	0.2931	0.1071
Avg. Stack AA%	0.0002	0.0144	0.0708	0.0168	0.0792
Avg. Bulge UG%	0.0002	0.0042	0.0566	0.0031	0.0398
GU%	0.0002	0.0709	0.0772	0.0736	0.0727
Avg. Internal Loop UG%	0.0002	0.0169	0.0933	0.0208	0.1229
Avg. Loop AC%	0.0001	0.0087	0.0526	0.0098	0.0521
Avg. Stack A%	0.0001	0.1339	0.1072	0.1316	0.0959
Avg. Stack G%	0.0001	0.3661	0.1072	0.3684	0.0959
Avg. Loop UG%	0.0001	0.0181	0.0946	0.0216	0.1218
Avg. Bulge CC%	0.0001	0.0034	0.0537	0.0027	0.0502
Avg. Stack CU%	0.0001	0.0578	0.1465	0.0559	0.1326
UA%	0.0001	0.0348	0.0737	0.0335	0.0643
Avg. Stack CG%	0.0001	0.0761	0.2178	0.0709	0.2092
Avg. Loop CA%	0.0001	0.0102	0.0603	0.0106	0.0656
Avg. Loop US%	0.0001	0.0072	0.0585	0.0073	0.0495
Avg. Stack GG%	0.0001	0.1278	0.1967	0.1296	0.1870
Avg. Bulge GA%	0.0001	0.0034	0.0533	0.0045	0.0555
AU% Bond	0.0001	0.2673	0.2109	0.2636	0.1906
Avg. Bulge AA%	0.0001	0.0073	0.0762	0.0085	0.0844

Avg. Internal Loop GU%	0.0001	0.0087	0.0606	0.0071	0.0566
Avg. Internal Loop CG%	0.0001	0.0081	0.0658	0.0079	0.0501
Avg. Internal Loop AU%	0.0001	0.0116	0.0736	0.0132	0.0671
Avg. Bulge C%	4e-05	0.0382	0.1754	0.0376	0.1756
Avg. Bulge AG%	3e-05	0.0047	0.0501	0.0035	0.0451
Avg. Stack GU%	3e-05	0.0822	0.1661	0.0797	0.1623
Avg. Loop GC%	3e-05	0.0075	0.0582	0.0068	0.0610
Avg. Bulge CG%	3e-05	0.0013	0.0224	0.0018	0.0375
Avg. Loop GU%	2e-05	0.0115	0.0692	0.0107	0.0751
Avg. Stack U%	2e-05	0.1881	0.1236	0.1899	0.1078
Avg. Stack C%	2e-05	0.3119	0.1236	0.3101	0.1078
Avg. Bulge GG%	2e-05	0.0040	0.0562	0.0028	0.0438
CU%	2e-05	0.0574	0.0712	0.0576	0.0646
Avg. Bulge GU%	1e-05	0.0037	0.0397	0.0046	0.0545
Avg. Stack GC%	1e-05	0.0590	0.1456	0.0565	0.1374
AU%	8e-06	0.0265	0.0625	0.0266	0.0551
A%	8e-06	0.1537	0.1104	0.1526	0.0988
Avg. Loop CG%	5e-06	0.0074	0.0555	0.0078	0.0484
Avg. Stack UU%	1e-06	0.0346	0.1107	0.0343	0.1083

Table E.21: Experiment 2 Bridge Metric Statistics. Ranks each bridge metric from the second experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
Avg. Stack Size	0.0114	8.5623	3.2362	7.9108	2.7886
Stack%	0.0082	0.6177	0.1193	0.5956	0.1258
PP	0.0079	0.6309	0.1003	0.6130	0.1022
Avg. Stack SLC	0.0048	0.4459	0.1060	0.4606	0.1059
GC% Bond	0.0045	0.5931	0.2174	0.5897	0.2222
Avg. Hairpin Loop SLC	0.0037	0.3801	0.0852	0.3697	0.0894
Avg. Hairpin Loop Size	0.0037	5.4184	2.2269	5.7052	2.6124
AU% Bond	0.0032	0.2831	0.1958	0.2852	0.1971
Avg. Hairpin Loop GG%	0.0026	0.0420	0.1186	0.0561	0.1399
Avg. Loop AA%	0.0024	0.0643	0.1599	0.0505	0.1274
Avg. Internal Loop SLC	0.0024	0.3527	0.3807	0.3907	0.3907
Avg. Hairpin Loop AG%	0.0021	0.0605	0.1192	0.0716	0.1263
Avg. Hairpin Loop AA%	0.0020	0.1586	0.2102	0.1401	0.2096
Avg. Loop FS	0.0020	4.5198	1.4320	4.4209	1.3659
Avg. Hairpin Loop GU%	0.0018	0.0609	0.1162	0.0509	0.1077
Avg. Internal Loop FS	0.0018	4.4589	1.4798	4.3759	1.4118
Avg. Internal Loop G%	0.0017	0.1410	0.2319	0.1617	0.2558
Avg. Loop CS	0.0017	5.3399	1.7486	5.1815	1.7553
Avg. Internal Loop NLC	0.0017	0.4507	0.4611	0.4895	0.4647
Avg. Internal Loop CS	0.0016	5.4066	1.8182	5.2587	1.8041
SLC	0.0015	0.8865	0.0470	0.8904	0.0507
Avg. Internal Loop GC% Bond	0.0015	0.2584	0.3404	0.2863	0.3540
Hairpin%	0.0015	0.2550	0.1456	0.2660	0.1560
Avg. Loop G%	0.0014	0.1428	0.2261	0.1614	0.2467
Avg. Loop SLC	0.0014	0.4575	0.3981	0.4877	0.4019
Avg. Hairpin Loop AU%	0.0014	0.0555	0.1081	0.0645	0.1166
Avg. Hairpin Loop UC%	0.0013	0.0529	0.1137	0.0450	0.1018
Avg. Stack AC%	0.0012	0.0338	0.0966	0.0413	0.1082
Avg. Bulge FS	0.0012	4.6087	1.5327	4.5194	1.4795

Avg. Internal Loop C%	0.0011	0.0833	0.1612	0.0940	0.1672
Avg. Hairpin Loop AG%	0.0011	0.0524	0.1100	0.0596	0.1148
Avg. Bulge AG%	0.0011	0.0093	0.0681	0.0057	0.0503
Avg. Bulge CS	0.0010	5.0720	1.8311	4.9229	1.7995
Avg. Stack GC%	0.0010	0.1382	0.2687	0.1223	0.2453
Avg. Loop GC% Bond	0.0009	0.3297	0.3528	0.3532	0.3602
Avg. Stack GG%	0.0009	0.1386	0.1864	0.1273	0.1751
Avg. Internal Loop CG%	0.0009	0.0094	0.0509	0.0128	0.0641
Avg. Loop NLC	0.0009	0.5588	0.4583	0.5872	0.4579
Avg. Loop C%	0.0009	0.1040	0.1912	0.1151	0.1981
Avg. Bulge AA%	0.0009	0.0215	0.1183	0.0152	0.0923
Avg. Stack GC%	0.0009	0.0480	0.1129	0.0551	0.1228
Avg. Internal Loop AA%	0.0008	0.0602	0.1543	0.0521	0.1345
GA%	0.0008	0.0660	0.0518	0.0689	0.0538
Avg. Stack C%	0.0008	0.2946	0.1107	0.2885	0.1148
Avg. Bulge AC%	0.0008	0.0084	0.0669	0.0053	0.0474
Avg. Stack CC%	0.0008	0.1014	0.1690	0.0921	0.1590
Loop%	0.0007	0.1141	0.1187	0.1211	0.1191
Avg. Stack G%	0.0007	0.3550	0.1034	0.3496	0.1074
Avg. Bulge GC% Bond	0.0007	0.1898	0.3335	0.2095	0.3466
Avg. Bulge UU%	0.0007	0.0100	0.0794	0.0064	0.0596
Avg. Stack UU%	0.0006	0.0385	0.1071	0.0432	0.1092
Avg. Loop AU%	0.0006	0.0163	0.0568	0.0196	0.0642
GU% Bond	0.0006	0.1237	0.1202	0.1251	0.1211
Avg. Internal Loop AU%	0.0006	0.0158	0.0574	0.0189	0.0651
GC%	0.0006	0.0815	0.0695	0.0783	0.0682
Internal Loop%	0.0006	0.0945	0.1141	0.1003	0.1137
Avg. Bulge SLC	0.0006	0.2721	0.4128	0.2939	0.4257
Avg. Loop UU%	0.0006	0.0249	0.1070	0.0202	0.0828
Avg. Bulge NLC	0.0005	0.3083	0.4499	0.3312	0.4614
Avg. Internal Loop U%	0.0005	0.0940	0.1913	0.1028	0.1952

Avg. Loop CG%	0.0005	0.0094	0.0494	0.0114	0.0559
Avg. Stack UA%	0.0004	0.0278	0.1162	0.0329	0.1309
CS	0.0004	29.3482	8.4823	29.3207	8.4559
Avg. Stack GU%	0.0004	0.0737	0.1363	0.0792	0.1431
Avg. Internal Loop AU% Bond	0.0004	0.1544	0.2557	0.1644	0.2568
Avg. Loop GU% Bond	0.0004	0.0908	0.1737	0.0836	0.1649
Avg. Hairpin Loop CG%	0.0004	0.0387	0.0974	0.0426	0.1015
Avg. Stack NLC	0.0004	0.8735	0.1266	0.8681	0.1384
AC%	0.0003	0.0497	0.0472	0.0515	0.0475
Avg. Loop US%	0.0003	0.0136	0.0621	0.0158	0.0669
Avg. Stack U%	0.0003	0.2007	0.1087	0.2042	0.1121
Avg. Hairpin Loop GC%	0.0003	0.0496	0.1114	0.0456	0.1042
Avg. Internal Loop SR	0.0003	0.9028	0.1874	0.8964	0.1886
Avg. Internal Loop CU%	0.0003	0.0138	0.0776	0.0163	0.0754
Avg. Loop AG%	0.0003	0.0295	0.0977	0.0267	0.0931
NLC	0.0003	0.9627	0.0284	0.9616	0.0324
AA%	0.0003	0.0718	0.0718	0.0693	0.0758
Avg. Stack CU%	0.0003	0.0566	0.1241	0.0606	0.1268
Avg. Loop A%	0.0003	0.2220	0.2876	0.2130	0.2699
Avg. Stack CG%	0.0003	0.0793	0.2115	0.0732	0.1993
Avg. Bulge C%	0.0003	0.0622	0.2041	0.0690	0.2182
Avg. Bulge AU%	0.0003	0.0070	0.0580	0.0092	0.0685
Size	0.0003	28.6532	17.7924	29.3289	18.6346
Avg. Stack A%	0.0002	0.1403	0.0983	0.1431	0.1001
Avg. Internal Loop UC%	0.0002	0.0144	0.0685	0.0166	0.0719
Avg. Stack FS	0.0002	2.9543	1.1815	2.9446	1.2240
Avg. Internal Loop CC%	0.0002	0.0116	0.0675	0.0139	0.0687
Avg. Hairpin Loop U%	0.0002	0.2296	0.2081	0.2237	0.2076
Avg. Bulge G%	0.0002	0.0468	0.1729	0.0524	0.1846
Avg. Hairpin Loop C%	0.0002	0.1862	0.1861	0.1903	0.1862
Avg. Stack AG%	0.0002	0.0458	0.1120	0.0484	0.1150

Avg. Bulge GU% Bond	0.0002	0.0391	0.1331	0.0357	0.1274
Bulge%	0.0002	0.0196	0.0379	0.0208	0.0392
Avg. Bulge CU%	0.0002	0.0037	0.0410	0.0048	0.0483
Avg. Bulge UC%	0.0002	0.0042	0.0471	0.0054	0.0530
Avg. Stack AA%	0.0002	0.0187	0.0746	0.0202	0.0767
Avg. Loop AU% Bond	0.0001	0.1837	0.2583	0.1894	0.2540
G%	0.0001	0.3104	0.0926	0.3088	0.0938
CG%	0.0001	0.0679	0.0617	0.0664	0.0632
Avg. Hairpin Loop NLC	0.0001	0.8754	0.1031	0.8773	0.1091
Avg. Hairpin Loop CU%	0.0001	0.0443	0.1030	0.0461	0.1032
UU%	0.0001	0.0513	0.0624	0.0525	0.0628
Avg. Internal Loop CA%	0.0001	0.0170	0.0673	0.0184	0.0779
Avg. Loop CU%	0.0001	0.0139	0.0728	0.0152	0.0680
Avg. Loop CC%	0.0001	0.0117	0.0710	0.0134	0.0645
Avg. Hairpin Loop UA%	0.0001	0.0736	0.1218	0.0717	0.1220
Avg. Hairpin Loop UU%	0.0001	0.0645	0.1516	0.0616	0.1449
Avg. Bulge GC%	0.0001	0.0031	0.0429	0.0024	0.0332
Avg. Loop AC%	0.0001	0.0173	0.0686	0.0162	0.0622
GC% Bond	0.0001	0.5931	0.2174	0.5897	0.2222
Avg. Loop GG%	0.0001	0.0217	0.0896	0.0199	0.0856
CC%	0.0001	0.0663	0.0636	0.0674	0.0683
Avg. Hairpin Loop CC%	0.0001	0.0427	0.1169	0.0444	0.1218
Avg. Hairpin Loop GA%	0.0001	0.0924	0.1448	0.0900	0.1426
Avg. Loop CA%	0.0001	0.0178	0.0728	0.0190	0.0757
Avg. Internal Loop Size	0.0001	2.1664	2.7048	2.2171	2.6181
Avg. Bulge Size	0.0001	0.5883	1.2033	0.6143	1.2500
GU%	0.0001	0.0658	0.0486	0.0649	0.0492
AU%	0.0001	0.0447	0.0518	0.0455	0.0508
Avg. Internal Loop UG%	0.0001	0.0219	0.1096	0.0201	0.0907
Avg. Loop UG%	0.0001	0.0208	0.1007	0.0191	0.0841
Avg. Internal Loop AC%	0.0001	0.0156	0.0626	0.0169	0.0677

Avg. Loop GA%	0.0001	0.0327	0.1059	0.0309	0.1055
U%	0.0001	0.2131	0.0997	0.2145	0.1009
UG%	0.0001	0.0636	0.0498	0.0645	0.0515
Avg. Internal Loop A%	0.0001	0.1756	0.2500	0.1722	0.2371
Avg. Loop UA%	0.0001	0.0226	0.0721	0.0239	0.0830
Avg. Loop GC%	0.0001	0.0111	0.0553	0.0103	0.0561
Avg. Bulge UA%	0.0001	0.0081	0.0620	0.0092	0.0704
Avg. Internal Loop UU%	5e-05	0.0227	0.1049	0.0213	0.0929
AU% Bond	5e-05	0.2831	0.1958	0.2852	0.1971
Avg. Internal Loop GG%	4e-05	0.0240	0.1006	0.0225	0.0969
GU% Bond	4e-05	0.1237	0.1202	0.1251	0.1211
GG%	4e-05	0.0993	0.0775	0.0990	0.0799
Avg. Bulge CA%	4e-05	0.0062	0.0590	0.0071	0.0586
Avg. Bulge A%	4e-05	0.1221	0.2865	0.1267	0.2926
Avg. Bulge AU%	3e-05	0.0944	0.2209	0.0972	0.2187
Avg. Internal Loop GA%	3e-05	0.0354	0.1175	0.0338	0.1198
Avg. Internal Loop GC%	3e-05	0.0116	0.0565	0.0111	0.0609
Avg. Bulge CC%	2e-05	0.0037	0.0519	0.0041	0.0458
Avg. Internal Loop GU%	2e-05	0.0118	0.0634	0.0123	0.0643
UA%	2e-05	0.0484	0.0534	0.0488	0.0533
CA%	2e-05	0.0499	0.0461	0.0495	0.0459
Avg. Bulge GG%	2e-05	0.0033	0.0414	0.0036	0.0457
Avg. Hairpin Loop G%	1e-05	0.2334	0.1893	0.2349	0.1973
Avg. Stack AU%	9e-06	0.0338	0.1315	0.0325	0.1253
Avg. Bulge GU%	9e-06	0.0050	0.0524	0.0052	0.0511
FS	8e-06	5.7058	4.4837	5.7958	4.6108
A%	7e-06	0.2292	0.0972	0.2296	0.1001
Avg. Bulge U%	7e-06	0.0922	0.2530	0.0942	0.2499
Avg. Hairpin Loop UG%	6e-06	0.0502	0.1092	0.0493	0.1072
Avg. Loop U%	6e-06	0.1354	0.2402	0.1366	0.2279
Avg. Stack CS	6e-06	5.9881	1.6992	5.7966	1.7371

Avg. Bulge UG%	6e-06	0.0051	0.0530	0.0052	0.0516
Avg. Bulge GA%	5e-06	0.0067	0.0559	0.0069	0.0548
Avg. Stack UG%	5e-06	0.0640	0.1330	0.0642	0.1289
Avg. Loop Size	4e-06	2.1307	2.3642	2.1239	2.2969
Avg. Internal Loop GU% Bond	4e-06	0.0812	0.1755	0.0800	0.1738
Avg. Stack UC%	3e-06	0.0546	0.1227	0.0551	0.1223
Avg. Bulge CG%	3e-06	0.0021	0.0296	0.0021	0.0272
CU%	1e-06	0.0542	0.0470	0.0544	0.0468
AG%	1e-06	0.0693	0.0500	0.0692	0.0528
Avg. Hairpin Loop CA%	1e-06	0.0612	0.1218	0.0609	0.1157
Avg. Loop GU%	0e+00	0.0128	0.0672	0.0126	0.0606
Avg. Hairpin Loop A%	0e+00	0.3509	0.2282	0.3510	0.2240
Avg. Stack CA%	0e+00	0.0377	0.1018	0.0379	0.1021
Avg. Internal Loop AG%	0e+00	0.0286	0.0975	0.0290	0.1008
C%	0e+00	0.2473	0.0882	0.2472	0.0922
Avg. Internal Loop UA%	0e+00	0.0223	0.0735	0.0224	0.0775
UC%	0e+00	0.0502	0.0461	0.0500	0.0456

