# 國 立 交 通 大 學

## 生物資訊及系統生物研究所

碩 士 論 文

搜尋相似的RNA 三級子結構

Searching for Similar RNA Tertiary Substructures

研究生：劉芸蓁
指導教授：盧錦隆 博士

中華民國 九十九 年 六 月

# 搜尋相似的RNA 三級子結構

# Searching for Similar RNA Tertiary Substructures

研究生：劉芸蓁　　Student：Yun-Chen Liu
指導教授：盧錦隆 博士　Advisor：Dr. Chin Lung Lu

國　立　交　通　大　學

生物資訊及系統生物研究所

碩　士　論　文

A Thesis Submitted to Institute of Bioinformatics
College of Biological Science and Technology
National Chiao Tung University in partial Fulfillment of the
Requirements for the Degree of Master in
Biological Science and Technology
June 2010
Hsinchu, Taiwan

# 中文摘要

近年來人們對非編碼RNA (ncRNAs)的興趣正快速地成長，因為他們在細胞內扮演著許多重要的角色，儘管這些 ncRNAs 不會被轉譯成蛋白質。事實上，大多數已有的 ncRNAs 的功能仍是未知。如同蛋白質，一個較為可靠去決定出 ncRNA 功能的方法便是去分析他們的三級結構，因為分子的結構通常會比他們的一級序列在演化上還來得保守。在這方面，最近一連串的努力與研究已使得存放在 PDB 資料庫裡頭的 RNA 三級結構在數量與大小上都大大地增加。因此，發展出一個能夠快速且正確地搜尋出 PDB 資料庫裡結構相似的 RNAs 的自動化工具就顯得愈來愈重要了。在這個研究中，我們利用結構字元的方法發展出一個名叫 R3D-BLAST 工具讓生物學家去搜尋 PDB 資料庫裡與某一個 RNA 三級結構相似的 RNAs。我們設計出 R3D-BLAST 背後的基本想法如下：首先，我們利用 RNA 核苷酸骨幹上的二個假的扭轉角(Pseudo-torsion Angles)以及親和性互動式(Affinity Propagation)的分群方法得到一個含有 23 個字母的結構字元集，然後再根據這個結構字元集把目前存放在 PDB 資料庫裡頭所有 RNA 三級結構編碼成一級的序列。接著我們再利用 BLAST 這個程式去搜尋出與 qury RNA 三級結構局部相似的 RNAs。我們實驗的結果最後證明：我們的 R3D-BLAST 在識別出與 qury RNA 三級結構有局部相似的 RNAs 這方面的表現確實比 BLAST 還要好，而且在找出與 qury RNA 三級結構有整體相似的 RNAs 這方面的表現也比 FASTR3D 還好。因此，我們相信 R3D-BLAST 在結構生物學的研究上可以充當一個有用的生物資訊工具。

# Abstract

In recent years, there is a fast growing interest in non-coding RNAs (ncRNAs) because they play a lot of essential roles in many cellular processes, even though the transcripts of these ncRNAs are not translated into proteins. Actually, the function of most available ncRNAs is still unknown. Likewise to proteins, a more reliable way for determining the functions of ncRNAs is to analyze their three-dimensional (3D, tertiary) structures, because structures of molecules are typically more evolutionarily conserved than their primary sequences. In this regard, a series of recent efforts and studies has led to a substantial increase in both the number and the size of solved RNA tertiary structures deposited in the PDB database. Therefore, it has become more and more crucial to develop automatic tools that are able to fast and accurately search the PDB database for structurally similar RNAs. In this study, we have used a structural-alphabet approach to develop a web server, called R3D-BLAST, that allows biologists to search the PDB database for structural similarities of an RNA 3D structure. The basic idea behind our R3D-BLAST is as follows. We first encode all the RNA 3D structures deposited in the PDB database as 1D sequences using the structural alphabet of 23 letters, which was obtained by using the two pseudo-torsion angles of RNA nucleotide backbones and the affinity propagation clustering approach. We then apply BLAST to searching for RNA molecules whose 3D structures are locally similar to that of the query RNA. Our experimental results have finally shown that our R3D-BLAST indeed has better performance than BLAST, a famous bioinformatics tool to find homologous proteins/RNAs only based on their sequence similarity, for identifying those RNA molecules whose tertiary substructures are locally similar to that of the query RNA, as well as FASTR3D for finding those

RNAs whose structures are entirely similar to that of the query RNA. Therefore, we believe that our R3D-BLAST can serve as a useful bioinformatics tool in the study of structural biology.

# Acknowledgement

兩年的研究所經歷，首先要感謝我的指導教授，盧錦隆教授，不僅在研究時不遺餘力幫助我，且教導我作研究時該有的態度和方法，面對困難和分析問題的能力也在這兩年得到很好的磨練。

感謝學姊，在研究上是個很好的討論對象，並且也是個很好的傾訴對象，每次和學姊討論過後不管是公私事，都能得到很好的舒解。

感謝已經畢業的學長們，在第一年的研究中，謝謝你們的照顧，帶我們四處去吃吃喝喝，一起打球、騎腳踏車、玩遊戲、訂便當、過節日。在研究方面有問題時也時常麻煩學長們。

感謝和我同屆的同學，一起修課，一起作研究，一起被電，一起度過這兩年。也感謝你們容忍我常常神經很大條的行為，還有忍受我很吵。這兩年有你們陪伴很開心。

感謝學弟們，負責實驗室的許多雜務，讓我們可以專心作研究，並且帶動實驗室的風氣，也常常為實驗室帶來歡樂。

感謝我的男朋友，在我疲憊、心情不好和遇到困難時，陪在我身邊。這段期間，非常謝謝你的支持和陪伴，讓我總是能夠繼續面對困境。

感謝我的家人們，在我作研究疲憊時，有個放鬆的地方，也能無後顧之憂地作研究。

芸蓁　20100728

# Contents

# List of tables

# List of figures

# Chapter 1

# Introduction

In recent years, there is a fast growing interest in noncoding RNAs (ncRNAs) because, they play essential roles in many cellular processes, including gene regulation, RNA modification and chromosome replication [9, 14, 25, 31], although their transcripts are not translated into proteins. However, the function of most available ncRNAs is unknown and needs to be determined. Likewise to proteins, a common and useful approach for annotating the function of an ncRNA is to search databases for similar RNA molecules whose functions are already known. For this purpose, several databases of ncRNAs have been proposed, such as NONCODE [19], RNAdb [27], miRBase [17], fRNAdb [20] and ncRNAdb [32]. For these databases, however, the search is performed solely by querying keywords, accession numbers, transcript/organism names and/or nucleotide sequences. Compared with the 20-letter protein alphabet, the 4-letter RNA alphabet is smaller and less informative, leading to that searching for similar RNA molecules based on sequence comparison/alignment is not as accurate and powerful as it does for proteins.

