STARS

University of Central Florida
STARS

Electronic Theses and Dissertations, 2004-2019

2016

Computational Methods for Comparative Non-coding RNA Analysis: from Secondary Structures to Tertiary Structures

Ping Ge
University of Central Florida

 Part of the Computer Engineering Commons
Find similar works at: https://stars.library.ucf.edu/etd
University of Central Florida Libraries http://library.ucf.edu

This Doctoral Dissertation (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Electronic Theses and Dissertations, 2004-2019 by an authorized administrator of STARS. For more information, please contact STARS@ucf.edu.

STARS Citation
Ge, Ping, "Computational Methods for Comparative Non-coding RNA Analysis: from Secondary Structures to Tertiary Structures" (2016). *Electronic Theses and Dissertations, 2004-2019*. 4958.
https://stars.library.ucf.edu/etd/4958

COMPUTATIONAL METHODS FOR COMPARATIVE NON-CODING RNA ANALYSIS: FROM SECONDARY STRUCTURES TO TERTIARY STRUCTURES

by

Ping Ge
B.S. Wuhan University, 2002
M.S. Huazhong University of Science and Technology, 2005
M.S. University of Central Florida, 2009

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Engineering
in the Department of Electrical Engineering and Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Spring Term
2016

Major Professor: Shaojie Zhang

© 2016 Ping Ge

ABSTRACT

Unlike message RNAs (mRNAs) whose information is encoded in the primary sequences, the cellular roles of non-coding RNAs (ncRNAs) originate from the structures. Therefore studying the structural conservation in ncRNAs is important to yield an in-depth understanding of their functionalities. In the past years, many computational methods have been proposed to analyze the common structural patterns in ncRNAs using comparative methods. However, the RNA structural comparison is not a trivial task, and the existing approaches still have numerous issues in efficiency and accuracy. In this dissertation, we will introduce a suite of novel computational tools that extend the classic models for ncRNA secondary and tertiary structure comparisons.

For RNA secondary structure analysis, we first developed a computational tool, named PhyloRNAalifold, to integrate the phylogenetic information into the consensus structural folding. The underlying idea of this algorithm is that the importance of a co-varying mutation should be determined by its position on the phylogenetic tree. By assigning high scores to the critical covariances, the prediction of RNA secondary structure can be more accurate. Besides structure prediction, we also developed a computational tool, named ProbeAlign, to improve the efficiency of genome-wide ncRNA screening by using high-throughput RNA structural probing data. It treats the chemical reactivities embedded in the probing information as pairing attributes of the searching targets. This approach can avoid the time-consuming

base pair matching in the secondary structure alignment. The application of ProbeAlign to the FragSeq datasets shows its capability of genome-wide ncRNAs analysis.

For RNA tertiary structure analysis, we first developed a computational tool, named STAR3D, to find the global conservation in RNA 3D structures. STAR3D aims at finding the consensus of stacks by using 2D topology and 3D geometry together. Then, the loop regions can be ordered and aligned according to their relative positions in the consensus. This stack-guided alignment method adopts the divide-and-conquer strategy into RNA 3D structural alignment, which has improved its efficiency dramatically. Furthermore, we also have clustered all loop regions in non-redundant RNA 3D structures to *de novo* detect plausible RNA structural motifs. The computational pipeline, named RNAMSC, was extended to handle large-scale PDB datasets, and solid downstream analysis was performed to ensure the clustering results are valid and easily to be applied to further research. The final results contain many interesting variations of known motifs, such as GNAA tetraloop, kink-turn, sarcin-ricin and t-loops. We also discovered novel functional motifs that conserved in a wide range of ncRNAs, including ribosomal RNA, sgRNA, SRP RNA, GlmS riboswitch and twister ribozyme.

For my parents.

ACKNOWLEDGMENTS

First and foremost, I would like to thank Dr. Shaojie Zhang for his support and encouragement throughout my Ph.D. study. I am very grateful to work with him because his persistence in facts and passion in work change my way of doing research. In addition, I would also like to thank him for his wise suggestions and meticulous supervision, which are essential to make my academic achievement possible.

I also wish to thank Dr. Ratan Guha, Dr. Sumit Jha, Dr. Kenneth Stanley, and Dr. Hojun Song for serving on my dissertation committee and reviewing my dissertation.

Finally, I want to thank my family for their patience and support along the way.

In the dissertation, Chapter 2, in part, is a reprint of the paper "Incorporating phylogenetic-based covarying mutations into RNAalifold for RNA consensus structure prediction", co-authored with Shaojie Zhang in BMC bioinformatics 14, no. 1 (2013): 142. The dissertation author was the primary investigator and author of the paper.

Chapter 3, in part, is a reprint of the paper "ProbeAlign: incorporating high-throughput sequencing-based structure probing information into ncRNA homology search", co-authored with Cuncong Zhong, and Shaojie Zhang in BMC bioinformatics 15, no. Suppl 9 (2014): S15. The dissertation author was the primary investigator and author of the paper.

Chapter 4, in part, is a reprint of the paper "STAR3D: a stack-based RNA 3D structural alignment tool", co-authored with Shaojie Zhang in Nucleic acids research 43, no. 20 (2015): e137-e137. The dissertation author was the primary investigator and author of the paper.

Chapter 5 is adapted from the material in submission, "De novo discovery of structural motifs in RNA 3D structures through clustering", co-authored with Shahidul Islam, Cuncong Zhong, and Shaojie Zhang. The dissertation author was the primary investigator and author of the paper.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

The classic central dogma of molecular biology explains that the genetic information encoded in deoxyribonucleic acid (DNA) flows in the biological system by transcribing to messenger ribonucleic acid (mRNA), and then translating into protein. This important mechanism ensures the building of body structure and the heredity to offspring. As a result, mRNAs, as well as proteins, attract most of the focus in the study of molecular biology. Recently, more and more research indicates that the non-coding RNAs (ncRNAs) are also important participators in the biological system [38, 110, 147]. One of the most well-known ncRNAs is the transfer RNA (tRNA), which helps to decode the codons of mRNAs in ribosome [127]. MicroRNA (miRNA) is another important type of ncRNA found in plant, animals and viruses [22]. It forms RNA-induced silencing complex to conduct the post-transcriptional regulation of gene expression by binding to the reverse complementary in untranslated regions (UTRs) of mRNAs. Long non-coding RNA (lncRNA) refers to the non-protein-coding transcripts larger than 200 nucleotides. Recent studies show that the mutations and the dysregulations of lncRNAs are closely related to many human diseases [43, 59, 120]. What's more, with the rapid development of next generation sequencing technique, substantial genome-wide datasets have been analyzed to annotate the transcriptomes. The results show that actually only a small fraction of transcripts are protein-coding, while most of the non-protein-coding transcripts are functional [153, 154]. Therefore, fully understanding any biological system is impossible without the thorough research on the ncRNAs in it.

Unlike mRNAs whose primary sequences contain all codes for protein synthesis and thus can be applied for function prediction directly, the functions of ncRNAs are determined by their high-order structures, including the secondary structures and tertiary structures (three-dimensional structures). Some well-known instances are the cloverleaf-like structure of tRNAs and the kink-turn structural motifs which serve as important sites for protein recognition. This specific feature posts challenges for the computational methods designed for the function inference of ncRNAs. Generally, there are two major issues: first, we need to predict the plausible secondary structures of ncRNAs; second, the comparison among ncRNAs must be primarily based on their structures, which may increase running time greatly.

Many computational methods have been proposed to solve the RNA structure folding problem. One type of algorithms makes use of Minimal Free Energy (MFE) model on a single sequence [74, 126, 191]. The negative stabilizing base stacking energies and the positive destabilizing loop energies for each possible secondary structure are computed and then summed up based on the experimentally measured parameters. By using dynamic programming, the most stable structure with the minimum free energy is chosen. However, the prediction accuracy of this approach is limited, especially for long ncRNA sequences [63]. What's more, single sequence folding can not be applied to discover new ncRNA families, because the statistical signals in the secondary structure of a ncRNA are not strong enough to distinguish itself from the stable structures folding from random sequences [128, 176]. To solve the problem, comparative methods are incorporated into the folding algorithms. The basic idea is that the secondary structures of ncRNAs are conserved through the evolution, and thus the consensus structure of RNA sequences in one family should be more accurate and significant. The first consensus folding algorithm was proposed by Sankoff [133]. It

computes the alignment of two RNAs and their consensus structure simultaneously, since constructing a precise structural alignment for folding is also a challenging problem. Excessive computational resources are required by this algorithm [$O(n^6)$]. Many other structural alignment methods try to limit the search space of the Sankoff algorithm [65, 68, 168], but they are still too computationally expensive to be applied to genome-wide datasets [$O(n^4)$]. Now, the most widely used consensus folding methods align the RNA sequences without considering their secondary structures, and then detect the conserved signals in the sequence alignment to infer the consensus secondary structure [73, 88].

On the other hand, comparing ncRNAs and their structures is the prerequisite for the functional annotation. However, as we have discussed, the existing structural alignment approaches are not efficient enough for the genome-wide annotation of known ncRNA families. Some approaches incorporate filters into homology search to remove the transcripts that lack strong sequence conservation with the query before the structural alignment [7, 49, 87]. For example, the most recent release of the widely used software package CMsearch adopts a pipeline consisting of five different filters which have significantly improved the computational efficiency of its previous version [112]. However, it still would take about 3 hours to annotate the 1Gbp chicken genome with known Rfam [19] families on a 100-CPU cluster with all filters and MPI applied. The sensitivity of CMsearch reaches a plateau about 87% without filters, indicating intrinsic difficulty in genome-wide ncRNA detection. What's more, ncRNAs in cells may fold significantly differently *in vivo* from their unconstrained *in silico* predictions.

Although the computational methods of RNA secondary structural alignment provide excellent starting point for the analysis of functional conservation, they are incapable of identifying the tertiary motifs and homologous 3D structures. To study the tertiary structures of ncR-

NAs, 3D structural alignment, which can take into full account the tertiary interactions, have been proposed. Traditionally, the alignment of 3D coordinates, including in proteins and in RNAs, is formulated as finding a rigid transformation that superimposes the largest number of atoms in one structure onto the atoms of the other structure with a predefined distance error [3]. The exact solution for this problem is impractical since the time complexity is $O(n^{32.5})$ [4]. Some approximate algorithms have been proposed [2, 55]. However, considering that the lengths of ncRNAs can be thousands of nucleotides, those methods are still not efficient enough for real applications. Therefore, most of the RNA 3D structural alignment methods simplify the model by combining the local similarity or integrating specific features of RNAs. For examples, R3D Align [124] assembles the alignments of local neighbor atoms and ARTS [35] uses base pairs as seeds for the global alignment. Although an increasing number of tools have been developed for the alignments of RNA 3D structures [75, 91, 161], there is still lacking an efficient and accurate solution for the problem.

What's more, the RNA 3D structures are actually interlinked by recurrent subcomponents, named RNA structural motifs, which may play important functional roles. Based on the existing knowledge about some motifs, many computational tools have been designed to search their homologies by using comparative approaches. However, in recent years, the traditional biological methods of detecting novel structural motifs can not keep the pace with rapidly increasing number of resolved RNA 3D structures. Therefore, it is an urgency to *de novo* discover new RNA structural motifs with computational tools directly. Several clustering pipelines, such as COMPADRES [160], LENCS [30] and RNA 3D Motif Atlas [122], have been proposed to categorize the conserved structural elements together. Although they have successfully suggested many potential novel motifs, the restriction of finding rigid 3D geometric similarity makes them impossible to consider complex structural motifs with base

pairing variations. Unlike these methods, RNAMSC evaluates the alignment of two motifs by using their base pairing patterns. Although the advantage of this strategy has already been demonstrated by clustering three rRNAs, it still needs to be extended to handle the large-scale dataset in the PDB database.

We are particularly interested in these existing problems of comparative methods for RNA structure analysis, such as consensus folding, secondary structural alignment, tertiary structural alignment, and the discovery of RNA structural motifs. The major issue of consensus folding comes from the underlying multiple sequence alignments, which treat all inputs equally. For secondary and tertiary structural alignments, the difficulty is the complicated comparison among the base pairs in RNAs. The existing approaches to improving their efficiency always result in the loss of accuracy. Last but not least, a clustering pipeline for detecting RNA structural motifs in the PDB database is required to study the relationship between RNA structures and their functionalities.

1.1 PhyloRNAalifold

RNAalifold is a popular computational method for RNA consensus structure prediction which incorporates covarying mutations into a thermodynamic model to fold the aligned RNA sequences. When quantifying covariance, it evaluates conserved signals of two aligned columns with base-pairing rules. This scoring scheme performs better than some other approaches, such as mutual information. However it ignores the phylogenetic history of the aligned sequences, which is an important criterion to evaluate the level of sequence covariance.

In order to improve the accuracy of consensus structure folding, we propose a novel approach named PhyloRNAalifold. It incorporates the number of covarying mutations on the phylogenetic tree of the aligned sequences into the covariance scoring of RNAalifold. The benchmarking results show that the new scoring scheme of PhyloRNAalifold can improve the consensus structure detection of RNAalifold.

In conclusion, incorporating additional phylogenetic information of aligned sequences into the covariance scoring of RNAalifold can improve its performance of consensus structures folding. This improvement is correlated with alignment characteristics, such as pair-wise identity and the number of sequences in the alignment.

1.2 ProbeAlign

Recent advances in RNA structure probing technologies, including the ones based on high-throughput sequencing, have improved the accuracy of thermodynamic folding with quantitative nucleotide-resolution structural information. We present a novel approach, ProbeAlign, to incorporate the reactivities from high-throughput RNA structure probing into ncRNA homology search for functional annotation. To reduce the overhead of structure alignment on large-scale data, the specific pairing patterns in the query sequences are ignored. On the other hand, the partial structural information of the target sequences embedded in probing data is retrieved to guide the alignment. Thus the structure alignment problem is transformed into a sequence alignment problem with additional reactivity information. The benchmark results show that the prediction accuracy of ProbeAlign outperforms filter-based CMsearch with high computational efficiency. The application of ProbeAlign to the FragSeq

data, which is based on genome-wide structure probing, has demonstrated its capability to search ncRNAs in a large-scale dataset from high-throughput sequencing.

In conclusion, by incorporating high-throughput sequencing-based structure probing information, ProbeAlign can improve the accuracy and efficiency of ncRNA homology search. It is a promising tool for ncRNA functional annotation on genome-wide datasets.

1.3 STAR3D

The various roles of versatile non-coding RNAs typically require the attainment of complex high- order structures. Therefore, comparing the 3D structures of RNA molecules can yield in-depth understanding of their function conservation and evolutionary history. Recently, many powerful tools have been developed to align the RNA 3D structures. Although some methods rely on both backbone conformations and base pairing interactions, none of them considers the entire hierarchical formation of the RNA secondary structure. One of the major reason for this problem is that applying the algorithms of matching the tree-like topology to the 3D coordinates directly is particularly time- consuming. In this article, we propose a novel RNA 3D structural alignment method named STAR3D to take into full account the stack relationship without complicated 2D structural alignment. It adopts a two-step strategy, which includes detection of the consensus of stacks and guided alignment of loops. The matching between 3D conserved stacks in the inputs is identified by joining small building components, and then combined into a tree-like consensus of secondary structures. After that, the unaligned loop regions are compared one- to-one in accordance with their relative positions in the common tree. To evaluate the performance of STAR3D, we tested it on

some RNAs in PDB with known 3D geometric information. The experimental results show that the prediction of STAR3D is highly accurate for both non-homologous and homologous RNAs. In addition, it is more efficient than the state-of-the-art tools by at least tens of orders of magnitude.

1.4 RNA Structural Motif Clustering

As recurrent components of three-dimensional conformations and functional roles in biological systems, the RNA structural motifs provide us an easy way to associate molecular architectures with their cellular mechanisms. In the past years, some computational tools have been developed to search motif instances by using the existing knowledge of well-studied families. Recently, with the rapidly increasing number of resolved RNA 3D structures, there is an urgency of discovering novel motifs with the newly presented information. In this work, we classify all the loops in non-redundant RNA 3D structures to de novo detect plausible RNA structural motif families by using a clustering pipeline. Compared with other clustering approaches, our method has two benefits: First, the underlying alignment algorithm is highly sensitive to the variations in 3D structures; Second, sophisticated downstream analysis has been performed to ensure the clusters are valid and easily applied to further research. The final results of the clustering contain many interesting variants of known motif families, such as GNAA tetraloop, kink turn, sarcin-ricin and T-loop. We also discover potential functional motifs that conserved in ribosomal RNA, sgRNA, SRP RNA, riboswitch, and ribozyme.

1.5 Overview of the Dissertation

In summary, we presented a suite of computational methods that target to solve the central problems of ncRNA structure analysis and function annotation. For the study of RNA secondary structures, we have developed computational methods to improve the consensus folding algorithm and the RNA secondary structural alignment algorithm. For the study of RNA tertiary structures, We have also developed a stack-based 3D structural alignment tool and an automated clustering pipeline for discovering RNA structural motifs. In Chapter 2, 3, 4 and 5, all four methods will be described in detail. Their basic mechanisms are summarized in the following.

1. PhyloRNAalifold is designed to incorporate phylogenetic information into the consensus folding of RNA secondary structures, as described in Chapter 2.

2. ProbeAlign aims at making use of the pairing attributes revealed by probing data to avoid the complex computation of secondary structure alignment without loss of accuracy, as shown in Chapter 3.

3. STAR3D is developed to align RNA 3D structures with the guide of stack configuration, as discussed in Chapter 4.

4. The RNA structural motif clustering pipeline is proposed to detect novel motif families, as discussed in Chapter 5.

The first three tools and the clustering results are available at the website of Computational Biology and Bioinformatics Group in the University of Central Florida (http://www.genome.ucf.edu/).

It is anticipated that our tools can improve the ncRNA structure analysis and function annotation in the future.

CHAPTER 2: RNA CONSENSUS STRUCTURE PREDICTION WITH PHYLOGENETIC-BASED COVARYING MUTATIONS

2.1 Background

The discovery of novel non-coding RNA (ncRNA) families expanded our understanding of RNAs, which not only carry genetic codes for protein synthesis but also participate in other functions, especially the regulatory processes, such as localization, replication, translation and degradation [38, 69, 107, 109]. In mammals, a substantial amount of transcripts (above 90%) are non-protein-coding, and most of them are functional [14, 154]. What's more, the non-coding regions in the human genome are crucially important. For example, microRNA (miRNA) is used as a marker to differ normal tissues from tumors [27, 42, 114]; long non-coding RNA (lncRNA) also contributes to human disease etiology [110]. These findings fuel the research of RNA and also pose new challenges.

Unlike protein-coding genes, whose primary sequences can be applied for accurate functional prediction with statistical signals, RNAs' functions depend on their secondary structures. Many computational methods have been proposed to fold RNA structures. One type of popular algorithms adopts Minimum Free Energy (MFE) model to fold a single RNA sequence, which has been implemented in Mfold [192] and RNAfold [74]. However, the structure prediction accuracy of this approach is limited. One major reason is that the precise energy

parameters are hard to obtain experimentally [80]; on the other hand, the functional RNA structure may not be the one with the minimum energy [107]. What's more, single sequence folding may not be applied to discover new RNA families even if the predicted structures are correct, because the statistical signals in an RNA secondary structure are not strong enough to distinguish itself from the stable structures folding from random sequences [128, 176].

Comparative methods can solve these problems by folding a consensus structure from multiple sequences, which not only improve the structure prediction accuracy, but also provide additional signals to discover novel RNAs [171]. The idea of this approach is that RNA secondary structures are conserved through evolution. Therefore, a consensus structure detected by comparing related RNA sequences should be more accurate and significant than the structure folded from a single sequence. With a consistent consensus structure, the specific structure of each sequence in the alignment can be obtained by constraint folding. A classic comparative method is the Sankoff algorithm [133]. Because constructing a precise structural alignment of RNA sequences is also a challenging problem, the Sankoff algorithm computes alignment and fold structure simultaneously. Excessive computational resources ($O(n^6)$) are required by the Sankoff algorithm for a large-scale problem. Some implementations of this approach, such as Dynalign [65], Foldalign [68], LocARNA [168] and Conan [33], attempt to restrict its solution space by limiting the number of possible sub-structures. However these methods are still computationally expensive ($O(n^4)$).

To reduce the computation complexity, comparative methods may align related sequences first and then detect conserved signals in the alignment to infer a consensus structure. One type of these methods extends the energy-based model from single sequences to alignments. Based on the assumption that high covariance of two aligned columns implies the conservation of pairing, all potential pairing columns in an alignment can be determined. After that

the optimal consensus structure with minimum average free energy can be folded just as a single sequence structure. An example of covariance scoring scheme is Mutual Information (MI), which can measure the dependence of two columns in the alignments [25, 60, 61]. RNAalifold [73] adopts the basic idea of MI scoring and imports the pairing rules of RNA into the measurement of covariance. Another type of comparative methods is evolution-based. In these methods, no thermodynamic parameters but statistical learning algorithms are used. The evolutionary history of the aligned sequences is reformed with probability theories [39, 121], and the RNA secondary structures are modeled as stochastic context-free grammar (SCFG) [88, 89, 132]. Both strategies have their own strengths and weaknesses [31]. Some methods try to integrate both approaches. For example, PETfold extends Pfold [89], an evolution-based algorithm, to consider the energetically favorable base-pairs [137]. However, PETfold utilizes a Nussinov style model [117], which does not make full use of the energy parameters. RNAalifold also tested two other covariance scoring schemes to incorporate evolutionary information [13, 73], but neither of them yielded a better result.

In this article, we propose a novel method called PhyloRNAalifold. It improves RNAalifold by explicitly incorporating the phylogenetic tree of the aligned sequences into the computation of covariance scores. Like RNAalifold, PhyloRNAalifold detects pairing columns by evaluating covarying mutations and folds RNA structures through an MFE model. Unlike RNAalifold, which does not consider the relative positions of sequences in the phylogeny, PhyloRNAalifold counts the number of covarying mutations on the phylogenetic tree for each pair of columns with a parsimony approach. What's more, comparing with PETfold, PhyloRNAalifold retains the Turner's model [190] in RNAalifold, which describes RNA structures with many thermodynamic parameters derived from physical studies. With the supports of

both energy-based model and evolution-based model, PhyloRNAalifold may detect consensus structures more precisely.

The rest of the article is organized as follows: in the methods section, we discuss the basic mechanism of RNAalifold, its shortcomings, and details of the PhyloRNAalifold algorithm. In the results section, we describe the benchmark datasets, experimental results, and the effect of parameters and alignment characteristics on our algorithm. In the discussion and conclusion section, we summarize our existing works and propose directions for future research.

2.2 Materials and Methods

2.2.1 Consensus folding energy and covariance score in RNAalifold

The basic approach of RNAalifold [73] is to integrate covarying mutation into the thermodynamic model to predict consensus structures. First, covariance scores are computed for all pairs of columns to determine possible pairing positions in the consensus structure. Then, based on the MFE model, the minimum average folding energy is computed with dynamic programming. Assume the given alignment is denoted by $\mathbb{A}$, which contains N sequences $\mathbb{A} = \{s^1, s^2, \ldots, s^N\}$. Each sequence contains L symbols, including nucleotides and gaps, and s_i^k represents the i^{th} symbol ($1 \leq i \leq L$) at the k^{th} ($1 \leq k \leq N$) RNA sequence. The

minimization of free energy is computed by using the following recursive functions:

$$
\begin{aligned}
F_{i,j} &= \min(F_{i+1,j}, \min_{i<k\leq j}(C_{i,k} + F_{k+1,j})) \\
C_{i,j} &= \phi_2\gamma_{i,j} + \min \begin{cases} \sum_{s^k\in\mathbb{A}} H(i,j,s^k) \\ \min_{i<p<q<j} \left(\sum_{s^k\in\mathbb{A}} I(i,j,p,q,s^k) + C_{p,q} \right) \\ \min_{i<p<j} (FM_{i,p} + FM1_{p+1,j} + M_a) \end{cases} \\
FM_{i,j} &= \min \begin{cases} FM_{i+1,j} + M_c \\ \min_{i<p<j} C_{i,p} + FM_{p+1,j} + M_b \\ FM1_{i,j} \end{cases} \\
FM1_{i,j} &= \min\left(FM1_{i,j-1} + M_c, C_{i,k}\right)
\end{aligned}
\tag{2.1}
$$

where $F_{i,j}, C_{i,j}, FM_{i,j}, FM1_{i,j}$ denote the minimum free energies for the region between i^{th} column and j^{th} column with unconstrained structure, with enclosed structure, with a multi-loop, and with a multi-loop containing a single branch, respectively. $H(i,j,s)$ is the free energy for a hairpin loop enclosed by s_i and s_j, and $I(i,j,p,q,s)$ is the free energy for an internal loop containing two base-pairs, one is between s_i and s_j and the other is between s_p and s_q. M_a, M_c are penalties for closing bases and non-pairing bases in multi-loops. M_b is the bonus for branch bases in multi-loops.

The recursive functions were derived from the Turner's model [190]. One major change made by RNAalifold for consensus folding is the usage of covariance score γ. It is not only a factor in the computing of free energy, but also determines the possible pairing columns in the alignment. Two parts, one is bonus and the other is penalty, are in this score. The

first part of the covariance score is called the conservation score. For (s_i^k, s_j^k) and (s_i^l, s_j^l), three levels of confidence for pairing are assessed: base-pairs without mutation, base-pairs with one mutation, and base-pairs with two mutations. In the latest version of Vienna RNA package (v2.0) [102], the recursive function for computing conservation score is:

$$V_{i,j} = \frac{1}{N} \sum_{1 \le k < l \le N} \begin{cases} h(s_i^k, s_i^l) + h(s_j^k, s_j^l) & \text{if } (s_i^k, s_j^k) \in \mathbb{B} \text{ and } (s_i^l, s_j^l) \in \mathbb{B} \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

where $h(x, y)$ is the Hamming distance between base x and base y, and $\mathbb{B}$ ={'AU', 'UA', 'CG', 'GC', 'GU', 'UG'} is the set of all possible base-pairs. The second part is the penalty score $Q_{i,j}$, which deals with a pair of symbols that cannot form a base-pair:

$$Q_{i,j} = \sum_{1 \le k \le N} \begin{cases} 0 & \text{if } (s_i^k, s_j^k) \in \mathbb{B} \\ 0.25 & \text{if } s_i^k \text{ and } s_j^k \text{ are gaps} \\ 1 & \text{otherwise} \end{cases} \tag{2.3}$$

Overall, the covariance score is:

$$\gamma_{i,j} = V_{i,j} - \phi_1 \times Q_{i,j} \tag{2.4}$$

where $\phi_1 = \phi_2 = 1$. A threshold value $\gamma_t = -2$ is defined for $\gamma_{i,j}$. If $\gamma_{i,j} > \gamma_t$, i^{th} column and j^{th} column are considered to be pairing columns. In the final output, the minimum average folding energy, including the covariance score, is normalized by dividing N.

2.2.2 Phylogenetic-based covarying mutation

RNAalifold incorporates covarying mutations into consensus folding to improve the detection of pairing columns. From Equation (2), it can be seen that RNAalifold counts the level of covariance by treating all sequences equally and try all possible combinations of base-pairs. In short, RNAalifold models the relationship of sequences as a complete graph. As a result, the specific evolutionary relationship among sequences in the phylogenetic history is ignored. Take the RNA structural alignment in Figure 2.2.2 as an example. The red and green columns achieve the same covariance score (2) in RNAalifold. However, as described in [58], the conservation evidence in Figure 2.2.2(c) is stronger than that in Figure 2.2.2(b) because at least two mutations occur at the green columns while only one is required to form the red ones.