Table E.22: Experiment 2 Stemloop Metric Statistics. Ranks each stemloop metric from the second experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		Nongene	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
MFE	0.0190	-22.8759	7.7273	-20.8579	6.8840
Bridge%	0.0061	0.0671	0.1150	0.0861	0.1274
NLC	0.0049	0.9871	0.0046	0.9877	0.0039
Stemloop%	0.0042	0.7495	0.1677	0.7269	0.1818
Multiloop%	0.0029	0.0753	0.1102	0.0874	0.1142
SLC	0.0027	0.9303	0.0213	0.9326	0.0223
Internal Loop%	0.0014	0.1237	0.0895	0.1305	0.0882
Bulge%	0.0013	0.0245	0.0321	0.0269	0.0355
AG%	0.0010	0.0733	0.0259	0.0716	0.0271
AA%	0.0009	0.0809	0.0440	0.0783	0.0440
CC%	0.0007	0.0593	0.0364	0.0612	0.0375
A%	0.0007	0.2570	0.0650	0.2536	0.0660
C%	0.0005	0.2309	0.0626	0.2336	0.0624
UC%	0.0004	0.0479	0.0242	0.0489	0.0240
UA%	0.0003	0.0552	0.0349	0.0540	0.0330
UG%	0.0003	0.0634	0.0285	0.0643	0.0276
Stem%	0.0002	0.6755	0.1239	0.6716	0.1222
GC%	0.0002	0.0721	0.0328	0.0711	0.0319
AC%	0.0002	0.0521	0.0242	0.0528	0.0245
CU%	0.0002	0.0537	0.0260	0.0544	0.0253
GU%	0.0002	0.0628	0.0269	0.0621	0.0252
Stack%	0.0002	0.5831	0.1247	0.5864	0.1315
UU%	0.0002	0.0543	0.0439	0.0554	0.0425
% Num. Bridge	0.0001	0.0001	0.0005	0.0001	0.0005
% Num. Multiloop	0.0001	0.0001	0.0005	0.0001	0.0005
GA%	0.0001	0.0672	0.0266	0.0677	0.0265
U%	0.0001	0.2213	0.0729	0.2225	0.0700
% Num. Hairpin Loop	2e-05	0.0003	0.0016	0.0003	0.0016
% Num. Stemloop	2e-05	0.0003	0.0016	0.0003	0.0016

G%	1e-05	0.2908	0.0722	0.2903	0.0706
GG%	1e-05	0.0889	0.0483	0.0892	0.0477
% Num. Internal Loop	1e-05	0.0004	0.0022	0.0004	0.0017
AU%	9e-06	0.0509	0.0335	0.0511	0.0322
CA%	7e-06	0.0538	0.0250	0.0539	0.0240
CG%	7e-06	0.0643	0.0331	0.0642	0.0333
Hairpin Loop%	7e-06	0.1411	0.0585	0.1414	0.0592
% Num. Joint	5e-06	0.1692	0.0046	0.1692	0.0039
% Num. Stack	0e+00	0.0009	0.0046	0.0010	0.0042
Tail%	0e+00	0.0275	0.2851	0.0308	0.2866
% Num. Stem	0e+00	0.0004	0.0020	0.0004	0.0020
Joint%	0e+00	0.0950	0.1166	0.0787	0.1137
% Num. Bulge	0e+00	0.0002	0.0013	0.0002	0.0010
% Num. Tail	0e+00	0.1624	0.0050	0.1624	0.0047
Size	0e+00	80.0000	0.0000	80.0000	0.0000

Table E.23: Experiment 2 Structure Metric Statistics. Ranks each structure metric from the second experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
GC% Bond	0.0045	0.6899	0.4626	0.6259	0.4840
SLC	0.0037	0.3801	0.0853	0.3697	0.0894
Size	0.0037	5.4158	2.2098	5.7052	2.6124
AU% Bond	0.0032	0.1978	0.3984	0.2444	0.4298
GG%	0.0026	0.0425	0.1204	0.0561	0.1399
AG%	0.0021	0.0611	0.1197	0.0716	0.1263
AA%	0.0020	0.1594	0.2112	0.1401	0.2096
GU%	0.0018	0.0605	0.1162	0.0509	0.1077
AU%	0.0014	0.0559	0.1085	0.0645	0.1166
UC%	0.0013	0.0528	0.1136	0.0450	0.1018
AC%	0.0011	0.0513	0.1091	0.0596	0.1148
FS	0.0008	25.0122	8.5746	24.7540	8.3869
GU% Bond	0.0006	0.1123	0.3158	0.1297	0.3361
CS	0.0005	29.4141	8.5949	29.3450	8.5556
CG%	0.0004	0.0386	0.0973	0.0426	0.1015
GC%	0.0003	0.0498	0.1116	0.0456	0.1042
U%	0.0002	0.2305	0.2088	0.2237	0.2076
C%	0.0002	0.1843	0.1854	0.1903	0.1862
NLC	0.0001	0.8749	0.1037	0.8773	0.1091
CU%	0.0001	0.0439	0.1024	0.0461	0.1032
UA%	0.0001	0.0747	0.1226	0.0717	0.1220
UU%	0.0001	0.0649	0.1522	0.0616	0.1449
CC%	0.0001	0.0418	0.1161	0.0444	0.1218
GA%	0.0001	0.0928	0.1449	0.0900	0.1426
G%	1e-05	0.2335	0.1898	0.2349	0.1973
UG%	6e-06	0.0498	0.1091	0.0493	0.1072
CA%	1e-06	0.0602	0.1209	0.0609	0.1157
A%	0e+00	0.3517	0.2286	0.3510	0.2240

Table E.24: Experiment 2 Hairpin Loop Metric Statistics. Ranks each hairpin loop metric from the second experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
AA%	0.0013	0.1204	0.2506	0.1037	0.2344
SLC	0.0009	0.6227	0.2776	0.6384	0.2860
NLC	0.0004	0.9128	0.1222	0.9174	0.1221
A%	0.0004	0.3657	0.3501	0.3528	0.3539
GA%	0.0004	0.0653	0.1799	0.0584	0.1683
CC%	0.0003	0.0249	0.1101	0.0291	0.1251
C%	0.0003	0.1733	0.2737	0.1830	0.2862
GU%	0.0003	0.0344	0.1201	0.0306	0.1170
UA%	0.0002	0.0508	0.1428	0.0468	0.1384
CG%	0.0002	0.0215	0.0990	0.0245	0.1074
CS	0.0002	4.0640	1.9349	4.0181	1.9878
GC%	0.0002	0.0298	0.1210	0.0271	0.1153
AG%	0.0001	0.0552	0.1610	0.0520	0.1571
UC%	0.0001	0.0324	0.1272	0.0299	0.1183
UU%	0.0001	0.0428	0.1534	0.0404	0.1470
Size	0.0001	3.5134	3.1360	3.4678	3.2259
GG%	0.0001	0.0353	0.1432	0.0375	0.1517
UG%	0.0001	0.0384	0.1447	0.0361	0.1397
CU%	4e-05	0.0277	0.1091	0.0292	0.1149
CA%	3e-05	0.0389	0.1270	0.0374	0.1259
AU%	3e-05	0.0421	0.1313	0.0433	0.1296
AC%	3e-05	0.0414	0.1359	0.0403	0.1307
G%	1e-05	0.2565	0.3215	0.2584	0.3339
U%	9e-06	0.2044	0.2879	0.2057	0.2951
FS	8e-06	1.0502	0.2917	1.0489	0.2683

Table E.25: Experiment 2 Unpaired Metric Statistics. Ranks each unpaired metric from the second experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		Nongene	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
Size	0.0026	8.1492	3.6413	7.7984	3.3268
SLC	0.0018	0.4662	0.1238	0.4764	0.1209
FS	0.0003	2.9625	1.3441	2.9370	1.3914
UU%	0.0003	0.0400	0.1357	0.0449	0.1429
GU% Bond	0.0002	0.1240	0.1642	0.1287	0.1691
CU%	0.0002	0.0582	0.1603	0.0622	0.1650
CC%	0.0002	0.0923	0.1940	0.0870	0.1896
U%	0.0002	0.2083	0.1329	0.2115	0.1358
C%	0.0002	0.2917	0.1329	0.2885	0.1358
GC% Bond	0.0002	0.5835	0.2658	0.5770	0.2715
GC%	0.0001	0.1344	0.3183	0.1278	0.3122
AG%	0.0001	0.0483	0.1477	0.0513	0.1517
UA%	0.0001	0.0295	0.1552	0.0331	0.1628
AC%	0.0001	0.0389	0.1340	0.0417	0.1382
NLC	0.0001	0.8858	0.1122	0.8838	0.1134
CG%	0.0001	0.0754	0.2494	0.0717	0.2420
AA%	4e-05	0.0195	0.0967	0.0209	0.1001
GG%	3e-05	0.1257	0.2169	0.1228	0.2152
CS	3e-05	5.8758	1.8524	5.7199	1.9134
CA%	2e-05	0.0402	0.1360	0.0391	0.1342
GA%	1e-05	0.0560	0.1577	0.0574	0.1594
AU% Bond	8e-06	0.2926	0.2444	0.2942	0.2478
G%	8e-06	0.3537	0.1222	0.3529	0.1239
A%	8e-06	0.1463	0.1222	0.1471	0.1239
AU%	7e-06	0.0338	0.1626	0.0331	0.1620
UC%	5e-06	0.0590	0.1613	0.0583	0.1605
UG%	2e-06	0.0671	0.1744	0.0675	0.1731
GU%	0e+00	0.0816	0.1849	0.0813	0.1847

Table E.26: Experiment 2 Stack Metric Statistics. Ranks each stack metric from the second experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
AC%	0.0028	0.0280	0.1376	0.0163	0.0907
AG%	0.0027	0.0279	0.1258	0.0162	0.0940
NLC	0.0021	0.9571	0.1036	0.9655	0.0912
AA%	0.0020	0.0605	0.2055	0.0436	0.1691
GC% Bond	0.0017	0.5937	0.4913	0.5877	0.4924
GU% Bond	0.0008	0.1084	0.3109	0.1116	0.3149
SLC	0.0007	0.8450	0.2347	0.8580	0.2365
UU%	0.0007	0.0270	0.1340	0.0211	0.1181
AU% Bond	0.0006	0.2980	0.4575	0.3007	0.4587
GC%	0.0005	0.0105	0.0866	0.0069	0.0624
G%	0.0005	0.1432	0.3020	0.1564	0.3234
AU%	0.0004	0.0213	0.1137	0.0257	0.1207
A%	0.0004	0.3794	0.4325	0.3624	0.4345
UA%	0.0003	0.0285	0.1312	0.0247	0.1202
CG%	0.0001	0.0059	0.0519	0.0072	0.0604
CU%	0.0001	0.0123	0.0885	0.0143	0.0914
CC%	0.0001	0.0112	0.0940	0.0134	0.0947
UG%	0.0001	0.0143	0.0950	0.0159	0.0968
GU%	0.0001	0.0148	0.0947	0.0156	0.0998
GU% Bond	2e-05	0.1084	0.3109	0.1116	0.3149
GC% Bond	9e-06	0.5937	0.4913	0.5877	0.4924
U%	9e-06	0.2806	0.4083	0.2824	0.4074
GA%	5e-06	0.0202	0.1098	0.0204	0.1036
GG%	5e-06	0.0111	0.0838	0.0122	0.0926
Size	4e-06	1.7961	1.6400	1.7948	1.7442
C%	1e-06	0.1968	0.3594	0.1989	0.3641
AU% Bond	0e+00	0.2980	0.4575	0.3007	0.4587
UC%	0e+00	0.0144	0.1005	0.0142	0.0926
CA%	0e+00	0.0214	0.1176	0.0207	0.1151

Table E.27: Experiment 2 bulge Metric Statistics. Ranks each bulge metric from the second experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
AA%	0.0031	0.1218	0.2300	0.0968	0.2073
SLC	0.0030	0.7132	0.2219	0.7392	0.2257
A%	0.0021	0.3514	0.2889	0.3243	0.2868
Size	0.0018	4.3922	2.6898	4.1626	2.7011
NLC	0.0009	0.9150	0.1077	0.9219	0.1083
AG%	0.0008	0.0646	0.1730	0.0548	0.1620
GA%	0.0005	0.0689	0.1877	0.0603	0.1806
CG%	0.0005	0.0193	0.0852	0.0233	0.1005
UA%	0.0005	0.0460	0.1203	0.0415	0.1174
C%	0.0004	0.1661	0.2219	0.1756	0.2347
GC%	0.0004	0.0262	0.1041	0.0219	0.1045
CC%	0.0004	0.0234	0.1066	0.0271	0.1196
CU%	0.0004	0.0239	0.1094	0.0282	0.1148
AU% Bond	0.0003	0.3194	0.4663	0.3041	0.4601
G%	0.0003	0.2984	0.3136	0.3072	0.3330
GC% Bond	0.0002	0.5276	0.4993	0.5411	0.4984
AC%	0.0002	0.0340	0.1112	0.0296	0.1029
U%	0.0001	0.1842	0.2650	0.1929	0.2742
UC%	0.0001	0.0328	0.1421	0.0290	0.1126
GC% Bond	0.0001	0.5276	0.4993	0.5411	0.4984
UG%	0.0001	0.0437	0.1734	0.0410	0.1655
GU%	0.0001	0.0236	0.0991	0.0221	0.1001
GU% Bond	0.0001	0.1530	0.3601	0.1548	0.3618
AU% Bond	3e-05	0.3194	0.4663	0.3041	0.4601
CA%	8e-06	0.0332	0.1081	0.0315	0.1171
GG%	7e-06	0.0434	0.1483	0.0433	0.1575
AU%	2e-06	0.0349	0.1021	0.0356	0.1057
UU%	1e-06	0.0409	0.1574	0.0416	0.1543
SR	1e-06	0.8074	0.2637	0.8090	0.2673
GU% Bond	0e+00	0.1530	0.3601	0.1548	0.3618

Table E.28: Experiment 2 Internal Loop Metric Statistics. Ranks each internal loop metric from the second experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
C%	0.0013	0.1776	0.2458	0.1949	0.2611
CC%	0.0012	0.0268	0.1070	0.0351	0.1293
CG%	0.0009	0.0240	0.0986	0.0300	0.1056
UU%	0.0007	0.0565	0.1585	0.0488	0.1403
UC%	0.0004	0.0320	0.1054	0.0364	0.1162
GA%	0.0003	0.0692	0.1779	0.0629	0.1592
CA%	0.0003	0.0467	0.1351	0.0422	0.1203
U%	0.0003	0.2231	0.2657	0.2138	0.2558
G%	0.0002	0.2649	0.3013	0.2579	0.2977
UG%	0.0002	0.0421	0.1302	0.0387	0.1148
GU%	0.0001	0.0488	0.1431	0.0461	0.1360
Size	0.0001	5.0588	4.8262	5.1731	4.7893
AG%	0.0001	0.0647	0.1636	0.0617	0.1506
AU%	5e-05	0.0567	0.1446	0.0548	0.1291
CU%	3e-05	0.0396	0.1163	0.0374	0.1175
NLC	3e-05	0.9162	0.1136	0.9173	0.1126
SLC	2e-05	0.5304	0.2850	0.5272	0.2918
UA%	2e-05	0.0543	0.1312	0.0558	0.1300
CS	2e-05	56.6958	4.0007	56.6463	4.0977
FS	2e-05	34.3916	8.0014	34.2927	8.1955
GC%	7e-06	0.0371	0.1412	0.0367	0.1287
A%	4e-06	0.3343	0.3072	0.3333	0.3064
AA%	2e-06	0.1045	0.2158	0.1041	0.2103
GG%	0e+00	0.0434	0.1508	0.0436	0.1476
AC%	0e+00	0.0498	0.1316	0.0498	0.1351

Table E.29: Experiment 2 Tail Metric Statistics. Ranks each tail metric from the second experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
G%	0.0013	0.2361	0.3075	0.2143	0.2914
CC%	0.0009	0.0260	0.1119	0.0334	0.1355
UU%	0.0008	0.0313	0.1165	0.0386	0.1326
C%	0.0006	0.1611	0.2480	0.1741	0.2574
U%	0.0006	0.1649	0.2371	0.1774	0.2537
UA%	0.0006	0.0607	0.1570	0.0536	0.1457
GC%	0.0003	0.0291	0.1227	0.0338	0.1390
AG%	0.0002	0.0621	0.1743	0.0566	0.1567
AU%	0.0002	0.0526	0.1506	0.0575	0.1522
AA%	0.0002	0.1560	0.2828	0.1480	0.2745
UC%	0.0002	0.0267	0.1026	0.0295	0.1091
CG%	0.0001	0.0180	0.0920	0.0200	0.0921
NLC	0.0001	0.9073	0.1256	0.9043	0.1282
CA%	0.0001	0.0421	0.1328	0.0444	0.1353
SLC	4e-05	0.5841	0.2798	0.5806	0.2749
AC%	3e-05	0.0615	0.1699	0.0598	0.1643
Size	2e-05	4.0579	3.6757	4.0193	3.6162
CU%	2e-05	0.0275	0.0985	0.0265	0.1038
GU%	2e-05	0.0345	0.1221	0.0334	0.1257
GA%	2e-05	0.0664	0.1769	0.0673	0.1773
A%	2e-05	0.4378	0.3437	0.4342	0.3446
UG%	1e-05	0.0257	0.0991	0.0265	0.1018
FS	8e-06	1.0315	0.5701	1.0348	0.5056
CS	3e-06	4.0812	2.1724	4.0723	2.1361
GG%	0e+00	0.0315	0.1347	0.0319	0.1359

Table E.30: Experiment 2 Joint Metric Statistics. Ranks each joint metric from the second experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
CC%	0.0011	0.0264	0.1095	0.0339	0.1311
C%	0.0009	0.1692	0.2471	0.1841	0.2590
G%	0.0006	0.2502	0.3048	0.2353	0.2955
CG%	0.0005	0.0209	0.0953	0.0248	0.0995
UC%	0.0003	0.0293	0.1040	0.0329	0.1126
AG%	0.0002	0.0634	0.1691	0.0593	0.1543
UA%	0.0001	0.0576	0.1449	0.0548	0.1383
AA%	0.0001	0.1307	0.2535	0.1264	0.2455
GU%	0.0001	0.0415	0.1330	0.0391	0.1297
GC%	0.0001	0.0331	0.1321	0.0352	0.1341
GA%	4e-05	0.0678	0.1774	0.0653	0.1692
SLC	3e-05	0.5578	0.2836	0.5551	0.2847
UG%	3e-05	0.0337	0.1157	0.0325	0.1092
CU%	3e-05	0.0334	0.1077	0.0323	0.1127
AU%	3e-05	0.0546	0.1477	0.0561	0.1410
CA%	2e-05	0.0443	0.1340	0.0431	0.1276
Size	1e-05	4.5483	4.3071	4.5792	4.2742
NLC	1e-05	0.9117	0.1200	0.9109	0.1206
A%	1e-05	0.3871	0.3304	0.3854	0.3305
U%	9e-06	0.1934	0.2532	0.1951	0.2552
AC%	8e-06	0.0557	0.1524	0.0549	0.1507
CS	6e-06	4.6283	2.5614	4.6402	2.5301
FS	4e-06	1.1428	0.9083	1.1453	0.8148
UU%	0e+00	0.0436	0.1392	0.0433	0.1358
GG%	0e+00	0.0374	0.1429	0.0372	0.1411

Table E.31: Experiment 2 Joint-Tail Metric Statistics. Ranks each joint-tail metric from the second experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

		Gene		**Nongene**	
Feature	**F-score**	**Mean**	**Std.**	**Mean**	**Std.**
AA%	0.0026	0.1019	0.2242	0.0801	0.1981
SLC	0.0019	0.7561	0.2344	0.7758	0.2359
A%	0.0012	0.3605	0.3425	0.3362	0.3407
NLC	0.0011	0.9287	0.1082	0.9358	0.1053
AG%	0.0010	0.0527	0.1601	0.0430	0.1453
Size	0.0007	3.5474	2.6898	3.4079	2.6680
AC%	0.0007	0.0320	0.1205	0.0268	0.1033
GC%	0.0004	0.0211	0.0990	0.0170	0.0941
GC% Bond	0.0004	0.5491	0.4976	0.5573	0.4967
FS	0.0004	4.4430	1.5795	4.3362	1.5561
UA%	0.0004	0.0403	0.1242	0.0356	0.1169
CG%	0.0004	0.0149	0.0762	0.0178	0.0891
G%	0.0003	0.2479	0.3183	0.2586	0.3364
CC%	0.0003	0.0195	0.1028	0.0230	0.1139
CU%	0.0003	0.0201	0.1032	0.0242	0.1101
GA%	0.0003	0.0530	0.1680	0.0477	0.1614
GU% Bond	0.0002	0.1385	0.3455	0.1409	0.3479
C%	0.0001	0.1761	0.2747	0.1835	0.2825
AU% Bond	0.0001	0.3124	0.4635	0.3018	0.4591
AU% Bond	0.0001	0.3124	0.4635	0.3018	0.4591
CS	0.0001	5.2517	1.9192	5.0648	1.9509
GC% Bond	0.0001	0.5491	0.4976	0.5573	0.4967
AU%	0.0001	0.0305	0.1062	0.0320	0.1104
UC%	0.0001	0.0268	0.1303	0.0246	0.1086
UU%	0.0001	0.0364	0.1503	0.0343	0.1431
U%	0.0001	0.2155	0.3219	0.2216	0.3261
UG%	2e-05	0.0342	0.1530	0.0335	0.1486
GU%	9e-06	0.0207	0.0978	0.0200	0.0997
GG%	8e-06	0.0329	0.1317	0.0336	0.1411
CA%	4e-06	0.0294	0.1114	0.0295	0.1192
GU% Bond	2e-06	0.1385	0.3455	0.1409	0.3479

Table E.32: Experiment 2 Loop Metric Statistics. Ranks each loop metric from the second experiment in descending value of F-score. The mean and standard deviation of the metrics for the SRNAG and non-SRNAG classes are also listed.

