Separating Structural Non-coding RNAs From Genomic Backgrounds

by

Yingfeng Wang

(Under the direction of Liming Cai and Russell L. Malmberg)

Abstract

Non-coding RNAs (ncRNAs) are RNA molecules without potential of producing proteins. The computational detection of ncRNA genes in genomic backgrounds requires capturing the signals of ncRNAs. However, unlike protein-coding genes, strong, universal sequential signals for identification ncRNAs have not yet been discovered.

RNA secondary structure has been widely applied as an exploitable feature for identifying ncRNA genes. Under the traditional Boltzmann ensemble of secondary structures, some structure based approaches have been developed, which have diverse performance across different ncRNA species. The mixed success of traditional structure based methods shows that some features of ncRNA sequences may not have been captured by the canonical secondary structure space defined with the Boltzmann structure ensemble.

In this dissertation, I introduce novel models of the RNA secondary structure by narrowing down the space with incorporating structural elements favored by tertiary structures. The significant performance improvement achieved by the new models in separating ncRNAs from other sequences suggests that investigating secondary structure space is a promising approach to design an effective ncRNA gene finding method.

Index words: RNA secondary structure, Boltzmann ensemble, RNA tertiary structure, Stochastic context-free grammar, Inside algorithm, Outside algorithm

Separating Structural Non-coding RNAs From Genomic Backgrounds

by

Yingfeng Wang

B.S., Hohai University, China, 2002

M.S., Hohai University, China, 2005

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2012

©2012

Yingfeng Wang

all rights reserved

Separating Structural Non-coding RNAs From Genomic Backgrounds

by

Yingfeng Wang

Approved:

Major Professors:	Dr. Liming Cai
	Dr. Russell L. Malmberg
Committee:	Dr. E. Rodney Canfield
	Dr. Tianming Liu

Electronic Version Approved:

Grasso, Maureen
Dean of the Graduate School
The University of Georgia
May 2012

Dedication

This dissertation is dedicated to my parents and my wife. Without their continuous support, I would not have been able to finish my Ph.D. study. This dissertation is also dedicated to my wonderful daughter, who brings a lot of joy.

Acknowledgments

I would like to thank my major advisor, Dr. Liming Cai, for his guidance and advice through my Ph.D. study, my co-major advisor, Dr. Russell L. Malmberg, for his advice and introducing knowledge of biology and bioinformatics, Dr. E. Rodney Canfield and Dr. Tianming Liu for serving my dissertation committee. I also would like to thank Dr. Khaled Rasheed, Dr. Michael Terns, and Dr. Rebecca Terns for their help.

My research has been supported in part by NSF MRI 0821263, NIH BISTI R01GM072080-01A1 grant, NIH ARRA Administrative Supplement to NIH BISTI R01GM072080-01A1, NSF IIS grant of award No: 0916250, and a dissertation completion award from the graduate school.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

RNA is the intermediate of the well known central dogma [4, 5]. In the central dogma, DNAs can be transcribed to RNAs, and RNAs can be translated to proteins. In the early 1980s, the catalytic functions of RNAs were discovered [31, 21]. Scientists started realizing the existence of some RNAs without the potential to be translated to proteins. This category of RNAs are called non-coding RNAs (ncRNAs), while DNAs, producing ncRNAs, were termed ncRNA genes.

In the past three decades, ncRNAs have been verified to play key functions in various biological processes such as gene regulation, catalysis, and RNA splicing [14, 62, 8]. The advent of new RNA families and functions is receiving more and more attention [14, 18, 16].

One critical mission of ncRNAs research is to identify ncRNA genes in newly sequenced genomes. The current sequencing technologies allow researchers to quickly produce an enormous amount of genome sequences [62]. However, finding ncRNAs in genomes is still expensive and time consuming through experiments. By far, there are just a small percentage of ncRNA genes found. For example, the current UCSC genome browser has thousands of

found ncRNA genes in humans, while it is suggested that there are about 16,000 or more ncRNAs in humans[60, 69]. This urgent need encourages computational scientists to develop novel effective computational tools for ncRNA gene finding.

1.2 Non-coding RNA Gene Finding

There are three types of computational ncRNA gene finding methods. The first is profile based gene finding. These methods detect ncRNA genes based on given profiles, e.g., sequence alignments [13, 50]. Since this type of approach requires known profiles, they are used to search for specific ncRNA genes of a known structure. The second is based on comparative analysis. Gene regions are usually more conserved than regions without genes. By comparing two or more related genomes, comparative analysis methods are able to identify ncRNA genes including novel genes [47, 45]. However, comparative analysis methods are limited by the availability of evolutionarily related genomes. The third type of methods are called *ab initio* methods; these don't require any priori knowledge for identifying ncRNA genes on a query genome. This dissertation is focused on *ab initio* methods to detect structural ncRNAs.

RNA secondary structure has been the most exploited feature in ncRNA gene finding [15, 38, 59]. In particular, a ncRNA sequence is expected to have a thermodynamically more stable secondary structure than one predicted from a non-structural sequence; this has energized some leading groups to develop structure-based ncRNA gene finding tools [47, 66, 65, 45]. Most of these tools compute the minimum free energy, using the thermodynamic energy model [58, 57, 71, 23], as the fold stability of a given sequence; however, they may also rely on multiple genome sequences to incorporate additional information into their models. For example, RNAz considers sequence stability and secondary structure conservation [66, 65]. Evofold uses not only a sequence alignment, but also phylogenetic information

incorporated into a stochastic context-free grammar (SCFG) [45]. However, the fold stability measure may underperform relative to expectations because of the small difference in free energy between native and non-native fold states. Random folds by chance also complicate the issue [68, 48, 2, 43, 17]. Additionally, real genome backgrounds tend to have higher levels of partial RNA-like structures than do purely random sequences.

Fold certainty, the Shannon entropy over alternative secondary structures defined with the Boltzmann ensemble [42], has also been used to characterize ncRNAs. The fold certainty is often approximated with the entropy defined over alternative base pairs [41, 26]. The certainty of base pairs is expected to be low for native ncRNA sequences. The fold certainty measure has shown a strong correlation with the fold stability measure [8, 16]; both gave diverse performances across different ncRNA data sets [16, 2]. For example, they perform very well on miRNA precursors but poorly on tRNAs tested against randomly shuffled sequences.

The mixed success of the traditional structure based measures poses the question whether we need new structural measures based on the traditional ensemble or a new structural ensemble. We believe the latter. The need of performance improvement motivated us to seek ways for narrowing down the RNA secondary structure space.

1.3 A New Structural Measure

Because ncRNAs functions may be determined by their tertiary structure [1, 37], incorporating tertiary structure characteristics of ncRNAs into structural measures may improve performance in ncRNA detection. Indeed, earlier work on RNA secondary structure prediction showed improved results when coaxial stacking of helices was incorporated [61]. Other tertiary motifs, for instance tetra-loops, have been considered in some of the newest versions of secondary structure prediction programs [23, 24, 39].

My first attempt is to require all stems are stable, as indicated in known RNA tertiary structures. This is enforced as a simple requirement that a stem should contain at least three consecutive canonical base pairs, reflecting the energetic stability of helices in the tertiary structure. With the Shannon base pair entropy measure, this method significantly improves Boltzmann ensemble-based programs (e.g., RNAfold and NUPACK) in their ability to distinguish all 13 native ncRNAs [16] from randomly shuffled sequences. Although this method is not able to distinguish native ncRNAs from genomic backgrounds, it demonstrates a potential to effectively detect ncRNAs with secondary structure models constrained with tertiary elements.

Based on the success of adding constraints for stable stems, I then built constraints on the loops. Since RNA junctions play important roles in RNA tertiary structures, constraints on the loops are derived from data on existing RNA junctions. In this work, I present a novel structure based measure for ncRNAs, which has achieved the following performances. (1) It can effectively distinguish many native structural ncRNAs from genomic sequences; (2) it has nearly the same performance across all the 13 ncRNAs datasets of Freyhult, et al. [16] and the 51 ncRNA benchmarks from Rfam selected by Nawrocki et al. [44].

The new structural measure computes an *inside probability-like weight* score for any given sequence with a weighted Context Free Grammar (CFG) model for RNA secondary structure. In a number of aspects, our model differs from previously used SCFG models [45, 11, 49] that underlie the thermodynamic energy based Boltzmann ensemble [42]. First, with a CFG, this model describes RNA secondary structure as a collection of k-way junctions, $k \geq 1$. Second, constraints on the junction structures were obtained from the native ncRNA tertiary database [46] and incorporated into the model. Third, the probability distribution for CFG rules was replaced with weights to avoid biases against any alternative structure. These novelties make it possible for the method to effectively distinguish native ncRNAs from genomic backgrounds. In particular, for almost every tested ncRNA dataset (of the

13 families [16] and the 51 families [44]), this method can detect more than 70% of the ncRNA sequences when they are compared to the *Pyrococcus furiosus* genomic background with about 85% specificity.

1.4 Dissertation Outline

The rest of this dissertation is organized as follows. Some RNA structure related concepts are briefly introduced in chapter 2. Chapter 3 presents a constrained RNA secondary structure space for stable stems, based on which our method successfully distinguishes non-coding RNAs from random backgrounds. Chapter 4 proposed a junction based constrained RNA secondary structure space. Our method with this space has a good performance for separating non-coding RNAs from genomic backgrounds. The theoretical analysis of the relationship between thresholds, GC contents, and window sizes are given in chapter 5. Chapter 6 concludes this dissertation and discusses the future work.

Chapter 2

RNA Structure Related Fundamentals

The present work detects ncRNAs according to RNA structures. This chapter introduces some RNA structure related fundamentals.

2.1 RNA Structures

An RNA primary sequence consists of four nucleotides, i.e., adenine, cytosine, guanine, and uracil, which are represented by A, C, G, and U, respectively. The counterpart of uracil in DNAs is thymine, since a thymine in DNAs is transcribed into a uracil in RNAs. Thymine is designated by T.

RNA secondary structure consists of unpaired nucleotides and canonical base pairs; the latter are interactions between nucleotides including Watson-Crick pair G-C and A-U, and Wobble pair G-U. Figure 2.1 shows a secondary structure of a tRNA, which can also be represented by the dot-bracket format as follows,

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

where unpaired nucleotides are represented by ".", and paired nucleotides are represented

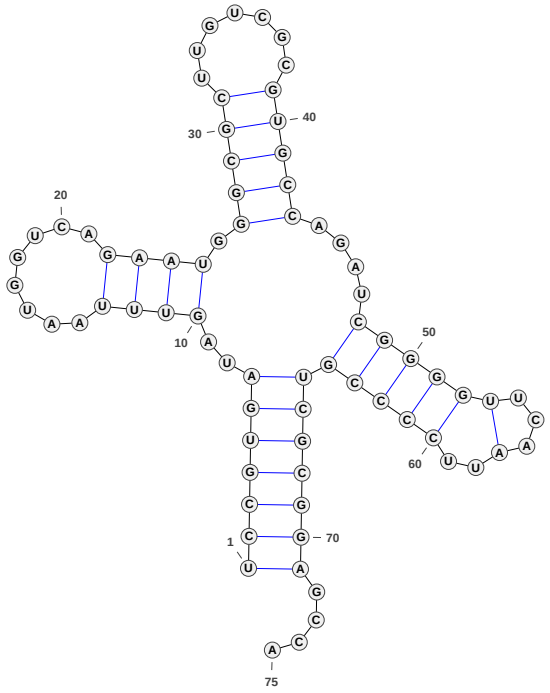

Figure 2.1: Example of a secondary structure of a tRNA. It consists of four stems, which are combined by a four-way junction. It also has three hairpin loops (drawn by VARNA [6]).

as "(" and ")". A pseudoknot is a special secondary structure containing at least two stems. In a pseudoknot, two halves of a stem are separated by half of anther stem, e.g., AAAA....(((..aaaa.))), where "(" and ")" represent paired nucleotides of a stem, and "A" and "a" are paired nucleotides of another stem.

Canonical base pairs are a signature of the secondary structure. RNA tertiary structure considers both canonical and non-canonical base pairs. All steric information of atoms of an RNA is included in its tertiary structure. Figure 2.2 illustrates the tertiary structure of a tRNA. RNA tertiary structures are expected to have a strong impact on RNA functions [22]. Since RNA secondary structures are the foundation of RNA tertiary structures, RNA secondary structures have been attracting much attention.

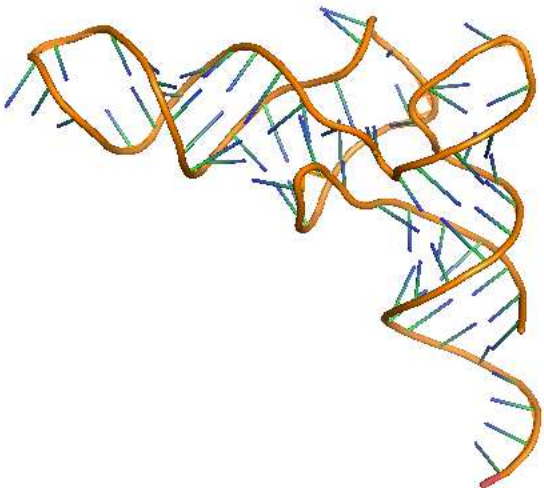

Figure 2.2: Example of a tertiary structure of the RNA molecule of PDB id 1VTQ (drawn with PyMOL [52]).