PhyloRNAalifold models the relationship of aligned sequences as a tree by introducing the phylogenetic history of the alignment into the computation of covariance scores. The level of structural conservation is measured by the number of covarying mutations on the tree. Our assumption is that more covarying mutations on the tree mean stronger evidence of conservation. In addition, PhyloRNAalifold does not discard the original scoring scheme of RNAalifold, because experimental results showed this scheme can infer significant RNA structural aspects with high sensitivity and selectivity [162]. Assume $m_{i,j}$ covarying mutations occur between i^{th} and j^{th} columns on the alignment $\mathbb{A}$'s phylogenetic tree and the number of base-pairs on those columns is $b_{i,j}$. The value of $m_{i,j}$ depends on the size of the alignment. Since our approach focuses on improving the bonus part of the covariance scores, the number of covarying mutations is normalized with its upper bound: $\frac{m_{i,j}}{b_{i,j}-1}$. A new factor

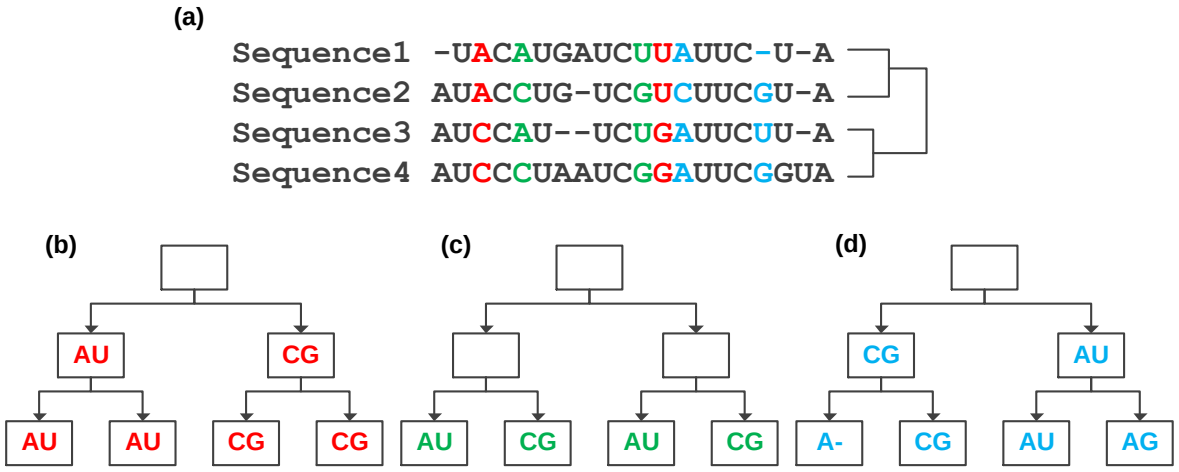

Figure 2.1: Covarying mutations in an RNA alignment. (a) A multiple RNA alignment and its phylogenetic tree. Three pairs of columns, which are marked with different colors, are analyzed in the following three sub-figures. (b) Possible covarying mutations in the red columns. In this case, only one pair-wise mutation is required at the root node. (c) Possible covarying mutations in the green columns. At least two pair-wise mutations occur at the internal nodes in this case; (d) Possible covarying mutations in the blue columns. There are non-pairing bases, 'AG' and 'A-'. The label inference of the internal nodes does not depend on them. So in this case, the number of mutations is one.

for the conservation score is proposed:

$$\epsilon_{i,j} = 1 + \beta \times \frac{m_{i,j}}{b_{i,j} - 1} \tag{2.5}$$

where β is the scale parameter for the normalized covarying mutation numbers. PhyloRNAalifold computes covariance scores with the following formula:

$$\gamma^p_{i,j} = \epsilon_{i,j} \times V_{i,j} - \phi_1 \times Q_{i,j} \tag{2.6}$$

All the other parameters and their default values in RNAalifold are retained. Due to the fact that $\gamma^p_{i,j} \geq \gamma_{i,j}$ ($\epsilon_{i,j} \geq 1$), two columns would be marked as pairing in PhyloRNAalifold if their covariance score in RNAalifold is greater than the threshold γ_t (the default value of γ_t is -2). Thus the advantage of PhyloRNAalifold is to import more potentially pairing positions with high mutation numbers.

2.2.3 Computing the number of covarying mutations

Given a phylogenetic tree and labels at its leaves, the Fitch algorithm can optimize nucleotide assignment of the internal nodes to minimize the number of mutations [48]. If we model solving phylogeny as a maximum parsimony problem, this number can be taken as the actual number of mutations. The Fitch algorithm consists of a forward phase and a backward phase. In the forward phase, all possible labels at each internal node are inferred. In addition, the number of mutations is estimated during a bottom-up traversal. In the backward phase, a top-down pass is performed to find the optimal label at each internal node. Only the forward algorithm is applied to PhyloRNAalifold, since we do not need the exact labels at the internal nodes, but only the number of mutations on the tree. Without loss of generality, we require $\mathbb{T}$ to be a rooted binary tree. r denotes the root of $\mathbb{T}$ and v, v_l, v_r denote a node, left child of v, and right child of v respectively. $F(v)$ is the set of possible labels at node v, and $cost(v)$ is the number of mutations on the sub-tree which is rooted at v. Then the

forward phase can be described with the following recursive functions:

$$
\begin{aligned}
F(v) &= \begin{cases} F(v_l) \cap F(v_r) & \text{if } F(v_l) \cap F(v_r) \neq \emptyset \\ F(v_l) \cup F(v_r) & \text{otherwise} \end{cases} \\
cost(v) &= \begin{cases} cost(v_l) + cost(v_r) & \text{if } F(v_l) \cap F(v_r) \neq \emptyset \\ cost(v_l) + cost(v_r) + 1 & \text{otherwise} \end{cases}
\end{aligned}
\tag{2.7}
$$

For each leaf, $F(v)$ is a base at the corresponding sequence. After the computation is finished, $cost(r)$ shows the minimum number of mutations on the phylogenetic tree. The optimization of this algorithm was proved in [164].

In Equation 2.5, the computation of $\epsilon_{i,j}$ does not depend on non-pairing bases. Therefore, in the revised Fitch algorithm non-pairing bases need not to be considered when the number of covarying mutations is computed. We changed the original Fitch algorithm in two ways: (1) at any leaf node, if $(s_i^k, s_j^k) \notin \mathbb{B}$, set $(s_i^k, s_j^k) =$ ('-', '-'); (2) for one internal node v, if the bases at $v_l(v_r)$ is ('-', '-'), v will obtain $F(v_r)(F(v_l))$ as its label. One example of this algorithm is shown in Figure 2.2.2(d). The revised Fitch algorithm can be described by using

the following functions.

$$
F(v) = \begin{cases} F(v_l) \cap F(v_r) & \text{if } F(v_l) \cap F(v_r) \neq \emptyset \text{ and } F(v_l) \neq (\text{`-'}, \text{`-'}) \text{ and } F(v_r) \neq (\text{`-'}, \text{`-'}) \\ F(v_l) & \text{if } F(v_r) = (\text{`-'}, \text{`-'}) \\ F(v_r) & \text{if } F(v_l) = (\text{`-'}, \text{`-'}) \\ F(v_l) \cup F(v_r) & \text{otherwise} \end{cases}
$$

$$
cost(v) = \begin{cases} cost(v_l) + cost(v_r) & \text{if } F(v_l) \cap F(v_r) \neq \emptyset \text{ and } F(v_l) \neq (\text{`-'}, \text{`-'}) \text{ and } F(v_r) \neq (\text{`-'}, \text{`-'}) \\ cost(v_l) & \text{if } F(v_r) = (\text{`-'}, \text{`-'}) \\ cost(v_r) & \text{if } F(v_l) = (\text{`-'}, \text{`-'}) \\ cost(v_l) + cost(v_r) + 1 & \text{otherwise} \end{cases} \tag{2.8}
$$

It is easy to see that our algorithm is optimal, because it only excludes non-pairing bases from the computation of the original Fitch algorithm.

In PhyloRNAalifold, the tree structure is an input variable and the clients can use any phylogenetic tree construction algorithm to build it. The time complexity of the original RNAalifold algorithm is $O(m \times n^2 + n^3)$ [177], where n is the length of the alignment and m is the number of sequences in the alignment. The extra computation in PhyloRNAalifold is caused by the revised Fitch algorithm, whose time complexity ranges from $O(\log m)$ to $O(m)$. In addition, PhyloRNAalifold needs to compute $\epsilon_{i,j}$ for each pair of columns in the alignment. Thus the overall time consumption of the revised Fitch algorithm is $O(\log m \times n^2)$ or $O(m \times n^2)$. Neither of them increases the time complexity of RNAalifold.

2.3 Results

2.3.1 Benchmarking datasets

The 19 Rfam [51] families used in the CMfinder paper [179] were selected as our first benchmarking dataset. It captures the diversity of known families by excluding highly conserved ones. Other programs, such as PETfold [137] and RNAalifold [13], also adopted it in their experiments. All the testing families came from Rfam version 11.0 and their seed alignments were used. In order to evaluate the dependence of our folding algorithm on the alignment quality, we also realigned the seeds with ClustalW [92] to generate the second benchmarking dataset. For this dataset, the predicted structure of the first sequence in each alignment was compared with its real consensus structure to measure the accuracy. Pair-wise identity and the number of sequences in an alignment are two important alignment characteristics. Pair-wise identity is related to the performance of consensus structure folding, while the number of members is important for inferring an accurate evolutionary history. To analyze these two factors, we generated the third benchmarking dataset which consisted of alignments with different number of sequences and identities. Member sequences were randomly picked from each seed alignment. For each family, we generated three sets. Each set included 50 alignments, and the alignments contained 5 sequences, 10 sequences and 20 sequences respectively. 7 families (ctRNA_pGA1, glmS, lin-4, Lysine, mir-10, s2m, Tymo_tRNA-like), whose sequences are less than 50, were excluded from this dataset because the diversity of generated alignments was too small.

PhyloRNAalifold requires a phylogenetic tree of the alignment to infer the consensus structural aspects. In our experiments, DNADIST and KITSCH in the PHYLIP package [44] were

used to generate phylogenies. First DNADIST computed a distance matrix of the sequences. After that, KITSCH estimated a phylogenetic tree from the output matrix of DNADIST. The reason of using KITSCH was that it can generate rooted binary trees, which were required by PhyloRNAalifold. Another notable issue in this process is that if two sequences differ in more than 75% of their positions, DNADIST would set the distance between them to -1, which represents infinity. KITSCH rejects negative distances. Thus -1 was replaced with 1000 in distance matrices. We have checked all the positive sequence distances in our experiments. None of them exceeded 10, so 1000 is large enough to represent infinity.

The implementation of PhyloRNAalifold was on the top of program RNAalifold in the Vienna RNA package 2.0.7 [102]. The major change made by PhyloRNAalifold is to incorporate our Fitch module into the scoring scheme of RNAalifold. To test our idea, Matthews correlation coefficient (MCC) [52] was used to measure the accuracy of consensus structure prediction. Its definition is:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.9}$$

where TP, TN, FP, FN represent the number of true positives, true negatives, false positives, and false negatives, respectively. Additional predicted base-pairs that are not in the reference structure fall into two categories. Some base-pairs contradict reference, the others are compatible with it. Compatible base-pairs can be inserted into the known consensus structure, while adding contradictory base pairs breaks the pairing rules. Only contradictory base-pairs were counted as false positive predictions.

Five algorithms, RNAalifold, RIBOSUM-based RNAalifold, PhyloRNAalifold, RIBOSUM-based PhyloRNAalifold, and PETfold, have been tested on the first two datasets. The first

four have also been benchmarked on the third dataset. The RIBOSUM scoring scheme [87] is used in the latest version of Vienna RNA package. In this scheme, the sum of Hamming distance $h(s_i^k, s_i^l) + h(s_j^k, s_j^l)$ in conservation score was replaced by an entry in the RIBOSUM matrix R: $R(s_i^k s_j^k, s_i^l s_j^l)$. The experiment results in [13] showed that RIBOSUM-based RNAalifold outperformed the orignal RNAalifold in most of cases. In addition, the authors of [13] used $\phi_1 = 0.6$ and $\phi_2 = 0.5$ as the default parameters in their experiments. We applied their settings to make the comparison fair. The performance of PETfold was tested in our experiments too. We used the web-server of PETfold [138] and its default parameters.

2.3.2 Effect of parameter β

In the first experiment, we compared the structure prediction results of PhyloRNAalifold with RNAalifold on the original CMfinder dataset. Default values for ϕ_1, ϕ_2 and γ_t were used, and β was a variable ranging from 1 to 15. Figure 2.3.2 shows that the novel scoring scheme of PhyloRNAalifold improves the performance of RNAalifold in nearly all cases, except $\beta = 1$. The differences of average MCC in both cases, with or without using RIBOSUM matrix, become larger when β is increased, and they are maximized at $\beta = 10$. The largest differences are 0.079 and 0.033 for RIBOSUM supported and non-RIBOSUM supported algorithms. After that, the performance of PhyloRNAalifold is not boosted with the increasing of β. In the following experiments, we select 10 as a default value for β.

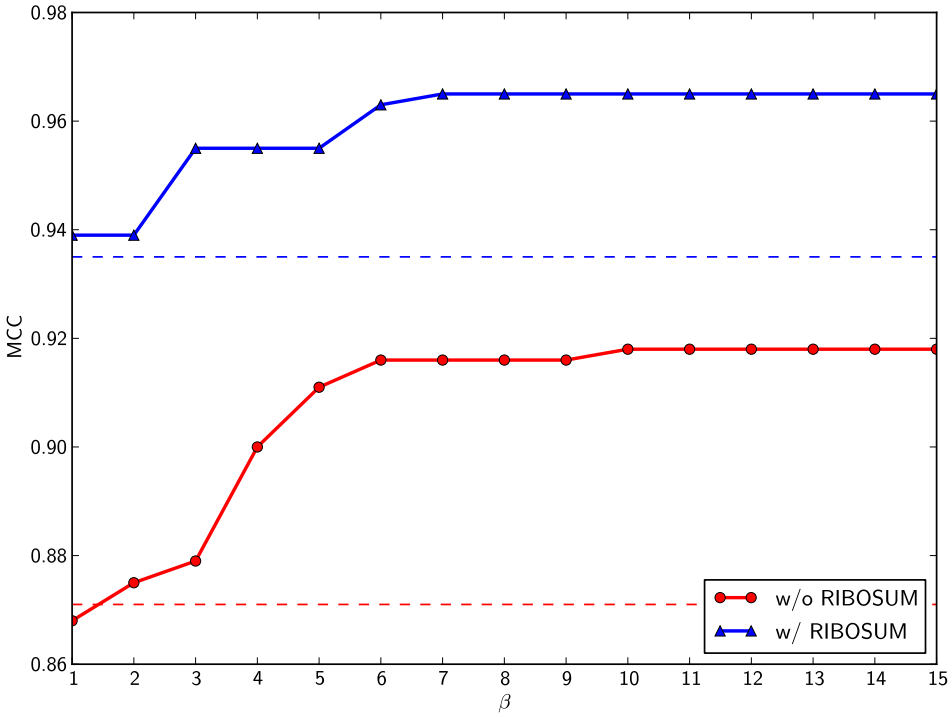

Figure 2.2: MCC on the CMfinder dataset as a function of the β parameter. The MCC results of PhyloRNAalifold, with and without RIBOSUM matrix support, are shown in this figure. The dash lines are references for the curves, which show the performance of RNAalifold on the same dataset. It can be seen that except for $\beta = 1$, the new phylogenetic-based covariance scoring scheme improves the performance of RNAalifold.

2.3.3 Benchmarking with other methods

Table 2.1 summarizes all results of the consensus structure predictions on the structural alignments. PhyloRNAalifold with RISOSUM support achieves the best average MCC score. When RIBOSUM matrix is incorporated, the score difference between PhyloRNAalifold and RNAalifold becomes smaller. This may suggest that by using RIBOSUM matrix, RNAalifold quantifies the conservation among base-pairs more precisely than its original solution. The advantage of using phylogenetic history is swamped by this strategy to some extent. The average specificity scores of five algorithms are the same, while PhyloRNAalifold and PET-

fold have two largest average sensitivity scores. It is evidence of evolutionary information helping the energy-based folding algorithms to detect more base-pairs with introducing very few errors. An interesting observation is that Cobalamin has relative low MCC scores for RNAalifold. PhyloRNAalifold improves the accuracy of the consensus structure prediction of Cobalamin greatly. In addition, PETfold has the best performance on this family among all five algorithms. We checked the consensus structure of Cobalamin and the predicted results of RNAalifold. One possible reason is that there are too many gaps and non-pairing bases at its pairing columns, which decrease the covariance scores of those columns in RNAalifold greatly. Without the bonus from evolutionary information, those columns cannot be detected by RNAalifold at all.

Table 2.2 shows the results on the non-structural alignments of the CMfinder dataset. In this case, RIBOSUM-based PhyloRNAalifold still achieves the highest average MCC score. The performance of RNAalifold with RIBOSUM support is almost the same as that of the top one algorithm. What's more, PETfold, which has a larger average MCC score in the previous experiment than RIBOSUM-based RNAalifold, falls to third place. This suggests that the evolutionary information at the pairing columns may be disrupted by ClustalW, whose alignment function does not consider secondary structures.

2.3.4 Effects of identity and the number of sequences

In this experiment, we try to analyze the correlation of two alignment characteristics, pairwise identity and the number of sequences, with the performance of PhyloRNAalifold. Figure 3 shows the experiment results. It can be seen that all four algorithms have similar

Table 2.1: The benchmarking results on the structural alignments of the CMfinder dataset ($\beta = 10$). The MCC on structural alignments of the CMfinder dataset is compared among PhyloRNAalifold, RNAalifold and PETfold. The parameter β of PhyloRNAalifold is 10. Best performance on the same family is set to **bold**.

Family	#seq	MPI	RNAalifold	PhyloRNAalifold	RNAalifold with RIBOSUM	PhyloRNAalifold with RIBOSUM	PETFold
Cobalamin	430	49.7	0.756	0.951	0.591	0.951	**0.976**
ctRNA_pGA1	15	73.0	0.979	0.979	0.979	0.979	**1.000**
Entero_CRE	56	81.7	0.912	0.912	0.916	0.916	**1.000**
Entero_OriR	60	87.4	0.47	0.681	0.703	0.703	**0.747**
glmS	18	57.4	0.973	0.987	**1.000**	**1.000**	0.987
Histone3	52	46.0	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Intron_gpII	98	52.3	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
IRE	62	76.8	0.814	0.854	**0.965**	**0.965**	**0.965**
let-7	67	66.4	0.861	0.967	**1.000**	**1.000**	0.915
lin-4	12	68.8	0.977	**1.000**	**1.000**	**1.000**	0.836
Lysine	47	48.4	0.952	**0.981**	**0.981**	**0.981**	0.952
mir-10	36	67.9	0.789	0.865	**0.957**	**0.957**	0.935
Purine	133	54.7	0.904	**1.000**	**1.000**	**1.000**	0.977
RFN	144	68.1	0.826	0.851	0.826	**1.000**	**1.000**
Rhino_CRE	12	81.4	0.581	0.581	**0.976**	**0.976**	0.750
S_box	433	62.9	0.883	0.924	0.963	**1.000**	0.944
s2m	38	78.3	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
SECIS	61	41.0	0.972	**1.000**	**1.000**	**1.000**	0.972
Tymo_tRNA-like	28	68.2	0.908	0.908	0.910	0.910	**0.975**
		Mean	0.871	0.918	0.935	**0.965**	0.944
		Specificity	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		Sensitivity	0.802	0.881	0.919	**0.952**	0.922

performance when the number of sequences in the alignments is 5. With the increasing of the members in the alignments, the average MCC difference between PhyloRNAalifold and RNAalifold becomes larger. Using more sequences provides a more precise phylogenetic history, so it is reasonable that PhyloRNAalifold achieves its best performance on alignments with 20 sequences. In addition, for experiments on the alignments of 10 sequences and 20 sequences, the maximum performance difference exists between the pair-wise identities 65 and 80. If the pair-wise identity of an alignment is small, the original covariance scoring scheme of RNAalifold works well enough because a large number of different base-pairs at the pairing columns provide substantial conservational signals. On the other hand, when

Table 2.2: The benchmarking results on the non-structural alignments of the CMfinder dataset ($\beta = 10$). The MCC on non-structural alignments of the CMfinder dataset is compared between PhyloRNAalifold, RNAalifold and PETfold. The parameter β of PhyloRNAalifold is 10. Best performance on the same family is set to **bold**.

Family	MPI	RNAalifold	PhyloRNAalifold	RNAalifold with RIBOSUM	PhyloRNAalifold with RIBOSUM	PETFold
Cobalamin	43.2	**-0.001**	**-0.001**	-0.002	-0.002	-0.002
ctRNA_pGA1	66.5	0.865	0.889	**0.979**	**0.979**	0.936
Entero_CRE	81.7	0.912	0.912	0.916	0.916	**1.000**
Entero_OriR	87.5	0.694	0.830	**0.965**	**0.965**	0.910
glmS	55.2	0.445	0.566	**0.873**	0.784	0.811
Histone3	45.1	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Intron_gpII	46.2	0.760	0.794	**0.826**	**0.826**	0.794
IRE	77.3	0.480	0.480	0.710	0.710	**0.816**
let-7	66.7	0.760	**0.761**	0.666	0.666	0.742
lin-4	64.6	0.523	0.521	0.712	0.712	**0.739**
Lysine	44.0	0.307	0.414	**0.513**	0.484	0.388
mir-10	68.4	0.741	0.820	**0.957**	**0.957**	0.935
Purine	53.8	0.852	**0.977**	**0.977**	**0.977**	0.953
RFN	64.2	0.342	0.309	0.302	0.433	**0.477**
Rhino_CRE	81.4	0.581	0.581	**0.976**	**0.976**	0.750
S_box	56.5	0.430	0.430	0.817	**0.860**	0.750
s2m	78.3	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
SECIS	36.5	**0.000**	**0.000**	**0.000**	**0.000**	-0.003
Tymo_tRNA-like	64.1	0.691	0.703	**0.768**	**0.768**	0.596
Mean		0.599	0.631	0.735	**0.737**	0.715
Specificity		0.947	0.947	0.947	0.947	**0.999**
Sensitivity		0.486	0.545	0.689	**0.704**	0.655

the alignment's pair-wise identity is too large, all the symbols at the pairing columns are almost the same. The effect of our new covariance scoring scheme is reduced due to the lack of evolutionary information.

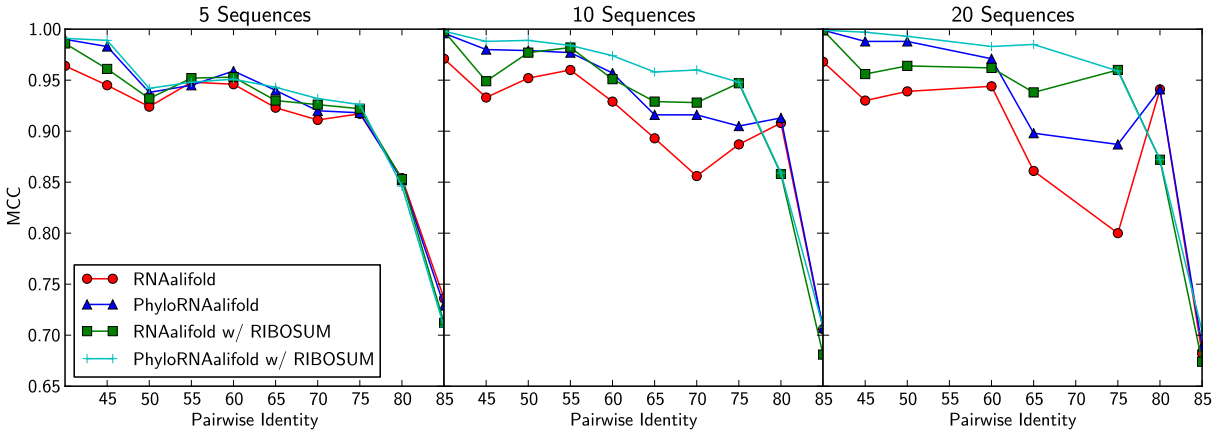

Figure 2.3: The effect of alignment pair-wise identity and sequence number on the structural prediction of PhyloRNAalifold ($\beta = 10$). The MCC results of PhyloRNAalifold and RNAalifold on the third benchmarking dataset are shown in this figure. It can be seen that the performance difference between PhyloRNAalifold and RNAalifold increases with the increasing of the sequence number in the alignments. In addition, the maximum MCC difference is achieved in the range of $65 \sim 75$ identitiess.

2.4 Discussion and Conclusion

We have proposed a novel approach, PhyloRNAalifold, to fold RNA consensus structures by evaluating the level of conservation in aligned RNA sequences. With an evolution-based covariance scoring scheme, PhyloRNAalifold can detect more potential pairing columns than RNAalifold. The benchmark testing shows that PhyloRNAalifold can improve the performance of RNAalifold, as well as PETfold.

There are two possible directions for further research. The first direction is to analyze the dependence of PhyloRNAalifold on the phylogenetic tree construction algorithms. Tree structures have great effect on the RNA structure prediction of PhyloRNAalifold. Besides DNADIST and KITSCH, there are other algorithms, such as UPGMA [145], PAUP [149] and MrBayes [77], which can construct alternative trees. Finding or design an optimal algorithm for detecting the phylogenetic information in the pairing columns is an open question. Ideally, structure conservation should be considered because it is crucial for evaluating the similarity between two RNA sequences. The second direction deals with incorporating the phylogenetic information of non-pairing bases into the folding algorithm. Only covarying mutations among base-pairs are considered in PhyloRNAalifold. In probabilistic methods, all the possible mutations, including mutations in loop regions and stack regions, are modeled with HMM. Our goal is to incorporate both types of mutations into PhyloRNAalifold, while still keep the simplicity of the scoring scheme.

CHAPTER 3: RNA HOMOLOGY SEARCH WITH STRUCTURE PROBING INFORMATION

3.1 Background

The non-coding RNAs (ncRNAs) play various vital roles in the biological systems [38, 110, 147], such as gene-expression regulation [157], catalysis [32], signal recognition [62], and ribosomal RNA modification [23]. Given the facts that most of the human genome (approximately 62% [14] to 93% [15]) is transcribed [174] while only a small fraction of it (about 3%) actually codes for proteins, it is tempting to hypothesize that the ncRNAs contribute enormously to the complex and elegant regulatory networks in human and other multicellular organisms. Therefore, fully understanding any biological system is impossible without the thorough research on the ncRNAs in it. However, annotating ncRNAs is more difficult than proteins, because ncRNAs with divergent sequences may fold into conserved secondary structures, and still perform similar biological functions. In this sense, secondary structure conservation is used as a better evidence for functional conservation than sequence similarity when conducting comparative ncRNA analysis.

Annotating the ncRNA secondary structure is a prerequisite for comparative ncRNA structural analysis. However, determining ncRNA secondary structure is a non-trivial task. The performance of the existing computational methods (such as mfold [191], RNAfold [74], and

RNAstructure [126]) for predicting secondary structure from a single ncRNA sequence is not satisfying, especially for long ncRNA sequences [63]. Although the prediction accuracy can be improved with evolutionary information from multiple sequence alignments [53, 56, 72, 121, 163], such information is not always available for every genome of interest. On the other hand, genome-wide annotation of known ncRNA families by homology search still appears as an open problem for lacking efficient and accurate computational pipelines. For example, the latest release of the widely used software CMsearch has significantly improved the computational efficiency of its previous versions [112]. However, it still would take about 3 hours to annotate the 1 Gbp chicken genome with known Rfam [19] families on a 100-CPU cluster even with filters and MPI applied [112]. The sensitivity of CMsearch reaches a plateau at $\sim$87% without filters, indicating intrinsic difficulty in detecting ncRNAs with diverse sequences. The difficulty of ncRNA annotation is partly due to the computational overhead of structure alignment, and partly due to the low information content given by the secondary structures [128].