Actually, a more reliable way for determining the functions of ncRNAs is to analyze their structures, since structures of molecules are typically more evolutionarily conserved than their sequences. In this regard, a series of recent efforts and studies has led to a substantial increase in both the number and the size of solved RNA structures deposited in the PDB and NDB databases [4, 3]. Therefore, it has become more and more crucial to develop automatic tools that are able to efficiently and accurately search for structurally similar RNA substructures and motifs against the

PDB/NDB database. Basically, detecting structural similarities in two RNA molecules at secondary structure level is an easy job, whereas it is intractable at tertiary structure level, because it has been shown to be an NP-hard problem even to find a constant ratio approximation algorithm for computing a pair of maximal substructures from two RNA (or protein) three-dimensional (3D) structures with exhibiting the highest degree of similarity [22]. Therefore, currently available tools, such as ARTS [10, 11], DIAL [15], SARSA [7], SARA [6] and iPARTS [35], are all based on some heuristic approaches for comparing the similarities of two RNA 3D structures.

ARTS is a web server for detecting maximum common substructures between two given RNA 3D structures, which was implemented by Dror *et al*. [10,11] based on a heuristic algorithm of cubic running time. By representing each RNA 3D structure by a set of its phosphate atoms, ARTS identifies all structurally similar quadrates (*i.e.,* four phosphate atoms located on two successive base pairs) between the two input RNA 3D structures and continues to extend them by using a greedy method for including additional coincident base pairs and unpaired nucleotides. ARTS is a good tool for detecting RNA structural motifs, but it is still time-consuming for ARTS to compare large RNA molecules (*e.g.,* ribosomal RNAs) because of its cubic time complexity and, as was pointed out in [15], the structural alignments produced by ARTS may be incorrect sometimes.

Later on, to overcome the inaccurate problems caused by ARTS, Ferre` *et al*. [15] implemented DIAL, a web server for aligning two RNA 3D structures, by using a dynamic programming algorithm of quadratic running time based on a scoring function that combines similarities of nucleotide sequences, base pairs, pseudo-torsion and torsion angles. DIAL is a

versatile web server by providing the user three types of alignments: (i) global alignment, (ii) local alignment and (iii) an extension of global-semiglobal alignment, that is, a global alignment of a motif *A* consisting of one or more contiguous segments is aligned to a contiguous sequence *B*; while gap penalties apply throughout for *A* (global alignment), gaps at the ends of *B* as well as between portions aligned to contiguous segments of *A* are not penalized (so-called middle gaps).

Next, we developed PARTS [7] for pairwise alignments of RNA tertiary structures based on a structural alphabet (SA)-based algorithm. Its basic idea is to reduce input RNA 3D structures to 1D sequences of SA letters using backbone torsion angles of constituent residues and continue to use algorithms of classical sequence alignments (including global, local, semiglobal and normalized local alignments) to compare these 1D SA-encoded sequences for determining their structural similarities. More recently, we have further derived a new SA of RNA nucleotide conformations using their pseudo-torsion angles. Based on this newly designed SA, we have re-implemented our PARTS as iPARTS [35] (short for improved PARTS) to make its structural alignments of two RNA molecules more accurate.

Recently, Capriotti and Marti-Renom [5] have proposed a new web server, called SARA, for globally aligning two RNA 3D structures based on the unit-vector approach and have further shown its ability in function assignment of RNA structures [6]. For each input RNA 3D structure, SARA first identifies an atom trace that consists of all contiguous atoms of user-defined type and also calculates all unit-vectors between any two consecutive atoms along this trace. For each nucleotide of an input RNA structure, it then groups a set of *k* consecutive unit-vectors starting from this nucleotide and places these *k* unit-vectors at the origin of a unit-sphere, where *k* is a user-defined positive integer. Finally, SARA applies a dynamic programming algorithm without

penalizing end gaps to the two sequences of unit-spheres to find an optimal semiglobal alignment between them.

However, all these methods mentioned above have at least quadratic-time complexity and hence are impractical for searching ever-increasing databases of RNA tertiary structures. Currently, there are several tools that can be used to search motifs in RNA structures, including FR3D [30], PRIMOS [13], RNAMotif [24], RNA FRABASE [28, 29], and FASTR3D [23]. FR3D uses a base-centered method to perform a geometric search of RNA local/composite 3D motifs. PRIMOS searches for locally structural similarities of consecutive RNA fragments by comparing their pseudo-torsion angles. RNAMotif finds the fragments of an RNA sequence that conform to a predefined descriptor of defining a particular motif of secondary structure. RNA FRABASE was developed on the basis of RNA primary sequences and/or secondary structures using the methods of regular expression and pattern recognition. FASTR3D was designed based on a hashing algorithm that is able to fast and accurately find structural similarities for a query of RNA molecule in the PDB database.

As mentioned above, RNA FRABASE and FASTR3D both find structurally similar RNAs whose secondary structures exactly match that of the query RNA. However, there are many examples of RNAs that share similar 3D structures but have only similar or even different 2D structures. For example, the two tRNA (PDB IDs: 2DR2 and 2JO2), as shown in Figure 1-1, have very similar 3D structures (Figure 1-1c), even though their 1D sequences and 2D structures look different. When querying one of these two tRNAs, RNA FRABASE and FASTR3D both fail to find the other one with any structural similarity. More often, some RNAs may share only similar local 3D substructures, rather than similar entire 3D structures (see Figure 1-2 for an example).

Therefore, in this study, we have developed a new web server, called R3D-BLAST, based on a structural-alphabet approach for fast and accurately searching for structural similarities for a query of RNA molecules.

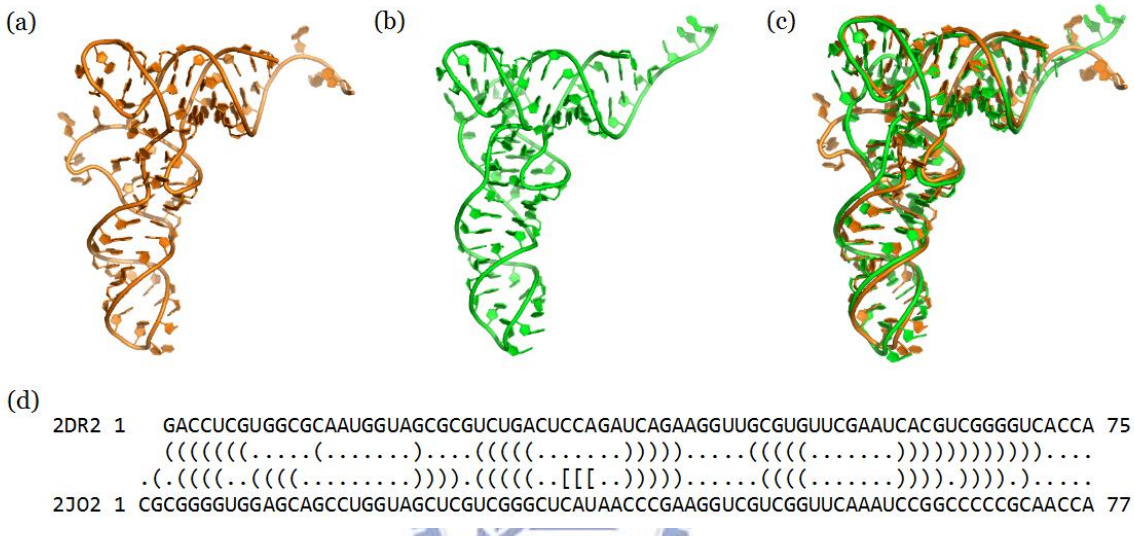



**Figure 1-1.** Two tRNAs: (a) PDB ID: 2DR2, chain ID: B and (b) PDB ID: 2JO2, chain ID: V (c) Superimposition of their 3D structures with RMSD of 5.541 angstrom, and (d) 1D sequences and 2D structures of the two tRNAs.