E.2 Individual Structural Element Models Gene Class Prediction Statistics

E.2.1 Experiment 1

Gene Type	**TP**	**FP**	**TN**	**FN**	**Acc.**	**Prec.**	**Recall**	**F-mea.**
16S rRNA	0.2993	0.0620	0.4380	0.2007	0.7372	0.8283	0.5985	0.6949
23S rRNA	0.4444	0.0417	0.4583	0.0556	0.9028	0.9143	0.8889	0.9014
5S rRNA	0.4531	0.0312	0.4688	0.0469	0.9219	0.9355	0.9062	0.9206
RNase P	0.4725	0.0275	0.4725	0.0275	0.9451	0.9451	0.9451	0.9451
SRP RNA	0.4744	0.0769	0.4231	0.0256	0.8974	0.8605	0.9487	0.9024
TmRNA	0.3740	0.0472	0.4528	0.1260	0.8268	0.8879	0.7480	0.8120
tRNA	0.5000	0.1471	0.3529	0.0000	0.8529	0.7727	1.0000	0.8718
All	0.4006	0.0541	0.4459	0.0994	0.8465	0.8811	0.8012	0.8392

Table E.33: External Loop Prediction Statistics for Experiment 1.

Gene Type	**TP**	**FP**	**TN**	**FN**	**Acc.**	**Prec.**	**Recall**	**F-mea.**
16S rRNA	0.4307	0.0146	0.4854	0.0693	0.9161	0.9672	0.8613	0.9112
23S rRNA	0.4583	0.0139	0.4861	0.0417	0.9444	0.9706	0.9167	0.9429
5S rRNA	0.4844	0.0781	0.4219	0.0156	0.9062	0.8611	0.9688	0.9118
RNase P	0.4451	0.0220	0.4780	0.0549	0.9231	0.9529	0.8901	0.9205
SRP RNA	0.4359	0.0449	0.4551	0.0641	0.8910	0.9067	0.8718	0.8889
TmRNA	0.1811	0.0315	0.4685	0.3189	0.6496	0.8519	0.3622	0.5083
tRNA	0.3824	0.1176	0.3824	0.1176	0.7647	0.7647	0.7647	0.7647
All	0.3764	0.0319	0.4681	0.1236	0.8446	0.9220	0.7529	0.8289

Table E.34: Structure Loop Prediction Statistics for Experiment 1.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4562	0.0146	0.4854	0.0438	0.9416	0.9690	0.9124	0.9398
23S rRNA	0.4861	0.0278	0.4722	0.0139	0.9583	0.9459	0.9722	0.9589
5S rRNA	0.3125	0.1562	0.3438	0.1875	0.6562	0.6667	0.6250	0.6452
RNase P	0.4231	0.0659	0.4341	0.0769	0.8571	0.8652	0.8462	0.8556
SRP RNA	0.3654	0.1154	0.3846	0.1346	0.7500	0.7600	0.7308	0.7451
TmRNA	0.3425	0.1220	0.3780	0.1575	0.7205	0.7373	0.6850	0.7102
tRNA	0.3824	0.1176	0.3824	0.1176	0.7647	0.7647	0.7647	0.7647
All	0.3996	0.0782	0.4218	0.1004	0.8214	0.8364	0.7992	0.8174

Table E.35: Stemloop Prediction Statistics for Experiment 1.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4745	0.0182	0.4818	0.0255	0.9562	0.9630	0.9489	0.9559
23S rRNA	0.4583	0.0000	0.5000	0.0417	0.9583	1.0000	0.9167	0.9565
5S rRNA	0.4219	0.0938	0.4062	0.0781	0.8281	0.8182	0.8438	0.8308
RNase P	0.4451	0.0879	0.4121	0.0549	0.8571	0.8351	0.8901	0.8617
SRP RNA	0.3782	0.0897	0.4103	0.1218	0.7885	0.8082	0.7564	0.7815
TmRNA	0.2953	0.0906	0.4094	0.2047	0.7047	0.7653	0.5906	0.6667
tRNA	0.3235	0.0294	0.4706	0.1765	0.7941	0.9167	0.6471	0.7586
All	0.4015	0.0627	0.4373	0.0985	0.8388	0.8649	0.8031	0.8328

Table E.36: Hairpin Prediction Statistics for Experiment 1.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.3467	0.1314	0.3686	0.1533	0.7153	0.7252	0.6934	0.7090
23S rRNA	0.2778	0.1528	0.3472	0.2222	0.6250	0.6452	0.5556	0.5970
5S rRNA	0.3594	0.1250	0.3750	0.1406	0.7344	0.7419	0.7188	0.7302
RNase P	0.4176	0.1538	0.3462	0.0824	0.7637	0.7308	0.8352	0.7795
SRP RNA	0.3397	0.1410	0.3590	0.1603	0.6987	0.7067	0.6795	0.6928
TmRNA	0.4094	0.1850	0.3150	0.0906	0.7244	0.6887	0.8189	0.7482
tRNA	0.4118	0.1765	0.3235	0.0882	0.7353	0.7000	0.8235	0.7568
All	0.3716	0.1525	0.3475	0.1284	0.7191	0.7090	0.7432	0.7257

Table E.37: Tail Prediction Statistics for Experiment 1.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4927	0.4964	0.0036	0.0073	0.4964	0.4982	0.9854	0.6618
23S rRNA	0.4861	0.4861	0.0139	0.0139	0.5000	0.5000	0.9722	0.6604
5S rRNA	0.4531	0.3906	0.1094	0.0469	0.5625	0.5370	0.9062	0.6744
RNase P	0.4945	0.4780	0.0220	0.0055	0.5165	0.5085	0.9890	0.6716
SRP RNA	0.3397	0.4295	0.0705	0.1603	0.4103	0.4417	0.6795	0.5354
TmRNA	0.5000	0.4921	0.0079	0.0000	0.5079	0.5040	1.0000	0.6702
tRNA	0.2941	0.2647	0.2353	0.2059	0.5294	0.5263	0.5882	0.5556
All	0.4624	0.4672	0.0328	0.0376	0.4952	0.4974	0.9247	0.6469

Table E.38: Joint Prediction Statistics for Experiment 1.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4964	0.4964	0.0036	0.0036	0.5000	0.5000	0.9927	0.6650
23S rRNA	0.5000	0.4861	0.0139	0.0000	0.5139	0.5070	1.0000	0.6729
5S rRNA	0.4219	0.3281	0.1719	0.0781	0.5938	0.5625	0.8438	0.6750
RNase P	0.4835	0.4176	0.0824	0.0165	0.5659	0.5366	0.9670	0.6902
SRP RNA	0.3846	0.4103	0.0897	0.1154	0.4744	0.4839	0.7692	0.5941
TmRNA	0.4843	0.4606	0.0394	0.0157	0.5236	0.5125	0.9685	0.6703
tRNA	0.3824	0.2941	0.2059	0.1176	0.5882	0.5652	0.7647	0.6500
All	0.4662	0.4431	0.0569	0.0338	0.5232	0.5127	0.9324	0.6616

Table E.39: Joint-Tail Prediction Statistics for Experiment 1.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4343	0.1204	0.3796	0.0657	0.8139	0.7829	0.8686	0.8235
23S rRNA	0.4444	0.0417	0.4583	0.0556	0.9028	0.9143	0.8889	0.9014
5S rRNA	0.1875	0.0469	0.4531	0.3125	0.6406	0.8000	0.3750	0.5106
RNase P	0.4231	0.1374	0.3626	0.0769	0.7857	0.7549	0.8462	0.7979
SRP RNA	0.1474	0.0962	0.4038	0.3526	0.5513	0.6053	0.2949	0.3966
TmRNA	0.4055	0.1260	0.3740	0.0945	0.7795	0.7630	0.8110	0.7863
tRNA	0.1765	0.0294	0.4706	0.3235	0.6471	0.8571	0.3529	0.5000
All	0.3591	0.1081	0.3919	0.1409	0.7510	0.7686	0.7181	0.7425

Table E.40: Bridge Prediction Statistics for Experiment 1.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4270	0.0620	0.4380	0.0730	0.8650	0.8731	0.8540	0.8635
23S rRNA	0.4722	0.0278	0.4722	0.0278	0.9444	0.9444	0.9444	0.9444
5S rRNA	0.2969	0.1719	0.3281	0.2031	0.6250	0.6333	0.5938	0.6129
RNase P	0.4066	0.1758	0.3242	0.0934	0.7308	0.6981	0.8132	0.7513
SRP RNA	0.2500	0.1795	0.3205	0.2500	0.5705	0.5821	0.5000	0.5379
TmRNA	0.3701	0.1102	0.3898	0.1299	0.7598	0.7705	0.7402	0.7550
tRNA	0.3529	0.1471	0.3529	0.1471	0.7059	0.7059	0.7059	0.7059
All	0.3755	0.1187	0.3813	0.1245	0.7568	0.7598	0.7510	0.7553

Table E.41: Stem Prediction Statistics for Experiment 1.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4927	0.4818	0.0182	0.0073	0.5109	0.5056	0.9854	0.6683
23S rRNA	0.5000	0.4861	0.0139	0.0000	0.5139	0.5070	1.0000	0.6729
5S rRNA	0.4219	0.3125	0.1875	0.0781	0.6094	0.5745	0.8438	0.6835
RNase P	0.4945	0.3901	0.1099	0.0055	0.6044	0.5590	0.9890	0.7143
SRP RNA	0.4359	0.3718	0.1282	0.0641	0.5641	0.5397	0.8718	0.6667
TmRNA	0.4528	0.3780	0.1220	0.0472	0.5748	0.5450	0.9055	0.6805
tRNA	0.5000	0.2647	0.2353	0.0000	0.7353	0.6538	1.0000	0.7907
All	0.4710	0.4064	0.0936	0.0290	0.5647	0.5369	0.9421	0.6840

Table E.42: Stack Prediction Statistics for Experiment 1.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.3978	0.0182	0.4818	0.1022	0.8796	0.9561	0.7956	0.8685
23S rRNA	0.4444	0.0000	0.5000	0.0556	0.9444	1.0000	0.8889	0.9412
5S rRNA	0.2188	0.0156	0.4844	0.2812	0.7031	0.9333	0.4375	0.5957
RNase P	0.3242	0.0165	0.4835	0.1758	0.8077	0.9516	0.6484	0.7712
SRP RNA	0.1538	0.0192	0.4808	0.3462	0.6346	0.8889	0.3077	0.4571
TmRNA	0.2953	0.0236	0.4764	0.2047	0.7717	0.9259	0.5906	0.7212
tRNA	0.0588	0.0000	0.5000	0.4412	0.5588	1.0000	0.1176	0.2105
All	0.3041	0.0174	0.4826	0.1959	0.7867	0.9459	0.6081	0.7403

Table E.43: Multiloop Prediction Statistics for Experiment 1.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4964	0.4964	0.0036	0.0036	0.5000	0.5000	0.9927	0.6650
23S rRNA	0.5000	0.4861	0.0139	0.0000	0.5139	0.5070	1.0000	0.6729
5S rRNA	0.4219	0.3281	0.1719	0.0781	0.5938	0.5625	0.8438	0.6750
RNase P	0.4835	0.4176	0.0824	0.0165	0.5659	0.5366	0.9670	0.6902
SRP RNA	0.3846	0.4103	0.0897	0.1154	0.4744	0.4839	0.7692	0.5941
TmRNA	0.4843	0.4606	0.0394	0.0157	0.5236	0.5125	0.9685	0.6703
tRNA	0.3824	0.2941	0.2059	0.1176	0.5882	0.5652	0.7647	0.6500
All	0.4662	0.4431	0.0569	0.0338	0.5232	0.5127	0.9324	0.6616

Table E.44: Junction Prediction Statistics for Experiment 1.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4307	0.2299	0.2701	0.0693	0.7007	0.6519	0.8613	0.7421
23S rRNA	0.4722	0.2083	0.2917	0.0278	0.7639	0.6939	0.9444	0.8000
5S rRNA	0.3750	0.1719	0.3281	0.1250	0.7031	0.6857	0.7500	0.7164
RNase P	0.3626	0.2473	0.2527	0.1374	0.6154	0.5946	0.7253	0.6535
SRP RNA	0.3782	0.1410	0.3590	0.1218	0.7372	0.7284	0.7564	0.7421
TmRNA	0.2717	0.2283	0.2717	0.2283	0.5433	0.5433	0.5433	0.5433
tRNA	0.2353	0.0882	0.4118	0.2647	0.6471	0.7273	0.4706	0.5714
All	0.3649	0.2095	0.2905	0.1351	0.6554	0.6353	0.7297	0.6792

Table E.45: Unpaired Prediction Statistics for Experiment 1.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4562	0.0985	0.4015	0.0438	0.8577	0.8224	0.9124	0.8651
23S rRNA	0.4722	0.0417	0.4583	0.0278	0.9306	0.9189	0.9444	0.9315
5S rRNA	0.3438	0.2031	0.2969	0.1562	0.6406	0.6286	0.6875	0.6567
RNase P	0.2527	0.1429	0.3571	0.2473	0.6099	0.6389	0.5055	0.5644
SRP RNA	0.3974	0.1282	0.3718	0.1026	0.7692	0.7561	0.7949	0.7750
TmRNA	0.2835	0.1457	0.3543	0.2165	0.6378	0.6606	0.5669	0.6102
tRNA	0.0882	0.0588	0.4412	0.4118	0.5294	0.6000	0.1765	0.2727
All	0.3514	0.1236	0.3764	0.1486	0.7278	0.7398	0.7027	0.7208

Table E.46: Internal Loop Prediction Statistics for Experiment 1.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4197	0.2190	0.2810	0.0803	0.7007	0.6571	0.8394	0.7372
23S rRNA	0.4861	0.1528	0.3472	0.0139	0.8333	0.7609	0.9722	0.8537
5S rRNA	0.3438	0.1875	0.3125	0.1562	0.6562	0.6471	0.6875	0.6667
RNase P	0.3077	0.2253	0.2747	0.1923	0.5824	0.5773	0.6154	0.5957
SRP RNA	0.3077	0.2115	0.2885	0.1923	0.5962	0.5926	0.6154	0.6038
TmRNA	0.3740	0.1654	0.3346	0.1260	0.7087	0.6934	0.7480	0.7197
tRNA	0.2059	0.2059	0.2941	0.2941	0.5000	0.5000	0.4118	0.4516
All	0.3649	0.1988	0.3012	0.1351	0.6660	0.6473	0.7297	0.6860

Table E.47: Loop Prediction Statistics for Experiment 1.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4635	0.4124	0.0876	0.0365	0.5511	0.5292	0.9270	0.6737
23S rRNA	0.4583	0.4028	0.0972	0.0417	0.5556	0.5323	0.9167	0.6735
5S rRNA	0.1875	0.1875	0.3125	0.3125	0.5000	0.5000	0.3750	0.4286
RNase P	0.3681	0.3297	0.1703	0.1319	0.5385	0.5276	0.7363	0.6147
SRP RNA	0.3397	0.2885	0.2115	0.1603	0.5513	0.5408	0.6795	0.6023
TmRNA	0.3268	0.2992	0.2008	0.1732	0.5276	0.5220	0.6535	0.5804
tRNA	0.0588	0.1471	0.3529	0.4412	0.4118	0.2857	0.1176	0.1667
All	0.3639	0.3282	0.1718	0.1361	0.5357	0.5258	0.7278	0.6105

Table E.48: Bulge Prediction Statistics for Experiment 1.