Recent advances in massive parallel sequencing make genome-wide probing of ncRNA secondary structures possible. Examples of technologies in this category include, but not limited to, PARS [84], FragSeq [159], and SHAPE-seq [103] (SHAPE-seq has not been applied in genome-wide study). With further improvements, such techniques are becoming more powerful for understanding the *in vitro* [99, 185], or even *in vivo* ncRNA structrome [29]. The information received from a typical genome-wide ncRNA secondary structure probing experiment is the *reactivity* for each site. As the probing reagents, such as 1M7 [28, 103, 167], DMS [29], or nuclease [159], are chosen to preferentially attack the paired/unpaired regions, the experimentally determined reactivity can be used to assess the probability of whether a specific site is paired. The reactivities can then be transformed into pseudo-

energy terms [28, 181], and be incorporated into existing RNA-folding algorithms to predict the optimal secondary structure that is compatible with both the RNA free energy models and the observed reactivity pattern. When the reactivity information derived from SHAPE technology [167] was incorporated, the 16s rRNA structure prediction accuracy was lifted from 47% to 97% [28]. This success implies great potential in using the structure probing information in other comparative genome-wide ncRNA analysis approaches.

Therefore, it is possible to improve the ncRNA annotation by incorporating the high-throughput RNA secondary structure probing information. First, the computational efficiency can be promoted by only focusing on transcribed regions revealed by the read-mapping pattern as used in standard RNA-seq experiments. In addition, the experimentally defined structural information can be used to reduce the search space of the alignment algorithms and lead to the development of a more efficient one. Meanwhile, we can also expect to improve the annotation accuracy because the experimentally determined structural information reflects the real RNA structures, and is much more accurate than the *in silico* predictions.

Here, we present a novel ncRNA annotation algorithm called ProbeAlign, which, to the best of our knowledge, is the first algorithm that considers high-throughput RNA structure probing information for the purpose of genome-wide ncRNA annotation. To make ProbeAlign feasible for large-scale sequencing data, the specific pairing relationships between bases in the query structures are discarded to achieve $O(n^2)$ time complexity. On the other hand, with the usage of structure probing data, the partial structural aspects of target sequences are introduced into the algorithm. Therefore, ProbeAlign tackles the homology search problem from another perspective with the support of new technologies. The benchmark results show that the prediction of ProbeAlign outperforms filter-based CMsearch with a shorter running time. Last but not least, the application of ProbeAlign to FragSeq data, which was

generated by high-throughput sequencing-based RNA structure probing technology, shows its capability of analyzing genome-wide datasets.

The rest of the paper is organized as follows: in the Methods section, we discuss the core algorithm of ProbeAlign and how to estimate the p-values for the alignment scores. In the Results section, we describe benchmark results, parameters optimization and an application of our algorithm to FragSeq data. In the Discussion section, we summarize our existing works, and propose possible directions for future research.

3.2 Materials and Methods

3.2.1 Algorithm design

ProbeAlign identifies the homologous ncRNAs in a profile-based search manner. The profile is generated by using the multiple sequence alignment of a given ncRNA family. The aligned columns formed by a majority of gap are excluded in the search profile. In addition, the consensus structure of the family is considered as the structural information of the profile. The targets of search are usually the genomic or transcriptomic sequences with experimentally determined reactivities. In the latest implementation of ProbeAlign, higher reactivity of a site indicates higher chance of being unpaired, and vice versa.

Assume the alphabet of RNA sequences is {A, C, G, U, X}, in which X represents all unknown nucleotides. First we denote the query of an ncRNA family as $Q = \{P, S\}$, where P is the sequence profile of the family and S is the pairing pattern in the corresponding consensus

structure. Let the length of the profile be n, then $P = \langle p_1, p_2, \ldots, p_n \rangle$ and $S = \langle s_1, s_2, \ldots, s_n \rangle$. Here, $p_i = [v_i^A, v_i^C, v_i^G, v_i^U, v_i^X, v_i^-]$, which is a vector that contains the frequency of the nucleotides and gap at site i. s_i is a boolean value indicating whether site i is paired in the consensus structure or not (0 means i is paired and vice versa). Note that the specific pairing relationship between sites in P is not considered in S, which is similar to the folded-BLAST [156]. For target T of length m, denote $B = \langle b_1, b_2, \ldots, b_m \rangle$ as the genomic sequence and $R = \langle r_1, r_2, \ldots, r_m \rangle$ as the observed reactivities. Denote $D_{i,j}, I_{i,j}, M_{i,j}$ as the optimal alignment scores for deleting, inserting and matching a column in the search profile, respectively. They can be computed using the following recursive functions:

$$\begin{aligned} D_{i,j} &= max\{M_{i-1,j} + \epsilon_0 + \epsilon_e, D_{i-1,j} + \epsilon_e\}, \\ I_{i,j} &= max\{M_{i,j-1} + \epsilon_0 + \epsilon_e, I_{i,j-1} + \epsilon_e\}, \\ M_{i,j} &= max\{D_{i,j}, I_{i,j}, M_{i-1,j-1} + \alpha \times \tau(s_i, r_j) + \beta \times \sigma(p_i, b_j)\}. \end{aligned} \tag{3.1}$$

Here, ϵ_0 and ϵ_e are the gap open penalty and the gap extension penalty, respectively. In our implementation, a "semi-global alignment" setting [70] is adopted. Therefore, these three functions are initialized with: $M_{0,0} = 0$, $M_{i,0} = \epsilon_0 + i \times \epsilon_e$, $M_{0,j} = 0$, and $D_{0,j} = I_{i,0} = -\infty$. τ and σ are functions to assess the structural and the sequence similarities between the queries and the targets, respectively. α and β are weights assigned to these two types of similarities.

The sequence similarity between the query profile and the target sequence is computed using the following formula:

$$\sigma(p_i, b_j) = \sum_{x \in \{\text{A, C, G, U, X, }-\}} v_i^x \times \text{m}(x, b_j), \tag{3.2}$$

where $\text{m}(x, y)$ is the substitution score between nucleotides x and y.

The general function to compute structural similarity is as follows:

$$\tau(s_i, r_j) = \begin{cases} 0 & \text{if } r_j \text{ is not defined,} \\ f(s_i, r_j) & \text{otherwise.} \end{cases} \tag{3.3}$$

Given the reactivity r_j, $p(\pi_j|r_j)$ is computed to compare the structural aspect of b_j with s_i. π_j is a random variable and $\pi_j \in \{0, 1\}$, 0 means b_j is paired and 1 means b_j is not paired. According to the Bayes' theorem, the probability can be computed as:

$$p(\pi_j|r_j) = \frac{p(r_j|\pi_j) \times p(\pi_j)}{\sum_{\pi_j} p(r_j|\pi_j) \times p(\pi_j)}. \tag{3.4}$$

The probabilities $p(r|\pi = 0)$ and $p(r|\pi = 1)$ can be inferred from the reactivities retrieved from the RNAs with known secondary structures [148]. To simplify the computation, we assume $p(\pi = 0)$ is equal to $p(\pi = 1)$ and then define the function f as:

$$\begin{aligned} f(s_i, r_j) &= \log p(\pi_j = s_i|r_j) - \log p(\pi_j \neq s_i|r_j) \\ &= \log p(r_j|\pi_j = s_i) - \log p(r_j|\pi_j \neq s_i). \end{aligned} \tag{3.5}$$

Note that the probability $p(r|\pi)$ varies among different probing techniques. Even for one protocol, the reactivity distributions may be different due to the distinct computational methods for transferring the chemical signals from biological experiments. Therefore, it may be hard to apply Equation 5 on some techniques whose statistical properties have not been studied yet. To overcome this limitation and make the implementation of ProbeAlign easy to use, a simplified scoring function is proposed:

$$f(s_i, r_j) = \begin{cases} 1 & \text{if } (r_j > r_c \text{ and } s_i = 1) \text{ or } (r_j < r_c \text{ and } s_i = 0), \\ -1 & \text{otherwise.} \end{cases} \tag{3.6}$$

In Equation 6, r_c is a cutoff value which is used to annotate the structural aspects of targets. Any site that has higher reactivity than r_c is deemed as unpaired, and vice versa ($r_j > r_c \Rightarrow p(\pi_j = 1|r_j) = 1; r_j <= r_c \Rightarrow p(\pi_j = 0|r_j) = 1$). We have compared two different types of structural similarity functions by taking SHAPE protocol as an example. The benchmark results show that the optimal performance of those two functions is comparable. Therefore, the simplified scoring function is practical for universal usage, while the protocol-specific scoring function may be a better option if the reactivity distribution is known.

The above described dynamic programming algorithm computes the optimal alignment between the query profile and the target sequence with the consideration of both structural and sequence similarity. After alignment, traceback is performed to check the base pairing consistency between the query structure and the target. Bonus scores are assigned to the possible pairing bases. Such additional information can be used to detect potential false positive hits that have high alignment scores but low structural consistency with the query. For example, in Figure 3.1, two alignment scores are the same. However, the target in alignment 1 is more conserved with the query, compared with the target in alignment 2, because the red letters can form canonical pairs, while the green ones can not. Structure consistency scores can help us distinguish these two targets.

3.2.2 P-value estimation

A robust scheme for evaluating the statistical significance of prediction results is important for the homology search applications. However, what statistical distribution the ncRNA alignment scores should follow is still unclear. In this case, we simulated the ProbeAlign

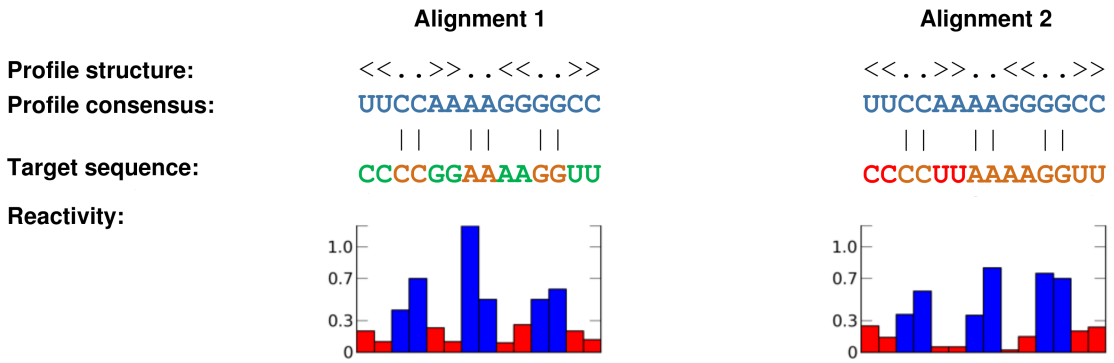

Figure 3.1: Two alignments with different structure consistency scores. The reactivities with red color are higher than r_c, while the reactivites with blue color are less than r_c. In Alignment 2, the red letters can not form canonical base pairs. The green letters in Alignment 1 can form canonical base pairs.

scores by searching 106 Rfam families (as defined by the Infernal RMARK3 dataset [112]) against five artificial sequences, whose GC content ranging from 30% to 70%. Each artificial sequence was constructed by concatenating random RNA sequences generated by GenRGenS [123] with a simple context-free grammar [37]. The secondary structure of each random sequence was predicted by mfold [191]. The corresponding reactivities of the secondary structure were simulated based on the SHAPE technology [148].

We fitted the ProbeAlign score density for each Rfam family with four different distributions: the normal, Gumbel, GEV (Generalized Extreme Value), and Gamma distributions. The goodness of fitting was measured with K-S test (Kolmogorov-Smirnov test). The fitting results on the five artificial sequences show that the ProbeAlign scores follow the Gamma distribution for most of Rfam families. Take the fitting on the artificial sequence with 50% GC content as an example. Out of 106 tested families, 103 families fit best with the Gamma distribution, and the other three families (bicoid_3, OLE, and rne5) fit best with the normal distribution. The score distribution fitting of the Corona_FSE family (which follows Gamma) and the rne5 family (which follows normal) is shown in Figure 3.2. It is clear that even rne5

fits better with the normal distribution, the Gamma distribution also fit the ProbeAlign score distribution well. So the Gamma distribution was chosen to evaluate the p-values in ProbeAlign.

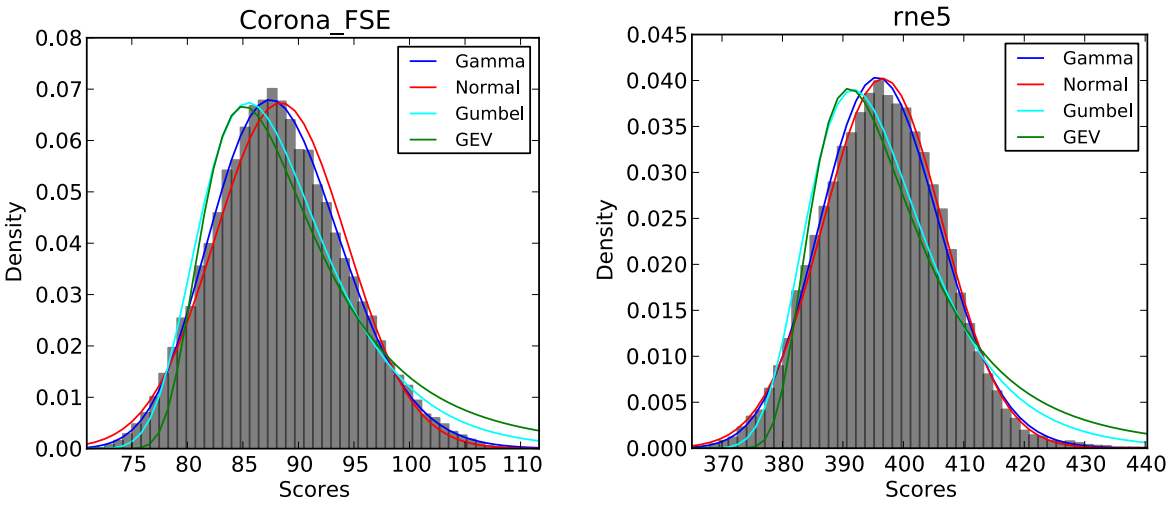

Figure 3.2: Fitting of the alignment score distributions for Corona_FSE and rne5 families.

3.3 Results

3.3.1 Benchmarks

In this section, we will compare the performance of ProbeAlign and CMsearch using the RMARK3 benchmark dataset. This dataset contains 106 families, and each family has a training alignment and a test set. The training alignments were employed to generate queries for both tools. The sequences in the test sets were concatenated together and served as the target in the experiments. The corresponding reactivities of the target were simulated based on the SHAPE technology [148]. To make the comparison between ProbeAlign and

CMsearch fair, for each family, we kept the number of predictions of these two programs the same. A server machine with 4 Xeon i7 CPUs (2.4 GHz) and 128 GB RAM was used for the benchmarking and subsequent experiments under single core configuration.

CMsearch adopts the covariance model to query against the target sequences to detect RNA homologs. The recent release of CMsearch is coupled with Hidden Markov Model (HMM)-based filters to improve its computational efficiency [112]. In the following experiments, CMsearch will be invoked with default setting, which means the filters are coupled and the default parameters are used. For ProbeAlign, the weights for the structural and sequence similarity, α and β, were set to 0.7 and 2.6, respectively. The simplified scoring function for structural similarity was used as default in the benchmarks. According to the property of the SHAPE reactivity data [28], r_c was set to 0.3. A detailed discussion of parameter selection will be presented in the following section.

The synthesized target contains 780 ncRNA sequences (160,390 bps) from the RMARK3 dataset. It takes 2.13 minutes CPU time for ProbeAlign to finish the search while it takes 6.85 minutes CPU time for CMsearch. Such improvement is expected, since ProbeAlign adopts an $O(mn)$ algorithm, while the core algorithm of CMsearch is from $O(mn^{1.3})$ to $O(mn^{2.4})$ [111], for a query with n sites and a target with m bases. In terms of prediction accuracy, the overall TP/FP ratio of CMsearch is 632/292, while that of ProbeAlign is 653/271. Of the 106 ncRNA families in RMARK3, ProbeAlign generates different prediction results with CMsearch in 27 families. Table 3.1 shows the performance difference of ProbeAlign and CMsearch on those families.

The search results for the tRNA and RNase_MRP families, whose ROC curves are shown in Figure 3.3, clearly demonstrate the advantage of using the probing information to detect

Table 3.1: Summary of the results of ProbeAlign and CMsearch on the RMARK3 dataset. Only the families with different results between the two programs are shown in the table.

Rfam ID	Name	Identity	# Tests	# Predictions	CMsearch		ProbeAlign	
					TP	FP	TP	FP
RF00005	tRNA	44%	20	61	10	51	**16**	45
RF00007	U12	61%	7	8	**7**	1	6	2
RF00013	6S	43%	38	24	21	3	**24**	0
RF00017	SRP_euk_arch	46%	21	24	19	5	**21**	3
RF00020	U5	52%	22	23	19	4	**22**	1
RF00023	tmRNA	48%	59	59	**58**	1	57	2
RF00028	Intron_gpI	34%	20	5	4	1	**5**	0
RF00030	RNase_MRP	44%	28	36	16	20	**22**	14
RF00066	U7	62%	2	1	**1**	0	0	1
RF00080	yybP-ykoY	46%	13	13	**13**	0	10	3
RF00104	mir-10	58%	2	1	0	1	**1**	0
RF00114	S15	61%	8	11	**8**	3	7	4
RF00140	Alpha_RBS	65%	3	4	1	3	**3**	1
RF00165	Corona_pk3	68%	1	4	0	4	**1**	3
RF00177	SSU_rRNA_5	49%	13	17	12	5	**13**	4
RF00230	T-box	46%	48	50	46	4	**47**	3
RF00504	Glycine	50%	14	14	**14**	0	13	1
RF00515	PyrR	46%	29	38	25	13	**28**	10
RF00534	SgrS	48%	3	2	0	2	**1**	1
RF00548	U11	57%	8	11	**7**	4	5	6
RF00640	MIR167_1	53%	10	9	**8**	1	7	2
RF00645	MIR169_2	52%	21	21	**21**	0	20	1
RF00661	mir-31	57%	3	3	2	1	**3**	0
RF01052	Arthropod_7SK	65%	2	3	0	3	**2**	1
RF01066	6C	67%	1	2	**1**	1	0	2
RF01069	purD	56%	8	9	**8**	1	7	2
RF01296	snoU85	62%	2	6	1	5	**2**	4
	Overall		406	459	322	137	**343**	116

remote homologous sequences. The sequence identities for these two families are 46% and 47%, which make it challenging for HMM-based filters to find the tested RNAs. Note that it would be possible for CMsearch to predict more low sequence identity hits by disabling the filters, but it would dramatically (by 10,000-fold) increase the running time [112]. On the other hand, when the probing information is considered, the high structural similarity is able to compensate the low sequence similarity, and lift the ranking of ncRNA sequences that are difficult to be detected by CMsearch. Figure 3.4 shows that a tRNA homolog (Accession ID: AY632242.1/10-80) missed by CMsearch was identified by ProbeAlign. Of 71

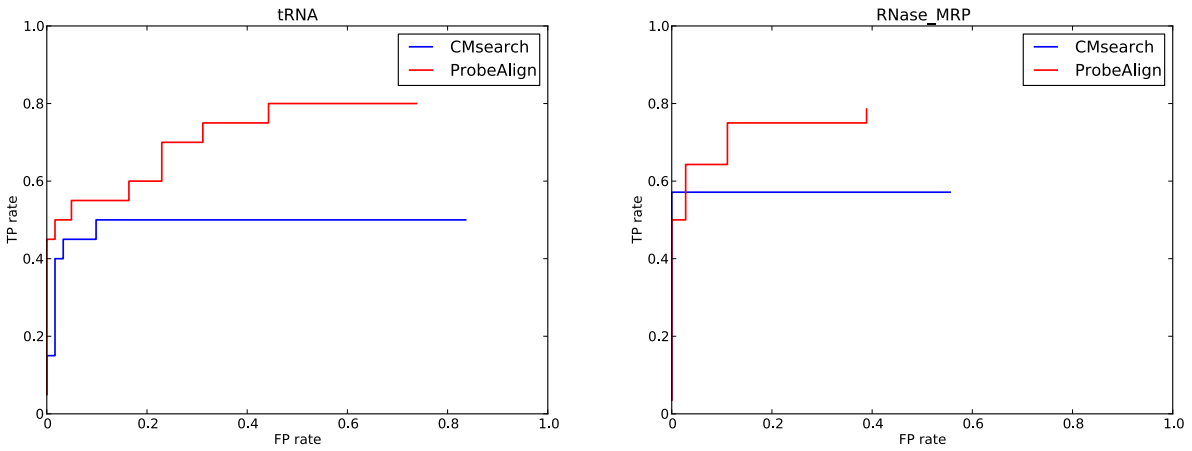

Figure 3.3: ROC plots for the performance of CMsearch and ProbeAlign in searching tRNA and RNase_MRP. CMsearch is invoked with the default parameters and filters. TP rate is computed by dividing the number of TP predictions by the size of the training set. FP rate is computed by dividing the number of FP predictions by the total number of all predictions.

sites in the profile of the training set, 13 sites have frequencies less than 12.5% and 22 sites have frequencies between 12.5% and 25%, which prevents the HMM filters to retrieve some tRNAs.

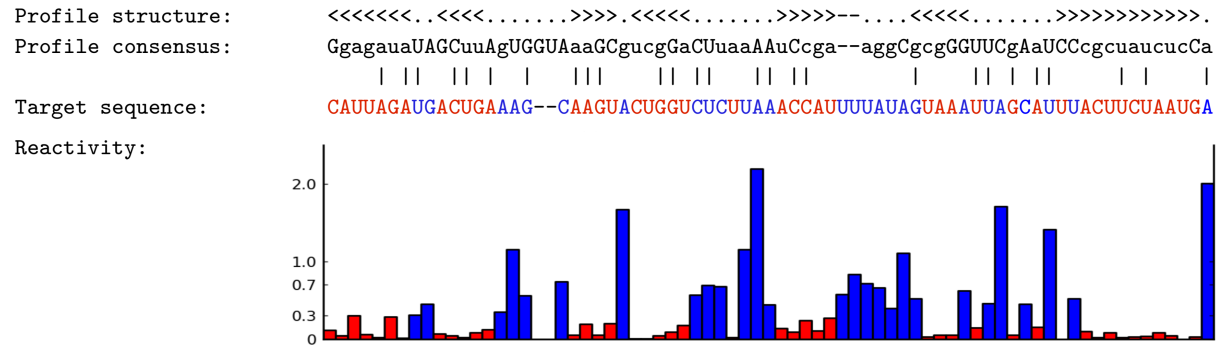

Figure 3.4: An alignment generated by ProbeAlign between the tRNA query profile and a target tRNA sequence. The accession ID of the target RNA sequence is AY632242.1/10-80. The red color in target sequence and bars indicates the sites with low reactive scores ($< r_c$). The blue color indicates the sites with high reactive scores ($> r_c$).

3.3.2 Optimizing the structure and sequence similarity weights

In the ProbeAlign algorithm, the parameters α and β indicate the weights for the structural and sequence similarity, and control how the two types of information are incorporated into the dynamic programming algorithm. The setting of these parameters should reflect how well the probing data would represent the actual secondary structure, as well as which information is more important in defining a specific ncRNA family. Ideally, the setting of the parameters should be family-specific to satisfy the structure and the sequence conservation patterns. However, it is tedious to define a set of parameters for each search profile, and more importantly, the overly tweaked parameters for the under-represented families would even bias the search. In this case, it is expected to apply a set of universal parameters for all families.

Three experiments have been conducted to analyze the effect of α and β on the performance of ProbeAlign by using the RMARK3 dataset. The value of α varied from 0 to 2 with an increasing step of 0.1, while the value of β varied from 4 to 0 with a decreasing step of -0.2. In the first experiment, we investigated the performance of ProbeAlign with the default setting under different combinations of α and β. In the second experiment, we excluded the structure consistency score to investigate its contribution to the overall performance. In the third experiment, the prediction was based upon the SHAPE-specific scoring function for structural similarity. Figure 3.5 shows the performance of ProbeAlign in these three experiments. For the first experiment (Figure 5, red line), the optimal performance is achieved at $\alpha = 0.7$ and $\beta = 2.6$, which is then taken as the default setting for the algorithm. For the second experiment (Figure 5, blue line), the optimal performance is achieved at $\alpha = 0.6$ and $\beta = 2.8$. The performance of ProbeAlign is higher than that without considering the struc-

ture consistency score. Such improvement is more significant when the structural weight is higher. Therefore structure consistency score is an effective way of improving the overall performance. In the last experiment, we adopted the SHAPE-specific scoring function to evaluate the structural similarity between the Rfam families and the target sequence. We can see that the optimal performance for the SHAPE-specific function (Figure 5, green line) and the default simplified function (Figure 5, red line) is comparable: 656/268 at $\alpha = 0.9$ and $\beta = 2.2$ for the SHAPE-specific scoring function; 653/271 at $\alpha = 0.7$ and $\beta = 2.6$ for original simplified scoring function. The performance difference is increased when raising the ratio of α/β. Therefore, the protocol-specific scoring function may be a better choice if the underlying reactivity distribution is known. The implementation of ProbeAlign allows users to decide which scoring function they use.

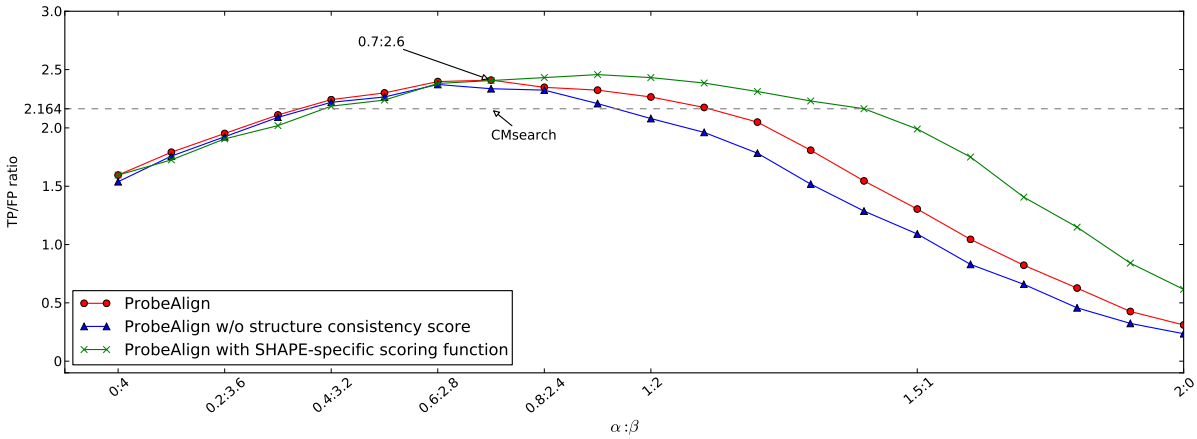

Figure 3.5: Performance of ProbeAlign with different structure and sequence similarity weights. The TP/FP ratio of CMsearch, 2.164, is represented as a dash line in the figure.

3.3.3 High-throughput sequencing-based RNA structure probing data

FragSeq (fragmentation sequencing) is a genome-wide RNA structure probing technique that has been applied to study the mouse nuclear transcriptome [159]. The RNA secondary structures in the KH2 undifferentiated mouse embryonic stem cells (undiff) and neural precursors cells (d5np) were probed. By analyzing nine snRNA families (U1, U2, U3, U4, U5, U6, U8, U11, and U12) in the mouse genome, the paper shows a good accordance between the probing data and the real secondary structures. New secondary structures have also been proposed to three other snRNA families (U15, U22, and U97), showing the ability of FragSeq to discover novel ncRNA transcripts and their secondary structures.