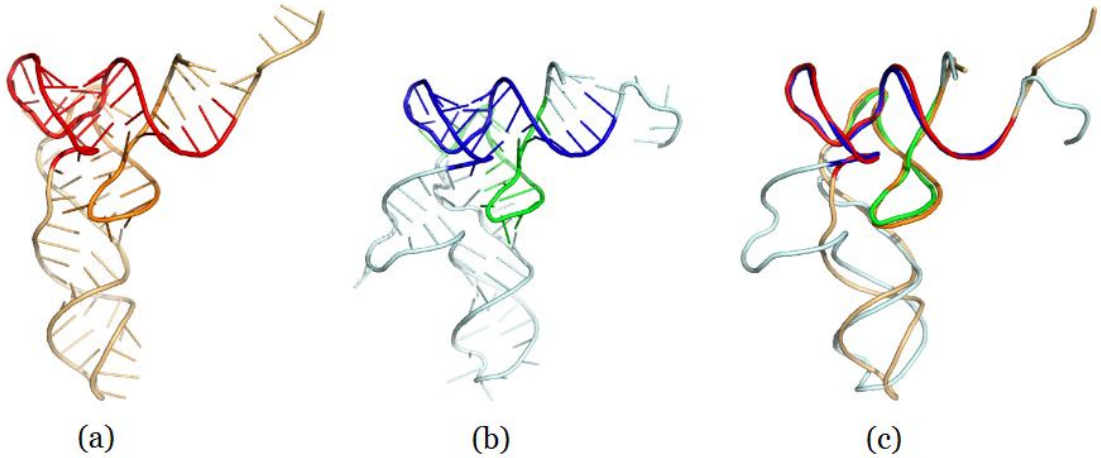


**Figure 1-2.** Two tRNAs: (a) PDB ID: 2J02, chain ID: V and (b) PDB ID: 1WZ2, chain ID: D, and (c) their two common similar substructures, one in orange and green colors (RMSD: 1.453 angstrom) and the other in red and blue colors (RMSD: 1.265 angstrom).

The basic idea behind our R3D-BLAST is as follows. First of all, we encode all the RNA 3D structures deposited in the PDB database as 1D sequences using the structural alphabet of 23 letters, which was obtained by using the two pseudo-torsion angles of RNA nucleotide backbones and the affinity propagation clustering approach. Next, we apply BLAST to searching for RNA sub-molecules whose 3D structures are similar to that of the query. Our experimental results have finally shown that our R3D-BLAST indeed has better performance than BLAST, a famous bioinformatics tool to find homologous proteins/RNAs only based on their sequence similarity, for identifying those RNA molecules whose tertiary substructures are locally similar to that of the query RNA, as well as FASTR3D for finding those RNAs whose structures are entirely similar to that of the query RNA.

# Chapter 2

# Materials and Methods

The basic idea we used to design our algorithm in this study is as follows. We first use the affinity propagation approach [16], an excellent method for clustering, to derive an RNA structural alphabet of 23 letters that represent distinct and most common backbone conformations of RNAs. According to this structural alphabet, we transform RNA 3D structures currently deposited in the PDB database into 1D sequences of SA-encoded letters. We then utilize BLAST [2] to search the collection of 1D SA-encoded sequences for RNA sub-molecules whose 3D structures are similar to substructures of the query RNA. In this chapter, we shall describe the details of (1) pseudo-torsion angles of RNAs, (2) how to use the affinity propagation approach to derive the structural alphabet and transform RNA 3D structures into 1D sequences, and (3) how to derive the substitution scoring matrix for aligning two 1D SA-encoded sequences, and (4) the details of our algorithm.

## 2.1 Pseudo-Torsion Angles and Ramachandran-like $\eta - \theta$ Plot

For protein backbones, two torsion angles ($\varphi$ and $\psi$) are sufficient to describe the backbone conformation of each amino acid residue. In contrast, RNA molecules have much higher dimensionality, since for each nucleotide residue there are six backbone torsion angles ($\alpha$, $\beta$, $\gamma$, $\delta$, $\varepsilon$ and $\zeta$) (see Figure 2-1a) and a torsion angle of the bond between base and ribose ring ($\chi$). This leads the analysis and classification of nucleotide conformation to be a high-dimensional problem

that is computationally intractable and cannot be evaluated visually. In addition, it is difficult to use these standard torsion angles to distinguish nucleotide conformations, because the so-called *crankshaft effect* usually results in that different combinations of stand torsion angles can describe identical nucleotide conformations. In fact, as was suggested by Duarte and Pyle [12], the pseudo-torsion angles ($\eta$ and $\theta$ as illustrated in Figure 2-1b) are at least as sensitive as standard torsion angles and even may be superior when specifying the backbone conformation of an individual nucleotide. Particularly, by representing the $\eta$ and $\theta$ pseudo-torsion angles of nucleotides on a 2D plot, one can obtain a Ramachandran-like diagram in which clusters of nucleotides appear at discrete regions and nucleotides in the same cluster have similar conformation [34, 12].

