The Dissertation Committee for Lei Shang certifies that this is the approved version of the following dissertation

Improving Secondary Structure Prediction with Covariation Analysis and Structure-based Alignment System of RNA sequences

Committee:

_____

Robin R. Gutell, Supervisor

_____

Pengyu Ren

_____

Rick Russell

_____

Scott Stevens

_____

Lydia Contreras

Improving Secondary Structure Prediction with Covariation Analysis and Structure-based Alignment System of RNA sequences

by

Lei Shang B.S; M.S

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2013

Dedication

I dedicate this dissertation to my wonderful family. Particularly to my precious daughter Shuiyi Shang, who is the joy of my life, my wife who has given me her fullest support, and my parents. Without their support, I could not succeed in this endeavor.

Acknowledgements

I would like to thank all those people whose help make this dissertation possible.

First, I wish to thank my advisor Dr. Robin Gutell for all his guidance, encouragement, support, and patience. His passion for science and education has been a great inspiration to me. Also, I would like to thank my committee members Dr. Pengyu Ren, Dr. Rick Russell, Dr. Scott Stevens, and Dr. Lydia Contreras for their advice and valuable suggestion.

I would also like to thank all the members of the Gutell laboratories for their help in my research: Jamie Cannone, David Gardner and Jung Lee.

Improving Secondary Structure Prediction with Covariation Analysis and Structure-based Alignment System of RNA sequences

Lei Shang, Ph.D.

The University of Texas at Austin, 2013

Supervisor: Robin R. Gutell

RNA molecules form complex higher-order structures which are essential to perform their biological activities. The accurate prediction of an RNA secondary structure and other higher-order structural constraints will significantly enhance the understanding of RNA molecules and help interpret their functions. Covariation analysis is the predominant computational method to accurately predict the base pairs in the secondary structure of RNAs. I developed a novel and powerful covariation method, Phylogenetic Events Count (PEC) method, to determine the positional covariation. The application of the PEC method onto a bacterial 16S rRNA sequence alignment proves that it is more sensitive and accurate than other mutual information based method in the identification of base-pairs and other structural constraints of the RNA structure. The analysis also discoveries a new type of structural constraint – neighbor effect, between sets of nucleotides that are in proximity in the three dimensional RNA structure with weaker but significant covariation with one another. Utilizing these covariation methods, a proposed secondary structure model of an entire HIV-1 genome RNA is evaluated. The results reveal that vast majority of the predicted base pairs in the proposed HIV-1 secondary structure model do not have covariation, thus lack the support from comparative analysis.

Generating the most accurate multiple sequence alignment is fundamental and essential of performing high-quality comparative analysis. The rapid determination of nucleic acid sequences dramatically increases the number of available sequences. Thus developing the accurate and rapid alignment program for these RNA sequences has become a vital and challenging task to decipher the maximum amount of information from the data. A template-based RNA sequence alignment system, CRWAlign-2, is

developed to accurately align new sequences to an existing reference sequence alignment based on primary and secondary structural similarity. A comparison of CRWAlign-2 with eight alternative widely-used alignment programs reveals that CRWAlign-2 outperforms other programs in aligning new sequences with higher accuracy. In addition to aligning sequences accurately, CRWAlign-2 also creates secondary structure models for each sequence to be aligned, which provides very useful information for the comparative analysis of RNA sequences and structures. The CRWAlign-2 program also provides opportunities for multiple areas including the identification of chimeric 16S rRNA sequences generated in microbiome sequencing projects.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Background

1. The Importance of RNA

RNA was once considered the transient and labile molecule whose primary function was to facilitate the translation of DNA sequences into proteins - the Robin to Batman's *more important role* as the stable genetic material in DNA and the enzymatic and functional proteins. The first three RNAs identified - messenger RNA (mRNA), ribosomal RNA (rRNA) and transfer RNA (tRNA) were associated with the protein synthesis. While DNA is known to transfer genetic information from one generation to the next, and proteins are capable of forming three-dimensional structures and perform various functions including catalyzing metabolic reactions, replicating DNA, responding to stimuli, and transporting molecules from one location to another, the function of RNA was perceived predominantly as the carrier of genetic information to code for amino acids in protein, be a scaffold for proteins in the ribosome, and catalyze the formation of bonds between adjacent amino acids and attach the new amino acid to the growing peptide chain during protein synthesis. All three of these RNA functions were considered to be labile and passive.

However this simple perspective of RNA has been undergoing a major transformation. RNA is capable of forming complex three-dimensional structure like proteins. And like proteins these higher-order structures catalyze reactions. Now hundreds, if not thousands, of different RNA families are being identified and characterized. Not only is RNA now implicated in nearly all of the cellular functions in the cell, but the analysis of RNA is revealing many new functions in the cell, including enzymatic activity, regulation of gene expression [1-3], facilitating epigenetics, and association with cancer and other diseases [4,5]. This major paradigm shift in molecular and cellular biology is dramatically changing our appreciation of the machinery, mechanisms, and regulation within cells, and providing a better understanding for the normal and aberrant physiological conditions in biological organisms [6-9].

2. Comparative Analysis

Darwin used comparative methods as the foundation of his theory on the evolution of biological species [10]. The identification of characterization of non-coding RNA molecules and their higher-order structures have utilized an important principle in molecular and evolutionary biology: homologous RNA sequences with different primary structures (or sequences) can form the same higher-order structure to maintain function [11]. The comparative analysis has been widely used in many research fields of RNA. One of the first structures determined with this process was the tRNA secondary structure which was verified by high resolution X-ray crystallography [12-15]. Over 97% of the predicted base pairs in the 16S and 23S ribosomal RNA secondary structure models predicted with comparative analysis were found in the crystal structures[16-18]. Thus comparative analysis has become the fundamentals for the computational analysis of the deluge of nucleic acid sequences that are determined with next-generation sequencing (NGS) methodology.

The sensitivity, accuracy and detail that can be achieved from a RNA comparative analysis is directly proportional to and dependent on: 1) the number and diversity of all sequences within the sequence alignment; 2) the quality/accuracy of the multiple sequence alignment; 3) the types of information used effectively; 4) the performance of the covariation methods that identify the structural constraints; 5) the computational tools that is capable of archiving and analyzing the sequences and structures.

Therefore, one of the core problems to computational comparative analysis is to utilize various types of information about RNAs most effectively. Recently, the Gutell lab developed a novel and sophisticated relational database system – RNA Comparative Analysis Database (rCAD) [19]. It integrates and cross-indexes four primary dimensions of data: (1) metadata, including functional information about sequences and structures; (2) raw sequences and sequence alignments; (3) higher-order structures and (4) evolutionary/phylogenetic relationships between the sequences and structures.

The rCAD system provides opportunities to develop new comparative analysis solutions that utilize multiple dimensional information of RNA. These new algorithms, methods, and programs are dedicated to improve the secondary structure prediction of RNAs and generate large sequence alignment more accurately and rapidly.

Overview of Dissertation

This dissertation has been focused on two major areas of RNA research: the prediction of RNA secondary structure with comparative analysis, and the automated sequence alignment of different RNA families.

Chapter 2 demonstrates that the accuracy and sensitivity of comparative analysis can be improved by utilizing multiple dimensional information of RNA. I developed a novel covariation method, Phylogenetic Events Counting (PEC) method, which used multiple sequence alignment and phylogenetic information to determine positional covariations. A general comparison revealed that the PEC method outperformed other statistics-based methods in the base pair identification of RNA secondary structure. The PEC method also identified a new type of structural constraint – neighbor effect.

Chapter 3 is engaged in the creation of the large multiple sequence alignments that are essential for comparative analysis. With the deluge of nucleic sequences determined with next-gen sequencing technology, it has been essential and challenging to develop computational programs that automatically align these sequences accurately and rapidly. With numerous properly aligned sequences and verified secondary structural information archived in rCAD, my approach utilizes these template sequence alignment and well-established structural information to align new RNA sequences. The automated alignment system I have developed, CRWAlign-2, retrieves template sequence alignment, secondary structure information, and phylogenetic information from rCAD, creates secondary structure models for every new sequence, and aligns the new sequence based on primary and secondary structural similarity. A comparison of CRWAlign-2 with

other existing sequence alignment programs reveals that CRWAlign-2 is more accurate than other alignment methods.

In Chapter 4, I used comparative methods to evaluate a secondary structure model of an entire HIV-1 RNA genome proposed by Weeks group. Every predicted base pair in the HIV-1 secondary structure model are evaluated with different covariation metrics of comparative analysis. The results show the proposed HIV-1 secondary structure model does not have support from comparative analysis. I also determined the positional covariations of HIV-1 genome sequences with mutual information based method, and identified the putative highly conserved base pairs with helix-extension strategy.

Chapter 2: Structural Constraints identified with Phylogenetic Events Counting Analysis in Ribosomal RNA

Abstract

Comparative analysis is able to identify a structure common to a set of sequences in the same RNA family. Covariation analysis, a specific type of comparative analysis is used to identify those positions in an alignment with similar patterns of sequence variation. These two positions usually form a base pair in a helix. While Mutual Information (MI) and its variants have been widely used to accurately predict an RNA secondary structure and a few higher-order structural constraints, early studies revealed that the integration of phylogenetic information improves the accuracy and sensitivity of the covariation analysis for the prediction of base pairs.

With the Gutell lab's new RNA Comparative Analysis Database (rCAD) system, we developed a novel and powerful Phylogenetic Events Counting (PEC) method for identifying and quantifying positional covariations. The application of the PEC method onto a bacterial 16S rRNA sequence alignment proves it is more sensitive and accurate in identifying base-pairs and other constraints in the RNA structure. The comparison between the PEC and MI-based methods reveals that each of these methods identifies unique base pairs, and jointly identifies many other base pairs. In summary, the combination of both methods with an N-best and helix-extension strategy identify the maximal number of base pairs.

While covariation methods have effectively predicted RNAs secondary structure with high accuracy, it only identified a small amount of tertiary structural base pairs. My analysis and the data presented at the Comparative RNA Web (CRW) Site reveal that the majority of these tertiary structural base pairs do not covary with another. However, our analysis discoveries a new type of structural constraint – neighbor effects, which occur between sets of nucleotides that are in proximity in the three dimensional RNA structure with weaker but significant covariation with one another.

Background

The computational prediction of an RNAs higher-order structure from nucleic acid sequences is usually determined by two significantly different methods. The first method attempts to predict the correct higher-order structure from fundamental principles of RNA structure. The primary knowledge used in the majority of these computational algorithms is the free-energy values for simple structural elements, such as two consecutive base pairs [20]. The accuracy of the predicted structure can be high, usually for shorter RNAs (e.g. tRNA – 76 nucleotides), and can be very low for other RNAs (e.g. some of the eukaryotic nuclear and mitochondrial small and large subunit rRNAs) [21,22]. This method is dependent on our understanding of the factors that transform a linear RNA molecule into a secondary and ultimately a three-dimensional structure. Thus, the more available knowledge about RNA structure and the dynamics associated with its folding into a higher-order structure, the more accurate this method is for a larger collection of diverse RNAs.

The second method – covariation analysis is one form of comparative analysis. With the underlying principle that the sequences in the same RNA family fold into the same higher-order structure, the covariation analysis identifies the positions in the RNA molecules that have similar patterns of variation, or covariation, for all or a subset of the sequences within the same RNA family. Covariation analysis was utilized to predict the secondary structure of many noncoding RNAs including tRNA, 5S, 16S, and 23S rRNA [17,18,23], group I introns [24-26], RNase P [27-29], tmRNA [30,31], U RNA [32,33], and SRP RNA [34-36]. For molecules like tRNA or the rRNAs that are known to form a common structure, the accuracy of the predicted RNA structure is or nearly 100% when the number and diversity of sequences within each RNA family is substantial [16]. These examples provide additional support that comparative analysis can identify the secondary structure for some RNAs with extremely high accuracy. The constraints identified with comparative analysis can be utilized to enhance our knowledge about the fundamental rules for RNA structure as well as functional and folding dynamics of these RNA molecules. Although

the underlying concepts of the two methods are significantly different; they each provide knowledge and insight to enhance the other method.

The search for a common structure with comparative analysis, does not in isolation determine an RNA structure. Comparative analysis provides different types of information that can be interpreted to infer: 1) RNA structure, 2) regions of the molecule with functional importance, 3) conserved RNA structural motifs, 4) phylogenetic relationships, 5) other constraints that establish the boundary conditions for the sequences and higher-order structure that have survived the process of evolutionary mutation, and 6) the fitness functions that dictate the options available to maintain the structural and functional integrity of the RNA molecule.

Starting with a multiple sequence alignment consisting of a set of evolutionary-related RNA sequences with sufficient sequence identity, covariation analysis is utilized to predict the early working models of the secondary structure that are subsequently used to refine the alignment in parallel with the addition of more sequences. Additional covariation analysis with more sophisticated algorithms are used to refine the secondary structure in the regions of the rRNA that are present in all sequences spanning the entire phylogenetic tree, regions only present in the three major phylogenetic domains (e.g. Archaea, Bacteria, and Eucarya), and regions only present in sub-branches within these three domains, etc. This iterative process of refinement results in secondary structure models that are very accurate. For 16S and 23S rRNAs, a total of 97-98% of the base pairs predicted with comparative analysis are in the high-resolution crystal structure [16]. The highly accurate secondary structure models substantiate the accuracy of the multiple sequence alignments and the subsequent covariation analysis. Several more detailed description of the RNA sequence alignment have been published [11,23,37].

As we learned from the RNA structure, there are two of the most fundamental principles of RNA structure – 1) the canonical base pair types initially determined by Chargaff [38,39] and Watson and Crick [40], and 2) the arrangement of these base pair types into regular nucleic acid helical structures [40]. While the earliest covariation analysis only searched for canonical base pairs (e.g. G:C, A:U and G:U) occur within a secondary

structural helix [12,41-43], newer mathematical-based and computational rigorous methods identify columns within an multiple sequence alignment with similar patterns of nucleotide variations, regardless of the base pair type and the location of putative base pairs [37,44-46]. The vast majority of all putative base pairs identified with the latest comparative analysis are canonical pairs (G:C, A:U and G:U), and these base pairs are consecutive and antiparallel with one another to form a regular helix. However, these covariation methods have also discovered a large amount of non-canonical structural constraints including pseudo-knots [47,48], base pair exchanges [11,48], base triples [49-51], and sets of positions with a weak network of covariations [46,49]. Therefore, while the vast majority of positions with strong covariations form canonical base pairs within a regular helix, a small portion of pairs with significant covariations are not part of a standard helix and do not exchange solely between canonical base pair types.

The traditional covariation methods identify positional covariation based on the nucleotide frequencies and mutual dependence. This approach has been successfully used in the secondary structure prediction of many RNAs including tRNAs and rRNAs. Recent studies revealed that the phylogenetic relationships between the sequences can enhance the sensitivity for the determination of the number of mutual changes that have occurred during the evolution of the RNA. For example, in determining the first putative helices that forms a pseudo knot, our confidence was significantly reinforced by observing several of the same base pair types (e.g. A:U, G:C) evolved multiple times through the evolutionary history of the 570:866 base pair in 16S rRNA [47] since it has a increasing likelihood that these two positions with similar patterns of variations did not occur by chance. Thus the phylogeny of the sequences is a new dimension of information that can enhance the resolution and alternative interpretations of the covariation analysis. For the early studies incorporating the phylogenetic information [47,48], the number of coordinated changes during the evolution of the RNA was counted from a visual inspection of the data. With the deluge of numerous nucleic acid sequences determined in modern days, developing novel automatic computational methods for the identification of covariations based on phylogenetic relationships has become an essential and challenging

task. Several research groups have presented new covariation methods based on modeling phylogenetic relationship [52-54].

The Gutell lab's RNA Comparative Analysis Database (rCAD) system [19] integrates and cross-indexes multiple dimensions of information for storage, retrieval, and analysis. While this infrastructure has many applications for the analysis of RNA structure, function and evolution, I developed a new Phylogenetic Events Counting (PEC) method that utilized rCAD to determine the coordinated changes at each pair of nucleotide positions in the RNA molecule during its evolution. The PEC method traversals the entire phylogenetic tree hierarchy from leaf node to root, and measures the significance of positional covariation. To augment the PEC method, a Joint N-Best method and a helix-extension procedure are utilized to enhance the identification and accuracy of identification of the structural constraints present in the sequence alignment. A comparison between the PEC based method and other Mutual Information (MI) based covariation methods reveals that while PEC outperform other covariation methods in the identification of base pairs, MI based methods also identify unique base pairs, and they jointly identify many other base pairs. The combination of both types of methods when applied simultaneously identifies more base pairs than either method by itself. And last, the process of applying these covariation methods also identifies other types of structural constraints – neighbor effect in an RNA molecule.

Methods

1. Phylogenetic Events Counting (PEC) Algorithm

Given a high quality multiple sequence alignment (MSA) consisting of a set of properly aligned sequences, and phylogenetic relationships between all of the sequences within the MSA, the PEC method gages the evolution of the RNA molecule to determine positions having similar patterns of variations. The phylogenetic information is obtained from taxonomy page at NCBI (ftp://ftp.ncbi.nih.gov/pub/taxonomy/).

The PEC algorithm maps the nucleotides of each pair of positions onto the phylogenetic tree according to the taxonomy information, and performs a tree-traversal from leaf nodes to root which counts all types of nucleotide changes. Since the NCBI taxonomy tree is not a binary tree (each node may have more than two child nodes), a standard variation of Fitch's maximum parsimony approach adapted for non-binary tree is used to determine the nucleotides of ancestor nodes (equality set). The equality set of each node is determined as the type of pair that occurs most frequently in all sequences within that node and its child nodes. The types of pairs that are different from the equality set will be counted as positive event (nucleotide changes at both positions) or negative event (nucleotide change at only one position) according to the definition. To avoid over-sampling of certain branches, we only consider the minimum number of variations - each type of pair will only be counted once regardless of its number of occurrence. The Pseudo code of PEC algorithm is shown in Figure 2.1.

Phylogenetic Events Counting

1. For each pair of positions, start with root of phylogenetic tree
2. For each node
3. Compute the *equality set* for the node as the set of pair types having the maximum count among the child sequences at this node
4. If the node is an internal node,
5. for each child node
6. do co-evolutionary events counting for each child node
7. add the positive and negative events count tag from child to current node respectively.
8. Sort each pair type into *pair_list* based on the number of occurrence among sequences directly within this node if it is at least one.
9. Add any pair type from child equality sets that is not included in the *pair_list* at the end of *pair_list*
10. Move the head of *pair_list* to the parent candidate set
11. For each pair type *p* in *pair_list*
12. if p is covary with parent set, move p into parent candidate set, and increase current positive_event_count by 1
13. else increase current negative_event_counts by 1

Definition of positive events: Given a pair of positions in a set of aligned sequences under the same ancestral node on the phylogenetic tree, a positive event is counted when both positions changed from the ancestral sequence. The type of pair having positive events will be accounted as one of ancestral pair collection.

Definition of negative events: Given a pair of positions in a set of aligned sequences under the same ancestral node on the phylogenetic tree, a negative event is counted when only one position changed from the ancestral sequence.

Figure 2.1: Pseudo code of Phylogenetic Event Counting (PEC) algorithm

After the complete tree-traversal, the Covariation Percentage of Events (CPE) is calculated as the sum of positive events divided by the sum of total events (both positive and negative). High CPE score indicates strong covariation between the two positions.

2. RNA Comparative Analysis Database (rCAD) System

The PEC method is implemented on the Gutell lab's RNA Comparative Analysis Database (rCAD) system. This system is built with a novel schema to stores and cross-indexes four primary dimensions of data: (1) metadata, including functional information about sequences and structures; (2) raw sequences and sequence alignment; (3) higher-order structure and (4) evolutionary/phylogenetic relationships between the sequences and structures. The system supports SQL queries accessing individual rows, columns and cells in multiple sequence alignments as well as RNA structures and taxonomy information. It provides the fundamentals for novel analysis of the sequence, structure, and function characterizations of RNAs, such as covariation analysis and structural statistics [55].

3. Other Covariation Methods

Standard Mutual Information (MIxy) measures the coordinated or compensatory variations between two positions. It has been utilized to successfully in several previous studies of RNA structures [45,46,53,56]. The MIxy value between column x and y in the alignment is calculated as

$$\boldsymbol{MI(x,y)} = \sum_{\boldsymbol{M,N\epsilon\{A,C,G,U\}}} \boldsymbol{Pr(M_x,N_y)} * \boldsymbol{ln}\frac{\boldsymbol{Pr(M_x,N_y)}}{\boldsymbol{Pr(M_x)} * \boldsymbol{Pr(N_y)}} \qquad (\mathbf{2.1})$$

where $Pr(M_x,N_y)$ is the joint probability of nucleotide M and N in column x and y , and $Pr(M_x)$ and $Pr(N_y)$ is the marginal probability for a nucleotide (M or N) in column x and y.

Dunn et al. developed a modified mutual information based method to estimates the background for each pair of positions in a given sequence alignment of RNA/protein [57]. Removal of this background generates a corrected mutual information metric, MIp, improves the base-pair identification. Here we repeated the calculation process of MIp as described in their paper.

Other covariation methods involved include OMES [58], McBASC [59] and ELSC [60]. OMES measures the difference between the expected and observed di-nucleotides frequency for a pair of positions (columns). It is calculated as

$$\boldsymbol{OMES} = \frac{-\sum_{i}^{N}(\boldsymbol{N_0} - \boldsymbol{N_e})^2}{\boldsymbol{N_t}} \qquad (\mathbf{2.2})$$

where N_0 is the observed number of di-nucleotides in a pair of positions, N_e is the expected number, N is the total number of possible di-nucleotide pairs, and N_t is the total number of sequences in the alignment. The calculation of McBASC and ELSC is implemented using the code provided by the authors (http://www.afodor.net/).

We also tried several other covariation methods including PSICov [61], Direct information (DI) [62], RNAalifold [63], RNAfold [64,65], Pfold [66,67] and Evofold [68]. However, due to the limitations on the molecule type and the size of input sequence alignments, none of these methods are applicable in this study, and therefore not included in this analysis.

4. N-Best Strategy

In 1992, a simple descending ranking of MIxy value for tRNA revealed that the top 19 pairings are real base pairs in the tRNA secondary structure while the 20th pairing was a tertiary base pair [46]. However, many pairs of positions that are not base-pairing in the tRNA higher-order structure have higher MIxy values than several of the base pairs present in the tRNA secondary structure model. It has been determined that the mutual information value is associated with Shannon's information entropy [69]. The Mixy score between two positions is the difference between the sums of the entropies for these two positions minus the joint entropy [http://sciencehouse.wordpress.com/2009/08/08/-information-theory/]. According to Shannon's entropy equation, highly conserved positions have the minimum entropy values, while highly variable positions have the maximum entropy values. Thus, the MIxy score of two positions with the identical patterns of variation (i.e. covariation) is greater when the entropy value is smaller (i.e. greater variation).

To correct this potential bias, a simple, although not the most mathematically eloquent solution is to determine the positions with the highest mutual information scores, or covariation for each individual position. This method, named N-Best was utilized to enhance the interpretation of base pairs from the MIxy scores [46]. The N-best score is measured as the ratio of the second highest covariation scores divided by the highest covariation score in the series of pairs (X_1:Y_1, X_1:Y_2, ..., X_1:Y_n).. The pairs with N-best score satisfying the threshold will also be considered as candidate base-pairs having significant covariations.

A variation of N-best method – Joint N-Best is used to determine the pairs of positions with the most significant covariation. For each pair (X_1:Y_1), the N-Best scores of position X_1 and Y1 are calculated separated. The pairs with both N-Best scores lower than the predefined threshold ($\leq$0.5) will be considered as candidate base-pairs having significant covariations.

5. Helix-extension strategy

The long term goal of comparative analysis is to identify every base pair in the RNAs higher-order structure with covariation analysis. An assessment of the conservation diagrams of the three primary forms of life – Bacteria, Archaea, and Eukaryotes [http://www.rna.ccbb.utexas.edu-/SAE/2B/ConsStruc/]] reveals a significant amount of sequence conservation within each major phylogenetic domain. Thus many positions in the bacterial 16S rRNA sequences that are base paired in the comparative structure model have no variation and thus no covariation. However, nearly every pair of positions that is base paired in the comparative structure model has covariation in alignments that include sequences from organisms spanning all or part of the phylogenetic tree of life [http://www.rna.ccbb.utexas.edu/SAE/2A/nt_Frequency/BP/-16S_Model]. When structural elements are conserved within the RNA family under study, and covariation analysis cannot identify the base pair or structural element, then we search for RNA structure elements that have been well characterized, such as the adjacent and antiparallel base pairs that form a secondary structure helix.

For example, base pairs in a helix can be identified when G:C, A:U, and G:U pairings are antiparallel and immediately adjacent to a putative base pair identified with the covariation analysis. With this helix-extension procedure, a collection of predicted covariant pairs are used as the "nucleation pairs". The corresponding columns of these nucleation points within the MSA are determined. When the Watson-Crick (G:C or A:U) or Wobble pair (G:U) percentage (WCWB%) of the neighboring columns is higher than a predefined threshold (85%), the neighboring columns in an alignment are considered to be base-paired.

6. Calculation of Conservation Score and Purity Score

Given a sequence alignment, the conservation score of column *i* (*Ci*) is calculated with

$$\boldsymbol{Ci} = \sum \boldsymbol{Pm} * \mathbf{log_2}(\mathbf{4} * \boldsymbol{Pm}) + \boldsymbol{P_\Delta} * \mathbf{log_2}\, \boldsymbol{P_\Delta} \qquad (\mathbf{2.3})$$

where *Pm* is the frequency of occurrence of nucleotide *m* at column *i* and P_Δ is the frequency of deletions (gaps) at column *i* [23].

The purity score measures the extent that one nucleotide (A, C, G, U) at column *i* is only associated with one other nucleotide at column *j*. For example, for a pair of columns in the alignment, the set of paired nucleotides {A:U; G:C; U:A; C:G} have the highest purity score – 100% since each nucleotide at one column is uniquely associated with one other nucleotide at the other column. The set {A:U; G:C; G:U; and C:G} would have a lower purity score since G is associated with C and U, and the set {C:A, C:C, C:G, C:U} would have the lowest purity score since nucleotide C at one column is associated with four different nucleotides at the other column. Higher purity score indicates that the two columns are more likely to have strong covariation. Figure 2.2 describes the procedure that defines the list of base pair types that have a covariation with one another, and the purity score calculated is the sum of the percentages of these base pair types.

Figure 2.2: Base pairs in the Bacterial 16S rRNA structure model that are identified with the helix extension method using different nucleation pairs. Red: true positive base-pairs identified in Joint N-Best method, which are used as nucleation points in the helix extension Magenta: false positives in the nucleation pairs; Blue: true positive base-pairs identified with the helix-extension method; Yellow: false-positive pairs identified with the helix-extension method. Secondary base-pairs are represented by closed circles while tertiary base-pairs are represented by open circle and highlighted with arrows. (A) Using pairs identified in PEC/JN-Best as the nucleation pairs. (B) Using pairs identified in MI/JN-Best as the nucleation pairs.

When the top two base pairs with the highest percentage are A:U and G:C, then a G:U pair is not a covariation type according to the above definition. However, G:U base pairs, also called the wobble base pair [70] occur within a regular helix. To accommodate this change, a GU-Plus purity score is calculated with a slightly modified procedure: the

base pairs G:U (or U:G) are counted as covariation with G:C (or C:G) and A:U (or U:A) for all of the known base pairs.

7. Identification of Neighbor Effects and Physical Distance Calculation

Most pairs with strong covariations are identified as base pair in the RNA higher-order structure. However, there are numerous pairs with significant covariations that have not been identified as a potential base pair. Those pairing are not necessarily indicative of a base pair, instead they comprise a structural constraint on the evolution of a set of nucleotides. The objective here is to quantitate the process of identifying these pairs, named "neighbor effects".

The neighbor effects are identified with the standard one-directional N-Best method with some constraints. Given a pair X_1:Y_1, the N-Best score of column X_1 is calculated as the ratio of the second highest CPE score divided by the highest CPE score in the series of pairs (X_1:Y_1, X_1:Y_2, ..., X_1:Y_n). When the covariation score (CPE in this case) is low (for example pair with CPE < 15%), the background noise could interfere with the covariation signal and lower the quality of the analysis. To remove this background noise, only those pairs have a minimum number of total changes during the evolution (total events) and have a CPE score higher than a predefined lowest cutoff value (25%) are included in this analysis. The pairs with: 1) N-Best scores exceeding the predefined threshold (0.85); 2) Covariation score (CPE) higher than a predefined lowest cutoff (25%); 3) Total events (positive plus negative) higher than a minimum event threshold, are considered as neighbor effects.

The two primary types of interactions between bases are hydrogen bonding and base stacking. While the latter contributes more to the stability of the RNA structure [71-73], the specificity of the interactions is dictated by the hydrogen bonding of the two nucleotides that form a base pair. For all identified neighbor effects, the physical distance at atomic level are estimated using the 3D high-resolution crystal structures (PDBID

1J5E for 16S rRNA; PDBID 2AW4 for 5S and 23S rRNA). The physical distance between two nucleotides (N_1 and N_2) is calculated with

$$\boldsymbol{Dist} = \sqrt{(\overline{\boldsymbol{x}}_{\boldsymbol{1}} - \overline{\boldsymbol{x}}_{\boldsymbol{2}})^{\boldsymbol{2}} + \left(\overline{\boldsymbol{y}}_{\boldsymbol{1}} - \overline{\boldsymbol{y}}_{\boldsymbol{2}}\right)^{\boldsymbol{2}} + (\overline{\boldsymbol{z}}_{\boldsymbol{1}} - \overline{\boldsymbol{z}}_{\boldsymbol{2}})^{\boldsymbol{2}}} \qquad (\mathbf{2.4})$$

where $\overline{\boldsymbol{x}}_{\boldsymbol{1}}$, $\overline{\boldsymbol{y}}_{\boldsymbol{1}}$, $\overline{\boldsymbol{z}}_{\boldsymbol{1}}$ are the coordinates for the center of atoms in N_1 that usually form the hydrogen bonds in two nucleotides that are base paired, and $\overline{\boldsymbol{x}}_{\boldsymbol{2}}$, $\overline{\boldsymbol{y}}_{\boldsymbol{2}}$, $\overline{\boldsymbol{z}}_{\boldsymbol{2}}$ are the coordinates for the center in N_2.