2.2 Boltzmann Ensemble and Energy Model

Thermodynamic free energy has been applied to measure fold stability of a RNA secondary structure. Most energy-based structure models of RNA are based on Boltzmann ensemble, which is a probabilistic space of all alternative RNA secondary structures of any given RNA sequence [42]. Figure 2.3 illustrates the Boltzmann ensemble of a given sequence. The Boltzmann equilibrium probability of an alternative secondary structure I with given sequence S is

$$P(I) = \frac{e^{-\frac{E(S,I)}{RT}}}{U} \tag{2.1}$$

where $E(S, I)$ is the free energy of the structure for the sequence, R is the gas constant, T is the absolute temperature, and U is the partition function for all admissible secondary

structures of the given RNA sequence, i.e.,

$$U = \sum_{I} e^{-\frac{E(S,I)}{RT}} \tag{2.2}$$

The Boltzmann ensemble is statistically characterized by the Boltzmann equilibrium distribution [8]. Borrowing formulas from [42, 8], we may explain how to calculate the probabilities of structures. Given sequence S with n nucleotides and S_{ij} is the segment of S from nucleotide i to j , where $1 \leq i < j \leq n$. Let I_{ij} be a secondary structure on S_{ij}, and C_{ij} be a secondary structure on S_{ij} with nucleotide i and j being paired, we are able to respectively calculate the summation of probabilities of all I_{ij} and all C_{ij} by the following two formulas,

$$u(i,j) = \sum_{I_{ij}} e^{-\frac{E(S_{ij},I_{ij})}{RT}} \tag{2.3}$$

and

$$p(i,j) = \sum_{C_{ij}} e^{-\frac{E(S_{ij},C_{ij})}{RT}} \tag{2.4}$$

where the summation for $u(i,j)$ is based on all alternative secondary structures I_{ij}, and the summation for $p(i,j)$ is based on all alternative secondary structures C_{ij}, while $E(S_{ij}, I_{ij})$ is the free energy of I_{ij} on S_{ij} and $E(S_{ij}, C_{ij})$ is the free energy of C_{ij} on S_{ij}. Therefore, given the primary sequence S_{1n}, the probability of secondary structure I_{1n} is,

$$P(I_{1n}|S_{1n}) = \frac{e^{-\frac{E(S_{1n},I_{1n})}{RT}}}{u(1,n)} \tag{2.5}$$

where $P(I_{1n}|S_{1n})$ is equal to $P(I)$. Furthermore, the probability of nucleotide i being paired with nucleotide j, represented by B_{ij}, can be calculated by the summation of all alternative

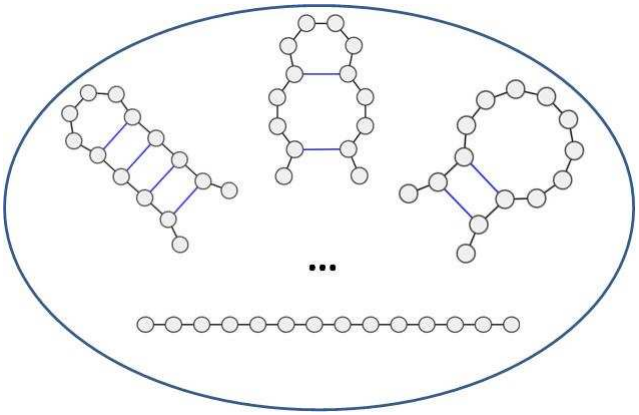

Figure 2.3: Example of a Boltzmann ensemble of a given sequence (drawn by R2R [67] and Visio).

secondary structures having base pair (i, j), i.e.,

$$B_{ij} = \sum_{base\ pair\ (i,j) \in I} P(I). \tag{2.6}$$

It can also be rewritten as[1],

$$B_{ij} = \frac{u(1, i-1)p(i, j)u(j+1, n)}{u(1, n)}. \tag{2.7}$$

where B_{ij} is based on S_{1n}, while $p(i, j)$ is based on S_{ij}. Formula 2.7 divides a secondary structure into three substructures that are the substructures on $S_{1(i-1)}$, S_{ij}, and $S_{(j+1)n}$. Although McCaskill's algorithm for calculating base pair probabilities [42] doesn't directly predict secondary structures, the Boltzmann ensemble approach has been successfully applied for RNA secondary structure prediction.

Based on the Boltzmann ensemble, some energy minimization based methods have been developed [70, 23, 24, 39, 9]. These methods calculate the free energy of given secondary

[1]We don't consider pseudoknots.

structures by dynamic programming with energy parameters associating structural elements such as base pairs, hairpin loops, bulges, internal loops, and multi-loops [18]. Since it was suggested that RNA secondary structures are stabler than random folds [36, 35, 34, 3], energy models have been used for secondary structure prediction by searching the structures with minimum free energy (MFE) [70, 23, 24, 39, 9]. However, free energy parameters are still not accurate by far. For instance, the free energy of a multi-loop is usually calculated by a linear function [24], which has been inconsistent with observations [7].

2.3 Context-Free Grammar Based Methods

A context free grammar (CFG) is a formal grammar system composed of production rules with the format as follows,

$$S \to \alpha$$

where S is a non-terminal and α is a string composed of terminals and non-terminals. A CFG can produce pseudoknot-free RNA secondary structures. For example, $S \to n$ can generate an unpaired nucleotide, and $S \to nS$ can recursively generate a loop composed of some unpaired nucleotides, where S is a non-terminal, and n is a terminal representing A, C, G, or U. Furthermore, production rules like $S \to lSr$ can recursively produce base pairs, where S is a non-terminal, l and r are terminals, respectively representing the left and right nucleotides of a base pair.

If each production rule is associated with a positive weight, it is weighted context-free grammar (WCFG). If the weights of rules of the same left-hand-side non-terminals sum to 1, a weight can be explained as a probability, and the grammar is stochastic context-free grammar (SCFG).

Inside and Outside algorithms have been developed for SCFG [12]. Given a non-terminal

S in a SCFG, the *inside probability* on the sequence x with given starting position i and ending position j is defined as

$$\alpha(S, i, j, x) = Prob(S \Rightarrow^* x_i x_{i+1} \cdots x_j) \tag{2.8}$$

while the corresponding *outside probability* is defined as

$$\beta(S, i, j, x) = Prob(S_0 \Rightarrow^* x_1 \cdots x_{i-1} S x_{j+1} \cdots x_n) \tag{2.9}$$

where S_0 is the starting non-terminal (see Figure 2.4 for illustration). By taking advantage of Inside and Outside algorithms, the probability of nucleotide i and j being paired, represented by $P_{i,j}(x)$, can be calculated by the following expression,

$$P_{i,j}(x) = \frac{\sum_{S \to aRbT} Prob(S \to aRbT, a = x_i, b = x_j)\, \gamma(R, S, T, i, j, x)}{\alpha(S_0, 1, n, x)} \tag{2.10}$$

where

$$\gamma(R, S, T, i, j, x) = \sum_{j < k \leq n} \alpha(R, i+1, j-1, x) \times \beta(S, i, k, x) \times \alpha(T, j+1, k, x)$$

in which S, R, T are non-terminals. If a rule doesn't have T, k is fixed as j and term $\alpha(T, j+1, k, x)$ is equal to 1.

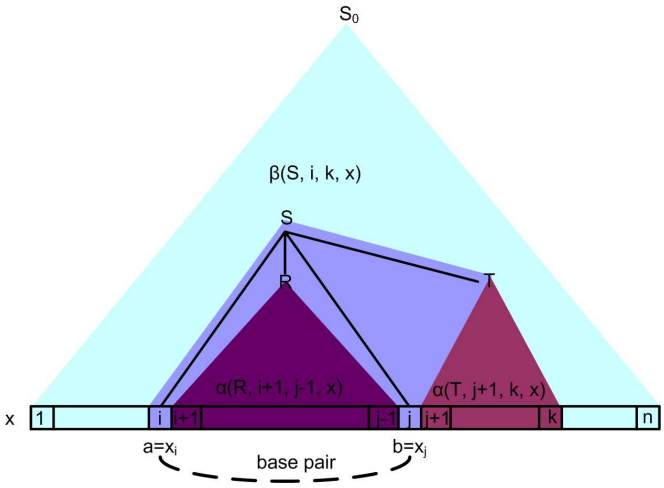

Figure 2.4: Illustration of the application of the generic production rule $S \rightarrow aRbT$ that produces a base pair between positions i and j for the query sequence x, provided that the start non-terminal S_0 derives $x_1 x_2 \ldots x_{i-1} S x_{k+1} \ldots x_n$. Note that given i and j, the position of k can vary (reproduced with permission from BioMed Central [64]).

Chapter 3

Distinguish Non-coding RNAs from Random Sequences [1]

3.1 Method and Model

Our method to distinguish ncRNAs from random sequences is based on measuring the base pairing Shannon entropy [41, 26] under a new RNA secondary structure model. The building blocks of this model are stems arranged in parallel and nested patterns connected by unpaired strand segments, similar to those permitted by a standard ensemble [71, 42, 24]. The new model is constrained, however, to contain a smaller space of equilibrium alternative structures, requiring there are only energetically stable stems (e.g., of free energy levels under a threshold) to occur in the structures. The constraint is basically to consider the effect of energetically stable stems on tertiary folding and to remove spurious structures that may not correspond to a tertiary fold. According to the RNA folding pathway theory and the hierarchical folding model [40, 56, 1], building block helices are first stabilized by canonical base pairings before being arranged to interact with each other or with unpaired strands

[1] Reprinted with permission from ©2011 IEEE [63] and BioMed Central [64]

through tertiary motifs (non-canonical nucleotide interactions). A typical example is the multi-loop junctions in which one or more pairs of coaxially stacked helices bring three or more regions together, further stabilized by the tertiary motifs at the junctions [37, 32]. The helices involved are stable before the junction is formed or any possible nucleotide interaction modifications are made to the helical base pairs at the junction [55].

3.1.1 Energetically stable stems

A stem is the atomic, structural unit of the new secondary structure space. To identify the energy levels of stems suitable to be included in this model, we conducted a survey on the 51 sets of ncRNA seed alignments, representatives of the ncRNAs in Rfam [20], which had been used with the software Infernal [44] as benchmarks. From each ncRNA seed structural alignment, we computed the thermodynamic free energy of every instance of a stem in the alignment data using various functions of the Vienna Package [23, 24] as follows. `RNAduplex` was first applied to the two strands of the stem marked by the annotation to predict the optimal base pairings within the stem, then, the minimum free energy of the predicted stem structure, with overhangs removed, was computed with `RNAeval`. Figures 3.1 and 3.2 respectively show plots of the percentages and cumulative percentages of free energy levels of stems in these 51 ncRNA seed alignments.

The peaks (with relatively high percentages) on the percentage curve of Figure 3.1 indicate concentrations of certain types of stems at energies levels around -4.5, -3.3, and -2.4 kcal/mol. Since a `G-U` pair is counted weakly towards the free energy contribution (by the Vienna package), we identified the peak value -4.5 kcal/mol to be the free energy of stems of three base pairs, with two `G-C` pairs and one `A-U` in the middle or two `A-U` pairs and one `G-C` in the middle. The value -3.3 kcal/mol is the free energy of stems containing exactly two `G-C` pairs or stems with one `G-C` pair followed by two `A-U` pairs. Values around -2.4 kcal/mol are stems containing one `G-C` and an `A-U` pair or simply four `A-U` pairs.

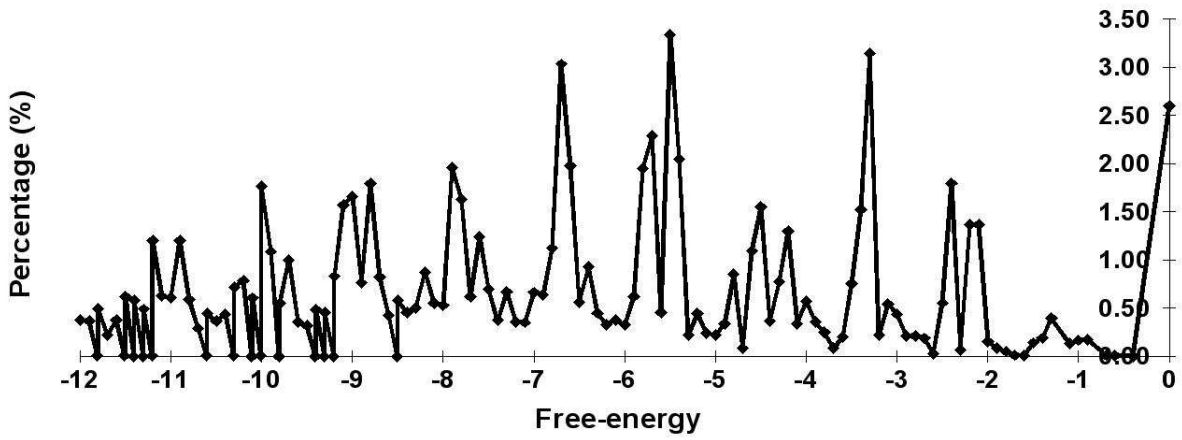

Figure 3.1: Percentages of free-energy of stems from 51 Rfam datasets (percentages of stems with free-energy less than -12 are not given in this figure)

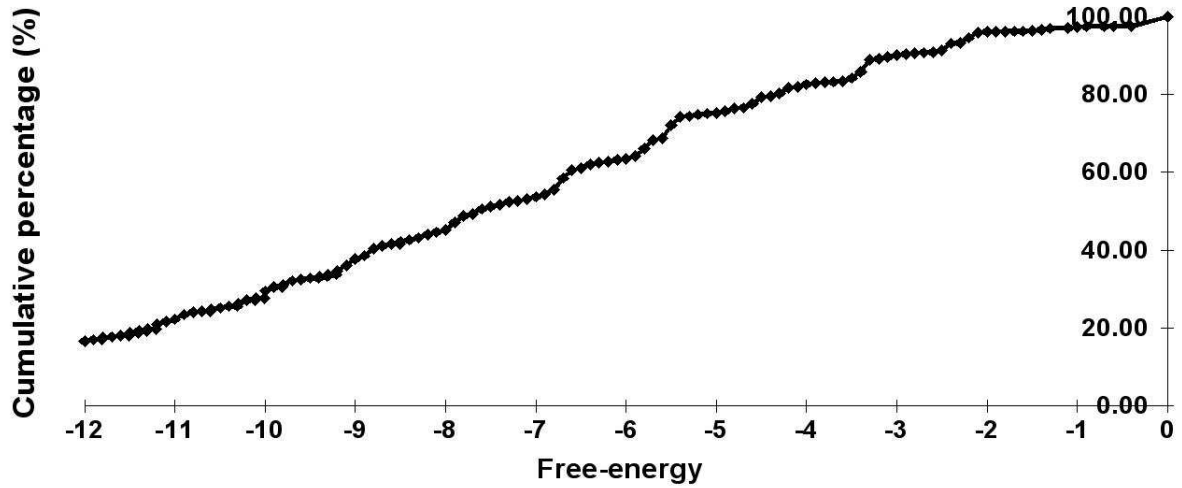

Figure 3.2: Cumulative percentages of free-energy of stems from 51 Rfam datasets (cumulative percentages of stems with free-energy less than -12 are not given in this figure). Note the step at -3.4.

Based on this survey, we were able to identify two energy thresholds: -3.4 and -4.6 kcal/mol for *semi-stable stems* and *stable stems* respectively. Both require at least three base pairs of which at least one is `G-C` pair. We further observed the difference between these two categories of stems on the 51 ncRNA datasets. In general, although levels of energy appear to be somewhat uniformly distributed (see Figure 3.1), an overwhelmingly large percentage of stems in both categories are located in the vicinity of other stems. In particular, 79.6% of stable stems (with a free energy -4.6 kcal/mol or lower) have 0 (number of nucleotides) distance from their closest neighbor stem and 16.5% of stable stems have distance 1 from their closest neighbors. For semi-stable stems, the group having zero distance to other stems is 85.6% of the total while the group having distance 1 is 10.6%. Since zero distance between two stems may reflect a contiguous strand connecting two coaxially stacked helices in tertiary structure, our survey suggests a semi-stable stem interacts with another stem to maintain even its own local stability. In the rest of this work, we do not distinguish between stable and semi-stable stems. In conducting this survey, we did not directly use the stem structures annotated in the seed alignments to compute their energies. Due to evolution, substantial structural variation may occur across species; one stem may be present in one sequence and absent in another but a structural alignment algorithm may try to align all sequences to the consensus stem, giving rise to "misalignments" which we have observed [25]. Most of such "malformed stems" mistakenly aligned to the consensus often contain bulges or internal loops and have higher free energies greater than the threshold -3.4 kcal/mol.

3.1.2 The RNA secondary structure model

In the present study, a secondary structure model is defined with a Stochastic Context Free Grammar (SCFG) [12]. Our model requires there are at least three consecutive base pairs in every stem; the constraint is described with the following seven generic production rules:

(1) $X \rightarrow a$ (2) $X \rightarrow aX$ (3) $X \rightarrow aHb$

(4) $X \rightarrow aHbX$ (5) $H \rightarrow aHb$ (6) $H \rightarrow aYb$

(7) $Y \rightarrow aXb$

where capital letters are non-terminal symbols that define substructures and low case letters are terminals, each being one of the four nucleotides `A`, `C`, `G`, and `U`.