We used ProbeAlign with default setting to search the nine snRNAs families against the FragSeq data to demonstrate its utility on experimentally determined reactivities. We were only interested in the genomic regions that were transcribed, i.e. being covered by more than 4 sequenced reads. There are 18,388 regions (32.2 Mbps) for the undiff cell line and 17,007 regions (29.0 Mbps) for the d5np cell line. The reactivities for these regions were computed using FragSeq v0.0.1, a supplementary software for the probing protocol [159]. Because FragSeq is a different technology than SHAPE, r_c was adjusted to 0.5 from 0.3. All other parameters remained the same as in the benchmark. A universal p-value cutoff (0.01) was set for all searches. The running time for the undiff dataset was 30.20 minutes CPU time, and for the d5np dataset was 26.97 minutes CPU time. During the analysis of the predicted results, we found some reads were mapped onto repeat regions in the genome. Those hits were removed by using Repbase database [82]. The final search results are summarized in Table 3.2. U11 and U12 have no record in Repbase. Only 17 and 21 U4 records in Repbase are covered by the transcribed regions of d5np and undiff cell lines, and all of them are top

ranked in the results of ProbeAlign. The corresponding sequences with their locations on the genome can be downloaded at http://genome.ucf.edu/ProbeAlign.

Table 3.2: Summary of the prediction results by ProbeAlign on the FragSeq data. The numbers in the brackets show that how many predictions by ProbeAlign are recorded in Repbase.

	RF00003 U1	RF00004 U2	RF00012 U3	RF00015 U4	RF00020 U5	RF00026 U6	RF00096 U8	RF00548 U11	RF00007 U12
d5np	46(46)	18(18)	11(11)	243(17)	12(12)	120(117)	2(2)	1	4
undiff	46(46)	19(19)	11(11)	302(21)	12(12)	146(134)	2(2)	1	4

One interesting observation from the ProbeAlign search results is that the transcription of U4 and U6 snRNA families are more active in undiff cells than in d5np cells. It is not surprising to see the potential correlation between the U4 and U6 transcription level, as they have been proposed to interact with each other in splicing control. In fact, it is hypothesized that they can bind with each other due to a long complementary sequence between them [78]. Recent experiments show that the snRNAs in un-proliferated stem cells have higher expression than in proliferated cells [101]. The observation is explained by the snRNAs playing an important role in ribosome biogenesis, cellular proliferation and pre-mRNA splicing [85]. From the ProbeAlign search results, we can further conclude that not only the expression level of the snRNA is higher in un-proliferated cells, there are actually more U4 and U6 snRNA genes being transcribed in un-proliferated stem cells.

3.4 Discussion and Conclusion

In this article, we have proposed a novel algorithm, ProbeAlign, for incorporating high-throughput sequencing-based RNA structure probing data into ncRNA homology search.

To our knowledge, this is the first application of structure probing information to RNA functional annotation. This integration makes the accuracy of ProbeAlign even higher than the CMsearch tool, especially for ncRNA homologs with low sequence identity. In addition, the time complexity of the algorithm is $O(n^2)$, which is feasible for handling genome-wide datasets.

ProbeAlign itself can also act as a filter for more detailed downstream alignment algorithms. Considering both ProbeAlign and the HMM filters in CMsearch being $O(n^2)$ time complexity algorithms, they should have comparable time efficiency if similarly optimized. It is clear that ProbeAlign guarantees higher sensitivity and specificity. In this case, ProbeAlign can be coupled with more accurate alignment algorithms such as CMsearch itself, or other structure-sequence alignment algorithms such as FastR [7], PFastR [183], and RSEARCH [87]. We are also developing a new structure-sequence alignment algorithm that takes into account the probing information, which can also be used as the downstream detailed alignment after ProbeAlign screening.

In conclusion, we present here an accurate and efficient RNA homology search algorithm, ProbeAlign, which incorporates the high-throughput sequencing-based RNA structure probing information. With the increasing requirement of genome-wide ncRNA annotation, we anticipate that more RNA transcripts, and their secondary structures and functionalities, will be annotated by using ProbeAlign.

CHAPTER 4: RNA TERTIARY STRUCTURE ALIGNMENT USING A STACK-BASED STRATEGY

4.1 Background

Non-coding RNAs (ncRNAs) play diverse cellular functions in biological systems [8, 38, 69, 147]. Unlike mRNAs whose primary sequences are genetic codes for protein synthesis, the regulatory information of most ncRNAs is encoded in their architectures: the secondary structures defined by the hierarchical assembly of double-stranded stacks, and higher-order three-dimensional (3D) structures consisting of packed secondary structure modules interlinked by tertiary interactions [79, 175]. Therefore, the structural alignments of such ncRNAs can provide essential insight to their functional and evolutionary relationships. However, compared to the development of the computational methods for RNA secondary structure analysis, the progress of RNA 3D structural alignment has been limited. Although the protein 3D structural alignment has been studied for years and many sophisticated methods have been proposed [40, 76, 105, 106, 180], it is hard to apply them directly to ncRNAs due to the different properties of their secondary structures.

Recently, with the rapid growth of RNA deposition in the Protein Data Bank (PDB) [12], a number of tools have been developed specifically for the alignments of RNA 3D structures. Generally, they can be categorized into two groups. In the first group, the base pairing

interactions in the inputs are ignored, or degraded into sequential information. Then the RNAs can be compared using the quadratic-time alignment algorithms. For example, both iPARTS [161] and LaJolla [9] represent RNA backbones as sequences of letters derived from the features of nucleotide torsion angles. iPARTS continues to apply conventional pairwise alignment methods to the encoded linear sequences, while LaJolla searches the similar "n-grams" (substrings of length n) in the RNAs by using hash tables. Similar to LaJolla, FRIEs [166] also uses the matching of k-mer RNA fragments. In this method, a large set of training fragments from the PDB are clustered into tens of classes based on their structural properties. Each k-mer in an RNA can be labelled with the probabilities in these classes, and thus the similarity of two fragments can be measured with the dot product of their probability vectors. Rclick [113] is another RNA 3D structural alignment tool based on the detection of local similarity. The matches between n-body cliques (in which n member nucleotides satisfy that all pair-wise spatial distances are within a threshold) are determined by the superimposition of their atomic coordinates. With this local structural equivalence, the optimal global alignment is generated by using 3D least squares fitting. Unlike the previously mentioned tools, DIAL [45] incorporates base pairing interactions into its dynamic programming scoring function, which also accounts for sequence and torsion angle information. A penalty is assigned if the pairing attributes (paired or unpaired) of two aligned nucleotides are different. ESA [91] models the RNA 3D structures not as sequences but as curves in a four-dimensional space: the atomic coordinates are in 3D space, and the sequence information is encoded as an additional dimension. Then the similarity between two RNAs can be evaluated by minimizing their geodesic distance with a quadratic-time dynamic programming algorithm.

The other group of RNA 3D structural alignment tools relies on the comparison of base pairing interactions in the molecules. In ARTS [35], two successive base pairs are used together as a seed. The optimal matching of seeds in two RNAs is extended globally to the unpaired regions, and the result is refined with the least squares fitting technique. Similar to ARTS, the final results of R3D Align [124] are assembled from the alignments of local neighborhoods. Neighbors are the spatially closest nucleotides in one RNA, which may imply interactions such as base pairs, tertiary interlinks and stacking contacts. The structurally similar neighbors in two RNAs are detected, and the optimal combination of these local alignments is determined by employing maximum clique finding algorithm on a compatibility graph. SARA [20] does not discriminate the paired and unpaired regions in RNAs. Inspired by a protein 3D structural alignment method named MAMMOTH [118], SARA describes the backbone of an RNA as a series of unit-vectors. The distances between the unit spheres of inputs can be measured with URMS (unit-vector root mean square), and the corresponding global alignment is identified by using dynamic programming. The same procedure is applied only to base pairs if the pairing information is provided. The 3D structural alignment of entire RNAs can be optimized based on the mapping of pairing interactions. SETTER [75] integrates stacks and loops into the RNA 3D structural alignment method. It splits the RNA sequences into GSSUs (generalized secondary structure units), each of which has a loop, a neck and a stem. The highly similar GSSU pairs are used as seeds to guide the alignment of other GSSUs. To simplify the computation, the exact mapping of nucleotides is ignored in this method.

It can be seen that the RNA secondary structural information, in particular the hierarchical topology of stacks, is not used in the reviewed methods. However, the enclosing and juxtaposing relations between stacks provide more detailed structural information than what

has been used in the existing tools, such as "paired" or "unpaired" attributes, base pairing interactions and stack positions. The issue is the difficulty of integrating the conventional RNA secondary structure alignment algorithms into the RNA 3D structural comparison. Given the high time complexity of these algorithms [at least $O(n^3)$] [68, 133, 168, 188], applying them directly to the relatively complicated atomic coordinates will increase the computational complexity significantly.

Here, we propose a novel RNA 3D structural alignment tool called STAR3D that explicitly makes use of the conservation of secondary structures with high efficiency. It aims at finding the consensus of stacks by using 2D topology and 3D geometry first, and then uses it to guide the alignments of the loop regions. To achieve this goal, first, the sub-stacks with similar 3D structures are detected and assembled into conserved stack pairs. Then, a compatible graph is constructed based on their secondary structural relations and spatial distances. In this graph, the maximum clique can be converted into a tree-like consensus structure of two RNAs. After that, the loop regions are ordered by the common tree. Each of them only needs to be compared with its partner by using 3D information. STAR3D has been implemented in Java. The benchmarking results show that STAR3D outperforms the state-of-the-art RNA 3D structural alignment tools with high efficiency.

4.2 Materials and Methods

4.2.1 Preprocessing

The inputs of STAR3D are the atomic coordinates of two polymer RNA chains, which are presented in the corresponding PDB files. They are preprocessed to obtain the RNA secondary structures to guide the 3D structural alignment. All plausible pairing interactions are identified by using MC-Annotate [54, 93]. Among them, the Watson-Crick base pairs (A↔U, C↔G) and wobble base pairs (G↔U) are retrieved to form the RNA secondary structures [34]. Other pairing interactions are considered during the loop alignment. In order to avoid excessive computation, we eliminate the crossing base pairs in the secondary structures by using the program RemovePseudoknots [144] in the RNAstructure package [126]. The discarded stems are used as pairing interactions in the loops.

4.2.2 Stack decomposition

Helical structured stacks are formed by consecutively nested Watson-Crick base pairs and wobble base pairs. To detect the 3D structural conservation in the stacks efficiently, the double-stranded regions in the pseudoknot-free secondary structures are decomposed into consecutive sub-stacks of size k, namely k-stacks. A stack with l base pairs ($l \geq k$) can be divided into $l - k + 1$ overlapping k-stacks. All the possible k-stacks are collected for further processing.

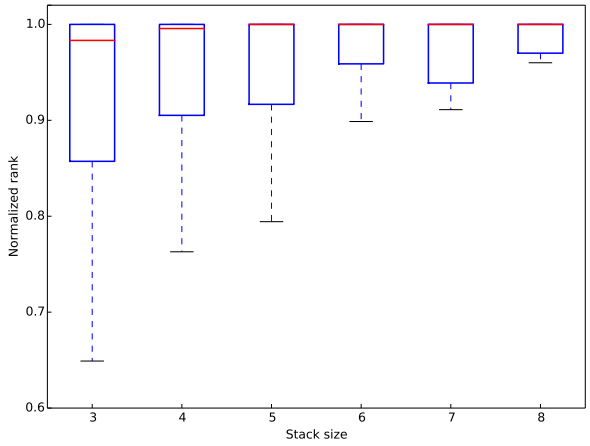

Figure 4.1: The normalized ranks of the matched stacks in two 23S rRNAs (PDB 2j01, chain A and PDB 2aw4, chain B). The structure similarity is measured with RMSD and the matching of stacks is determined by a hand-crafted base pair alignment [146]. For a stack of size k in 2j01, its RMSDs to all the size-k stacks in 2aw4 are computed.

Based on the definition of k-stack, we introduce some basic notations. Given an RNA A, the 3D coordinates of the i-th residue are denoted as $A[i]$. At the secondary structure level, the set of k-stacks in the pseudoknot-free structure is denoted as $\mathcal{P}^A$. For a specific k-stack $p^A \in \mathcal{P}^A$, the index of the leftmost base (5' end) is represented as $b(p^A)$ and the index of the rightmost base (3' end) is represented as $e(p^A)$. Thus the 3D coordinates of the double-stranded subsequences in p^A are $A[b(p^A) \dots b(p^A) + k - 1]$ and $A[e(p^A) - k + 1 \dots e(p^A)]$, which are defined as $3D(p^A)$.

4.2.3 Detecting the conserved stack regions

STAR3D identifies the stack components conserved in 3D structures as anchors and uses them to constrain the global alignment. Similar approaches have been applied in numerous computation-efficient tools for genome alignment [18, 100] and RNA secondary structure

alignment [6, 81]. The difference is that STAR3D uses the 3D coordinates of atoms to detect the potential homologous regions. Given the fact that RNA stacks adopt an A-form helical conformation, a major issue needs to be addressed: whether the 3D structural similarity of conserved stack regions is significant enough to distinguish them from the random ones. The survey results in Figure 4.1 indicate that the orthologous sub-stacks have highly similar 3D structures, and they can be detected by evaluating RMSD. In our method, the k-stacks (default value of k is 3) in the inputs are retrieved as the building blocks for the larger conserved regions. Shorter helices are not considered because of their low occurrence in the real RNAs.

Given two input RNAs A and B, the two sets of k-stacks $\mathcal{P}^A$ and $\mathcal{P}^B$ are sorted in ascending order by the leftmost bases. The three-dimensionally conserved k-stacks in A and B are determined by their RMSDs. $C_{i,j}$, the indicator of conservation for p_i^A and p_j^B, is computed using the following function:

$$C_{i,j} = \begin{cases} 1 & RMSD(3D(p_i^A), 3D(p_j^B)) < r_c \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

The function $RMSD$ measures the average spatial distance between the superimposed residues in $3D(p_i^A)$ and $3D(p_j^B)$, with r_c being the RMSD cutoff (default value is 4 Å) [20]. In our implementation, the RMSD values are computed with the Kabsch method [83] by using the geometric center of six backbone atoms C3', C4', C5', O3', O5' and P [125]. The indicators for all pairs of k-stacks ($\mathcal{P}^A \times \mathcal{P}^B$) are stored in a matrix. Then, we extend the consecutive matches of k-stacks to form larger ungapped alignments. For instance, p_i^A, p_{i+1}^A and p_j^B, p_{j+1}^B can be merged into two aligned stacks of size $k+1$, if $C_{i,j} = C_{i+1,j+1} = 1$, $b(p_i^A) = b(p_{i+1}^A) - 1$, $e(p_i^A) = e(p_{i+1}^A) + 1$, $b(p_j^B) = b(p_{j+1}^B) - 1$ and $e(p_j^B) = e(p_{j+1}^B) + 1$. This

procedure continues through the diagonals of the matrix until all the constructed alignments can not be extended any further. The two stack components in an assembled alignment are called extended stacks, written shortly as e-stacks. Correspondingly, the aligned e-stacks form e-stack pairs. We define the sets of e-stacks in A and B as $\mathcal{Q}^A$ and $\mathcal{Q}^B$, and the set of e-stack pairs as $\mathcal{S}$. According to the definition of e-stack, the cardinalities of $\mathcal{Q}^A$, $\mathcal{Q}^B$ and $\mathcal{S}$ are identical. As a result, we denote the members of a specific e-stack pair $s_i(\in \mathcal{S})$ as $q_i^A(\in \mathcal{Q}^A)$ and $q_i^B(\in \mathcal{Q}^B)$ $(s_i = (q_i^A, q_i^B))$. Note that e-stacks may overlap with each other [see Figure 4.2(a)]. Unlike k-stacks, the sizes of e-stacks are not fixed. Therefore, we define a new notation $l(q^A)$ to represent the number of base pairs in q^A. Hence $3D(q^A)$ are $A[b(q^A) \ldots b(q^A) + l(q^A) - 1]$ and $A[e(q^A) - l(q^A) + 1 \ldots e(q^A)]$.

For some large RNAs, the numbers of e-stack pairs are too large for computation. To determine the highly significant ones, we consider two criteria: the RMSD between two e-stacks and their size. The significant scores of e-stack pairs are defined using the formula $RMSD(3D(q_i^A), 3D(q_i^B)) - 0.1 \times l(q_i^A)$. They are sorted in ascending order and only 200 top-ranked pairs are retained for further processing. Based on our study, 200 high-scoring e-stack pairs are sufficient to cover most of the conserved helical regions in 23S rRNAs, the largest RNAs in PDB. More e-stack pairs may be used by setting the parameter if more complex structures are presented.

4.2.4 Assembling the consensus of stacks

To generate a consensus of stacks, the positions of e-stack pairs in the secondary structures and 3D space are analyzed. In the pseudoknot-free secondary structure of A, two e-stacks q_i^A

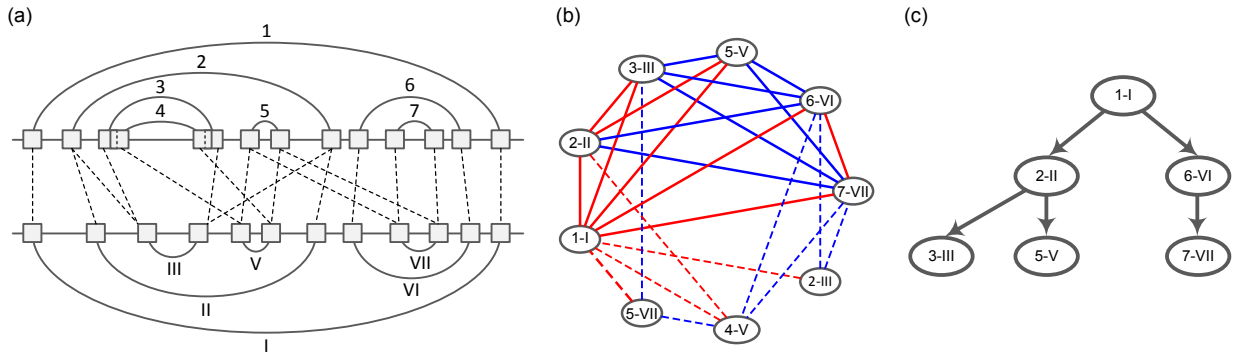

Figure 4.2: A description of the basic data structures used in STAR3D. (a) The e-stack pairs in two artificial RNAs. The gray boxes show the e-stacks and the dashed lines show the matching between them. e-stack 3 and e-stack 4 are overlapped with each other. (b) The compatible graph of e-stack pairs in (a). Red color marks the enclosing relations and blue color marks the juxtaposing relations. The solid lines show the edges in the maximum clique. To simplify the presentation, the 3D similarity requirement is not considered in the figure. (c) The tree-like consensus of e-stacks obtained from the clique in (b).

and q_j^A may have one of the three following relations: (i) q_i^A and q_j^A are overlapping (denoted by $q_i^A \otimes q_j^A$); (ii) q_i^A encloses q_j^A (denoted by $q_i^A \prec_E q_j^A$); (iii) q_i^A is before and juxtaposed to q_j^A (denoted by $q_i^A \prec_J q_j^A$). In our algorithm, q_i^A *directly* encloses q_j^A if $q_i^A \prec_E q_j^A$ and $\nexists k(q_i^A \prec_E q_k^A \prec_E q_j^A)$ (denoted by $q_i^A <_E q_j^A$). Similarly, we say q_i^A is *directly* before and juxtaposed to q_j^A if $q_i^A \prec_J q_j^A$ and $\nexists k(q_i^A \prec_J q_k^A \prec_J q_j^A)$ (denoted by $q_i^A <_J q_j^A$).

Notice that both $\prec_E$ and $\prec_J$ are strict partial orders, so the non-overlapping e-stacks in an RNA can form a directed acyclic graph (DAG). It is well-known that the RNA secondary structures have a tree-like topology [71, 104, 135, 184]. Thus we model the non-overlapping relations of e-stacks in A as a tree:

1. Assign a pseudo stack $q_\bullet^A$ $(b(q_\bullet^A) = 0, e(q_\bullet^A) = |A| + 1, l(q_\bullet^A) = 0)$ to the root node.

2. Connect q_i^A to q_j^A if $q_i^A <_E q_j^A$.

3. Order the children nodes of q_i^A in ascending order based on $\prec_J$.

We also define the compatible e-stack pairs: s_i and s_j are compatible if $(q_i^A, q_j^A) \in R$, $(q_i^B, q_j^B) \in R$, and $R \in \{\prec_E, \succ_E, \prec_J, \succ_J\}$. The non-compatible e-stack pairs can not be in the consensus together because their members are disordered in the secondary structures.

Furthermore, we can prove the following lemma:

Lemma 1. *For a non-empty set $\mathcal{S}' \subseteq \mathcal{S}$, if any two of e-stack pairs $s_i = (q_i^A, q_i^B)$ and $s_j = (q_j^A, q_j^B) \in \mathcal{S}'$ are compatible, the corresponding two e-stack sets have the same tree structure.*

Proof. Without loss of generality, we assume that q_i^A is a child of $q_\bullet^A$ and q_i^B is not a child of $q_\bullet^B$. Then q_i^B must be on a subtree rooted at one child of $q_\bullet^B$. Thus at least two stacks, $q_\bullet^B$ and q_j^B, enclose q_i^B. However q_i^A has only one ancestor $q_\bullet^A$. It is a contradiction to the conditions, because s_i and s_j are not compatible. So the children of $q_\bullet^A$ and $q_\bullet^B$ are from the same set of e-stack pairs. Based on Step 3 of the tree construction procedure, the orders of their children should be the same. Then it is proved that the lemma holds for the top two levels of the trees. Assume it is also true for the top n levels. Then for an e-stack $q_{i'}^A$ at n-th level, its partner $q_{i'}^B$ must also be at n-th level, and their relative positions on the trees are the same. Assume an e-stack $q_{j'}^A$ is a child of $q_{i'}^A$ and $q_{j'}^B$ is not a child of $q_{i'}^B$. First $q_{j'}^B$ must be on a subtree rooted at $q_{i'}^B$. Otherwise $s_{i'}$ and $s_{j'}$ are not compatible. Second, $q_{j'}^B$ can not be at $(n+2)$-th or lower levels, otherwise the numbers of the ancestors of $q_{j'}^A$ and $q_{j'}^B$ are different, which is a contradiction to the conditions. So the children of $q_{i'}^A$ and $q_{i'}^B$ are also from the same set of e-stack pairs, and they can be sorted by the juxtaposing relation. By induction, we know the lemma is true. □

Lemma 1 indicates how to find the 3D structural consensus of the stack regions in two input RNAs. Thus, to detect the e-stack configuration for the consensus, we construct a compatible graph. The vertices are the e-stack pairs. A vertex s_i is connected to another one s_j if they meet two requirements. First, s_i and s_j are compatible, which ensures the e-stacks in them are well-ordered in the secondary structures. Second, s_i and s_j must satisfy $RMSD(3D(q_i^A) \circ 3D(q_j^A), 3D(q_i^B) \circ 3D(q_j^B)) < r_c$, which implies that s_i and s_j share similar rigid transformation ("$\circ$" means the concatenation operation which joins the lists of 3D coordinates end-to-end). Based on the graph properties, the optimal stack configuration can be inferred from the maximum clique in the graph, which is detected by using the Bron-Kerbosch algorithm [17]. After that, the 3D structural alignment in the double-stranded regions is determined by the topology of these vertices in the clique. Note that the e-stacks are the maximal 3D conversed regions in the helices (they can not be extended any more). Therefore, the 3D similarity requirement will filter most of the improper edges, and make the compatible graph very sparse. Although normally finding the maximum clique takes exponential time, it is solved very efficiently in our method. Figure 4.2(b) and 4.2(c) show a compatible graph and the corresponding consensus of stacks. The detected consensus is the "core" of the 3D structural conservation, and it will work as an anchor for the following loop alignment. The double-stranded regions not in the consensus, such as stack 4 in Figure 4.2(a), are considered as loops in the following computation. The corresponding Watson-Crick base pairs and wobble base pairs are also used as interactions in the loop regions to assist the alignment.

4.2.5 Loop alignment using 3D information

With the tree-like consensus of stacks, all the other regions not in it can be divided into ordered loops. For one leaf node, two hairpin loops enclosed in two e-stacks can be identified. For the internal nodes, their enclosed regions are split by their children nodes into internal loops, bulges, or multi-loops. Hence, the numbers of loops in the inputs are the same, and we can find the mapping of them by traversing the tree. This approach has two benefits. First, the computational efficiency of loop alignment can be improved significantly for large RNAs, because only the matched loops need to be aligned together. Second, the superimposition of stack regions can be used to guide the 3D structural alignment of loop regions. For the functional RNAs, the stack regions are more conserved than the loop regions. Thus, any RMSD computation during the loop alignment uses the rotation and translation of the stack alignment.

A dynamic programming algorithm with quadratic-time complexity is applied to the 3D structural alignment of two loops. Assume the 3D structures of k-th pair of matched loops are $A[i_k \dots i_k + n_{k_1} - 1]$ and $B[j_k \dots j_k + n_{k_2} - 1]$. To simplify the description and computation, we denote them as $L_k^A[1 \dots n_{k_1}]$ and $L_k^B[1 \dots n_{k_2}]$, whose starting index is 1. Thus, the recursive function is ($1 \le i \le n_{k_1}$, $1 \le j \le n_{k_2}$):

$$
\begin{aligned}
I_{i,j} &= \max\{M_{i-1,j} + \epsilon_o + \epsilon_e, I_{i-1,j} + \epsilon_e, D_{i-1,j} + \epsilon_o + \epsilon_e\} \\
D_{i,j} &= \max\{M_{i,j-1} + \epsilon_o + \epsilon_e, I_{i,j-1} + \epsilon_o + \epsilon_e, D_{i,j-1} + \epsilon_e\} \\
M_{i,j} &= \max\{I_{i-1,j-1}, D_{i-1,j-1}, M_{i-1,j-1}\} + \alpha(i,j) + \beta(i,j)
\end{aligned}
\tag{4.2}
$$

Here, ϵ_o and ϵ_e are the gap open penalty and gap extension penalty. I, D, M denote the optimal alignment scores for insertions, deletions and substitutions, respectively. These functions are initialized with $M_{0,0} = I_{0,0} = D_{0,0} = 0$, $M_{i,0} = M_{0,j} = -\infty$, $I_{i,0} = \epsilon_o + \epsilon_e \times i$, $D_{0,j} = \epsilon_o + \epsilon_e \times j$, $I_{0,j} = D_{i,0} = -\infty$. The optimal score is $\max(I_{n_{k_1},n_{k_2}}, D_{n_{k_1},n_{k_2}}, M_{n_{k_1},n_{k_2}})$ and the exact 3D structural alignment for the two loops can be found by using traceback.

The scores for substitution contain two parts: $\alpha(i,j)$ and $\beta(i,j)$. The function $\alpha(i,j)$ is based on the 3D distance between two bases. The corresponding formula is:

$$\alpha(i,j) = \begin{cases} -\infty & d_{i,j} \geq 2 \cdot r_c \\ mismatch\_score & 2 \cdot r_c > d_{i,j} \geq r_c \\ 0.5 \times match\_score & r_c > d_{i,j} \geq 0.5 \cdot r_c \\ match\_score & 0.5 \cdot r_c > d_{i,j} \end{cases} \tag{4.3}$$

where $d_{i,j}$ denotes the RMSD between two nucleotides $L_k^A[i]$ and $L_k^B[j]$. Note that they are superimposed with the transformation of aligned stack regions. To capture the backbone conformation, STAR3D uses 3-nucleotide regions, $L_k^A[i-1, i, i+1]$ for $L_k^A[i]$ and $L_k^B[j-1, j, j+1]$ for $L_k^B[j]$, in the computation of $d_{i,j}$. The possible values of $d_{i,j}$ can be categorized into three groups. The two nucleotides are not allowed to be aligned if the spatial distance is too large ($\geq 2 \cdot r_c$). Otherwise, they are defined to be "matched" or "mismatched", and the matched nucleotides may be assigned with two different scores.