8. Dataset and Filtration Algorithm

Three data sets are used in this analysis: a bacterial 16S rRNA sequence alignment containing 4142 sequences with 3236 Columns; a bacteria 5S rRNA alignment containing 2088 sequences with 333 columns; and a bacteria 23S rRNA alignment containing 2339 sequences with 7330 columns. The sequences in this analysis include organisms from most of the major branches of the bacterial phylogenetic tree.

Considering a MSA consisting of *m* columns and *n* rows, the total amount of column pairwise comparison is m*(m-1)/2, and the time complexity of PEC algorithm is in the order of $O(m^2n)$. Given the finding that positions with similar conservation values have the potential to have a higher covariation score (Figure 2.3), the number of column pairwise comparison can be reduced significantly by only analyzing those sets of positions with similar conservation values. A coarse filter based on relative entropy and the MIxy is implemented to eliminate the unnecessary comparisons between two columns that unlikely to have a significant covariation score. The PEC analysis is only performed on those pairwise sets of columns with: 1) the relative entropy score lower than a predefined threshold (0.2), and 2) MIxy value of column X and column Y are among the top 100 for both column X (with any other column) and column Y (with any other column) [74]. This filtration step significantly reduces the computational cost by over 300 times. For example, in the 16S rRNA MSA, the course filter reduces the total of ~5,234,230 pairwise comparisons to 14,276 pairings. This smaller number of pairings is

analyzed in the subsequent PEC analysis. Among the 608 secondary and tertiary base pairs present in the *T. thermophilus* 16S rRNA high resolution crystal structure (PDB ID 1J5E), 218 are eliminated in the filtration step. None of these eliminated base pairs have significant covariations except one of them can be identified with PEC method. Therefore the coarse filter effectively reduces the computational cost with a minor decrease in sensitivity. The same filtration procedures are applied in analyzing the 5S rRNA and 23S rRNA MSA.

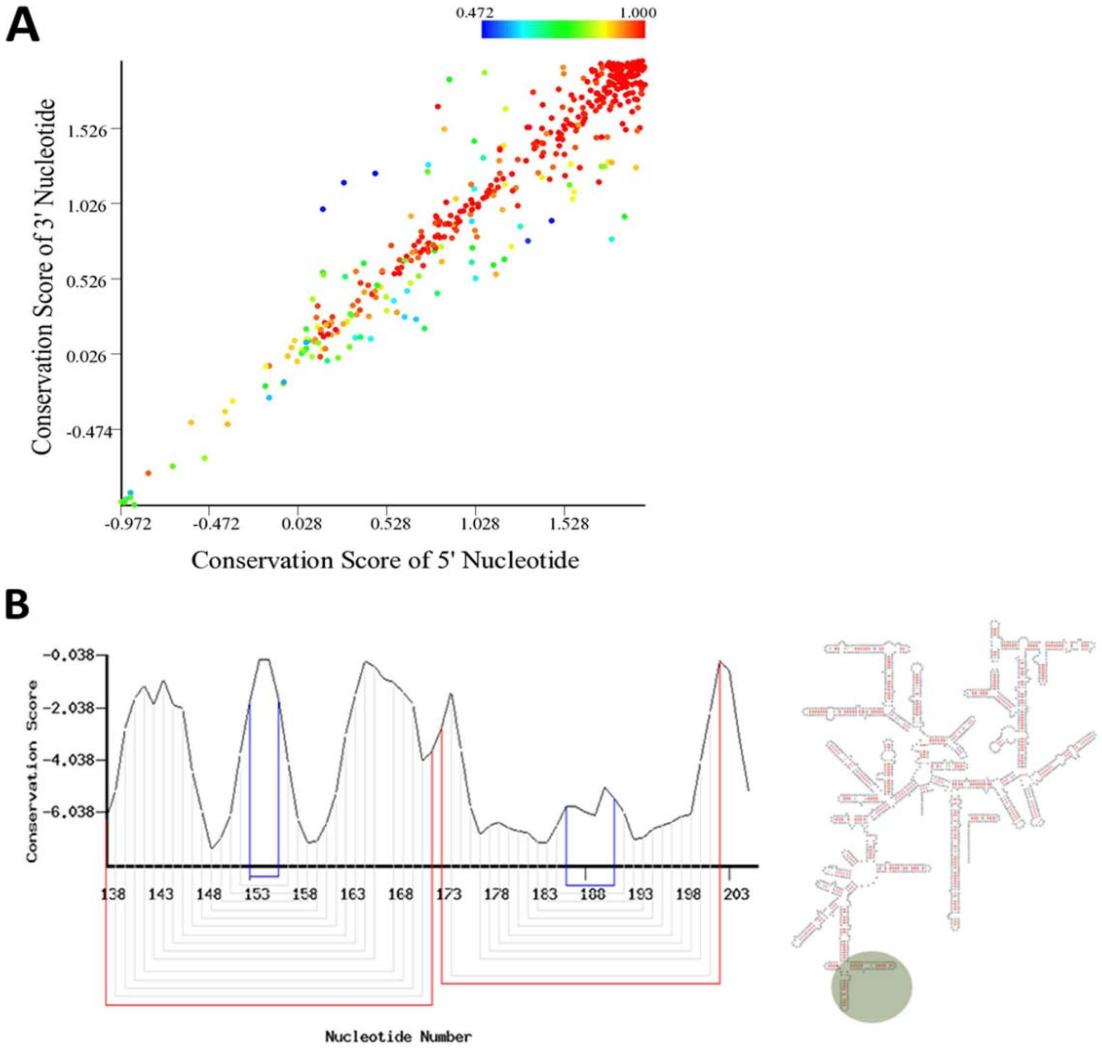


Figure 2.3: The underlying principle of coarse filter that reduce the number of pairwise comparison. (A) The conservation scores for all nucleotides that are base paired in the 16S rRNA comparative structure model. Each base pair is represented with a colored circle, where the color indicates the purity score (minimal value: 0.472; maximum value: 1). The vast majority of the dots representing base pairs are close to the diagonal. (B) The conservation scores for each nucleotide position from 138 to 205 which is under the shadow on the entire *Escherichia coli* 16S rRNA secondary structure (right). The red and blue lines indicate the outer and inner boundaries of the helices respectively while grey lines connect the positions that form a base pair.

Results

1. Conceptual Overview of the Methods

1.1. Phylogenetic Events Counting Method

Figure 2.4 shows the overall analysis workflow of the Phylogenetic Event Counting method (PEC). The program retrieves four primary dimensions of data including 1) raw sequences (unaligned) and sequence alignment; 2) higher-order structural information; 3) sequence and structure metadata; 4) evolutionary/phylogenetic relationships between sequences and structures that are stored and analyzed in rCAD (Figure 2.4A & 2.4B). The di-nucleotides of each pair of positions to be processed are mapped onto the phylogenetic tree (Figure 2.4C). A tree-traversal from leaf nodes to root counts two types of phylogenetic events: positive event and negative event (Figure 2.4D).

Definition 1 Positive event: Given a pair of positions on a sequence, a positive event is observed when both positions are changed from its direct ancestral sequence.

Definition 2 Negative event: Given a pair of positions on a sequence, a negative event is observed when only one position is changed from its direct ancestral sequence.

In practice, there are usually no actual ancestral sequences at every internal node of a phylogenetic tree. Therefore, given a set of sequences under a node, we define an equality set to represent the nucleotides of ancestor nodes using maximum parsimony strategy.

Definition 3 Equality set: Given a set of sequences under a node of phylogenetic tree, an equality set is defined as the type of pair that occurs most frequently in all sequences within that node and its child nodes.

To avoid bias caused by over sampling under certain nodes of the phylogenetic tree, each type of pair of child nucleotides is counted only once. For example, ancestor node is G:C, child nodes contain G:C which occurs 10 times, A:U which occurs 2 times, A:C which occurs 1 time. The A:U pair will be counted only once as positive event regardless of its actual occurrence. Thus the observed events are minimized to assure

high confidence in this research. The covariation between two positions is determined by calculating the Covariation Percentage of Events (CPE), which is the ratio of positive events to the total number of events (both positive and negative) (Details in Method section).

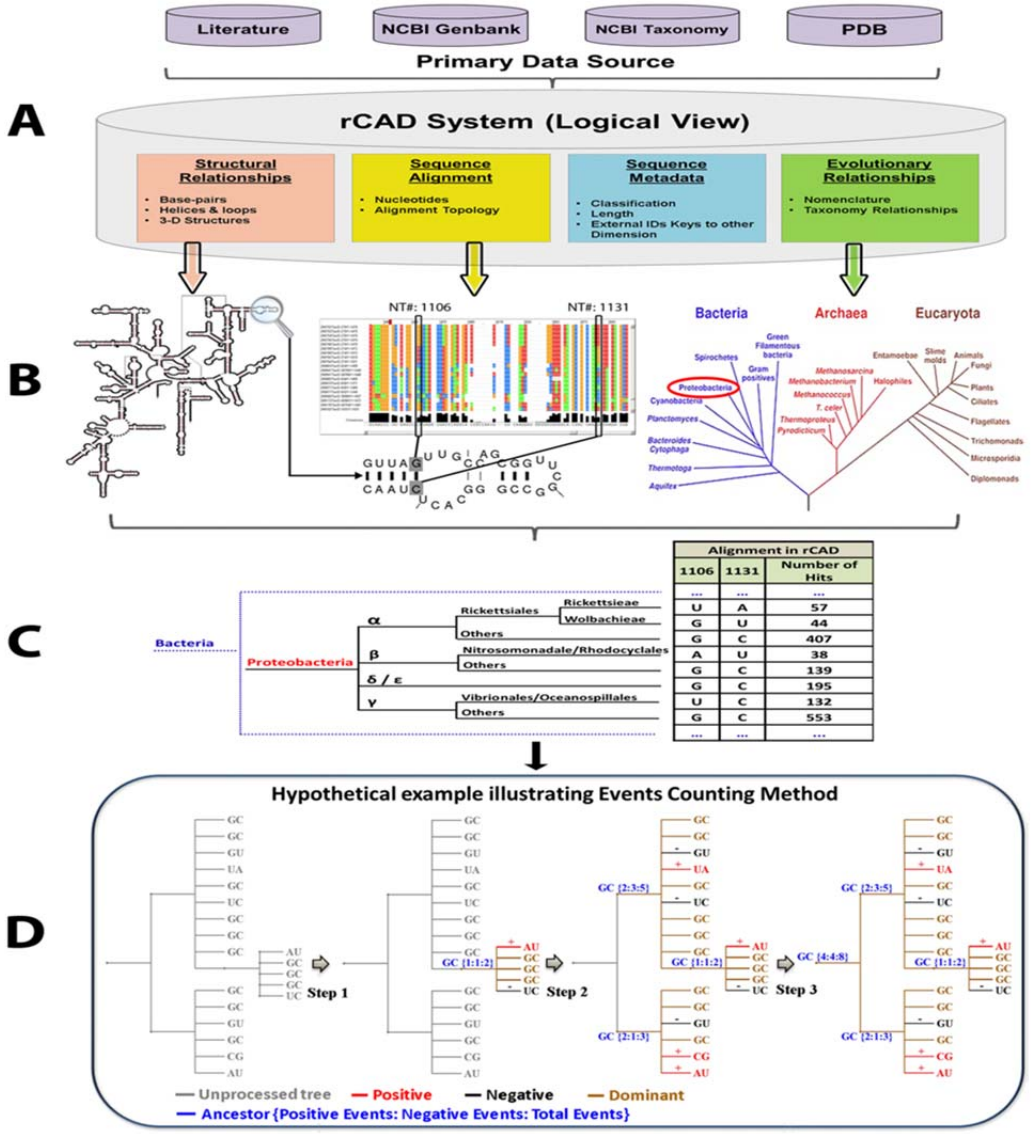


Figure 2.4: The highlight and underlying concepts of the PEC based covariation analysis: Data source (A); multi-dimensional data (B); mapping the substitutions (C); counting the positive and negative events (D).

1.2. Base Pair Identification Process

Figure 2.5 shows the analysis procedure that reveals higher-order structural constraints of RNA molecules. Joint N-Best strategy is used to measure the significance of covariation score (i.e. CPE, MIxy, MIp) between two positions. Pairs of positions satisfying predefined thresholds are identified as putative base pairs, and used as the nucleation points in the following helix-extension procedure to further improve the sensitivity (Process colored blue in Figure 2.5) (Details in Methods section).

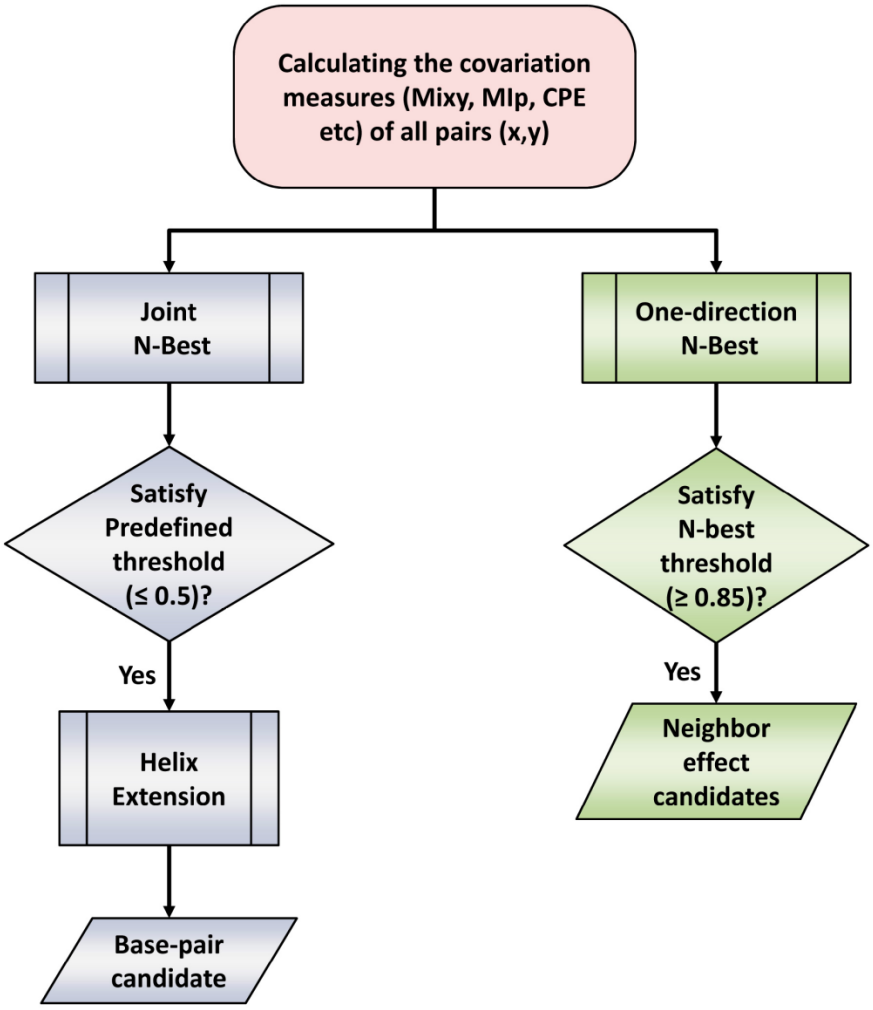


Figure 2.5: The flowchart of analysis in the identification of base-pairs and neighbor effects.

The N-Best strategy was initially used with mutual information (MIxy) on a set of tRNA sequences [46]. Since the Mixy values increase for similar extents of covariation as the entropy value decreases (ie. increases in variation), the Mixy values should be standardized for the different entropy values. To approximate this, a simple solution is to rank the positions with the highest covariation scores for each individual position. The previous study [46] revealed that for the most majority of base pairs in the comparative structures of the tRNAs, the positions forming a base pair with cardinal position number usually had a MIxy value significantly higher than the Mixy values for the other ranked positions.

This N-Best strategy standardizes the covariation scores by first ranking the positions in descending order with their covariation scores (i.e. MIxy, CPE), followed by calculating the ratio of the second highest covariation score to the highest score. For position X and position Y, the likelihood that they form a base pair is further enhanced when the position with the highest score with X is Y, and the position with the highest score for Y is X. Thus this Joint N-Best strategy is applied to the covariation scores with a predefined N-Best threshold. While our confidence in the prediction of a base pair is proportional to the difference between the two positions with the highest covariation values, here we set a predefined N-Best threshold as 0.5. The pairs of positions satisfying this threshold are considered as base pair candidates with significant covariations. The implementation of Joint N-Best with PEC method (PEC/JN-Best) improves the sensitivity and accuracy for the identification of base pairs.

The three-dimensional high-resolution crystal structure of *T. thermophilus* 30S ribosomal subunit (PDBID 1J5E) which contains the 16S rRNA, and *E.coli* 50S ribosomal subunit (PDBID 2AW4) which contains the 5S rRNA and 23S rRNA are used as the reference structures for this study. All identified putative base pairs are categorized as true positives (annotated in the reference structures) or false positives (not annotated in the reference structures).

Given a sequence alignment, the amount of covariation is directly proportional to the amount of variation. For the bacterial 16S rRNA alignment used in this study, Figure

2.6 shows the relationships between the overall variation and the amount of variation in three categories in the secondary structure: 1) both positions forming base pairs undergo changes; 2) one of the two base paired positions changes, and 3) the unpaired positions.

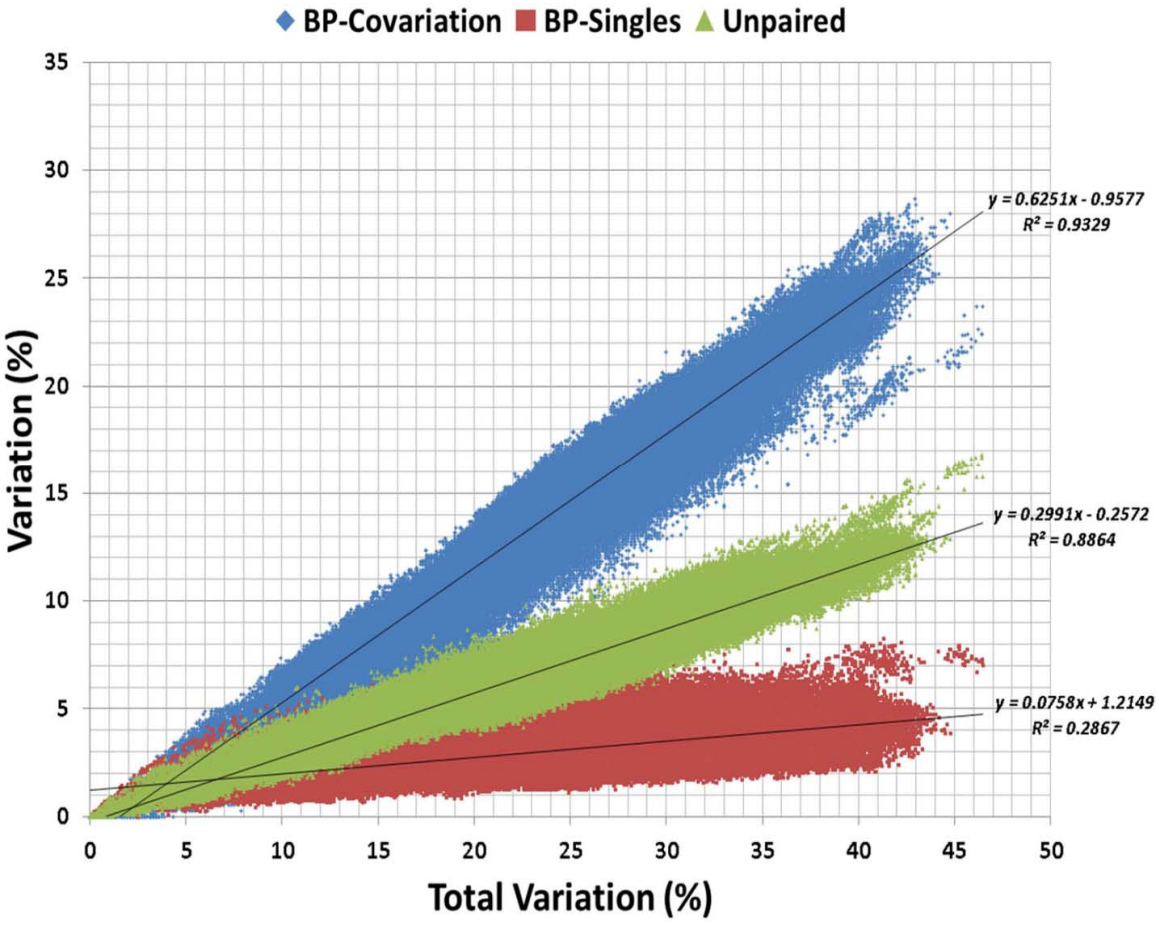


Figure 2.6: Variation/covariation analysis of the secondary structure of the bacterial 16S rRNA sequence alignment. Total variation in each pairwise set of sequences (X-direction) is plotted vs. (1) the amount of variation in that set of sequences for the two positions that are base paired in the secondary structure (blue), (2) only one position of the two that are base paired in the secondary structure (red), and (3) variation in the unpaired region of the second structure (green) (Y-direction). The slope, Y-intercept, and R^2 co-efficiency values of the linear regression line for each of the three analyses are at the right side of the line.

This variation/covariation analysis reveals that highly conserved positions are less likely to be identified as a base pair with covariation methods (i.e. no variation, no covariation). Since the ultimate objective is to identify every base pair in the secondary and higher-order structure, a helix-extension method was developed to identify those highly conserved base pairs and improve the sensitivity of this analysis. The putative base pairs identified with Joint N-Best method are used as the nucleation pairs in the helix-extension process. The helix-extension algorithm seeks to increase the length of a putative helix composed of canonical base pairs (G:C, A:U, and G:U) that are 1) adjacent and antiparallel with the nucleation pair and 2) occur in at least 85% of the sequences. A primitive and less quantitative version of helix extension was first applied in building the original 16S and 23S rNRA secondary structure models [41,75]. As more 16S and 23S rRNA sequences were determined, the putative extended base pairs were verified with covariation criteria: some of the extended base pairs were removed when the two positions did not have similar patterns of variation, while the most majority of the extended base pairs did have similar patterns of variation in alignments that contained more sequences [23,37]. Since our confidence in a predicted base pair is directly proportional to the amount of covariation, we have less confidence in those extended base pairs that have minimal or no covariation.

1.3. Neighbor Effects Identification Process

Previous analysis has shown that as the extent of positional covariation decrease, some pairs with lower covariation scores form base pairs, and others do not. As shown in Figure 2.7, for the majority of all positions that are base paired, the highest covariation score is significantly higher than the position with the second highest score (example of nucleotides 3 in tRNA are presented in Figure 2.7A left side, while the overall picture are shown in Figure 2.7B). However, the highest covariation score for some base pairs is lower, while the set of next highest positions are closer to the highest (see Figure 2.7A right side and Figure 2.8).

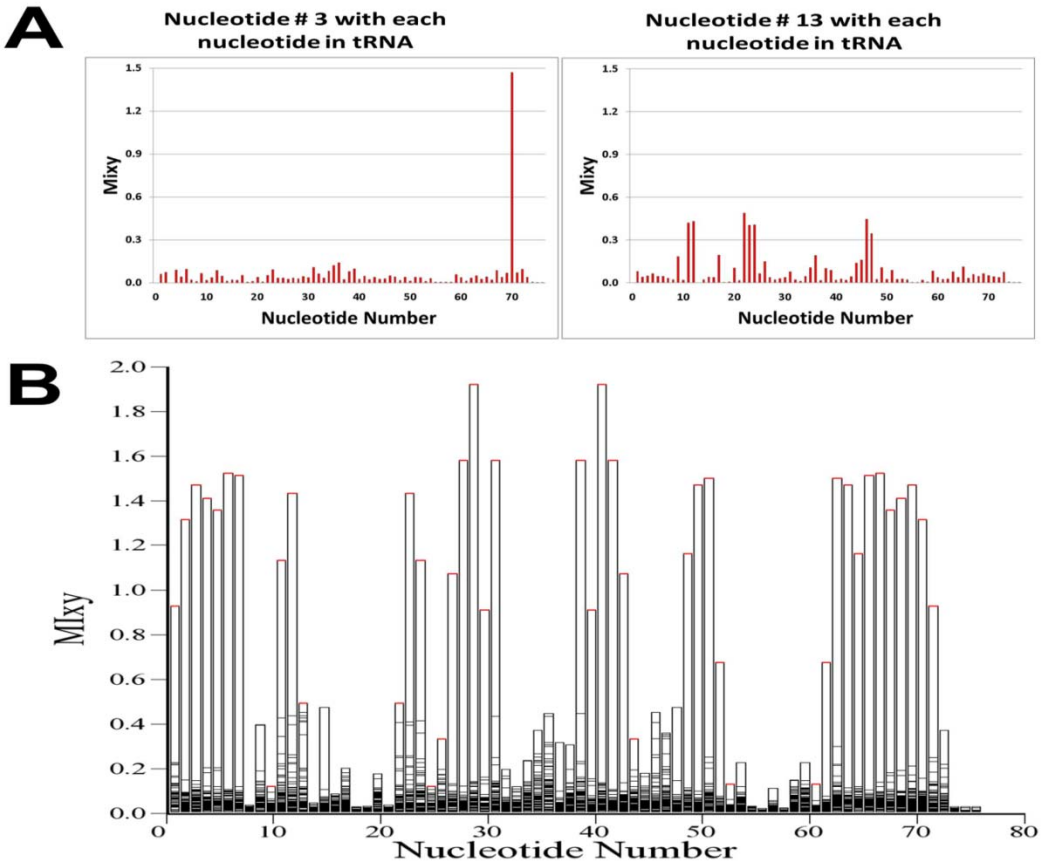


Figure 2.7: Graphical representation of N-Best method. While the mutual-information (MIxy) covariation method compares all positions against all other positions, the N-best method ranks covariation scores for two positions for each individual position. The position numbers are in the X-axis and the MIxy values are in the Y-axis. (A) Left: The MIxy scores for position 3 with all 76 positions in tRNA; Right: The MIxy values for position 13 with all 76 positions are also displayed in the right side with the same manner. (B) Each nucleotide position in a tRNA is shown in the X-axis while the MIxy score are displayed in the Y-axis. The vertical bar is the MIxy value for position Z and each of the individual positions in the X-axis. When the positions with the best covariation scores for each position are base paired in the tRNA structure, that vertical bar is shown in red. The positions with lower MIxy values are shown as black vertical lines. This diagram illustrates that the majority of all positions that are base paired has a MIxy value significantly higher than the MIxy value for all of the other positions.

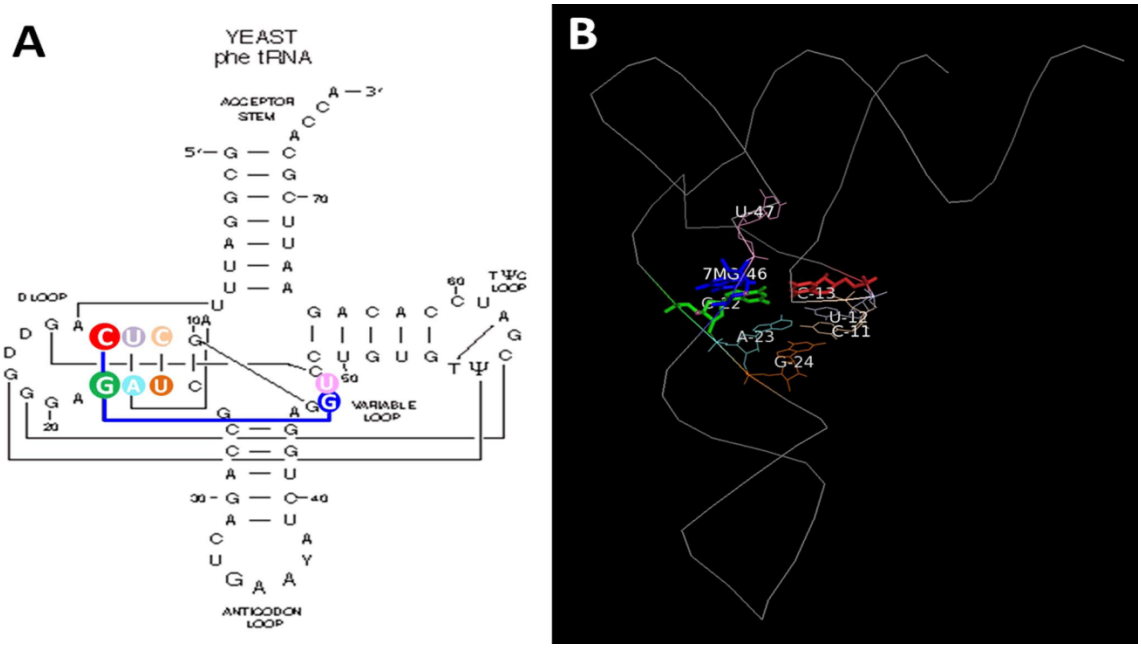


Figure 2.8: The secondary (A) and three-dimensional structure (B) of *S. cerevisiae Phe* tRNA with neighbor effect identified in 1992.