The starting non-terminal, X, can generate an unpaired nucleotide or a base pair with the first three rules. The fourth rule generates two parallel substructures. Non-terminal H is used to generate consecutive base pairs with non-terminal Y to generate the closing base pair. Essentially, the process of generating a stem needs to recursively call production rules with the left-hand-side non-terminals X, H and Y each at least once. This constraint guarantees that every stem has at least three consecutive base pairs, as required by our secondary structure model.

3.1.3 Probability parameter calculation

There are two sets of probability parameters associated with the induced SCFG. First, we used a simple scheme of probability settings for the unpaired bases and base pairs, with a uniform 0.25 probability for every base. The probability distribution of $\{0.25, 0.25, 0.17, 0.17, 0.08, 0.08\}$ is given to the six canonical base pairs `G-C`, `C-G`, `A-U`, `U-A`, `G-U`, and `U-G`; a probability of zero is given to all non-canonical base pairs. Alternatively, probabilities for unpaired bases and base pairs may be estimated from available RNA datasets with known secondary structures [20], as has been done in some of the previously work with SCFGs [28, 29].

Second, we computed the probabilities for the production rules of the model as follows. To allow our method to be applicable to all structural ncRNAs, we did not estimate the probabilities based on a training data set. In fact, we believe that the probability parameter setting of an SCFG for the fold certainty measure should be different from that for fold

stability measure (i.e., folding). Based on the principle of maximum entropy, we developed the following approach to calculate the probabilities for the rules in our SCFG model.

Let p_i be the probability associated with the production rule i, for $i = 1, 2, \ldots, 7$, respectively. Since the summation of probabilities of rules with the same non-terminal on the left-hand-side is required to be 1, we can establish the following equations:

$$\begin{cases} p_1 + p_2 + p_3 + p_4 & = 1 \\ p_5 + p_6 & = 1 \\ p_7 & = 1 \end{cases}$$

Let

$$q_{bp} = \sqrt[6]{0.25 \times 0.25 \times 0.17 \times 0.17 \times 0.08 \times 0.08}$$

be the geometric average of the six base pair probabilities. According to the principle of maximum entropy, given we have no prior knowledge of a probability distribution, the assumption of a distribution with the maximum entropy is the best choice, since it will take the smallest risk [27]. If we apply this principle to our problem, the probability contribution from a base pair should be close to the contribution from unpaired bases. Rule probabilities can be estimated to satisfy following equations:

$$\begin{cases} p_1 & = p_2 \\ p_3 & = p_4 \\ (q_{bp})^3 \times p_3 \times p_6 \times p_7 & = (0.25 \times p_1)^6 \\ (q_{bp})^4 \times p_3 \times p_5 \times p_6 \times p_7 & = (0.25 \times p_1)^8 \end{cases}$$

From above equations, it follows that

$$p_1 = 0.499 \quad p_2 = 0.499 \quad p_3 = 0.001$$
$$p_4 = 0.001 \quad p_5 = 0.103 \quad p_6 = 0.897$$
$$p_7 = 1$$

3.1.4 Computing base pairing Shannon Entropy

Based on the new RNA secondary structure model, we can compute the fold certainty of any given RNA sequence, which is defined as the Shannon entropy measured on base pairings formed by the sequence over the specified secondary structure space Ω. Specifically, let the sequence be $x = x_1 x_2 \dots x_n$ of n nucleotides. For indexes $i < j$, the probability $P_{i,j}$ of base pairing between bases x_i and x_j is computed with

$$P_{i,j}(x) = \sum_{s \in \Omega} p(s, x)\, \delta(x)^s_{i,j} \tag{3.1}$$

where $p(s, x)$ is the probability of x being folded into to the structure s in the space Ω and $\delta(x)^s_{i,j}$ is a binary value indicator for the occurrence of base pair (x_i, x_j) in structure s. The Shannon entropy of $P_{i,j}(x)$ is computed as [41, 26]

$$Q(x) = -\frac{1}{n} \sum_{i<j} P_{i,j}(x) \log P_{i,j}(x) \tag{3.2}$$

To compute the expected frequency of the base pairing, $P_{i,j}(x)$, with formula 3.1, we take advantage of the Inside and Outside algorithms developed for SCFG [12]. Given any nonterminal symbol S in the grammar, the *inside probability* is defined as

$$\alpha(S, i, j, x) = Prob(S \Rightarrow^* x_i x_{i+1} \cdots x_j) \tag{3.3}$$

i.e., the total probability for the sequence segment $x_i x_{i+1} \cdots x_j$ to adopt alternative substructures specified by S. Assume S_0 to be the initial nonterminal symbol for the SCFG model. Then $\alpha(S_0, 1, n, x)$ is the total probability of the sequence x's folding under the model.

The *outside probability* is defined as

$$\beta(S, i, j, x) = Prob(S_0 \Rightarrow^* x_1 \cdots x_{i-1} S x_{j+1} \cdots x_n) \tag{3.4}$$

i.e., the total probability for the whole sequence $x_1 \cdots x_n$ to adopt all alternative substructures that allow the sequence segment from position i to position j to adopt any substructure specified by S (see Figure 2.4 for illustration).

$P_{i,j}(x)$ then can be computed as the normalized probability of the base pair (x_i, x_j) occurring in all valid alternative secondary structures of x:

$$P_{i,j}(x) = \frac{\sum_{S \to aRbT} Prob(S \to aRbT, a = x_i, b = x_j)\, \gamma(R, S, T, i, j, x)}{\alpha(S_0, 1, n, x)} \tag{3.5}$$

where

$$\gamma(R, S, T, i, j, x) = \sum_{j < k \leq n} \alpha(R, i+1, j-1, x) \times \beta(S, i, k, x) \times \alpha(T, j+1, k, x)$$

in which variables S, R, T are for non-terminals and variable production $S \to aRbT$ represents rules (3)$\sim$(7) which involve base pair generations. For rules where T is empty, the summation and term $\alpha(T, j+1, k, x)$ do not exist and k is fixed as j.

The efficiency to compute $P_{i,j}(x)$ mostly depends on computing the *Inside* and *Outside* probabilities, which can be accomplished with dynamic programming and has the time complexity $O(mn^3)$ for a model of m nonterminals and rules and sequence length n.

3.2 Experimental Results

We implemented the algorithm for Shannon base pairing entropy calculation into a program named TRIPLE. We tested it on ncRNA datasets and compared its performance on these ncRNAs with the performance achieved by the software NUPACK [9] and RNAfold [23, 24] developed under the Boltzmann standard secondary structure ensemble [42, 10].

3.2.1 Data preparation

We downloaded the 13 ncRNA datasets previously investigated in Table 1 of [16]. They are of diverse functions, including pre-cursor microRNAs, group I and II introns, RNase P and MRP, bacterial and eukaryotic signal recognition particle (SRP), ribosomal RNAs, small nuclear spliceosomal RNAs, riboswitches, tmRNAs, regulatory RNAs, tRNAs, telomerase RNAs, small nucleolar RNAs, and Hammerhead ribozymes.

The results from using these datasets were analyzed with 6 different types of measures, including Z-score and p-value of minimal free energy (MFE), and Shannon base pairing entropy [16], in comparisons with random sequences. The six measures correlate to varying degrees, hence using MFE Z-score and Shannon base pairing entropy may be sufficient to cover the other measures. However, these two measures, as the respective indicators for the fold stability and fold certainty of ncRNA secondary structure, have varying performances on the 13 ncRNA datasets.

For our tests, we also generated random sequences as control data. For every ncRNA sequence, we randomly shuffled it to produce two sets of 100 random sequences each; one set was based upon single nucleotide shuffling, the other was based upon di-nucleotide shuffling. In addition, all ncRNA sequences containing nucleotides other than `A`, `C`, `G`, `T`, and `U` were removed for the reason that NUPACK [9] doesn't accept sequences containing wildcard symbols.

3.2.2 Shannon entropy distribution of random sequences

Two energy model based softwares, NUPACK (with the pseudoknot function turned off) and RNAfold, and our program TRIPLE computed base pairing probabilities on ncRNA sequences and on random sequences. In particular, for every ncRNA sequence $\mathbf{x}$ and its associated randomly shuffled sequence set $S_{\mathbf{x}}$, the Shannon entropies of these sequences were computed.

A Kolmogorov-Smirnov test (KS test) [30] was applied to verify the normality of the entropy distributions from all randomly shuffled sequence sets. The results show that for 99% of the sequence sets we fail to reject the hypothesis that entropies are normally distributed with 95% confidence level. This indicates that we may use a Z-score to measure performance.

3.2.3 Z-score scores and comparisons

For each ncRNA, the average and standard deviation of Shannon entropies of the randomly shuffled sequences were estimated. The Z-score of the Shannon entropy $Q(\mathbf{x})$ of ncRNA sequence $\mathbf{x}$ is defined as follows:

$$Z(\mathbf{x}) = \frac{\mu(Q(S_{\mathbf{x}})) - Q(\mathbf{x})}{\sigma(Q(S_{\mathbf{x}}))} \tag{3.6}$$

where $\mu(Q(S_{\mathbf{x}}))$ and $\sigma(Q(S_{\mathbf{x}}))$ respectively denote the average and standard deviation of the Shannon entropies of the random sequences in set $S_{\mathbf{x}}$. The Z-Score measures how well entropies may distinguish the real ncRNA sequence $\mathbf{x}$ from their corresponding randomly shuffled sequences in $S_{\mathbf{x}}$. Figure 3.3 compares the averages of the Z-scores of Shannon base pairing entropies computed by NUPACK, RNAfold, and TRIPLE on each of the 13 ncRNA datasets. It shows that TRIPLE significantly improved the Z-scores over NUPACK and RNAfold across all the 13 datasets.

To examine how the Z-scores might have been improved by TRIPLE, we designated four

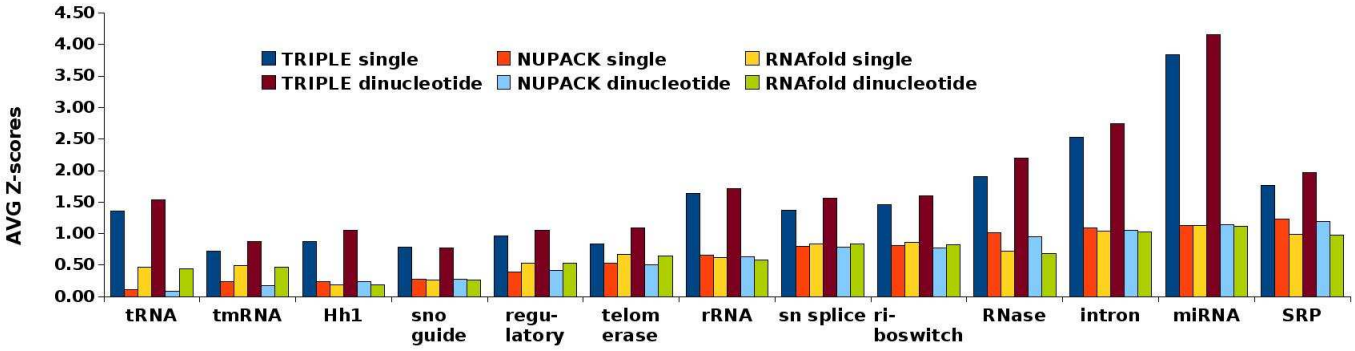

Figure 3.3: Comparisons of averaged Z-score of Shannon base pairing entropies computed by NUPACK, RNAfold, and TRIPLE for each of the 13 ncRNA datasets downloaded from [16]

thresholds for Z-scores, which are 2, 1.5, 1, and 0.5. The percentages of sequences of each dataset with Z-score greater than or equal to the thresholds were computed.

Table 3.1 shows details of the Z-score improvements over NUPACK when di-nucleotide shuffling was used. With a threshold 2 or 1.5, our method performed better than NUPACK in all datasets. With the threshold 1 and 0.5, our method improved upon NUPACK in 12 and 10 datasets, respectively. The results of TRIPLE and NUPACK using a single nucleotide random shuffling are given in Table 3.2, which shows that our method also performs better than NUPACK in the majority of datasets. In particular, TRIPLE performed better than NUPACK in all datasets with threshold of 2; with threshold equal to 1.5 or 1, our method had better results than NUPACK in 12 datasets and in 9 datasets with threshold equal of 0.5.

The results of RNAfold using the default setting are given in Table 3.3 and 3.4. Table 3.3 shows results on di-nucleotide shuffling datasets. TRIPLE works better in the majority of datasets. It outperforms RNAfold in all datasets with threshold equal to 2 and 1.5. With threshold of 1 and 0.5, TRIPLE wins 12 (tie 1) and 8 (tie 1) datasets, respectively. In

Table 3.1: Comparisons of TRIPLE and NUPACK by the percentages of sequences falling in each category of a Z-score range. Random sequences were obtained with di-nucleotide shuffling of the real ncRNA sequences.

ncRNA	Method	Z $\geq$ 2.0	Z $\geq$ 1.5	Z $\geq$ 1.0	Z $\geq$ 0.5
Hh1	TRIPLE	26.67	40.00	53.33	73.33
	NUPACK	0.00	0.00	20.00	53.33
sno_guide	TRIPLE	14.43	24.45	38.39	58.19
	NUPACK	0.73	8.80	27.63	45.23
sn_splice	TRIPLE	40.51	50.63	60.76	65.82
	NUPACK	3.80	18.99	48.10	70.89
SRP	TRIPLE	35.06	44.16	59.74	67.53
	NUPACK	3.90	36.36	72.73	85.71
tRNA	TRIPLE	29.56	51.33	70.97	86.02
	NUPACK	0.00	2.30	12.04	32.21
intron	TRIPLE	60.75	69.16	78.50	85.98
	NUPACK	1.87	19.63	61.68	85.05
riboswitch	TRIPLE	34.64	48.37	60.13	78.43
	NUPACK	1.96	18.95	45.75	69.28
miRNA	TRIPLE	81.48	88.89	94.07	97.04
	NUPACK	0.00	12.59	68.15	97.78
telomerase	TRIPLE	29.41	35.29	41.18	58.82
	NUPACK	11.76	17.65	35.29	47.06
RNase	TRIPLE	50.70	70.42	81.69	92.25
	NUPACK	5.63	23.94	48.59	72.54
regulatory	TRIPLE	22.41	24.14	32.76	56.90
	NUPACK	1.72	3.45	18.97	51.72
tmRNA	TRIPLE	18.64	32.20	45.76	55.93
	NUPACK	1.69	8.47	27.12	37.29
rRNA	TRIPLE	36.16	50.62	70.87	83.06
	NUPACK	4.75	21.07	42.56	61.16

Table 3.2: Comparisons of TRIPLE and NUPACK by the percentages of sequences falling in each category of a Z-score range. Random sequences were obtained with single nucleotide shuffling of the real ncRNA sequences.