The second function $\beta(i,j)$ calculates the bonus scores for the base pairs in loop regions. Pseudoknots, non-canonical base pairs and canonical base pairs in the unaligned stack regions are considered in the computation. Due to the potential crossing in pseudoknots and non-canonical base pairs, finding the optimal matching of these pairing interactions is an NP-hard

problem. To reduce the running time, we propose a heuristic algorithm to solve the problem. Generally, each base has three possible pairs: Watson-Crick base pair, Hoogsteen base pair and Sugar base pair [97]. All the predicted base pairs of two nucleotides $L_k^A[i]$ and $L_k^B[j]$ are compared in 3D space by using a similar approach of comparing nucleotides in $\alpha(i, j)$. The match of two base pairs is valid if the corresponding RMSD is less than r_c. The maximum number of matched pairs is returned as the result of $\beta(i, j)$. Thus the problem is converted into bipartite graph matching, which can be solved by dynamic programming.

4.3 Results

4.3.1 Benchmarking tools

STAR3D is benchmarked with ARTS, LaJolla (v2.2), SARA (v1.0.7) and R3D Align in this section. Their batch programs are available and widely used for performance testing. In addition, they can output the exact one-to-one mapping of nucleotides, which is important for the analyzing of specific alignments of homologous and non-homologous RNAs. R3D Align is dedicated to homologous RNAs. To make the comparison fair, it is only used in the experiments for homologous rRNAs. An in-house modification of LaJolla is implemented to output not only the rigid transformation but also the exact alignments. All the experimental results were performed with default parameters. To evaluate the secondary structure similarity and optimize the superimposition, "-b" and "-s" options are specified for SARA. Both ARTS and LaJolla generate "disordered alignments", e.g. a_i is aligned to b_j, a_k is aligned to b_l, while $i < k$ and $j > l$. For ARTS, the largest proper alignment is retrieved;

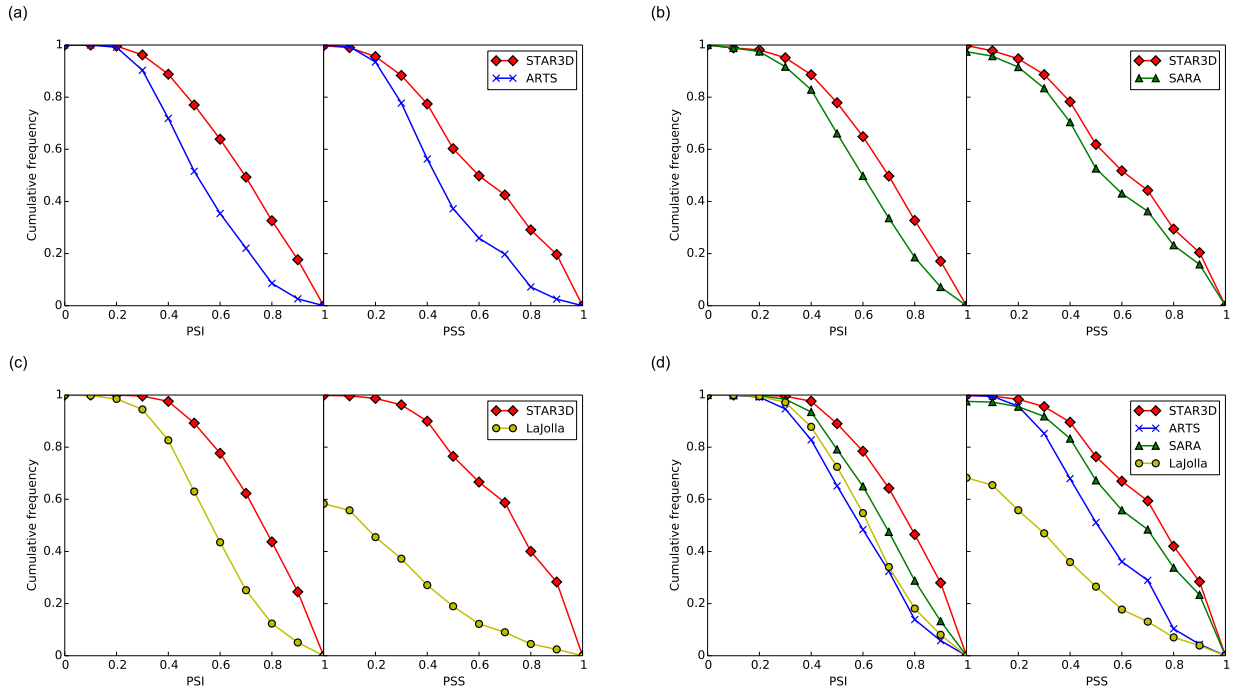

Figure 4.3: The cumulative frequencies of the PSI and PSS values of STAR3D, ARTS, SARA and LaJolla in different experiments. (a) STAR3D vs. ARTS. (b) STAR3D vs. SARA. (c) STAR3D vs. LaJolla. (d) All four tools.

for LaJolla, the improper alignment is discarded since only one result is returned from the modified implementation.

4.3.2 Alignment quality assessment with R-FSCOR dataset

The R-FSCOR dataset [21] contains 192 chains collected from the SCOR database [151]. In SCOR, the chains with at least three base pairs and unique function annotations are clustered at 90% identity. The representative in each cluster is selected into the R-FSCOR dataset. The performance of four tools is compared by calculating PSI (percentage of structural iden-

Table 4.1: The comparison of mean PSI and PSS values between STAR3D and three other tools by using the R-FSCOR dataset.The total number of inputs is 18336. ARTS, SARA, LaJolla and STAR3D output 11385, 18335 , 7771 and 17455 alignments, respectively. Best performance is set to **bold**.

	# of overlapped alignments	ARTS		SARA		LaJolla		STAR3D	
		PSI	PSS	PSI	PSS	PSI	PSS	PSI	PSS
ARTS vs. STAR3D	11054	0.538	0.485	-	-	-	-	**0.682**	**0.632**
SARA vs. STAR3D	17454	-	-	0.601	0.580	-	-	**0.679**	**0.638**
LaJolla vs. STAR3D	7397	-	-	-	-	0.580	0.251	**0.754**	**0.729**
Consensus	4451	0.600	0.549	0.683	0.668	0.627	0.318	**0.764**	**0.729**

tity) and PSS (percentage of aligned secondary structure) values of the all-to-all alignments for the R-FSCOR dataset. PSI is defined as the percentage of aligned nucleotides in 4Å with respect to the length of the shorter sequence. PSS is defined as the percentage of aligned base pairs in 4Å with respect to the smaller number of base pairs of two aligned RNA sequences. PSI and PSS have been used as replacement for RMSD to evaluate the quality of the 3D structural alignment [20, 75]. The base pairs in the tested chains, including both canonical base pairs and non-canonical base pairs, are predicted using MC-Annotate. All programs in this experiment were executed on a CentOS cluster with 100 nodes. None of the tools can find alignments for all the inputs. ARTS outputs 11385 proper alignments, LaJolla outputs 7771 proper alignment, SARA outputs 18335 alignments and STAR3D outputs 17455 alignments, respectively. For STAR3D, no alignment is generated if the sizes of all potential e-stacks in the inputs are less than $k(= 3)$. However, the alignments for these inputs can be detected if a smaller k (e.g. 2) is specified. In addition, RNAMotifScanX [189], which is also designed by our lab for searching RNA 3D structural motifs in the single-stranded regions, can be applied since those RNAs are relatively short and are dominated by loops. STAR3D was compared with ARTS, SARA, and LaJolla one by one. To make the comparison fair, the inputs are not considered if STAR3D or the corresponding benchmarking tool can not generate alignments for them. Table 4.1 summarizes the mean PSI and PSS values of four

tools in the experiments. It can be seen that STAR3D outperforms the other three tools by a large margin: the PSIs are increased by 13% to 30%, and the PSSs are increased by 10% to 190%. The low PSS values of LaJolla may be caused by ignoring of the secondary structural features. ARTS and SARA have relatively high PSS values because the base pairing information is integrated. For SARA, the optimization step after the backbone alignment may contribute to its better performance than ARTS. By considering the secondary structures of two input RNAs, STAR3D accurately predicts the matching of the stack regions, which is demonstrated by the high PSS values. And guided by the consensus of stacks, STAR3D provides best global alignments in all four tools without an optimization step, which is shown by the high PSI values. We also find that the running time of STAR3D for the whole procedure is much shorter (at least 1/10) than the other three tools. A detailed discussion about the computational efficiency will be shown in a later section. Figure 4.3 shows the cumulative frequencies of the PSI and PSS values in different comparisons. Figure 4.3(d) is based on the valid inputs for all tools. It can be seen that some alignments of SARA and LaJolla may not contain any base pair. On the other hand, the PSS values of ARTS and STAR3D are all greater than zero, because ARTS extends the base pair mapping and STAR3D relies on the stack mapping. What's more, from Figure 4.3(d), we can see that PSI curves between 0.0 to 0.2 are very similar for all tools. The major performance difference between STAR3D and the other three tools is at the range from 0.4 to 0.7, which indicates STAR3D may be more sensitive to the local conservation of RNAs.

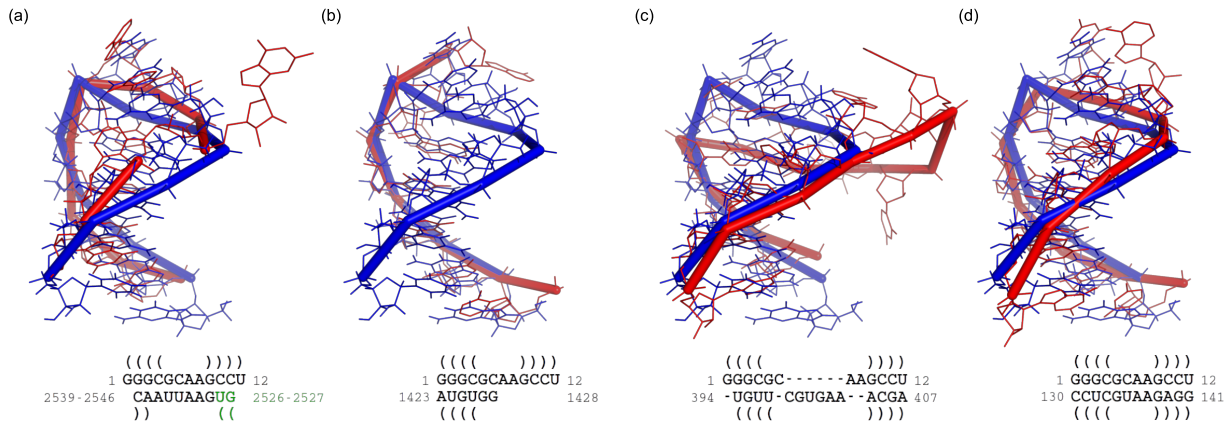

Figure 4.4: The alignment results for the GNRA motif (PDB: 1zih, chain A) and the 23S rRNA (PDB: 1njo, chain 0). (a) The result of LaJolla. (b) The result of SARA. (c) The result of ARTS. (d) The result of STAR3D. The blue ribbons show the 3D structure of the GNRA motif and the red ribbons show the 3D structures of the aligned regions in the 23S rRNA. The secondary structural alignments are listed below the 3D structure figures and the base pairs are predicted by MC-Annotate. The green letters in the LaJolla alignment mark the disordered nucleotides (2526-2527).

4.3.3 Structural alignments of non-homologous RNAs

Identifying the conserved regions in non-homologous RNAs is a major aim of the RNA 3D structural alignment tools. In this section, we will analyze the different strategies of STAR3D and three other tools by showing the alignments of non-homologous RNAs. The RNAs in the examples are obtained from the R-FSCOR dataset.

The first example is the alignment between a GNRA motif (PDB 1zih, chain A) and a *Deinococcus radiodurans* 23S rRNA (PDB 1njo, chain 0). The aligned regions and the corresponding secondary structures are shown in Figure 4.4. Although it has a decent stack mapping (residue 2526 is paired with residue 2540 and residue 2527 is paired with residue 2539), the alignment produced by LaJolla is disordered: residue 2526-2527 should be at the 5' side of residue 2539. For SARA, the aligned region of the rRNA is highly conserved with a

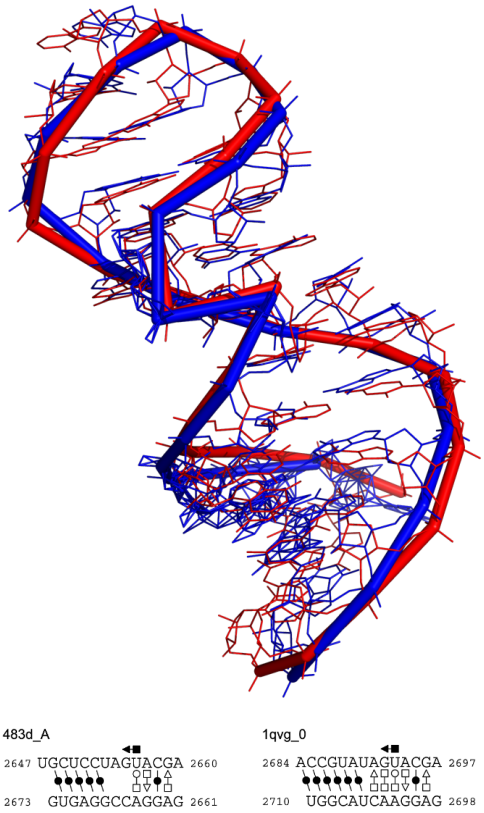

Figure 4.5: The alignment result of STAR3D for the sarcin-ricin motif (PDB: 483d, chain A) and the 23S rRNA (PDB: 1qvg, chain 0). The blue ribbon shows the 3D structure of the sarcin-ricin motif and the red ribbon shows the 3D structure of the 23S rRNA. The base pairs in the aligned regions are predicted by MC-Annotate.

segment of the motif. However, only one strand of the motif is aligned and the corresponding loop regions are very different. One possible reason is that the unit-vectors used by SARA only describe the conformation of the backbone. Furthermore, the alignments of base pairs and the whole 3D structures are computed separately. Thus it may overlook the pairing information if the partial structure alignment achieves the maximum score. ARTS finds the matching of base pairs first and then extends it to both 5' and 3' directions of the RNA strand. In Figure 4.4(c), it can be seen that three base pairs are matched very well, while the 3D structures of the loop regions are distinct. This may be caused by the different treatment of stacks and the corresponding loops in the computation of ARTS. For STAR3D, the entire

motif (residue 1-12) is aligned to the residues 130-141 in the rRNA. The tetraloop of the rRNA has the common structural characteristics of the GNRA motif: the four residues are "GUAA" and the loop is closed by a "C↔G" pair. The 3D structural alignment in Figure 4.4(d) also shows that the detected region in the 23S rRNA has a very high probability to be a GNRA motif. Similarly to the strategy of ARTS, the conserved stack regions are detected first in STAR3D. However, STAR3D ensures that the corresponding loops should have similar rigid transformation with the stacks, otherwise, the entire alignment will be assigned a low score.

The sarcin-ricin motif is an important structural motif involved in the interaction between rRNAs and the elongation factors [150]. In the R-FSCOR dataset, there is one chain of 23S sarcin-ricin motif (PDB: 483d, chain A) and 22 23S rRNAs, 11 from *Haloarcula marismortui* (*H.m.*) and 11 from *Deinococcus radiodurans* (*D.r.*). The 3D structural alignments of the motif and all the 23S rRNAs are analyzed. Compared with the GNRA motif, sarcin-ricin is more complex: it contains 27 residues, 6 canonical base pairs and 4 non-canonical base pairs. For ARTS and LaJolla, no highly conserved region is found in those rRNAs. SARA can detect potential motifs in all the *H.m.* 23S rRNAs, but none in the *D.r.* 23S rRNAs. STAR3D not only finds the motif candidates in *H.m.* 23S rRNAs, but also in 6 of 11 *D.r.* 23S rRNAs. To verify the detected motifs, the docking results of the alignments are analyzed. An example alignment of the sarcin-ricin motif and one *H.m.* 23S rRNA is shown in Figure 4.5. From the base pair profiles and the 3D structures, it can be seen that the hairpin loop (residues 2684-2710) has a high probability to be a sarcin-ricin motif. By checking the base pair annotation of all the *D.r.* 23S rRNA, we can find the structural variance in the motif regions. For the five rRNAs in which STAR3D can not detect the motifs, only four base pairs are annotated in the helix of the motif regions. The different annotation between these

regions and the sarcin-ricin motif, which has five base pairs in the stack, disallows STAR3D to make the correct prediction.

4.3.4 Structural alignments of homologous rRNAs

We also tested the performance of STAR3D on aligning the 3D structures of homologous 16S and 23S rRNAs. The benchmarking dataset includes three 23S rRNA chains from three different species: *Haloarcula marismortui* (PDB 1s72, chain 0), *Escherichia coli* (*E.coli*; PDB 2aw4, chain B), and *Thermus thermophilus* (*T.th*; PDB 2j01, chain A); and two 16S rRNA chains from two species: *T.th.* (PDB 2avy, chain A) and *E.coli* (PDB 1j5e, chain A). We also used the two manually generated alignments of these 16S rRNAs as references. The first one is the Crystallographer alignment, which is implied in the numbering system used by the crystallographers; the second one is the Composite alignment, which is hand-crafted and based on comparative analysis. They have been used as benchmarking dataset before [124, 166] (note that no such alignments are available for 23S rRNAs). R3D Align is also included in this benchmarking. All the tools are installed locally on a DELL XPS Desktop with Intel i7-4770 CPU at 3.40 GHz with 16 GB of RAM. To make the comparison fair, only one thread is allowed in the experiments.

First, we compare the running time of five tools for the rRNA alignments (see Table 4.2). It can be seen that STAR3D improves the time efficiency of the other tools by ten to a thousand folds. The adoption of the MaxSub algorithm [143] to refine the original 3D structural alignments may cause the huge time consumption in SARA. For STAR3D, the major running time reduction comes from the computation of loop alignments. Assuming

Table 4.2: Running time (in seconds) of ARTS, LaJolla, SARA, R3D Align and STAR3D for the homologous alignments of 16S and 23S rRNAs. Best performance is set to **bold**. The preprocessing time is not included for ARTS, SARA, R3D Align, and STAR3D.

rRNAs	ARTS	LaJolla	SARA	R3D Align	STAR3D
H.m. and *E.coli* 23S	117.2	119886.7	27035.2	751.7	**1.7**
H.m. and *T.th.* 23S	98.5	125835.9	26184.8	573.4	**2.0**
E.coli and *T.th.* 23S	79.7	152635.6	27467.3	653.2	**1.4**
E.coli and *T.th.* 16S	20.1	16209.3	4714.4	308.9	**1.1**

the total lengths of loop regions for two RNA sequences are n_1 and n_2, the time complexity of loop alignment is $O(n_1 \times n_2)$. In STAR3D, one loop region only needs to be compared with another one marked by the e-stacks. Thus the time complexity is $O((n_1 \times n_2)/m)$, where m denotes the number of e-stack pairs in the consensus. With the relatively large number of stacks in RNAs with complex structures, our method can significantly improve the efficiency of the loop alignments.

Table 4.3: Summary of alignments between the *T.th.* and *E.coli* 16S RNAs (PDB: 1j5e, chain A and PDB: 2avy, chain A). Best performance is set to **bold**.

	Manual		3D Structural Alignment				
	Crystallographer	Composite	ARTS	LaJolla	SARA	R3D Align	STAR3D
Number of aligned nucleotides	1488	1414	1116	1106	1343	1400	1466
Agreeing with Composite	1401	1414	1056	1101	1240	**1362**	**1362**
Agreeing with Crystallographer	1488	1401	1081	1030	1276	1354	**1414**

Based on manually generated alignments of two 16S rRNAs, the accuracy of five tools is also examined. The results are shown in Table 4.3. It can be seen that both R3D Align and STAR3D achieve the maximum true positive number if the background dataset is the Composite dataset. The accuracy of STAR3D is slightly lower than R3D Align because it detects more nucleotide matches in the two sequences. However, for the crystallographer dataset, STAR3D outperforms the other three tools. So STAR3D is not only highly efficient but also an accurate algorithm when it is used to align large homologous RNA molecules.

4.4 Discussion and Conclusion

In this article, we have proposed a novel tool, named STAR3D, for RNA 3D structural alignment. First it detects the conserved double-stranded regions in two input RNAs by joining the matches of small stack components. Then the consensus of stacks is assembled based on the 3D structural similarity and 2D compatible relationship. Its underlying tree-like topology leads to the ordering of loop regions. In addition, the rigid transformation of the aligned stacks can guide the 3D alignment of the loop regions. As a result, each loop only needs to be compared with its partner in the other sequence by using the superimposition of the conserved stacks. Finally, we combine the stack alignment and all the loop alignments as the final result. This "two-step" strategy is derived on the basis of three observations. First, insertions and deletions are rarely seen in the conserved helical regions, which means that the ungapped extension is applicable to the stacks; second, the 3D structural similarity of conserved stacks is higher than that of random stacks; third, the stack regions are easier to annotate, even for low resolution PDB structures, so the stack alignment can be used as an anchor for the loop alignment. By integrating these properties into the design, STAR3D avoids the complex computation of secondary structure comparison. Furthermore, the one-to-one loop alignments, which replace the all-to-all base matching in entire single-stranded regions, reduce the running time of STAR3D for large RNAs significantly. The benchmark results show that the prediction accuracy of STAR3D outperforms the state-of-the-art tools, and does so with higher efficiency. What's more, STAR3D can be easily implemented with multi-thread support. The detection of e-stack pairs depends on ungapped alignment. The computation at each diagonal can be performed at an individual thread. For maximum clique finding, the Bron-Kerbosch algorithm can be implemented in parallel too. In the last step,

the alignments of loop regions are independent, and can be deployed in different threads as well.

A potential expansion of STAR3D is to implement a local alignment version of the tool. From the experiment of aligning the GNRA motif and the 23S rRNA, it can be seen that STAR3D is sensitive to local similarities in RNA 3D structures. On the other hand, it is natural to convert STAR3D into finding local alignments. In the original method, only the maximum clique in the compatible graph is chosen to build the structural tree. To develop a local alignment approach, We can change STAR3D to deal with multiple cliques. For each one, a local alignment can be generated by only comparing the loops covered by the aligned stacks. With this new method, we anticipate that more structural motifs will be found in the functional ncRNAs.

Another direction for future study is to incorporate comparative methods into STAR3D. There are two approaches. The first one is to use the comparative methods to improve the alignments of the RNAs with low resolution coordinates. The excessive flexibility of atomic positions challenges the prediction of base pairing interactions, which may affect the performance of STAR3D. To solve the problem, we plan to design an iterative pipeline to find the base pairs in the low resolution RNA structures. First, we need a homology of the target that has high resolution 3D structural data. Hence a better annotation of base pairs for the target can be inferred by aligning two RNAs with STAR3D. In the following run, these predicted base pairs can be used as the secondary structural information in STAR3D to generate a more precise alignment. This procedure is continued until no new base pairs can be detected for the target. Considering the high efficiency of STAR3D, the time consumption of the pipeline should be practical. In addition, the low resolution RNAs can be aligned to other RNAs more accurately with the inferred base pairs. The second way is to find the 3D

structural conservation among the RNAs in one family by using comparative methods. A hierarchical clustering based method, which is similar to CLUSTALW [155], is adopted. A 3D consensus structure of two RNAs can be constructed by connecting the centroids of the mapped nucleotides. Then, by merging the sub-clusters we can find the consensus for the whole family and its corresponding multiple sequence alignment.

CHAPTER 5: DE NOVO DISCOVERY OF STRUCTURAL MOTIFS IN RNA 3D STRUCTURES THROUGH CLUSTERING

5.1 Background

Non-coding RNAs (ncRNAs) achieve their specific cellular functions by folding into three-dimensional (3D) structures interlinked by numerous locally stable components. Among them, some highly abundant building blocks called "RNA structural motifs", are found to play important roles which may determine the behaviors of the molecules. For examples, the kink-turn motifs are the important binding sites for 9 proteins in the bacterial 23S ribosomal RNAs (rRNAs) [86]; The cleavage of sarcin-ricin motifs led by the toxin proteins may result in the shutdown of protein synthesis in ribosome completely [50]. Therefore, the identification and understanding of these recurrent structural components are indispensable for the study of RNA molecules. Considering that the number of resolved RNA 3D structures is rapidly increased in recent years, thorough analysis of structural motifs is expected to extend our knowledge of the relationship between RNA architectures and functionalities.

One major computational approach for studying RNA structural motifs is to search homologous instances of known motifs by using comparative methods. Traditionally, similar to the proteins in tertiary structural alignments, the motifs are modeled with their 3D geometric

features, such as backbone conformations or torsion angles. NASSAM [66] and PRIMOS [36] are typical tools which primarily rely on the 3D atomic coordinates. They perform well for some simple motifs, but may not work for complex ones since the underlying computational methods are too rigid to identify the flexible variations in structures. Unlike these two methods, FR3D integrates base pairing interactions into the alignments for RNA structural motifs [134]. The evaluation of 3D spatial distances is constrained by the pairwise interactions which improves the computational efficiency and accuracy dramatically. However, as the most critical characteristics of RNAs, the base-base interactions should be used as key factors in the assessment of structural discrepancy directly [119]. Based on the idea, RNAMotifScan is proposed to search new motif candidates that share highly conserved secondary structural patterns with the query [186]. The benchmarking results show that RNAMotifScan outperforms other state-of-the-art RNA structural motif searching tools, especially for the instances with distinct geometric variations caused by insertions or deletions.

The issue of search tools is that they are based on the existing knowledge of RNA structural motifs, and thus can not be applied to detect new families. To solve the problem, the comparative methods for searching are incorporated into clustering pipelines for the de novo discovery of conserved structural elements. One example is COMPADRES [160], which makes use of PRIMOS to categorize RNA structural motifs in the database of existing RNA 3D structures. Its performance is limited by the rigid alignments, and the clustering results are hard to be applied to the further research due to the complex models only covering 3D geometric information. LENCS (longest extensible non-canonical substructure) adopts a much simpler model which defines the RNA structural motifs as graphs of nucleotides interconnected by base pairs [30]. Thus the structural similarity of two motifs can be evaluated by the maximum common subset in their graphs. With this measurement of similarity, a hi-

erarchical clustering tree is built, and the homologous motifs are classified by cutting it with a universal threshold. LENCS has successfully identified several putative new motifs in three rRNAs without using any tertiary information directly. But its sensitivity to the potential structural variations is relatively low due to the strict matching only allowing the same types of base pairs. A recent approach of classifying RNA structural motifs takes into account all the hairpin and internal loops in the non-redundant RNA 3D structures [122]. Based on FR3D, this pipeline aims at grouping the loop regions conserved in 3D space together with the help of pairing interaction constraints. All the annotated motif instances and families are well organized in an online database named RNA 3D Motif Atlas. Due to the rigid restriction on the 3D geometric discrepancy among cluster members, the method intends to categorize the highly similar components into numerous small groups. On the other hand, it may lose insights of the relationship among motif variations with structural difference. We also developed a clustering framework named RNAMSC for *de novo* RNA structural motif identification in rRNAs [187]. To ensure the high coverage of base pairing information on the RNA sequences, the base-pair annotation of two different tools, MC-Annotate [93] and RNAView [178], are combined. Then the non-canonical base pairs in the loops are compared according to their isostericity [96], and the statistically significant alignments are determined by using P-values which are inferred from the background simulated data. After that, the conserved candidate pairs with high P-values are summarized into a graph, in which the strongly connected subgraphs are retrieved. The experimental results show that RNAMSC not only outperforms LENCS in the recovery of known motifs, but also discovers several novel motif families. Compared with RNA 3D Motif Atlas, our approach assumes that the base pairing interactions, which are the direct indicators of cellular functions, should be adopted to measure structural similarity in the clustering. As a result, RNAMSC can detect the potential motif variations whose 3D structures are distinct from the majority of instances.