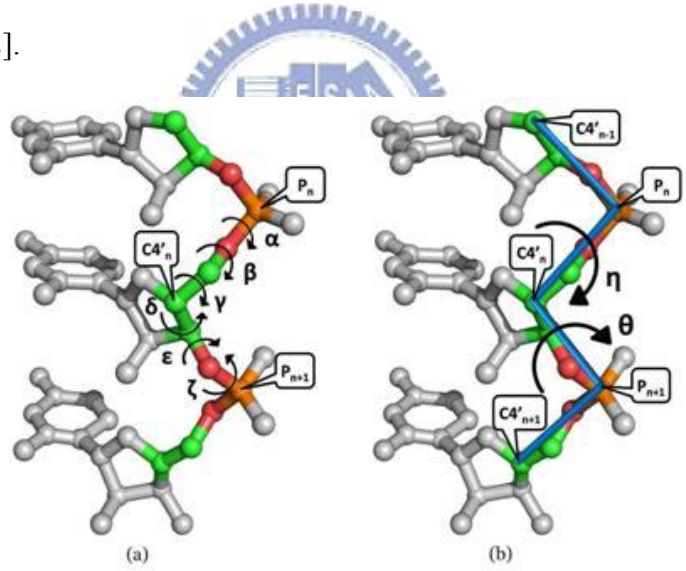



**Figure 2-1.** (a) Six standard backbone torsion angles of $\alpha$, $\beta$, $\gamma$, $\delta$, $\varepsilon$ and $\zeta$ and (b) two backbone pseudo-torsion angles of $\eta$ and $\theta$ for a nucleotide (denoted by n), where $\eta$ is defined by the atoms $C4'_{n-1}$, $P_n$, $C4'_n$ and $P_{n+1}$, while $\theta$ is defined by $P_n$, $C4'_n$, $P_{n+1}$ and $C4'_{n+1}$.

To depict this $\eta$–$\theta$ plot, we prepared a dataset that includes non-redundant crystal structures with minimum resolution of 3.0 ˚A from the PDB database [4]. This dataset finally contains 117

crystal RNA structures, particularly including 74 structures used by Wadley *et al*. [34], with 9,527 nucleotides in total. We then used AMIGOS that was developed by Duarte and Pyle [12] to calculate the $\eta$ and $\theta$ pseudo-torsion angles for all non-terminal nucleotides (9,267 nt in total) from all RNA molecules in the above dataset and plotted these calculated pseudo-torsion angles on the axes of a 2D plot as illustrated in Figure 2-2.

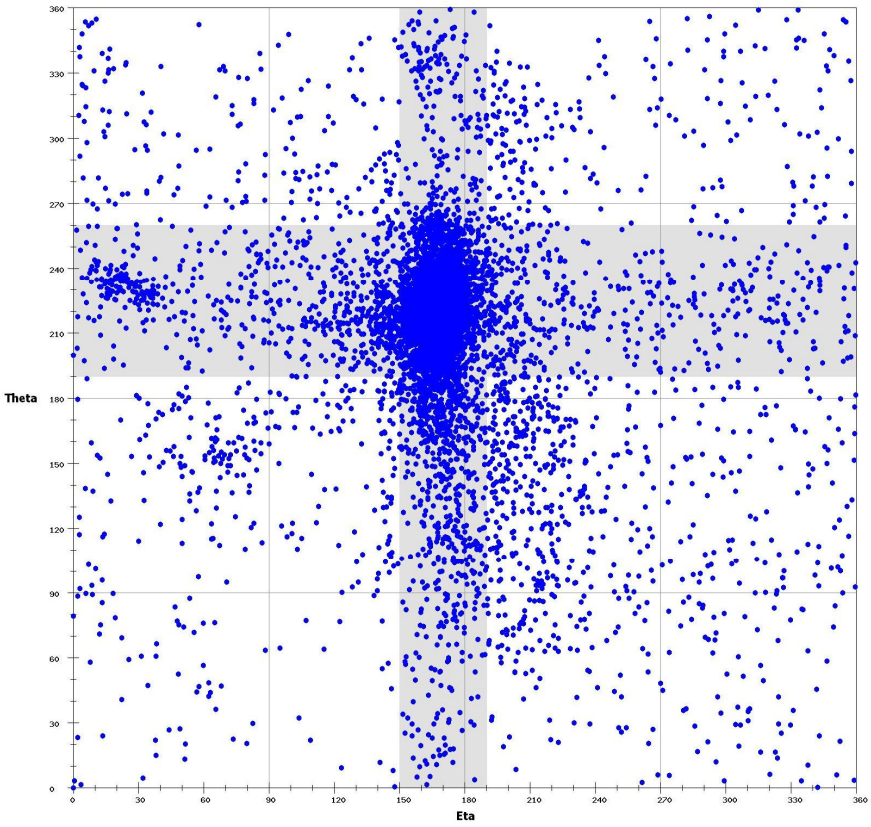



**Figure 2-2**. An $\eta$-$\theta$ plot of all non-terminal nucleotides from all RNA molecules in the dataset, where the intersection of the perpendicular gray regions ($150^{\circ} \leq \eta \leq 190^{\circ}$ and $190^{\circ} \leq \theta \leq 260^{\circ}$) is designated the helical region.

## 2.2 Affinity Propagation and Structural Alphabet

We here applied the so-called *affinity propagation* (AP) clustering algorithm, introduced by Frey and Dueck recently [16], to classify all the non-terminal nucleotides in our prepared dataset according to their $\eta$ and $\theta$ pseudo-torsion angles. Like *k*-means clustering algorithms, the VQ approaches usually find locally optimum clusters and are sensitive to outliers and noise [36], although it can be used to classify high dimensional data points. Besides, the VQ methods need to keep track of a fixed set of candidate centers (or exemplars) while searching for good solutions.

Basically, the AP algorithm is an *exemplar-based* clustering method for approximately solving the *exemplar learning problem* that aims to identify a set of data points as exemplars and assign every data point to an exemplar so as to maximize a fitness function, where notably the exemplar learning problem has been shown to be NP-hard [8]. Denote the input data points by $x_1$, $x_2$, . . . , $x_n$, the exemplar assigned to $x_i$ by $c_i$, and the similarity between $x_i$ and $c_i$ by $s(x_i, c_i)$. Then the *fitness function* mentioned above is defined to be $\sum_{i=1}^{n} s(x_i, c_i)$. Notably, if $x_i$ is an exemplar (*i.e.*, $c_i = x_i$), then this *fitness function* includes the term $s(x_i, c_i)$. Basically, the AP algorithm operates by simultaneously considering all input data points $x_1$, $x_2$, . . . , $x_n$ as potential exemplars and exchanging messages between data points until a good set of exemplars and clusters emerges. For simplicity, the similarity $(x_i, x_j)$ between two points $x_i$ and $x_j$ is also denoted as $s(i, j)$. In each iteration, two kinds of messages, called responsibility and availability, respectively, are exchanged between data points. The *responsibility* $r(i, k)$, which is sent from point $x_i$ to point $x_k$, indicates the accumulated evidence for how proper it would be for $x_k$ to serve as the exemplar of $x_i$ with taking into account other potential exemplars for $x_i$. Before being sent, the value of $r(i, k)$ is updated according to the following rule: $r(i, k) = s(i, k) - \max_{k' \text{ s.t. } k' \neq k}\{a(i, k') + s(i, k')\}$. The