We utilize a standard one-directional N-Best strategy with some covariation constraints to identify a set of "neighbor effects" (Process colored green in Figure 2.5, details in the Method section). The physical distance between the positions forming the neighbor effect is determined using the reference crystal structure.

2. Application of Methods on Datasets

2.1. Datasets and the filtration process

The accuracy of the sequence alignment will influence the quality and significance of subsequent covariation analysis. The sequence alignments used in this study are generated from manual curation of more than twenty years of refinement. The data sets used in this analysis are bacterial 16S rRNA multiple sequence alignment (MSA) consisting of 4142 sequences, bacterial 5S rRNA MSA consisting of 2088 sequences, and bacterial 23S rRNA MSA consisting of 2339 sequences (details in the Methods section).

As mentioned above, the high-resolution three-dimensional crystal structure of *Thermus thermophiles* 30S ribosomal subunit [18] is utilized as the reference in the analysis of the 16S rRNA, while the high-resolution structure for *Escherichia coli* 50S ribosomal subunit [76] is used in the analysis of the 5S and 23S rRNA. The sequences in these crystal structures are used as the reference sequences.

Like most other covariation methods, PEC method performs exhaustive pairwise comparison: every column in the alignment is analyzed with every other column. For the 16S rRNA data set, the reference sequence has 1521 nucleotides, while the alignment contains 3,236 columns. Thus the total number of pairwise comparisons is 5,234,230. The time complexity of PEC algorithm on this dataset scales up to $O(4.4x10^{10})$. The PEC algorithm requires a significant amount of time to transverse the entire phylogenetic tree and count the number of changes during the evolution of the RNA. Since the positions with similar conservation scores have the higher likelihood to have good covariation score (Figure 2.3, details in Methods section), we used a coarse filter to eliminate those pairwise positions that were unlikely to have a significant covariation [74], and speed up the calculation process of PEC method. The coarse filter reduced the amount of pairwise comparison calculations to 14,276, which were processed by PEC method (details in Method section).

2.2. Performance Comparison of Different Covariation Methods in the Identification of Base Pairs

The performance of the PEC method in the identification of real base pairs in the bacterial 5S, 16S, and 23S rRNA alignment data sets was compared with other covariation methods including MIxy [45,46], MIp [57], OMES [58], ELSC [60], and McBASC [59]. The percentage of predicted base pairs that are present in the crystal structures are measured as a function of rank order. In addition to the covariation methods used here to evaluate the performance, we also tried to evaluate several other programs including PSICov [61], RNAfold [64,65], Direct information (DI) [62], RNAalifold [63], Evofold [68] and Pfold

[66,67]. However, these programs are either not suitable for the prediction of higher-order structure of RNAs with covariation analysis, or they are unable to operate on the large alignments used in our study.

The precision of top N ranked prediction plot reveals the fraction of pairs with ranked N or higher in each data set that are the contacting base pairs in the crystal structures. It has been utilized in several studies to gauge the precision of several covariation methods [57,61,77,78]. As shown in Figure 3, the PEC method performs better than Mixy and MIp, and significantly better than ELSC, OMES, and McBASC on the 16S rRNA alignment (Figure 3B). For the 5S and 23S rRNA alignments, PEC and the MIp achieve similar accuracies, which is significantly better than other methods, while ELSC, OMES, and McBASC methods are considerably lower (Figure 3A and 3C). The total event (positive events plus negative events) in PEC method measures the total amount of changes on a pair of positions throughout its evolution. Adding the total event threshold (e.g. >= 10) helps reduce the background noise and improves the accuracy of PEC method. PEC with total events threshold achieved higher accuracy than PEC without total events threshold in the 5S and 16S rRNA alignments (Figure 2.9A and 2.9B). However, that performance of PEC with or without total events threshold is exactly the same on the 23S rRNA data set (Figure 2.9C). Overall, the PEC method outperforms other covariation methods in the identification of base pairs, while MIp is the second best method.

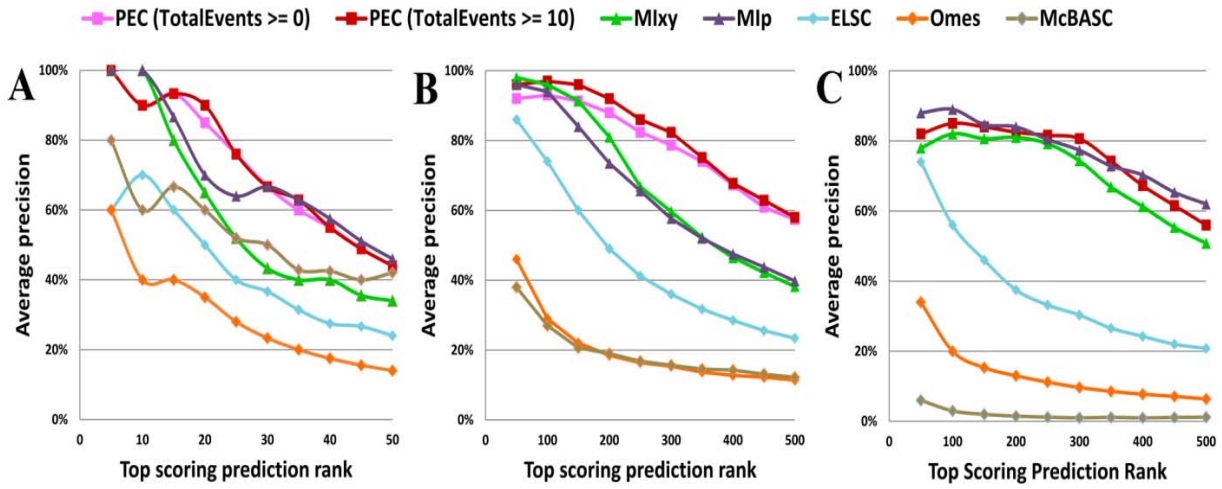


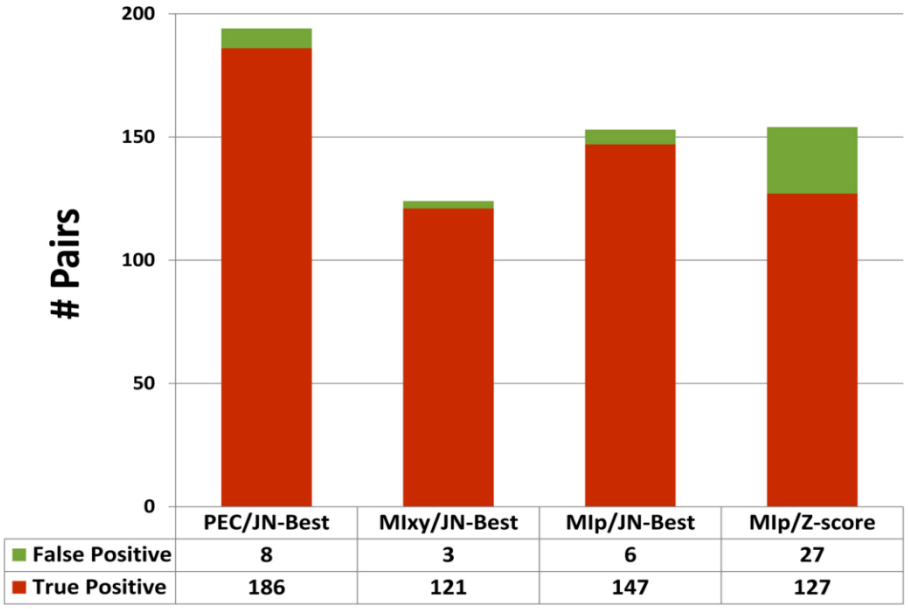
Figure 2.9: The precision of top N ranked prediction plot with different covariation methods in the identification of base pairs using different data sets: 5S rRNA data set (A), 16S rRNA data set (B), and 23S rRNA data set (C).

2.3. Application of Joint N-Best

The precision of top N ranked curve plot in Figure 3 reveals that the PEC, MIp, and MIxy methods are the top 3 methods in the identification of base pairs for the data sets. Mutual information (MIxy) measures the dependence of one position on another in the RNA sequence alignment. This measure was first introduced for identification of covariations in RNA [45,46]. In 2006, Lindgreen et al. evaluated 10 various MIxy-based covariation methods for the identification of covariations in RNA alignments [79]. Their results demonstrated that the standard MIxy is a good metrics for the prediction of base pairs in the RNA secondary structure, while several variations of MIxy improved the performance in the identification of base pairs. Dunn et al. developed an improved implementation of MIxy, named MIp, which estimated the level of background noise for each position [57]. After the removal of background and conversion to Z-Score (MIp/Z-Score), they determined that the MIp/Z-Score method identified substantially more co-varying positions than other existing MIxy-based methods.

Here we utilize the Joint N-Best strategy to measure the significance of the covariation scores calculated in different methods (details in Methods section). The Joint N-Best algorithm is applied onto PEC, MIp, and MIxy methods (PEC/JN-Best, MIp/JN-Best, MIxy/JN-Best) with the recommended (default) cutoff value of N-best score 0.5. We also used Z-Score conversion on MIp with the recommended Z-Score cutoff as comparison [57].

The PEC/JN-Best, MIxy/JN-Best and MIp/JN-Best methods are utilized on the 5S, 16S and 23S rRNA data sets to identify base pairs. The number of true positives (putative base pairs present in the reference crystal structure) and false positives (putative base pairs not present in the crystal structure) obtained by different methods on the 16S rRNA dataset are shown in Figure 2.10. The PEC/JN-Best method identifies 186 true positives with only 8 false positives (95.9% accuracy), while the MIxy/JN-Best achieves similar accuracy but much lower sensitivity (121 true positives, 3 false positives, 97.6% accuracy). The MIp/JN-Best obtains 147 true positives and 6 false positives (96% accuracy), and it identifies all but one pair found by MIxy/JN-Best. The MIp/Zscore method identifies 127 true positives, however the number of false positives – 27 decreases the accuracy (82.5%). In comparison to Z-Score conversion (MIp/Zscore), the utilization of Joint N-Best strategy with MIp (MIp/JN-Best) increases the number of true positives and decreases the number of false positives, thus improves both accuracy and sensitivity.



	PEC/JN-Best	MIxy/JN-Best	MIp/JN-Best	MIp/Z-score
False Positive	8	3	6	27
True Positive	186	121	147	127

Figure 2.10: The number of true positives and false positives identified with different covariation methods.

Since MIp/JN-Best method identifies all of the pairs found by the MIxy/JN-Best method except for the pair 150:159 (*Thermus thermophiles* numbering), we combine the non-redundant putative pairs identified in both methods. These pairs are referred as identified by Mutual Information Based Measure with Joint N-Best (MI/JN-Best).

The real base-pairs (true positives) identified by PEC/JN-Best and MI/JN-Best methods are plotted onto the *T. thermophilus* 16S rRNA secondary structure diagram (Figure 2.11). The total number of base pairs identified by both types of methods is 243, while the number of real base pairs identified only by PEC/JN-Best, only by MI/JN-Best, and by both methods are: 95 (red), 57 (green) and 91 (yellow). The ratio of the number of base pairs that are uniquely identified with PEC/JN-Best and MI/JN-Bes is 62.5%. Table S3 contains the detail results of these methods for 5S, 16S and 23S rRNA data sets.

Our results of the general comparison of these methods reveals: 1) with the default N-best cutoff (0.5), the PEC/JN-Best method has higher accuracy and sensitivity than MIxy/JN-Best and MIp/JN-Best in detecting covariant base pairs, 2) while both PEC/JN-Best and MI/JN-Best uniquely identifies base pairs that are not identified with

the other method, both methods also identified many of the same base pairs, 3) MIp/JN-Best was superior to the MIp/Z-score in detecting covariant base pairs for the 16S rRNA, and 4) MIp/JN-Best identifies a larger percentage of the base pairs found with by MIxy/JN-Best.

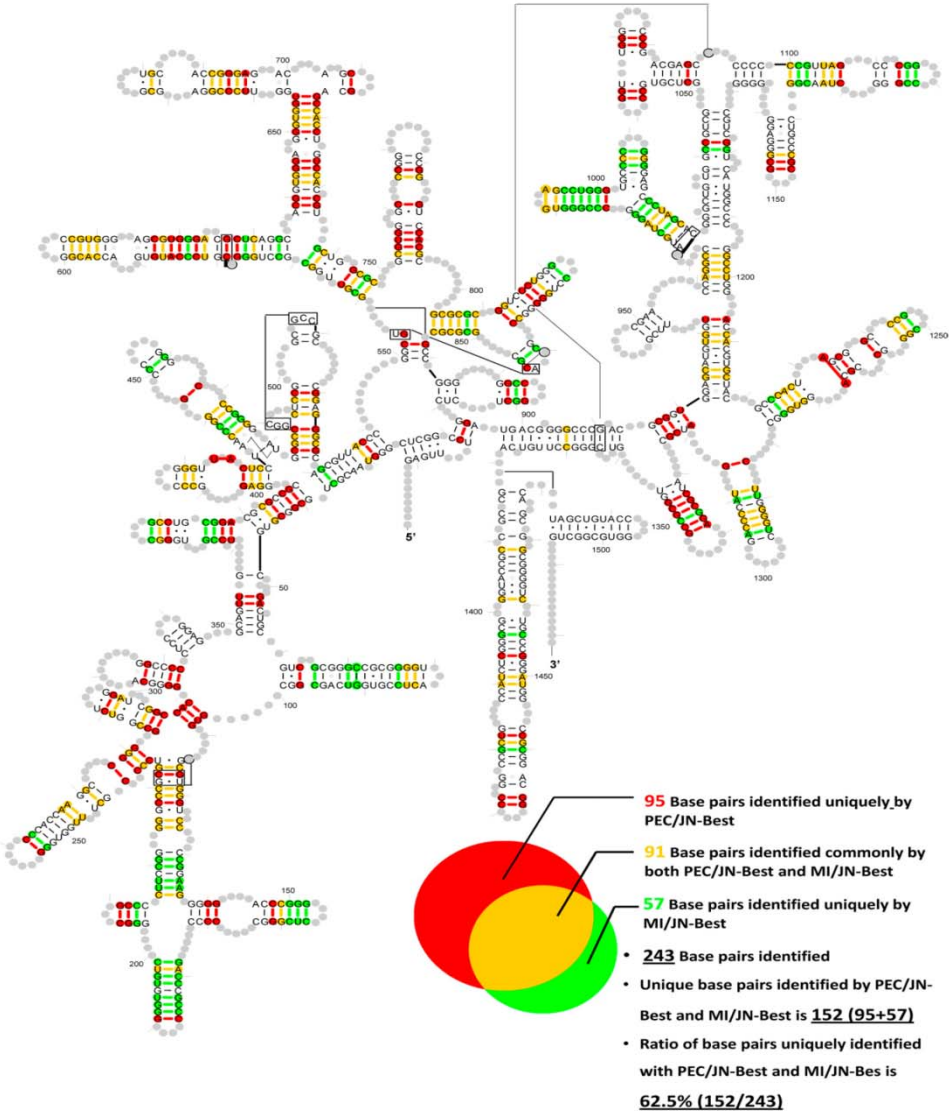
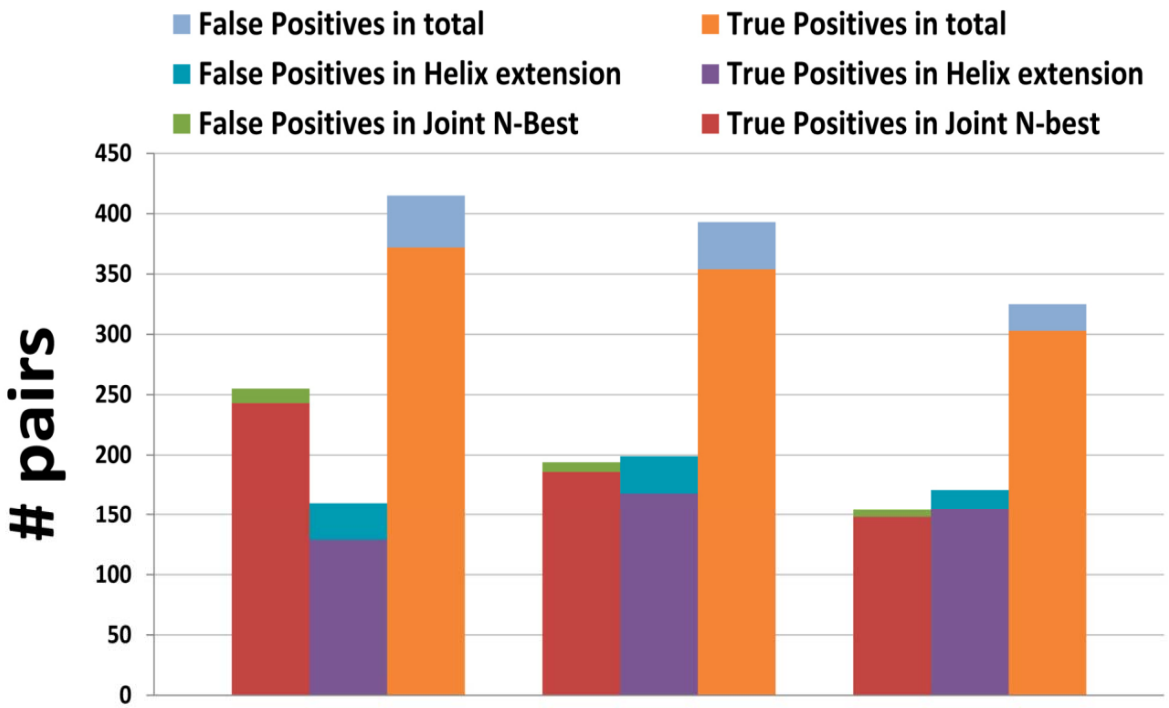


Figure 2.11: The base pairs (true positives) identified by PEC/JN-Best and MI/JN-Best are plotted onto the *T. thermophiles* 16S rRNA secondary structure diagram. Red: base pairs only identified by PEC/JN-Best; Green: base pairs only identified by MI/JN-Best; Yellow: base pairs identified by both methods.

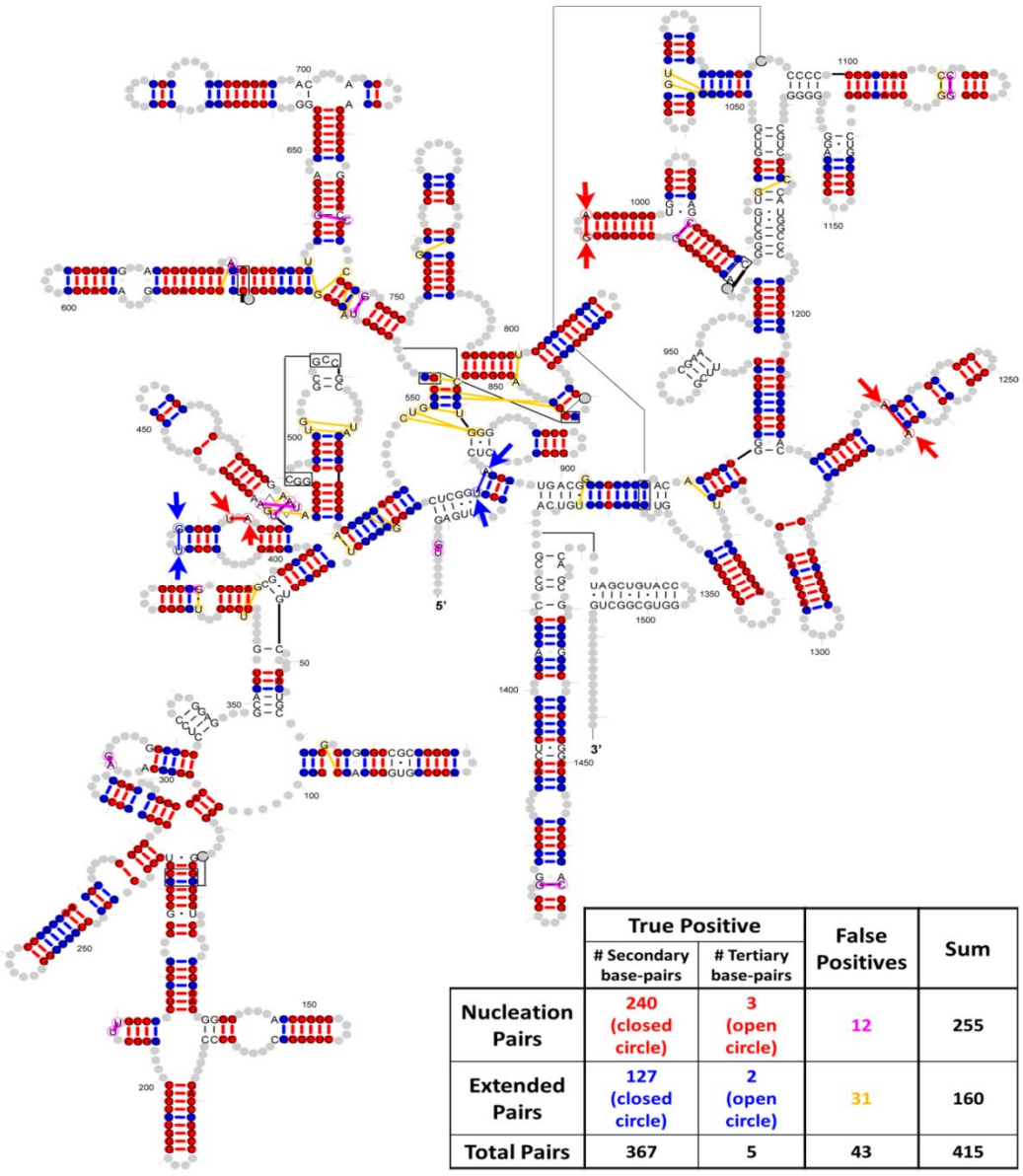
2.4. Identification of Highly Conserved Base Pairs with Helix-extension Strategy

All non-redundant predicted base pairs by PEC/JN-Best and MI/JN-Best methods are used as nucleation pairs in the helix-extension procedure. The set of adjacent and antiparallel nucleotides to the nucleation base pair with more than 85% WC/Wobble base-pairs in the alignment are considered an extended base pair. Additional base pairs that satisfy this helix extension threshold continue to be added to this extending helix until they fail the extending threshold. Figure 2.12 shows the number of nucleation pairs and extended pairs obtained in our helix extension analysis of 16S rRNA data set. When using the sum of predicted base pairs by both PEC/JN-Best and MI/JN-Best methods as nucleation pairs (255 pairs: 243 true positives plotted on Figure 2.11 and 12 false positives not plotted, Figure 2.12 left), the total number of extended pairs added with the helix extension is 160; 129 of these are true positives (present in the crystal structure), while the 31 false positives primarily occur at the end of helices. These nucleation and extended pairs are mapped onto the secondary structure diagram of *T. thermophilus* 16S rRNA in Figure 2.13. The number of nucleation pairs with PEC/JN-Best and MI/JN-Best, and the extended pairs in the helix extensions are also shown in Figure 2.12 (middle and right). This result reveals that with a set of nucleation pairs with high quality, the helix-extension strategy is able to identify those highly-conserved base pairs accurately and sensitively. The successful application of this helix-extension method onto the 5S and 23S rRNA data sets further substantiates this conclusion.



	Sum			PEC/JN-Best			MI/JN-Best		
	PEC/JN-Best + MI/JN-Best	Helix Extension	Total	PEC/JN-best	Helix Extension	Total	MI-JN-Best	Helix Extension	Total
True Positive	243	129	372	186	168	354	148	155	303
False Positive	12	31	43	8	31	39	6	16	22

Figure 2.12: For each method, the number of true positives and false positives identified in the Joint N-Best calculation (nucleation pairs), following helix extension procedure (extended pairs), and sum of them are shown as a stacked histogram.



	True Positive		False Positives	Sum
	# Secondary base-pairs	# Tertiary base-pairs		
Nucleation Pairs	240 (closed circle)	3 (open circle)	12	255
Extended Pairs	127 (closed circle)	2 (open circle)	31	160
Total Pairs	367	5	43	415

Figure 2.13: Base pairs in the Bacterial 16S rRNA structure model that are identified with the helix extension method. Red: true positive base-pairs identified as the sum of PEC/JN-Best and MIxy/JN-Best methods, which are used as nucleation points in the helix extension Magenta: false positives in the nucleation pairs; Blue: true positive base-pairs identified with the helix-extension method; Yellow: false-positive pairs identified with the helix-extension method. Secondary base-pairs are represented by closed circles while tertiary base-pairs are represented by open circle and highlighted with arrows.

2.5. The Purity and Conservation Scores of the Secondary and Tertiary Structure Base Pairs

Our results suggest that most of the identified base pairs are part of the secondary structure (represented as closed circle in Figure 2.13), while only a few tertiary structure base pairs are identified (represented as open circle and get highlighted by arrows in Figure 2.13): the Joint N-Best analysis identifies 240 secondary structure base pairs but only 3 tertiary structure base pairs; the helix extension procedure identifies 127 secondary base pairs but only 2 tertiary base pairs.

A quantitative and graphical analysis of 16S rRNA comparative secondary structure and the high resolution crystal structure for *Thermus thermophilus* 16S rRNA demonstrates the general observation noted in the previous paragraph – secondary structure base pairs usually have strong covariation between the two positions that form that interaction while the majority of the tertiary structure base pairs have weak or no covariation. For every pair of positions that form a base pair, the purity score which measures the precision of covariation (details in Method section and Figure 2.2), is plotted against the conservation score (details in Method section) (Figure 2.14). For both of comparative and crystal structures, two plots were created, the first for the standard purity score (Figure 2.14 left) and the second for purity scores adjusted for G:U base pairs (Figure 2.14 right, details in Methods section). The overall results from these plots are consistent with our base pair prediction as expected: 1) though base pairs in the bacterial 16S rRNA dataset range from highly conserved to highly variable, the most majority of the secondary structure base pairs are at or very close to a purity score of 1; 2) Many of the base pairs with a lower standard purity score increase their GU-plus score close to 1, which indicates the base pairs associated with these lower purity scores involve a G:U base pair; 3) The majority of tertiary structure base pairs do not have the highest purity scores, indicating that many of positions that form tertiary base pairs have no covariation, or some weak covariation with many exceptions, consistent with our previous observation [16] [http://www.rna.ccbb.utexas.edu/SAE/2A/xtal_Info/16S/Index].

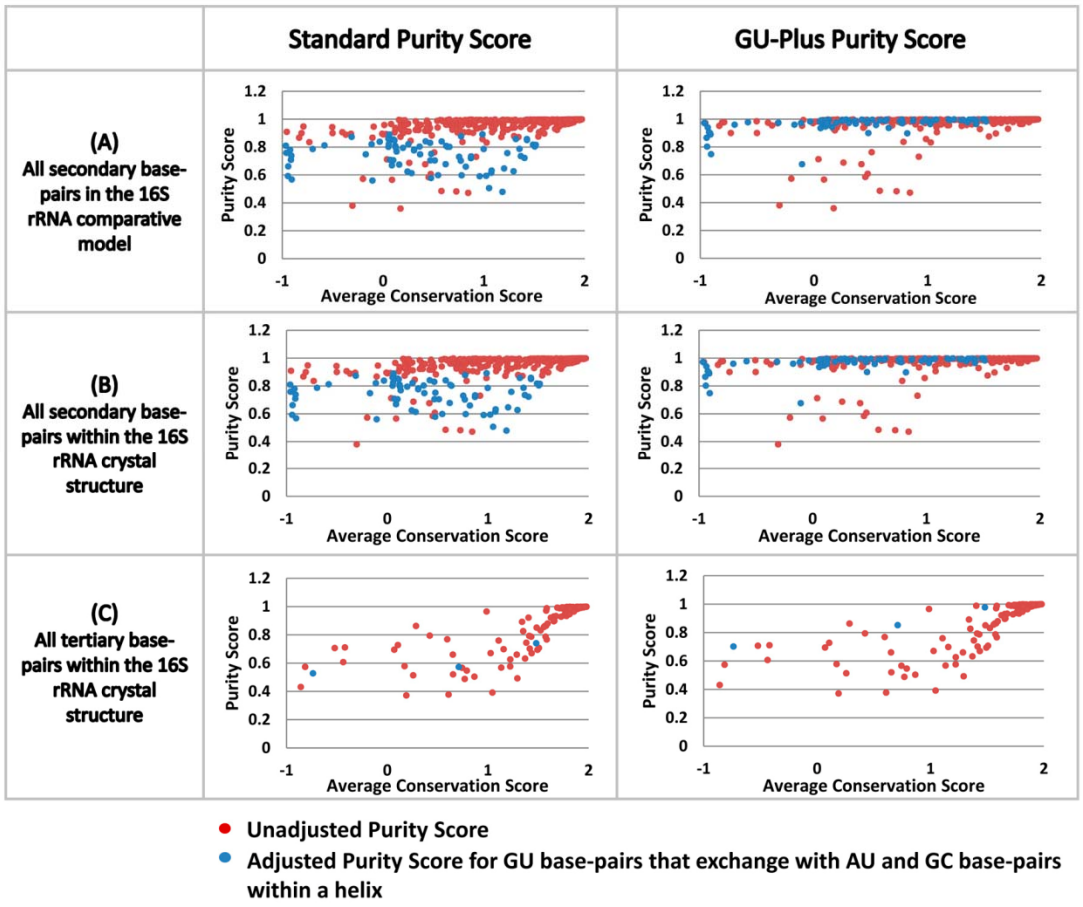


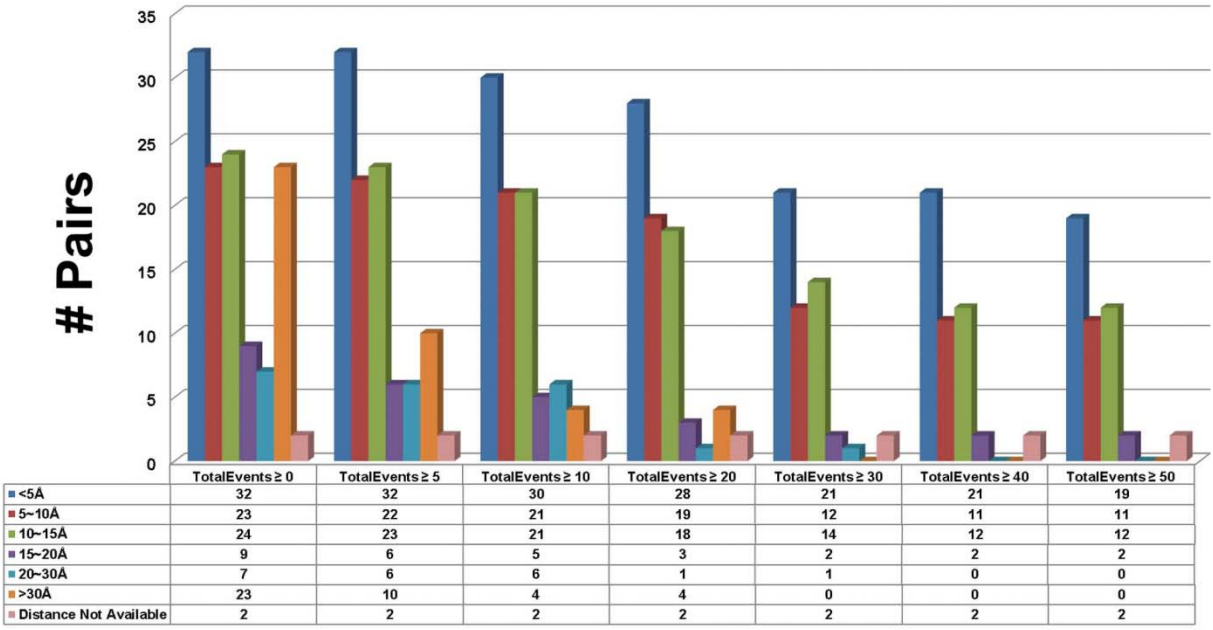
Figure 2.14: The distribution of purity score and average conservation (or informational entropy) for the two nucleotides that form a base pair in the 16S rRNA comparative structure model (A), secondary structure base pairs in crystal structure (B), and tertiary interactions in crystal structure (C).