Here we propose a new clustering pipeline to automatically detect novel RNA structural motifs by extending RNAMSC. There are three major differences between this framework with the original RNAMSC. First, the new pipeline is optimized for the large-scale inputs, such as the non-redundant RNA structure dataset; Second, all the single-stranded regions in the RNA molecules, including the multi-way junctions, are considered in the classification; Third, the clustering results are post-processed to analyze their functionalities and relationships. By using this new clustering approach, we have identified totally 192 motif families, 68 from hairpin loops, 79 from internal loops, and 45 from multi-loops. Generally, the large clusters contain the pervasive motifs in RNA 3D structures, such as GNRA tetraloop, T-loop, kink-turn, sarcin-ricin, etc. The variations in some motif families which are accidentally separated from the majority can be retrieved back based on checking their secondary and tertiary structural patterns in the downstream analysis. Furthermore, we also discover some novel motifs conserved in both rRNAs and non-rRNAs, such as single guide RNA (sgRNA) in Cas9 complex, Alu domain in the signal recognition particle RNA (SRP RNA), GlmS riboswitch and twister ribozyme. All the clusters and the corresponding annotation are available at the web-site http://genome.ucf.edu.

5.2 Materials and Methods

5.2.1 Data preparation

Our clustering approach is based on the known knowledge of RNA 3D structures deposited in the PDB database [12]. As the paper is written, there are over 3000 experimentally

solved macro-molecular structures containing RNAs. To avoid the possibly statistical biases caused by over similar ones, the Non-redundant List (of RNA-containing PDB structures) from BGSU RNA group [98] was adopted. This dataset eliminated the redundancy both in a single PDB file and among multiple PDB files, while keeping sufficiently diverged homologous structures. The selected 876 PDB files (including 1307 RNA chains) at 4.0 Å resolution threshold in v1.89 NR list were downloaded.

After that, all the plausible pairing interactions in the RNA 3D structures were identified by using MC-Annotate [93] and RNAView [178]. Their annotation results were merged, and the conflicts were solved by taking the MC-Annotation predictions. For each chain, the predicted *cis* Watson-Crick base pairs were retrieved to reveal the A-form helices in the RNA secondary structure. The pseudoknots in the structure were recognized by the program K2N [144] and then eliminated. In the pseudoknot-free secondary structure, the single-stranded region was decomposed into hairpin loops, internal loops (including bulges), and multi-loops by the consecutively nested *cis* Watson-Crick base pairs (≥ 2). The loops without non-canonical base pairing interaction were removed to refined the dataset. Given the fact that some known structural motifs were closed by *cis* Watson-Crick base pairs, the helix ends were also retained. Finally, the orders of the strands in loops were considered to generalize motif candidate instances for the alignment. The two strands in the hairpin loops were concatenated in both ascending and descending orders. For loops with more than two strands, we avoided the excessive number of strand permutations. The loops were converted into circle forms by connecting the 5' ends of left most strands and 3' ends of the rightmost strands, and then only the permutations in the cyclic orders were used in the further computation.

5.2.2 Loop alignment and clustering

All the motif candidates were grouped into three different datasets: HL (from hairpin loops), IL (from internal loops and bulges), and ML (from multi-loops). The HL dataset contained 1036 instances, the IL dataset contained 1868 instances, and the ML dataset contained 2778 instances. In each dataset, an all-to-all alignment of the loops was performed using RNAMotifScan. Because RNAMotifScan was developed for searching which treated queries and targets differently, any loop was aligned twice to its partner, as the query in first one and as the target in the second one. The two corresponding Z-scores were computed with the alignment score distributions of queries, and the smaller one was assigned to the candidate pair as the numerical measurement of their tertiary structural similarity.

After that, three weighted graphs for different loops were constructed from the alignment results. In these graphs, the vertices represented the loops and the edges were labeled with the standardized scores. Noted that internal loops and multi-loops had multiple candidates with different orientations of strands. The maximum Z-score of all the candidate alignments for two loops was chosen as the weight of the edge, and a cutoff was set to determine whether it should be removed or not. The strongly connected sub-graphs were identified with a CAST-like clique finding algorithm [10] in the processed unweighted graph.

During the alignment and the clustering, the parameters of the pipeline were tuned to generate the most reliable results for the motif discovery. For RNAMotifScan, 30 sets of parameters were used: the weights for sequence similarity and structural similarity can be (0.2, 0.8) and (0.4, 0.6); the gap start and extend penalties can be (3, 2), (6, 2) and (6, 4); the penalty of missing one base pair in two inputs can be 1 to 5. Different Z-score cutoffs,

ranging from 1.0 to 3.0 with step of 0.1, were also applied in the graphs. Therefore, there were 930 (30 $\times$ 31) different clustering results for each loop dataset. To benchmark their accuracy, we compute the sensitivity and specificity by using the known motifs in rRNAs [187]. The one with highest sensitivity and 0 specificity was chosen for the further analysis. Based on these criteria, we chose the five parameters and the cutoff to be 0.2, 0.8, 6, 4, 3, and -1.1 for the HL dataset; 0.2, 0.8, 6, 2, 2, and -2.2 for the IL dataset. Since no enough known motif instances in multi-loops to conduct the evaluation, the parameters for IL dataset were used in the clustering for ML dataset directly.

5.2.3 Motif family identification

We extract the potentially conserved motif families from the clusters for HL, IL, and ML datasets by using both computational methods and visual inspection. For two members of a cluster, the aligned regions enclosed by base pairs were detected, and the fragments outside of all base pairs were filtered out. This "core" of the alignment must satisfy two requirements to keep itself for the further analysis. First, the length of the base-pair surrounded alignment must be greater than 3; Second, their root-mean-square deviation (RMSD) must be greater than 4Å. As a result, the edges for the false conservations which did not meet the two conditions were deleted from the graph. After that, the dangling loops were believed to be the false predictions and removed from the cluster. Considering different motifs may be grouped into the same cluster if they share common patterns in their secondary structures [189]. The 3D structures of the remained loops in clusters were manually checked to categorize them into sub-groups. Then the secondary structural consensus and the key 3D structural features of sub-groups were extracted. With these critical properties, the possible functionality of the

motifs in the sub-groups were obtained by literature research, and the relationship among sub-groups in different clusters was observed by comparison. Finally, we designed an ID system to refer the clusters and sub-groups. The cluster ID contains two fields: a loop type prefix and a cluster index suffix (e.g. IL1). Based on that, the sub-group ID is defined as cluster ID followed by the sub-group index, separated by an underscore character (e.g. IL1_1).

5.3 Results

5.3.1 Summary of the clustering results

To evaluate the clustering results, 10 well-studied motif families are analyzed. The clusters containing the maximum numbers of instances for these motifs are chosen as representatives. All the other motif instances not in the representatives are annotated as plausible variations. Table 5.1 summarizes the clustering results for the 10 motif families, including the prediction accuracy and the numbers of variations. The precision is computed by dividing the number of true motifs with the size of the cluster. Note that if one loop consists of several motifs, it will be counted multiple times. From the table, we can see the clustering results of GNAA and GNGA motifs are accurate because they are highly conserved in both sequences and secondary structures. The related variations mainly come from the combination with sarcin-ricin motifs, which will be discussed later. T-loops are relatively hard to be clustered together, due to their low sequence identity and simple base pairing patterns. It indicates the weight of structural similarity should be set much greater than the weight of sequence

Table 5.1: The clustering results of 10 well-known motif families. The numbers in the brackets show the variations detected from the datasets not containg the clusters.

Motif name	Cluster ID	# of true instances	size of cluster	precision	# of variations
GNAA	HL1	85	87	98%	5
GNGA	HL3	45	45	100%	14
T-loop	HL4	29	31	94%	55(6)
Sarcin-ricin	IL3	47	56	84%	18(14)
Kink-turn	IL5	25	39	64%	38
Hook-turn	IL6	26	31	84%	0
C-loop	IL8	16	21	76%	7
E-loop	IL9	16	21	76%	7
Tandem shear	IL13	22	30	73%	5
Reverse kink-turn	IL21	19	20	95%	6

similarity when searching T-loops. Both sarcin-ricin and kink-turn have lots of variantions, which only share the key 3D structural features but not the secondary structural patterns with the majority of instances in the representative clusters. One possible reason is that the binding activity may disturb the base pairing interactions in them, and we will show several examples in the later sections. The precision of the kink-turn cluster (IL5) is relatively low because E-loops have two common non-canonical base pairing interactions with kink-turn motifs. Hook-turn has a unique base pairing pattern, so it is easy to identify. Although C-loop is hard to detect due to its crossing base pairs, our pipeline still achieves acceptable results for it. All the other three motifs, E-loop, tandem shear, and reverse kink-turn, consist of tandem non-canonical base pairs. E-loop and tandem shear have similar 3D structures, so we mainly use their secondary structural features to distinguish them.

Besides the motifs in the table, we have discovered other functional ones in the clustering results. The first example is the well-known tetraloop receptor in group I intron (IL4_1 and IL22_1) [1]. Some of them are used in the target molecules to maximize its crystallizability [46]. The L1 protuberance of 50S rRNA and mRNA are also clustered together in IL18. It has already been proved that they have both similar 3D structures and binding activities

[115]. We also detected the motifs that are conserved both in mitochondrial 16S rRNAs and bacterial 23S rRNAs [140]. The identification of these known functional motifs indicate that the clustering results can be applied to further analysis for new motifs.

5.3.2 Novel instances of known motifs

5.3.2.1 Tetraloops

Tetraloops are basic building blocks of RNA 3D structures which are important for thermodynamic stability and binding activity of the molecules [47, 141]. The most frequent two types of tetraloops are GNRA loops [172] and UUCG loops [41]. In addition, GNRAs can be categorized into GNGA loops and GNAA loops. In our clustering results, the majority of GNAA, GNGA, and UUCG motifs are in IL1, IL3 and IL6. Some other GNAAs and GNGAs are found to be linked with sarcin-ricin motifs. One instance of this motif module is shown in Figure 5.1 (a). This loop is from the region C3120-A3136 in the *Homo sapiens* (*H. sapiens*) mitochondrial 16S rRNA. It can be seen that the 3D structure of A3125-G3131 is highly conserved to a GNAA reference. We also find that the corresponding region in the *Haloarcula marismortui* (*H. marismortui*) 23S rRNA contains a GNGA motif, which indicates GNAA and GNGA are interchangeable in this module. Similar modules for UUCG have been also detected. One example is shown in Figure 5.1 (b), which is in the *H. marismortui* 23S rRNA. The "U-shape" turn in this loop is docked with the blue UUCG tetraloop precisely in the 3D space. Base on the observation, We may hypothesize that the combination of sarcin-ricin and tetraloop should be a very common module in RNA 3D structures.

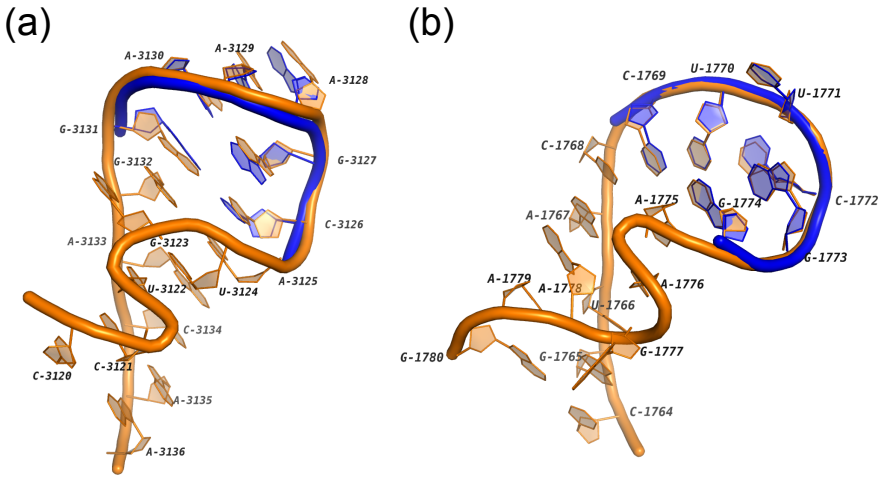

Figure 5.1: The 3D structures of two RNA motifs containing both tetraloops and sarcin-ricins. (a) The hairpin loop in the *H. sapiens* mitochondrial 16S rRNA (PDB: 3J7Y, chain: A) which contains a GNAA tetraloop. The blue tube shows a superimposed GNAA tetraloop in the *H. marismortu* 23S rRNA (PDB: 4BW0, chain: A). (b) The hairpin loop in the *H. marismortui* 23S rRNA (PDB: 1S72, chain: 0) which contains a UUCG tetraloop. The blue tube shows a superimposed UUCG tetraloop in the *Methanococcus vannielii* mRNA fragment.

5.3.2.2 T-loops

T-loop is a compact U-turn-like loop which was originally discovered in transfer RNA (tRNA) [129]. After that, many T-loop instances have been identified in a variety of ncRNAs, ranging from rRNA to riboswitch [24]. Our clustering results cover almost all the known T-loops in the hairpin loops. What's more, We also find two instances of T-loop in the internal loop cluster IL26. One of them is in the Thi-box (thiamine pyrophosphate sensing) riboswitch and known for the ligand 4-amino-5-hydroxymethyl-2-methylpyrimidine (HMP) [139]. The other one is discovered in a T-box stem I RNA. Figure 5.2 shows its structures and the 3D alignment to the T-loop in a tRNA(leu). Note that their secondary structure consensus consists of one *trans* S/H and one *trans* W/H base pairing interactions. The difference is that in the tRNA the base pairs exist in a hairpin loop, while the two interactions in the T-box

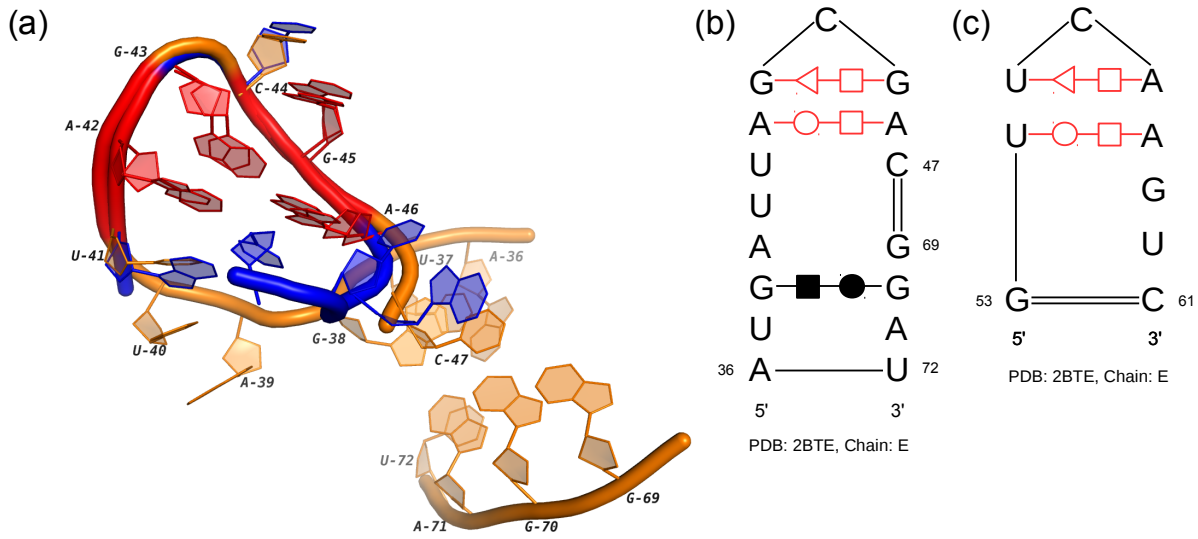

Figure 5.2: The 3D and secondary structures of an internal loop in T-box stem I RNA (*Oceanobacillus iheyensis*) and a hairpin loop in tRNA (*Thermus thermophilus*). (a) The 3D docking of two loops. The orange tube represents the internal loop in the T-box stem I RNA (PDB: 4TZZ, chain: C) and the blue tube represents the hairpin loop in the tRNA (PDB: 2BTE, chain: E). (b) The secondary structure of the internal loop. (c) The secondary structure of the hairpin loop. In (a), (b), and (c), red color marks the structural consensus of the two loops.

stem I RNA bend one strand of the internal loop to a U-shape turn. Considering relatively large size of the twisted strand, the third interaction at G38/G70 should be important to the stability of the entire loop. This T-loop also works with another T-loop (the homology is 4MGN_C:51-63 in HL8) in the same RNA to stack on tRNA elbow [182]. The similar binding behavior is also in RNase P and ribosomal RNA, so the study of this T-loop and its partner may provide useful information for searching more homologous modules.

5.3.2.3 Kink-turn

Kink-turn is a motif in the internal loop region with an asymmetrical architecture [86]. Its key feature is the tight kink at the backbone of the longer strand, which causes the axes of the two helical stems differ by about 120°. In a real cellular environment, kink-turn may adopt a dynamic conformation [136]. To maintain the k-shape geometry, the motifs require the presence of metal ions [108], or the binding with proteins [158]. We detected two kink-turn-like motif instances in the cluster IL37. Note that the loop in the 16S rRNA was detected in our previous work [187]. With the newly discovered instance, we can analyze their conserved patterns and the related functions. Figure 5.3 shows their secondary and 3D structures. It can be seen that all base pairs can be matched together if the red nucleotides are ignored. Compared with the consensus base pairing pattern of common kink-turns [189], these two instances form the 3D kinks by three base pairs (G247/A282, A246/G278, A246/G281 in 1FJG and G18/A48, A17/A44, A17/G47 in 3RW6). In the common kink-turns, the Watson-Crick base pairs, C242/G284 in 1FJG and U11/G50 in 3RW6, should be followed by two continuous non-Watson-Crick base pairs. However, in these two loops, the two interactions are separated by the nucleotides marked with red color. In Figure 5.3, these red nucleotides form the bulges at the shorter strand, which do not exist in the common kink-turns. What's more, both red regions have binding functions. According to the results of MC-Annotate, the nucleotide U244 is paired with A893 in another loop region. On the other hand, the large bulge containing the flipped out nucleotides, A13, G14, and A15, is the binding site of the TAP protein and critical to the formation of CTE-TAP complex [152]. So based on the function similarity, we may suggest that the secondary structural pattern is important to the binding activity in the kink-turn motifs.

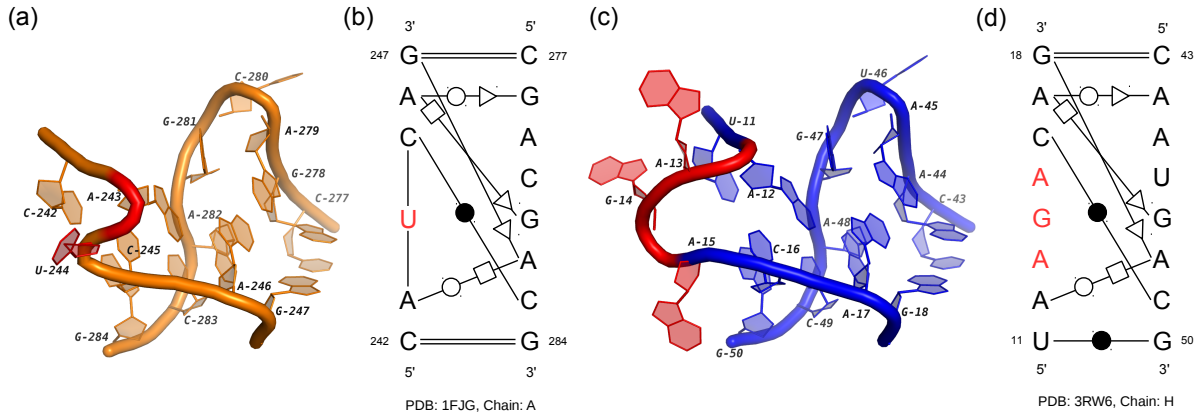

Figure 5.3: The 3D and secondary structures of two internal loops in a 16S rRNA (*Thermus thermophilus*) and a CTE rRNA. (a) and (c) are the superimposed 3D structures of two loops in the 16S rRNA (PDB: 1FJG, chain: A) and the CTE RNA (PDB: 3RW6, chain: H). (b) and (d) are the secondary structures of (a) and (b). In (a), (b), (c), and (d), red color marks the nucleotides with binding activities.

5.3.2.4 Sarcin-ricin

Sarcin-ricin motif is first found in the large ribosomal subunit as the attacking site of two protein toxins, ricin and α-sarcin. The catalyzation among them will impact the binding between elongation factors and ribosome, which may result in the cessation of the protein synthesis [67]. More sarcin-ricin instances with similar structural features have been discovered in other RNAs, including 5S and 16S rRNAs, by using computational methods [95, 186]. In our clustering results, the majority of sarcin-ricins are also detected in rRNAs (see the cluster IL3). Their secondary structures are almost the same as the widely used consensus [95], and the 3D structures are highly conserved with the known instances. On the other hand, we also find some functional loops that share structural features with sarcin-ricin. Here, we present two possible variations of sarcin-ricin whose secondary and 3D structures are shown in Figure 5.4. The first loop is in the cluster IL38. Based on the secondary structure, the S-shape turn in its 3D structure is mainly supported by two non-canonical

base pairs (A415/G428 and A414/A430) and one outward stacking interaction (G428/A430). All these three pairing interactions are in the consensus of sarcin-ricins [189]. However, in common sarcin-ricins, the *cis* H/W base pair U429/A431 should be a *cis* H/S interaction between U429/A430. The possible reason for this difference is that the strand U427-C433 is longer than in the consensus. This motif instance also exhibits a special structural property: the bulge at the strand G409-G416. Actually, it has important molecular functions that S4 protein interacts with the flipped out A412 and its backbone contacts to G410 and A411 [16]. So the two base pairing interactions not in the consensus, A411/A430 and G413/G428, may be important for the structural stability disrupted by the long range linkages. We may hypothesize that this motif is a sarcin-ricin disturbed by the protein binding activity. And the comparison of its secondary structure pattern with the sarcin-ricin consensus may help us to detect potential RNA-protein interactions.

Another interesting loop is in the cluster IL62. We call it "double S-turns" because there are two symmetrical S-shape turns in its 3D structure (see Figure 5.4 (c)). In the existing model for ligand-induced folding of the TPP riboswitch, this loop is the TPP-bind pocket which is critical for the ligand recognition [139]. The two nucleotides, U62 and U79, shape the pocket by protruding into solution and losing the stacking effects to the adjacent bases. From Figure 5.4 (d), it can be seen that there are two stacking interactions, A61/C63 and G78/A80, to enforce the local stability around these two nucleotides. They also cause the large turns in the S-shape structures. On the other hand, the other two non-canonical base pairs tight the two strands together. The analysis of this internal loop indicates that the stacking effect between discontinuous bases is important evidence of detecting specific structural motifs, such as bulge and S-turn. In addition, this specific organization of pair-

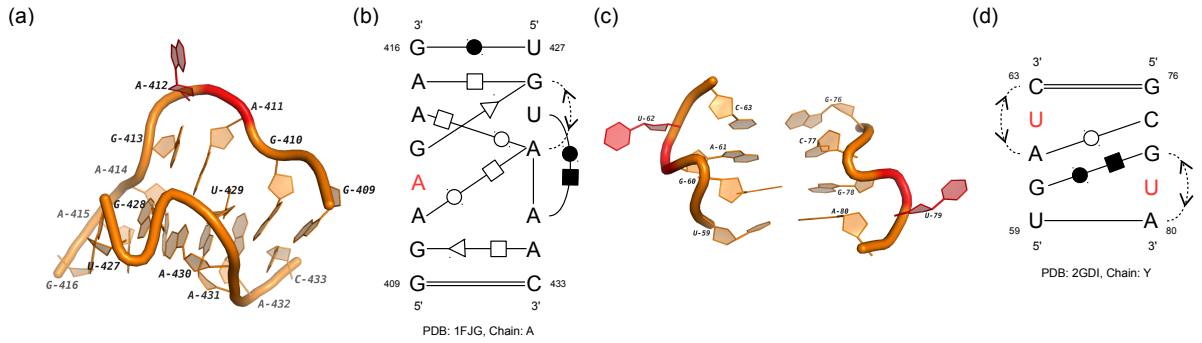

Figure 5.4: The 3D and secondary structures of two internal loops in a 16S rRNA (*Thermus thermophilus*) and a TPP riboswitch (*Escherichia coli*). (a) and (c) are the 3D structures of two loops in the 16S rRNA (PDB: 1FJG, chain: A) and the TPP riboswitch (PDB: 2GDI, chain Y). (b) and (d) are the secondary structures of (a) and (c). In (a), (b), (c), and (d), red color marks the nucleotides with binding activities.

pair ineractions, including pairing interactions and stacking interactions, may be important to form pocket-like 3D structures.

5.3.3 Novel motif families

5.3.3.1 Novel motif families in the hairpin loop regions

The first potential motif family mainly contains four different instances from HL2_1 and HL53_1. One of them is the loop 10 in the yeast 18S rRNA [94]. The other three are the "stem loop 1" in the sgRNA of the Cas9-sgRNA-DNA ternary complex [5, 116]. It is newly discovered in the complex by studying the crystal structure of Cas9 [116]. The mutation of residues interacted with stem loop 1 results in decreased DNA cleavage activity of the CRISPR-Cas system, which indicates the loop is essential for the formation of the functional

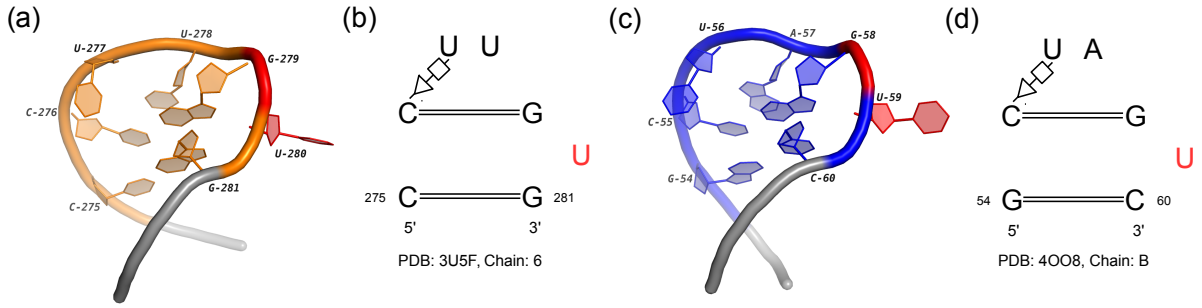

Figure 5.5: The 3D and secondary structures of two hairpin loops in an 18S rRNA (*Saccharomyces cerevisiae*) and a Cas9-sgRNA-DNA complex. (a) and (c) are the superimposed 3D structures of the loops in the 18S rRNA (PDB: 3U5F, chain: 6) and the sgRNA (PDB: 4OO8, chain: B). The extension of two loops is shown in gray. (b) and (d) are the secondary structures of (a) and (c). In (a), (b), (c), and (d), the red color marks the nucleotides binding with the proteins.

Cas9-sgRNA complex. Figure 5.5 shows the high similarity between these two internal loops in terms of both geometric and base pairing patterns. Except the helix ends, all the other interacted bases are identical in two loops. The continuity of the stacks is broken by U280 and U59 that are marked with red color in the figures. Both of them flip out from the stems and cause the turns in the backbone of two loops. The most critical feature is that they have similar functional roles: U280 interacts with L24e protein through the eB13 bridge in the hyper-rotated state [11, 57]; U59 in the sgRNA hydrogen bonds with Asn77 in the bridge helix of Cas9 [116]. So these loops are not only conserved in 3D structures but also the functions, which implies the potential closely relationship between the base pairing pattern and the protein binding activity.