*availability* $a(i, k)$, which is sent from point $x_k$ to point $x_i$, indicates the accumulated evidence for how proper it would be for $x_i$ to choose $x_k$ as its exemplar with taking into account the support from other points that $x_k$ should be an exemplar. The value of $a(i, k)$ is updated as follows: if $i \neq k$, then $a(i,k) = \min\left\{0, r(k,k) + \sum_{i' s.t. i' \notin i,k} \max\{0, r(i',k)\}\right\}$; otherwise, $a(k,k) = \sum_{i' s.t. i' \notin k} \max\{0, r(i',k)\}$. It should be noted that numerical oscillations may arise in some circumstances when updating the above two messages. To avoid such oscillations, therefore, each message is set to $\lambda$ times its value from the previous iteration plus $1 - \lambda$ times its currently prescribed updated value, where $\lambda$ is a *damping factor* whose value is between 0 and 1. In this study, we used a default damping factor of $\lambda = 0.5$. The above message-passing scheme is therefore referred to as *affinity propagation*. At any point during the affinity propagation, responsibilities and availabilities are combined to identify exemplars. That is, for data point $x_i$, the $k$ that maximizes $r(i, k)+a(i, k)$ indicates that $x_k$ is the exemplar of $x_i$. Finally, the message-passing procedure may be terminated after a fixed number of iterations (or after the changes in the messages fall below a threshold or the local decisions stay constant for some number of iterations).

Note that each data point in this study corresponds to a non-terminal nucleotide of an RNA 3D structure on the 2-dimensional $\eta$-$\theta$ plot and, therefore, the similarity between data point $x_i$ and its exemplar $c_i$ defined in this study is the negative squared Euclidean distance (that is, $s(x_i, c_i) = -\|x_i - c_i\|^2$), if $x_i \neq c_i$. As to $x_i = c_i$, the value of $s(x_i, x_i)$ represents the *a priori preference* for $x_i$ to serve as an exemplar and, therefore, it is not calculated in the same way as $s(x_i, x_k)$, where $x_i \neq x_k$, because it does not represent an assignment similarity. As suggested in [16], the preference values can be set to a global (shared) value, or customized for particular data

points. Particularly, moreover, high values of the preferences will cause the AP algorithm to find many exemplars (clusters), while low values will lead to a small number of exemplars. Here, we set a global value to $s(x_i, x_i)$ for all $1 \leq i \leq n$ such that a total of 9,267 non-terminal nucleotides on the $\eta$-$\theta$ plot is classified into 23 conformation clusters, as was illustrated in Figure 2-3. The 3D conformations of these 23 exemplar nucleotides are shown in Figure 2-4.

For our purpose of transforming RNA 3D structures into 1D sequences, we further assigned a letter to each of 23 clusters, as named in Table 2-1. We used the set of these 23 letters as a *structural alphabet* (SA) and encoded RNA 3D structures as 1D sequences of SA letters by using the *nearest neighbor rule*, by which each nucleotide in an RNA molecule is assigned with the letter of the cluster whose exemplar (center) is nearest to the nucleotide being encoded. In this study, we chose 23 as the number of the clusters on the $\eta$-$\theta$ plot based on the following two reasons. First, over 60% of nucleotides on the $\eta$-$\theta$ plot fall within the helical region (defined by the intersection of the two perpendicular gray regions in Figure 2-2). As illustrated in Figure 2-3, the helical region is partitioned into four clusters when $N$ = 23. However, if $N$ = 46, then an overpartitioning (with more than 10 clusters) in this helical region can be observed. This overpartitioning results was actually due to the fact that the helical region is so highly populated in the dataset of currently collected RNA structures that any clustering algorithm may tend to divide it into a lot of clusters. In fact, according to our experiments (data not shown), the value of the AUC obtained using our testing dataset with $N$ = 46 is not better than that with $N$ = 23. Second, choosing $N$ = 23 will allows one to apply BLAST, the most widely used tool of sequence homology search, for efficiently performing the structurally similar search on the database consisting of the SA-encoded sequences of RNA 3D structures.

**Table 2-1.** The structural alphabet of 23 conformational clusters classified by the AP algorithm with their associated letters and the $\eta$ and $\theta$ pseudo-torsion angles of their exemplars.

| Number | Letter | Pseudo-torsion angle | | Number | Letter | Pseudo-torsion angle | |
|---|---|---|---|---|---|---|---|
| | | $\eta$ | $\theta$ | | | $\eta$ | $\theta$ |
| 1 | A | 168.7 | 221.4 | 13 | M | 203.8 | 307.5 |
| 2 | B | 169.1 | 205.7 | 14 | N | 92.5 | 232.2 |
| 3 | C | 167.3 | 235.1 | 15 | Y | 69.6 | 153.8 |
| 4 | D | 163.7 | 257.1 | 16 | P | 310.6 | 220.1 |
| 5 | E | 169.4 | 179.5 | 17 | Q | 162.5 | 1.4 |
| 6 | F | 139.7 | 216.6 | 18 | R | 248.7 | 218.9 |
| 7 | G | 194.1 | 227.2 | 19 | S | 318.9 | 127.7 |
| 8 | H | 173.3 | 125.9 | 20 | T | 299.4 | 3.2 |
| 9 | I | 208.5 | 167.9 | 21 | Z | 88.3 | 292.5 |
| 10 | J | 23.1 | 228.9 | 22 | V | 48.3 | 52.5 |
| 11 | K | 229.4 | 104.9 | 23 | W | 5.9 | 314.3 |
| 12 | L | 179.8 | 71.4 | | | | |