E.2.2 Experiment 2

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.0414	0.0207	0.4793	0.4586	0.5207	0.6667	0.0828	0.1472
23S rRNA	0.0366	0.0122	0.4878	0.4634	0.5244	0.7500	0.0732	0.1333
5S rRNA	0.1774	0.0161	0.4839	0.3226	0.6613	0.9167	0.3548	0.5116
RNase P	0.0851	0.0426	0.4574	0.4149	0.5426	0.6667	0.1702	0.2712
SRP RNA	0.0506	0.0253	0.4747	0.4494	0.5253	0.6667	0.1013	0.1758
TmRNA	0.0931	0.0069	0.4931	0.4069	0.5862	0.9310	0.1862	0.3103
tRNA	0.3750	0.0000	0.5000	0.1250	0.8750	1.0000	0.7500	0.8571
All	0.0742	0.0204	0.4796	0.4258	0.5538	0.7843	0.1484	0.2496

Table E.49: Junction Prediction Statistics for Experiment 2.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4690	0.4586	0.0414	0.0310	0.5103	0.5056	0.9379	0.6570
23S rRNA	0.5000	0.4634	0.0366	0.0000	0.5366	0.5190	1.0000	0.6833
5S rRNA	0.5000	0.4839	0.0161	0.0000	0.5161	0.5082	1.0000	0.6739
RNase P	0.4734	0.4840	0.0160	0.0266	0.4894	0.4944	0.9468	0.6496
SRP RNA	0.4747	0.4747	0.0253	0.0253	0.5000	0.5000	0.9494	0.6550
TmRNA	0.4862	0.4586	0.0414	0.0138	0.5276	0.5146	0.9724	0.6730
tRNA	0.5000	0.3750	0.1250	0.0000	0.6250	0.5714	1.0000	0.7273
All	0.4796	0.4666	0.0334	0.0204	0.5130	0.5069	0.9592	0.6632

Table E.50: Stack Prediction Statistics for Experiment 2.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.3966	0.3103	0.1897	0.1034	0.5862	0.5610	0.7931	0.6571
23S rRNA	0.4146	0.3049	0.1951	0.0854	0.6098	0.5763	0.8293	0.6800
5S rRNA	0.4516	0.2903	0.2097	0.0484	0.6613	0.6087	0.9032	0.7273
RNase P	0.4521	0.2979	0.2021	0.0479	0.6543	0.6028	0.9043	0.7234
SRP RNA	0.3987	0.3481	0.1519	0.1013	0.5506	0.5339	0.7975	0.6396
TmRNA	0.3897	0.3241	0.1759	0.1103	0.5655	0.5459	0.7793	0.6420
tRNA	0.3750	0.2500	0.2500	0.1250	0.6250	0.6000	0.7500	0.6667
All	0.4091	0.3154	0.1846	0.0909	0.5937	0.5647	0.8182	0.6682

Table E.51: Stemloop Prediction Statistics for Experiment 2.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.2448	0.1690	0.3310	0.2552	0.5759	0.5917	0.4897	0.5358
23S rRNA	0.2439	0.1707	0.3293	0.2561	0.5732	0.5882	0.4878	0.5333
5S rRNA	0.1452	0.2097	0.2903	0.3548	0.4355	0.4091	0.2903	0.3396
RNase P	0.3138	0.1436	0.3564	0.1862	0.6702	0.6860	0.6277	0.6556
SRP RNA	0.2152	0.1646	0.3354	0.2848	0.5506	0.5667	0.4304	0.4892
TmRNA	0.2241	0.1586	0.3414	0.2759	0.5655	0.5856	0.4483	0.5078
tRNA	0.3750	0.2500	0.2500	0.1250	0.6250	0.6000	0.7500	0.6667
All	0.2421	0.1642	0.3358	0.2579	0.5779	0.5959	0.4842	0.5343

Table E.52: Stem Prediction Statistics for Experiment 2.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.3276	0.1586	0.3414	0.1724	0.6690	0.6738	0.6552	0.6643
23S rRNA	0.2927	0.1707	0.3293	0.2073	0.6220	0.6316	0.5854	0.6076
5S rRNA	0.2419	0.0645	0.4355	0.2581	0.6774	0.7895	0.4839	0.6000
RNase P	0.3617	0.1543	0.3457	0.1383	0.7074	0.7010	0.7234	0.7120
SRP RNA	0.2405	0.1392	0.3608	0.2595	0.6013	0.6333	0.4810	0.5468
TmRNA	0.2966	0.2517	0.2483	0.2034	0.5448	0.5409	0.5931	0.5658
tRNA	0.0000	0.1250	0.3750	0.5000	0.3750	0.0000	0.0000	0.0000
All	0.3024	0.1753	0.3247	0.1976	0.6271	0.6330	0.6048	0.6186

Table E.53: Structure Prediction Statistics for Experiment 2.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.3379	0.1483	0.3517	0.1621	0.6897	0.6950	0.6759	0.6853
23S rRNA	0.2683	0.1829	0.3171	0.2317	0.5854	0.5946	0.5366	0.5641
5S rRNA	0.4194	0.0968	0.4032	0.0806	0.8226	0.8125	0.8387	0.8254
RNase P	0.3830	0.1702	0.3298	0.1170	0.7128	0.6923	0.7660	0.7273
SRP RNA	0.3101	0.1582	0.3418	0.1899	0.6519	0.6622	0.6203	0.6405
TmRNA	0.2828	0.1655	0.3345	0.2172	0.6172	0.6308	0.5655	0.5964
tRNA	0.3750	0.0000	0.5000	0.1250	0.8750	1.0000	0.7500	0.8571
All	0.3265	0.1568	0.3432	0.1735	0.6698	0.6756	0.6531	0.6642

Table E.54: Hairpin Prediction Statistics for Experiment 2.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.0483	0.0103	0.4897	0.4517	0.5379	0.8235	0.0966	0.1728
23S rRNA	0.0366	0.0366	0.4634	0.4634	0.5000	0.5000	0.0732	0.1277
5S rRNA	0.0323	0.0484	0.4516	0.4677	0.4839	0.4000	0.0645	0.1111
RNase P	0.0851	0.0266	0.4734	0.4149	0.5585	0.7619	0.1702	0.2783
SRP RNA	0.0316	0.0253	0.4747	0.4684	0.5063	0.5556	0.0633	0.1136
TmRNA	0.0586	0.0241	0.4759	0.4414	0.5345	0.7083	0.1172	0.2012
tRNA	0.2500	0.0000	0.5000	0.2500	0.7500	1.0000	0.5000	0.6667
All	0.0547	0.0232	0.4768	0.4453	0.5315	0.7024	0.1095	0.1894

Table E.55: Bridge Prediction Statistics for Experiment 2.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4931	0.4828	0.0172	0.0069	0.5103	0.5053	0.9862	0.6682
23S rRNA	0.4512	0.4878	0.0122	0.0488	0.4634	0.4805	0.9024	0.6271
5S rRNA	0.5000	0.5000	0.0000	0.0000	0.5000	0.5000	1.0000	0.6667
RNase P	0.4947	0.4787	0.0213	0.0053	0.5160	0.5082	0.9894	0.6715
SRP RNA	0.4810	0.4873	0.0127	0.0190	0.4937	0.4967	0.9620	0.6552
TmRNA	0.4828	0.4828	0.0172	0.0172	0.5000	0.5000	0.9655	0.6588
tRNA	0.5000	0.5000	0.0000	0.0000	0.5000	0.5000	1.0000	0.6667
All	0.4861	0.4842	0.0158	0.0139	0.5019	0.5010	0.9722	0.6612

Table E.56: Unpaired Prediction Statistics for Experiment 2.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.3379	0.3276	0.1724	0.1621	0.5103	0.5078	0.6759	0.5799
23S rRNA	0.3049	0.3780	0.1220	0.1951	0.4268	0.4464	0.6098	0.5155
5Sr RNA	0.4516	0.3871	0.1129	0.0484	0.5645	0.5385	0.9032	0.6747
RNase P	0.3723	0.3245	0.1755	0.1277	0.5479	0.5344	0.7447	0.6222
SRP RNA	0.3544	0.3038	0.1962	0.1456	0.5506	0.5385	0.7089	0.6120
TmRNA	0.3310	0.3103	0.1897	0.1690	0.5207	0.5161	0.6621	0.5801
tRNA	0.1250	0.2500	0.2500	0.3750	0.3750	0.3333	0.2500	0.2857
All	0.3469	0.3256	0.1744	0.1531	0.5213	0.5159	0.6939	0.5918

Table E.57: Loop Prediction Statistics for Experiment 2.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.2207	0.2241	0.2759	0.2793	0.4966	0.4961	0.4414	0.4672
23S rRNA	0.1707	0.2439	0.2561	0.3293	0.4268	0.4118	0.3415	0.3733
5S rRNA	0.3710	0.2581	0.2419	0.1290	0.6129	0.5897	0.7419	0.6571
RNase P	0.2872	0.1862	0.3138	0.2128	0.6011	0.6067	0.5745	0.5902
SRP RNA	0.2215	0.2089	0.2911	0.2785	0.5127	0.5147	0.4430	0.4762
TmRNA	0.3034	0.2138	0.2862	0.1966	0.5897	0.5867	0.6069	0.5966
tRNA	0.5000	0.2500	0.2500	0.0000	0.7500	0.6667	1.0000	0.8000
All	0.2616	0.2161	0.2839	0.2384	0.5455	0.5476	0.5232	0.5351

Table E.58: External Loop Prediction Statistics for Experiment 2.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.2207	0.1655	0.3345	0.2793	0.5552	0.5714	0.4414	0.4981
23S rRNA	0.1341	0.2195	0.2805	0.3659	0.4146	0.3793	0.2683	0.3143
5S rRNA	0.3548	0.0968	0.4032	0.1452	0.7581	0.7857	0.7097	0.7458
RNase P	0.1755	0.1489	0.3511	0.3245	0.5266	0.5410	0.3511	0.4258
SRP RNA	0.2658	0.1392	0.3608	0.2342	0.6266	0.6562	0.5316	0.5874
TmRNA	0.1862	0.1621	0.3379	0.3138	0.5241	0.5347	0.3724	0.4390
tRNA	0.1250	0.1250	0.3750	0.3750	0.5000	0.5000	0.2500	0.3333
All	0.2106	0.1577	0.3423	0.2894	0.5529	0.5718	0.4212	0.4850

Table E.59: Internal Loop Prediction Statistics for Experiment 2.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.3000	0.2828	0.2172	0.2000	0.5172	0.5148	0.6000	0.5541
23S rRNA	0.2561	0.2805	0.2195	0.2439	0.4756	0.4773	0.5122	0.4941
5S rRNA	0.1774	0.2419	0.2581	0.3226	0.4355	0.4231	0.3548	0.3860
RNase P	0.3617	0.2713	0.2287	0.1383	0.5904	0.5714	0.7234	0.6385
SRP RNA	0.2278	0.2532	0.2468	0.2722	0.4747	0.4737	0.4557	0.4645
TmRNA	0.3069	0.2448	0.2552	0.1931	0.5621	0.5563	0.6138	0.5836
tRNA	0.2500	0.2500	0.2500	0.2500	0.5000	0.5000	0.5000	0.5000
All	0.2913	0.2635	0.2365	0.2087	0.5278	0.5251	0.5826	0.5523

Table E.60: Joint Prediction Statistics for Experiment 2.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4310	0.4207	0.0793	0.0690	0.5103	0.5061	0.8621	0.6378
23S rRNA	0.4024	0.3780	0.1220	0.0976	0.5244	0.5156	0.8049	0.6286
5S rRNA	0.3387	0.4194	0.0806	0.1613	0.4194	0.4468	0.6774	0.5385
RNase P	0.4574	0.3936	0.1064	0.0426	0.5638	0.5375	0.9149	0.6772
SRP RNA	0.3861	0.4114	0.0886	0.1139	0.4747	0.4841	0.7722	0.5951
TmRNA	0.4276	0.3655	0.1345	0.0724	0.5621	0.5391	0.8552	0.6613
tRNA	0.5000	0.3750	0.1250	0.0000	0.6250	0.5714	1.0000	0.7273
All	0.4212	0.3961	0.1039	0.0788	0.5250	0.5153	0.8423	0.6394

Table E.61: Joint-Tail Prediction Statistics for Experiment 2.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.0586	0.0759	0.4241	0.4414	0.4828	0.4359	0.1172	0.1848
23S rRNA	0.0854	0.1463	0.3537	0.4146	0.4390	0.3684	0.1707	0.2333
5S rRNA	0.0968	0.0484	0.4516	0.4032	0.5484	0.6667	0.1935	0.3000
RNase P	0.1330	0.0957	0.4043	0.3670	0.5372	0.5814	0.2660	0.3650
SRP RNA	0.1139	0.0759	0.4241	0.3861	0.5380	0.6000	0.2278	0.3303
TmRNA	0.1069	0.0759	0.4241	0.3931	0.5310	0.5849	0.2138	0.3131
tRNA	0.0000	0.0000	0.5000	0.5000	0.5000	0.0000	0.0000	0.0000
All	0.0965	0.0826	0.4174	0.4035	0.5139	0.5389	0.1929	0.2842

Table E.62: Tail Prediction Statistics for Experiment 2.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.0345	0.0000	0.5000	0.4655	0.5345	1.0000	0.0690	0.1290
23S rRNA	0.0244	0.0000	0.5000	0.4756	0.5244	1.0000	0.0488	0.0930
5S rRNA	0.0968	0.0161	0.4839	0.4032	0.5806	0.8571	0.1935	0.3158
RNase P	0.0319	0.0106	0.4894	0.4681	0.5213	0.7500	0.0638	0.1176
SRP RNA	0.0190	0.0000	0.5000	0.4810	0.5190	1.0000	0.0380	0.0732
TmRNA	0.0345	0.0103	0.4897	0.4655	0.5241	0.7692	0.0690	0.1266
tRNA	0.3750	0.0000	0.5000	0.1250	0.8750	1.0000	0.7500	0.8571
All	0.0371	0.0056	0.4944	0.4629	0.5315	0.8696	0.0742	0.1368

Table E.63: Multiloop Prediction Statistics for Experiment 2.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.1069	0.1034	0.3966	0.3931	0.5034	0.5082	0.2138	0.3010
23S rRNA	0.0976	0.1341	0.3659	0.4024	0.4634	0.4211	0.1951	0.2667
5S rRNA	0.1935	0.1290	0.3710	0.3065	0.5645	0.6000	0.3871	0.4706
RNase P	0.1117	0.0904	0.4096	0.3883	0.5213	0.5526	0.2234	0.3182
SRP RNA	0.0759	0.1076	0.3924	0.4241	0.4684	0.4138	0.1519	0.2222
TmRNA	0.1000	0.0931	0.4069	0.4000	0.5069	0.5179	0.2000	0.2886
tRNA	0.0000	0.0000	0.5000	0.5000	0.5000	0.0000	0.0000	0.0000
All	0.1048	0.1020	0.3980	0.3952	0.5028	0.5067	0.2096	0.2966

Table E.64: Bulge Prediction Statistics for Experiment 2.

E.3 Progressively Inclusive Structural Element Voting Results

E.3.1 Experiment 1

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.2993	0.0620	0.4380	0.2007	0.7372	0.8283	0.5985	0.6949
23S rRNA	0.4444	0.0417	0.4583	0.0556	0.9028	0.9143	0.8889	0.9014
5S rRNA	0.4531	0.0312	0.4688	0.0469	0.9219	0.9355	0.9062	0.9206
RNase P	0.4725	0.0275	0.4725	0.0275	0.9451	0.9451	0.9451	0.9451
SRP RNA	0.4744	0.0769	0.4231	0.0256	0.8974	0.8605	0.9487	0.9024
TmRNA	0.3740	0.0472	0.4528	0.1260	0.8268	0.8879	0.7480	0.8120
tRNA	0.5000	0.1471	0.3529	0.0000	0.8529	0.7727	1.0000	0.8718
All	0.4006	0.0541	0.4459	0.0994	0.8465	0.8811	0.8012	0.8392

Table E.65: Greedy Voting Size 1 Prediction Statistics for Experiment 1. The progressive voting group contains the external loop.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4854	0.0693	0.4307	0.0146	0.9161	0.8750	0.9708	0.9204
23S rRNA	0.4722	0.0556	0.4444	0.0278	0.9167	0.8947	0.9444	0.9189
5S rRNA	0.5000	0.1094	0.3906	0.0000	0.8906	0.8205	1.0000	0.9014
RNase P	0.4945	0.0440	0.4560	0.0055	0.9505	0.9184	0.9890	0.9524
SRP RNA	0.5000	0.1218	0.3782	0.0000	0.8782	0.8041	1.0000	0.8914
TmRNA	0.3937	0.0787	0.4213	0.1063	0.8150	0.8333	0.7874	0.8097
tRNA	0.5000	0.2353	0.2647	0.0000	0.7647	0.6800	1.0000	0.8095
All	0.4672	0.0820	0.4180	0.0328	0.8851	0.8506	0.9344	0.8905

Table E.66: Greedy Voting Size 2 Prediction Statistics for Experiment 1. The progressive voting group contains the external loop and structure.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4818	0.0438	0.4562	0.0182	0.9380	0.9167	0.9635	0.9395
23S rRNA	0.4722	0.0139	0.4861	0.0278	0.9583	0.9714	0.9444	0.9577
5S rRNA	0.5000	0.0469	0.4531	0.0000	0.9531	0.9143	1.0000	0.9552
RNase P	0.4945	0.0385	0.4615	0.0055	0.9560	0.9278	0.9890	0.9574
SRP RNA	0.4808	0.0577	0.4423	0.0192	0.9231	0.8929	0.9615	0.9259
TmRNA	0.3701	0.0472	0.4528	0.1299	0.8228	0.8868	0.7402	0.8069
tRNA	0.5000	0.1176	0.3824	0.0000	0.8824	0.8095	1.0000	0.8947
All	0.4575	0.0463	0.4537	0.0425	0.9112	0.9080	0.9151	0.9115

Table E.67: Greedy Voting Size 3 Prediction Statistics for Experiment 1. The progressive voting group contains the external loop, structure, and stemloop.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4854	0.0182	0.4818	0.0146	0.9672	0.9638	0.9708	0.9673
23S rRNA	0.4722	0.0139	0.4861	0.0278	0.9583	0.9714	0.9444	0.9577
5S rRNA	0.5000	0.0312	0.4688	0.0000	0.9688	0.9412	1.0000	0.9697
RNase P	0.4945	0.0165	0.4835	0.0055	0.9780	0.9677	0.9890	0.9783
SRP RNA	0.4679	0.0577	0.4423	0.0321	0.9103	0.8902	0.9359	0.9125
TmRNA	0.3819	0.0433	0.4567	0.1181	0.8386	0.8981	0.7638	0.8255
tRNA	0.5000	0.0882	0.4118	0.0000	0.9118	0.8500	1.0000	0.9189
All	0.4595	0.0328	0.4672	0.0405	0.9266	0.9333	0.9189	0.9261