ncRNA	Method	Z $\geq$ 2	Z $\geq$ 1.5	Z $\geq$1	Z $\geq$ 0.5
Hh1	TRIPLE	6.67	33.33	53.33	73.33
	NUPACK	0.00	0.00	20.00	60.00
sno_guide	TRIPLE	14.91	25.43	41.10	57.95
	NUPACK	0.98	9.05	28.85	45.72
sn_splice	TRIPLE	31.65	43.04	56.96	65.82
	NUPACK	5.06	26.58	51.90	69.62
SRP	TRIPLE	32.47	45.45	55.84	68.83
	NUPACK	3.90	37.66	72.73	87.01
tRNA	TRIPLE	24.07	45.31	64.25	79.47
	NUPACK	0.00	2.12	14.69	33.45
intron	TRIPLE	59.81	68.22	74.77	84.11
	NUPACK	1.87	22.43	66.36	85.98
riboswitch	TRIPLE	32.03	44.44	56.86	71.90
	NUPACK	1.96	21.57	46.41	69.28
miRNA	TRIPLE	75.56	81.48	90.37	93.33
	NUPACK	0.00	9.63	70.37	98.52
telomerase	TRIPLE	23.53	29.41	41.18	58.82
	NUPACK	5.88	29.41	29.41	52.94
RNase	TRIPLE	38.03	56.34	72.54	87.32
	NUPACK	10.56	26.06	52.11	76.06
regulatory	TRIPLE	18.97	25.86	31.03	51.72
	NUPACK	0.00	1.72	24.14	50.00
tmRNA	TRIPLE	15.25	27.12	38.98	57.63
	NUPACK	3.39	6.78	27.12	42.37
rRNA	TRIPLE	34.09	47.31	64.88	79.96
	NUPACK	6.40	21.69	43.19	60.74

Table 3.3: Comparisons of TRIPLE and RNAfold by the percentages of sequences falling in each category of a Z-score range. Random sequences were obtained with di-nucleotide shuffling of the real ncRNA sequences.

dataset	method	≥2 (%)	≥1.5 (%)	≥1(%)	≥0.5 (%)
Hh1	TRIPLE	26.67	40.00	53.33	73.33
	RNAfold	0.00	0.00	20.00	53.33
sno_guide	TRIPLE	14.43	24.45	38.39	58.19
	RNAfold	1.71	7.82	23.96	43.03
sn_splice	TRIPLE	40.51	50.63	60.76	65.82
	RNAfold	6.33	21.52	54.43	69.62
SRP	TRIPLE	35.06	44.16	59.74	67.53
	RNAfold	5.19	24.68	58.44	71.43
tRNA	TRIPLE	29.56	51.33	70.97	86.02
	RNAfold	0.18	4.25	24.78	47.96
intron	TRIPLE	60.75	69.16	78.50	85.98
	RNAfold	2.80	17.76	60.75	84.11
riboswitch	TRIPLE	34.64	48.37	60.13	78.43
	RNAfold	0.65	17.65	47.06	70.59
miRNA	TRIPLE	81.48	88.89	94.07	97.04
	RNAfold	0.00	7.41	65.93	97.78
telomerase	TRIPLE	29.41	35.29	41.18	58.82
	RNAfold	0.00	23.53	41.18	58.82
RNase	TRIPLE	50.70	70.42	81.69	92.25
	RNAfold	1.41	12.68	34.51	59.15
regulatory	TRIPLE	22.41	24.14	32.76	56.90
	RNAfold	0.00	6.90	27.59	63.79
tmRNA	TRIPLE	18.64	32.20	45.76	55.93
	RNAfold	1.69	10.17	33.90	50.85
rRNA	TRIPLE	36.16	50.62	70.87	83.06
	RNAfold	1.45	15.70	35.33	56.82

Table 3.4: Comparisons of TRIPLE and RNAfold by the percentages of sequences falling in each category of a Z-score range. Random sequences were obtained with single nucleotide shuffling of the real ncRNA sequences.

dataset	method	$\geq$2 (%)	$\geq$1.5 (%)	$\geq$1 (%)	$\geq$0.5 (%)
Hh1	TRIPLE	6.67	33.33	53.33	73.33
	RNAfold	0.00	0.00	20.00	53.33
sno_guide	TRIPLE	14.91	25.43	41.10	57.95
	RNAfold	1.47	7.33	24.21	44.01
sn_splice	TRIPLE	31.65	43.04	56.96	65.82
	RNAfold	6.33	24.05	53.16	68.35
SRP	TRIPLE	32.47	45.45	55.84	68.83
	RNAfold	5.19	29.87	59.74	77.92
tRNA	TRIPLE	24.07	45.31	64.25	79.47
	RNAfold	0.00	6.19	26.19	48.85
intron	TRIPLE	59.81	68.22	74.77	84.11
	RNAfold	1.87	16.82	58.88	85.98
riboswitch	TRIPLE	32.03	44.44	56.86	71.90
	RNAfold	1.31	20.92	49.67	71.24
miRNA	TRIPLE	75.56	81.48	90.37	93.33
	RNAfold	0.74	10.37	69.63	97.78
telomerase	TRIPLE	23.53	29.41	41.18	58.82
	RNAfold	5.88	17.65	35.29	58.82
RNase	TRIPLE	38.03	56.34	72.54	87.32
	RNAfold	1.41	15.49	35.92	61.27
regulatory	TRIPLE	18.97	25.86	31.03	51.72
	RNAfold	0.00	5.17	32.76	67.24
tmRNA	TRIPLE	15.25	27.12	38.98	57.63
	RNAfold	0.00	11.86	35.59	45.76
rRNA	TRIPLE	34.09	47.31	64.88	79.96
	RNAfold	1.86	17.98	37.60	57.64

Table 3.4, TRIPLE shows similar performance on single nucleotide shuffling datasets. It has better scores than RNAfold in 13, 13, 11, and 7 (tie 1) datasets with threshold of 2, 1.5, 1, and 0.5, respectively. In addition, RNAfold was tested with the available program options (tables not shown). With option "noLP" on RNAfold, TRIPLE performs better in 13, 13, 11 (tie 1), and 9 di-nucleotide shuffling datasets in terms of threshold of 2, 1.5, 1, and 0.5, respectively. In single nucleotide shuffling datasets, TRIPLE wins 13, 13, 12 and 8 datasets separately with threshold of 2, 1.5, 1, and 0.5. When we specify "noLP" and "noCloseGU" on RNAfold, TRIPLE beats RNAfold in 13, 13, 12, and 11 di-nucleotide shuffling datasets, and 13, 13, 13, and 11 single nucleotide shuffling datasets with threshold 2, 1.5, 1, and 0.5, respectively. If we specify "noLP" and "noGU" on RNAfold, our method performs better on all di-nucleotide shuffling and single nucleotide shuffling datasets with all four thresholds.

We also compared TRIPLE, NUPACK, and RNAfold on some real genome background tests. Several genome sequences from bacteria, archaea, and eukaryotes were retrieved from the NCBI database. Using these genome sequences, we created genome backgrounds for the 13 ncRNA data sets. In particular, for each RNA sequence from 13 ncRNA data sets, 100 sequence segments of the same length were sampled from each genome sequence and used to test against the RNA sequence to calculate base pairing entropies and Z-score. With such genome backgrounds, the overall performance of TRIPLE on the 13 ncRNA data sets is mixed and is close to that of NUPACK and RNAfold (data not shown). This performance of TRIPLE on real genomes indicates that there is still a gap between the ability of our method and successful ncRNA gene finding. Nevertheless, the test results reveal that the constrained "triple base pairs" model is necessary but still not sufficient enough. This suggests incorporating further structural constraints will improve the effectiveness for ncRNA search on real genomes.

To roughly evaluate the speed of the three tools, the running time for 101 sequences, including 1 real miRNA sequence and its 100 single nucleotide shuffled sequences, was mea-

sured on a Linux machine with an Intel dual-core CPU (E7500 2.93GHz). Each sequence has 100 nucleotides. TRIPLE, NUPACK, and RNAfold spent 20.7 seconds, 36.2 seconds and 3.4 seconds, respectively. We point out that TRIPLE has the potential to be optimized for each specific grammar to improves its efficiency.

3.3 Discussion

This work introduced a modified ensemble of ncRNA secondary structures with the constraint of requiring only canonical base pairs to only occur and that stems must be energetically stable in all the alternative structures. The comparisons of performances between our program TRIPLE and energy model based software (NUPACK and RNAfold) implemented based on the canonical structure ensemble have demonstrated a significant improvement in the entropy measure for ncRNA fold certainty by our model. In particular, an improvement of the entropy Z-scores was shown across almost all 13 tested ncRNAs datasets previously used to test various ncRNA measures [16].

We note that there is only one exceptional case observed from Table 3.1-3.4: SRP whose entropy Z-score performance was not improved (as much as other ncRNAs) when $Z < 1.5$. The problem might have been caused by the implementation technique rather than the methodology. Most of the tested SRP RNA sequences (Eukaryotic and archaeal 7S RNAs) are of length around 300 and contain about a dozen stems. In many of them, consecutive base pairs are broken by internal loops into small stem pieces, some having only two consecutive canonical pairs; whereas, in our SCFG implementation we simply required three consecutive base pairs as a must in a stem, possibly missing the secondary structure of many of these sequences. This issue with the SCFG can be easily fixed, e.g., by replacing the SCFG with one that better represents the constrained Boltzmann ensemble in which stems are all energetically stable.

To ensure that the performance difference between TRIPLE and energy model based software (NUPACK and RNAfold) was *not* due to the difference in the thermodynamic energy model (Boltzmann ensemble) and the simple statistical model (SCFG) with stacking rules, we also constructed two additional SCFG models, one for unconstrained base pairs and another requiring at least two consecutive canonical base pairs in stems. Tests on these two models over the 13 ncRNA data set resulted in entropy Z-scores (data not shown) comparable to those obtained by NUPACK and RNAfold but inferior to the performance of TRIPLE. We attribute the impressive performance by TRIPLE to the constraint of "triple base pairs" satisfied by real ncRNA sequences but which is hard to achieve for random sequences.

Since the entropy Z-score improvement by our method was not uniform across the 13 ncRNAs, one may want to look into additional other factors that might have contributed to the under-performance of certain ncRNAs. For example, the averaged `GC` contents are different in these 13 datasets, with SRP RNAs having 58% GC and standard deviation of 10.4%. A sequence with a high `GC` content is more likely to produce more spurious, alternative structures, possibly resulting in a higher base pairing entropy. However, since randomly shuffled sequences would also have the same `GC` content, it becomes very difficult to determine if the entropies of these sequences have been considerably affected by the `GC` bias. Indeed, previous investigations [53] have revealed that, while the base composition of a ncRNA is related to the phylogenetic branches on which the specific ncRNA may be placed, it may not fully explain the diverse performances of structure measures on various ncRNAs. Notably it has been discovered that base compositions are distinct in different parts of rRNA secondary structure (stems, loops, bulges, and junctions) [54], suggesting that an averaged base composition may not suitably represent the global structural behavior of an ncRNA sequence.

Technically the TRIPLE program was implemented with an SCFG that assumes stems to have at least three consecutive canonical base pairs. Yet, as we pointed out earlier, the

performance results should hold for a constrained Boltzmann ensemble in which stems are required to be energetically stable. This constraint of stable stems was intended to capture the energetic stability of helical structures in the native tertiary fold [40, 56]. Since the ultimate distinction between a ncRNA and a random sequence lies in its function (thus tertiary structure); additional, critical tertiary characteristics may be incorporated into the structure ensemble to further improve the fold certainty measure. In our testing of stem stability (see section "Energetically stable stems"), ncRNA sequences from the 51 datasets demonstrated certain sequential properties that may characterize tertiary interactions, e.g., coaxial stacking of helices. However, to computationally model tertiary interactions, a model beyond a context-free system would be necessary; thus it would be difficult to use an SCFG or a Boltzmann ensemble for this purpose. We need to develop methods to identify tertiary contributions critical to the Shannon base pairing entropy measure and to model such contributions. Although this method and technique have been developed with reference to non-coding RNAs, it is possible that protein-coding mRNAs would display similar properties, when sufficient structural information about them has been gathered.

Chapter 4

Separating Non-coding RNAs from Genomic Backgrounds [1]

4.1 Method

Our method constrains the RNA secondary structure space by incorporating elements based upon RNA junctions. Consider an RNA secondary structure as consisting of interconnected junctions. A junction can be a k-way junction, for some $k \geq 1$, which is a closed loop enclosed by k helices. For example, one-way junction, two-way junction, and k-way junction ($k \geq 3$) correspond to stem-loop, internal loop, and multi-loop in an secondary structure. Thus, our secondary structure space was defined to contain various loops of junctions. A k-way junction is composed of a leading helix, k loops, and $k-1$ entries for other junctions. Figre 4.1 illustrates a RNA secondary structure starting with a two-way junction (red), then a three-way junction (blue) followed by a two-way junction (green) and a one-way junction (black); the green two-way junction also ends with a one-way junction (yellow).

[1] Yingfeng Wang, Russell L. Malmberg, and Liming Cai, A Novel Structural Measure Separating Non-Coding RNAs from Genomic Backgrounds, submitted to Proceedings of 11th European Conference on Computational Biology (ECCB)

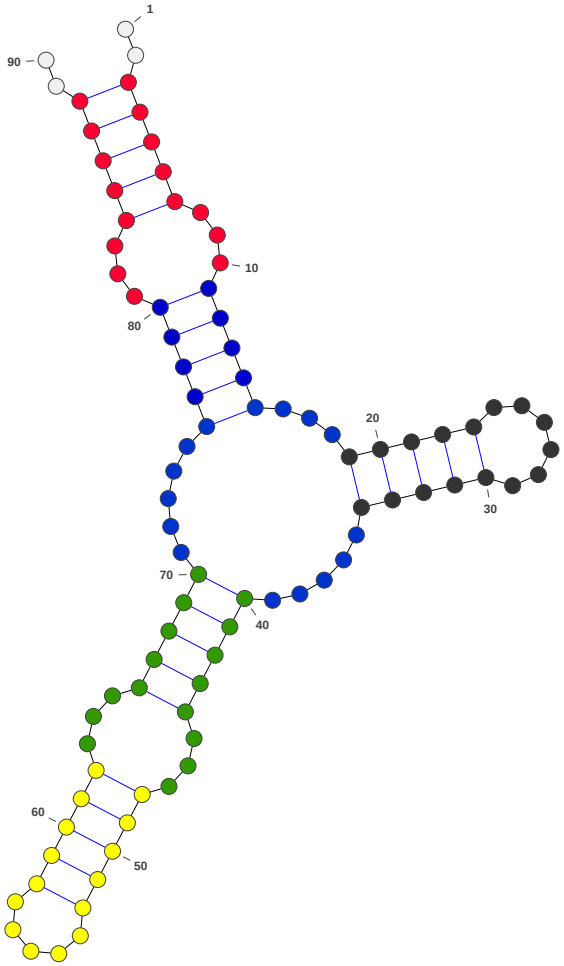

Figure 4.1: Illustration of a RNA secondary structure starting with a two-way junction (red), then a three-way junction (blue) followed by a two-way junction (green) and a one-way junction (black); the green two-way junction also ends with a one-way junction (yellow). (This figure was made using VARNA [6])

Table 4.1: Constraints of loop lengths of two-way junctions

left loop length	right loop length
0	1,2,3,4
1	0,1,2,3
2	0,1,2,3,4
3	0,1,2,3,4,5,6,7
4	1,2,3,4,5,6,7
5	1,2,3,4,5,6,7
6	3,4,5
7	2,3,4,5,6,7

Known tertiary structures were investigated by retrieving their preferences for loop lengths. We obtained the numbers of junctions with given fixed loop lengths by submitting queries to an available native ncRNA database RNA FRABASE 2.0 [46]. Based on the distributions of loop lengths, the constraints of loop lengths of k-way junctions ($1 \leq k \leq 5$) were set separately so as to cover more than 90% cases on each category. Our investigation shows that the loops of one-way junctions usually have 3 to 15 unpaired nucleotides, while most four-way and five-way junctions have at least two loops with up to 2 unpaired nucleotides, and other loops contain up to 7 nucleotides. The constraints we observed on loop lengths of two-way and three-way junctions are given in Table 4.1 and 4.2, respectively. Additionally, some constraints of helices were also included based on an energy model [23, 24] to guarantee all detected helices are thermodynamically stable.