2.6. The Identification of Neighbor Effects

Previous analysis has revealed that when two positions in a sequence alignment have very similar patterns of variation, as gauged with a high covariation score, those positions usually form a base pair in the RNA higher-order structure. However as the extent of positional covariation decreases, our observations here and in our previous analysis [46,49] reveals that some pairs with lower covariation scores form base pairs, and

others do not. While the full significance of these observations have not been determined, we have observed that the positions in these clusters of significant but lower covariation scores are usually very close with one another in the three-dimensional structure with the traditional, covariation methods, hereafter named neighbor effects [49,80].

The covariation scores (e.g. CPE, MIxy, MIp) of the highest and second highest positions for the base pairs identified in our PEC/JN-Best method are significantly different (threshold value of 0.5, see "The Joint N-Best strategy" in the Methods section). These putative base pairs are analogous to the tRNA base pair 3:70 as shown in Figure 2.7A left side. However the difference between the highest and the set of next highest positions in our Bacterial 16S rRNA dataset are smaller for numerous positions, analogous to Figure 2.7A right side and Figure 2.8. As shown in earlier sections of this manuscript and previous studies [52,53], phylogenetic event based covariation methods have the potential to identify covariations that are not observed with the traditional methods. Thus we use PEC method to identify the neighbor effects. The positions with the N-best scores exceeding a predefined threshold of ≥0.85 (see Methods section for details) and in close proximity are considered as neighbor effects. For this analysis, the physical distance is minimal for those positions that are defined to be a neighbor effect. This criterion is satisfied for those positions with at least 10 phylogenetic events (Figure 2.15).



	TotalEvents ≥ 0	TotalEvents ≥ 5	TotalEvents ≥ 10	TotalEvents ≥ 20	TotalEvents ≥ 30	TotalEvents ≥ 40	TotalEvents ≥ 50
<5Å	32	32	30	28	21	21	19
5~10Å	23	22	21	19	12	11	11
10~15Å	24	23	21	18	14	12	12
15~20Å	9	6	5	3	2	2	2
20~30Å	7	6	6	1	1	0	0
>30Å	23	10	4	4	0	0	0
Distance Not Available	2	2	2	2	2	2	2

Figure 2.15: The maximal distance between the positions defined to be a neighbor effect is determined from a comparison of the number of phylogenetic events. Different phylogenetic events and their number of positions with different physical distances were calculated. Those positions with at least 10 phylogenetic events contain a large number of positions that are very close in three-dimensional space and a very small number of positions with larger physical distances.

There are 89 neighbor-effect pairs identified and plotted onto the *T. thermophilus* 16S rRNA secondary structure diagram in Figure 2.16. Among these neighbor-effect pairs, 15 are annotated as known nucleotide interactions in the 16S *T. Thermophilus* rRNA crystal structure including 8 secondary base-pairs, 4 tertiary base-pairs and 3 base-triples (colored green in Figure 2.16). The remaining 74 pairs do not form hydrogen bonds between the bases (colored red in Figure 2.16). The average physical distance between these 89 neighbor effects is 8.82 ±5.91Å, while only four pairs (686:905, 686:930, 686:1209 and 686:1371, *T. thermophiles* numbering) are separated by more than 30Å. Most of these neighbor effects involve nucleotides that are either each nucleotide of the pair are on opposite sides of a helix, consecutive on the sequence, , adjacent to two nucleotides that form a base pair at the end of a helix, or involve a nucleotide in a loop

and a nucleotide in a helix that is very close to the loop. Neighbor effects are also identified on 5S and 23S rRNA datasets using the same parameter setting.

The observation of neighbor effects suggests that nucleotides that do not form a base pair can influence the evolution of other nucleotides that are physically in proximity. The complete structural and functional significance of these neighbor effects has not been fully determined. Several studies have revealed that: 1) nucleotides associated with base triples in and near the D stem in tRNA have moderately high covariation values [46,49] (Figure 2.8), 2) recent experimental studies of the ribosome discover that the D stem of tRNA is dynamic during protein synthesis [81,82].

Two other research groups have determined covariations by modeling phylogenetic relationships in bacterial 16S rRNA [52,53]. A detailed assessment of the similarities and differences of my results with their new covariations revealed that: 1) Both methods identifies a few new pairings with significant covariations; 2) Some of the nucleotides with a covariant pair identified with their methods are separated by a minimal distance (ie. neighbor effect), while many other nucleotides are separated by a much larger distance in the high-resolution crystal structure.

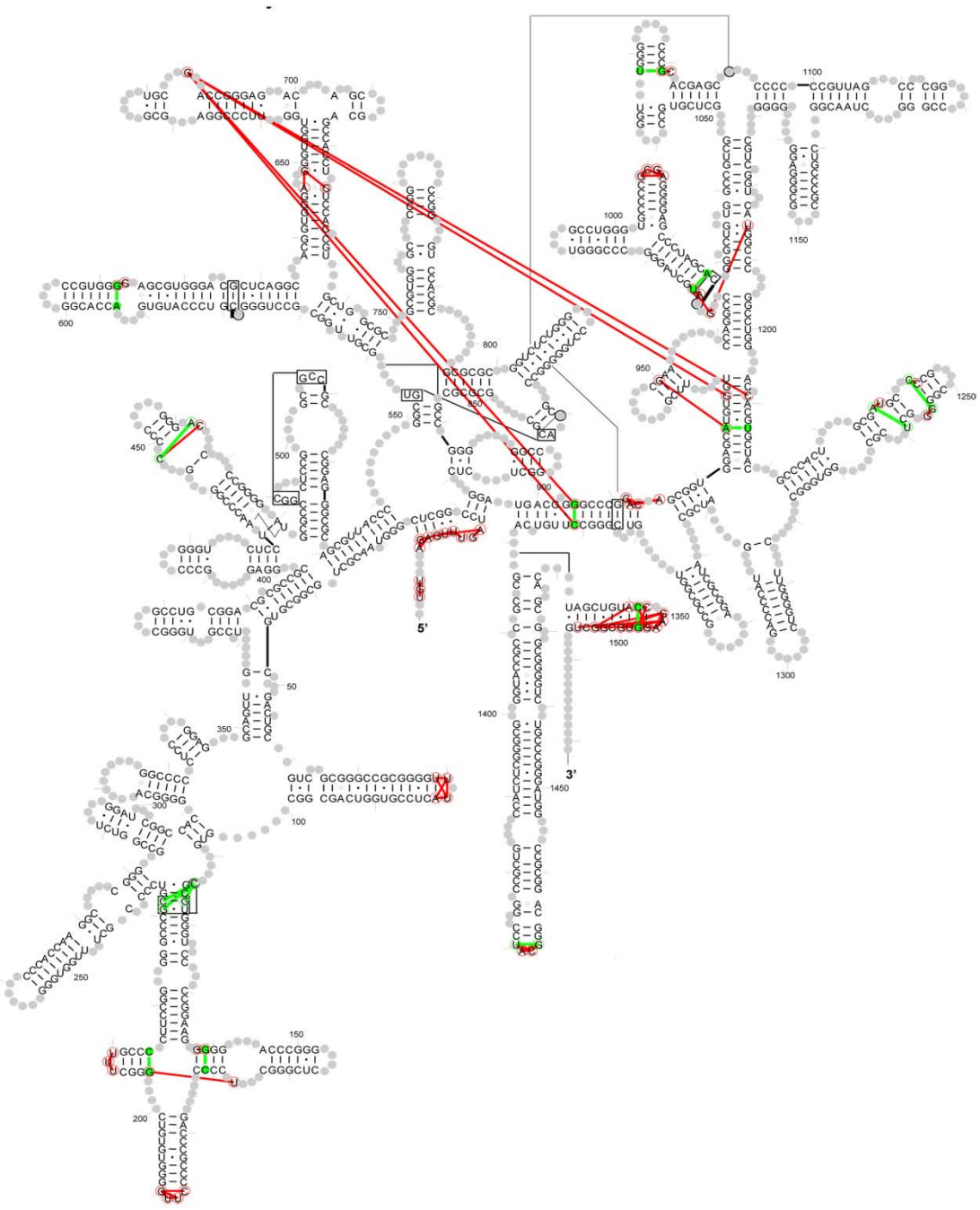


Figure 2.16: The secondary structural diagram of *T. thermophilus* 16S rRNA reveals all identified neighbor effects. Red lines connecting nucleotides indicate non-base-pairing interactions. Green lines represent the base-pairs or base-triples identified as neighbor effects.

Discussion

Improve the covariation methods using the evolution of the RNA structures

Previous research has revealed that the sensitivity and accuracy of the covariation analysis can be enhanced by integrating the evolutionary history of the RNA [47]. Our analysis of tetraloops in 16S rRNA discovered that this four-nucleotides hairpin loop that caps a helix can evolve from one common form to another many times during the evolution of the 16S rRNA [83]. For these studies, the number of times these positions changed during their evolution was determined after the base pairs and tetraloop were identified. The evolutionary dimension of the RNA structure provides temporal information to distinguish divergent and convergent evolution for specific positions and regions of the RNA. While our preference is to utilize the evolutionary history of the positions in the RNA to identify these base pairs and other higher-order structural constraints, monitoring these temporal changes is a significant computational challenge.

The Gutell lab's new RNA Comparative Analysis Database (rCAD) system cross-indexes multiple dimensional data of RNAs [19,74], which creates the opportunity to perform several types of novel analysis, including the phylogenetic event counting (PEC) covariation analysis..

The implementation of Phylogenetic Event Counting method (PEC) method and performance comparison with other covariation methods

The analysis reveals that overall the PEC method is superior to other covariation methods for the identification of base pairs in both sensitivity and accuracy (Figure 2.9). With the complementation of Joint N-Best strategy, PEC/JN-Best is more sensitive and accurate than the mutual information based methods that do not utilize the evolution of the RNA in its calculation (see Figures 2.10). The variation of standard MIxy method – MIp, when integrated with the JN-Best method, improves the standard MIxy method. Both PEC/JN-Best and MI/JN-Best method uniquely identifies many base pairs, and together identifies many other base pairs. The ratio of the number of base pairs that are

uniquely identified with PEC/JN-Best and MI/JN-Bes is 62.5% in the 16S rRNA data set (Figure 2.11) and 76.0% for the three rRNAs. Thus the combination of these two covariation methods significantly increases the number of identified base pairs.

The result also demonstrates that the sensitivity and accuracy of the covariation analysis is improved with the Joint N-Best. The vast majority of the base pairs with covariation analysis occur in secondary structure helices, while only a few tertiary base pairs are identified non-canonical base pairs, psueudoknots, and base pairs that begin to fold the secondary structure into a three-dimensional structure [11].

Prediction of base pairs with empirical rules for RNA secondary structure – Helix Extend

The helix extension method was initially used when the first 16S and 23S rRNA secondary structure diagrams were proposed from the analysis of the first few complete 16S and 23S rRNA sequences and many partial sequences [41,42]. However, as the number of sequences, and the diversity among those sequences increased, we have determined that nearly every base pair does have a covariation for datasets that include the Bacteria, Archaea, Eukaryotic nuclear encoded and the two Eukaryotic organelles. An assessment of the nucleotide conservation of the three primary domains of life – Bacteria, Archaea, and Eukaryotes reveals a significant amount of sequence conservation within each major phylogenetic domain [http://www.rna.ccbb.utexas.edu/SAE/2B/ConsStruc/]. For the later studies, different sets of alignments – (1) Bacteria, (2) Archaea, (3) Eukaryotes, (4) Bacteria, Archaea, Eukaryotic nuclear encoded, (5) nuclear encoded Bacteria, Archaea, Eukaryotes plus their two organelles – Mitochondria and Chloroplasts – were analyzed to identify covariation for nearly every base pair in the 16S and 23S rRNA structure model [16,23] [http://www.rna.ccbb.utexas.edu/SAE/2A/nt_Frequency-/BP/16S_Model].

In this analysis, many of the base paired positions in the bacterial 5S, 16S and 23S rRNA comparative model analyzed in study have no variation and no covariation, thus the rationale for the helix extension method (see Figure 2.12 and 2.13). The helix-

extension method facilitates the identification of many highly conserved or invariant positions in the bacterial rRNA helices.

The purity of the covariation between the two positions that form a base pair, and the identification neighbor effects

The purity of the covariations that underlies the prediction of a base pair range from an absolute 1:1 relationship (i.e. only base pairs with a strict covariation are found at a specific location in the structure, e.g. 70% A:U and 30% G:C) to base pairs with an increased number and types of exceptions (e.g. 40% A:U, 35% G:C, 15% G:U, 5% A:A, 3% A:C and 2% G:G). While we have higher confidence in the prediction of a base pair when its covariation is very pure, the prediction of a base pair becomes increasingly more difficult as the purity of the covariation decreases (see Figure 2.14).

The vast majority of pairs of positions with the strongest covariation scores are base paired in the RNAs higher-order structure. As the covariation scores decreases, many pairwise positions with lower covariation scores are still base paired, while some other pairs of positions with similar covariation scores do not form a base pair. Most of these positions are in close proximity in high-resolution three-dimensional structure, thus form neighbor effects [46,49] (Figure 2.16). While a complete understanding are still not known, these neighbor effects have been observed getting involved in base triple interactions in tRNA and group I introns [46,49] and could be involved in the fine tuning of tRNA structure in protein synthesis [81].

The majority of the tertiary structure base pairs do not covary with one another

The prediction of an RNA structure with comparative analysis has one primary underlying principle – the sequences of the same RNA family folds into a common secondary and three-dimensional structure. In other word, when base pairs are predicted by determining same pattern of variation of both positions in an alignment, it is implicitly

assumed that the sets of nucleotides that are base paired in an RNAs secondary and higher-order structure will have similar patterns of variation. Previous analysis of the high-resolution three-dimensional crystal structure of rRNAs revealed that the majority of the sets of nucleotides that form tertiary structure base pairs do not have similar patterns of variation – no covariation (details at the CRW site http://www.rna.ccbb.utexas.edu/-SAE/2A/nt_Frequency/BP/). This observation was substantiated by more recent studies [81]. The major reasons of forming these non-covariant tertiary structure base pairs include 1) While the different covariant base pair types can form similar conformations when two positions in an alignment have similar patterns of variation (e.g. G:C <-> A:U <-> U:A <-> C:G; C:C <-> U:U; A:G <-> G:A; etc.), non-covariant base pair types (e.g. G:A <-> A:A) can also form a similar conformation [49,84-86]. In a secondary structure helix, the non-covariant base pair types are unable to form similar base pair conformation due to their non-helical backbone conformation. However in the local structure flanking most of the tertiary structure, non-covariant base pair types can accommodate the non-helical backbone confirmation and maintain a similar base pair conformation. 2) Analysis of various tRNA high-resolution crystal structures revealed that different sets of tertiary structure interactions could form the same or very similar three-dimensional structures of the tRNA [49]. Thus sets of analogous positions of RNAs in the same family do not always form tertiary structure interactions, while sets of analogous positions usually form base pairs in a secondary structure helix,; 3) Analysis of the high-resolution crystal structures of ribosome reveals that though the ribosome (and rRNAs for this study) is dynamic, the secondary structure of the rRNAs remains the same during different stages of protein synthesis. The movement is primarily associated with changes in the tertiary structure interactions [87]. Thus, while our ultimate goal is to identify every base pair in an RNAs higher-order structure with comparative analysis, the current covariation analysis will not identify a high percentage of the tertiary structure base pairs.

In conclusion, utilizing the Gutell lab's new rCAD system, I have developed a more sophisticated covariation method based on phylogenetic events counting algorithm, This PEC method in combination with the enhanced mutual information, joint N-Best

and helix-extension methods creates a pipeline of programs that are superior to other existing covariation programs. This method has greater sensitivity and accuracy for the identification of the maximum number of secondary and other higher-order structural constraints including neighbor effects.

Chapter 3: CRWAlign-2: An Accurate Structure Template-based RNA Alignment and its application

Abstract

RNA has been discovered to be implicated in many more functions within the cell than just the message carrier between DNA and protein. The analysis of ribosomal RNA sequences is revealing more about the microbial ecology within all biological and environmental systems. The rapid determination of nucleic acid sequences dramatically increases the number of sequences that are available. Developing accurate and rapid alignment programs for these RNA sequences has been essential to decipher the maximum amount of information from this data. A template-based computational system, CRWAlign-2, that utilizes the Gutell lab's RNA Comparative Analysis Database (rCAD) is developed to align new sequences to an existing template sequence alignment. CRWAlign-2 retrieves multiple dimensions of information from rCAD, creates a profile based on sequence information, secondary structure, and phylogenetic relationships, and aligns new sequences into the template alignment using the generated profile.

The performance of CRWAlign-2 is compared with six widely-used template-based rRNA alignment programs and two best *de-novo* alignment programs on different sets of 16S rRNA sequence alignments with sequence identity ranging from 50% to 100%. The results reveal that CRWAlign-2 outperforms other alignment programs in aligning new sequences with higher accuracy. CRWAlign-2 also creates secondary structure models for each sequence to be aligned, which is very useful for the comparative analysis of RNA structures and sequences. Thus CRWAlign-2 can be used to align the very extensive amount of sequences determined by next-generation sequencing technology, which creates opportunities for numerous types of large-scale data analysis, such as the identification of the chimeric sequences generated in microbiome research projects.

Background

The comparative method is widely used in many research areas for RNAs, and is fundamental for the computational analysis of large-scale sequencing data analysis. Many of the comparative analysis utilize a alignment of homologous RNA sequences, which juxtapose similar structural and/or functional elements into the same set of columns. The analysis of these alignment are used to discover the secondary and higher-order structure, patterns of structural variation and conservation, evolutionary relationships, and association between RNA's structure and function. Thus accuracy of the alignment will determine the quality of the subsequent analysis. A few of the seminal discoveries include: the determination of the phylogenetic relationships for organisms that span the entire tree of life and the identification of the third kingdom of life – the Archaea [88], the accurate prediction of RNA secondary structure and constraints in the higher-order structure [16,37], the identification of new structural motifs [11], the creation of pseudo-energies for many RNA structural elements and their utility in improving the accuracy of folding an RNA sequence into its secondary structure [89], and the identification of the Microbiome - the collection of microbes in different ecological environments, using 16S rRNA [90-92].

The advent of next-generation sequencing (NGS) method brings a deluge of nucleic acid sequences and rapidly enhances our understanding of many different biological systems. Thus the development of more accurate and faster automated alignment methods has become an essential and challenging task for optimal analysis and interpretation of the results. The two most widely-used alignment strategies are *de novo* alignment, and template-based alignment.

De novo alignment programs, such as CLUSTAL [93,94], MAFFT [95], and SATe [96], generate multiple sequence alignments without the guide of any pre-refined alignment (seed/template alignment). Template-based alignment programs use a seed/template alignment as the reference to facilitate the alignment of new sequences. The seed alignment is usually manually curated to optimize its accurate juxtaposition of nucleotides. Several research groups have developed automated template-based sequence

alignment as web servers for different RNAs: Silva [97] aligns 16S and 23S rRNA, Greengenes [98] only aligns 16S rRNA, and RDP [99] only aligns 16S rRNA. Silva utilizes SINA (SILVA Incremental Aligner) which is implemented with a variant of the Needleman-Wunsch algorithm [100]. It uses a maximal of 40 various seed sequences, and switch between them then aligning various regions. Greengenes aligns new sequences with the Nearest Alignment Space Termination (NAST) algorithm [98], which performs BLAST [101] to identify the most closely matched seed sequence and then do a pairwise alignment. RDP is a secondary-structure based aligner, which switched to Infernal [102] from release 10.

Several other stand-alone template-based alignment programs are available for download: Infernal, ssu-align [103], and HMMER [103-105]. Both infernal and ssu-align build consensus secondary structure profiles for the template alignment which guide the alignment of new sequences, while ssu-align is implemented with additional integrative profile hidden Markov models (profile HMMs) on the consensus structure profiles. HMMER aligns new sequences with profile HMMs without creating the consensus secondary structures. While both infernal and HMMER are capable of aligning any type of RNA, ssu-align is currently limited to the 16S rRNA.

Another approach of template-based alignment utilizes a seed/template alignment with the correct secondary structure of that RNA molecule to generate a descriptor that defines the primary and secondary structural constraints. The sequences that satisfy all conditions of the descriptor are identified with candidate structural models. This procedure has been implemented in several previous programs. RNAMot [106] is one of the first in this family of programs. It is developed with a simple descriptor syntax that facilitates manual generation of the descriptor, but only captures limited details in the structural constraints. RNAMotif [107] has a richer descriptor syntax with greater specificity and complexity of the RNA structural constraints that can be distinguished and identified. Locomotif [108] is developed with a graphical descriptor editor and dynamic search algorithm. While these programs are an improvement over the original RNAMot and provide some provisions to align new sequences, their performance are not adequate for

aligning large numbers of large RNA sequences with great specificity due to lack of several essential functions. First, these programs are unable to identify larger RNA molecules (e.g. 16S rRNA, 23S rRNA). For example, RNAMotif program can search a structural descriptor with a maximum of 100 structural elements, while 16S rRNA consists of over 400 structural elements. Second, the descriptor syntax implemented within these programs is unable to encapsulate all possible variations in the RNA molecules, results in non-optimal candidates of search process. And third, these programs require the descriptor to be generated manually, which could cost substantial amount of time and effort.

When sequences to be aligned have maximum identity with one another, both de novo and template-base alignment methods align sequences with high accuracies. However, for sequences have minimal identity, *de novo* alignment algorithms are unable to placing the nucleotides sharing common structural/functional features within each sequence into the correct columns of the alignment. In contrast, the template-based alignment algorithms utilize the previously determined seed alignment that has been refined to maximize the correct juxtaposition of structural, functional, and evolutionary relationships of the sequences. Until we are able to capture all constraints encrypted in the large RNA molecules into de novo alignment algorithms, the template-based alignment algorithms will be more accurate in the generation of new multiple sequence alignments.

I developed a new template-based alignment system, CRWAlign-2, which utilizes the Gutell lab's RNA Comparative Analysis Database (rCAD) relational database management system that cross-indexes multiple dimensions of information, including sequence alignments, comparative secondary structures, and phylogenetic relationships [109]. CRWAlign-2 1) analyzes the seed/template alignment with secondary structural information, and automatically generates the structural profile/descriptor containing sophisticated sequence and structural constraints for specific and generalized phylogenetic groups, 2) searches for and creates complete structural models satisfying this profile/descriptor, and 3) aligns the new sequences against the template alignment.

The structural information used by CRWAlign-2 is obtained from rCAD system and the CRW Site (http://www.rna.icmb.utexas.edu/) [23]. The phylogenetic information in rCAD is obtained from the NCBI Taxonomy database (http://www.ncbi.nlm.nih.gov/-Taxonomy/taxonomyhome.html/). CRWAlign-2 is capable of aligning sequences for any type of RNA molecule that has an existing high-quality template alignment and an accurate secondary structure model regardless of the size of the molecule.

The primary objective of this study is to measure the performance of CRWAlign-2 in aligning new sequences, and compare with six widely-used template-based alignment programs (three web-based aligners: Silva, GreenGenes, and RDP; three stand-alone aligners: Infernal, ssu-align, and HMMER), and two de novo alignment programs (SATe and MAFFT). For a rigorous assessment across all alignment programs, the bacterial 16S rRNA is selected as the test set, which has at least 1,400 nucleotides per sequence in length. The results reveal that CRWAlign-2 is superior to other programs in aligning new sequences with higher accuracy and generating more useful structural models besides the new alignment.

The superior performance of CRWAlign-2 provides extensive research opportunities for multiple areas. One of the benefitted hot areas is microbiome research. The term "microbiome" was originally defined as "the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space" [110]. More research has revealed that microorganisms inhabiting inside the human body play essential roles in health and disease. The advent of next-generation sequencing (NGS) method brings a deluge of nucleic acid sequences and enhances our understanding of the bacterial and archaeal world around us. Many scientists dedicating in microbiome research isolates and identifies the collection of microbes in different ecological environments.

The 16S rRNA is the primary sequence to analyze and evaluate the microbial composition in the microbiome research. Despite the fair amount of effort spent on removing low quality sequences, more recent analysis suggest that a significant

percentage of these 16S rRNA sequences from microbiome research are likely to be artifacts rather than real biological diversity. One major source causing this dilemma is the formation of chimeric sequence during PCR amplification [111,112]. A single-strand incompletely extended sequence from an earlier PCR cycle can work as a primer in the subsequent extension cycle. When there are more than one type of template sequences exists, these aborted extension product can anneal to an improper template, and form a chimeric sequence. Several previous studies suggested that current curated sequence database may contain up to 45% chimeric sequences [113-115]. Therefore it is essential and challenging to identify these chimeric 16S rRNA sequences.

The two strategies most widely used in 16S rRNA chimera detection are 1) aligning query sequence onto a chimera-free reference alignment and calculating pairwise evolutionary distance; and 2) using BLAST to search NCBI database for taxonomic anomalies. Pintail [113] and Mallard [116] use Clustal [117] to align the query sequence to all or all pairs of sequences in a trusted chimera-free reference sequences. The evolutionary distance is calculated across the query sequence while large deviation from the expected evolutionary rate indicates a chimera. Bellerophon aligns query sequence using GreenGenes [98], and calculates the a evolutionary distance matrix between every pair of sequences for the left and right fragments at an assumed break point [115]. ChimeraChecker [118] utilizes BLAST to search the closest match of different regions. When the closest match for region one is different that of region two, the query sequence is marked as potential chimera. While these chimera-detection methods are used widely in diverse research, their sensitivity and accuracy are not satisfactory.

The performance of the chimera checking programs is affected by two essential factors: 1) the accuracy of the generated sequence alignment, and 2) the approach used in sequence comparison. The 16S rRNA sequence has 1542 nucleotides (*E. coli* number), and contains multiple highly variable regions which is the major source of inaccurate alignment. The existing sequence alignment programs, such as Silva, GreenGenes, RDP, SATe, and Clsutal can not align 16S rRNA sequences with satisfactory accuracy [119].

Thus, the generation of the most accurate sequence alignments is absolutely essential for the subsequent chimera detection analysis. The other factor, the approach used in sequence comparison, is also critical to assure high accuracy and sensitivity in the detection of chimeric sequences. Both BLAST search and evolutionary distance calculation that are widely used in existing chimera detection algorithms are based on pairwise comparison between a query sequence and a subject (reference) sequence. Though the pairwise comparison reveals useful information by calculating the mismatches between the query and subject sequences, it does not use the multi-dimensional sequence information as effectively as it could. The result could be biased by incomplete or false taxonomy information associated with reference sequences (in BLAST-based methods). Therefore, creating a more accurate sequence alignment and implementing a more sophisticated sequence comparison strategy incorporating more dimensions of sequence information is vital and necessary to improve the accuracy and sensitivity of the chimera detection.