Table E.68: Greedy Voting Size 4 Prediction Statistics for Experiment 1. The progressive voting group contains the external loop, structure, stemloop, and hairpin.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4818	0.0219	0.4781	0.0182	0.9599	0.9565	0.9635	0.9600
23S rRNA	0.4583	0.0278	0.4722	0.0417	0.9306	0.9429	0.9167	0.9296
5S rRNA	0.5000	0.0156	0.4844	0.0000	0.9844	0.9697	1.0000	0.9846
RNase P	0.4890	0.0275	0.4725	0.0110	0.9615	0.9468	0.9780	0.9622
SRP RNA	0.4744	0.0513	0.4487	0.0256	0.9231	0.9024	0.9487	0.9250
TmRNA	0.3819	0.0315	0.4685	0.1181	0.8504	0.9238	0.7638	0.8362
tRNA	0.4706	0.0882	0.4118	0.0294	0.8824	0.8421	0.9412	0.8889
All	0.4566	0.0319	0.4681	0.0434	0.9247	0.9348	0.9131	0.9238

Table E.69: Greedy Voting Size 5 Prediction Statistics for Experiment 1. The progressive voting group contains the external loop, structure, stemloop, hairpin, and tail.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4708	0.0146	0.4854	0.0292	0.9562	0.9699	0.9416	0.9556
23S rRNA	0.4583	0.0139	0.4861	0.0417	0.9444	0.9706	0.9167	0.9429
5S rRNA	0.5000	0.0000	0.5000	0.0000	1.0000	1.0000	1.0000	1.0000
RNase P	0.4890	0.0220	0.4780	0.0110	0.9670	0.9570	0.9780	0.9674
SRP RNA	0.4615	0.0449	0.4551	0.0385	0.9167	0.9114	0.9231	0.9172
TmRNA	0.3780	0.0276	0.4724	0.1220	0.8504	0.9320	0.7559	0.8348
tRNA	0.4706	0.0882	0.4118	0.0294	0.8824	0.8421	0.9412	0.8889
All	0.4508	0.0251	0.4749	0.0492	0.9257	0.9473	0.9015	0.9238

Table E.70: Greedy Voting Size 6 Prediction Statistics for Experiment 1. The progressive voting group contains the external loop, structure, stemloop, hairpin, tail, and joint.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4818	0.0255	0.4745	0.0182	0.9562	0.9496	0.9635	0.9565
23S rRNA	0.4583	0.0139	0.4861	0.0417	0.9444	0.9706	0.9167	0.9429
5S rRNA	0.5000	0.0156	0.4844	0.0000	0.9844	0.9697	1.0000	0.9846
RNase P	0.4945	0.0275	0.4725	0.0055	0.9670	0.9474	0.9890	0.9677
SRP RNA	0.4744	0.0449	0.4551	0.0256	0.9295	0.9136	0.9487	0.9308
TmRNA	0.3819	0.0354	0.4646	0.1181	0.8465	0.9151	0.7638	0.8326
tRNA	0.4706	0.0882	0.4118	0.0294	0.8824	0.8421	0.9412	0.8889
All	0.4575	0.0319	0.4681	0.0425	0.9257	0.9349	0.9151	0.9249

Table E.71: Greedy Voting Size 7 Prediction Statistics for Experiment 1. The progressive voting group contains the external loop, structure, stemloop, hairpin, tail, joint, and joint-tail.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4745	0.0146	0.4854	0.0255	0.9599	0.9701	0.9489	0.9594
23S rRNA	0.4583	0.0139	0.4861	0.0417	0.9444	0.9706	0.9167	0.9429
5S rRNA	0.4844	0.0156	0.4844	0.0156	0.9688	0.9688	0.9688	0.9688
RNase P	0.4890	0.0220	0.4780	0.0110	0.9670	0.9570	0.9780	0.9674
SRP RNA	0.4679	0.0513	0.4487	0.0321	0.9167	0.9012	0.9359	0.9182
TmRNA	0.3819	0.0315	0.4685	0.1181	0.8504	0.9238	0.7638	0.8362
tRNA	0.4706	0.0882	0.4118	0.0294	0.8824	0.8421	0.9412	0.8889
All	0.4527	0.0280	0.4720	0.0473	0.9247	0.9418	0.9054	0.9232

Table E.72: Greedy Voting Size 8 Prediction Statistics for Experiment 1. The progressive voting group contains the external loop, structure, stemloop, hairpin, tail, joint, joint-tail, and bridge.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4745	0.0146	0.4854	0.0255	0.9599	0.9701	0.9489	0.9594
23S rRNA	0.4583	0.0278	0.4722	0.0417	0.9306	0.9429	0.9167	0.9296
5S rRNA	0.5000	0.0156	0.4844	0.0000	0.9844	0.9697	1.0000	0.9846
RNase P	0.4890	0.0220	0.4780	0.0110	0.9670	0.9570	0.9780	0.9674
SRP RNA	0.4679	0.0513	0.4487	0.0321	0.9167	0.9012	0.9359	0.9182
TmRNA	0.3819	0.0236	0.4764	0.1181	0.8583	0.9417	0.7638	0.8435
tRNA	0.4706	0.0882	0.4118	0.0294	0.8824	0.8421	0.9412	0.8889
All	0.4537	0.0270	0.4730	0.0463	0.9266	0.9438	0.9073	0.9252

Table E.73: Greedy Voting Size 9 Prediction Statistics for Experiment 1. The progressive voting group contains the external loop, structure, stemloop, hairpin, tail, joint, joint-tail, bridge, and stem.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4781	0.0146	0.4854	0.0219	0.9635	0.9704	0.9562	0.9632
23S rRNA	0.4583	0.0278	0.4722	0.0417	0.9306	0.9429	0.9167	0.9296
5S rRNA	0.5000	0.0312	0.4688	0.0000	0.9688	0.9412	1.0000	0.9697
RNase P	0.4890	0.0220	0.4780	0.0110	0.9670	0.9570	0.9780	0.9674
SRP RNA	0.4679	0.0513	0.4487	0.0321	0.9167	0.9012	0.9359	0.9182
TmRNA	0.3819	0.0236	0.4764	0.1181	0.8583	0.9417	0.7638	0.8435
tRNA	0.4706	0.0882	0.4118	0.0294	0.8824	0.8421	0.9412	0.8889
All	0.4546	0.0280	0.4720	0.0454	0.9266	0.9420	0.9093	0.9253

Table E.74: Greedy Voting Size 10 Prediction Statistics for Experiment 1. The progressive voting group contains the external loop, structure, stemloop, hairpin, tail, joint, joint-tail, bridge, stem, and stack.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4708	0.0146	0.4854	0.0292	0.9562	0.9699	0.9416	0.9556
23S rRNA	0.4583	0.0139	0.4861	0.0417	0.9444	0.9706	0.9167	0.9429
5S rRNA	0.4844	0.0156	0.4844	0.0156	0.9688	0.9688	0.9688	0.9688
RNase P	0.4890	0.0220	0.4780	0.0110	0.9670	0.9570	0.9780	0.9674
SRP RNA	0.4744	0.0513	0.4487	0.0256	0.9231	0.9024	0.9487	0.9250
TmRNA	0.3819	0.0157	0.4843	0.1181	0.8661	0.9604	0.7638	0.8509
tRNA	0.4706	0.0882	0.4118	0.0294	0.8824	0.8421	0.9412	0.8889
All	0.4527	0.0241	0.4759	0.0473	0.9286	0.9494	0.9054	0.9269

Table E.75: Greedy Voting Size 11 Prediction Statistics for Experiment 1. The progressive voting group contains the external loop, structure, stemloop, hairpin, tail, joint, joint-tail, bridge, stem, stack, and multiloop.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4818	0.0182	0.4818	0.0182	0.9635	0.9635	0.9635	0.9635
23S rRNA	0.4583	0.0278	0.4722	0.0417	0.9306	0.9429	0.9167	0.9296
5S rRNA	0.5000	0.0469	0.4531	0.0000	0.9531	0.9143	1.0000	0.9552
RNase P	0.4890	0.0220	0.4780	0.0110	0.9670	0.9570	0.9780	0.9674
SRP RNA	0.4872	0.0577	0.4423	0.0128	0.9295	0.8941	0.9744	0.9325
TmRNA	0.3858	0.0236	0.4764	0.1142	0.8622	0.9423	0.7717	0.8485
tRNA	0.5000	0.0882	0.4118	0.0000	0.9118	0.8500	1.0000	0.9189
All	0.4604	0.0309	0.4691	0.0396	0.9295	0.9371	0.9208	0.9289

Table E.76: Greedy Voting Size 12 Prediction Statistics for Experiment 1. The progressive voting group contains the external loop, structure, stemloop, hairpin, tail, joint, joint-tail, bridge, stem, stack, multiloop, and junction.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4854	0.0219	0.4781	0.0146	0.9635	0.9568	0.9708	0.9638
23S rRNA	0.4583	0.0139	0.4861	0.0417	0.9444	0.9706	0.9167	0.9429
5S rRNA	0.5000	0.0469	0.4531	0.0000	0.9531	0.9143	1.0000	0.9552
RNase P	0.4890	0.0220	0.4780	0.0110	0.9670	0.9570	0.9780	0.9674
SRP RNA	0.4872	0.0577	0.4423	0.0128	0.9295	0.8941	0.9744	0.9325
TmRNA	0.3858	0.0236	0.4764	0.1142	0.8622	0.9423	0.7717	0.8485
tRNA	0.5000	0.0882	0.4118	0.0000	0.9118	0.8500	1.0000	0.9189
All	0.4614	0.0309	0.4691	0.0386	0.9305	0.9373	0.9228	0.9300

Table E.77: Greedy Voting Size 13 Prediction Statistics for Experiment 1. The progressive voting group contains the external loop, structure, stemloop, hairpin, tail, joint, joint-tail, bridge, stem, stack, multiloop, junction, and unpaired.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4599	0.0146	0.4854	0.0401	0.9453	0.9692	0.9197	0.9438
23S rRNA	0.4583	0.0139	0.4861	0.0417	0.9444	0.9706	0.9167	0.9429
5S rRNA	0.4844	0.0000	0.5000	0.0156	0.9844	1.0000	0.9688	0.9841
RNase P	0.4890	0.0165	0.4835	0.0110	0.9725	0.9674	0.9780	0.9727
SRP RNA	0.4679	0.0385	0.4615	0.0321	0.9295	0.9241	0.9359	0.9299
TmRNA	0.3780	0.0039	0.4961	0.1220	0.8740	0.9897	0.7559	0.8571
tRNA	0.5000	0.0000	0.5000	0.0000	1.0000	1.0000	1.0000	1.0000
All	0.4488	0.0145	0.4855	0.0512	0.9344	0.9688	0.8977	0.9319

Table E.78: Greedy Voting Size 14 Prediction Statistics for Experiment 1. The progressive voting group contains the external loop, structure, stemloop, hairpin, tail, joint, joint-tail, bridge, stem, stack, multiloop, junction, unpaired, and internal loop.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4672	0.0146	0.4854	0.0328	0.9526	0.9697	0.9343	0.9517
23S rRNA	0.4583	0.0139	0.4861	0.0417	0.9444	0.9706	0.9167	0.9429
5S rRNA	0.4844	0.0000	0.5000	0.0156	0.9844	1.0000	0.9688	0.9841
RNase P	0.4890	0.0165	0.4835	0.0110	0.9725	0.9674	0.9780	0.9727
SRP RNA	0.4679	0.0513	0.4487	0.0321	0.9167	0.9012	0.9359	0.9182
TmRNA	0.3819	0.0039	0.4961	0.1181	0.8780	0.9898	0.7638	0.8622
tRNA	0.5000	0.0000	0.5000	0.0000	1.0000	1.0000	1.0000	1.0000
All	0.4517	0.0164	0.4836	0.0483	0.9353	0.9649	0.9035	0.9332

Table E.79: Greedy Voting Size 15 Prediction Statistics for Experiment 1. The progressive voting group contains the external loop, structure, stemloop, hairpin, tail, joint, joint-tail, bridge, stem, stack, multiloop, junction, unpaired, and internal loop, and loop.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4672	0.0146	0.4854	0.0328	0.9526	0.9697	0.9343	0.9517
23S rRNA	0.4583	0.0139	0.4861	0.0417	0.9444	0.9706	0.9167	0.9429
5S rRNA	0.4844	0.0000	0.5000	0.0156	0.9844	1.0000	0.9688	0.9841
RNase P	0.4890	0.0165	0.4835	0.0110	0.9725	0.9674	0.9780	0.9727
SRP RNA	0.4744	0.0513	0.4487	0.0256	0.9231	0.9024	0.9487	0.9250
TmRNA	0.3858	0.0079	0.4921	0.1142	0.8780	0.9800	0.7717	0.8634
tRNA	0.5000	0.0000	0.5000	0.0000	1.0000	1.0000	1.0000	1.0000
All	0.4537	0.0174	0.4826	0.0463	0.9363	0.9631	0.9073	0.9344

Table E.80: Greedy Voting Size 16 Prediction Statistics for Experiment 1. The progressive voting group contains the external loop, structure, stemloop, hairpin, tail, joint, joint-tail, bridge, stem, stack, multiloop, junction, unpaired, and internal loop, loop, and bulge.

E.3.2 Experiment 2

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.3276	0.1586	0.3414	0.1724	0.6690	0.6738	0.6552	0.6643
23S rRNA	0.2927	0.1707	0.3293	0.2073	0.6220	0.6316	0.5854	0.6076
5S rRNA	0.2419	0.0645	0.4355	0.2581	0.6774	0.7895	0.4839	0.6000
RNase P	0.3617	0.1543	0.3457	0.1383	0.7074	0.7010	0.7234	0.7120
SRP RNA	0.2405	0.1392	0.3608	0.2595	0.6013	0.6333	0.4810	0.5468
TmRNA	0.2966	0.2517	0.2483	0.2034	0.5448	0.5409	0.5931	0.5658
tRNA	0.0000	0.1250	0.3750	0.5000	0.3750	0.0000	0.0000	0.0000
All	0.3024	0.1753	0.3247	0.1976	0.6271	0.6330	0.6048	0.6186

Table E.81: Greedy Voting Size 1 Prediction Statistics for Experiment 2. The progressive voting group contains the structure.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4517	0.3207	0.1793	0.0483	0.6310	0.5848	0.9034	0.7100
23S rRNA	0.4268	0.3415	0.1585	0.0732	0.5854	0.5556	0.8537	0.6731
5S rRNA	0.4516	0.2581	0.2419	0.0484	0.6935	0.6364	0.9032	0.7467
RNase P	0.4628	0.3032	0.1968	0.0372	0.6596	0.6042	0.9255	0.7311
SRP RNA	0.4367	0.3354	0.1646	0.0633	0.6013	0.5656	0.8734	0.6866
TmRNA	0.4379	0.3690	0.1310	0.0621	0.5690	0.5427	0.8759	0.6702
tRNA	0.1250	0.3750	0.1250	0.3750	0.2500	0.2500	0.2500	0.2500
All	0.4434	0.3312	0.1688	0.0566	0.6122	0.5725	0.8868	0.6958

Table E.82: Greedy Voting Size 2 Prediction Statistics for Experiment 2. The progressive voting group contains the structure and stemloop.

Gene Type	**TP**	**FP**	**TN**	**FN**	**Acc.**	**Prec.**	**Recall**	**F-mea.**
16S rRNA	0.4103	0.1966	0.3034	0.0897	0.7138	0.6761	0.8207	0.7414
23S rRNA	0.4024	0.2683	0.2317	0.0976	0.6341	0.6000	0.8049	0.6875
5S rRNA	0.4032	0.1290	0.3710	0.0968	0.7742	0.7576	0.8065	0.7812
RNase P	0.4362	0.2234	0.2766	0.0638	0.7128	0.6613	0.8723	0.7523
SRP RNA	0.3671	0.1962	0.3038	0.1329	0.6709	0.6517	0.7342	0.6905
TmRNA	0.3759	0.2966	0.2034	0.1241	0.5793	0.5590	0.7517	0.6412
tRNA	0.3750	0.1250	0.3750	0.1250	0.7500	0.7500	0.7500	0.7500
All	0.3980	0.2291	0.2709	0.1020	0.6688	0.6346	0.7959	0.7062

Table E.83: Greedy Voting Size 3 Prediction Statistics for Experiment 2. The progressive voting group contains the structure, stemloop, and hairpin.

Gene Type	**TP**	**FP**	**TN**	**FN**	**Acc.**	**Prec.**	**Recall**	**F-mea.**
16S rRNA	0.4379	0.3103	0.1897	0.0621	0.6276	0.5853	0.8759	0.7017
23S rRNA	0.4024	0.3293	0.1707	0.0976	0.5732	0.5500	0.8049	0.6535
5S rRNA	0.4194	0.2581	0.2419	0.0806	0.6613	0.6190	0.8387	0.7123
RNase P	0.4628	0.2713	0.2287	0.0372	0.6915	0.6304	0.9255	0.7500
SRP RNA	0.3987	0.2722	0.2278	0.1013	0.6266	0.5943	0.7975	0.6811
TmRNA	0.4207	0.3448	0.1552	0.0793	0.5759	0.5495	0.8414	0.6649
tRNA	0.3750	0.2500	0.2500	0.1250	0.6250	0.6000	0.7500	0.6667
All	0.4276	0.3052	0.1948	0.0724	0.6224	0.5835	0.8553	0.6938

Table E.84: Greedy Voting Size 4 Prediction Statistics for Experiment 2. The progressive voting group contains the structure, stemloop, hairpin, and external loop.