The constrained secondary structure space was modeled using a context-free grammar (CFG). To fit the junction based structure space, five subsets of production rules were used to describe k-way junctions ($1 \leq k \leq 5$) separately. This CFG had weights assigned to production rules which were not probabilities, such as are used in a stochastic context free grammar (SCFG). More details about setting the weights are given in the next section.

Table 4.2: Constraints of loop lengths of three-way junctions

left loop length	middle loop length	right loop length
0	2,3,4,5	2,3
1	3,4,5,6,7,8,9	2,3,4,5,6,7,8
2	1,2,3,4,5,15	1,2,3,4,5,6
3	0,1,2,3,4,5,6	3,4,5,6,7
4	0,1,2,3	1,2,3
5	0,1,2	2,3
5	3,4	6,7
6	1,4,10,11	3,4
7	3,4	2,3

4.2 Model and Algorithm

Our secondary structure space model was defined with a weighted context-free grammar (WCFG), where each production rule is associated with a weight which is not constrained to be a probability. A context-free grammar (CFG) can be used to generate RNA secondary structures [12]. In practice, the weighted CFG approach allows us more space to set weights than a strict SCFG would. In our grammar, we use production rules such as $T \Rightarrow LJO$ to start a new junction based on the current k-way junction ($k \geq 2$), where T, L, J, and O are all non-terminals representing the starting position, e.g. the end of the leading helix of the current junction, the first loop, the new junction, and other parts to the right of the new junction, respectively. If $k = 2$, the current junction is a two-way junction, and O will generate a loop. Otherwise, O will generate other new junctions by recursively using the previous rule format.

In our approach, all unpaired nucleotides were treated equally. Meanwhile, our model roughly assigned weights 0.25, 0.25, 0.17, 0.17, 0.08, and 0.08 to canonical base pairs C-G, G-C, A-U, U-A, G-U, and U-G, respectively. All other non-canonical base pairs were assigned zero weight. In addition, T was considered to be the same as U whether paired or unpaired.

With this setting, the weights of production rules were set following the principle that the weight gain of producing two unpaired nucleotides should be equal to that of generating one base pair, on average. This principle guaranteed our weight setting to show no preference between folding and non-folding substructures. In our current grammar, the weight gain of generating an unpaired nucleotide is 1. So the gain of two unpaired nucleotides is $1 \times 1 = 1$. If there is a production rule producing a canonical base pair within C-G, G-C, A-U, U-A, G-U, and U-G, then the average weight gain of this base pair is $w \times q_{bp}$, where q_{bp} is the geometric average of base pair weights, and can be calculated by the following expression,

$$q_{bp} = \sqrt[6]{0.25 \times 0.25 \times 0.17 \times 0.17 \times 0.08 \times 0.08}.$$

According to our principle, we have

$$w \times q_{bp} = 1.$$

So the weight w of this rule is 6.65. Weights of other production rules generating base pairs were calculated similarly.

Under this structure model, we are able to compute the summation of weights of all alternative structures fitting our structure space by adopting the inside algorithm, which was originally developed for stochastic context-free grammar (SCFG) [12]. For a given nonterminal S, the *inside probability-like weight* is defined as

$$\alpha(S, i, j, x) = Weight(S \Rightarrow^* x_i x_{i+1} \cdots x_j)$$

i.e., the summation of weights of alternative substructures specified by S for the sequence segment $x_i x_{i+1} \cdots x_j$. So, $\alpha(S_0, 1, n, x)$ was the summation of weights of all sequence x's alternative structures under the model, where S_0 was the initial nonterminal. The inside

algorithm can efficiently compute $\alpha(S_0, 1, n, x)$ based upon dynamic programming with time complexity $O(mn^3)$ for a grammar containing m nonterminals and rules and a sequence having n nucleotides.

We hypothesized that the *inside probability-like weight* of real ncRNAs would be significantly higher than that of genomic backgrounds, since real ncRNAs are assumed to have more alternative structures fitting our structure space than does the background of genomic sequence.

4.3 Results

A program named JUNCTION was implemented for the *inside probability-like weight* calculation. This program was tested on ncRNA datasets, and its performance was compared with that of the minimum free energy (MFE) calculated by RNAFOLD [23, 24], which was built based on the Boltzmann secondary structure ensemble [42].

Our basic experiment was to obtain known ncRNAs, compare their scores with equivalent genomic sequences of the same length, then determine the detection accuracy.

4.3.1 Data preparation

We chose standard ncRNA sequence families which have been used by others for similar tests. All 13 ncRNAs datasets of Freyhult, et al. [16] and 51 ncRNA benchmarks from Rfam selected by Nawrocki et al. [44] were downloaded. All sequences having nucleotides other than A, C, G, U, and T were removed. For each of the Freyhult 13 ncRNA datasets, we randomly picked 50 sequences, if there were more than 50 sequences. For each dataset of Nawrocki 51 ncRNA datasets, we randomly picked 100 sequences, if there were more than 100 sequences.

We prepared genomic background sequences using a *Pyrococcus furiosus* sequence from

which we removed the annotated genes. For each native ncRNA sequence, 100 genomic background segments with the same length were randomly obtained from the *Pyrococcus furiosus* genome. All genomic background segments were further queried against the sequences in Rfam [19, 20] to guarantee they had no match with default setting.

4.3.2 Interquartile range based score

The tests generated 101 values for each ncRNA sequence being examined, in which one value was from the ncRNA sequence itself, while the other 100 values were based on corresponding genomic backgrounds. We needed a measurement to evaluate the performance of the two methods in distinguishing real ncRNAs from its genomic backgrounds. Since the values of our *inside probability-like weight* were not normally distributed, we used interquartile range (IQR) based scores, which do not rely on a normal distribution, to evaluate the performance of JUNCTION and RNAFOLD. High values of an *inside probability-like weight* indicated likely real ncRNAs, so we sorted all values of these weights for the 100 genomic background segments in ascending order. Further, low MFEs may indicate likely real ncRNAs. Therefore, we sorted all MFEs of all 100 genomic backgrounds in descending order. Let Q_1 be the value of the first quartile and Q_3 be the value of the third quartile. The IQR is then $|Q_3 - Q_1|$. We can compute the IQR based Score by

$$Score = \frac{real - Q_3}{Q_3 - Q_1},$$

where *real* is either the *inside probability-like weight* or the MFE. If IQR is zero, the Score is set to be 10000, when $real > Q_3$ for the *inside probability-like weight*, or $real < Q_3$ for the MFE. If IQR is zero, the Score is set to be -10000, when $real \leq Q_3$ for the *inside probability-like weight*, or $real \geq Q_3$ for the MFE. High IQR based scores on real ncRNAs mean high sensitivities of distinguishing ncRNAs, while low IQR based scores on genomic backgrounds

indicates high specificities.

4.3.3 Comparison

Both the Freyhult 13 and Nawrocki 51 datasets were tested with JUNCTION and RNAFOLD. To evaluate performance, we set 5 thresholds of IQR based scores, which are 0, 1.5, 5, 10, and 100. Tables 4.3 and 4.4 shows details of the percentages of IQR based scores with JUNCTION and RNAFOLD on the 13 datasets, while tables 4.5, 4.6, 4.7, 4.8, 4.9 and 4.10 show details of the percentages on 51 datasets.

JUNCTION consistently received high IQR scores on real ncRNAs. We use 5 as the threshold for distinguishing ncRNAs from backgrounds. JUNCTION has 2 datasets over 90% of the sequences detected, 7 datasets over 80%, 10 datasets over 70%, and 11 datasets over 65% on the 13 data; it has 15 datasets with over 90% of the sequences detected, 27 datasets over 80%, 32 datasets over 70%, and 33 datasets over 65% on the 51 RNAs. The average sensitivities of JUNCTION are 78% and 71% on the 13 and 51 datasets, respectively. In comparison, RNAFOLD has 12 datasets with below 20% detection on the Freyhult 13 RNAs, and its highest percentage is 58.82% ; RNAFOLD also has 47 datasets with below 20% detection on the Nawrocki 51 RNAs, and its highest percentage is 54.05%. The average sensitivities of RNAfold are 9% and 4% on the 13 and 51 datasets, respectively. So JUNCTION significantly improves sensitivity in comparison of RNAFOLD.

We also use 5 as the threshold for IRQ based Scores on genomic backgrounds. On the Freyhult 13 RNAs, JUNCTION has 11 datasets with percentages between 10% and 20%; the lowest percentage is 9.32%, and the highest is 20.94%. On the Nawrocki 51 RNAs, JUNCTION has 45 datasets with percentages between 10% and 20%; the lowest percentage is 2.36%, and the highest is 20.70%. The average specificities of JUNCTION are 85% and 85% on the 13 and 51 datasets, respectively. On the 13 RNAs, RNAFOLD has 12 datasets with percentages below 0.5%; the highest percentage is 0.53%. With the 51 RNAs, RNAFOLD

Table 4.3: Comparison of percentages of IQR based scores with real ncRNAs on 13 datasets [16]

		real ncRNAs (higher is better)				
Species	Method	≥100	≥10	≥5	≥1.5	≥0
Hh1	JUNCTION	26.67%	46.67%	46.67%	46.67%	46.67%
	RNAFOLD	0.00%	0.00%	0.00%	13.33%	86.67%
RNase	JUNCTION	88.00%	90.00%	92.00%	92.00%	92.00%
	RNAFOLD	0.00%	0.00%	12.00%	88.00%	98.00%
SRP	JUNCTION	82.00%	86.00%	88.00%	96.00%	100.00%
	RNAFOLD	0.00%	0.00%	18.00%	90.00%	100.00%
intron	JUNCTION	48.00%	68.00%	72.00%	76.00%	82.00%
	RNAFOLD	0.00%	0.00%	2.00%	50.00%	94.00%
miRNA	JUNCTION	60.00%	80.00%	82.00%	84.00%	86.00%
	RNAFOLD	0.00%	0.00%	2.00%	98.00%	100.00%
rRNA	JUNCTION	78.00%	82.00%	82.00%	82.00%	86.00%
	RNAFOLD	0.00%	0.00%	8.00%	70.00%	94.00%
regulatory	JUNCTION	30.00%	56.00%	68.00%	76.00%	82.00%
	RNAFOLD	0.00%	0.00%	6.00%	50.00%	98.00%
riboswitch	JUNCTION	72.00%	80.00%	86.00%	92.00%	98.00%
	RNAFOLD	0.00%	0.00%	4.00%	58.00%	96.00%
sn_splice	JUNCTION	52.00%	74.00%	76.00%	78.00%	82.00%
	RNAFOLD	0.00%	0.00%	0.00%	58.00%	92.00%
sno_guide	JUNCTION	16.00%	42.00%	50.00%	68.00%	78.00%
	RNAFOLD	0.00%	0.00%	0.00%	20.00%	86.00%
tRNA	JUNCTION	62.00%	76.00%	78.00%	80.00%	84.00%
	RNAFOLD	0.00%	0.00%	0.00%	60.00%	86.00%
telomerase	JUNCTION	100.00%	100.00%	100.00%	100.00%	100.00%
	RNAFOLD	0.00%	0.00%	58.82%	100.00%	100.00%
tmRNA	JUNCTION	78.00%	88.00%	88.00%	90.00%	90.00%
	RNAFOLD	0.00%	0.00%	6.00%	44.00%	92.00%

has all datasets with percentages no greater than 0.5%;the highest is 0.5%. The average specificities of RNAFOLD are 99% and 99% on 13 and 51 datasets, respectively. Thus, RNAFOLD has a better specificity than JUNCTION.

Table 4.4: Comparison of percentages of IQR based scores with genomic backgrounds on 13 datasets [16]

		genomic backgrounds (lower is better)				
Species	Method	≥100	≥10	≥5	≥1.5	≥0
Hh1	JUNCTION	5.20%	9.20%	11.40%	13.67%	32.13%
	RNAFOLD	0.00%	0.00%	0.00%	2.47%	26.33%
RNase	JUNCTION	11.64%	17.80%	19.46%	22.10%	26.00%
	RNAFOLD	0.00%	0.00%	0.16%	1.92%	26.08%
SRP	JUNCTION	5.76%	11.82%	14.60%	19.30%	26.02%
	RNAFOLD	0.00%	0.00%	0.12%	2.30%	26.24%
intron	JUNCTION	4.28%	10.48%	13.16%	18.32%	26.02%
	RNAFOLD	0.00%	0.00%	0.16%	1.82%	26.18%
miRNA	JUNCTION	2.42%	8.60%	12.12%	17.62%	26.00%
	RNAFOLD	0.00%	0.00%	0.02%	2.00%	26.10%
rRNA	JUNCTION	12.66%	17.62%	19.24%	22.12%	26.00%
	RNAFOLD	0.00%	0.00%	0.18%	2.06%	26.08%
regulatory	JUNCTION	2.20%	6.86%	9.32%	14.32%	27.84%
	RNAFOLD	0.00%	0.00%	0.04%	2.40%	26.34%
riboswitch	JUNCTION	5.88%	13.38%	16.08%	20.10%	26.00%
	RNAFOLD	0.00%	0.00%	0.22%	2.48%	26.06%
sn_splice	JUNCTION	3.90%	11.22%	14.30%	18.78%	26.00%
	RNAFOLD	0.00%	0.00%	0.14%	2.08%	26.12%
sno_guide	JUNCTION	2.62%	8.78%	12.06%	17.92%	26.02%
	RNAFOLD	0.00%	0.00%	0.04%	2.06%	26.28%
tRNA	JUNCTION	1.34%	6.96%	10.24%	16.70%	26.04%
	RNAFOLD	0.00%	0.00%	0.00%	2.32%	26.12%
telomerase	JUNCTION	16.06%	19.94%	20.94%	22.94%	26.00%
	RNAFOLD	0.00%	0.00%	0.53%	2.65%	26.00%
tmRNA	JUNCTION	12.90%	17.94%	19.86%	22.08%	26.00%
	RNAFOLD	0.00%	0.02%	0.30%	2.20%	26.12%

4.4 Discussion

This study introduced a modified secondary structure ensemble with constraints based on RNA junctions. The measure based on this secondary structure can efficiently distinguish native structural ncRNA sequences from genomic backgrounds and obtained a consistent good performance on two standard ncRNA datasets. The comparison between our software