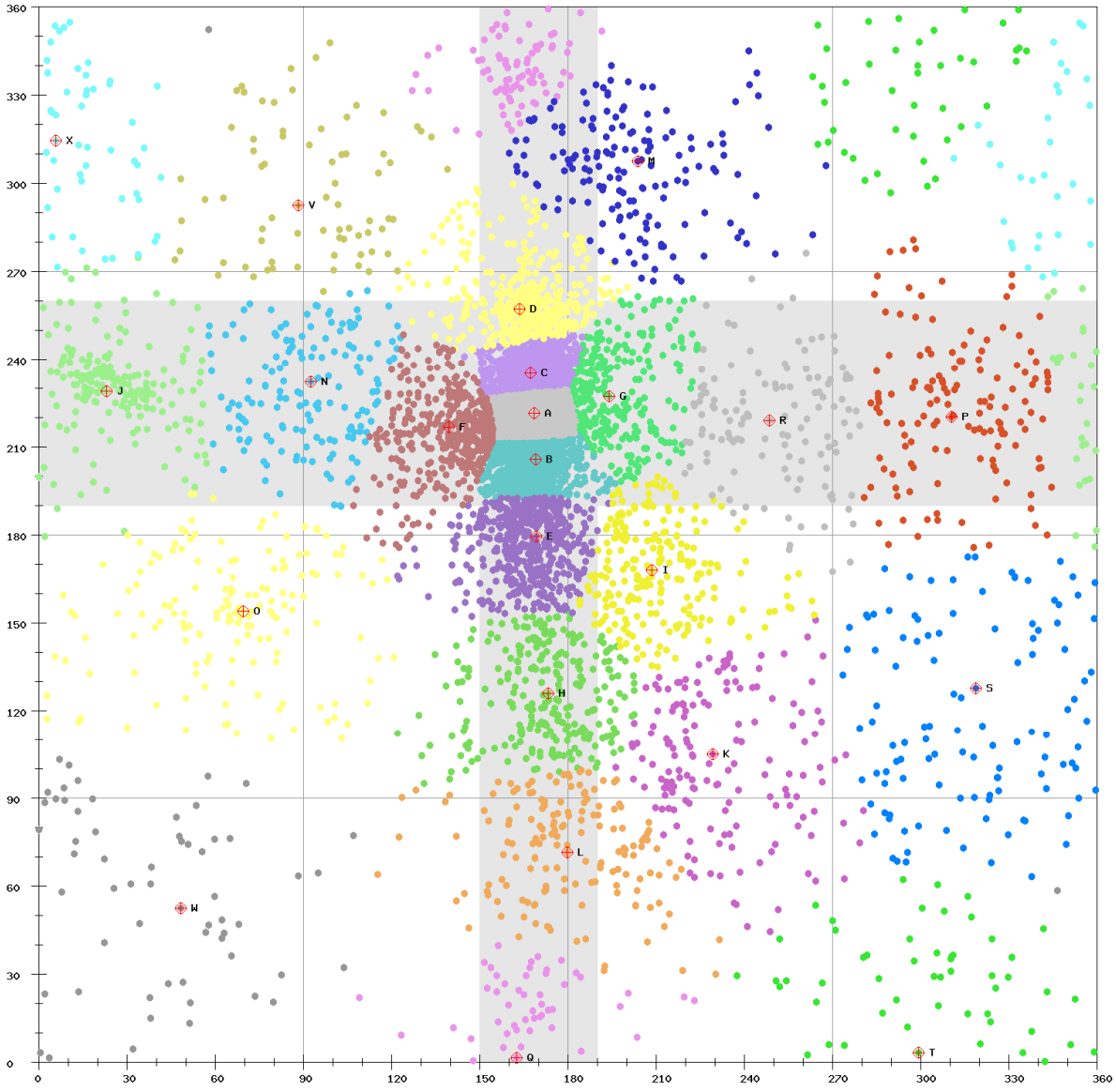



**Figure 2-3.** Twenty-three clusters classified by the AP algorithm.

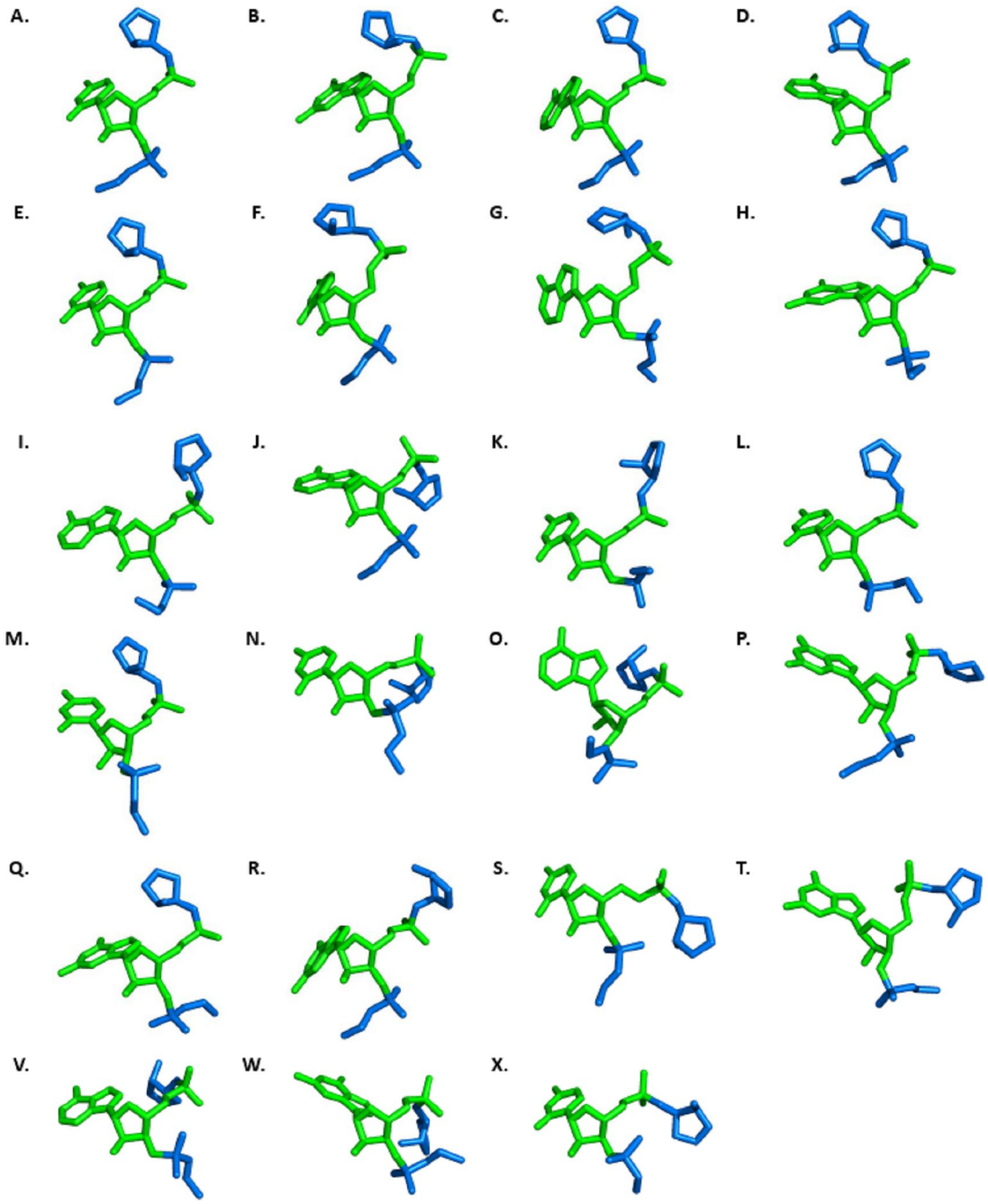
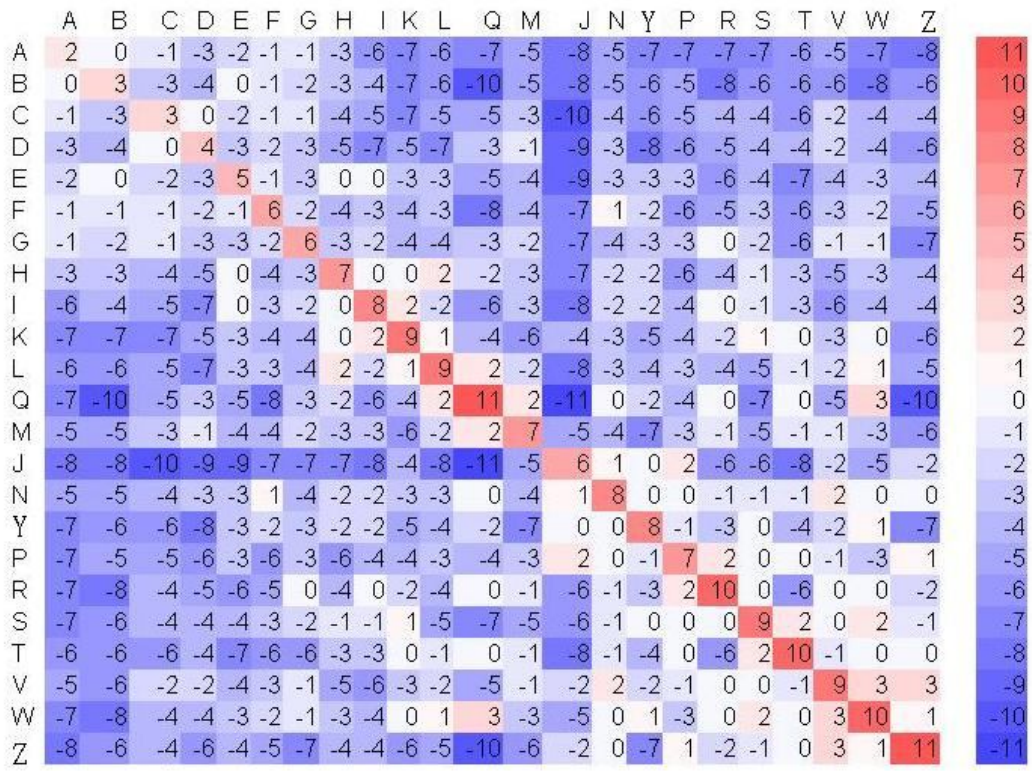