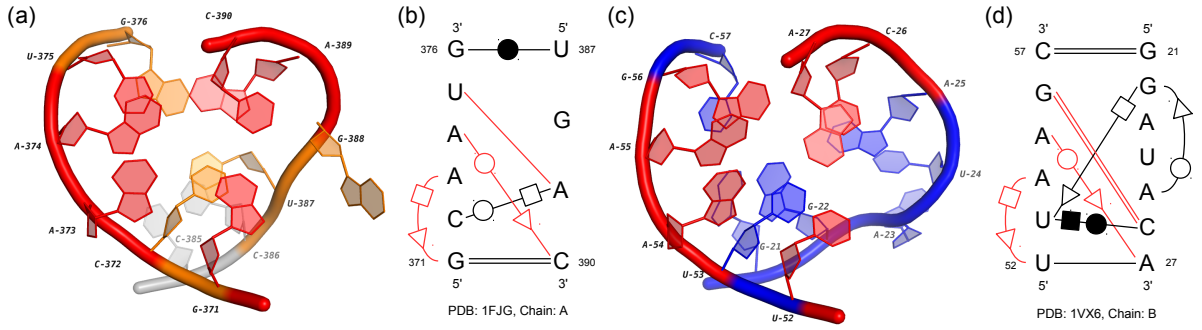

Figure 5.6: The 3D and secondary structures of two internal loops in a 16S rRNA (*Thermus thermophilus*) and a 5S rRNA (*Plasmodium falciparum*). (a) and (c) are the superimposed 3D structures of two loops in the 16S rRNA (PDB: 1FJG, chain: A) and the 5S rRNA (PDB: 1VX6, chain: B). (b) and (d) are the secondary structures of (a) and (c). In (a), (b), (c), and (d), the conserved base pairing interactions are marked as red. The extension in (a) is marked with gray color.

5.3.3.2 Novel motif families in the internal loop regions

One interesting highly significant cluster for internal loops is IL16_1 which contains 6 instances. Four of them are conserved regions in 16S rRNAs, and two of them are the loop B in the 5S rRNAs. We choose representatives from two sub-groups and describe their 3D and secondary structures in Figure 5.6. From the results of superimposition, we can see that the consensus of interactions in two loops, which are shown in red, are highly conserved in 3D space. The corresponding base pairs in the secondary structures are from the same groups in the isostericity matrices [96]: U375-A389 and G56-C26 belong to *cis* W/W I_1; A374/C390 and A55/A27 belong to *trans* W/S I_1; A373/G371 and A54/U52 belong to *trans* H/S I_1. So they are co-variations, and the interchange between them will maintain the 3D structures of the loops. What's more, although the interactions of C372/A389 (*trans* W/H) and U53/C26 (*cis* H/W) are not the same, the geometric relationship of bases in them are quite similar. Therefore, these two base pairs may also contribute to the structural similarity of these two internal loops.

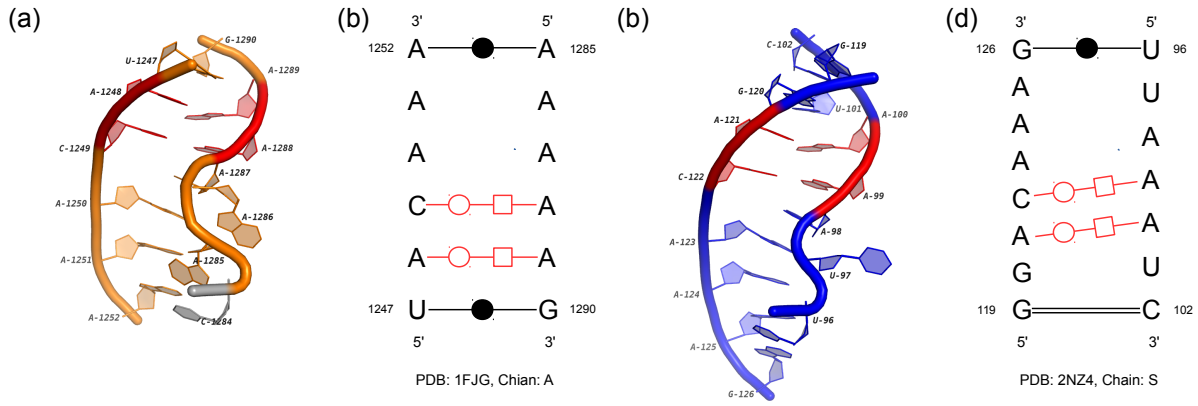

Figure 5.7: The 3D and secondary structures of two internal loops regions in a 16S rRNA (*Thermus thermophilus*) and a GlmS riboswitch (*Bacillus anthracis*). (a) and (c) are the superimposed 3D structures of two loops in the 16S rRNA (PDB: 1FJG, chain: A) and the GlmS riboswitch (PDB: 2NZ4, chain: S). (b) and (d) are the secondary structures of (a) and (c). The conserved base pairing interactions are marked as red. The extension in (a) is marked with gray color.

Then the major difference between two motif instances comes from the regions U387-G388 and G21-A25. First, the lengths of two regions are not the same, which suggests a potential insertion in the loop of 5S rRNA. It also can be seen from the Figure 5.6 that a significant feature shared between them is the turn on the phosphate backbone. However, the backbone of the internal loop in 5S rRNA (the blue one) turns with a slightly large angle. The reason may be the *trans* H/S base pairing interaction between G22 and U53. Although the 3D structures of two regions are not totally the same, they actually may have similar molecular functions. Based on the results of MC-Annotate, the nucleotide G388 in the 16S rRNA, which flips out from the stem, is interacted with C58. For the region in the 5S rRNA, a possible contact to helix 89 in 23S rRNA has been identified by a SELEX (systematic evolution of ligands by exponential enrichment) experiment [90]. It is hypothesized that A23 is the possible binding site due to its base twisting further than the backbone. Moreover, the base pairing consensus we detected here should be very critical for their interlinking functions.

Another possible functional motif is discovered in the cluster IL42. One instance in this cluster is from the 16S rRNA of *Thermus thermophilus*, while the other two are actually the same internal loop in the GlmS riboswitch of *Bacillus*. Riboswitches are metabolite-sensing RNAs that can directly control the synthesis of downstream genes [169]. By binding to specific ligands, their structures are rearranged to terminate the transcription or hinder the translation . However, unlike other riboswitches, the GlmS riboswitch does not alternate its structure upon the binding of glucosamine-6-phosphate (GlcN6P) [64]. Instead, the binding activity results in a cleavage on the GlmS mRNA which reduces the GlcN6P synthetase production greatly [170]. So it is also called "GlmS ribozyme". The internal loop studied here is interlinking two helices, P4 and P4.1, in the GlmS riboswitch. Its secondary and 3D structures are aligned with those of the loop in 16S rRNA, and the results are shown in Figure 5.7. We can see that although both strands of the loop in Figure 5.7 (a) are shorter than those of the loop in Figure 5.7 (b), the consensus marked by red color is highly conserved in sequences, base pairing interactions, and 3D structures. The "S-shape" turns in the regions C1284-A1287 and U96-A98 are important common features of two loops too. In the GlmS riboswitch, the turn is supposed to pack obliquely into the minor groove of P2.1 helix, which is important for the GlcN6P binding [26]. On the other hand, we also find that the flipped out nucleotide A1287 in the 16S rRNA also forms two interactions with A1353 and A1370. So the bulge-like structures may be indicators for the long-range tertiary interactions. The discovered motif may also be critical for the stability of the large internal loops whose structures are disturbed by intra-molecular linkages.

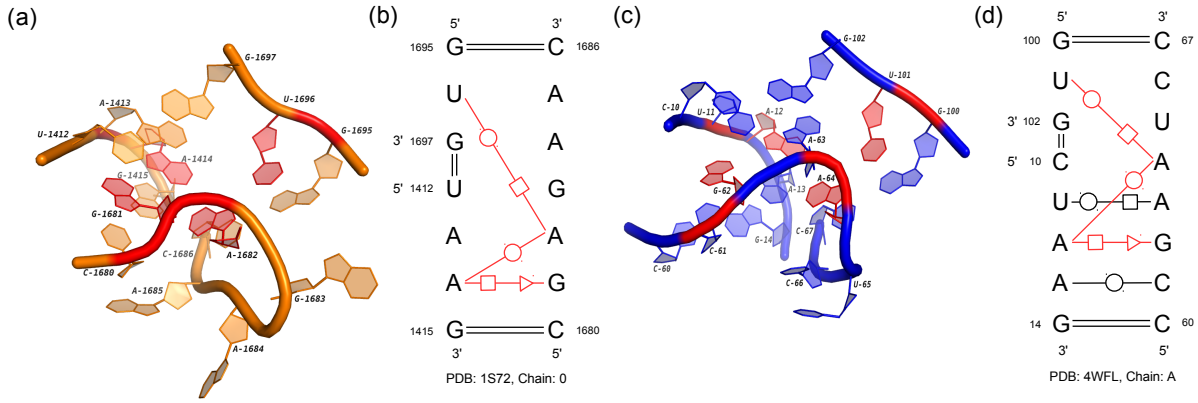

Figure 5.8: The 3D and secondary structures of two multi-loops in a 23S rRNA (*Haloarcula marismortui*) and the Alu domain of a SRP RNA (*Bacillus subtilis*). (a) and (c) are the superimposed 3D structures of two loops in the 23S rRNA (PDB: 1S72, chain: 0) and the SRP RNA (PDB: 4WFL, chain: A). The conserved base pairing interactions are marked as red. (b) and (d) are the secondary structures of (a) and (c).

5.3.3.3 Novel motif families in the multi-loop regions

The first potential novel family in multi-loops is obtained from the sub-group ML2_1. Ten members are conversed regions from 21S, 23S, 25S and 28S rRNAs, and the last one comes from the Alu domain of a signal recognition particle (SRP) RNA (*Bacillus subtilis*). SRP is a highly diverse ribonucleoprotein complex existing in all three kingdoms of life [130]. The RNA in it can be divided into two functional domains, and one of them, the Alu domain, arrests protein biosynthesis by blocking the elongation factor entry site [142, 173]. Then by hindering the translation, SRP can prevent membrane proteins from being prematurely released from the ribosome. The multi-loop in the cluster is the one interlinking helix 1, helix 2 and helix 5a in the SRP RNA. Figure 5.8 shows the comparison of its secondary and 3D structures with those of the loop in 1S72. Both loops have three strands, which are inter-connected by three highly conserved non-canonical base pairs: G1681/A1414 (*trans* S/H), A1414/A1682 (*trans* W/W) and A1682/U1696 (*trans* H/W) in 1S72, G62/A12 (*trans*

S/H), A12/A64 (*trans* W/W) and A64/U101 (*trans* H/W) in 4WFL. The eight interacted nucleotides are marked with red color in Figure 5.8 (a) and (c). Note that the 3D geometric patterns of the consensus are quite similar in two loops, except there is an insertion A63 between G62 and A64 in 4WFL. The structural difference between the Alu domain of SRP RNA in mammalian and bacteria may explain the potential function of the motif. The G-A-A-U 4 base platform observed in *Bacillus subtilis* (bacteria) is absented from the Alu domain in eukaryota. Previous experiments have already shown that the 5' region of human Alu domain is very flexible and SRP9/14 proteins are required to stabilize the conformation and induce the binding to 50S rRNA [165]. On the other hand, the bacterial Alu domain adopts a closed conformation directly with the help of the 4 base platform. This evidence may suggest that the discovered motif is critical to the stabilization of the local structure that binds to proteins.

Another interesting sub-group for multi-loop is ML17_1, which contains three conserved regions in 23S rRNAs and one instance in *env22* (type P1) twister ribozyme. As a small self-cleaving ribozyme, twister presents in many species of bacteria and eukaryota. It has been identified by using bioinformatics method recently, and the name comes from the ancient Egyptian hieroglyph "twisted flax" which resembles the 3D structure of the molecule [131]. The further research shows that twister may play a similar role as the hammerhead ribozyme in the biological systems. Moreover, the instances of twister are categorized into three groups, type P1, type P3, and type P5, which can circularly permute to each others. The crystal structure of the twister used here comes from a type 1 instance. To compare it with the multi-loop in 23S rRNAs, we pick the one in 1S72 as a representative. Figure 5.9 shows the secondary structures of two loops and the 3D superimposition of their extensions. One common feature is that both of them are interlinked by *trans* S/S base pairing interactions

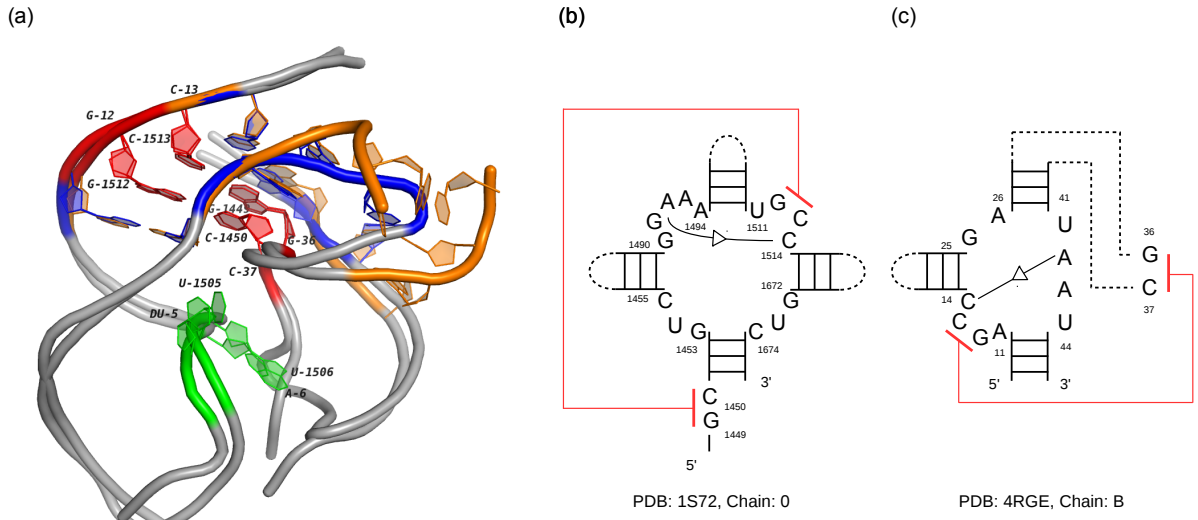

Figure 5.9: The 3D and secondary structures of two multi-loops in a 23S rRNA (*Haloarcula marismortui*) and a *env22* twister ribozyme (synthesized). (a) The 3D docking of two loops. The yellow tube shows the multi-loop in the 23S rRNA (PDB: 1S72, chain: 0) and the blue tube shows the multi-loop in the twister ribozyme (PDB: 4RGE, chain: B). The extended regions of both loops are shown in gray. The nucleotides involved in the pseudoknots are labeled (G1512-C1450 and C1513-G1449 in 1S72, G12-C37 and C13-G36 in 4RGE). (b) The secondary structure of the multi-loop in the 23S rRNA. (c) The secondary structure of the multi-loop in the twister ribozyme. The pseudoknots in (b) and (c) are marked with red lines.

(A1492/C1514 and A42/C14). What's more, the neighbors of the paired bases (G1512 and C1513 in 1S72, G12 and C13 in 4RGE) form pseudoknots with nucleotides outside of the multi-loops (C1450 and G1449 in 1S72, C37 and G36 in 4RGE). From the Figure 5.9 (a), it can be seen that their 3D structures, which are marked with red color, are highly conserved. Note that the orange multi-loop in 1S72 has four strands, while blue one only has three strands. The extension of the $4th$ strand in the orange loop (C1455→G1453) involves in the formation of the pseudoknot. Although not direct substitution, the blue loop has one sharply bent strand (G25→A26) who makes an 180-degree turn to serve the interaction. This interesting case that different secondary structures result in highly homologous 3D structures may suggest the tertiary structural pattern is highly important.

We also extend the 3D alignment to the P2 and P4 helices of the twister ribozyme to study its local structural similarity with the 23S rRNA. Figure 5.9 (a) shows that the two RNAs are quite conserved in these 40-nt regions. The self-cleavage sites in the twister, dU5 and A6, are highlighted with green color. During the transcription, guanosine and Mg^{2+} are coordinated to the non-bridging phosphate oxygen at the U-A step for cleavage catalysis and structural integrity. We also mark the corresponding nucleotides, U1505 and U1506, in 1S72 with green color. It can be seen that they share a similar splayed-apart conformation with the cleavage sites in the twister ribozyme. With so many common features, these two regions should be further studied with the experimental effort to confirm their functional correlation.

5.4 Discussion and Conclusion

In this paper, we studied the RNA structural motifs in non-redundant RNA 3D structures by using a *de novo* clustering approach. The single-stranded regions in the corresponding secondary structures are extracted and categorized into hairpin loops, internal loops, and multi-loops. The base pairing patterns in the same type of loops are compared by RNAMotifScan, and then the significant conservations are assembled into a graph. The densely connected sub-graphs are retrieved to form the clusters in which the members share common secondary structural features. In each cluster, by evaluating the alignments, the loops not close to any others in 3D space are removed. The remained loops in the clusters are further analyzed, and then classified into different sub-groups if their 3D structures are distinguishable from critical conformations. Finally, we try to detect the homologous sub-groups in different clusters by measuring the similarity of their secondary and 3D structural

patterns. The clustering results for the known motifs indicate the high prediction accuracy of this new pipeline. Some interesting instances, which not only maintain the key features of known motifs but also exhibit specific structural variations, are found in the downstream analysis. We also identify numerous novel motif families, even in the multi-loop regions.

The in-depth investigation of the clusters provides directions for the further research. First, RNA structural motifs may work together as a "module", such as the hairpin loops containing sarcin-ricins and tetraloops (Figure 5.1), and the two T-loops in the T-box stem I RNA (Figure 5.2). However, all the existing searching tools, no matter what models they use, only focus on detecting the single motifs in isolation. Therefore, a new tool for discovering motif modules may provide essential evidence of the relationship among RNA structural motifs, which is important for the study of RNA structures and their functions. Another problem is to use base pairing interactions to infer the potential binding activities between RNAs and other molecules. The disturbed secondary structures of the kink-turn and sarcin-ricin variations (Figure 5.3 and Figure 5.4) reveal that they may be the indicators of the long range interlinkages. Furthermore, the affected base pairs also have specific patterns which can be easily integrated into computational methods. This approach should be more accurate than many other methods based on indirect measurements, such as using the distances between atoms.

CHAPTER 6: CONCLUSION

Recently, non-coding RNAs are attracting increasing research focus due to their abundance in living cells and versatile biological roles. Generally, their molecular functions are largely determined by the ability to fold into high-order structures. Therefore, the understanding of the ncRNAs' underlying structures is critical to the study of many cellular processes, such as catalysis, regulation, and host defense. In this dissertation, we have presented a suite of computational methods for analyzing the secondary and tertiary structures of RNAs by using comparative approaches. It is anticipated that our computational methods will promote the function annotation and discovery of ncRNAs.

Folding and searching are two major problems in the computational analysis of RNA secondary structures. Due to its limitation to distinguish random sequences and the restriction of the energy parameters, the single-sequence folding will be replaced by the consensus folding if the alignment of homologous RNAs is available. One approach to consensus folding is based on the assumption that high covariance of two sites in a multiple alignment indicates the conservation of base pairing interaction. To evaluate the possibilities of potential base pairs, the numbers of the mutations at the pairing columns are counted. However, these covariances should not be treated equally because their relative positions on the phylogenetic tree are different. For ncRNA searching, one of the major issues is the low efficiency when applying it to the genome-wide dataset. Although many heuristic optimizations have been

purposed to solve the problem, they are either still too complex to annotate all the known ncRNA families or too simplified to make accuracy prediction.

The tool named PhyloRNAalifold is developed to aim at improving the performance of consensus folding algorithm that adopts the covariance model. It counts the potential covarying mutations following the structure of phylogenetic trees, which avoids the over-emphasizing of some insignificant ones. We incorporated the idea into the widely used consensus folding tool RNAalifold, and the benchmarking results show its superior capability of detecting the conserved base pairs. On the other hand, to solve the efficiency issue of ncRNA homology search, the high-throughput structural probing data is incorporated to indicate the pairing attributes of targets. By using this partial information, we can ignore the matching of base pairs in the secondary structural alignment, which is the most time-consuming part of the algorithm. The idea has been implemented as a tool named ProbeAlign by using C++ and benchmarked with the other state-of-the-art ncRNA homology search tool CMsearch. It can be seen that its prediction outperforms CMsearch with a shorter running time, even with filters applied. We also used ProbeAlign to search ncRNAs in the mouse genome with the help of FragSeq probing data.

At the tertiary structure level, both global and local conservation are important. The global 3D structural similarity between two RNA molecules can provide crucial evidence of their evolutionary history. On the other hand, the local stable structural components in RNAs are essential for specific cellular functions, such as RNA-RNA and RNA-protein interactions. In recent years, many computational tools have been designed to compare RNA tertiary structures. Some of them ignore the base pairing interactions, which results in inaccurate predictions, while others rely on the inter-linkages among nucleotides, which causes the high overload in computation. For searching known RNA structural motifs, there are also some

tools to align the local components in RNA 3D structures. Now, the large size of RNA deposition in the PDB database provides us plenty of resources to discover the novel motifs computationally. Some clustering pipelines have been proposed to identify potential motif families automatically by categorizing the conserved structural components together. The concentration on 3D geometric similarity and overlook of base pairing patterns may lead to rigid clustering results which lose lots of plausible instances for the motifs.

STAR3D is a stack-based RNA 3D structural alignment tool that achieved both accuracy and efficiency by adopting the divide-and-conquer strategy. We found that there is small diversity between the 3D structures of the helical regions in RNAs. Then by comparing the distances between superimposed stacks, the non-homologous ones can be determined and filtered out. Combining this information with the topologies of RNA secondary structures, the core consensus of the stack regions in 3D space are extracted. After that, the loops in two RNAs are ordered and aligned one-by-one. From the benchmarking results, it can be seen that STAR3D not only runs faster than other tools but also generates better alignments accurate at both secondary and tertiary structural levels. Besides finding global similarity in RNA 3D structures, we also attempt to detect valuable local building blocks. The original RNAMSC clustering pipeline has been extended to process the large-scale loop dataset retrieved from the non-redundant RNA 3D structures. The highly conserved loop regions are grouped and analyzed to determine the positive motif family members. Based on the clustering results, function annotation is performed to find potential variations of know motifs and discover novel ones with new features.

In conclusion, our works contain four parts. The first tool, named PhyloRNAalifold, distinguishes the covarying mutations on the phylogenetic tree of homologous RNAs and then use the information in the classic energy model for RNA secondary structure folding. The sec-

ond tool, named ProbeAlign, integrates the pairing information embedded in probing data into genome-wide ncRNA homology search to reduce the overhead of structural alignment. The third tool, named STAR3D, uses the stacks in RNA secondary structures to guide the tertiary structural alignment which results in dramatical improvement of performance. We also classify the single-stranded regions in non-redundant RNA 3D structures to de novo discover structural motifs. The underlying algorithms of these tools are expected to inspire more advanced computational methods, and we also hope the downstream findings can help the biological experiment studies of RNA structures.

LIST OF REFERENCES

[1] P. L. Adams, M. R. Stahley, A. B. Kosek, J. Wang, and S. A. Strobel. Crystal structure of a self-splicing group I intron with both exons. *Nature*, 430(6995):45–50, Jul 2004.

[2] T. Akutsu. Protein structure alignment using dynamic programing and iterative improvement. *IEICE TRANSACTIONS on Information and Systems*, 79(12):1629–1636, 1996.

[3] H. Alt and L. J. Guibas. Discrete geometric shapes: Matching, interpolation, and approximation. *Handbook of computational geometry*, 1:121–153, 1999.

[4] C. Ambühl, S. Chakraborty, and B. Gärtner. Computing largest common point sets under approximate congruence. In *Algorithms-ESA 2000*, pages 52–64. Springer, 2000.

[5] C. Anders, O. Niewoehner, A. Duerst, and M. Jinek. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature*, 513(7519):569–573, Sep 2014.

[6] V. Bafna, H. Tang, and S. Zhang. Consensus folding of unaligned RNA sequences revisited. *J. Comput. Biol.*, 13(2):283–295, 2006.

[7] V. Bafna and S. Zhang. FastR: fast database search tool for non-coding RNA. *Proc IEEE Comput Syst Bioinform Conf*, pages 52–61, 2004.

[8] P. J. Batista and H. Y. Chang. Long noncoding RNAs: cellular address codes in development and disease. *Cell*, 152(6):1298–1307, 2013.

[9] R. A. Bauer, K. Rother, P. Moor, K. Reinert, T. Steinke, J. M. Bujnicki, and R. Preissner. Fast structural alignment of biomolecules using a hash table, n-grams and string descriptors. *Algorithms*, 2(2):692–709, 2009.

[10] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *J. Comput. Biol.*, 6(3-4):281–297, 1999.

[11] A. Ben-Shem, N. Garreau de Loubresse, S. Melnikov, L. Jenner, G. Yusupova, and M. Yusupov. The structure of the eukaryotic ribosome at 3.0 resolution. *Science*, 334(6062):1524–1529, Dec 2011.

[12] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28(1):235–242, Jan 2000.

[13] S. H. Bernhart, I. L. Hofacker, S. Will, A. R. Gruber, and P. F. Stadler. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 9:474, 2008.

[14] B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.

[15] E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, Jun 2007.

[16] D. E. Brodersen, W. M. Clemons, A. P. Carter, B. T. Wimberly, and V. Ramakrishnan. Crystal structure of the 30 S ribosomal subunit from Thermus thermophilus: structure of the proteins and their interactions with 16 S RNA. *J. Mol. Biol.*, 316(3):725–768, Feb 2002.

[17] C. Bron and J. Kerbosch. Algorithm 457: Finding all cliques of an undirected graph. *Commun. ACM*, 16(9):575–577, 1973.

[18] M. Brudno, A. Poliakov, A. Salamov, G. M. Cooper, A. Sidow, E. M. Rubin, V. Solovyev, S. Batzoglou, and I. Dubchak. Automated whole-genome multiple alignment of rat, mouse, and human. *Genome Res.*, 14(4):685–692, 2004.

[19] S. W. Burge, J. Daub, R. Eberhardt, J. Tate, L. Barquist, E. P. Nawrocki, S. R. Eddy, P. P. Gardner, and A. Bateman. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, 41(Database issue):D226–232, Jan 2013.

[20] E. Capriotti and M. A. Marti-Renom. RNA structure alignment by a unit-vector approach. *Bioinformatics*, 24(16):i112–118, 2008.

[21] E. Capriotti and M. A. Marti-Renom. SARA: a server for function annotation of RNA structures. *Nucleic Acids Res.*, 37(Web Server issue):W260–265, 2009.

[22] J. C. Carrington and V. Ambros. Role of microRNAs in plant and animal development. *Science*, 301(5631):336–338, Jul 2003.

[23] J. Cavaille and J. P. Bachellerie. SnoRNA-guided ribose methylation of rRNA: structural features of the guide RNA duplex influencing the extent of the reaction. *Nucleic Acids Res.*, 26(7):1576–1587, Apr 1998.

[24] C. W. Chan, B. Chetnani, and A. Mondragon. Structure and function of the T-loop structural motif in noncoding RNAs. *Wiley Interdiscip Rev RNA*, 4(5):507–522, 2013.

[25] D. K. Y. Chiu and T. Kolodziejczak. Inferring consensus structure from nucleic acid sequences. *Computer applications in the biosciences : CABIOS*, 7(3):347–352, 1991.

[26] J. C. Cochrane, S. V. Lipchock, and S. A. Strobel. Structural investigation of the GlmS ribozyme bound to Its catalytic cofactor. *Chem. Biol.*, 14(1):97–105, Jan 2007.

[27] C. M. Croce. Causes and consequences of microRNA dysregulation in cancer. *Nat. Rev. Genet.*, 10(10):704–714, Oct 2009.

[28] K. E. Deigan, T. W. Li, D. H. Mathews, and K. M. Weeks. Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U.S.A.*, 106(1):97–102, Jan 2009.

[29] Y. Ding, Y. Tang, C. K. Kwok, Y. Zhang, P. C. Bevilacqua, and S. M. Assmann. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*, Nov 2013.

[30] M. Djelloul and A. Denise. Automated motif extraction and classification in RNA tertiary structures. *RNA*, 14(12):2489–2497, Dec 2008.

[31] C. B. Do, D. A. Woods, and S. Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–98, Jul 2006.

[32] E. A. Doherty and J. A. Doudna. Ribozyme structures and mechanisms. *Annu Rev Biophys Biomol Struct*, 30:457–475, 2001.

[33] R. D. Dowell and S. R. Eddy. Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinformatics*, 7:400, 2006.