Table 4.5: Comparison of percentages of IQR based scores with real ncRNAs on 51 datasets [44] (part 1)

		real ncRNAs (higher is better)				
Rfam ID	Method	≥100	≥10	≥5	≥1.5	≥0
RF00002	JUNCTION	41.94%	77.42%	82.26%	83.87%	87.10%
	RNAFOLD	0.00%	0.00%	0.00%	43.55%	96.77%
RF00003	JUNCTION	99.00%	100.00%	100.00%	100.00%	100.00%
	RNAFOLD	0.00%	0.00%	0.00%	98.00%	100.00%
RF00004	JUNCTION	73.00%	83.00%	86.00%	89.00%	91.00%
	RNAFOLD	0.00%	0.00%	0.00%	67.00%	100.00%
RF00005	JUNCTION	68.00%	84.00%	84.00%	85.00%	86.00%
	RNAFOLD	0.00%	0.00%	0.00%	57.00%	84.00%
RF00008	JUNCTION	73.81%	89.29%	89.29%	92.86%	92.86%
	RNAFOLD	0.00%	0.00%	0.00%	85.71%	100.00%
RF00009	JUNCTION	99.00%	99.00%	99.00%	99.00%	99.00%
	RNAFOLD	0.00%	0.00%	11.00%	95.00%	99.00%
RF00010	JUNCTION	99.00%	99.00%	99.00%	99.00%	99.00%
	RNAFOLD	0.00%	0.00%	42.00%	92.00%	100.00%
RF00011	JUNCTION	80.00%	84.00%	85.00%	85.00%	88.00%
	RNAFOLD	0.00%	0.00%	1.00%	55.00%	92.00%
RF00012	JUNCTION	85.19%	96.30%	96.30%	96.30%	100.00%
	RNAFOLD	0.00%	0.00%	0.00%	81.48%	100.00%
RF00015	JUNCTION	33.00%	52.00%	54.00%	60.00%	76.00%
	RNAFOLD	0.00%	0.00%	0.00%	21.00%	93.00%
RF00017	JUNCTION	99.00%	99.00%	99.00%	100.00%	100.00%
	RNAFOLD	0.00%	0.00%	31.00%	99.00%	100.00%
RF00018	JUNCTION	85.71%	85.71%	85.71%	100.00%	100.00%
	RNAFOLD	0.00%	0.00%	0.00%	78.57%	100.00%
RF00019	JUNCTION	9.00%	34.00%	41.00%	54.00%	60.00%
	RNAFOLD	0.00%	0.00%	0.00%	40.00%	97.00%
RF00020	JUNCTION	20.00%	42.00%	50.00%	65.00%	79.00%
	RNAFOLD	0.00%	0.00%	0.00%	28.00%	91.00%
RF00023	JUNCTION	87.00%	90.00%	90.00%	91.00%	92.00%
	RNAFOLD	0.00%	0.00%	2.00%	65.00%	93.00%
RF00024	JUNCTION	100.00%	100.00%	100.00%	100.00%	100.00%
	RNAFOLD	0.00%	0.00%	54.05%	100.00%	100.00%
RF00025	JUNCTION	4.17%	37.50%	45.83%	45.83%	50.00%
	RNAFOLD	0.00%	0.00%	0.00%	16.67%	62.50%

Table 4.6: Comparison of percentages of IQR based scores with real ncRNAs on 51 datasets [44] (part 2)

		real ncRNAs (higher is better)				
Rfam ID	Method	≥100	≥10	≥5	≥1.5	≥0
RF00028	JUNCTION	26.67%	33.33%	40.00%	50.00%	53.33%
	RNAFOLD	0.00%	0.00%	0.00%	26.67%	73.33%
RF00029	JUNCTION	56.00%	81.00%	86.00%	91.00%	97.00%
	RNAFOLD	0.00%	0.00%	4.00%	76.00%	99.00%
RF00030	JUNCTION	87.64%	89.89%	92.13%	92.13%	94.38%
	RNAFOLD	0.00%	0.00%	7.87%	83.15%	95.51%
RF00031	JUNCTION	19.67%	60.66%	73.77%	85.25%	86.89%
	RNAFOLD	0.00%	0.00%	0.00%	45.90%	98.36%
RF00033	JUNCTION	0.00%	0.00%	0.00%	0.00%	25.00%
	RNAFOLD	0.00%	0.00%	0.00%	0.00%	25.00%
RF00037	JUNCTION	41.03%	41.03%	41.03%	41.03%	41.03%
	RNAFOLD	0.00%	0.00%	0.00%	15.38%	84.62%
RF00040	JUNCTION	83.33%	83.33%	83.33%	83.33%	83.33%
	RNAFOLD	0.00%	0.00%	0.00%	66.67%	100.00%
RF00054	JUNCTION	0.00%	25.00%	37.50%	50.00%	50.00%
	RNAFOLD	0.00%	0.00%	0.00%	0.00%	100.00%
RF00055	JUNCTION	22.22%	33.33%	44.44%	66.67%	77.78%
	RNAFOLD	0.00%	0.00%	0.00%	11.11%	88.89%
RF00059	JUNCTION	68.00%	78.00%	81.00%	84.00%	85.00%
	RNAFOLD	0.00%	0.00%	1.00%	58.00%	94.00%
RF00066	JUNCTION	6.38%	48.94%	63.83%	68.09%	82.98%
	RNAFOLD	0.00%	0.00%	0.00%	44.68%	85.11%
RF00067	JUNCTION	33.33%	66.67%	77.78%	83.33%	100.00%
	RNAFOLD	0.00%	0.00%	0.00%	27.78%	100.00%
RF00080	JUNCTION	68.00%	84.00%	92.00%	92.00%	96.00%
	RNAFOLD	0.00%	0.00%	0.00%	60.00%	100.00%
RF00096	JUNCTION	91.67%	95.83%	95.83%	95.83%	95.83%
	RNAFOLD	0.00%	0.00%	0.00%	68.75%	100.00%
RF00101	JUNCTION	38.46%	92.31%	92.31%	92.31%	100.00%
	RNAFOLD	0.00%	0.00%	0.00%	69.23%	100.00%
RF00104	JUNCTION	36.36%	81.82%	81.82%	90.91%	90.91%
	RNAFOLD	0.00%	0.00%	0.00%	100.00%	100.00%
RF00114	JUNCTION	3.75%	27.50%	37.50%	56.25%	62.50%
	RNAFOLD	0.00%	0.00%	0.00%	1.25%	82.50%

Table 4.7: Comparison of percentages of IQR based scores with real ncRNAs on 51 datasets [44] (part 3)

		real ncRNAs (higher is better)				
Rfam ID	Method	≥100	≥10	≥5	≥1.5	≥0
RF00163	JUNCTION	16.22%	31.08%	36.49%	39.19%	48.65%
	RNAFOLD	0.00%	0.00%	0.00%	22.97%	90.54%
RF00165	JUNCTION	0.00%	14.29%	42.86%	71.43%	78.57%
	RNAFOLD	0.00%	0.00%	0.00%	7.14%	100.00%
RF00167	JUNCTION	6.00%	35.00%	44.00%	63.00%	75.00%
	RNAFOLD	0.00%	0.00%	0.00%	6.00%	70.00%
RF00168	JUNCTION	53.19%	65.96%	68.09%	78.72%	82.98%
	RNAFOLD	0.00%	0.00%	2.13%	42.55%	87.23%
RF00169	JUNCTION	70.00%	90.00%	94.00%	96.00%	99.00%
	RNAFOLD	0.00%	0.00%	1.00%	94.00%	100.00%
RF00170	JUNCTION	10.00%	50.00%	50.00%	60.00%	70.00%
	RNAFOLD	0.00%	0.00%	0.00%	50.00%	100.00%
RF00174	JUNCTION	88.00%	93.00%	94.00%	94.00%	94.00%
	RNAFOLD	0.00%	0.00%	9.00%	82.00%	99.00%
RF00177	JUNCTION	64.00%	68.00%	71.00%	73.00%	75.00%
	RNAFOLD	0.00%	0.00%	3.00%	55.00%	88.00%
RF00206	JUNCTION	4.55%	27.27%	36.36%	45.45%	54.55%
	RNAFOLD	0.00%	0.00%	0.00%	0.00%	81.82%
RF00213	JUNCTION	5.26%	26.32%	26.32%	36.84%	47.37%
	RNAFOLD	0.00%	0.00%	0.00%	15.79%	57.89%
RF00230	JUNCTION	33.85%	50.77%	58.46%	66.15%	70.77%
	RNAFOLD	0.00%	0.00%	1.54%	7.69%	60.00%
RF00234	JUNCTION	38.89%	61.11%	72.22%	83.33%	88.89%
	RNAFOLD	0.00%	0.00%	0.00%	27.78%	100.00%
RF00373	JUNCTION	90.28%	95.83%	95.83%	97.22%	97.22%
	RNAFOLD	0.00%	0.00%	31.94%	94.44%	100.00%
RF00379	JUNCTION	77.00%	90.00%	93.00%	99.00%	99.00%
	RNAFOLD	0.00%	0.00%	1.00%	48.00%	97.00%
RF00380	JUNCTION	45.83%	66.67%	70.83%	80.21%	85.42%
	RNAFOLD	0.00%	0.00%	2.08%	58.33%	98.96%
RF00448	JUNCTION	100.00%	100.00%	100.00%	100.00%	100.00%
	RNAFOLD	0.00%	0.00%	0.00%	80.00%	100.00%
RF00504	JUNCTION	47.17%	75.47%	81.13%	84.91%	90.57%
	RNAFOLD	0.00%	0.00%	0.00%	66.04%	98.11%

Table 4.8: Comparison of percentages of IQR based scores with genomic backgrounds on 51 datasets [44] (part 1)

		genomic backgrounds (lower is better)				
Rfam ID	Method	≥100	≥10	≥5	≥1.5	≥0
RF00002	JUNCTION	4.87%	12.18%	15.15%	19.89%	26.00%
	RNAFOLD	0.00%	0.00%	0.16%	1.68%	26.10%
RF00003	JUNCTION	5.58%	13.23%	16.06%	20.00%	26.00%
	RNAFOLD	0.00%	0.00%	0.14%	2.05%	26.11%
RF00004	JUNCTION	6.88%	14.28%	16.69%	20.80%	26.00%
	RNAFOLD	0.00%	0.00%	0.21%	2.05%	26.08%
RF00005	JUNCTION	1.55%	7.43%	10.77%	16.68%	26.01%
	RNAFOLD	0.00%	0.00%	0.01%	2.42%	26.22%
RF00008	JUNCTION	2.13%	9.27%	12.69%	18.77%	28.36%
	RNAFOLD	0.00%	0.00%	0.02%	2.38%	26.31%
RF00009	JUNCTION	10.80%	16.54%	18.52%	21.56%	26.00%
	RNAFOLD	0.00%	0.02%	0.30%	2.11%	26.04%
RF00010	JUNCTION	12.45%	18.06%	19.69%	22.41%	26.00%
	RNAFOLD	0.00%	0.02%	0.30%	1.87%	26.02%
RF00011	JUNCTION	12.64%	17.92%	19.51%	22.13%	26.00%
	RNAFOLD	0.00%	0.00%	0.18%	1.85%	26.06%
RF00012	JUNCTION	7.63%	14.56%	17.07%	20.81%	26.00%
	RNAFOLD	0.00%	0.00%	0.19%	1.93%	26.07%
RF00015	JUNCTION	4.41%	11.95%	14.43%	19.59%	26.00%
	RNAFOLD	0.00%	0.01%	0.15%	2.24%	26.16%
RF00017	JUNCTION	11.03%	17.10%	19.01%	21.77%	26.00%
	RNAFOLD	0.00%	0.00%	0.25%	1.85%	26.10%
RF00018	JUNCTION	13.29%	17.00%	18.64%	21.79%	26.00%
	RNAFOLD	0.00%	0.00%	0.14%	2.00%	26.00%
RF00019	JUNCTION	2.49%	8.95%	12.18%	17.97%	26.00%
	RNAFOLD	0.00%	0.00%	0.06%	2.00%	26.20%
RF00020	JUNCTION	3.16%	10.44%	13.70%	18.85%	26.00%
	RNAFOLD	0.00%	0.00%	0.09%	1.78%	26.15%
RF00023	JUNCTION	12.85%	18.09%	19.81%	22.25%	26.00%
	RNAFOLD	0.00%	0.02%	0.30%	1.79%	26.01%
RF00024	JUNCTION	14.57%	19.41%	20.70%	22.68%	26.00%
	RNAFOLD	0.00%	0.00%	0.35%	2.24%	26.00%
RF00025	JUNCTION	6.92%	14.17%	16.46%	20.25%	26.00%
	RNAFOLD	0.00%	0.00%	0.29%	2.33%	26.08%

Table 4.9: Comparison of percentages of IQR based scores with genomic backgrounds on 51 datasets [44] (part 2)

		genomic backgrounds (lower is better)				
Rfam ID	Method	≥100	≥10	≥5	≥1.5	≥0
RF00028	JUNCTION	11.57%	17.17%	18.83%	22.10%	26.00%
	RNAFOLD	0.00%	0.03%	0.40%	2.57%	26.13%
RF00029	JUNCTION	2.08%	7.85%	11.10%	16.77%	26.01%
	RNAFOLD	0.00%	0.00%	0.06%	2.21%	26.23%
RF00030	JUNCTION	11.02%	16.96%	19.01%	21.97%	26.01%
	RNAFOLD	0.00%	0.02%	0.35%	2.55%	26.06%
RF00031	JUNCTION	1.16%	6.69%	9.95%	16.00%	26.00%
	RNAFOLD	0.00%	0.00%	0.02%	1.89%	26.31%
RF00033	JUNCTION	1.00%	8.75%	11.25%	16.75%	26.00%
	RNAFOLD	0.00%	0.00%	0.00%	0.50%	26.00%
RF00037	JUNCTION	2.36%	2.36%	2.36%	2.36%	41.03%
	RNAFOLD	0.00%	0.00%	0.08%	2.97%	26.46%
RF00040	JUNCTION	12.17%	16.50%	18.33%	21.67%	26.00%
	RNAFOLD	0.00%	0.00%	0.50%	3.00%	26.00%
RF00054	JUNCTION	1.25%	7.25%	10.25%	16.00%	26.00%
	RNAFOLD	0.00%	0.00%	0.00%	2.13%	26.13%
RF00055	JUNCTION	1.56%	8.11%	10.78%	17.67%	26.00%
	RNAFOLD	0.00%	0.00%	0.00%	1.00%	26.11%
RF00059	JUNCTION	3.18%	9.78%	12.67%	18.17%	26.00%
	RNAFOLD	0.00%	0.00%	0.09%	1.88%	26.13%
RF00066	JUNCTION	1.79%	8.23%	11.38%	17.19%	26.02%
	RNAFOLD	0.00%	0.00%	0.04%	2.47%	26.28%
RF00067	JUNCTION	4.22%	10.17%	13.17%	18.83%	26.00%
	RNAFOLD	0.00%	0.00%	0.17%	2.17%	26.11%
RF00080	JUNCTION	3.44%	10.60%	13.92%	18.92%	26.00%
	RNAFOLD	0.00%	0.00%	0.04%	1.64%	26.36%
RF00096	JUNCTION	4.13%	11.38%	14.35%	18.85%	26.02%
	RNAFOLD	0.00%	0.00%	0.08%	2.02%	26.10%
RF00101	JUNCTION	4.00%	11.69%	14.15%	19.08%	26.00%
	RNAFOLD	0.00%	0.00%	0.00%	1.46%	26.31%
RF00104	JUNCTION	1.36%	7.64%	10.73%	16.73%	26.00%
	RNAFOLD	0.00%	0.00%	0.00%	1.82%	26.27%
RF00114	JUNCTION	3.08%	10.83%	14.21%	19.18%	26.00%
	RNAFOLD	0.00%	0.00%	0.09%	2.05%	26.16%