**Figure 2-4.** Three-dimensional conformations of 23 exemplar nucleotides, where the exemplar nucleotides are shown in green, whereas the portions of the previous and next nucleotides that affect the pseudo-torsions are shown in blue.

## 2.3 BLOSUM-Like Scoring Matrices

For the accuracy of aligning two SA-encoded sequences, we derived a 23 × 23 log-odds matrix for SA-letter substitution using the statistical method proposed by Henikoff and Henikoff [18]. Let $\{a_1, a_2, \ldots, a_{23}\}$ denote the structural alphabet of 23 SA letters and $f_{ij}$ be the observed substitution frequency of SA-letter pair ($a_i$, $a_j$ ). Then the relative frequency $q_{ij}$ of an SA-letter pair ($a_i$, $a_j$) is $q_{ij} = \frac{f_{ij}}{\sum_{k=1}^{23}\sum_{l=1}^{k} f_{kl}}$, and the frequency of occurrence of SA letter *ai* in an SA-letter pair ($a_i$, $a_j$) is $p_{ij} = q_{ii} + \frac{\sum_{k=1,k\neq i}^{23} q_{ik}}{2}$. The expected frequency $e_{ij}$ for a substitution between two SA-letters ($a_i$, $a_j$) is $p_ip_j$ for $i = j$ and $p_ip_j + p_jp_i = 2p_ip_j$ for $i \neq j$. The logarithm of the odds matrix is finally calculated by $score(a_i, a_j) = \lambda \log_2\left(\frac{q_{ij}}{e_{ij}}\right)$, where $\lambda$ is a positive scale factor. For the purpose of constructing this BLOSUM-like matrix, a dataset of structurally similar RNA pairs was obtained from the DARTS database [1], which used an automated method to classify 1,333 RNA tertiary structures into 244 groups of highly identical structures, and the SCOR database [21, 33], which organized many RNA structural motifs in a hierarchical classification system similar to the SCOP database for protein domains. From the initial dataset of 1,333 high-resolution RNA 3D structures, the DARTS database first selected 244 representative structures based on RNA sequence and 3D structure resemblances and then marked each of the remaining structures as either a highly identical structure or a highly identical fragment of a representative structure. A highly identical structure is defined as a structure that is globally almost identical (*i.e.*, with at least 90% sequence or 3D structure identity) to some other structure of similar size (*i.e.*, size ratio is between 1 and 1.5), while a highly identical fragment is defined as a structure that is almost

identical to only a small substructure of a larger structure (*i.e.*, size ratio is greater than 1.5). Note that 101 out of 244 representative structures have no highly identical structure. For our purpose, we used only the remaining 143 representative structures and their highly identical structures to construct our BLOSUM-like matrix. In addition, a set of structurally similar RNA motif pairs was obtained from the SCOR database based on the following criteria: (1) motifs must belong to a structural family, (2) motifs must have length greater than 3 nt, (3) motifs must have specified starting and ending positions in the chain, and (4) motif pairs must have no 100% sequence identity. In total, 3,391 RNA structural alignment pairs from 143 DARTS groups of 686 high-resolution RNA 3D structures and 430,628 RNA motif pairs from 334 SCOR classes of 6,220 structural motifs were analyzed, which together accounted for 8,500,322 SA-letter pairs. The $\lambda$ value used in this study was set to 1.6 for the best performance, by testing various values ranging from 1 to 2. Figure 2-5 illustrates the BLOSUM-like substitution matrix for the 23 SA-letters we derived in this study.