I developed a chimera-checking program that utilizes the most accurate sequence alignment algorithm, CRWAlign-2, and a more sophisticated strategy for sequence comparison. With a well-aligned reference sequence alignment and taxonomy information associated with each reference sequence, the comparison between the query sequence and the entire set of reference sequences at a taxonomy branch can provide more useful information to discover the features of phylogenetic groups and the similarity between the query sequence and the specific phylogenetic group. My chimera-checking program uses both high-quality reference sequence alignment and taxonomy information to generate the statistical characteristics of different phylogenetic groups and analyze the query sequences with the generated statistical signatures. The query sequences are first aligned onto the reference sequences alignment utilizing CRWAlign-2. Then all reference sequences are mapped onto the phylogenetic tree based on the known taxonomy information, which generates a tree representing the phylogenetic relationships of the reference sequences. The chimera-checking program traverses from top (root node) to bottom (leaf node) of the phylogenetic tree. At each node of the tree, the signature

difference between the query sequence and the reference sequence alignment is measured with multiple statistical metrics. Based on the calculated signature difference, the query sequence is categorized as a member of that phylogenetic branch, or a chimeric sequence consisting of multiple fragments from different phylogenetic branches.

Methods

1. CRWAlign-2

The CRWAlign-2 system is an strongly enhanced version with numerous expanded and novel functions of the RNAMotif program [107] that was developed primarily to identify sequences that satisfy the secondary structure constraints in the descriptor. The enhanced CRWAlign-2 program 1) has a richer and more sophisticated descriptor syntax that provides greater specificity; 2) analyze the template alignment and automatically generates a descriptor; 3) is capable of operating on much larger RNA molecules (e.g. 16S and 23S rRNA); 4) searches and creates secondary structure models for each sequence; 5) aligns new sequences automatically based on analogous primary and secondary structural similarity; 6) is written in C# and directly exchange data with MS SQL (rCAD).

Stage1: Computer Generated Secondary Structural Descriptor

The first stage is to automatically create a structural descriptor that contains information describing various constraints applied to the canonical, regular, or standard RNA secondary structure and relevant taxonomy. The descriptor syntax is based on the original RNAMotif program [107] but enhanced significantly to improve detail of the encapsulated structural constraints and specificity. The most important enhancements include: a) each structural constraint (*e.g.* the length of helix/unpaired region, the mispair (or non-canonical base pair) number for helix) are described more accurately with an assigned weight score which indicates the frequency/occurrence of the variable; b) to

reduce the running time of structure identification process, a weight score cutoff section is constructed at the end of each descriptor that defines the lowest score of an elongating structural model; and c) the program will automatically generate descriptors at each phylogenetic branch to describe the most relevant structural constraints applying on that phylogenetic group. These enhancements in the descriptor significantly improve the specificity of the descriptor or different RNA molecules and phylogenetic groups.

As shown in Figure 3.1C, the major fields in descriptor include

- Params: this section define the base pairing rule, e.g. "wc" just consider Watson Crick base pair, while "wc += gu" consider both Watson Crick and Wobble base pair type.
- Descr: the main section of the descriptor. Two major types of structural elements are defined – helix (h5/h3) and single strand (ss). The format for **h5/3** and **ss** are:
    - h5(tag =helix name, {len1 = x1: weight = w1; len2= x2, weight=w2}, {mispair = m_a: weight = w_a; mispair = m_b: weight = w_b} where:
        - Length and mispair are discrete numbers defining the allowed lengths and mispairs. These numbers are not defining the minimum and maximum lengths and mispairs. At least one set of lengths and mispairs are required.
        - Weights (column) are the frequency for each length and mispair, where the frequencies for all of the lengths (and mispair) sum to one.
        - The presence or absence of mispairs at the end of a helix is defined with the **ends** variable ("mm" allows mispairs at both 5' and 3' end of helix; "pm" only allows mispair at 3' end of helix; "mp" only allows mispair at 5' end of helix; "pp" disallows any mispair at neither 5' nor 3' end of helix").
    - h3 (tag = helix name) is the sequence associated with the 3' half of helix.

- ss(tag = single strand name, {len1 = x1: weight = w1; len2 = x2: weight = w2}, seq = regular expression of nucleotide pattern)
    - Similar to h5, length are discrete number defining allowed lengths with weights.
    - Allowed sequence e.g. seq = "^UG*AG$" defines a single strand sequence that starts at the 5' end with "UG", "*" designates an insertion or deletion with no sequence specificity, and terminates at the 3' end with "AG". While "^" and "$" force the UG and AG to be adjacent to the end of the flanking structural element, their absence allows the UG and AG sequences to occur anywhere within the single stranded region.
  - The example in Fig. 3.1A, B, and C has one hairpin composed of h1 (h5,h3) and ss1 (ss). The Descr for this simple secondary structure has only three lines, one for each structure element, in the order they appear in the sequence 5' to 3' - h5, ss, and h3.
  - The Descr and secondary structure diagram for tRNA – phe secondary structure with four helices (eight structural elements) and five single stranded, is illustrated in Fig. 3.1D.
- **Sites**: define the allowed base pair types (e.g. only {A:U, U:A, C:G} at base pair 1 in helix 1) at specific positions in a helix (pos=1/pos=$-0, pos=2/pos=$-1, pos=3/pos=$-2, … pos=x/pos=$-(x-1)).
- **Matrix**: SxE - defines the number of sequences in an alignment (S) and number of structural elements defined in the descriptor (E). For example "7x3" is for an alignment with 7 sequences in the alignment and 3 structural elements defined in the descriptor (h5 of helix 1, single strand 1, and h3 of helix 1).
- **Weights**: Each sequence in the alignment has a weight (row) for each structural element (E).
  - For the example (Fig 3.1A), the first sequence (seq1) in the alignment has all three structural elements (h5, singleStrand1, h3). The weights of H5/3

are derived from the length and mispair weights. Seq1 in the alignment (Fig. 3.1A) is “AUCGU”:”ACGAU” in helix1, the length (# bps) is 5 and the mispair is 0. The weight for the first element h5 of helix1 is 0.714 x 0.857 = 0.612. Seq1 is “UUAG” in ss1, the length is 4. Thus the sequence weight for ss1 is 0.571. Therefore, "seqIndex:0 0.612, 0.571, 0.612" means that the first sequence in the alignment has all three structural elements (h5, singleStrand1, h3), and according to the constraints identified, it has weight score 0.612 for h5 of helix 1, 0.571 for single strand 1, and 0.612 for h3 of helix 1.

- **Cutoffs**: the search for complete structural model could start at any structural element. As the search is proceeding and the structural model is extending, the program will keep checking the overall weight score of the extending model. If the model is abnormal (say it is consisting of consecutive structural elements with very low weight score), the extension will be terminated since this model is very likely to be false. For example, if the extending model has a helix with 3 base pairs, and a loop (single strand) with 7 nucleotides, while the descriptor defines that helix with length = 3 is very rare (say its weight is 0.01), and the loop with length 7 is also very rare (say its weight is also 0.01), then this model is very likely caused by chance because its overall weight score is very low 0.01(h5)x0.01(ss)x0.01(h3)=0.000001. The program will check the matrix defined in Cutoffs section, and determine if that 0.000001 is significant. If it is not significant, the extension will be terminated.

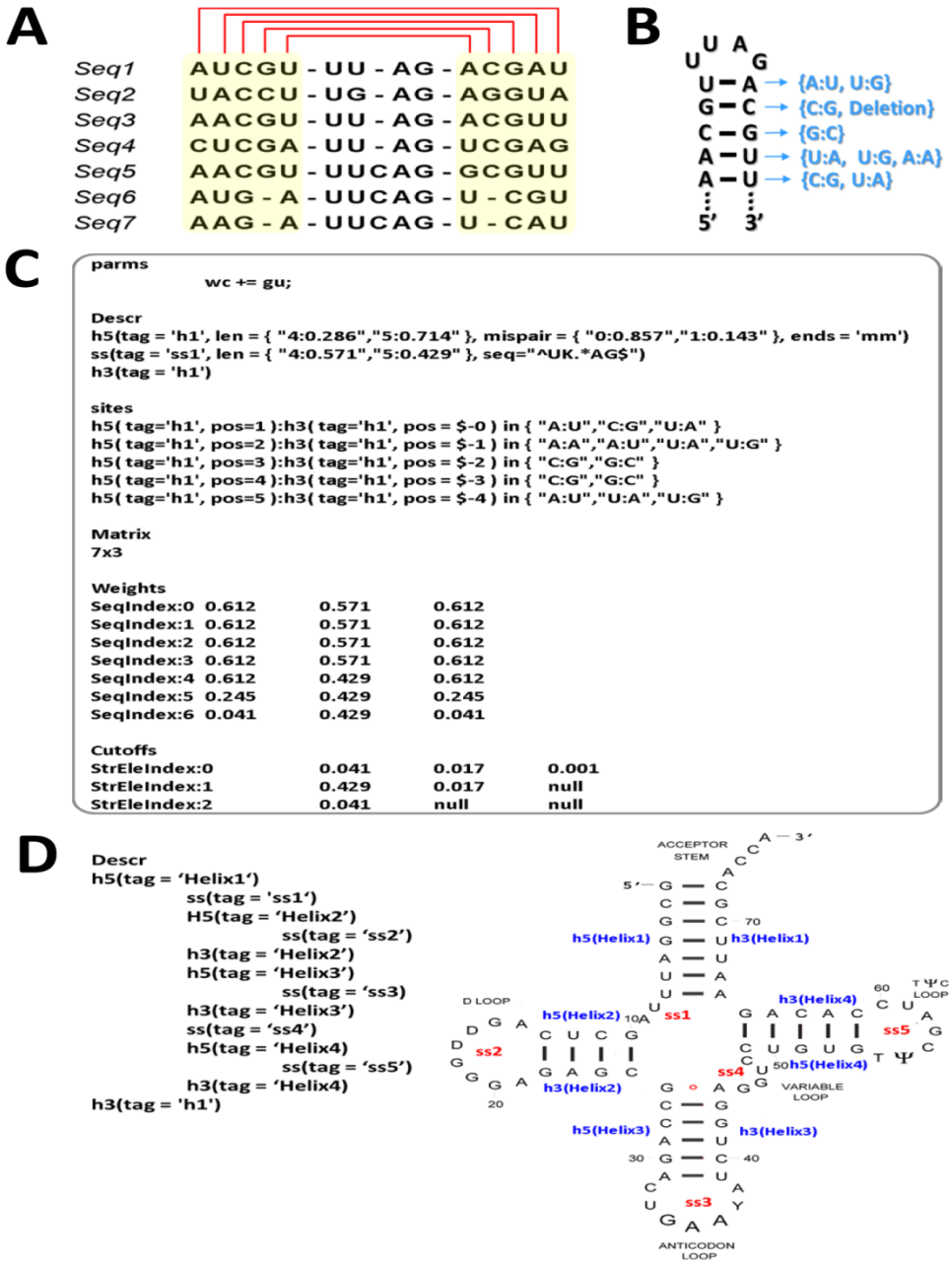


Figure 3.1: The generation of structural descriptor. The sequence alignment (A) and a common secondary structure (B) are used to generate the structural descriptor (C). For RNA molecules like tRNA (D), the main section ("Descr" section) is consisting of multiple descriptive lines while each line describes a structural element and the constraints applied.

Figure 3.2 illustrates the overall process of generating descriptors for secondary structural elements that occur within the analogous region of an RNA molecule from the template sequence alignment. As shown in Figure 3.2a, three phylogenetic nodes are under consideration, and for each of the three nodes, variation in nucleotide composition is shown in the secondary structure diagram. The first sequence in each node is shown in black, while accommodated variations within each node are shown in blue. The aligned sequences in the same phylogenetic nodes are grouped together in the sequence alignment, *i.e.* seq1-2 under node 1, seq3-4 under node 2, and seq5-7 under node 3 (Figure 3.2b). The positions that form base pairs with one another are highlighted and connected with red lines. At the onset, a generalized structure descriptor (Fig. 3.2c) that contains the structural constraints, such as the length of helix and unpaired region, the mispair number, the nucleotide conservation, *etc,* for every sequence in the template alignment is created. In the subsequent process, the structural descriptors for all phylogenetic nodes (Nodes 1, 2, 3 in Fig. 3.2c) are generated to provide the most relevant structural constraints for each of the phylogenetic nodes. To simplify the example shown here, only the main body of the descriptor is shown in Figure 3.2c.

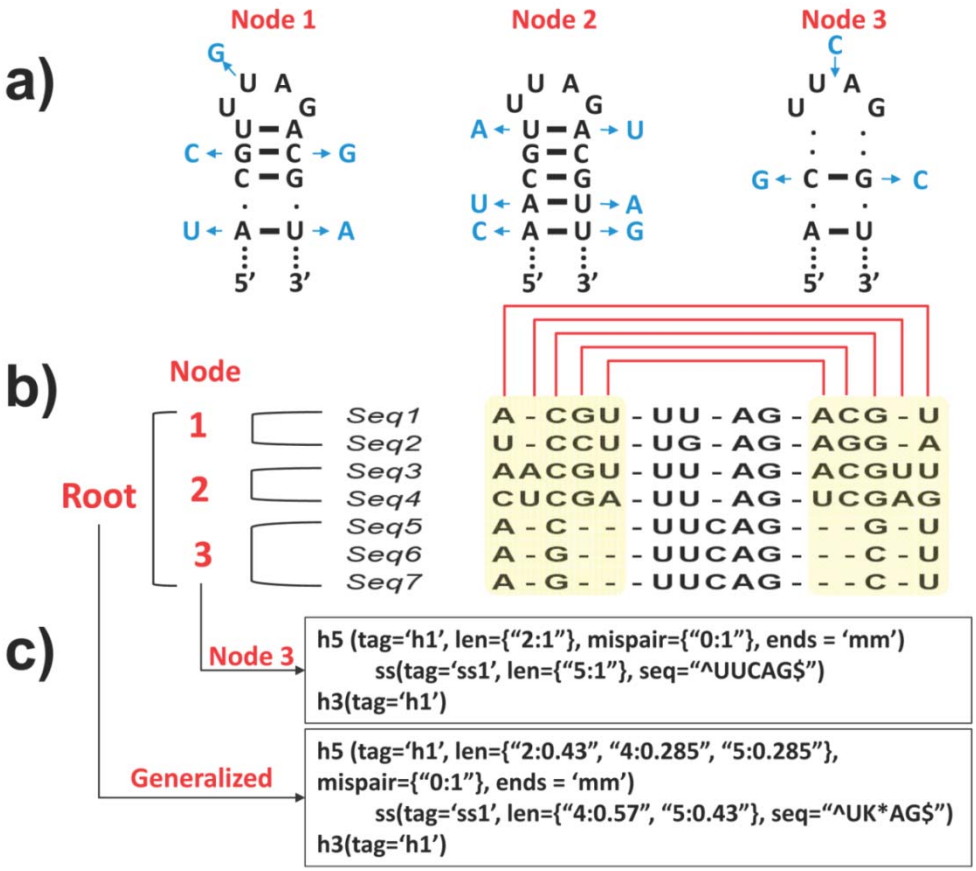


Figure 3.2: For CRWAlign-2, (a) secondary structure diagrams for each of the three phylogenetic nodes. *Seq1* in node 1, *Seq3* in node 3, and *Seq5* in node 3 are shown with black nucleotides; Blue nucleotides reveal the differences between the first sequence in each node (Seq1, Seq3, and Seq5) and the other sequences within each node; (b) Template sequence alignment with seven sequences distributed over three phylogenetic nodes. Red lines above the alignment indicate columns in alignment that form a base pair; (c) RNAMotif structural descriptors for node 3 and for all seven sequences (root).

Stage 2: Identifying Secondary Structural Elements and Creating Secondary Structure Models

After generating structural descriptors for all relevant taxonomic nodes that are present in the template alignment, CRWAlign-2 searches a sequence to identify the

structural elements defined within the descriptor on a sequence, and build the candidate structural models. The flowchart in Figure 3.3 illustrates the overall process of searching structural elements and creating complete structural model.

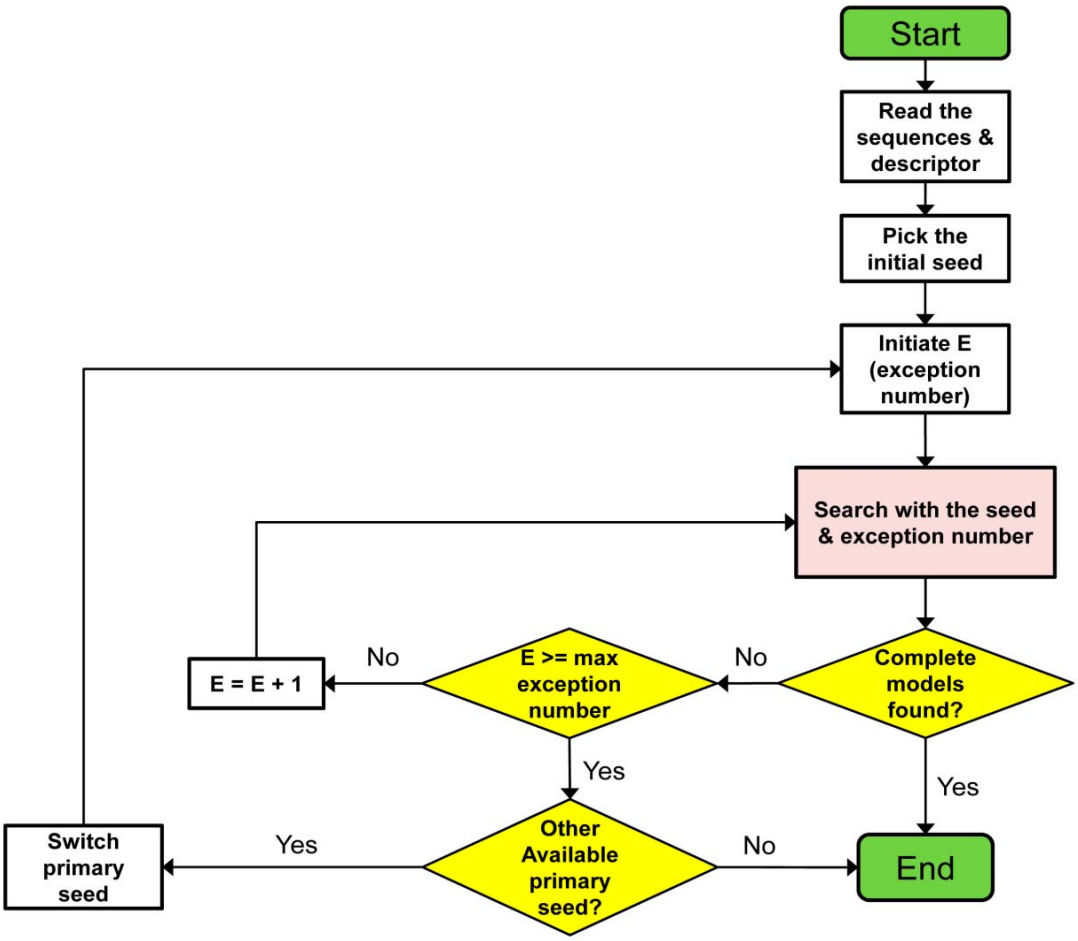


Figure 3.3: The flowchart of the complete structural model identification process for CRWAlign-2. The program reads the structural descriptor and sequences to be aligned, prioritizes structural elements in the descriptor to build seed points, and iteratively searches for complete structural models on the sequences that satisfy all structural constraints defined in the descriptor (see text in Methods section for details).

The first step is to read the new sequences (*e.g.* GenBank entries) and the descriptor profile that was generated in Stage 1 into the memory, and determine the most relevant descriptors for new sequences. For each sequence to be searched that contains

phylogenetic information, the most closely related phylogenetic node with existing descriptor is determined. Each structural element defined within the descriptor is formatted with all constraints, and organized with its relative order in the reference secondary structure.

The next step prioritizes or ranks the structural elements in the descriptor by their stringency, as measured with a probability score for each structural element. The more stringent, the less likely it will occur by chance, and thus has a lower probability score. The identification of the structural model starts with the structural elements with the lowest probability scores, which serve as the initial seed or nucleation points for further structural extension.

The program then attempts to identify the structural elements that are adjacent to these initial seeds. When there is more than one structural element flanking the seed, the one with lower probability score will be extended first. The adjacent structural element can be either a base paired region (*e.g.* a helix strand) or an unpaired region (*e.g.* a hairpin, internal, or multi-stem loop). When the program is able to identify one or multiple different subsequences with acceptable matches for the structural element under search, CRWAlign-2 carries all of the matches forward to create all possible structural model candidates. The growing seed structure models are used as the seeds for the next round. This extension of the structural model iterates either to a set of complete structural models, as defined by the identification of all of the structural elements in the descriptor. Or when the extension on both 5' and 3' ends of the seeds cannot be extended, and then the process is terminated, resulting in a partial structure model.

During the structural extension, when there is a helix enclose the seed with the one strand (5' strand) located adjacently to the 5' end of the seed but the other strand (3' strand) tens of nucleotides away from the 3' end of the seed, the program searches for matches of the entire helix (both strands), add the matches for 5' strand to the growing seeds, and put the matches for 3' strand to a temporary list. Thus, for a candidate structure model, it is possible that a putative structural element identified in say round 3 conflicts with structural elements identified in round 7. To ensure the consistence of each

structural element and terminate the extension of structural models with potential inconsistency, the program performs a consistency test at the end of each round of extension to evaluate the fitness of the constraints for each structural element with the identified structural model. The consistency test checks 1) the compatibility between the growing seed structural model and every potential match in the temporary list, 2) the overall weight score of the seed structural model. The candidate structure models that fail this test result in a partial solution.

Abnormal or aberrant insertions/deletions or nucleotide composition can occur in some sequences, which stops the continuous elongation. While the original RNAMotif would abort without reporting partial structure models, CRWAlign-2 allows users to permit exceptions to specific structural elements defined in the descriptor. The number of allowed exceptions represents the variance from the canonical or regular structure model. As the number of allowed exceptions increases, the program is able to identify more complete or partial structure models with lower specificity. Therefore, while this number should be minimal to avoid over-loosing structural constraints and causing long running time, it is extremely useful to have this option to permit the identification of structural models that are truly exceptions to the norm.

Since the descriptor of the most relevant phylogenetic node has more specific structural constraints for that branch, the specificity and resolving power for this identification process is greatest when the phylogenetic information for each sequence is known. In contrast, the generalized descriptor without the phylogenetic information identifies more sequences and requires more computational cost.

It is important to note that CRWAlign-2 not only identifies sequences that contain the structure model in the descriptor, but also creates a structure model for those sequences that have one. This feature thus allows a very large number of comparative structure models to be generated automatically. These comparative structure models can be used to in multiple applications including evaluate the accuracy of RNA folding algorithms[22,89], and identify structural motifs for different phylogenetic groups.

Stage 3: Aligning Sequences Based on Similar Primary and Secondary Structural Elements

As noted earlier, CRWAlign-2 is capable of aligning new sequences based on a common secondary structure. With the complete structural models determined in stage 2, the sequence will be aligned based on primary and secondary structural similarity with the template sequence alignment. When there are multiple complete structure models identified, the one with highest overall weight score is used to align the sequence. According to the relative order and boundary of structural elements identified in the structure model, the sequence to be aligned is split into multiple fragments. For each fragment that represents a specific structural element in the secondary structure and the descriptor, the alignment program identifies the previously aligned template sequence that is most similar to the sequence to be aligned, based on the length of the fragment and the sequence conservation. The alignment of the new sequence against the template will be performed to maximize the correct juxtaposition of the nucleotides in the new sequence to the analogous nucleotides in the template sequence. When aligning the pairing regions (*e.g.* helices), the program first attempts to juxtapose nucleotides within the same length, regardless of sequence conservation. In contrast, sequence conservation in the unpaired regions is the primary factor in the juxtaposition of sequences.

2. Chimera-checking Procedures

The creation of the reference sequence alignment and aligning query sequences

Template-based sequence alignment programs utilize a reference alignment that is usually manually curated to optimize its accurate juxtaposition of nucleotides regarding the similarity in nucleotide sequence, higher-order structure, and evolutionary relationships. Thus the creation and maintenance of a most accurate reference sequence alignment is essential for performing the subsequent chimera-checking procedures. The reference sequence alignment used in this study is manual checked to assure it is chimera free and reliable. This process requires significant amount of manual effort. The relevant

metadata of each sequence within the reference alignment, including the taxonomy information, are stored and cross-indexed in our RNA Comparative Analysis Database (rCAD) system. As shown in Figure 3.4, the query sequences are aligned onto the reference sequence alignment utilizing CRWAlign-2.

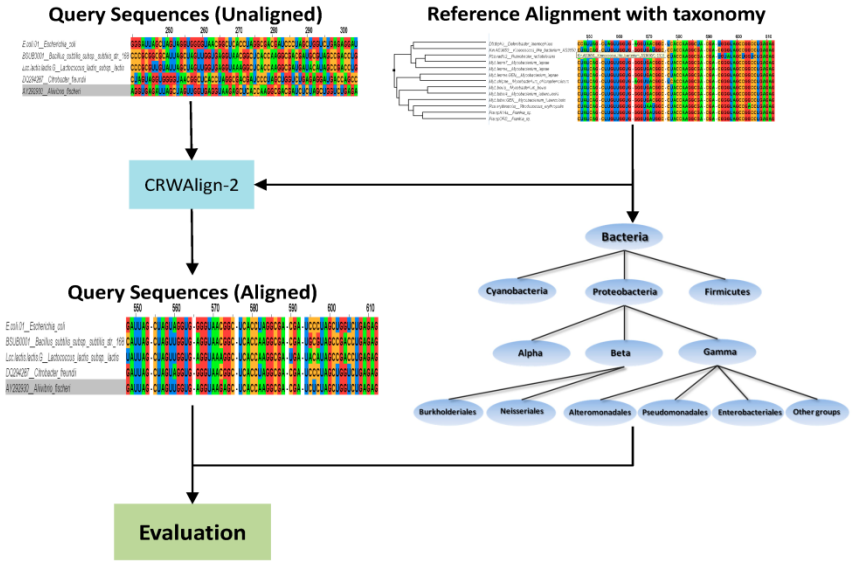


Figure 3.4: The alignment of query sequences using CRWAlign2 and generation of the phylogenetic tree contains all valid taxons with sufficient amount of aligned sequences in the reference sequence alignment.

Evaluation of query sequences

The taxonomy information of each sequence in the reference sequence alignment is obtained from NCBI (ftp://ftp.ncbi.nih.gov/pub/taxonomy/). Based on these phylogenetic relationships, all reference sequences are mapped onto certain branch of the taxonomy tree. Given a taxon (a node of the phylogenetic tree), if it contains a minimal amount of reference sequences (at least 10 sequences), the taxon will be marked as a valid node which will be used in the following analysis. As shown in Figure 3.4, with a reference alignment that consists of 1750 bacteria 16S rRNA sequences, all valid taxons

are displayed as a partial phylogenetic tree which means all these nodes contain at least 10 sequences.

The program traverses through the taxonomy tree from root to leaf nodes, and determines if the query sequence is purebred or chimeric. As shown in Figure 3.4, the root node Bacteria has three valid child nodes with sufficient amount of sequences aligned in the reference alignment: Cyanobacteria, Proteobacteria, and Firmicutes. For each child node, the signature nucleotide frequency of each position is calculated. Given a position within the alignment, a weighted nucleotide frequency (F_w) is calculated as

$$F_w = \sum_{N\in\{A,C,G,U,-\}} (\frac{P_N}{S})^2 \tag{3.1}$$

where P_N is the number of sequences having a specific nucleotide N (N in {A, C, G, G and deletion "-"}) at this position, and S is the number of sequences having nucleotide at the position. The consensus score (C_{cons}) of the position is calculated as

$$C_{Cons} = \sum_{N\in\{A,C,G,U,-\}} \frac{P_N * ABS\left(\frac{P_N}{S} - F_w\right)}{S} \tag{3.2}$$

A higher F_w value indicates more conservative the position is. For example, given a position (*E. coli* nucleotide number 80 at TaxID 1224, Proteobacteria) has {A: 990, C: 252, G:201, U:225, deletion: 50}, the Fw at this position is calculated as $(990/1718)^2+(252/1718)^2+(201/1718)^2+(225/1718)^2+(50/1718)^2=0.385$, and its C_{Cons} is 0.22. For another position (E. coli nucleotide number 39 at Proteobacteria node) with {A: 0, C:1718, G:0, U:0, deletion: 0}, its *Fw* is $(1718/1718)^2=1$, and the C_{cons} equals 0.

For the position i in a given sequence, the column difference score (CDS_i) is calculated as

$$CDS_i = ABS(C_{Cons} - ABS(F_i - F_w)) \tag{3.3}$$

where *Fi* is the corresponding nucleotide frequency of column i in the alignment. With the reference alignment mentioned above, given a query sequence (Accession Number

AB015574) that has C at position 39 and A at position 80, its CDS_{39} is 0 while CDS_{80} is 0.029. This example indicates that CDS inversely correlate with the similarity between the query sequence and the reference alignment: larger *CDS* value indicates more difference. The entire sequence difference score (*SDS*) is calculated as

$$SDS = \sum_{i=1}^{N} \frac{CDS_i^2}{N} \tag{3.4}$$

where N is the total number of positions under check within the test sequence. Larger SDS value indicates more difference between the entire query sequence and the reference alignment. The query sequence is considered to be homologous when its SDS is lower than any aligned sequence in the reference alignment.

Results

The CRWAlign-2 has been evaluated in 1) the accuracies of the alignment results, 2) the running time of the program executions, and 3) the scalability for large data sets. The results have been compared to eight existing widely-used automatic alignment programs.