Gene Type	**TP**	**FP**	**TN**	**FN**	**Acc.**	**Prec.**	**Recall**	**F-mea.**
16S rRNA	0.4517	0.3483	0.1517	0.0483	0.6034	0.5647	0.9034	0.6950
23S rRNA	0.4146	0.3537	0.1463	0.0854	0.5610	0.5397	0.8293	0.6538
5S rRNA	0.4516	0.2903	0.2097	0.0484	0.6613	0.6087	0.9032	0.7273
RNase P	0.4734	0.3085	0.1915	0.0266	0.6649	0.6054	0.9468	0.7386
SRP RNA	0.4177	0.3101	0.1899	0.0823	0.6076	0.5739	0.8354	0.6804
TmRNA	0.4310	0.3552	0.1448	0.0690	0.5759	0.5482	0.8621	0.6702
tRNA	0.5000	0.2500	0.2500	0.0000	0.7500	0.6667	1.0000	0.8000
All	0.4425	0.3340	0.1660	0.0575	0.6085	0.5699	0.8850	0.6933

Table E.85: Greedy Voting Size 5 Prediction Statistics for Experiment 2. The progressive voting group contains the structure, stemloop, hairpin, external loop, and joint-tail.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4103	0.2724	0.2276	0.0897	0.6379	0.6010	0.8207	0.6939
23S rRNA	0.3537	0.2927	0.2073	0.1463	0.5610	0.5472	0.7073	0.6170
5S rRNA	0.4355	0.2097	0.2903	0.0645	0.7258	0.6750	0.8710	0.7606
RNase P	0.4521	0.2394	0.2606	0.0479	0.7128	0.6538	0.9043	0.7589
SRP RNA	0.3608	0.2468	0.2532	0.1392	0.6139	0.5938	0.7215	0.6514
TmRNA	0.4069	0.3241	0.1759	0.0931	0.5828	0.5566	0.8138	0.6611
tRNA	0.3750	0.2500	0.2500	0.1250	0.6250	0.6000	0.7500	0.6667
All	0.4063	0.2746	0.2254	0.0937	0.6317	0.5967	0.8126	0.6881

Table E.86: Greedy Voting Size 6 Prediction Statistics for Experiment 2. The progressive voting group contains the structure, stemloop, hairpin, external loop, joint-tail, and joint.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.3759	0.1966	0.3034	0.1241	0.6793	0.6566	0.7517	0.7010
23S rRNA	0.3537	0.2195	0.2805	0.1463	0.6341	0.6170	0.7073	0.6591
5S rRNA	0.4032	0.1290	0.3710	0.0968	0.7742	0.7576	0.8065	0.7812
RNase P	0.4362	0.2021	0.2979	0.0638	0.7340	0.6833	0.8723	0.7664
SRP RNA	0.3228	0.1962	0.3038	0.1772	0.6266	0.6220	0.6456	0.6335
TmRNA	0.3724	0.2966	0.2034	0.1276	0.5759	0.5567	0.7448	0.6372
tRNA	0.3750	0.1250	0.3750	0.1250	0.7500	0.7500	0.7500	0.7500
All	0.3776	0.2217	0.2783	0.1224	0.6558	0.6300	0.7551	0.6869

Table E.87: Greedy Voting Size 7 Prediction Statistics for Experiment 2. The progressive voting group contains the structure, stemloop, hairpin, external loop, joint-tail, joint, and stack.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.3655	0.1828	0.3172	0.1345	0.6828	0.6667	0.7310	0.6974
23S rRNA	0.3293	0.2073	0.2927	0.1707	0.6220	0.6136	0.6585	0.6353
5S rRNA	0.4032	0.1613	0.3387	0.0968	0.7419	0.7143	0.8065	0.7576
RNase P	0.4202	0.1915	0.3085	0.0798	0.7287	0.6870	0.8404	0.7560
SRP RNA	0.3291	0.1709	0.3291	0.1709	0.6582	0.6582	0.6582	0.6582
TmRNA	0.3828	0.2793	0.2207	0.1172	0.6034	0.5781	0.7655	0.6588
tRNA	0.3750	0.0000	0.5000	0.1250	0.8750	1.0000	0.7500	0.8571
All	0.3738	0.2078	0.2922	0.1262	0.6660	0.6427	0.7477	0.6913

Table E.88: Greedy Voting Size 8 Prediction Statistics for Experiment 2. The progressive voting group contains the structure, stemloop, hairpin, external loop, joint-tail, joint, stack, and loop.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.3862	0.2414	0.2586	0.1138	0.6448	0.6154	0.7724	0.6850
23S rRNA	0.3415	0.2683	0.2317	0.1585	0.5732	0.5600	0.6829	0.6154
5S rRNA	0.4355	0.1613	0.3387	0.0645	0.7742	0.7297	0.8710	0.7941
RNase P	0.4574	0.2074	0.2926	0.0426	0.7500	0.6880	0.9149	0.7854
SRP RNA	0.3354	0.2152	0.2848	0.1646	0.6203	0.6092	0.6709	0.6386
TmRNA	0.4069	0.3172	0.1828	0.0931	0.5897	0.5619	0.8138	0.6648
tRNA	0.3750	0.1250	0.3750	0.1250	0.7500	0.7500	0.7500	0.7500
All	0.3961	0.2486	0.2514	0.1039	0.6475	0.6144	0.7922	0.6921

Table E.89: Greedy Voting Size 9 Prediction Statistics for Experiment 2. The progressive voting group contains the structure, stemloop, hairpin, external loop, joint-tail, joint, stack, loop, and unpaired.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4034	0.2862	0.2138	0.0966	0.6172	0.5850	0.8069	0.6783
23S rRNA	0.3780	0.3049	0.1951	0.1220	0.5732	0.5536	0.7561	0.6392
5S rRNA	0.4355	0.1935	0.3065	0.0645	0.7419	0.6923	0.8710	0.7714
RNase P	0.4681	0.2340	0.2660	0.0319	0.7340	0.6667	0.9362	0.7788
SRP RNA	0.3734	0.2468	0.2532	0.1266	0.6266	0.6020	0.7468	0.6667
TmRNA	0.4172	0.3310	0.1690	0.0828	0.5862	0.5576	0.8345	0.6685
tRNA	0.3750	0.1250	0.3750	0.1250	0.7500	0.7500	0.7500	0.7500
All	0.4137	0.2783	0.2217	0.0863	0.6354	0.5979	0.8275	0.6942

Table E.90: Greedy Voting Size 10 Prediction Statistics for Experiment 2. The progressive voting group contains the structure, stemloop, hairpin, external loop, joint-tail, joint, stack, loop, unpaired, and stem.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.3897	0.2241	0.2759	0.1103	0.6655	0.6348	0.7793	0.6997
23S rRNA	0.3415	0.2439	0.2561	0.1585	0.5976	0.5833	0.6829	0.6292
5S rRNA	0.4032	0.1774	0.3226	0.0968	0.7258	0.6944	0.8065	0.7463
RNase P	0.4415	0.1968	0.3032	0.0585	0.7447	0.6917	0.8830	0.7757
SRP RNA	0.3608	0.2025	0.2975	0.1392	0.6582	0.6404	0.7215	0.6786
TmRNA	0.3897	0.2931	0.2069	0.1103	0.5966	0.5707	0.7793	0.6589
tRNA	0.3750	0.2500	0.2500	0.1250	0.6250	0.6000	0.7500	0.6667
All	0.3915	0.2338	0.2662	0.1085	0.6577	0.6261	0.7829	0.6958

Table E.91: Greedy Voting Size 11 Prediction Statistics for Experiment 2. The progressive voting group contains the structure, stemloop, hairpin, external loop, joint-tail, joint, stack, loop, unpaired, stem, and internal loop.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.3931	0.2172	0.2828	0.1069	0.6759	0.6441	0.7862	0.7081
23S rRNA	0.3293	0.2561	0.2439	0.1707	0.5732	0.5625	0.6585	0.6067
5S rRNA	0.4194	0.1935	0.3065	0.0806	0.7258	0.6842	0.8387	0.7536
RNase P	0.4415	0.2021	0.2979	0.0585	0.7394	0.6860	0.8830	0.7721
SRP RNA	0.3608	0.1962	0.3038	0.1392	0.6646	0.6477	0.7215	0.6826
TmRNA	0.3862	0.2862	0.2138	0.1138	0.6000	0.5744	0.7724	0.6588
tRNA	0.3750	0.2500	0.2500	0.1250	0.6250	0.6000	0.7500	0.6667
All	0.3915	0.2319	0.2681	0.1085	0.6596	0.6280	0.7829	0.6969

Table E.92: Greedy Voting Size 12 Prediction Statistics for Experiment 2. The progressive voting group contains the structure, stemloop, hairpin, external loop, joint-tail, joint, stack, loop, unpaired, stem, internal loop, and bridge.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4034	0.2655	0.2345	0.0966	0.6379	0.6031	0.8069	0.6903
23S rRNA	0.3415	0.2927	0.2073	0.1585	0.5488	0.5385	0.6829	0.6022
5S rRNA	0.4355	0.1935	0.3065	0.0645	0.7419	0.6923	0.8710	0.7714
RNase P	0.4681	0.2234	0.2766	0.0319	0.7447	0.6769	0.9362	0.7857
SRP RNA	0.3797	0.2468	0.2532	0.1203	0.6329	0.6061	0.7595	0.6742
TmRNA	0.4103	0.3207	0.1793	0.0897	0.5897	0.5613	0.8207	0.6667
tRNA	0.3750	0.2500	0.2500	0.1250	0.6250	0.6000	0.7500	0.6667
All	0.4100	0.2681	0.2319	0.0900	0.6419	0.6047	0.8200	0.6961

Table E.93: Greedy Voting Size 13 Prediction Statistics for Experiment 2. The progressive voting group contains the structure, stemloop, hairpin, external loop, joint-tail, joint, stack, loop, unpaired, stem, internal loop, bridge, and bulge.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4034	0.2655	0.2345	0.0966	0.6379	0.6031	0.8069	0.6903
23S rRNA	0.3537	0.2927	0.2073	0.1463	0.5610	0.5472	0.7073	0.6170
5S rRNA	0.4677	0.1935	0.3065	0.0323	0.7742	0.7073	0.9355	0.8056
RNase P	0.4681	0.2234	0.2766	0.0319	0.7447	0.6769	0.9362	0.7857
SRP RNA	0.3797	0.2468	0.2532	0.1203	0.6329	0.6061	0.7595	0.6742
TmRNA	0.4103	0.3241	0.1759	0.0897	0.5862	0.5587	0.8207	0.6648
tRNA	0.3750	0.2500	0.2500	0.1250	0.6250	0.6000	0.7500	0.6667
All	0.4128	0.2690	0.2310	0.0872	0.6438	0.6054	0.8256	0.6986

Table E.94: Greedy Voting Size 14 Prediction Statistics for Experiment 2. The progressive voting group contains the structure, stemloop, hairpin, external loop, joint-tail, joint, stack, loop, unpaired, stem, internal loop, bridge, bulge, and multiloop.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16S rRNA	0.4000	0.2483	0.2517	0.1000	0.6517	0.6170	0.8000	0.6967
23S rRNA	0.3415	0.2683	0.2317	0.1585	0.5732	0.5600	0.6829	0.6154
5S rRNA	0.4516	0.1935	0.3065	0.0484	0.7581	0.7000	0.9032	0.7887
RNase P	0.4521	0.2128	0.2872	0.0479	0.7394	0.6800	0.9043	0.7763
SRP RNA	0.3797	0.2215	0.2785	0.1203	0.6582	0.6316	0.7595	0.6897
TmRNA	0.4034	0.3034	0.1966	0.0966	0.6000	0.5707	0.8069	0.6686
tRNA	0.3750	0.2500	0.2500	0.1250	0.6250	0.6000	0.7500	0.6667
All	0.4054	0.2514	0.2486	0.0946	0.6540	0.6172	0.8108	0.7009

Table E.95: Greedy Voting Size 15 Prediction Statistics for Experiment 2. The progressive voting group contains the structure, stemloop, hairpin, external loop, joint-tail, joint, stack, loop, unpaired, stem, internal loop, bridge, bulge, multiloop, and junction.

Gene Type	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
16SrRNA	0.4069	0.2828	0.2172	0.0931	0.6241	0.5900	0.8138	0.6841
23S rRNA	0.3659	0.3049	0.1951	0.1341	0.5610	0.5455	0.7317	0.6250
5S rRNA	0.4677	0.1935	0.3065	0.0323	0.7742	0.7073	0.9355	0.8056
RNase P	0.4681	0.2287	0.2713	0.0319	0.7394	0.6718	0.9362	0.7822
SRP RNA	0.3924	0.2532	0.2468	0.1076	0.6392	0.6078	0.7848	0.6851
TmRNA	0.4138	0.3345	0.1655	0.0862	0.5793	0.5530	0.8276	0.6630
tRNA	0.3750	0.2500	0.2500	0.1250	0.6250	0.6000	0.7500	0.6667
All	0.4174	0.2792	0.2208	0.0826	0.6382	0.5992	0.8349	0.6977

Table E.96: Greedy Voting Size 16 Prediction Statistics for Experiment 2. The progressive voting group contains the structure, stemloop, hairpin, external loop, joint-tail, joint, stack, loop, unpaired, stem, internal loop, bridge, bulge, multiloop, junction, tail.

E.4 Structural Element Paired Voting

Elements	TP	FP	TN	FN	Acc.	Prec.	Recall	F-mea.
Hairpin & Junction	0.4604	0.0589	0.4411	0.0396	0.9015	0.8866	0.9208	0.9034
Junction & Structure	0.4585	0.0598	0.4402	0.0415	0.8986	0.8845	0.9170	0.9005
External Loop & Hairpin	0.4575	0.0647	0.4353	0.0425	0.8929	0.8762	0.9151	0.8952
Multiloop & Structure	0.4440	0.0512	0.4488	0.0560	0.8929	0.8967	0.8880	0.8923
External Loop & Structure	0.4672	0.0820	0.4180	0.0328	0.8851	0.8506	0.9344	0.8905
Bridge & Structure	0.4508	0.0724	0.4276	0.0492	0.8784	0.8616	0.9015	0.8811
Junction & Stemloop	0.4459	0.0695	0.4305	0.0541	0.8764	0.8652	0.8919	0.8783
External Loop & Multiloop	0.4459	0.0705	0.4295	0.0541	0.8755	0.8636	0.8919	0.8775
External Loop & Stemloop	0.4846	0.1226	0.3774	0.0154	0.8620	0.7981	0.9691	0.8753
Hairpin & Structure	0.4431	0.0695	0.4305	0.0569	0.8736	0.8644	0.8861	0.8751
Bridge & Hairpin	0.4344	0.0589	0.4411	0.0656	0.8755	0.8806	0.8687	0.8746
External Loop & Junction	0.4469	0.0801	0.4199	0.0531	0.8668	0.8480	0.8938	0.8703
Hairpin & Multiloop	0.4208	0.0463	0.4537	0.0792	0.8745	0.9008	0.8417	0.8703
Stemloop & Structure	0.4266	0.0550	0.4450	0.0734	0.8716	0.8858	0.8533	0.8692
External Loop & Internal Loop	0.4556	0.1014	0.3986	0.0444	0.8542	0.8180	0.9112	0.8621
Hairpin & Stemloop	0.4546	0.1004	0.3996	0.0454	0.8542	0.8191	0.9093	0.8618
Stem & Structure	0.4440	0.0869	0.4131	0.0560	0.8571	0.8364	0.8880	0.8614
Multiloop & Junction	0.4112	0.0454	0.4546	0.0888	0.8658	0.9006	0.8224	0.8597
Junction & Stem	0.4112	0.0463	0.4537	0.0888	0.8649	0.8987	0.8224	0.8589

Structure & Tail	0.4479	0.0965	0.4035	0.0521	0.8514	0.8227	0.8958	0.8577
Internal Loop & Junction	0.4160	0.0550	0.4450	0.0840	0.8610	0.8832	0.8320	0.8569
Bridge & Stemloop	0.4421	0.0917	0.4083	0.0579	0.8504	0.8282	0.8842	0.8553
Hairpin & Stem	0.4102	0.0502	0.4498	0.0898	0.8600	0.8910	0.8205	0.8543
Hairpin & Tail	0.4151	0.0569	0.4431	0.0849	0.8581	0.8793	0.8301	0.8540
External Loop & Unpaired	0.4324	0.0820	0.4180	0.0676	0.8504	0.8405	0.8649	0.8525
External Loop & Stem	0.4257	0.0772	0.4228	0.0743	0.8485	0.8464	0.8514	0.8489
Multiloop & Stemloop	0.4344	0.0907	0.4093	0.0656	0.8436	0.8272	0.8687	0.8475
Bridge & External Loop	0.4363	0.0936	0.4064	0.0637	0.8427	0.8233	0.8726	0.8472
Hairpin & Internal Loop	0.4218	0.0743	0.4257	0.0782	0.8475	0.8502	0.8436	0.8469
Hairpin & Stack	0.4180	0.0695	0.4305	0.0820	0.8485	0.8574	0.8359	0.8465
Junction & Unpaired	0.3996	0.0454	0.4546	0.1004	0.8542	0.8980	0.7992	0.8458
Hairpin & Loop	0.4112	0.0618	0.4382	0.0888	0.8494	0.8694	0.8224	0.8452
Junction & Stack	0.4054	0.0560	0.4440	0.0946	0.8494	0.8787	0.8108	0.8434
Internal Loop & Structure	0.4112	0.0647	0.4353	0.0888	0.8465	0.8641	0.8224	0.8427
External Loop & Loop	0.4064	0.0589	0.4411	0.0936	0.8475	0.8734	0.8127	0.8420
External Loop & Stack	0.3996	0.0512	0.4488	0.1004	0.8485	0.8865	0.7992	0.8406
External Loop & Joint	0.4006	0.0541	0.4459	0.0994	0.8465	0.8811	0.8012	0.8392
External Loop & Joint-Tail	0.4006	0.0541	0.4459	0.0994	0.8465	0.8811	0.8012	0.8392
Bulge & External Loop	0.4006	0.0541	0.4459	0.0994	0.8465	0.8811	0.8012	0.8392
External Loop & Tail	0.4006	0.0541	0.4459	0.0994	0.8465	0.8811	0.8012	0.8392

Hairpin & Unpaired	0.4025	0.0569	0.4431	0.0975	0.8456	0.8761	0.8050	0.8390
Joint-Tail & Hairpin	0.3986	0.0531	0.4469	0.1014	0.8456	0.8825	0.7973	0.8377
Bridge & Junction	0.4102	0.0724	0.4276	0.0898	0.8378	0.8500	0.8205	0.8350
Loop & Structure	0.3919	0.0473	0.4527	0.1081	0.8446	0.8923	0.7838	0.8345
Loop & Junction	0.4025	0.0627	0.4373	0.0975	0.8398	0.8651	0.8050	0.8340
Structure & Unpaired	0.4189	0.0859	0.4141	0.0811	0.8330	0.8298	0.8378	0.8338
Internal Loop & Stemloop	0.4266	0.0975	0.4025	0.0734	0.8292	0.8140	0.8533	0.8332
Stack & Structure	0.3793	0.0338	0.4662	0.1207	0.8456	0.9182	0.7587	0.8309
Stem & Stemloop	0.4054	0.0705	0.4295	0.0946	0.8349	0.8519	0.8108	0.8309
Joint & Structure	0.3938	0.0560	0.4440	0.1062	0.8378	0.8755	0.7876	0.8293
Joint-Tail & Structure	0.3764	0.0319	0.4681	0.1236	0.8446	0.9220	0.7529	0.8289
Bulge & Structure	0.3764	0.0319	0.4681	0.1236	0.8446	0.9220	0.7529	0.8289
Joint-Tail & Junction	0.4035	0.0705	0.4295	0.0965	0.8330	0.8513	0.8069	0.8285
Loop & Stemloop	0.4160	0.0888	0.4112	0.0840	0.8272	0.8241	0.8320	0.8280
Joint & Stemloop	0.4122	0.0840	0.4160	0.0878	0.8282	0.8307	0.8243	0.8275
Stemloop & Tail	0.4421	0.1264	0.3736	0.0579	0.8156	0.7776	0.8842	0.8275
Hairpin & Joint	0.4122	0.0849	0.4151	0.0878	0.8272	0.8291	0.8243	0.8267
Bulge & Hairpin	0.4025	0.0714	0.4286	0.0975	0.8311	0.8493	0.8050	0.8266
Joint-Tail & Stemloop	0.4276	0.1129	0.3871	0.0724	0.8147	0.7911	0.8552	0.8219
Stack & Stemloop	0.4160	0.0965	0.4035	0.0840	0.8195	0.8117	0.8320	0.8217
Joint & Junction	0.3996	0.0734	0.4266	0.1004	0.8263	0.8449	0.7992	0.8214
Stemloop & Unpaired	0.4131	0.0946	0.4054	0.0869	0.8185	0.8137	0.8263	0.8199
Bulge & Stemloop	0.4276	0.1197	0.3803	0.0724	0.8079	0.7813	0.8552	0.8166
Junction & Tail	0.4044	0.0888	0.4112	0.0956	0.8156	0.8200	0.8089	0.8144
Internal Loop & Multiloop	0.3880	0.0685	0.4315	0.1120	0.8195	0.8499	0.7761	0.8113
Bulge & Junction	0.4122	0.1071	0.3929	0.0878	0.8050	0.7937	0.8243	0.8087