Table 4.10: Comparison of percentages of IQR based scores with genomic backgrounds on 51 datasets [44] (part 3)

		genomic backgrounds (lower is better)				
Rfam ID	Method	≥100	≥10	≥5	≥1.5	≥0
RF00163	JUNCTION	5.27%	10.78%	13.32%	17.03%	31.84%
	RNAFOLD	0.00%	0.00%	0.01%	2.35%	26.42%
RF00165	JUNCTION	1.57%	6.86%	10.36%	16.64%	26.00%
	RNAFOLD	0.00%	0.00%	0.07%	2.36%	26.07%
RF00167	JUNCTION	2.33%	9.42%	12.40%	18.21%	26.01%
	RNAFOLD	0.00%	0.00%	0.01%	1.61%	26.09%
RF00168	JUNCTION	6.15%	14.11%	16.57%	20.57%	26.00%
	RNAFOLD	0.00%	0.00%	0.21%	2.32%	26.04%
RF00169	JUNCTION	2.67%	9.17%	12.35%	18.04%	26.01%
	RNAFOLD	0.00%	0.00%	0.11%	1.98%	26.22%
RF00170	JUNCTION	0.60%	5.30%	8.60%	15.50%	26.00%
	RNAFOLD	0.00%	0.00%	0.00%	2.20%	26.20%
RF00174	JUNCTION	7.17%	14.42%	17.13%	20.74%	26.00%
	RNAFOLD	0.00%	0.00%	0.28%	2.19%	26.07%
RF00177	JUNCTION	14.40%	19.24%	20.57%	22.74%	26.00%
	RNAFOLD	0.00%	0.02%	0.31%	2.11%	26.01%
RF00206	JUNCTION	1.91%	7.82%	10.77%	17.23%	26.00%
	RNAFOLD	0.00%	0.00%	0.00%	2.45%	26.23%
RF00213	JUNCTION	2.00%	7.05%	9.84%	17.32%	26.05%
	RNAFOLD	0.00%	0.00%	0.05%	1.95%	26.21%
RF00230	JUNCTION	8.15%	14.92%	17.28%	20.74%	26.00%
	RNAFOLD	0.00%	0.00%	0.25%	2.28%	26.05%
RF00234	JUNCTION	7.11%	14.39%	16.72%	20.00%	26.00%
	RNAFOLD	0.00%	0.00%	0.17%	1.72%	26.00%
RF00373	JUNCTION	11.47%	17.29%	19.24%	22.08%	26.00%
	RNAFOLD	0.00%	0.01%	0.31%	2.39%	26.04%
RF00379	JUNCTION	4.44%	11.66%	14.89%	19.54%	26.00%
	RNAFOLD	0.00%	0.01%	0.10%	2.03%	26.15%
RF00380	JUNCTION	5.88%	13.93%	16.19%	20.25%	26.00%
	RNAFOLD	0.00%	0.01%	0.15%	2.04%	26.10%
RF00448	JUNCTION	5.80%	13.00%	17.20%	21.80%	26.00%
	RNAFOLD	0.00%	0.00%	0.20%	1.60%	26.20%
RF00504	JUNCTION	2.53%	8.98%	11.70%	17.45%	26.00%
	RNAFOLD	0.00%	0.00%	0.02%	1.74%	26.32%

JUNCTION and RNAFOLD that is based on the traditional Boltzmann secondary structure ensemble, shows a significant improvement from our approach to identify ncRNA sequences.

We note that there are some exceptional cases without high sensitivities. One possible reason is that their sequences are incomplete. For example, JUNCTION can't capture whole structures of some sequences of Hammerhead ribozyme type I (Hh1) due to the lack of one hairpin loop in their incomplete sequences. Additionally, short sequences, e.g. Iron response element (RF00037), are unlikely to have complex structures and many RNA junctions; this lack of structure may also confuse our program. We also note that the specificities of JUNCTION on most ncRNA datasets are worse than that of RNAFOLD. This problem could be caused by the contribution from the genomic background of previously non-annotated genes. One genomic background with high *inside probability-like weight* was subsequently identified as a non-annotated real ncRNA (Dr.Michael Terns and Dr. Rebecca Terns, personal communication).

Technically, JUNCTION was implemented with a WCFG on an RNA junction based structure space. The constraints were intended to capture the tertiary structure preferences in RNA junctions. The improvement of performance was due to the constrained secondary structure space. Our results show there is a potential for developing reliable programs for effective ncRNA gene finding. However, we may need to develop a new model beyond a context-free grammar for incorporating more tertiary features, some of which are context-sensitive. Although our approach is designed to detect ncRNA genes, it may also detect genes for protein-coding mRNAs, which may also have distinctive structures.

Chapter 5

Theoretical Analysis of Threshold Setting

Based on our method presented in chapter 4, a high *inside probability-like weight* indicates that the corresponding sequence is likely to be a ncRNA. To distinguish ncRNA genes from genomic backgrounds, we need to set a threshold. Sequences with an *inside probability-like weight* above or equal to the threshold are considered to be potential ncRNAs. This chapter theoretically analyzes the threshold setting under different conditions.

GC content and window size may both influence the *inside probability-like weight*. A high GC content gives more chances to form GC pairs, which increases the number of alternative folds. Larger window size also has more alternative folds. Therefore, increasing GC content and window size may potentially increase *inside probability-like weight*. These two factors will be investigated in the rest of this chapter.

Table 5.1: Statistics of *inside probability-like weights* with window size 100 (IQR score threshold is 4)

GC content	1st quartile	3rd quartile	max	min	average	specificity
35%	1.00E+000	2.30E+002	1.43E+004	1.00E+000	3.80E+002	95.00%
40%	6.81E+000	4.80E+002	7.81E+004	1.00E+000	2.65E+003	87.00%
45%	2.06E+001	1.15E+003	4.22E+004	1.00E+000	2.71E+003	85.00%
50%	4.84E+001	5.51E+003	9.71E+004	1.00E+000	6.07E+003	94.00%
55%	3.00E+002	1.35E+004	7.00E+005	1.00E+000	2.93E+004	90.00%
60%	2.47E+003	8.02E+004	6.76E+006	1.00E+000	1.76E+005	91.00%
65%	1.09E+004	2.61E+005	1.66E+007	3.14E+000	4.87E+005	93.00%

5.1 Dataset Preparation

We generated random DNA sequences by a online random DNA sequence generator that was at `http://users-birc.au.dk/biopv/php/fabox/random_sequence_generator.php`. Sequence lengths were set to be 100, 150, 200, and 250. For each length, seven different GC contents 35%, 40%, 45%, 50%, 55%, 60%, and 65% were used. With each sequence length and each GC content, 100 random sequences were generated.

5.2 Threshold Setting

The *inside probability-like weights* of all random sequences were computed by JUNCTION. The statistics of *inside probability-like weights* of random sequences with lengths 100, 150, 200, and 250 are given in tables 5.1, 5.2, 5.3, and 5.4, respectively.

For each length, I set the IQR score thresholds to ensure that the average specificity is above 90%. In this way, the IQR score thresholds is independent of GC content. To reach this goal, the IQR score thresholds of sequence lengths 100, 150, 200, and 250 are 4, 6, 7, and 14, respectively. The corresponding statistics based on different GC contents are given in tables 5.1, 5.2, 5.3, and 5.4.

Table 5.2: Statistics of *inside probability-like weights* with window size 150 (IQR score threshold is 6)

GC content	1st quartile	3rd quartile	max	min	average	specificity
35%	1.68E+001	1.89E+004	4.81E+006	1.00E+000	1.41E+005	83.00%
40%	1.57E+002	2.80E+004	3.41E+006	1.00E+000	9.79E+004	91.00%
45%	1.72E+003	2.22E+005	3.20E+008	1.00E+000	4.83E+006	88.00%
50%	1.72E+004	2.31E+006	2.12E+009	1.00E+000	3.43E+007	89.00%
55%	1.85E+005	3.51E+007	2.28E+009	2.72E+002	6.43E+007	95.00%
60%	1.59E+006	1.57E+008	3.94E+009	4.14E+002	2.71E+008	94.00%
65%	1.07E+007	1.53E+009	1.40E+011	3.76E+003	4.70E+009	93.00%

Table 5.3: Statistics of *inside probability-like weights* with window size 200 (IQR score threshold is 7)

GC content	1st quartile	3rd quartile	max	min	average	specificity
35%	4.83E+002	5.09E+005	2.66E+008	1.00E+000	7.18E+006	88.00%
40%	1.08E+003	4.15E+006	4.58E+009	1.00E+000	1.05E+008	87.00%
45%	1.70E+005	5.62E+007	4.37E+009	2.46E+000	2.10E+008	90.00%
50%	2.84E+006	7.84E+008	9.06E+010	1.00E+000	2.10E+009	93.00%
55%	3.94E+007	1.64E+010	8.69E+011	9.95E+002	4.76E+010	90.00%
60%	1.55E+009	3.29E+011	2.13E+014	1.33E+006	2.79E+012	94.00%
65%	2.89E+010	4.79E+012	1.05E+016	4.73E+006	1.45E+014	91.00%

Table 5.4: Statistics of *inside probability-like weights* with window size 250 (IQR score threshold is 14)

GC content	1st quartile	3rd quartile	max	min	average	specificity
35%	9.82E+003	1.77E+007	1.46E+012	1.00E+000	2.13E+010	85.00%
40%	3.39E+005	6.69E+008	5.21E+011	3.00E+001	1.12E+010	89.00%
45%	1.03E+007	1.28E+010	3.06E+013	2.07E+002	4.21E+011	93.00%
50%	6.79E+008	1.89E+011	3.23E+014	6.08E+002	5.23E+012	92.00%
55%	4.98E+010	7.31E+012	1.36E+016	1.78E+006	3.44E+014	88.00%
60%	9.49E+011	4.60E+014	8.71E+016	3.81E+009	2.82E+015	92.00%
65%	2.54E+013	1.59E+016	1.35E+019	1.43E+007	2.79E+017	91.00%

5.3 Discussion

This chapter analyzes the relationship between IQR score thresholds, GC contents, and window sizes. With fixed IQR score thresholds, standard derivations of specificities with

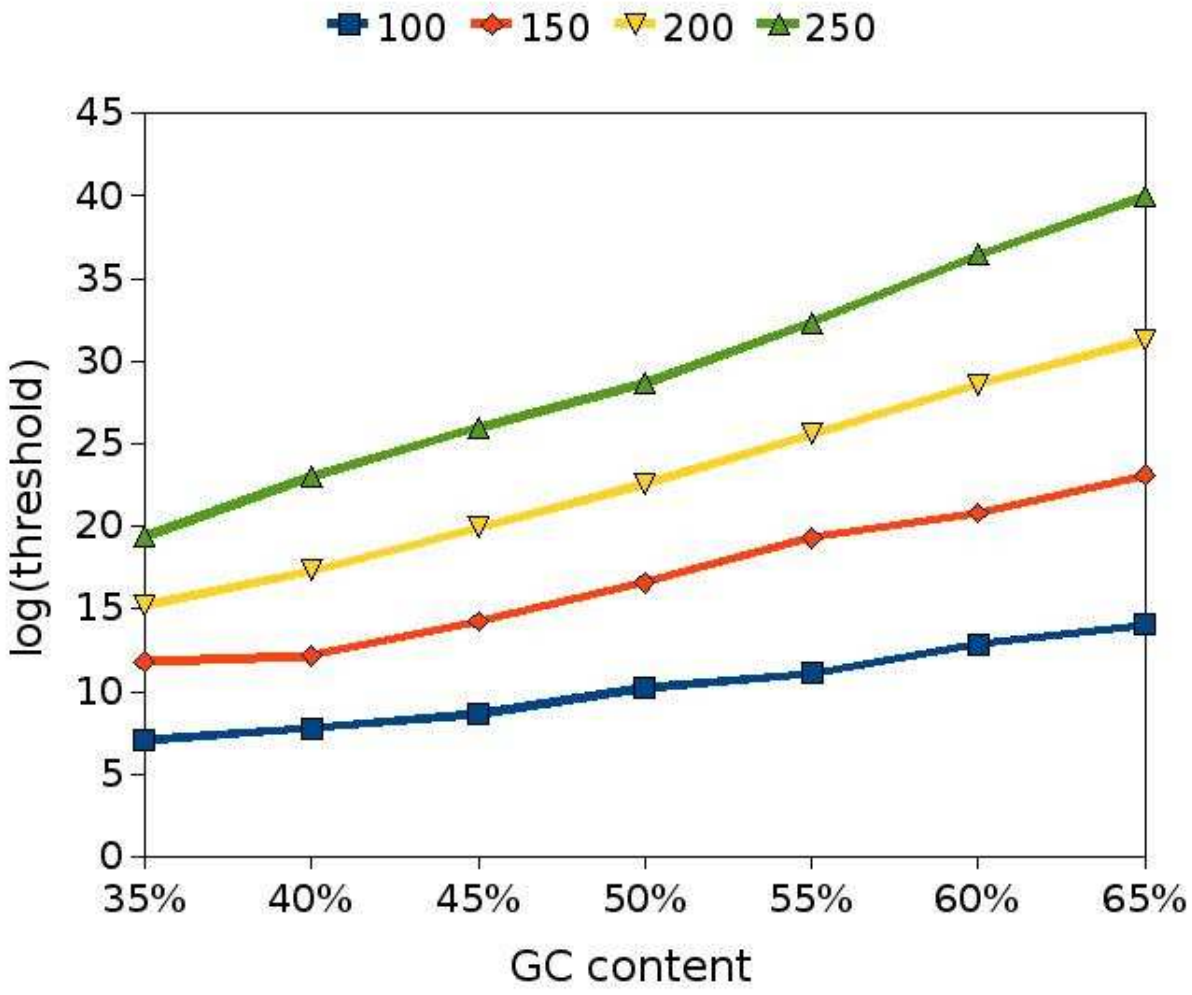

Figure 5.1: Comparison of log(threshold) of *inside probability-like weight*s under different window sizes and GC contents.

window sizes 100, 150, 200, and 250 are 0.0368, 0.0416, 0.0251, and 0.0283, respectively. These statistics suggest that the IQR score threshold is invariant to the GC content. Unlike GC content, the window size shows a strong correlation with the IQR score threshold. A larger window size needs a larger IQR score threshold.