1. Alignment programs compared

Eight alignment programs are included in the comparison with CRWAlign-2: ssu-align, infernal, HMMER, RDP, Silva, GreenGenes, MAFFT and SATe. Six of them are implemented with template-based alignment algorithm: ssu-align, infernal, HMMER, Silva, RDP, and GreenGenes, while the other two (MAFFT and SATe) are *de novo* alignment programs. Among the six template-based alignment programs, ssu-align, Infernal and HMMER are stand-alone and available for download (http://selab.janelia.org/software.html). The other three (Silva, RDP, and Greengenes) do not provide download, thus are only available as web-servers. In this analysis, all programs and web-servers run with the default parameters.

Regarding the type of RNAs each program are able to align, CRWAlign-2, ssu-align, Infernal, HMMER, MAFFT, and SATe are capable of aligning any type of RNA sequences, while Silva, RDP, and Greengenes are able to only align 16S rRNA. To perform a robust performance comparison, 16S rRNA is used in this study, since it is the only RNA that can be aligned by all nine programs. Table 3.1 shows the details information about sequences in the test and template data sets. Both test and template data sets are random subsets of a large bacterial 16S rRNA alignment available at the CRW site. There is no overlap between a test and template set (i.e. none of the sequences in the test set are present in the template set. In the measurement of the template-size effect, small template alignments are always subsets of any larger template alignment (e.g the 500 16S rRNA template alignment was a subset of the 2000 16S rRNA template alignment).

Table 3.1: Sequences in template alignments and used for testing. No overlap occurs between sequences tested and sequences in template alignments.

RNA Molecule	**Template Sequences**			**Unaligned Test Sequences**		
	Count	Avg Length	# of Taxonomic Leafs	Count	Avg. Length	# of Taxonomic Leafs
16S Bacterial rRNA	250	1447.3	188	500	1446.1	320
	500	1447.4	324			
	1000	1449.4	593	1000	1448.1	598
	2000	1449.2	1154			

2. Evaluating the Accuracy of an Alignment

The accuracies of the sequence alignments generated for this analysis are evaluated through pairwise sequence comparisons with the correct alignment. Given a pair of sequences *i* and *j*, the pairwise sequence identity for sequences *i* and *j* is defined as

$$PSI_{ij} = \frac{|\mathbf{B}|}{|\mathbf{E}|} \tag{3.1}$$

where B is the set of columns that contain nucleotides from both sequences *i* and *j*, and E is the set of columns that contain nucleotides from either sequence *i* or *j*. The pairwise sequence accuracy is defined as

$$\text{Accuracy} = \frac{|\mathbf{S}|}{|\mathbf{E}|} \tag{3.2}$$

where *S* is the set of columns in the test alignment that have an identical stack relative to the correct alignment. For example, if nucleotide 45 (G) of sequence *i* is stacked with nucleotide 53 (C) of sequence *j* in the correct alignment, then the test alignment must have nucleotide 45 stacked with nucleotide 53 and not with a C nucleotide at any position in sequence *j* other than nucleotide 53. If a nucleotide from either sequence is stacked with a gap, the test alignment must have the nucleotide stacked with a gap.

3. Accuracy Comparison with Other Methods

A test set consisting of 1000 bacterial 16S rRNA sequences is aligned by each alignment program. The accuracies of the generated alignments are calculated based upon the pairwise sequence identity ranges (Fig. 3.3). Each of the four programs (CRWAlign-1, CRWAlign-2, HMMER and Infernal) that accept template alignments are given three template alignments with different size (250, 500, and 2000 sequences), and the best results of each are presented in Fig. 3.4.

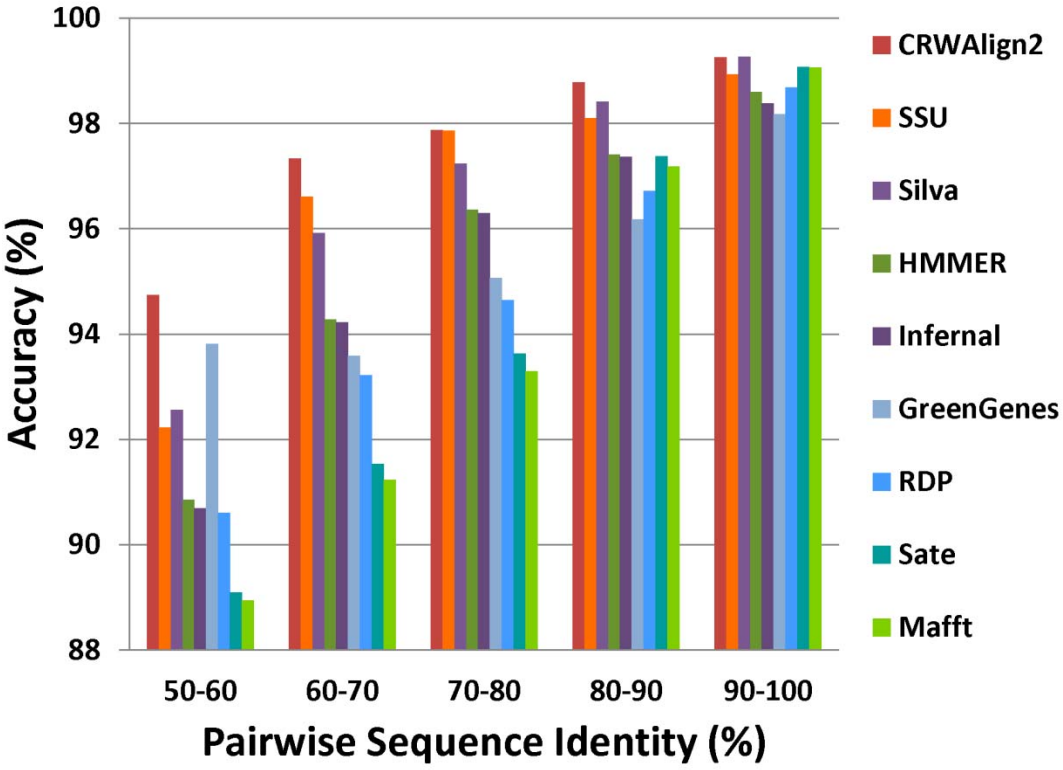


Figure 3.5: The pairwise sequence accuracies for alignments generated with CRWAlign-1, CRWAlign-2, and eight other alignment programs. Accuracies were evaluated for sequences with five pairwise sequence identities, 50-60%, 60-70%, 70-80%, 80-90%, and 90-100%. Alignments contain 1,000 bacterial 16S rRNA sequences.

All programs have very high accuracies (>98%) in the 90-100% pairwise identity range. CRWAlign-2 (dark red bar in Figure 3.5) and Silva outperform the other programs with ~99.25% accuracy, which is 0.2% (for SATe, Mafft) to 1.0% (for GreenGenes) higher than the other seven programs. In the 80-90% identity range, CRWAlign-2 achieves 98.8% accuracy, which is superior to other eight programs, including Silva, by 0.5% (for Silva) to 2.7% (for GreenGenes). In the 70-80% sequence identity range, CRWAlign-2 and ssu-align are the top two programs with 97.9% accuracy, which lead other programs by 1.8% (Silva) to 5.7% (Mafft). In the 60-70% and 50-60% identity ranges, CRWAlign-2 beats other programs again by at least 0.7% accuracy. While the de novo alignment programs (Mafft, SATe) are able to obtain similar accuracy as the

template-based program in the high sequence identity ranges (80-90% and 90-100%), their accuracies are remarkably lower than the template-base programs for sequences with lower pairwise identity.

4. Effect of Template Size on Accuracy

CRWAlign-2, HMMER and Infernal are able to accept user-defined template alignment with different sizes. To gauge the influence of the template size on the accuracy for these three programs, each program is analyzed to align a test set consisting of 1000 bacterial 16S rRNA sequences with three different template alignments containing 250, 500, and 2000 bacterial 16S rRNA sequences. As shown in Figure 3.6, all three programs achieve nearly identical accuracies with all three template alignments, while CRWAlign-2 outperforms HMMER and Infernal in every case.

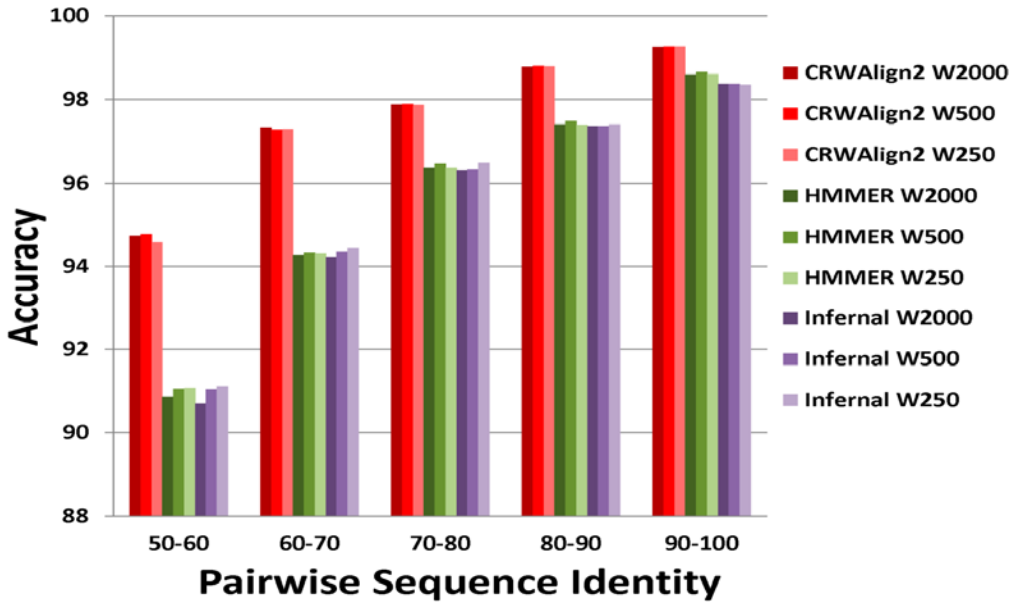


Figure 3.6: The pairwise sequence accuracies for alignments generated with CRWAlign-1, CRWAlign-2, HMMER, and Infernal were determined. The alignments contain 1,000 bacterial 16S rRNA sequences. Three different template sizes (250, 500, and 2,000 sequences) were evaluated for five pairwise sequence identities, 50-60%, 60-70%, 70-80%, 80-90%, and 90-100%.

5. Comparison of the Run Time and Scalability

The execution time of aligning 1000 bacterial 16S rRNA sequences is determined for CRWAlign-2, and the three stand-alone programs, HMMER, Infernal and ssu-align. To determine the effect of template size on the execution time, the three programs (CRWAlign-2, HMMER, and Infernal) that are able to accept user-defined template alignment are check with three data points, while there is only one data point for ssu-align since it integrated the profile (default template) into the program. Due to platform requirements and software dependencies, CRWAlign-1 and CRWAlign-2 were tested on Windows Server 2008 R2 Enterprise (64 bit) with an Intel Xeon x7550 @ 2GHz. HMMER and Infernal were run on a Linux platform (Ubuntu 11.10, 32 bit) with an Intel Core i7 920 @2.67GHz. The ssu-align program was run on Solaris 10.0 with an Intel Xeon processor 5400. These three server configurations have very comparable speeds. Figure 3.7A shows HMMER and Infernal run faster than CRWAlign-2 and ssu-align with a tradeoff in lower accuracies (as shown in Figure 3.6). CRWAlign-2 creates the complete secondary structure models for each sequence to be aligned. The identification of structural models is an iterative process and requires significant amount of computational time, which is still faster than ssu-align. While the comparative structural models generated with CRWAlign-2 are essential for the alignment of sequences, these structure models can be used for other applications, *e.g.* to improve and evaluate RNA folding algorithms [89].

As mentioned previously, the operating process of the CRWAlign-2 program consists of three phases: 1) the generation of structural descriptor, 2) the identification of complete structural models for each sequence, and 3) aligning sequences. Thus the total running time of CRWAlign-2 is expected to be sensitive to the template size as well as the number of sequences to be aligned.

The complete execution time of CRWAlign-2 for aligning two test sets (500 and 1,000 16S rRNA sequences) using 3 different template alignment (250, 500, and 2,000 sequences) is determined (Figure 3.7B). The computational time of generating the structural descriptor increases linearly with the number of sequences in the template

alignment, while the execution time of two latter stages (identifying structural models and aligning sequences) is linear to the number of sequences to be aligned. In addition, larger template alignment requires more computational cost in the stage of descriptor generation, but helps to speed up the identification of structural models which is the most time-consuming step.

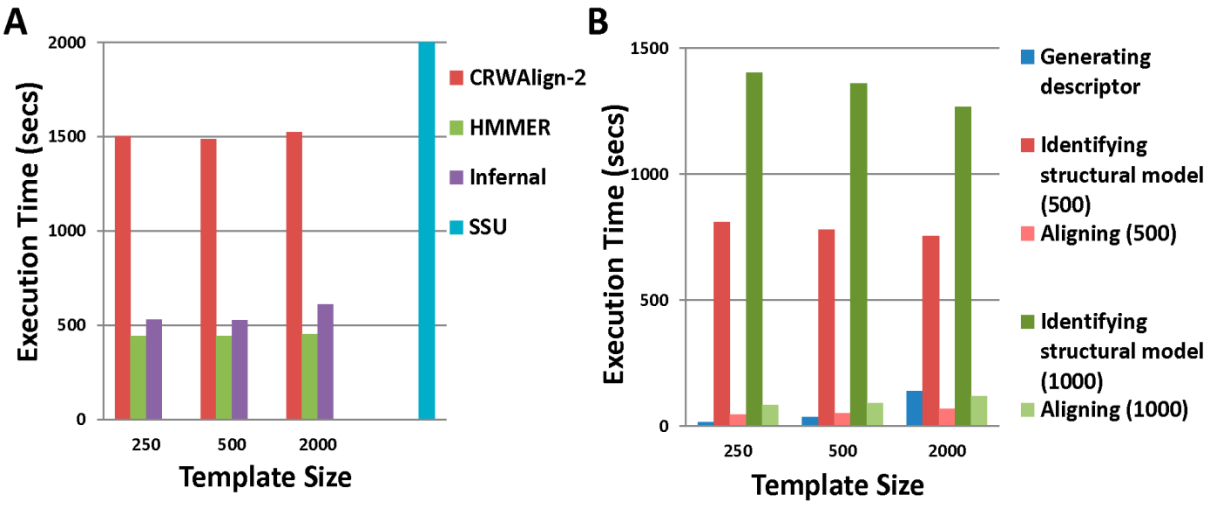


Figure 3.7: A) The total execution time of aligning 1,000 bacterial 16S rRNA sequences for four alignment programs with three different template sizes (250, 500, and 2,000 sequences). B) The execution time of the different phases for CRWAlign-2 programs in aligning two test sets (500 and 1,000 bacterial 16S rRNA sequences) with three different template sizes (250, 500 and 2000 sequences).

6. Identification of chimeric sequences

As shown in Figure 3.8, the query sequence (Accession No. AB015574) is evaluated at three taxonomy nodes: Cyanobacteria, Proteobacteria, and Firmicutes. As shown in Figure 3.8, the signature nucleotide frequency at each node is computed. At each node, the CDS of every position within the test sequence are calculated with equation 3.3 and plotted: peaks indicate significant difference while valleys represent high similarity. The plots show that the query sequence is quite different from Cyanobacteria and Firmicutes branches at most positions since high peaks of CDS are observed (Figure 3.8A, Figure 3.8C), while it is very similar to Proteobacteria with very few exceptions (Figure 3.8B). The SDS of the query sequence at Cyanobacteria, Proteobacteria, and Firmicutes nodes are 0.14, 0.007, and 0.87 respectively, which further confirmed that the query sequence is a member of Proteobacteria (Figure 3.8 highlight in red).

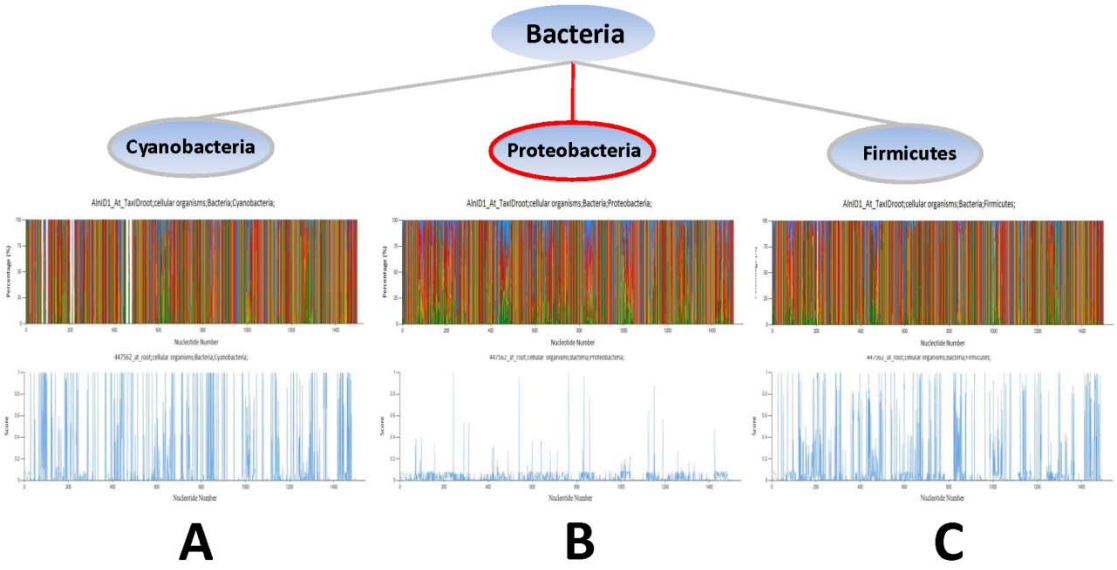


Figure 3.8: Scenarios at three child nodes of Bacteria: Cyanobacteria (A), Proteobacteria (B), Firmicutes (C).

The next step starts at the three child nodes of Proteobacteria: Alphaproteobacteria, Betaproteobacteria, and Gammaproteobacteria (Figure 3.9). Similar

computational process is fulfilled at each of these three nodes. The SDS values of the query sequence are 0.093 at Alphaproteobacteria node, 0.063 at Betaproteobacteria, and 0.008 at Gammaproteobacteria. The result indicates the test sequence is purebred Gammaproteobacteria (Figure 3.9 highlight in red).

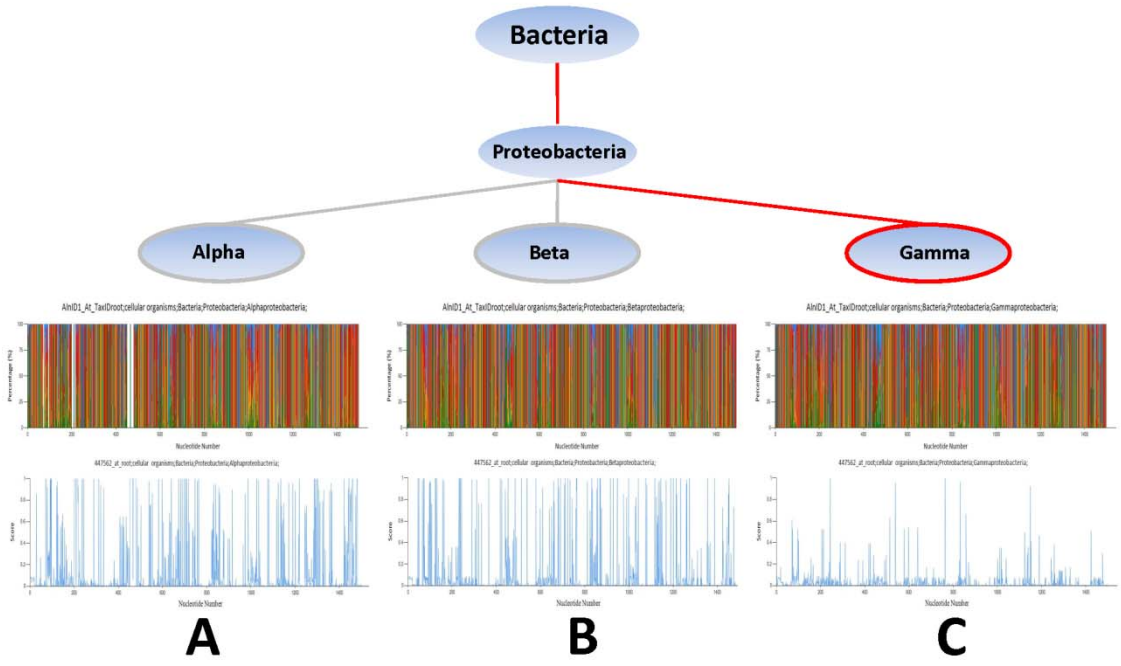


Figure 3.9: Scenarios at three child nodes of Proteobacteria: Alphaproteobacteria (A), Betaproteobacteria (B), Gammaproteobacteria (C).

The four child nodes of Gammaproteobacteria are under investigation in the next step (Figure 3.10). The SDS values of the query sequence at these four nodes are: 0.019 at Alteromonadales, 0.037 at Pseudomonadales, 0.066 at Enterobacteriales, and 0.134 at other groups. None of these SDS values satisfy the threshold to be considered as purebred at any of these four phylogenetic nodes. However, the plot reveals that the 500 nucleotides at the 5' end is closely related to Pseudomonadales (Figure 3.10B), while the rest 1000 nucleotides at the 3' end is very similar to Alteromonadales (Figure 3.10A). The SDS value by sliding window (100 nucleotide window size) confirms this identification. Therefore, the test sequence is identified as a chimeric sequence:

Pseudomonadales at nucleotide 1-500, and Alteromonadales at nucleotide 501-1542 (Figure 3.9 highlight in dark blue).

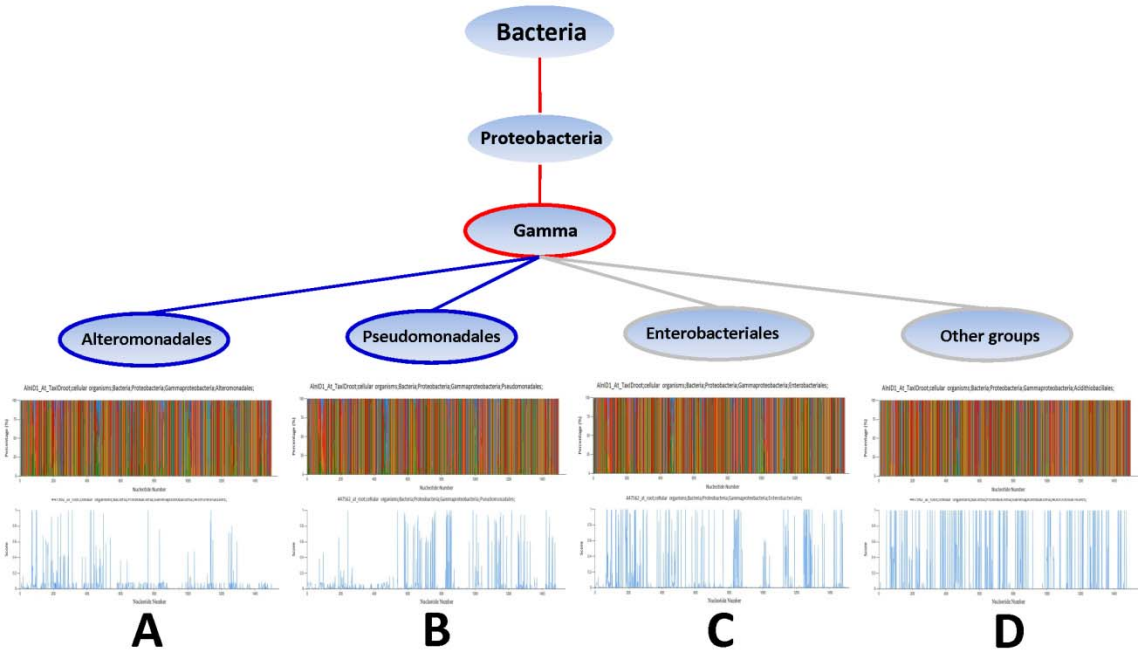


Figure 3.10: Scenarios at four child nodes of Gammaproteobacteria: Alteromonadales (A), Pseudomonadales (B), Enterobacteriales (C), other groups (D).

The chimera-checking program is also tested with a set of query 16S rRNA sequences consisting of 100 artificial chimeric sequences. These artificial chimeric sequences are created by breaking reference 16S rRNA sequences with known taxonomy information, and switching the fragments between different sequences. The program detects 99 out of 100 artificial chimeric sequences with correct taxonomy groups identified. This result suggests that my chimera-checking program is accurate and sensitive in the identification of chimeric 16S rRNA sequences.

The same underlying principle used in chimera-check program is also utilized to implement an alignment evaluating program.

Discussion

Manual curation has been used to create the most accurate RNA sequence alignment available at the CRW Site [23]. While this process maximizes the optimal juxtaposition of similar structural and facilitates the generation of highly accurate large sequence alignments, it requires a significant amount of time and manual effort. With the development of next gen sequencing technology that generates exceedingly large amounts of sequencing data, the traditional manual curation process is not feasible. Thus it has been essential and challenging to develop alignment programs that will create highly accurate sequence alignment quickly.

I have developed a template-based alignment program – CRWAlign-2 that utilizes sequence composition, secondary structural information and phylogenetic information to align sequences based on primary and secondary structural similarity. CRWAlign-2 retrieves the required information from rCAD, and uses it to create the structural descriptor which helps the identification of structural models and the alignment of new sequences.

The accuracy of CRWAlign-2 is tested on a set of 16S bacterial rRNA sequences and compared with various template-based and de novo sequence alignment programs (Figure 3.4). The result reveals that CRWAlign-2 significantly outperforms eight analogous alignment programs in accuracy. When the sequence identity range is 90-100%, the competing programs have similar accuracy to CRWAlign-2. However, for lower sequence identity, the other programs are considerably less accurate than CRWAlign-2. Even for the sequence identity range of 50-60%, CRWAlign-2 is able to align sequences with ~95% accuracy, which is significant higher than other existing alignment programs. The computation cost of CRWAlign-2 scale linearly with the number of sequences 1) to be aligned, and 2) in the template alignment (Figure 3.6).

In addition to aligning sequence accurately, CRWAlign-2 creates secondary structural models for each sequence to be aligned. These secondary structure information is very valuable for the program development of RNA secondary structure prediction

[89,120-123], and for the determination of structural statistics (http://www.rna.ccbb.-utexas.edu/SAE/2D/index.php). The structure models created by CRWAlign-2 can be easily converted into various formats of RNA secondary structure file including bpseq, alden, rnaml, ct, and bracket (http://www.rna.icmb.utexas.edu/DAT/3C/SBPI/, [23]), which further increases their utility. Currently the Gutell lab's Comparative RNA Web (CRW) Site has nearly 55,000 structure files in these multiple formats. The CRWAlign-2 system has the potential to increase the number of comparative structure model files to more than 1,000,000.

The highly-accurate sequence alignment generated by CRWAlign-2 also brings opportunities for other research fields including the identification of chimeric 16S rRNA sequences generated in microbiome research projects. I have developed a chimera-checking program utilizing a well-aligned reference sequence alignment and taxonomy information. The preliminary results suggest that, with a high-quality chimera-free reference sequence alignment, our strategy is sensitive and accurate in the identification of 16S rRNA chimeric sequences.

The deluge of nucleic acid sequences that are determined with next-generation sequencing technology increases the scale of sequencing data faster than Moore's law. Given that multiple-dimensions of information are available in systems like rCAD [109], I have developed the automated alignment system – CRWAlign-2 to address this challenge and need.

Chapter 4: Evaluation of the HIV secondary structure model

Abstract

Human immunodeficiency virus (HIV) that causes acquired immunodeficiency syndrome (AIDS), has become one of the world's most serious health and development challenges. The secondary structure of HIV RNA genome plays central role in the replication and metabolism. In 2009, a secondary structure model of an entire HIV RNA genome was proposed using high-throughput selective 2' OH acylation analyzed by primer extension (SHAPE) technology. This working model is useful to help elucidate the three dimensional structures of the small fragments in the HIV RNA genome and aid drug development against HIV. However, due to the limitation of SHAPE technology and thermodynamic-based algorithms, a large percentage of the predicted base pairs in the SHAPE-directed HIV secondary structural model could have low level of confidence. Utilizing comparative analysis methods, the proposed SHAPE-directed HIV secondary structure model is evaluated with multiple covariation metrics. Only a small portion of the predicted base pairs in the SHAPE-directed model are verified with covariation analysis. The overall results suggest that about 46.7% of the predicted base pairs in this model have very low confidence level, which require intensive improvement and correction. There are 52.4% of the predicted base pairs highly conserved which require additional information to validate. In addition to evaluating the predicted base pairs in the SHAPE-directed model, the comparative analysis also predicts 71 potential helices that are not present in the SHAPE-directed model but have strong support from comparative analysis.

Background

Human immunodeficiency virus (HIV) is a member of *Retroviridae* family that causes acquired immunodeficiency syndrome (AIDS), a disease of human immune system resulting in progressive immune failure and allow numerous life-threatening infections to thrive. HIV has infected more than 30 million people worldwide up to date. The mechanistic and therapeutic insights of HIV have been under intense research for more than 25 years.

As the predominant type of HIV, the human immunodeficiency virus type 1 (HIV-1) is more virulent than its less widespread cousin HIV-2, thus is the cause of the majority of HIV infections globally [124]. The genome of HIV-1 is composed of a ~9kb RNA which contains nine open reading frames that encodes fifteen proteins [125].