Multiloop & Stem	0.3909	0.0888	0.4112	0.1091	0.8021	0.8149	0.7819	0.7980
Multiloop & Tail	0.4421	0.1660	0.3340	0.0579	0.7761	0.7270	0.8842	0.7979
Bridge & Internal Loop	0.4170	0.1313	0.3687	0.0830	0.7857	0.7606	0.8340	0.7956
Bridge & Multiloop	0.3793	0.0763	0.4237	0.1207	0.8031	0.8326	0.7587	0.7939
Bridge & Tail	0.4054	0.1236	0.3764	0.0946	0.7819	0.7664	0.8108	0.7880
Bridge & Stem	0.3755	0.0782	0.4218	0.1245	0.7973	0.8277	0.7510	0.7874
Internal Loop & Stem	0.3986	0.1178	0.3822	0.1014	0.7809	0.7720	0.7973	0.7844
Multiloop & Unpaired	0.3494	0.0415	0.4585	0.1506	0.8079	0.8938	0.6988	0.7844
Internal Loop & Tail	0.4469	0.1959	0.3041	0.0531	0.7510	0.6952	0.8938	0.7821
Multiloop & Stack	0.3716	0.0792	0.4208	0.1284	0.7925	0.8244	0.7432	0.7817
Loop & Multiloop	0.3880	0.1081	0.3919	0.1120	0.7799	0.7821	0.7761	0.7791
Bridge & Loop	0.3822	0.1014	0.3986	0.1178	0.7809	0.7904	0.7645	0.7772
Stem & Tail	0.4469	0.2046	0.2954	0.0531	0.7423	0.6859	0.8938	0.7762
loop & Stem	0.4064	0.1409	0.3591	0.0936	0.7654	0.7425	0.8127	0.7760
Stem & Unpaired	0.3716	0.0946	0.4054	0.1284	0.7770	0.7971	0.7432	0.7692
Stack & Stem	0.3871	0.1197	0.3803	0.1129	0.7674	0.7638	0.7741	0.7689
Tail & Unpaired	0.4440	0.2133	0.2867	0.0560	0.7307	0.6755	0.8880	0.7673
Loop & Tail	0.4469	0.2191	0.2809	0.0531	0.7278	0.6710	0.8938	0.7666
Joint-Tail & Stem	0.3929	0.1332	0.3668	0.1071	0.7597	0.7468	0.7857	0.7658
Bridge & Stack	0.3774	0.1120	0.3880	0.1226	0.7654	0.7712	0.7548	0.7629
Bridge & Unpaired	0.3900	0.1351	0.3649	0.1100	0.7548	0.7426	0.7799	0.7608
Bridge & Joint-Tail	0.3678	0.1004	0.3996	0.1322	0.7674	0.7856	0.7355	0.7597
Bridge & Joint	0.3649	0.1081	0.3919	0.1351	0.7568	0.7714	0.7297	0.7500
Joint-Tail & Multiloop	0.3465	0.0792	0.4208	0.1535	0.7674	0.8141	0.6931	0.7487
Joint & Stem	0.3591	0.1042	0.3958	0.1409	0.7548	0.7750	0.7181	0.7455
Stack & Tail	0.4517	0.2635	0.2365	0.0483	0.6882	0.6316	0.9035	0.7434
Bulge & Tail	0.4054	0.1892	0.3108	0.0946	0.7162	0.6818	0.8108	0.7407
Internal Loop & Stack	0.3716	0.1342	0.3658	0.1284	0.7375	0.7347	0.7432	0.7390
Bridge & Bulge	0.3764	0.1467	0.3533	0.1236	0.7297	0.7196	0.7529	0.7358

Loop & Unpaired	0.3842	0.1612	0.3388	0.1158	0.7230	0.7044	0.7683	0.7350
Bulge & Stem	0.3755	0.1486	0.3514	0.1245	0.7268	0.7164	0.7510	0.7333
Bulge & Multiloop	0.3639	0.1293	0.3707	0.1361	0.7346	0.7378	0.7278	0.7328
Joint-Tail & Tail	0.3542	0.1129	0.3871	0.1458	0.7413	0.7583	0.7085	0.7325
Joint & Tail	0.4102	0.2104	0.2896	0.0898	0.6998	0.6610	0.8205	0.7321
Joint & Multiloop	0.3282	0.0685	0.4315	0.1718	0.7597	0.8273	0.6564	0.7320
Internal Loop & Loop	0.3407	0.0927	0.4073	0.1593	0.7481	0.7862	0.6815	0.7301
Internal Loop & Unpaired	0.3552	0.1187	0.3813	0.1448	0.7365	0.7495	0.7104	0.7294
Loop & Stack	0.3629	0.1458	0.3542	0.1371	0.7172	0.7135	0.7259	0.7196
Stack & Unpaired	0.4025	0.2317	0.2683	0.0975	0.6708	0.6347	0.8050	0.7098
Joint-Tail & Loop	0.3716	0.1805	0.3195	0.1284	0.6911	0.6731	0.7432	0.7064
Joint-Tail & Internal Loop	0.3784	0.2037	0.2963	0.1216	0.6747	0.6501	0.7568	0.6994
Bulge & Internal Loop	0.3330	0.1236	0.3764	0.1670	0.7095	0.7294	0.6660	0.6963
Internal Loop & Joint	0.3542	0.1660	0.3340	0.1458	0.6882	0.6809	0.7085	0.6944
Joint-Tail & Stack	0.3803	0.2317	0.2683	0.1197	0.6486	0.6215	0.7606	0.6840
Joint-Tail & Unpaired	0.3716	0.2162	0.2838	0.1284	0.6554	0.6322	0.7432	0.6832
Bulge & Stack	0.4913	0.4662	0.0338	0.0087	0.5251	0.5131	0.9826	0.6742
Bulge & Unpaired	0.4836	0.4566	0.0434	0.0164	0.5270	0.5144	0.9672	0.6716
Joint & Stack	0.4981	0.4865	0.0135	0.0019	0.5116	0.5059	0.9961	0.6710
Bulge & Loop	0.4131	0.3214	0.1786	0.0869	0.5917	0.5624	0.8263	0.6693
Joint & Loop	0.3436	0.1844	0.3156	0.1564	0.6593	0.6508	0.6873	0.6685
Joint & Unpaired	0.4981	0.4971	0.0029	0.0019	0.5010	0.5005	0.9961	0.6662
Bulge & Joint-Tail	0.4817	0.4681	0.0319	0.0183	0.5135	0.5071	0.9633	0.6644
Joint-Tail & Joint	0.4218	0.3533	0.1467	0.0782	0.5685	0.5442	0.8436	0.6616
Bulge & Joint	0.4759	0.4778	0.0222	0.0241	0.4981	0.4990	0.9517	0.6547

Table E.97: Structural Element Voting Pair Statistics for Experiment 1. Lists every combination of structural element voting pairs, sorted by descending F-measure.

Elements	**TP**	**FP**	**TN**	**FN**	**Acc.**	**Prec.**	**Recall**	**F-mea.**
Hairpin & Structure	0.4119	0.2653	0.2347	0.0881	0.6466	0.6082	0.8237	0.6998
Stemloop & Structure	0.4573	0.3562	0.1438	0.0427	0.6011	0.5621	0.9147	0.6963
Hairpin & Stack	0.4202	0.3052	0.1948	0.0798	0.6150	0.5793	0.8404	0.6858
Stack & Stemloop	0.4917	0.4555	0.0445	0.0083	0.5362	0.5191	0.9833	0.6795
Joint-Tail & Hairpin	0.3952	0.2681	0.2319	0.1048	0.6271	0.5958	0.7904	0.6794
Hairpin & Stemloop	0.4267	0.3340	0.1660	0.0733	0.5928	0.5610	0.8534	0.6770
Loop & Structure	0.4276	0.3358	0.1642	0.0724	0.5918	0.5601	0.8553	0.6769
Internal Loop & Stemloop	0.4416	0.3636	0.1364	0.0584	0.5779	0.5484	0.8831	0.6766
Hairpin & Internal Loop	0.3766	0.2375	0.2625	0.1234	0.6391	0.6133	0.7532	0.6761
Hairpin & Unpaired	0.4276	0.3377	0.1623	0.0724	0.5900	0.5588	0.8553	0.6760
Stack & Structure	0.4731	0.4267	0.0733	0.0269	0.5464	0.5258	0.9462	0.6759
Stemloop & Unpaired	0.4583	0.3980	0.1020	0.0417	0.5603	0.5352	0.9165	0.6758
External Loop & Hairpin	0.4100	0.3043	0.1957	0.0900	0.6058	0.5740	0.8200	0.6753
Bulge & Stemloop	0.4212	0.3265	0.1735	0.0788	0.5946	0.5633	0.8423	0.6751
Loop & Stemloop	0.4712	0.4249	0.0751	0.0288	0.5464	0.5259	0.9425	0.6751
Hairpin & Loop	0.3599	0.2069	0.2931	0.1401	0.6531	0.6350	0.7199	0.6748
External Loop & Stemloop	0.4527	0.3942	0.1058	0.0473	0.5584	0.5345	0.9054	0.6722
Joint-Tail & Stemloop	0.4221	0.3340	0.1660	0.0779	0.5881	0.5583	0.8442	0.6721
Junction & Stemloop	0.4147	0.3219	0.1781	0.0853	0.5928	0.5630	0.8293	0.6707
Hairpin & Stem	0.3340	0.1623	0.3377	0.1660	0.6716	0.6729	0.6679	0.6704
Stemloop & Tail	0.4249	0.3432	0.1568	0.0751	0.5816	0.5531	0.8497	0.6701
Bridge & Hairpin	0.3386	0.1725	0.3275	0.1614	0.6660	0.6624	0.6772	0.6697

Joint & Stack	0.4917	0.4768	0.0232	0.0083	0.5148	0.5077	0.9833	0.6696
Stack & Unpaired	0.4981	0.4898	0.0102	0.0019	0.5083	0.5042	0.9963	0.6696
Stem & Stemloop	0.4035	0.3024	0.1976	0.0965	0.6011	0.5716	0.8071	0.6692
Joint-Tail & Stack	0.4981	0.4907	0.0093	0.0019	0.5074	0.5038	0.9963	0.6692
Joint & Stemloop	0.4230	0.3414	0.1586	0.0770	0.5816	0.5534	0.8460	0.6691
Multiloop & Stemloop	0.4100	0.3163	0.1837	0.0900	0.5937	0.5645	0.8200	0.6687
Bulge & Hairpin	0.3581	0.2134	0.2866	0.1419	0.6447	0.6266	0.7161	0.6684
Internal Loop & Structure	0.3813	0.2597	0.2403	0.1187	0.6215	0.5948	0.7625	0.6683
Stem & Structure	0.3905	0.2783	0.2217	0.1095	0.6122	0.5839	0.7811	0.6683
Bridge & Stack	0.4796	0.4573	0.0427	0.0204	0.5223	0.5119	0.9592	0.6675
Loop & Stack	0.4805	0.4592	0.0408	0.0195	0.5213	0.5114	0.9610	0.6675
Hairpin & Multiloop	0.3312	0.1614	0.3386	0.1688	0.6698	0.6723	0.6623	0.6673
Bridge & Stemloop	0.4109	0.3210	0.1790	0.0891	0.5900	0.5615	0.8219	0.6672
Joint-Tail & Structure	0.4647	0.4286	0.0714	0.0353	0.5362	0.5202	0.9295	0.6671
Joint-Tail & Unpaired	0.4972	0.4935	0.0065	0.0028	0.5037	0.5019	0.9944	0.6671
Internal Loop & Stack	0.4833	0.4666	0.0334	0.0167	0.5167	0.5088	0.9666	0.6667
External Loop & Stack	0.4889	0.4777	0.0223	0.0111	0.5111	0.5058	0.9777	0.6667
Hairpin & Junction	0.3321	0.1651	0.3349	0.1679	0.6670	0.6679	0.6642	0.6660
Structure & Unpaired	0.4954	0.4926	0.0074	0.0046	0.5028	0.5014	0.9907	0.6658
Stack & Stem	0.4833	0.4685	0.0315	0.0167	0.5148	0.5078	0.9666	0.6658
Joint-Tail & Loop	0.4852	0.4731	0.0269	0.0148	0.5121	0.5063	0.9703	0.6654
Joint & Unpaired	0.4954	0.4944	0.0056	0.0046	0.5009	0.5005	0.9907	0.6650
Stem & Unpaired	0.4944	0.4926	0.0074	0.0056	0.5019	0.5009	0.9889	0.6650
Stack & Tail	0.4842	0.4731	0.0269	0.0158	0.5111	0.5058	0.9685	0.6645
Bulge & Stack	0.4842	0.4759	0.0241	0.0158	0.5083	0.5043	0.9685	0.6633
Junction & Stack	0.4805	0.4685	0.0315	0.0195	0.5121	0.5064	0.9610	0.6633
Multiloop & Stack	0.4796	0.4666	0.0334	0.0204	0.5130	0.5069	0.9592	0.6632

Hairpin & Joint	0.3256	0.1568	0.3432	0.1744	0.6688	0.6750	0.6512	0.6629
Hairpin & Tail	0.3553	0.2171	0.2829	0.1447	0.6382	0.6207	0.7106	0.6626
Internal Loop & Unpaired	0.4917	0.4926	0.0074	0.0083	0.4991	0.4995	0.9833	0.6625
External Loop & Unpaired	0.4926	0.4954	0.0046	0.0074	0.4972	0.4986	0.9852	0.6621
Junction & Unpaired	0.4870	0.4842	0.0158	0.0130	0.5028	0.5014	0.9740	0.6620
Loop & Unpaired	0.4917	0.4954	0.0046	0.0083	0.4963	0.4981	0.9833	0.6613
Bulge & Unpaired	0.4889	0.4898	0.0102	0.0111	0.4991	0.4995	0.9777	0.6612
Multiloop & Unpaired	0.4889	0.4898	0.0102	0.0111	0.4991	0.4995	0.9777	0.6612
Bridge & Unpaired	0.4861	0.4842	0.0158	0.0139	0.5019	0.5010	0.9722	0.6612
Tail & Unpaired	0.4879	0.4917	0.0083	0.0121	0.4963	0.4981	0.9759	0.6596
External Loop & Structure	0.4026	0.3200	0.1800	0.0974	0.5826	0.5571	0.8052	0.6586
Joint-Tail & Internal Loop	0.4564	0.4304	0.0696	0.0436	0.5260	0.5146	0.9128	0.6582
Joint & Loop	0.4555	0.4304	0.0696	0.0445	0.5250	0.5141	0.9109	0.6573
External Loop & Joint-Tail	0.4564	0.4332	0.0668	0.0436	0.5232	0.5130	0.9128	0.6569
Loop & Stem	0.4397	0.4007	0.0993	0.0603	0.5390	0.5232	0.8794	0.6561
Joint-Tail & Stem	0.4406	0.4156	0.0844	0.0594	0.5250	0.5146	0.8813	0.6498
Junction & Structure	0.3293	0.1855	0.3145	0.1707	0.6438	0.6396	0.6586	0.6490
Joint-Tail & Junction	0.4314	0.3980	0.1020	0.0686	0.5334	0.5201	0.8627	0.6490
Joint-Tail & Joint	0.4360	0.4119	0.0881	0.0640	0.5241	0.5142	0.8720	0.6469
Bulge & Joint-Tail	0.4406	0.4249	0.0751	0.0594	0.5158	0.5091	0.8813	0.6454
Internal Loop & Joint	0.3989	0.3377	0.1623	0.1011	0.5612	0.5416	0.7978	0.6452
Joint & Structure	0.4045	0.3506	0.1494	0.0955	0.5538	0.5356	0.8089	0.6445
Multiloop & Structure	0.3210	0.1781	0.3219	0.1790	0.6429	0.6431	0.6419	0.6425
Joint-Tail & Tail	0.4249	0.3989	0.1011	0.0751	0.5260	0.5158	0.8497	0.6419

Bridge & Joint-Tail	0.4239	0.3970	0.1030	0.0761	0.5269	0.5164	0.8479	0.6419
External Loop & Joint	0.3998	0.3488	0.1512	0.1002	0.5510	0.5341	0.7996	0.6404
Joint-Tail & Multiloop	0.4212	0.3970	0.1030	0.0788	0.5241	0.5147	0.8423	0.6390
Bulge & Structure	0.3386	0.2254	0.2746	0.1614	0.6132	0.6003	0.6772	0.6364
External Loop & Loop	0.4082	0.3748	0.1252	0.0918	0.5334	0.5213	0.8163	0.6363
Bridge & Structure	0.3191	0.1874	0.3126	0.1809	0.6317	0.6300	0.6382	0.6341
Internal Loop & Stem	0.3534	0.2644	0.2356	0.1466	0.5891	0.5721	0.7069	0.6324
Structure & Tail	0.3367	0.2301	0.2699	0.1633	0.6067	0.5941	0.6735	0.6313
External Loop & Stem	0.3692	0.3015	0.1985	0.1308	0.5677	0.5505	0.7384	0.6307
Internal Loop & Loop	0.3692	0.3256	0.1744	0.1308	0.5436	0.5314	0.7384	0.6180
Bridge & Loop	0.3701	0.3367	0.1633	0.1299	0.5334	0.5236	0.7403	0.6134
External Loop & Internal Loop	0.3534	0.3006	0.1994	0.1466	0.5529	0.5404	0.7069	0.6125
Loop & Tail	0.3738	0.3488	0.1512	0.1262	0.5250	0.5173	0.7477	0.6115
Loop & Junction	0.3646	0.3367	0.1633	0.1354	0.5278	0.5198	0.7291	0.6069
Loop & Multiloop	0.3590	0.3293	0.1707	0.1410	0.5297	0.5216	0.7180	0.6042
Joint & Stem	0.3488	0.3126	0.1874	0.1512	0.5362	0.5273	0.6976	0.6006
Bulge & Joint	0.3479	0.3210	0.1790	0.1521	0.5269	0.5201	0.6957	0.5952
Bulge & Loop	0.3581	0.3460	0.1540	0.1419	0.5121	0.5086	0.7161	0.5948
Bulge & Stem	0.3080	0.2449	0.2551	0.1920	0.5631	0.5570	0.6160	0.5850
Joint & Junction	0.3135	0.2681	0.2319	0.1865	0.5455	0.5391	0.6271	0.5798
Joint & Tail	0.3219	0.3006	0.1994	0.1781	0.5213	0.5171	0.6438	0.5736
External Loop & Tail	0.3024	0.2597	0.2403	0.1976	0.5427	0.5380	0.6048	0.5694
Stem & Tail	0.2848	0.2171	0.2829	0.2152	0.5677	0.5675	0.5696	0.5685
Bulge & External Loop	0.3015	0.2681	0.2319	0.1985	0.5334	0.5293	0.6030	0.5637
Bridge & External Loop	0.2876	0.2328	0.2672	0.2124	0.5547	0.5526	0.5751	0.5636