[34] D. E. Draper. The rna-folding problem. *Accounts of Chemical Research*, 25(4):201–207, 1992.

[35] O. Dror, R. Nussinov, and H. Wolfson. ARTS: alignment of RNA tertiary structures. *Bioinformatics*, 21 Suppl 2:47–53, 2005.

[36] C. M. Duarte, L. M. Wadley, and A. M. Pyle. RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. *Nucleic Acids Res.*, 31(16):4755–4761, Aug 2003.

[37] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK, 1998.

[38] S. R. Eddy. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, 2(12):919–929, Dec 2001.

[39] S. R. Eddy and R. Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Res.*, 22(11):2079–2088, Jun 1994.

[40] I. Eidhammer, I. Jonassen, and W. R. Taylor. Structure comparison and structure patterns. *J. Comput. Biol.*, 7(5):685–716, 2000.

[41] E. Ennifar, A. Nikulin, S. Tishchenko, A. Serganov, N. Nevskaya, M. Garber, B. Ehresmann, C. Ehresmann, S. Nikonov, and P. Dumas. The crystal structure of UUCG tetraloop. *J. Mol. Biol.*, 304(1):35–42, Nov 2000.

[42] A. Esquela-Kerscher and F. J. Slack. Oncomirs - microRNAs with a role in cancer. *Nat. Rev. Cancer*, 6(4):259–269, Apr 2006.

[43] M. A. Faghihi, F. Modarresi, A. M. Khalil, D. E. Wood, B. G. Sahagan, T. E. Morgan, C. E. Finch, G. St Laurent, P. J. Kenny, and C. Wahlestedt. Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat. Med.*, 14(7):723–730, Jul 2008.

[44] J. Felsenstein. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, 5:164–166, 1989.

[45] F. Ferre, Y. Ponty, W. A. Lorenz, and P. Clote. DIAL: a web server for the pairwise alignment of two RNA three-dimensional structures using nucleotide, dihedral angle and base-pairing similarities. *Nucleic Acids Res.*, 35(Web Server issue):W659–668, 2007.

[46] A. R. Ferre-D'Amare, K. Zhou, and J. A. Doudna. A general module for RNA crystallization. *J. Mol. Biol.*, 279(3):621–631, Jun 1998.

[47] J. L. Fiore and D. J. Nesbitt. An RNA folding motif: GNRA tetraloop-receptor interactions. *Q. Rev. Biophys.*, 46(3):223–264, Aug 2013.

[48] W. M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155(3760):279–284, Jan 1967.

[49] E. K. Freyhult, J. P. Bollback, and P. P. Gardner. Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res.*, 17(1):117–125, Jan 2007.

[50] L. Garcia-Ortega, E. Alvarez-Garcia, J. G. Gavilanes, A. Martinez-del Pozo, and S. Joseph. Cleavage of the sarcin-ricin loop of 23S rRNA differentially affects EF-G and EF-Tu binding. *Nucleic Acids Res.*, 38(12):4108–4119, Jul 2010.

[51] P. P. Gardner, J. Daub, J. Tate, B. L. Moore, I. H. Osuch, S. Griffiths-Jones, R. D. Finn, E. P. Nawrocki, D. L. Kolbe, S. R. Eddy, and A. Bateman. Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res.*, 39(Database issue):D141–145, Jan 2011.

[52] P. P. Gardner and R. Giegerich. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, 5:140, Sep 2004.

[53] P. Ge and S. Zhang. Incorporating phylogenetic-based covarying mutations into RNAalifold for RNA consensus structure prediction. *BMC Bioinformatics*, 14:142, 2013.

[54] P. Gendron, S. Lemieux, and F. Major. Quantitative analysis of nucleic acid three-dimensional structures. *J. Mol. Biol.*, 308(5):919–936, 2001.

[55] M. T. Goodrich, J. B. Mitchell, and M. W. Orletsky. Practical methods for approximate geometric pattern matching under rigid motions:(preliminary version). In *Proceedings of the tenth annual symposium on Computational geometry*, pages 103–112. ACM, 1994.

[56] A. R. Gruber, S. Findeiss, S. Washietl, I. L. Hofacker, and P. F. Stadler. RNAZ 2.0: improved noncoding RNA detection. *Pac Symp Biocomput*, 15:69–79, 2010.

[57] S. Gulay. *Building a map of the dynamic ribosome*. PhD thesis, University of Maryland, 2015.

[58] B. Gulko and D. Haussler. Using multiple alignments and phylogenetic trees to detect RNA secondary structure. In Lawrence Hunter and Teri Klein, editors, *Biocomputing: Proceedings of the 1996 Pacific Symposium*, pages 350–367. World Scientific Publishing Co, Singapore, 1996.

[59] R. A. Gupta, N. Shah, K. C. Wang, J. Kim, H. M. Horlings, D. J. Wong, M. C. Tsai, T. Hung, P. Argani, J. L. Rinn, Y. Wang, P. Brzoska, B. Kong, R. Li, R. B. West, M. J. van de Vijver, S. Sukumar, and H. Y. Chang. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, 464(7291):1071–1076, Apr 2010.

[60] R. R. Gutell, A. Power, G. Z. Hertz, E. J. Putz, and G. D. Stormo. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res.*, 20(21):5785–5795, Nov 1992.

[61] R. R. Gutell and C. R. Woese. Higher order structural elements in ribosomal RNAs: pseudo-knots and the use of noncanonical pairs. *Proc. Natl. Acad. Sci. U.S.A.*, 87(2):663–667, Jan 1990.

[62] T. Hainzl, S. Huang, and A. E. Sauer-Eriksson. Structural insights into SRP RNA: an induced fit mechanism for SRP assembly. *RNA*, 11(7):1043–1050, Jul 2005.

[63] M. Hajiaghayi, A. Condon, and H. H. Hoos. Analysis of energy-based algorithms for RNA secondary structure prediction. *BMC Bioinformatics*, 13:22, 2012.

[64] K. J. Hampel and M. M. Tinsley. Evidence for preorganization of the glmS ribozyme ligand binding pocket. *Biochemistry*, 45(25):7861–7871, Jun 2006.

[65] A. O. Harmanci, G. Sharma, and D. H. Mathews. Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. *BMC Bioinformatics*, 8:130, 2007.

[66] A. M. Harrison, D. R. South, P. Willett, and P. J. Artymiuk. Representation, searching and discovery of patterns of bases in complex RNA structures. *J. Comput. Aided Mol. Des.*, 17(8):537–549, Aug 2003.

[67] T. P. Hausner, J. Atmadja, and K. H. Nierhaus. Evidence that the G2661 region of 23S rRNA is located at the ribosomal binding sites of both elongation factors. *Biochimie*, 69(9):911–923, Sep 1987.

[68] J. H. Havgaard, E. Torarinsson, and J. Gorodkin. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Comput. Biol.*, 3(10):1896–1908, Oct 2007.

[69] L. He and G. J. Hannon. MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.*, 5(7):522–531, Jul 2004.

[70] J. Hertel, D. de Jong, M. Marz, D. Rose, H. Tafer, A. Tanzer, B. Schierwater, and P. F. Stadler. Non-coding RNA annotation of the genome of Trichoplax adhaerens. *Nucleic Acids Res.*, 37(5):1602–1615, Apr 2009.

[71] M. Hochsmann, T. Toller, R. Giegerich, and S. Kurtz. Local similarity in RNA secondary structures. *Proc IEEE Comput Soc Bioinform Conf*, 2:159–168, 2003.

[72] I. L. Hofacker. RNA consensus structure prediction with RNAalifold. *Methods Mol. Biol.*, 395:527–544, 2007.

[73] I. L. Hofacker, M. Fekete, and P. F. Stadler. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, 319(5):1059–1066, Jun 2002.

[74] I. L. Hofacker, W. Fontana, P. F. Stadler, S. L. Bonhoeffer, M. Tacker, and P. Schuster. Fast Folding and Comparison of RNA Secondary Structures. *Monatsh. Chem.*, 125:167–188, 1994.

[75] D. Hoksza and D. Svozil. Efficient RNA pairwise structure comparison by SETTER method. *Bioinformatics*, 28(14):1858–1864, 2012.

[76] L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, 233(1):123–138, 1993.

[77] J. P. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, Aug 2001.

[78] J. M. Izquierdo and J. Valcarcel. A simple principle to explain the evolution of pre-mRNA splicing. *Genes Dev.*, 20(13):1679–1684, Jul 2006.

[79] A. Jacquier. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat. Rev. Genet.*, 10(12):833–844, 2009.

[80] J. A. Jaeger, D. H. Turner, and M. Zuker. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. U.S.A.*, 86(20):7706–7710, Oct 1989.

[81] Y. Ji, X. Xu, and G. D. Stormo. A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics*, 20(10):1591–1602, 2004.

[82] J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, 110(1-4):462–467, 2005.

[83] W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 34(5):827–828, 1978.

[84] M. Kertesz, Y. Wan, E. Mazor, J. L. Rinn, R. C. Nutter, H. Y. Chang, and E. Segal. Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 467:103–107, 2010.

[85] S. Kishore, A. Khanna, Z. Zhang, J. Hui, P. J. Balwierz, M. Stefan, C. Beach, R. D. Nicholls, M. Zavolan, and S. Stamm. The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing. *Hum. Mol. Genet.*, 19(7):1153–1164, Apr 2010.

[86] D. J. Klein, T. M. Schmeing, P. B. Moore, and T. A. Steitz. The kink-turn: a new RNA secondary structure motif. *EMBO J.*, 20(15):4214–4221, Aug 2001.

[87] R. J. Klein and S. R. Eddy. RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics*, 4:44, Sep 2003.

[88] B. Knudsen and J. Hein. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15(6):446–454, Jun 1999.

[89] B. Knudsen and J. Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.*, 31(13):3423–3428, Jul 2003.

[90] J. Ko, Y. Lee, I. Park, and B. Cho. Identification of a structural motif of 23S rRNA interacting with 5S rRNA. *FEBS Lett.*, 508(3):300–304, Nov 2001.

[91] J. Laborde, D. Robinson, A. Srivastava, E. Klassen, and J. Zhang. RNA global alignment in the joint sequence-structure space using elastic shape analysis. *Nucleic Acids Res.*, 41(11):e114, 2013.

[92] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948, Nov 2007.

[93] S. Lemieux and F. Major. RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire. *Nucleic Acids Res.*, 30(19):4250–4263, 2002.

[94] L. Lempereur, M. Nicoloso, N. Riehl, C. Ehresmann, B. Ehresmann, and J. P. Bachellerie. Conformation of yeast 18S rRNA. Direct chemical probing of the 5' domain in ribosomal subunits and in deproteinized RNA by reverse transcriptase mapping of dimethyl sulfate-accessible. *Nucleic Acids Res.*, 13(23):8339–8357, Dec 1985.

[95] N. B. Leontis, J. Stombaugh, and E. Westhof. Motif prediction in ribosomal RNAs Lessons and prospects for automated motif prediction in homologous RNA molecules. *Biochimie*, 84(9):961–973, Sep 2002.

[96] N. B. Leontis, J. Stombaugh, and E. Westhof. The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.*, 30(16):3497–3531, Aug 2002.

[97] N. B. Leontis and E. Westhof. Geometric nomenclature and classification of RNA base pairs. *RNA*, 7(4):499–512, 2001.

[98] N. B. Leontis and C. L. Zirbel. Nonredundant 3d structure datasets for rna knowledge extraction and benchmarking. In Neocles Leontis and Eric Westhof, editors, *RNA 3D Structure Analysis and Prediction*, volume 27 of *Nucleic Acids and Molecular Biology*, pages 281–298. Springer Berlin Heidelberg, 2012.

[99] F. Li, Q. Zheng, P. Ryvkin, I. Dragomir, Y. Desai, S. Aiyer, O. Valladares, J. Yang, S. Bambina, L. R. Sabin, J. I. Murray, T. Lamitina, A. Raj, S. Cherry, L. S. Wang, and B. D. Gregory. Global analysis of RNA secondary structure in two metazoans. *Cell Rep*, 1(1):69–82, Jan 2012.

[100] R. A. Lippert, X. Zhao, L. Florea, C. Mobarry, and S. Istrail. Finding anchors for genomic sequence comparison. *J. Comput. Biol.*, 12(6):762–776, 2005.

[101] I. Livyatan, A. Harikumar, M. Nissim-Rafinia, R. Duttagupta, T. R. Gingeras, and E. Meshorer. Non-polyadenylated transcription in embryonic stem cells reveals novel non-coding RNA related to pluripotency and differentiation. *Nucleic Acids Res.*, 41(12):6300–6315, Jul 2013.

[102] R. Lorenz, S. H. Bernhart, C. Honer Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. ViennaRNA Package 2.0. *Algorithms Mol Biol*, 6:26, 2011.

[103] J. B. Lucks, S. A. Mortimer, C. Trapnell, S. Luo, S. Aviran, G. P. Schroth, L. Pachter, J. A. Doudna, and A. P. Arkin. Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl. Acad. Sci. U.S.A.*, 108:11063–11068, 2011.

[104] B. Ma, L. Wang, and K. Zhang. Computing similarity between RNA structures. *Theoretical Computer Science*, 276(12):111 – 132, 2002.

[105] T. Madej, J. F. Gibrat, and S. H. Bryant. Threading a database of protein cores. *Proteins*, 23(3):356–369, 1995.

[106] N. Malod-Dognin and N. Pržulj. GR-Align: fast and flexible alignment of protein 3D structures using graphlet degree similarity. *Bioinformatics*, 30(9):1259–1265, May 2014.

[107] M. Mandal and R. R. Breaker. Gene regulation by riboswitches. *Nat. Rev. Mol. Cell Biol.*, 5(6):451–463, Jun 2004.

[108] S. Matsumura, Y. Ikawa, and T. Inoue. Biochemical characterization of the kink-turn RNA motif. *Nucleic Acids Res.*, 31(19):5544–5551, Oct 2003.

[109] J. S. Mattick. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays*, 25(10):930–939, Oct 2003.

[110] T. R. Mercer, M. E. Dinger, and J. S. Mattick. Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.*, 10(3):155–159, Mar 2009.

[111] E. P. Nawrocki. *Structural RNA Homology Search and Alignment Using Covariance Models*. PhD thesis, Wathington University, 2009. Paper 256.

[112] E. P. Nawrocki and S. R. Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935, Nov 2013.

[113] M. N. Nguyen and C. Verma. Rclick: a web server for comparison of RNA 3D structures. *Bioinformatics*, 31(6):966–968, 2015.

[114] M. S. Nicoloso, R. Spizzo, M. Shimizu, S. Rossi, and G. A. Calin. MicroRNAs–the micro steering wheel of tumour metastases. *Nat. Rev. Cancer*, 9(4):293–302, Apr 2009.

[115] A. Nikulin, I. Eliseikina, S. Tishchenko, N. Nevskaya, N. Davydova, O. Platonova, W. Piendl, M. Selmer, A. Liljas, D. Drygin, R. Zimmermann, M. Garber, and S. Nikonov. Structure of the L1 protuberance in the ribosome. *Nat. Struct. Biol.*, 10(2):104–108, Feb 2003.

[116] H. Nishimasu, F. A. Ran, P. D. Hsu, S. Konermann, S. I. Shehata, N. Dohmae, R. Ishitani, F. Zhang, and O. Nureki. Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell*, 156(5):935–949, Feb 2014.

[117] R. Nussinov, G. Pieczenik, J. Griggs, and D. Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, 35(1):68–82, 1978.

[118] A. R. Ortiz, C. E. Strauss, and O. Olmea. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, 11(11):2606–2621, 2002.

[119] M. Parisien, J. A. Cruz, E. Westhof, and F. Major. New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA*, 15(10):1875–1885, Oct 2009.

[120] E. Pasmant, A. Sabbagh, M. Vidaud, and I. Bieche. ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB J.*, 25(2):444–448, Feb 2011.

[121] J. S. Pedersen, G. Bejerano, A. Siepel, K. Rosenbloom, K. Lindblad-Toh, E. S. Lander, J. Kent, W. Miller, and D. Haussler. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, 2(4):e33, Apr 2006.

[122] A. I. Petrov, C. L. Zirbel, and N. B. Leontis. Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. *RNA*, 19(10):1327–1340, Oct 2013.

[123] Y. Ponty, M. Termier, and A. Denise. Genrgens: Software for generating random genomic sequences and structures. *Bioinformatics*, 22(12):1534–1535, June 2006.

[124] R. R. Rahrig, N. B. Leontis, and C. L. Zirbel. R3D Align: global pairwise alignment of RNA 3D structures using local superpositions. *Bioinformatics*, 26(21):2689–2697, 2010.

[125] T. H. Reijmers, R. Wehrens, and L. M. Buydens. The influence of different structure representations on the clustering of an RNA nucleotides data set. *J Chem Inf Comput Sci*, 41(5):1388–1394, 2001.

[126] J. S. Reuter and D. H. Mathews. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 11:129, 2010.

[127] A. Rich and U. L. RajBhandary. Transfer RNA: molecular structure, sequence, and properties. *Annu. Rev. Biochem.*, 45:805–860, 1976.

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

[129] J. D. Robertus, J. E. Ladner, J. T. Finch, D. Rhodes, R. S. Brown, B. F. Clark, and A. Klug. Structure of yeast phenylalanine tRNA at 3 A resolution. *Nature*, 250(467):546–551, Aug 1974.

[130] M. A. Rosenblad, N. Larsen, T. Samuelsson, and C. Zwieb. Kinship in the SRP RNA family. *RNA Biol*, 6(5):508–516, 2009.

[131] A. Roth, Z. Weinberg, A. G. Chen, P. B. Kim, T. D. Ames, and R. R. Breaker. A widespread self-cleaving ribozyme class is revealed by bioinformatics. *Nat. Chem. Biol.*, 10(1):56–60, Jan 2014.

[132] Y. Sakakibara, M. Brown, R. Hughey, I. S. Mian, K. Sjolander, R. C. Underwood, and D. Haussler. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res.*, 22(23):5112–5120, Nov 1994.

[133] D. Sankoff. Simultaneous solution of the rna folding, alignment and protosequence problems. *SIAM Journal on Applied Mathematics*, 45(5):810–825, 1985.

[134] M. Sarver, C. L. Zirbel, J. Stombaugh, A. Mokdad, and N. B. Leontis. FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J Math Biol*, 56(1-2):215–252, Jan 2008.

[135] S. Schirmer and R. Giegerich. Forest alignment with affine gaps and anchors, applied in RNA structure comparison. *Theoretical Computer Science*, 483(0):51 – 67, 2013.

[136] K. T. Schroeder, S. A. McPhee, J. Ouellet, and D. M. Lilley. A structural database for k-turn motifs in RNA. *RNA*, 16(8):1463–1468, Aug 2010.

[137] S. E. Seemann, J. Gorodkin, and R. Backofen. Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Res.*, 36(20):6355–6362, Nov 2008.

[138] S. E. Seemann, P. Menzel, R. Backofen, and J. Gorodkin. The PETfold and PETcofold web servers for intra- and intermolecular structures of multiple RNA sequences. *Nucleic Acids Res.*, 39(Web Server issue):W107–111, Jul 2011.

[139] A. Serganov, A. Polonskaia, A. T. Phan, R. R. Breaker, and D. J. Patel. Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature*, 441(7097):1167–1171, Jun 2006.

[140] M. R. Sharma, E. C. Koc, P. P. Datta, T. M. Booth, L. L. Spremulli, and R. K. Agrawal. Structure of the mammalian mitochondrial ribosome reveals an expanded functional role for its component proteins. *Cell*, 115(1):97–108, Oct 2003.

[141] J. P. Sheehy, A. R. Davis, and B. M. Znosko. Thermodynamic characterization of naturally occurring RNA tetraloops. *RNA*, 16(2):417–429, Feb 2010.

[142] V. Siegel and P. Walter. Removal of the Alu structural domain from signal recognition particle leaves its protein translocation activity intact. *Nature*, 320(6057):81–84, 1986.

[143] N. Siew, A. Elofsson, L. Rychlewski, and D. Fischer. MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, 16(9):776–785, 2000.

[144] S. Smit, K. Rother, J. Heringa, and R. Knight. From knotted to nested RNA structures: a variety of computational methods for pseudoknot removal. *RNA*, 14(3):410–416, 2008.

[145] R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 28:1409–1438, 1958.

[146] J. Stombaugh, C. L. Zirbel, E. Westhof, and N. B. Leontis. Frequency and isostericity of RNA base pairs. *Nucleic Acids Res.*, 37(7):2294–2312, 2009.

[147] G. Storz. An expanding universe of noncoding RNAs. *Science*, 296(5571):1260–1263, May 2002.

[148] Z. Sukosd, M. S. Swenson, J. Kjems, and C. E. Heitsch. Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions. *Nucleic Acids Res.*, 41(5):2807–2816, Mar 2013.

[149] D. L. Swofford. *PAUP: Phylogenetic Analysis Using Parsimony (and other Methods)*. Sinauer Associates, Sunderland, Massachusettss, 2002.

[150] A. A. Szewczak, P. B. Moore, Y. L. Chang, and I. G. Wool. The conformation of the sarcin/ricin loop from 28S ribosomal RNA. *Proc. Natl. Acad. Sci. U.S.A.*, 90(20):9581–9585, 1993.

[151] M. Tamura, D. K. Hendrix, P. S. Klosterman, N. R. Schimmelman, S. E. Brenner, and S. R. Holbrook. SCOR: Structural Classification of RNA, version 2.0. *Nucleic Acids Res.*, 32(Database issue):D182–184, 2004.

[152] M. Teplova, L. Wohlbold, N. W. Khin, E. Izaurralde, and D. J. Patel. Structure-function studies of nucleocytoplasmic transport of retroviral genomic RNA by mRNA export factor TAP. *Nat. Struct. Mol. Biol.*, 18(9):990–998, Sep 2011.

[153] The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, Jun 2007.

[154] The FANTOM Consortium. The transcriptional landscape of the mammalian genome. *Science*, 309(5740):1559–1563, Sep 2005.

[155] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22(22):4673–4680, 1994.

[156] H. H. Tseng, Z. Weinberg, J. Gore, R. R. Breaker, and W. L. Ruzzo. Finding noncoding RNAs through genome-scale clustering. *J Bioinform Comput Biol*, 7(2):373–388, Apr 2009.

[157] B. J. Tucker and R. R. Breaker. Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.*, 15:342–348, 2005.

[158] B. Turner, S. E. Melcher, T. J. Wilson, D. G. Norman, and D. M. Lilley. Induced fit of RNA on binding the L7Ae protein to the kink-turn motif. *RNA*, 11(8):1192–1200, Aug 2005.

[159] J. G. Underwood, A. V. Uzilov, S. Katzman, C. S. Onodera, J. E. Mainzer, D. H. Mathews, T. M. Lowe, S. R. Salama, and D. Haussler. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods*, 7:995–1001, 2010.

[160] L. M. Wadley and A. M. Pyle. The identification of novel RNA structural motifs using COMPADRES: an automated approach to structural discovery. *Nucleic Acids Res.*, 32(22):6650–6659, 2004.

[161] C. W. Wang, K. T. Chen, and C. L. Lu. iPARTS: an improved tool of pairwise alignment of RNA tertiary structures. *Nucleic Acids Res.*, 38(Web Server issue):W340–347, 2010.

[162] S. Washietl and I. L. Hofacker. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol.*, 342(1):19–30, Sep 2004.

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

[164] M. S. Waterman. *Introduction to Computational Biology: Maps, sequences and genomes*. Chapman and Hall, London, 1995.

[165] O. Weichenrieder, K. Wild, K. Strub, and S. Cusack. Structure and assembly of the Alu domain of the mammalian signal recognition particle. *Nature*, 408(6809):167–173, Nov 2000.

[166] T. Wiegels, S. Bienert, and A. E. Torda. Fast alignment and comparison of RNA structures. *Bioinformatics*, 29(5):588–596, 2013.

[167] K. A. Wilkinson, E. J. Merino, and K. M. Weeks. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat Protoc*, 1(3):1610–1616, 2006.

[168] S. Will, K. Reiche, I. L. Hofacker, P. F. Stadler, and R. Backofen. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, 3(4):e65, Apr 2007.

[169] W. C. Winkler and R. R. Breaker. Regulation of bacterial gene expression by riboswitches. *Annu. Rev. Microbiol.*, 59:487–517, 2005.

[170] W. C. Winkler, A. Nahvi, A. Roth, J. A. Collins, and R. R. Breaker. Control of gene expression by a natural metabolite-responsive ribozyme. *Nature*, 428(6980):281–286, Mar 2004.

[171] C. R Woese and N. R Pace. 4 probing RNA structure, function, and history by comparative analysis. *Cold Spring Harbor Monograph Archive*, 24:91–117, 1993.

[172] C. R. Woese, S. Winker, and R. R. Gutell. Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proc. Natl. Acad. Sci. U.S.A.*, 87(21):8467–8471, Nov 1990.

[173] S. L. Wolin and P. Walter. Signal recognition particle mediates a transient elongation arrest of preprolactin in reticulocyte lysate. *J. Cell Biol.*, 109(6 Pt 1):2617–2622, Dec 1989.

[174] G. K. Wong, D. A. Passey, and J. Yu. Most of the human genome is transcribed. *Genome Res.*, 11(12):1975–1977, Dec 2001.

[175] S. A. Woodson. Compact intermediates in RNA folding. *Annu Rev Biophys*, 39:61–77, 2010.

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

[177] F. Xia, Y. Dou, X. Zhou, X. Yang, J. Xu, and Y. Zhang. Fine-grained parallel RNAalifold algorithm for RNA secondary structure prediction on FPGA. *BMC Bioinformatics*, 10 Suppl 1:S37, 2009.

[178] H. Yang, F. Jossinet, N. Leontis, L. Chen, J. Westbrook, H. Berman, and E. Westhof. Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res.*, 31(13):3450–3460, Jul 2003.

[179] Z. Yao, Z. Weinberg, and W. L. Ruzzo. CMfinder–a covariance model based RNA motif finding algorithm. *Bioinformatics*, 22(4):445–452, Feb 2006.

[180] Y. Ye and A. Godzik. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, 19 Suppl 2:i246–255, 2003.

[181] K. Zarringhalam, M. M. Meyer, I. Dotu, J. H. Chuang, and P. Clote. Integrating chemical footprinting data into RNA secondary structure prediction. *PLoS ONE*, 7(10):e45160, 2012.

[182] J. Zhang and A. R. Ferre-D'Amare. Co-crystal structure of a T-box riboswitch stem I domain in complex with its cognate tRNA. *Nature*, 500(7462):363–366, Aug 2013.

[183] S. Zhang, I. Borovok, Y. Aharonowitz, R. Sharan, and V. Bafna. A sequence-based filtering method for ncRNA identification and its application to searching for riboswitch elements. *Bioinformatics*, 22(14):e557–565, Jul 2006.

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

[185] Q. Zheng, P. Ryvkin, F. Li, I. Dragomir, O. Valladares, J. Yang, K. Cao, L. S. Wang, and B. D. Gregory. Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in Arabidopsis. *PLoS Genet.*, 6(9):e1001141, Sep 2010.

[186] C. Zhong, H. Tang, and S. Zhang. RNAMotifScan: automatic identification of RNA structural motifs using secondary structural alignment. *Nucleic Acids Res.*, 38(18):e176, Oct 2010.

[187] C. Zhong and S. Zhang. Clustering RNA structural motifs in ribosomal RNAs using secondary structural alignment. *Nucleic Acids Res.*, 40(3):1307–1317, Feb 2012.

[188] C. Zhong and S. Zhang. Efficient alignment of RNA secondary structures using sparse dynamic programming. *BMC Bioinformatics*, 14:269, 2013.

[189] C. Zhong and S. Zhang. RNAMotifScanX: a graph alignment approach for RNA structural motif identification. *RNA*, 21(3):333–346, 2015.

[190] M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244(4900):48–52, Apr 1989.

[191] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31(13):3406–3415, 2003.

[192] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9(1):133–148, Jan 1981.