For all positive-strand RNA virus, the secondary structures of the viral RNA genome play critical roles in the viral replication cycle, while HIV-1 is no exception. It has been discovered that a variety of discrete steps in HIV-1 replication cycle, including RNA transcription, dimerization of the RNA genome, and incorporation of RNA genome into virion are under regulation by the integrity of some secondary structural motifs of the viral RNA genome [126]. Previous research has identified several secondary structural motifs that are critical for viral replication: the trans-activation region (TAR) which is the Tat–binding site [127,128], the primer binding site which is important to initiate reverse transcription, the packaging signal that binds NC and is critical for incorporation of genomic RNA into the virion [129,130], the dimerization site (DIS) with a "kissing loop" hairpin [131,132], the Rev response element (REV) [133,134], and the major splice donor site which is used to generate all sub-genomic spliced mRNAs [135,136]. While the complete significance has not been fully understood, more evidence discovers that the HIV-1 genome forms extensive secondary structures whose functions are associated with different process at stages of HIV-1 viral replication cycle.

The advancement of RNA three-dimension structure determination has significantly increased the number of known RNA structures in the PDB

(http://www.rcsb.org/pdb/home-/home.do). The most widely-used experimental techniques for structure determination of biological macromolecules are X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR). While the 3D structures of multiple small fragments at 5' UTR of the HIV-1 RNA genome have been determined [137-142], the traditional techniques of 3D structure determination are not well suited to elucidate the structure of the entire HIV-1 RNA genome due to the complexity and flexibility of large RNA molecules [143]. Thus over 90% of the HIV-1 genome has not been structurally characterized.

In 2009, Watts *et al.* proposed a secondary structure model of an entire HIV-1 genomic RNA with the high-throughput selective 2' OH acylation analyzed by primer extension (SHAPE) technology [144]. This method measured SHAPE reactivity of the 9173 nucleotides in the NL4-3 HIV-1 genomic RNA sequence, converted the SHAPE reactivity value to free-energy change terms, and built a thermodynamically favored structural model. This proposed SHAPE-directed secondary structural model indicated that the genome of HIV-1 contains higher-order structural elements throughout the entire sequence. However, there are several concerning facts about the proposed HIV-1 secondary structure model. The structures of large RNAs, like HIV-1 RNA genome, are too large and complex to be predicted with sufficient confidence from first principles or thermodynamic-based algorithms alone with a single sequence. The SHAPE-directed HIV-1 secondary structure model is built based on analyzing only one HIV-1 complete genome sequence (Accession: AF324493). Recent benchmark of SHAPE technology on six small RNA molecules (tRNA-phe, 5S rRNA, the P4−P6 domain of the *Tetrahymena* group I ribozyme, and ligand-bound domains from riboswitches for adenine, cyclic di-GMP, and glycine) revealed that SHAPE-directed modeling structures of these small RNAs gave 17% false negative rate and 21% false positive rate [145]. Their bootstrapping calculation suggested that the overall accuracy of the SHAPE-directed HIV-1 secondary structure model was lower than 50%. Moreover, the accuracy of the thermodynamics-based secondary structure model decrease as the length of the input RNA sequence

increase [22]. Thus the proposed SHAPE-directed secondary structural model of an entire HIV-1 RNA genome has ambiguities which require further validation and correction.

As emphasized in previous chapters, the comparative analysis has been utilized successfully to decipher the secondary structures of many RNAs including tRNAs and rRNAs. The accuracies of the 16S and 23S rRNAs secondary structure models predicted with comparative analysis are over 97%. In this research, the predicted HIV-1 RNA secondary structure model is evaluated and improved with comparative analysis. A pre-made multiple sequence alignment consisting of over 2000 HIV-1 complete genome sequences is used for the subsequent comparative analysis. All predicted base pairs in the SHAPE-directed HIV-1 secondary structure model are evaluated with multiple covariation metrics. With the complementation of helix-extension strategy, the predicted base pairs in the proposed HIV-1 secondary structure model are categorized with various confidence levels. Moreover, this analysis also identifies many potential helices that are not present in the proposed SHAPE-directed HIV-1 secondary structure model but with strong support of comparative analysis.

Methods

1. Calculation of characteristic covariation metrics (Conservation score, purity score, and confidence score)

The conservation score measures the extent of how conserved a position of RNA molecule is throughout the evolutionary history. The possible value of conservation score varies between -1 and 2, while higher value indicates stronger conservation. Given a pair of columns *i* and *j* in a multiple sequence alignment, the conservation scores are calculated with equation 2.3.

Purity score provides a quick measurement of covariation between two positions. Given a pair of column I and j in a multiple sequence alignment, the purity score is determined with the procedure described in Figure 2.2. However, highly conserved positions could have very good purity score, for example a pair of columns having {G:C

60%, A:U 30%, C:G 7%, U:A 3%} could be assigned with the identical purity score as the pair of columns with {G:C 100%}. To distinguish highly conserved pairs from variable pairs, a new metric – confidence sore is introduced which integrates both conservation level and purity extent for a pair of positions.

Given a pair of columns *i* and *j* in the sequence alignment, the variation score (V_{ij}) is defined as

$$V_{ij} = \frac{Max - Csa}{3} \tag{4.1}$$

where *Max* is the maximum value of the conservation score (which is 2 by definition), and *Csa* is the average of the conservation scores for column *i* and *j*. The confidence score (C_{ij}) is calculated as

$$C_{ij} = P_{ij} + V_{ij}^2 \tag{4.2}$$

where *Pij* is the purity score of column I and j and *Vij* is the variation score calculated with equation 4.1.

Given the HIV-1 sequence alignment and the SHAPE-directed secondary structure model, variation/covariation analysis calculates the total number of variation in each pairwise set of sequences (sequence *i*, *j*) versus the amount of variation for sequence *i* and *j* at 1) the positions that form a predicted base pair and undergo a covary, 2) the positions that form a predicted base pair but only one position change, and 3) the positions in the unpaired region (do not form any base pair).

2. Measurement of Covariation with Mutual Information Based Method

The standard mutual information score (*MIxy*) between column *i* and column *j* within the sequence alignment is calculated with equation 2.1. The higher MI value explicitly indicates greater statistical dependence between the two positions. The MIxy values for each pair of positions within the HIV-1 alignment are calculated -- every position is compared against every other positions. The overall complexity of this

calculation is N*(N-1)/2 where N is the total number of columns within the alignment. The HIV-1 alignment used in this analysis consists of 25132 columns, thus the overall complexity is 315,796,146 (25132*25131/2).

The corrected mutual information method (*MIp*) is an variant of standard *MIxy* [57]. Given column *i* and *j* within the alignment, the background mutual information value or Average Product Correction (*APC*) is determined with

$$APC(i,j) = \frac{MI(i,\overline{x}) * MI(j,\overline{x})}{\overline{MI}} \tag{4.3}$$

where $MI(i,\overline{x})$ is the average *MIxy* value for position *i* with every other positions in the alignment, $MI(j,\overline{x})$ is the average MIxy value for position *j* with every other positions in the alignment, and $\overline{MI}$ is the average *MIxy* value for all positions within the alignment. The corrected mutual information score (*MIp*) is calculated as

$$MI_p(i,j) = MI(i,j) - APC(i,j) \tag{4.4}$$

where $MI(i,j)$ is the standard *MIxy* score for column *i* and *j*.

3. Helix Extension

Due to relatively short evolutionary history, many positions (columns) within the HIV-1 alignment used in this study are highly conserved. Classical covariation analysis methods (e.g. *MIxy*, *MIp*) cannot identify these highly conserved base pairs due to lack of variation. The helix-extension strategy has been proved very sensitive and accurate in identifying the highly conserved base pairs and extending the helix in the bacterial 16S rRNA secondary structure (details in Chapter 2). Thus the helix extension strategy is used with the HIV-1 sequence alignment.

The nucleation pairs are selected, and the corresponding columns of these nucleation pairs within the HIV-1 alignment are determined. For each nucleation pair, the adjacent and antiparallel columns with a percentage of canonical pairs (Watson-Crick {G:C or A:U} or Wobble pair {G:U}) higher than a predefined threshold (85%) are

considered as highly conserved base pairs and added to the extending helix. The extension process is terminated when the adjacent and antiparallel columns fail the extending threshold.

Results

1. Evaluation of the Proposed SHAPE-Directed Secondary Structure Model of an entire HIV-1 RNA genome

The proposed SHAPE-directed HIV-1 secondary structure model is evaluated with comparative analysis. The HIV-1 sequence alignment used in this analysis is obtained from HIV database maintained by Los Alamos national laboratory (http://www.hiv.lanl.gov). The alignment consists of 2025 non-redundant HIV-1 whole genome sequences and 25132 columns. The sequence "B.FR.1985.NL43 pNL43-NL4 3" was selected as the reference since it shares >99.95% similarity with the sequence used by Weeks's group (Accession: AF324493). All predicted base pairs in the SHAPE-directed HIV-1 secondary structure model are evaluated with multiple characteristic covariation metrics.

1.1 Percentage of Canonical type of Predicted Base Pairs

While comparative analysis searches for all positional dependence regardless of pair type, the standard Watson-Crick base pairs (G:C and U:A) and G:U wobble base pairs are the predominant pair type identified in the comparative models of tRNAs, ribosomal RNAs, and other non-coding RNAs. As shown in Table 4.1, among the 454 predicted base pairs in the comparative secondary structure of bacteria 16S rRNA, 423 pairs (or 423/454 = 93.17%) have 85% or higher canonical base pair percentage. Similarity results are obtained on other non-coding RNAs including tRNAs, 5S rRNA, and 23S rRNA, which substantiate that the vast majority of base pairs in the RNA secondary structures are canonical pair types The SHAPE-directed HIV-1 secondary

structural model includes 1891 predicted base pairs and all of them are canonical pair type {CG, UA, GU} on the single HIV-1 complete genome sequence (Accession: AF324493). Using the HIV-1 sequence alignment consisting of over 2000 sequences, the percentages of canonical pairs for all these 1891 predicted base pairs are calculated (Table 4.1).

Table 4.1. The percentage of canonical base pairs of proposed base pairs in the SHAPE-directed secondary structure model of the entire HIV-1 genome RNA

	HIV			16S rRNA Comparative Structure		
WC/WB Percentage	# of base pairs	Sum	Sum Pct	# of base pairs	Sum	Sum Pct
<10%	21	21	1.11%	14	14	3.08%
10%~20%	12	33	1.75%	2	16	3.52%
20%~30%	22	55	2.91%	1	17	3.74%
30%~40%	59	114	6.03%	2	19	4.19%
40%~50%	57	171	9.04%	3	22	4.85%
50%~60%	50	221	11.69%	2	24	5.29%
60%~70%	81	302	15.97%	2	26	5.73%
70%~80%	91	393	20.78%	4	30	6.61%
80%~85%	76	469	24.80%	1	31	6.83%
85%~90%	99	568	30.04%	7	38	8.37%
90%~95%	223	791	41.83%	16	54	11.89%
>95%	1100	1891	100.00%	400	454	100.00%
Total	1891			454		

There are 469 predicted base pairs of the HIV-1 model (or 469/1891 = 24.8%) have 85% or lower canonical base pair composition in the alignment (Table 4.1), which suggests that non-canonical pair types occur much more frequently at these pairwise positions. Although non-canonical base pairs were observed in 16S and 23S rRNA crystal structures, generally they only take a very small portion of all base pairs (e.g. 6.83% in the 16S rRNA as shown in the right side of Table 1). Therefore, these 24.8% (or 469 out of 1891) of the predicted base pairs in the SHAPE-directed HIV-1 secondary structure model are under suspicion.

1.2 Characteristic Covariation Metrics

As shown in Methods section of Chapter 2, in the sequence alignment of an RNA molecule, the positions with similar conservation values are more likely to have a higher MIxy score (Figure 2.3). Most of the base pairs in the 16S secondary structures have similar conservations values for the two positions that form the base pair, and thus are close to the diagonal of the plot in Figure 4.1A. The conservation values of the positions that form the putative base pairs in the HIV-1 secondary structure model are calculated and plotted in Figure 4.1B. The plot shows there is no significant correlation of conservation values between the two paired positions for most of the predicted base pairs in the HIV-1 secondary structure model.

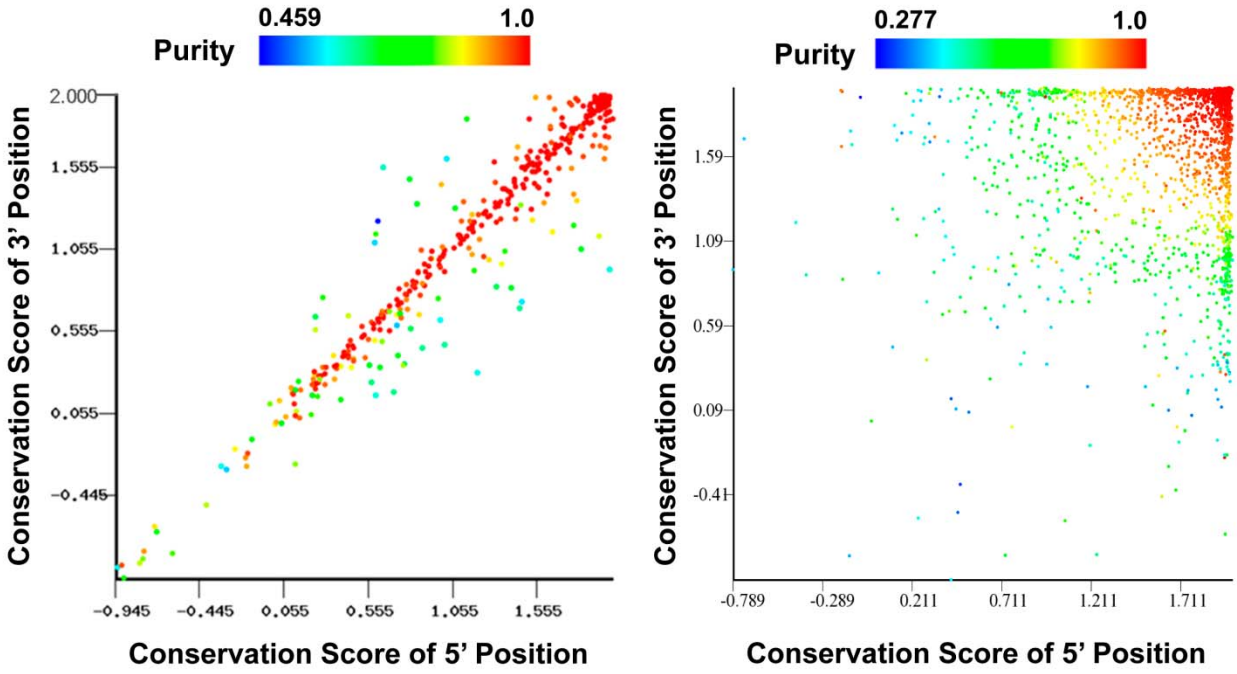


Figure 4.1: The plot of the conservation values for the two paired positions in bacterial 16S rRNA (A) and the proposed HIV-1 secondary structure model (B). The color of each data point represents its purity score with a scale shown on top.

The purity score measures the precision of covariation for a pair of positions (details in the Methods section of Chapter 2). Higher purity score indicates that the two

positions have higher probability to covary with one another. For every pair of positions that form a base pair in the bacterial 16S rRNA comparative structure, bacterial 16S rRNA crystal structure, and the HIV-1 secondary structure model, the standard purity score and the GU-Plus purity score is calculated and plotted against the average conservation score (Figure 4.2). Though the base pairs in the bacterial 16S rRNA comparative model (Figure 4.2A) and crystal structure (Figure 4.2B) range from highly conservative to highly variable, the vast majority of them have high purity score close to 1, which indicates the base pairs associated with these data points have strong covariation.

As shown in Figure 4.2C, most of the predicted base pairs in the HIV-1 secondary structure model are highly conserved (1175 out of 1891 pairs have average conservation score of 1 or lower). When the predicted HIV-1 base pair are highly conserved, they tend to have good purity score since the algorithm of purity score calculation will assign identical purity scores for pair 1 with {G:C 60%, U:A:20%, C:G 5%, A:U 5%} and pair 2 with {G:C 100%}. However, as the conservation level decreases, most of the variable base pairs in the HIV-1 secondary structure model have significantly lower purity scores than the analogous in the 16S rRNA secondary structure. This results indicates the positions involved in the HIV-1 predicted base pairs are not

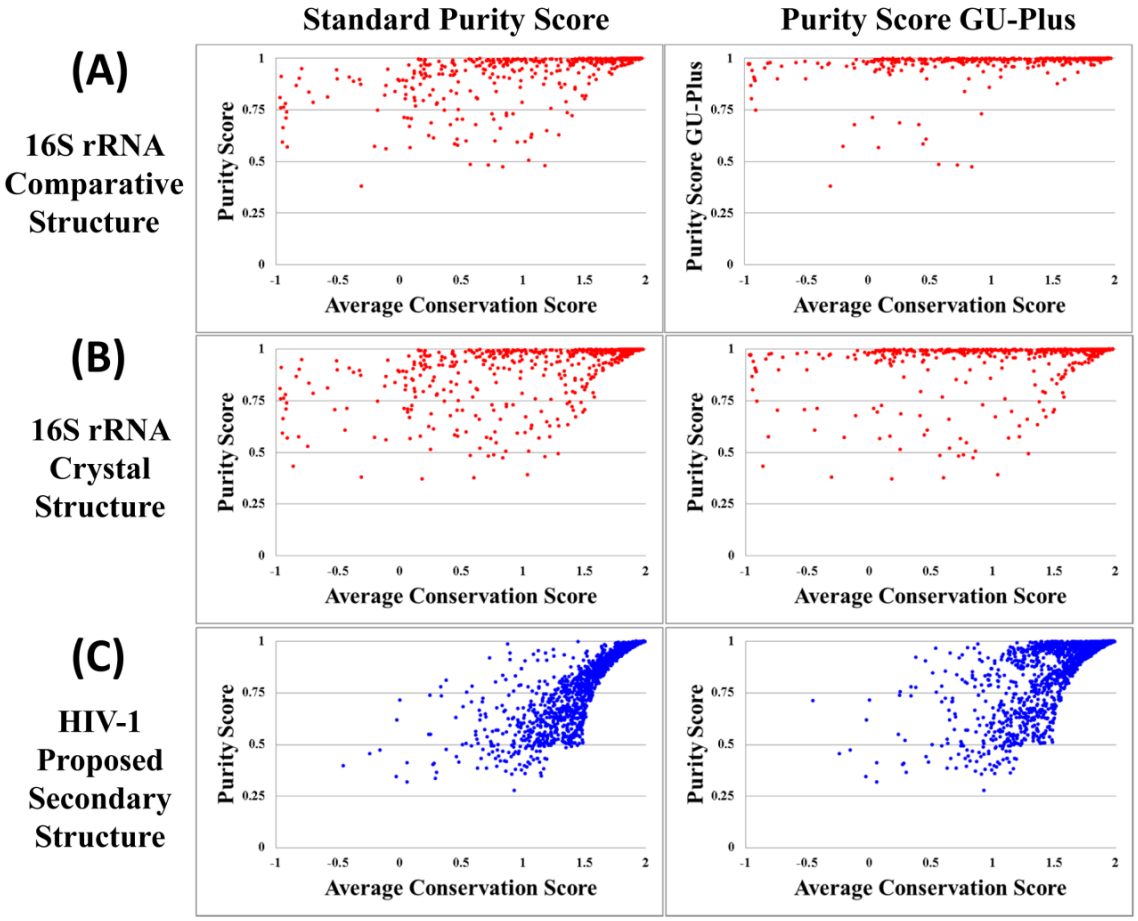


Figure 4.2: The distribution of purity score and average conservation score for the two positions that form a base pair in the 16S rRNA comparative structure model (A), secondary structure base pairs in crystal structure (B), and the predicted HIV-1 secondary structure model (C).

To distinguish the highly conserved pairs from the highly variable pairs when their purity scores are identical, a new covariation metric – confidence score is introduced. It measures the confidence level for a pair of positions with significant covariation or good purity to be a potential base pair (details in Methods section). Higher confidence score indicates the two positions are more likely to be covariant with one another. The confidence score for every base pair in the bacterial 16S rRNA secondary structure and the HIV-1 secondary structure model is calculated and plotted in Figure 4.3.

While the vast majority of the base pairs in the bacterial 16S rRNA have a good confidence score of 1 or higher (Figure 4.3A), most of the predicted base pairs in the HIV-1 secondary structure model have low confidence scores below 1 (Figure 4.3B).

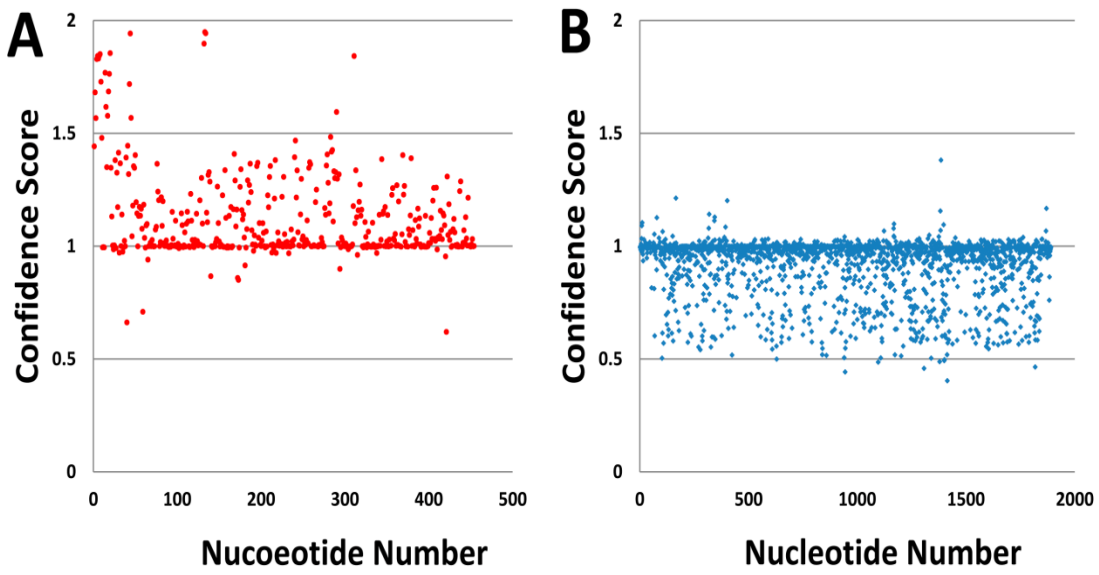


Figure 4.3: The distribution of confidence score for the two positions that form a base pair in the bacterial 16S rRNA secondary structure (A), and the proposed HIV-1 secondary structure model (B).

As shown in Figure 2.6 in Chapter 2 and empirical knowledge, the amount of covariation is directly proportional to the amount of variation within a multiple sequence alignment. The variation/covariation analysis of the bacterial 16S rRNA secondary structure (Figure 4.4A) and the predicted HIV-1 secondary structure model (Figure 4.4B) reveals that (1) while the base pair covariation is one of the major source of variation in the bacterial 16S rRNA, the predicted base pairs in the HIV-1 secondary structure model have very few covariations; (2) in SHAPE-directed HIV-1 secondary structure model, most of the variation is caused by the nucleotide change in the unpaired regions.

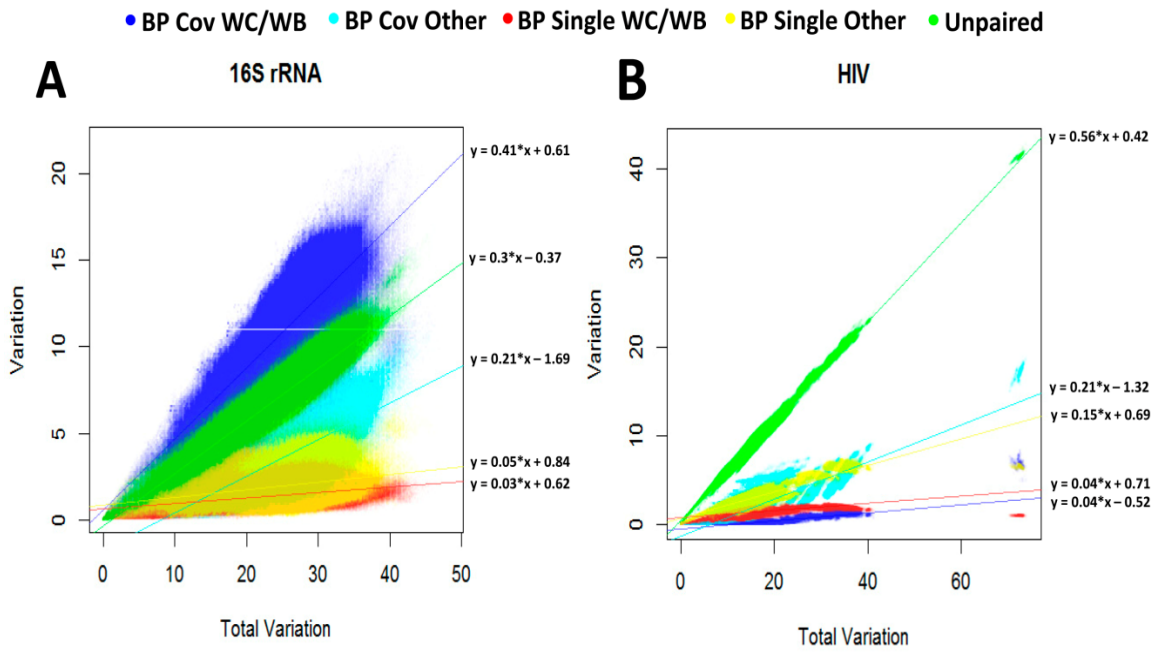


Figure 4.4: Variation/covariation analysis of the bacteria 16S rRNA secondary structure (A) and the predicted HIV-1 secondary structure model (B). Total variation in each pairwise set of sequences (X-direction) is plotted vs. (1) the amount of canonical (Watson-Crick or Wobble) variation (deep blue) and non-canonical variation (light blue) in that set of sequences for the two positions that are base paired in the secondary structure, (2) the amount of canonical variation (red) and non-canonical variation (yellow) only occur at one position of the two that are base paired in the secondary structure , and (3) variation in the unpaired region of the second structure (green) (Y-direction). The slope, Y-intercept, and R^2 co-efficiency values of the linear regression line for each of the three analyses are at the right side of the line.

The overall results of the HIV-1 secondary structural model evaluation suggest that: 1) a large percentage of the predicted base pairs are highly conserved, which cannot be fully evaluated from the perspective of comparative analysis; 2) for most predicted variable base pairs, the two positions that form the base pair do not covary with one another. Therefore, though the authors claimed that the model has been verified with comparative analysis, the SHAPE-directed secondary structure model of the entire HIV-1 RNA genome do not have the support of comparative analysis.

2. Base Pair Prediction with MI-based Methods and Helix-extension

In my previous research project, the performance of Phylogenetic Events Counting (PEC) method is compared with Mutual Information based methods including standard MI (*MIxy*) and corrected MI (*MIp*) (Details in Chapter 2). The results revealed that while PEC identified more real base pairs, mutual information based method (*MIxy*) and its variants (*MIp*) were capable of identifying significant amount of real base pairs with high accuracy. Because of the short evolutionary history and the lack of reliable taxonomy information for HIV-1, the PEC method is not suitable to predict the secondary structure of HIV-1 RNA genome. Therefore, *MIp* is utilized to determine the positional covariation between pairwise columns within the HIV-1 multiple sequence alignment.

The *MIp* scores of every pairwise positions are determined with standard calculation (details in the Methods section). The top 500 pairs of positions with the highest *MIp* scores are considered to have significant covariation. Only three (out of 1891) predicted base pairs (7443:7459, 9087:9122, 9085:9124) in the SHAPE-directed HIV-1 secondary structure model have strong covariation (high *MIp* value) and are selected as the top 500 pairs. Thus these three pairs are named "True Covariant Base Pairs". The rest 497 pairs (top 500 – 3 True Covariant Base Pairs) that are not present in the proposed HIV-1 secondary structure model are categorized as "Extra Covariant Pairs".

As described in chapter 2, helix-extension strategy has been substantiated sensitive and accurate in the identification of highly-conserved base pairs. Thus the helix-extension procedure is utilized to identify the highly-conserved base pairs in HIV-1 RNA genome, and measure the confidence level of the predicted highly-conserved base pairs in the SHAPE-directed HIV-1 secondary structure model. The helix consisting of four of more consecutive and antiparallel WC/Wobble base pairs assures over 99% probability that the helix is not formed by chance or random change. Therefore the extended helices with four or more base pairs are considered to be trustworthy. Using the top 500 pairs of positions with the highest *MIp* scores as the nucleation pairs, the helix-extension procedure identifies 31 helices consisting of at least four canonical base pairs. Two out of

the 31 extended helices are in the SHAPE-directed HIV-1 secondary structure model (helix 7443-7448:7454-7459, 9080-9090:9119-9129), while both helices are nucleated from the three "True Covariant Base Pair" (7443:7459, 9087:9122, 9085:9124). The 14 extended pairs in these two helices are present in the SHAPE-directed HIV-1 secondary structure model, thus are categorized as "True Extended Base Pairs". The rest 29 extended helices involve 121 extended pairs that are not present in the SHAPE-directed HIV-1 secondary structure model, thus named "Extra Extended Pairs". This result suggests that only small portion of the predicted base pairs in the SHAPE-directed HIV-1 secondary structure model can be validated with significant covariation (3 "True Covariant Base Pairs" and 14 "True Conserved Extended Base pairs"), while comparative analysis identifies 29 potential helices with strong covariant nucleation pair and considerable extended length of helix that are not in the SHAPE-directed HIV-1 secondary structure model.