External Loop & Multiloop	0.2783	0.2189	0.2811	0.2217	0.5594	0.5597	0.5566	0.5581
Joint & Multiloop	0.2950	0.2635	0.2365	0.2050	0.5315	0.5282	0.5900	0.5574
Bridge & Joint	0.2941	0.2672	0.2328	0.2059	0.5269	0.5240	0.5881	0.5542
Junction & Stem	0.2579	0.1735	0.3265	0.2421	0.5844	0.5978	0.5158	0.5538
External Loop & Junction	0.2746	0.2208	0.2792	0.2254	0.5538	0.5543	0.5492	0.5517
Internal Loop & Tail	0.2681	0.2041	0.2959	0.2319	0.5640	0.5678	0.5362	0.5515
Bulge & Internal Loop	0.2718	0.2301	0.2699	0.2282	0.5417	0.5416	0.5436	0.5426
Bridge & Stem	0.2495	0.1753	0.3247	0.2505	0.5742	0.5873	0.4991	0.5396
Multiloop & Stem	0.2458	0.1660	0.3340	0.2542	0.5798	0.5968	0.4917	0.5392
Bridge & Internal Loop	0.2449	0.1763	0.3237	0.2551	0.5686	0.5815	0.4898	0.5317
Internal Loop & Junction	0.2384	0.1735	0.3265	0.2616	0.5649	0.5788	0.4768	0.5229
Internal Loop & Multiloop	0.2273	0.1623	0.3377	0.2727	0.5649	0.5833	0.4545	0.5109
Bulge & Tail	0.1818	0.1688	0.3312	0.3182	0.5130	0.5185	0.3636	0.4275
Bulge & Junction	0.1642	0.1178	0.3822	0.3358	0.5464	0.5822	0.3284	0.4199
Junction & Tail	0.1512	0.1011	0.3989	0.3488	0.5501	0.5993	0.3024	0.4020
Bridge & Bulge	0.1558	0.1215	0.3785	0.3442	0.5343	0.5619	0.3117	0.4010
Bridge & Tail	0.1456	0.1011	0.3989	0.3544	0.5445	0.5902	0.2913	0.3901
Bulge & Multiloop	0.1401	0.1058	0.3942	0.3599	0.5343	0.5698	0.2801	0.3756
Multiloop & Tail	0.1271	0.0881	0.4119	0.3729	0.5390	0.5905	0.2542	0.3554
Bridge & Junction	0.1020	0.0399	0.4601	0.3980	0.5622	0.7190	0.2041	0.3179
Multiloop & Junction	0.0807	0.0232	0.4768	0.4193	0.5575	0.7768	0.1614	0.2673
Bridge & Multiloop	0.0705	0.0269	0.4731	0.4295	0.5436	0.7238	0.1410	0.2360

Table E.98: Structural Element Voting Pair Statistics for Experiment 2. Lists every combination of structural element voting pairs, sorted by descending F-measure.

Bibliography

[1] S. F. Altschul and B. W. Erickson. Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. *Mol Biol Evol*, 2(6):526–538, November 1985.

[2] Athanasius F Bompfünewerer. RNAs everywhere: genome-wide annotation of structured RNAs. *Journal of experimental zoology. Part B, Molecular and developmental evolution*, 308(1):1–25, January 2007.

[3] Tomas Babak, Benjamin J. Blencowe, and Timothy R. Hughes. Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC Bioinformatics*, 8:33+, January 2007.

[4] B. G. Barrell, A. T. Bankier, and J. Drouin. A different genetic code in human mitochondria. *Nature*, 282:189–194, 1979.

[5] Eugene Berezikov, Edwin Cuppen, and Ronald H. A. Plasterk. Approaches to microRNA discovery. *Nature Genetics*, 38 Suppl:S2–S7, May 2006.

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed. 2006. corr. 2nd printing edition, October 2007.

[7] Terry Brown. *Genomes 3*, chapter 1. Garland Science, third edition, May 2006.

[8] Christopher J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[9] Richard J. Carter, Inna Dubchak, and Stephen R. Holbrook. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucl. Acids Res.*, 29(19):3928–3938, October 2001.

[10] Chih-chung Chang and Chih-jen Lin. LIBSVM: a Library for Support Vector Machines, 2001.

[11] Chun-Long L. Chen, Hui Zhou, Jian-You Y. Liao, Liang-Hu H. Qu, and Laurence Amar. Genome-wide evolutionary analysis of the noncoding RNA genes and noncoding DNA of Paramecium tetraurelia. *RNA (New York, N.Y.)*, February 2009.

[12] Jih-H Chen, Shu-Yun Le, Bruce Shapiro, Kathleen M. Currey, and Jacob V. Maizel. A computational procedure for assessing the significance of RNA secondary structure. *Comput. Appl. Biosci.*, 6(1):7–18, January 1990.

[13] Y. W. Chen and C. J. Lin. *Combining SVMs with various feature selection strategies.* 2005.

[14] Peter Clote, Fabrizio Ferré, Evangelos Kranakis, and Danny Krizanc. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, 11(5):578–591, May 2005.

[15] Rita R. Colwell, David G. Swartz, and Michael T. MacDonell, editors. *Biomolecular data: a resource in transition*, pages 67+. Oxford science publications. Oxford University Press, Oxford, England, 1989.

[16] Rita R. Colwell, David G. Swartz, and Michael T. Macdonell, editors. *Biomolecular Data: A Resource in Transition*, pages 67+. Oxford University Press, June 1989.

[17] Sean R. Eddy. Computational genomics of noncoding RNA genes. *Cell*, 109(2):137–140, April 2002.

[18] Nicholas Erho and Kay Wiese. An Exploration of Individual RNA Structural Elements in RNA Gene Finding. In *Computational Intelligence in Bioinformatics and Computational Biology 2010*, pages 203–211. IEEE Computational Intelligence Society, May 2010.

[19] Paolo Frasconi and Ron Shamir, editors. *Artificial intelligence and heuristic methods in bioinformatics*, volume 138 of *NATO science series. Series III, Computer and systems sciences.* IOS Press, Washington, DC, 2003.

[20] N. Galtier and J. R. Lobry. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *Journal of molecular evolution*, 44(6):632–636, June 1997.

[21] T. R. Gingeras and R. J. Roberts. Steps toward Computer Analysis of Nucleotide Sequences. *Science*, 209:1322–1328, September 1980.

[22] Jiawei Han and Micheline Kamber. *Data mining: concepts and techniques.* The Morgan Kaufmann series in data management systems. Morgan Kaufmann Publishers, second edition, 2006.

[23] Shunmin He, Changning Liu, Geir Skogerbo, Haitao Zhao, Jie Wang, Tao Liu, Baoyan Bai, Yi Zhao, and Runsheng Chen. NONCODE v2.0: decoding the non-coding. *Nucl. Acids Res.*, pages gkm1011+, November 2007.

[24] D. K. Hendrix, S. E. Brenner, and S. R. Holbrook. RNA structural motifs: building blocks of a modular biomolecule. *Q Rev Biophys*, 38(3):221–243, August 2005.

[25] I. L. Hofacker, B. Priwitzer, and P. F. Stadler. Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, 20(2):186–190, January 2004.

[26] Ivo L. Hofacker, Walter Fontana, Peter F. Stadler, Sebastian L. Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast Folding and Comparison of RNA Secondary Structures. *Monatsh. Chem.*, 125:167–188, 1994.

[27] M. E. Hudson. Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, 8(1):3–17, January 2008.

[28] Robert J. Klein, Ziva Misulovin, and Sean R. Eddy. Noncoding RNA genes identified in AT-rich hyperthermophiles. *PNAS*, 99(11):7542–7547+, May 2002.

[29] S. Y. Le, J. H. Chen, and J. V. Maizel. Thermodynamic stability and statistical significance of potential stem-loop structures situated at the frameshift sites of retroviruses. *Nucleic Acids Res*, 17(15):6143–6152, August 1989.

[30] Shu-Yun Le, Jih-H Chen, Michael J. Braun, Matthew A. Gonda, and Jacob V. Maizel. Stability of RNA stem-loop structure and distribution of non-random structure in the human immunodeficiency virus (HIV-I). *Nucl. Acids Res.*, 16(11):5153–5168, June 1988.

[31] Shu-Yun Le, Jih-Hsiang Chen, Kathleen M. Currey, and Jacob V. Maizel. A program for predicting significant RNA secondary structures. *Comput. Appl. Biosci.*, 4(1):153–159, March 1988.

[32] Shuyun Le, Jih-Hsiang Chen, Ruth Nussinov, and Jacob V. Maizel. An improved secondary structure computation method and its application to intervening sequences in the human alpha-like globin mRNA precursors. *Comput. Appl. Biosci.*, 4(3):337–344, August 1988.

[33] N. B. Leontis, A. Lescoute, and E. Westhof. The building blocks and motifs of RNA architecture. *Curr Opin Struct Biol*, 16(3):279–287, June 2006.

[34] Arthur M. Lesk, editor. *Computational Molecular Biology: Sources and Methods for Sequence Analysis*. Oxford University Press, February 1989.

[35] Xiaoou Li and Kang Li. *Detecting RNA Sequences Using Two-Stage SVM Classifier*, volume 4689 of *Lecture Notes in Computer Science*. Springer Berlin, Heidelberg, Germany, 2007.

[36] T. M. Lowe and S. R. Eddy. A computational screen for methylation guide snoRNAs in yeast. *Science*, 283(5405):1168–1171, February 1999.

[37] J. S. Mccaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, May 1990.

[38] Richard F. Meraz and Stephen R. Holbrook. Classification of Non-coding RNA Using Graph Representations of Secondary Structure. *Pac Symp Biocomput.*, pages 4–15, 2004.

[39] Irmtraud M. Meyer. A practical guide to the art of RNA gene prediction. *Briefings in Bioinformatics*, 8(6):396–414, November 2007.

[40] Naila Mimouni. NcRNAs: Lost in Translation. *IBS Journal of Science*, 2(2):27–34, 2007.

[41] K. Missal, X. Zhu, D. Rose, W. Deng, G. Skogerbø, R. Chen, and P. F. Stadler. Prediction of structured non-coding RNAs in the genomes of the nematodes Caenorhabditis elegans and Caenorhabditis briggsae. *J Exp Zoolog B Mol Dev Evol*, 306(4):379–392, July 2006.

[42] Tobias Mourier, Celine Carret, Sue Kyes, Zoe Christodoulou, Paul P. Gardner, Daniel C. Jeffares, Robert Pinches, Bart Barrell, Matt Berriman, Sam Griffiths-Jones, Alasdair Ivens, Chris Newbold, and Arnab Pain. Genome-wide discovery and verification of novel structured RNAs in Plasmodium falciparum. *Genome research*, 18(2):281–292, February 2008.

[43] William S. Noble. What is a support vector machine? *Nature Biotechnology*, 24(12):1565–1567, December 2006.

[44] Kirt Noël. Examining Stem-Loops as a Sequence Signal for Identifying Structural RNA Genes. Master's thesis, Simon Fraser University, Canada, 2005.

[45] Kirt Noël and Kay C. Wiese. Exploring the Use of Stem-Loop Characteristics for Pinpointing Structural RNA Genes. *Computational Systems Bioinformatics Conference, International IEEE Computer Society*, 0:533–534, 2004.

[46] Kirt Noel and Kay C. Wiese. *Considering Stem-loops as Sequence Signals for Finding Structural RNA Genes*, chapter 14, pages 337–358. Springer Studies on Computational Intelligence. 2008.

[47] Jakob S. Pedersen, Gill Bejerano, Adam Siepel, Kate Rosenbloom, Kerstin Lindblad-Toh, Eric S. Lander, Jim Kent, Webb Miller, and David Haussler. Identification and Classification of Conserved RNA Secondary Structures in the Human Genome. *PLoS Comput Biol*, 2(4):e33+, April 2006.

[48] Thomas D. Pollard and William C. Earnshaw. *Cell Biology.* Saunders, August 2008.

[49] E. Rivas and S. R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of molecular biology*, 285(5):2053–2068, February 1999.

[50] E. Rivas and S. R. Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics (Oxford, England)*, 16(7):583–605, July 2000.

[51] E. Rivas and S. R. Eddy. Noncoding RNA gene detection using comparative sequence analysis. *BMC bioinformatics*, 2(1):8+, 2001.

[52] E. Rivas, R. J. Klein, T. A. Jones, and S. R. Eddy. Computational identification of noncoding RNAs in E. coli by comparative genomics. *Current biology : CB*, 11(17):1369–1373, September 2001.

[53] M. Rychetsky. *Algorithms and Architectures for Machine Learning Based on Regularized Neural Networks and Support Vector Approaches*. Shaker Verlag GmbH, Germany, December 2001.

[54] Peter Schattner. Searching for RNA genes using base-composition statistics. *Nucleic Acids Research*, 30(9):2076–2082, 2002.

[55] Susan J. Schroeder, Mark E. Burkard, and Douglas H. Turner. The Energetics of Small Internal Loops in RNA. *Biopolymers (Nucleic Acid Sciences)*, 52:157–167, 1999.

[56] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: a case study in RNA secondary structures. *Proc R Soc Lond B Biol Sci*, 255(1344):279–284, March 1994.

[57] Jennifer A. Smith. Computational Intelligence Method to Find Generic Non-coding RNA Search Models. In *CIBCB 2010 Proceedings*, pages 198–202. IEEE Computational Intelligence Society, May 2010.

[58] R. Staden. A Computer Program to Search for tRNA Genes. *Nucleic Acids Research*, 8(4):817–826, 1980.

[59] Tai Te Wu, editor. *Analytical molecular biology*. Kluwer Academic Publishers, 2001.

[60] Olga G. Troyanskaya, Ora Arbell, Yair Koren, Gad M. Landau, and Alexander Bolshoy. Sequence complexity profiles of prokaryotic genomic sequences: A fast algorithm for calculating linguistic complexity. *Bioinformatics*, 18(5):679–688, May 2002.

[61] Andrew V. Uzilov, Joshua M. Keegan, and David H. Mathews. Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics*, 7(1), March 2006.

[62] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[63] Wang. *Support Vector Machines: Theory and Applications*, volume 177 of *Studies in Fuzziness and Soft Computing*. Springer Berlin, Heidelberg, Germany, 2005.

[64] C. Wang, C. Ding, R. F. Meraz, and S. R. Holbrook. PSoL: a positive sample only learning algorithm for finding non-coding RNA genes. *Bioinformatics*, 22(21):2590–2596, November 2006.

[65] Jason T. L. Wang, Dongrong Wen, and Jianghui Liu. *On comparing and visualizing RNA secondary structures*. Wiley-Interscience, Hoboken, N.J, 2007.

[66] S. Washietl, I. L. Hofacker, and P. F. Stadler. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A*, 102(7):2454–2459, February 2005.

[67] Stefan Washietl, Jakob S. Pedersen, Jan O. Korbel, Claudia Stocsits, Andreas R. Gruber, Jorg Hackermuller, Jana Hertel, Manja Lindemeyer, Kristin Reiche, Andrea Tanzer, Catherine Ucla, Carine Wyss, Stylianos E. Antonarakis, France Denoeud, Julien Lagarde, Jorg Drenkow, Philipp Kapranov, Thomas R. Gingeras, Roderic Guigo, Michael Snyder, Mark B. Gerstein, Alexandre Reymond, Ivo L. Hofacker, and Peter F. Stadler. Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res.*, 17(6):852–864, June 2007.

[68] Stefan M. Washietl. *Prediction of Structural Non-coding RNAs by Comparative Sequence Analysis*. PhD thesis, Universität Wien, 2005.

[69] Zasha Weinberg and Walter L. Ruzzo. Exploiting conserved structure for faster annotation of non-coding RNAs without loss of accuracy. *Bioinformatics (Oxford, England)*, 20 Suppl 1:334–341, August 2004.

[70] Eric Westhof and Pascal Auffinger. *RNA Tertiary Structure*, pages 5222–5232. John Wiley & Sons Ltd., Chichester, 2000.

[71] C. Workman and A. Krogh. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res*, 27(24):4816–4822, December 1999.

[72] Yair and Ron Unger. RNAslider: a faster engine for consecutive windows folding and its application to the analysis of genomic folding asymmetry. *BMC Bioinformatics*, 10:76+, March 2009.

[73] Peter Yakovchuk, Ekaterina Protozanova, and Maxim D. Frank-Kamenetskii. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Research*, 34(2):564–574, 2006.

[74] Z. Yao, J. Barrick, Z. Weinberg, S. Neph, R. Breaker, M. Tompa, and W. L. Ruzzo. A computational pipeline for high- throughput discovery of cis-regulatory noncoding RNA in prokaryotes. *PLoS computational biology*, 3(7):e126+, July 2007.

[75] S. Zhang, B. Haas, E. Eskin, and V. Bafna. Searching genomes for noncoding RNA using FastR. *IEEE/ACM Trans Comput Biol Bioinform*, 2(4):366–379, 2005.

[76] Jizhen Zhao, Russell L. Malmberg, and Liming Cai. *Rapid ab initio RNA Folding Including Pseudoknots Via Graph Tree Decomposition*, volume 4175 of *Lecture Notes in Computer Science*, pages 262–273. Springer, Berlin, Germany, 2006.

[77] M. Zuker, D. H. Mathews, and D. H. Turner. *Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide.* NATO ASI Series. Kluwer Academic Publishers, 1999.

[78] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic acids research*, 9(1):133–148, January 1981.