It is also interesting to investigate the trend of thresholds of *inside probability-like weights* under different window sizes and GC contents. Figure 5.1 shows a comparison of the

log(threshold) of *inside probability-like weight*s under different window sizes and GC contents. It also confirms a high threshold of *inside probability-like weights* is correlated with a high GC content and a big window size.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Identifying ncRNAs has been receiving much attention in recent years. This dissertation presented structural methods for detecting ncRNAs against genomic backgrounds based on a constrained RNA secondary structure space.

The mixed success of traditional energy based approaches in different ncRNA species indicates that some structural features have not been captured. To address this problem, the work presented in this dissertation redefined the RNA secondary structure space through extensively investigating RNA native folds. Structural elements favored by tertiary structures were incorporated into the constrained structure space. The novel space is composed of RNA junctions. The constraints on loops are based on the preferences of known RNA tertiary structures, while the constraints on stems are based on the energy model.

Experiments with the novel method based on the novel constrained structure space show encouraging results of detecting structural ncRNAs from genomic backgrounds. These results suggest the potential with further investigations of RNA secondary structure space for designing effective ncRNA gene finding approaches.

6.2 Future Work

In this section, I list some open questions about applying our approach in ab initio ncRNA gene finding.

6.2.1 Improve Performance on Small RNAs

By far, our approach has not performed well on some small RNAs, e.g. the C/D box snoRNAs. Our approach can capture significant structures that usually exist in long ncRNAs, e.g. the performance on vertebrate telomerase RNAs (RF00024) was very good, whose lengths are from 380 to 560. However, short ncRNAs usually have just simple structures. For instance, C/D box snoRNAs are composed of a big hairpin loop with an enclosing stem containing about five base pairs.

To solve this problem, the potential solution could be structural features correlated to specific functions of small ncRNAs. To stabilize structures and perform functions, many small ncRNAs may have special structural motifs. One hope is that in the future having more known tertiary structures of small ncRNAs will provide more structural hints for RNAs with apparently simple structures. In addition, I also believe that small RNAs with significant conserved sequence motifs can be better recognized by sequence profile based search tools (e.g. [33, 51]) used in conjunction with our *ab initio* method.

6.2.2 NcRNA Gene Finding on Real Genomes

To apply our method in ncRNA gene finding on real genomes, a sliding window can be used for screening the given genome sequences. A size fixed sliding window may move forward one nucleotide each time. In this approach, two consecutive sliding windows share a big overlap. This gives us a chance to speed up our method.

The time complexity of our method is $O(mk^3)$ for a model of m nonterminals and rules

and window size k. In the real genome scan, we are able to skip the repeat calculation for the *alpha table* due to the overlap of two consecutive frames of the sliding windows. Once we move the sliding window with one nucleotide, only cells of the *alpha table* related to the last nucleotide need to be updated. Therefore, the time complexity of screening a genome with n nucleotides is $O(mnk^2)$ while the space complexity is still $O(mk^2)$.

Another problem is the window size setting. *Ab initio* methods can't know the lengths of ncRNA genes. If the window size is too small, structures crossing big regions could be missed. If window size is too big, the significance of *inside probability-like weights* of small ncRNA genes could be buried by their upstream and downstream regions. One possible solution to this problem is to scan the genome with multiple sliding window sizes.

In summary, my work provides a useful *ab initio* ncRNA gene finding method. I hope tools based on this method can help biologists to quickly detect ncRNA genes in genomes, especially novel ncRNA genes.

Bibliography

[1] Rt Batey, Rp Rambo, and Ja Doudna. Tertiary Motifs in RNA Structure and Folding. *Angewandte Chemie*, 38(16):2326–2343, August 1999.

[2] Eric Bonnet, Jan Wuyts, Pierre Rouzé, and Yves Van de Peer. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, 20(17):2911–7, November 2004.

[3] JH Chen, SY Le, and B Shapiro. A computational procedure for assessing the significance of RNA secondary structure. *Computational Applications in the Biosciences*, 1990.

[4] Francis Crick. On protein synthesis. *Symposia of the Society for Experimental Biology*, (12):139–163, 1958.

[5] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[6] K Darty and Alain Denise. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 7(2):309–22, 2009.

[7] J M Diamond, D H Turner, and D H Mathews. Thermodynamics of three-way multibranch loops in RNA. *Biochemistry*, 40(23):6971–81, June 2001.

[8] Ye Ding and Charles Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 31(24):7280–7301, December 2003.

[9] RM Dirks, JS Bois, and JM Schaeffer. Thermodynamic analysis of interacting nucleic acid strands. *SIAM Review*, 49(1):65–88, 2007.

[10] Robert M Dirks and Niles a Pierce. An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. *Journal of computational chemistry*, 25(10):1295–1304, July 2004.

[11] Robin D Dowell and Sean R Eddy. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC bioinformatics*, 5:71, June 2004.

[12] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge UK, 1998.

[13] Sean Eddy and Richard Durbin. RNA sequence analysis using covariance models. *Nucleic acids research*, 22(11):2079–2088, 1994.

[14] Sean R. Eddy. Non-coding RNA genes and the modern RNA world. *Nature reviews. Genetics*, 2(12):919–29, December 2001.

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

[16] Eva Freyhult, Paul P Gardner, and Vincent Moulton. A comparison of RNA folding measures. *BMC bioinformatics*, 6:241, January 2005.

[17] Jan Gorodkin and Ivo L. Hofacker. From Structure Prediction to Genomic Screens for Novel Non-Coding RNAs. *PLoS Computational Biology*, 7(8):e1002100, August 2011.

[18] Jan Gorodkin, Ivo L Hofacker, Elfar Torarinsson, Zizhen Yao, Jakob H Havgaard, and Walter L Ruzzo. De novo prediction of structured RNAs from genomic sequences. *Trends in biotechnology*, 28(1):9–19, January 2010.

[19] Sam Griffiths-Jones, Alex Bateman, Mhairi Marshall, Ajay Khanna, and Sean R. Eddy. Rfam: an RNA family database. *Nucleic Acids Research*, 31(1):439–441, January 2003.

[20] Sam Griffiths-Jones, Simon Moxon, Mhairi Marshall, Ajay Khanna, Sean R Eddy, and Alex Bateman. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic acids research*, 33(Database issue):D121–124, January 2005.

[21] Cecilia Guerrier-Takada, Katheleen Gardiner, Terry Marsh, Norman Pace, and Sidney Altman. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3):849–857, 1983.

[22] Donna K Hendrix, Steven E Brenner, and Stephen R Holbrook. RNA structural motifs: building blocks of a modular biomolecule. *Quarterly reviews of biophysics*, 38(3):221–43, August 2005.

[23] I.L. L. Hofacker, W. Fontana, P.F. F. Stadler, L.S. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie / Chemical Monthly*, 125(2):167–188, February 1994.

[24] Ivo L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Research*, 31(13):3429–3431, July 2003.

[25] Z. Huang, R. Malmberg, M. Mohebbi, and L. Cai. RNAv: Non-coding RNA secondary structure variation search via graph Homomorphism. In *Proceedings of Computational Systems Bioinformatics 2010*, pages 56–68, 2010.

[26] Martijn Huynen, Robin Gutell, and Danielle Konings. Assessing the reliability of RNA folding using statistical mechanics. *Journal of molecular biology*, 267(5):1104–12, April 1997.

[27] E.T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4(3):227–241, 1968.

[28] Robert J Klein and Sean R Eddy. RSEARCH: finding homologs of single structured RNA sequences. *BMC bioinformatics*, 4:44, September 2003.

[29] B. Knudsen. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Research*, 31(13):3423–3428, July 2003.

[30] A N Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *G Inst Ital Attuari*, 4(1):83–91, 1933.

[31] Kelly Kruger, Paula J. Grabowski, Arthur J. Zaug, Julie Sands, Daniel E. Gottschling, and Thomas R. Cech. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell*, 31(1):147–57, November 1982.

[32] Christian Laing and Tamar Schlick. Analysis of four-way junctions in RNA structures. *Journal of molecular biology*, 390(3):547–59, July 2009.

[33] Uri Laserson, Hin Hark Gan, and Tamar Schlick. Predicting candidate genomic sequences that correspond to synthetic functional RNA motifs. *Nucleic acids research*, 33(18):6057–69, January 2005.

[34] S Le, JH Chen, and KM Currey. A program for predicting significant RNA secondary structures. *Computer Applications in the Biosciences*, 1988.

[35] Shu-Yun Le, Jih-H Chen, and Jacob V. Maizel. Thermodynamic stability and statistical significance of potential stem-loop structures situated at the frameshift sites of retroviruses. *Nucleic acids research*, 17(15):6143, 1989.

[36] S.Y. Le, J.H. Chen, and J. Maizel. Efficient searches for unusual folding regions in RNA sequences. In R.H. Sarma and M.H. Sarma, editors, *Structure and Methods: Human Genome Initiative and DNA Recombination*, pages 127–136. Adenine Press, 1990.

[37] Aurélie Lescoute and Eric Westhof. Topology of three-way junctions in folded RNAs. *RNA*, 12(1):83–93, January 2006.

[38] Ariane Machado-Lima, Hernando A. del Portillo, and Alan Mitchell Durham. Computational methods in noncoding RNA research. *Journal of mathematical biology*, 56(1-2):15–49, January 2008.

[39] Nicholas R. Markham and Michael Zuker. UNAFold: Software for Nucleic Acid Folding and Hybridization. In *Bioinformatics Methods in Molecular Biology*, chapter 1, pages 3–31. Humana Press Inc, 2008.

[40] Benoît Masquida and Eric Westhof. A Modular and Hierarchical Approach for All-Atom RNA Modeling. In Ray Gesteland, Tom Cech, and John Atkins, editors, *The RNA World*, pages 659–681. Cold Spring Harbor Laboratory Press, 3rd edition, 2006.

[41] David H. Mathews. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, 10(8):1178–1190, 2004.

[42] J S McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–19, 1990.

[43] Vincent Moulton. Tracking down noncoding RNAs. *Proceedings of the National Academy of Sciences of the United States of America*, 102(7):2269, February 2005.

[44] Eric P Nawrocki, Diana L Kolbe, and Sean R Eddy. Infernal 1.0: inference of RNA alignments. *Bioinformatics*, 25(10):1335–1337, May 2009.

[45] Jakob Skou 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 computational biology*, 2(4):e33, April 2006.

[46] Mariusz Popenda, Marta Szachniuk, Marek Blazewicz, Szymon Wasik, Edmund K Burke, Jacek Blazewicz, and Ryszard W Adamiak. RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. *BMC bioinformatics*, 11:231, January 2010.

[47] Elena Rivas and Sean R Eddy. Noncoding RNA gene detection using comparative sequence analysis. *BMC bioinformatics*, 2:8, January 2001.

[48] Elena Rivas and S.R. Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 16(7):583, 2000.

[49] Elena Rivas, Raymond Lang, and Sean R Eddy. A range of complex probabilistic models for RNA secondary structure prediction that include the nearest-neighbor model and more. *RNA*, December 2011.

[50] Yasubumi Sakakibara, Michael Brown, Richard Hughey, I.Saira Mian, Kimmen Sjölander, Rebecca C. Underwood, and David Haussler. Stochastic context-free grammars for tRNA modeling. *Nucleic acids research*, 22(23):5112–20, November 1994.

[51] Raheleh Salari, Cagri Aksay, Emre Karakoc, Peter J Unrau, Iman Hajirasouliha, and S Cenk Sahinalp. smyRNA: a novel Ab initio ncRNA gene finder. *PloS one*, 4(5):e5433, January 2009.

[52] Schrödinger, LLC. The {PyMOL} Molecular Graphics System, Version˜1.3r1. August 2010.

[53] E.A. Schultes, P.T. Hraber, and T.H. LaBean. Estimating the contributions of selection and self-organization in RNA secondary structure. *Journal of molecular evolution*, 49(1):76–83, July 1999.

[54] Sandra Smit and Michael Yarus. Natural selection is not required to explain universal compositional patterns in rRNA secondary structure categories. *RNA*, pages 1–14, 2006.

[55] D Thirumalai. Native secondary structure formation in RNA may be a slave to tertiary folding. *Proceedings of the National Academy of Sciences*, 95(20):11506–11508, 1998.

[56] I. Tinoco, Carlos Bustamante, and Others. How RNA folds. *Journal of molecular biology*, 293(2):271–281, 1999.

[57] D.H. Turner, N. Sugimoto, and S.M. Freier. RNA structure prediction. *Annual Review of Biophysics and Biophysical Chemistry*, 17(1):167–192, 1988.

[58] Douglas H. Turner, Naoki Sugimoto, Ryszard Kierzek, and Scott D. Dreiker. Free energy increments for hydrogen bonds in nucleic acid base pairs. *Journal of the American Chemical Society*, 109(12):3783–3785, 1987.

[59] 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*, 7:173, January 2006.

[60] Harm van Bakel, Corey Nislow, Benjamin J Blencowe, and Timothy R Hughes. Most "dark matter" transcripts are associated with known genes. *PLoS biology*, 8(5):e1000371, May 2010.

[61] Amy E. Walter, Douglas H. Turner, James Kim, Matthew H. Lyttle, Peter Muller, David H. Mathews, and Michael Zuker. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proceedings of National Academy of Sciences*, 91(20):9218–22, September 1994.

[62] Yue Wan, Michael Kertesz, Robert C. Spitale, Eran Segal, and Howard Y. Chang. Understanding the transcriptome through RNA structure. *Nature Reviews Genetics*, 12(9):641–655, August 2011.

[63] Yingfeng Wang, Amir Manzour, Pooya Shareghi, Timothy I. Shaw, Ying-Wai Li, Russell L. Malmberg, and Liming Cai. Stable stem enabled shannon entropies distinguish non-coding RNAs from random backgrounds. In *The Proceedings of IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, pages 184–189. Ieee, February 2011.

[64] Yingfeng Wang, Amir Manzour, Pooya Shareghi, Timothy I Shaw, Ying-Wai Li, Russell L Malmberg, and Liming Cai. Stable stem enabled Shannon entropies distinguish non-coding RNAs from random backgrounds. *BMC Bioinformatics*, 13(Suppl 5):S1, 2012.

[65] Stefan Washietl, Ivo L Hofacker, Melanie Lukasser, Alexander Hüttenhofer, and Peter F Stadler. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nature biotechnology*, 23(11):1383–90, November 2005.

[66] Stefan Washietl, Ivo L Hofacker, and Peter F Stadler. Fast and reliable prediction of noncoding RNAs. *Proceedings of the National Academy of Sciences of the United States of America*, 102(7):2454–9, February 2005.

[67] Zasha Weinberg and Ronald R Breaker. R2R–software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC bioinformatics*, 12(1):3, January 2011.

[68] Christopher Workman and Anders Krogh. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic acids research*, 27(24):4816–22, December 1999.

[69] Augix Guohua Xu, Liu He, Zhongshan Li, Ying Xu, Mingfeng Li, Xing Fu, Zheng Yan, Yuan Yuan, Corinna Menzel, Na Li, Mehmet Somel, Hao Hu, Wei Chen, Svante Pääbo, and Philipp Khaitovich. Intergenic and repeat transcription in human, chimpanzee and macaque brains measured by RNA-Seq. *PLoS computational biology*, 6(7):e1000843, January 2010.

[70] Michael Zuker. Calculating nucleic acid secondary structure. *Current opinion in structural biology*, 10(3):303–310, June 2000.

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