While over 75% of the predicted base pairs in the SHAPE-directed HIV-1 structural model are highly conserved (Table 1), the positions forming these putative highly-conserved base pairs are not likely to have significant covariations, thus won't be selected as the nucleation pairs for helix-extension procedure. To check the credentials of these highly-conserved predicted base pairs, 1874 predicted base pairs in the SHAPE-directed HIV-1 structural model (1891 – 3 True Covariant Base Pairs – 14 True Extended Base Pairs) are utilized as the nucleation pairs to perform helix-extension procedure. This extension identifies 207 putative helices consisting of four or more canonical base pairs. Among the 1874 predicted base pairs in the SHAPE-directed HIV-1 secondary structure model that are used as the nucleation pairs, 990 pairs can be successfully extended, thus named "Neutral Extendable Base Pairs", while 884 pairs fails the helix-extension criteria (named "False Non-extendable Base Pairs").

Discussion

The RNA structure of HIV-1 has been discovered to either critical for or directly regulate diverse functions in viral life cycle including the synthesis of viral DNA, RNA splicing, genome packaging, and interactions with both viral and cellular proteins. While the 3D structure of entire HIV-1 genomic RNA is hard to be obtained with X-ray or NMR due to the size and flexibility of the large RNA molecule, accurate secondary structure models can help to elucidate the 3D structure and reveal conservative structural motifs which is usually tied with functions.

Single-nucleotide resolution chemical mapping (foot-printing) for highly-structured RNA has been rapidly advanced with multiple new technologies including novel chemical modification strategies and faster data analysis algorithms. Selective 2′-hydroxyl acylation by primer extension (SHAPE) technology measures local backbone flexibility in RNA molecule and scores the pairing probability of single nucleotide. Watts et al. proposed a SHAPE-directed secondary structure model of an entire HIV-1 RNA genome [144], and have proven useful in determining functions of RNA regions . However, recent benchmark of SHAPE technology on six small RNA molecules indicates that the accuracy of the proposed HIV-1 structural model could be lower than 50%, which is much lower than expected.

I evaluated the SHAPE-directed HIV-1 secondary structure model using comparative analysis methods. There are ~25% of the predicted base pairs in the SHAPE-directed model have low percentage of canonical base pair in the sequence alignment (Table 4.1). The covariation metrics including conservation scores, purity scores, confidence scores and variation/covariation plot reveal that vast majority of predicted base pairs in the SHAPE-directed HIV-1 structural model do not have significant covariation, thus lack the support of comparative analysis (Figure 4.1 – 4.4). A *de novo MIp* calculation and subsequent helix-extension measures the confidence levels of the 1891 proposed base pairs in the SHAPE-directed HIV-1 structural model. Two proposed helices (totally 17 base pairs) are identified with covariation methods, and therefore are assigned with high confidence level (True Covariant Base Pairs and True Conserved

Extended Base Pairs). 990 out of 1891 proposed based pairs in the HIV-1 structural model are highly-conserved but extendable to form a helix with a minimal of four base pairs. The amount of variations at these 990 pairs of positions is too low to distinguish true base pairs from false base pairs. Therefore, theses 990 proposed base pairs are marked as "Neutral Conserved Extendable Base Pairs". It will be a doable task to tell the quality of these "Neutral Conserved Extendable Base Pairs" with a larger multiple sequence alignment consisting of more HIV-1 complete genome sequences with more diversity. The rest 884 proposed base pairs neither have significant covariation scores, nor can be extended to form a helix with statistical significant length. These 884 pairs (named "False Non-extendable Base Pairs") are of very low confidence level, and will cause ambiguity in the structure determination. These "False Non-extendable Base Pairs" could be caused by two possible scenarios: 1) the pairing probabilities obtained from SHAPE reactivity values and the thermodynamic-based conversion algorithm are misinterpreted to build the secondary structure model; 2) these predicted base pairs are not crucial for the viral propagation, thus HIV-1 can tolerate any types of variations at these positions and evolve into different viral strands.

Overall, only a small portion of the base pairs in the SHAPE-directed secondary structure model of HIV-1 genome RNA are supported by comparative analysis. The structures of the entire viral RNA genomes are too large and complex to be predicted with a single approach and very limited number of sequences. The current SHAPE-directed HIV-1 RNA secondary structure requires additional information, such as evidence from experiments and comparative analysis, to validate true base pairs and eliminate the possible false base pairs and other ambiguities, especially in the regions marked with low confidence levels (False Non-extendable Base pairs and Neutral Conserved Extendable Base Pairs).

Chapter 5: Summary and Future Work

The accurate prediction of RNA secondary structure using comparative analysis is essential to decipher the secondary structure and other higher-order structural constraints of RNA molecules. In my first project, I developed a novel and powerful covariation method – Phylogenetic Events Counting (PEC) method for the identification of positional covariations. The PEC method utilizes phylogenetic information of sequences within the sequence alignment, and traverses through the phylogenetic tree to count the mutual changes on a pair of positions. The comparison between PEC and other statistics-based methods reveals that PEC is more sensitive and accurate in the identification of base pairs and other constrains in the RNA structure. With the complementation of joint N-Best and helix-extension strategy, PEC method is able to identify the maximal number of base pairs. In addition to the identification of base pair in the RNA higher-order structure, the analysis discovers a new type of structural constraint – neighbor effects which generally occur between sets of positions that are in proximity in the three-dimensional structure of RNAs. The neighbor effects have weaker but significant covariation with one another and possibly cause fitness function for a local cluster of nucleotides in the RNA structure.

The comparative methods are used to evaluate the proposed SHAPE-directed secondary structure model of entire HIV-1 RNA genome. Various covariation metrics reveals that the vast majority of the predicted base pairs in the HIV-1 secondary structure model do not have support from comparative analysis. In parallel, a *de novo* covariation analysis with mutual information based method and helix-extension procedure identifies 73 putative helices containing at least three base pairs. The 1891 predicted base pairs in the SHAPE-directed HIV-1 secondary structure model are categorized into four classes with different confidence levels. The results suggests that 17 predicted base pairs are supported by covariation analysis thus have high confidence level, 884 predicted base pairs have very low confidence level which require intensive improvement and correction, and 990 predicted base pairs are highly conserved which require additional information to verify.

The creation of well-aligned multiple RNA sequence alignments are essential for the subsequent comparative analysis. The traditional manual curation requires a significant amount of time and effort, which is not feasible for the extensive amount of nucleic acid sequences determined by next-gen sequencing technology. I developed a template-based alignment program package -- CRWAlign-2, which utilizes multiple dimensions of information about RNAs in rCAD. The program generates the structural descriptor for a RNA molecule at different phylogenetic nodes, searches for structure models satisfying conditions defined in the descriptor, and align the new sequences based on the primary and secondary structural similarity. When compared with eight other RNA sequence alignment programs, CRWAlign-2 is more accurate than other programs. Even for sequences with pairwise identity below 80%, CRWAlign-2 is still able to maintain a very high accuracy (> 95%). This improvement will significantly reduce the amount of time required for manual curation, especially in the variable regions of RNA molecules. CRWAlign-2 also generates the entire secondary structure model for each sequence to be aligned. This feature enables numerous biological applications. Several future tasks include 1) generating a set of distinguishable structural descriptors of different tRNAs, rRNAs and other RNA molecules for the purpose of sequence annotation, and 2) using this system to align large amount of RNA sequences to improve the data curation of sequence alignment.

References

1 Frohlich, K. S. & Vogel, J. Activation of gene expression by small RNA. *Curr. Opin. Microbiol.* 12, 674-682 (2009).

2 Georg, J. *et al.* Evidence for a major role of antisense RNAs in cyanobacterial gene regulation. *Mol. Syst. Biol.* 5, 305 (2009).

3 Khraiwesh, B. *et al.* Transcriptional control of gene expression by microRNAs. *Cell* 140, 111-122 (2010).

4 Sahoo, T. *et al.* Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. *Nat. Genet.* 40, 719-721 (2008).

5 Peters, J. Prader-Willi and snoRNAs. *Nat. Genet.* 40, 688-689 (2008).

6 Sabin, Leah R., Delás, M. J. & Hannon, Gregory J. Dogma Derailed: The Many Influences of RNA on the Genome. *Mol Cell* 49, 783-794 (2013).

7 Ball, P. DNA: Celebrate the unknowns. *Nature* 496, 419-420 (2013).

8 Mattick, J. S. Rocking the foundations of molecular genetics. *Proc Natl Acad Sci U S A* 109, 16400-16401 (2012).

9 Wang, K. C. *et al.* A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* 472, 120-124 (2011).

10 Darwin, C. *On the Origin of Species* 502 (John Murray, 1859).

11 Gutell, R. R., Larsen, N. & Woese, C. R. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiol. Rev.* 58, 10-26 (1994).

12 Holley, R. W. *et al.* Structure of a ribonucleic acid. *Science* 147, 1462-1465 (1965).

13 Levitt, M. Detailed molecular model for transfer ribonucleic acid. *Nature* 224, 759-763 (1969).

14 Robertus, J. D. *et al.* Structure of yeast phenylalanine tRNA at 3 A resolution. *Nature* 250, 546-551 (1974).

15 Suddath, F. L. *et al.* Three-dimensional structure of yeast phenylalanine transfer RNA at 3.0angstroms resolution. *Nature* 248, 20-24 (1974).

16 Gutell, R. R., Lee, J. C. & Cannone, J. J. The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct. Biol.* 12, 301-310 (2002).

17 Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science* 289, 905-920 (2000).

18 Wimberly, B. T. *et al.* Structure of the 30S ribosomal subunit. *Nature* 407, 327-339 (2000).

19 Ozer, S., Doshi, K. J., Xu, W. & Gutell, R. R. rCAD: A Novel Database Schema for the Comparative Analysis of RNA. *7th IEEE International Conference on e-Science* (2011).

20 Freier, S. M. *et al.* Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci. USA* 83, 9373-9377 (1986).

21 Konings, D. A. & Gutell, R. R. A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs. *RNA* 1, 559-574 (1995).

22 Doshi, K. J., Cannone, J. J., Cobaugh, C. W. & Gutell, R. R. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics* 5, 105 (2004).

23 Cannone, J. J. *et al.* The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* 3, 2 (2002).

24 Cate, J. H. *et al.* Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science* 273, 1678-1685 (1996).

25 Adams, P. L., Stahley, M. R., Kosek, A. B., Wang, J. & Strobel, S. A. Crystal structure of a self-splicing group I intron with both exons. *Nature* 430, 45-50 (2004).

26 Burke, J. M. *et al.* Structural conventions for group I introns. *Nucleic Acids Res.* 15, 7217-7221 (1987).

27 Kazantsev, A. V. *et al.* Crystal structure of a bacterial ribonuclease P RNA. *Proc. Natl. Acad. Sci. USA* 102, 13392-13397 (2005).

28 Torres-Larios, A., Swinger, K. K., Krasilnikov, A. S., Pan, T. & Mondragon, A. Crystal structure of the RNA component of bacterial ribonuclease P. *Nature* 437, 584-587 (2005).

29 Pace, N. R., Smith, D. K., Olsen, G. J. & James, B. D. Phylogenetic comparative analysis and the secondary structure of ribonuclease P RNA--a review. *Gene* 82, 65-75 (1989).

30 Bessho, Y. *et al.* Structural basis for functional mimicry of long-variable-arm tRNA by transfer-messenger RNA. *Proc. Natl. Acad. Sci. USA* 104, 8293-8298 (2007).

31 Williams, K. P. & Bartel, D. P. Phylogenetic analysis of tmRNA secondary structure. *RNA* 2, 1306-1310 (1996).

32 Vidovic, I., Nottrott, S., Hartmuth, K., Luhrmann, R. & Ficner, R. Crystal structure of the spliceosomal 15.5kD protein bound to a U4 snRNA fragment. *Mol. Cell* 6, 1331-1342 (2000).

33 Guthrie, C. & Patterson, B. Spliceosomal snRNAs. *Annu. Rev. Genet.* 22, 387-419 (1988).

34 Hainzl, T., Huang, S. & Sauer-Eriksson, A. E. Structure of the SRP19 RNA complex and implications for signal recognition particle assembly. *Nature* 417, 767-771 (2002).

35 Batey, R. T., Rambo, R. P., Lucast, L., Rha, B. & Doudna, J. A. Crystal structure of the ribonucleoprotein core of the signal recognition particle. *Science* 287, 1232-1239 (2000).

36 Zwieb, C. The secondary structure of the 7SL RNA in the signal recognition particle: functional implications. *Nucleic Acids Res.* 13, 6105-6124 (1985).

37 Gutell, R. R., Weiser, B., Woese, C. R. & Noller, H. F. Comparative anatomy of 16-S-like ribosomal RNA. *Prog. Nucleic Acid Res. Mol. Biol.* 32, 155-216 (1985).

38 Chargaff, E. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia* 6, 201-209 (1950).

39 Chargaff, E. Some recent studies on the composition and structure of nucleic acids. *J. Cell. Physiol. Suppl.* 38, 41-59 (1951).

40 Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171, 737-738 (1953).

41 Woese, C. R. *et al.* Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. *Nucleic Acids Res.* 8, 2275-2293 (1980).

42 Noller, H. F. *et al.* Secondary structure model for 23S ribosomal RNA. *Nucleic Acids Res.* 9, 6167-6189 (1981).

43 Fox, G. E. & Woese, C. R. 5S RNA secondary structure. *Nature* 256, 505-507 (1975).

44 Olsen, G. J. *Comparative analysis of nucleotide sequence data*, University of Colorado, (1983).

45 Chiu, D. K. & Kolodziejczak, T. Inferring consensus structure from nucleic acid sequences. *Comput. Appl. Biosci.* 7, 347-352 (1991).

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

47 Gutell, R. R., Noller, H. F. & Woese, C. R. Higher order structure in ribosomal RNA. *EMBO J.* 5, 1111-1113 (1986).

48 Gutell, R. R. & Woese, C. R. Higher order structural elements in ribosomal RNAs: pseudo-knots and the use of noncanonical pairs. *Proc. Natl. Acad. Sci. USA* 87, 663-667 (1990).

49 Gautheret, D., Damberger, S. H. & Gutell, R. R. Identification of base-triples in RNA using comparative sequence analysis. *J. Mol. Biol.* 248, 27-43 (1995).

50 Michel, F., Ellington, A. D., Couture, S. & Szostak, J. W. Phylogenetic and genetic evidence for base-triples in the catalytic domain of group I introns. *Nature* 347, 578-580 (1990).

51 Conn, G. L., Gutell, R. R. & Draper, D. E. A functional ribosomal RNA tertiary structure involves a base triple interaction. *Biochem.* 37, 11980-11988 (1998).

52 Yeang, C. H., Darot, J. F., Noller, H. F. & Haussler, D. Detecting the coevolution of biosequences--an example of RNA interaction prediction. *Mol. Biol. Evol.* 24, 2119-2131 (2007).

53 Dutheil, J., Pupko, T., Jean-Marie, A. & Galtier, N. A model-based approach for detecting coevolving positions in a molecule. *Mol. Biol. Evol.* 22, 1919-1928 (2005).

54 Tuffery, P. & Darlu, P. Exploring a Phylogenetic Approach for the Dectection of Correlated Substitutions in Proteins. *Mol. Biol. Evol.* 17, 1753-1759 (2000).

55 Wu, J. C., Gardner, D. P., Ozer, S., Gutell, R. R. & Ren, P. Correlation of RNA secondary structure statistics with thermodynamic stability and applications to folding. *J. Mol. Biol.* 391, 769-783 (2009).

56 Buck, M. J. & Atchley, W. R. Networks of coevolving sites in structural and functional domains of serpin proteins. *Mol. Biol. Evol.* 22, 1627-1634 (2005).

57 Dunn, S. D., Wahl, L. M. & Gloor, G. B. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 24, 333-340 (2008).

58 Kass, I. & Horovitz, A. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins* 48, 611-617 (2002).

59 Fodor, A. A. & Aldrich, R. W. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* 56, 211-221 (2004).

60 Dekker, J. P., Fodor, A., Aldrich, R. W. & Yellen, G. A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics* 20, 1565-1572 (2004).

61 Jones, D. T., Buchan, D. W., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28, 184-190 (2012).

62 Sadowski, M. I., Maksimiak, K. & Taylor, W. R. Direct correlation analysis improves fold recognition. *Computational Biology and Chemistry* 35, 323-332 (2011).

63 Bernhart, S. H., Hofacker, I. L., Will, S., Gruber, A. R. & Stadler, P. F. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* 9 (2008).

64 Hofacker, I. L. *et al.* Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte f. Chemie* 125, 167-188 (1994).

65 Hofacker, I. L. & Stadler, P. E. Memory Efficient Folding Algorithms for Circular RNA Secondary Structures. *Bioinformatics* 22, 1172-1176 (2006).

66 Knudsen, B. & Hein, J. Using stochastic context free grammars and molecular evolution to predict RNA secondary structure. *Bioinformatics* 15, 446-454 (1999).

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

68 Pedersen, J. S. *et al.* Identification and Classification of Conserved RNA Secondary Structures in the Human Genome. *PLoS Comput Biol* 2, e33 (2006).

69 Shannon, C. E. A Mathematical theory of communication. *Bell System Tech. J.* 27, 379–423 (1948).

70 Crick, F. H. Codon--anticodon pairing: the wobble hypothesis. *J. Mol. Biol.* 19, 548-555 (1966).

71 Yakovchuk, P., Protozanova, E. & Frank-Kamenetskii, M. D. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res.* 34, 564-574 (2006).

72 Zhou, H. & Zhang, Y. Pulling hairpinned polynucleotide chains: Does base-pair stacking interaction matter? *J. Chem. Phys.* 114, 8694-8700 (2001).

73 Zhou, H., Zhang, Y. & Ou-Yang, Z. C. Stretch-induced hairpin-coil transitions in designed polynucleotide chains. *Phys. Rev. Lett.* 86, 356-359 (2001).

74 Xu, W., Ozer, S. & Gutell, R. R. Covariant Evolutionary Event Analysis for Base Interaction Prediction Using a Relational Database Management System for RNA *Lecture Notes in Computer Science* 5566, 200-216 (2009).

75 Noller, H. F. *et al.* Secondary structure model for 23S ribosomal RNA. *Nucleic Acids Res.* 9, 6167-6189 (1981).

76 Schuwirth, B. S. *et al.* Structures of the bacterial ribosome at 3.5 A resolution. *Science* 310, 827-834 (2005).

77 Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 6, e28766, doi:10.1371/journal.pone.0028766 (2011).

78 Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci U S A* 106, 67-72 (2009).

79 Lindgreen, S., Gardner, P. P. & Krogh, A. Measuring covariation in RNA alignments: physical realism improves information measures. *Bioinformatics* 22, 2988-2995 (2006).

80 Gutell, R. Comparative studies of RNA: inferring higher-order structure from patterns of sequence variation. *Curr. Opin. Struct. Biol.* 3, 313-322 (1993).

81 Cochella, L. & Green, R. An active role for tRNA in decoding beyond codon:anticodon pairing. *Science* 308, 1178-1180 (2005).

82 Schmeing, T. M., Voorhees, R. M., Kelley, A. C. & Ramakrishnan, V. How mutations in tRNA distant from the anticodon affect the fidelity of decoding. *Nat. Struct. Mol. Biol.* 18 (2011).

83 Woese, C. R., Winker, S. & Gutell, R. R. Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proc. Natl. Acad. Sci. USA* 87, 8467-8471 (1990).

84 Gautheret, D. & Gutell, R. R. Inferring the conformation of RNA base pairs and triples from patterns of sequence variation. *Nucleic Acids Res.* 25, 1559-1564 (1997).

85 Elgavish, T., Cannone, J. J., Lee, J. C., Harvey, S. C. & Gutell, R. R. AA.AG@helix.ends: A:A and A:G base-pairs at the ends of 16 S and 23 S rRNA helices. *J. Mol. Biol.* 310, 735-753 (2001).

86 Lee, J. C. & Gutell, R. R. Diversity of base-pair conformations and their occurrence in rRNA structure and RNA structural motifs. *J. Mol. Biol.* 344, 1225-1249 (2004).

87 Schmeing, T. M. & Ramakrishnan, V. What recent ribosome structures have revealed about the mechanism of translation. *Nature* 461, 1234-1242 (2009).

88 Woese, C. R. Bacterial evolution. *Microbiol. Rev.* 51, 221-271 (1987).

89 Gardner, D. P., Ren, P., Ozer, S. & Gutell, R. R. Statistical potentials for hairpin and internal loops improve the accuracy of the predicted RNA structure. *J. Mol. Biol.* 413, 473-483, doi:S0022-2836(11)00931-4 [pii]

10.1016/j.jmb.2011.08.033 (2011).

90 Group, J. C. H. M. P. D. G. W. & Ravel, J. e. Evaluation of 16S rDNA-Based Community Profiling for Human Microbiome Research. *PLoS One* 7, e39315-e39315 (2012).

91 Robertson, C. E., Harris, J. K., Spear, J. R. & Pace, N. R. Phylogenetic diversity and ecology of environmental Archaea. *Curr. Opin. Microbiol.* 8, 638-642 (2005).

92 Huse, S. M., Ye, Y., Zhou, Y. & Fodor, A. A. A Core Human Microbiome as Viewed through 16S rRNA Sequence Clusters. *PLoS One* 7, e34242-e34242 (2012).

93 Higgins, D. G. & Sharp, P. M. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* 73, 237-244 (1988).

94 Thompson, J. D., Higgins, D. G. & Gibson, T. J. 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, 4673-4680 (1994).

95 Katoh, K. & Toh, H. Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. *BMC Bioinformatics* 9, 212 (2008).

96 Liu, K., Raghavan, S., Nelesen, S., Linder, C. R. & Warnow, T. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* 324, 1561-1564 (2009).

97 Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* 35, 7188-7196 (2007).

98 DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069-5072 (2006).

99 Cole, J. R. *et al.* The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.* 33 Database Issue, D294-296 (2005).

100 Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48, 443-453, doi:0022-2836(70)90057-4 [pii] (1970).

101 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* 215, 403-410 (1990).

102 Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* 25, 1335-1337 (2009).

103 Nawrocki, E. P. Ph.D. Dissertation, Washington University in St. Louis. (2009).

104 Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* 235, 1501-1531 (1994).

105 Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. J. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.*, (Cambridge University Press, 1998).

106 Laferriere, A., Gautheret, D. & Cedergren, R. An RNA pattern matching program with enhanced performance and portability. *Comput. Appl. Biosci.* 10, 211-212 (1994).

107 Macke, T. J. *et al.* RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.* 29, 4724-4735 (2001).

108 Reeder, J. & Giegerich, R. Locomotif: from graphical motif description to RNA motif search. *Bioinformatics* 23, i392-400, doi:10.1093/bioinformatics/btm179 (2007).

109 Ozer, S., Doshi, K. J., Xu, W. & Gutell, R. R. in *7th IEEE International Conference on e-Science* 15-22 (Stockholm, Sweden, 2011).

110 Lederberg, J. 'Ome Sweet 'Omics-- A Genealogical Treasury of Words. *The Scientist* 15 (2001).

111 Wintzingerode, F. V. Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS MICROBIOL Reviews* 21, 213-229 (1997).

112 Haas, B. J. *et al.* Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res* 21, 494-504 (2011).

113 Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J. & Weightman, A. J. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol.* 71, 7724-7736 (2005).

114 Quince, C., Lanzen, A., Davenport, R. J. & Turnbaugh, P. J. Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 12, 38-38 (2011).

115 Huber, T., G., F. & P., H. Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* 20, 2317-2319 (2004).

116 Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J. & Weightman, A. J. New Screening Software Shows that Most Recent Large 16S rRNA Gene Clone Libraries Contain Chimeras. *Appl. Envir. Microbiol.* 72, 5734-5741 (2006).

117 Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G. & Gibson, T. J. Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* 23, 403-405 (1998).

118 Nilsson, R. H. *et al.* An open source chimera checker for the fungal ITS region. *Molecular Ecology Resources* 10, 1076-1081 (2010).

119 Gardner, D. P. *et al.* in *Proceedings of 2012 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2012).* 237-243.

120 Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* 288, 911-940 (1999).

121 Zuker, M. & Jacobson, A. B. "Well-determined" regions in RNA secondary structure prediction: analysis of small subunit ribosomal RNA. *Nucleic Acids Res.* 23, 2791-2798 (1995).

122 Do, C. B., Woods, D. A. & Batzoglou, S. CONTRAfolH: RNA secondary structure prediction without physics-based models. *Bioinformatics* 22, e90-98 (2006).

123 Andronescu, M., Condon, A., Hoos, H. H., Mathews, D. H. & Murphy, K. P. Computational approaches for RNA energy parameter estimation. *RNA* 16, 2304-2318 (2010).

124 Gilbert, B. P. *et al.* Comparison of HIV-1 and HIV-2 infectivity from a prospective cohort study in Senegal. *Statist. Med.* 22, 573-593 (2003).

125 Frankel, A. D. & Young, J. A. HIV-1: fifteen proteins and an RNA. *Annu. Rev. Biochem.* 67, 1-25 (1998).

126 Damgaard, C. K., Andersen, E. S., Knudsen, B., Gorodkin, J. & Kjems, J. RNA interactions in the 5' region of the HIV-1 genome. *J. Mol. Biol.* 336, 369-379 (2004).

127 Puglisi, J. D., Tan, R., Calnan, B. J., Frankel, A. D. & Williamson, J. R. Conformation of the TAR RNA-arginine complex by NMR spectroscopy. *Science* 257, 76-80 (1992).

128 Aboul-ela, F., Karn, J. & Varani, G. The Structure of the Human Immunodeficiency Virus Type-1 TAR RNA Reveals Principles of RNA Recognition by Tat Protein. *J. Mol. Biol.* 253, 313-332 (1995).

129 Harrison, G. P. & Lever, A. The human immunodeficiency virus type 1 packaging signal and major splice donor region have a conserved stable secondary structure. *Journal of Virology* 66, 4144-4153 (1992).

130 Hayashi, T., Shioda, T., Iwakura, Y. & Shibuta, H. RNA Packaging Signal of Human lmmunodeficiency Virus Type 1. *Virology* 188, 590-599 (1992).

131 Paillart, J. C., Skripkin, E., Ehresmann, C. & Marquet, R. A loop-loop "kissing" complex is the essential part of the dimer linkage of genomic HIV-1 RNA. *Proc. Natl. Acad. Sci. USA* 93, 5572-5577 (1996).

132 Berkhout, B., Van Wamel, J. L., Ehresmann, B. & Ehresmann, C. Role of the DIS hairpin in replication of human immunodeficiency virus type 1. *J. Virol.* 70, 6723-6732 (1996).

133 Zemmel, R. Z., Kelley, A. C., Karn, J. & Butler, J. G. P. Flexible Regions of RNA Structure Facilitate Co-operative Rev Assembly on the Rev-response Element. *J. Mol. Biol.* 258, 763-777 (1996).

134 Laughrea, M. *et al.* Mutations in the kissing-loop hairpin of human immunodeficiency virus type 1 reduce viral infectivity as well as genomic RNA packaging and dimerization. *J. Virol.* 71, 3397-3406 (1997).

135 Wang, Q., Barr, I., Guo, F. & Lee, C. Evidence of a novel RNA secondary structure in the coding region of HIV-1 pol gene. *RNA* 14, 2478-2488 (2008).

136 Baudin, F. *et al.* Functional sites in the 5’ region of human immunodeficiency virus type 1 RNA form defined structural domains. *J. Mol. Biol.* 229, 382-397 (1993).

137 Pappalardo, L., Kerwood, D. J., Pelczer, I. & Borer, P. N. Three-dimensional folding of an RNA hairpin required for packaging HIV-1. *J. Mol. Biol.* 282, 801-818 (1998).

138 Amarasinghe, G. K., De Guzman, R. N., Turner, R. B. & Summers, M. F. NMR structure of stem-loop SL2 of the HIV-1 psi RNA packaging signal reveals a novel A-U-A base-triple platform. *J. Mol. Biol.* 299, 145-156 (2000).

139 Zeffman, A., Hassard, S., Varani, G. & Levelr, A. The major HIV-1 packaging signal is an extended bulged stem loop whose structure is altered on interaction with the Gag polyprotein. *J. Mol. Biol.* 297, 877-893 (2000).

140 Greatorex, J., Gallego, J., Varani, G. & Lever, A. Structure and stability of wild-type and mutant RNA internal loops from the SL-1 domain of the HIV-1 packaging signal. *J. Mol. Biol.* 322, 543-557 (2002).

141 Lu, K. *et al.* NMR Detection of Structures in the HIV-1 5′-Leader RNA that Regulate Genome Packaging. *Science* 334, 242-245 (2011).

142 Stephenson, J. D., Li, H., Kenyon, J. C., Symmons, M. & Klenerman, D. Three-Dimensional RNA Structure of the Major HIV-1 Packaging Signal Region. *Structure* 21, 951-962 (2013).

143 Shapiro, B. A., Yingling, Y. G., Kasprzak, W. & Bindewald, E. Bridging the gap in RNA structure prediction. *Curr. Opin. Struct. Biol.* 17, 157-165 (2007).

144 Watts, J. M. *et al.* Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* 460, 711-716 (2009).

145 Kladwang, W., VanLang, C. C., Cordero, P. & Das, R. Understanding the errors of SHAPE-directed RNA structure modeling. *Biochemistry* 50, 8049-8056, doi:10.1021/bi200524n (2011).