| | A | B | C | D | E | F | G | H | I | K | L | Q | M | J | N | Y | P | R | S | T | V | W | Z | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 2 | 0 | -1 | -3 | -2 | -1 | -1 | -3 | -6 | -7 | -6 | -7 | -5 | -8 | -5 | -7 | -7 | -7 | -7 | -6 | -5 | -7 | -8 | 11 |
| B | 0 | 3 | -3 | -4 | 0 | -1 | -2 | -3 | -4 | -7 | -6 | -10 | -5 | -8 | -5 | -6 | -5 | -8 | -6 | -6 | -6 | -8 | -6 | 10 |
| C | -1 | -3 | 3 | 0 | -2 | -1 | -1 | -4 | -5 | -7 | -5 | -5 | -3 | -10 | -4 | -6 | -5 | -4 | -4 | -6 | -2 | -4 | -4 | 9 |
| D | -3 | -4 | 0 | 4 | -3 | -2 | -3 | -5 | -7 | -5 | -7 | -3 | -1 | -9 | -3 | -8 | -6 | -5 | -4 | -4 | -2 | -4 | -6 | 8 |
| E | -2 | 0 | -2 | -3 | 5 | -1 | -3 | 0 | 0 | -3 | -3 | -5 | -4 | -9 | -3 | -3 | -3 | -6 | -4 | -7 | -4 | -3 | -4 | 7 |
| F | -1 | -1 | -1 | -2 | -1 | 6 | -2 | -4 | -3 | -4 | -3 | -8 | -4 | -7 | 1 | -2 | -6 | -5 | -3 | -6 | -3 | -2 | -5 | 6 |
| G | -1 | -2 | -1 | -3 | -3 | -2 | 6 | -3 | -2 | -4 | -4 | -3 | -2 | -7 | -4 | -3 | -3 | 0 | -2 | -6 | -1 | -1 | -7 | 5 |
| H | -3 | -3 | -4 | -5 | 0 | -4 | -3 | 7 | 0 | 0 | 2 | -2 | -3 | -7 | -2 | -2 | -6 | -4 | -1 | -3 | -5 | -3 | -4 | 4 |
| I | -6 | -4 | -5 | -7 | 0 | -3 | -2 | 0 | 8 | 2 | -2 | -6 | -3 | -8 | -2 | -2 | -4 | 0 | -1 | -3 | -6 | -4 | -4 | 3 |
| K | -7 | -7 | -7 | -5 | -3 | -4 | -4 | 0 | 2 | 9 | 1 | -4 | -6 | -4 | -3 | -5 | -4 | -2 | 1 | 0 | -3 | 0 | -6 | 2 |
| L | -6 | -6 | -5 | -7 | -3 | -3 | -4 | 2 | -2 | 1 | 9 | 2 | -2 | -8 | -3 | -4 | -3 | -4 | -5 | -1 | -2 | 1 | -5 | 1 |
| Q | -7 | -10 | -5 | -3 | -5 | -8 | -3 | -2 | -6 | -4 | 2 | 11 | 2 | -11 | 0 | -2 | -4 | 0 | -7 | 0 | -5 | 3 | -10 | 0 |
| M | -5 | -5 | -3 | -1 | -4 | -4 | -2 | -3 | -3 | -6 | -2 | 2 | 7 | -5 | -4 | -7 | -3 | -1 | -5 | -1 | -1 | -3 | -6 | -1 |
| J | -8 | -8 | -10 | -9 | -9 | -7 | -7 | -7 | -8 | -4 | -8 | -11 | -5 | 6 | 1 | 0 | 2 | -6 | -6 | -8 | -2 | -5 | -2 | -2 |
| N | -5 | -5 | -4 | -3 | -3 | 1 | -4 | -2 | -2 | -3 | -3 | 0 | -4 | 1 | 8 | 0 | 0 | -1 | -1 | -1 | 2 | 0 | 0 | -3 |
| Y | -7 | -6 | -6 | -8 | -3 | -2 | -3 | -2 | -2 | -5 | -4 | -2 | -7 | 0 | 0 | 8 | -1 | -3 | 0 | -4 | -2 | 1 | -7 | -4 |
| P | -7 | -5 | -5 | -6 | -3 | -6 | -3 | -6 | -4 | -4 | -3 | -4 | -3 | 2 | 0 | -1 | 7 | 2 | 0 | 0 | -1 | -3 | 1 | -5 |
| R | -7 | -8 | -4 | -5 | -6 | -5 | 0 | -4 | 0 | -2 | -4 | 0 | -1 | -6 | -1 | -3 | 2 | 10 | 0 | -6 | 0 | 0 | -2 | -6 |
| S | -7 | -6 | -4 | -4 | -4 | -3 | -2 | -1 | -1 | 1 | -5 | -7 | -5 | -6 | -1 | 0 | 0 | 0 | 9 | 2 | 0 | 2 | -1 | -7 |
| T | -6 | -6 | -6 | -4 | -7 | -6 | -6 | -3 | -3 | 0 | -1 | 0 | -1 | -8 | -1 | -4 | 0 | -6 | 2 | 10 | -1 | 0 | 0 | -8 |
| V | -5 | -6 | -2 | -2 | -4 | -3 | -1 | -5 | -6 | -3 | -2 | -5 | -1 | -2 | 2 | -2 | -1 | 0 | 0 | -1 | 9 | 3 | 3 | -9 |
| W | -7 | -8 | -4 | -4 | -3 | -2 | -1 | -3 | -4 | 0 | 1 | 3 | -3 | -5 | 0 | 1 | -3 | 0 | 2 | 0 | 3 | 10 | 1 | -10 |
| Z | -8 | -6 | -4 | -6 | -4 | -5 | -7 | -4 | -4 | -6 | -5 | -10 | -6 | -2 | 0 | -7 | 1 | -2 | -1 | 0 | 3 | 1 | 11 | -11 |

**Figure 2-5.** The BLOSUM-like substitution matrix for the 23 SA-letters we derived in this study.

## 2.4 Implementation of R3D-BLAST

Our R3D-BLAST was implemented based on a structural-alphabet approach whose procedure flowchart, as shown in Figure 2-6, consists of three major procedures. The first procedure is a preprocessing job to derive the tertiary structure information (*i.e.*, pseudo-torsion angles η and θ) of all the RNAs in the PDB database, where the η and θ values were derived using the AMIGOS program [12]. We then encode each RNA 3D structure as a 1D sequence of structural-alphabet (SA) letters according to its pseudo-torsion angles, and store these SA-encoded sequences in a local database. Similarly, the second procedure is to encode the RNA queried by the user as a 1D sequence of SA-letters. Notice that the user currently can query our R3D-BLAST by a PDB or NDB code of an RNA tertiary structure optionally with specified residue range. The third procedure is to use BLAST to search the local database of SA-encoded sequences for those RNA molecules whose tertiary substructures are locally similar to that of the query RNA. One of the benefits of using BLAST is that it can provide each identified RNA an *E*-value to show its statistical significance. Basically, the *E*-value is defined as $E = Kmne^{-\lambda S}$, indicating the expected number of HSPs (High Scoring Pairs) with score at least *S* by chance, where *m* is the number of letters in the query RNA and *n* is the number of letters in the database, and *K* and $\lambda$ are two constants related to the searching space (*i.e.*, $m \times n$) and the scoring matrix, respectively. Since the database we maintain in the R3D-BLAST and the used BLOSUM-like scoring matrix are different from those originally used in BLAST, we need to re-estimate *K* and $\lambda$ values so that our R3D-BLAST can return correct E-values. Here, we utilize the island method that was proposed by Olsen *et al.* [26] for estimating our $\lambda$ and *K* values according to different set of gap open and extension penalties, whose results are shown in the Table 2-2. In addition, the user can decide on

whether or not to calculate the RMSD between the query RNA and each of identified RNAs, which may cost our R3D-BLAST a few minutes to finish its jobs. It is inevitable that some RNAs identified in this step may not similar to the query RNA in tertiary structure. Therefore, we design a filter based on a geometric match measure, called *structural alignment score* (SAS), to further screen out those RNAs that are structurally non-similar to the query RNA, when their SAS score with respect to the query RNA is greater than a pre-defined threshold, where SAS=100×RMSD/(number of aligned residues) and its default value is 25 in our R3D-BLAST.

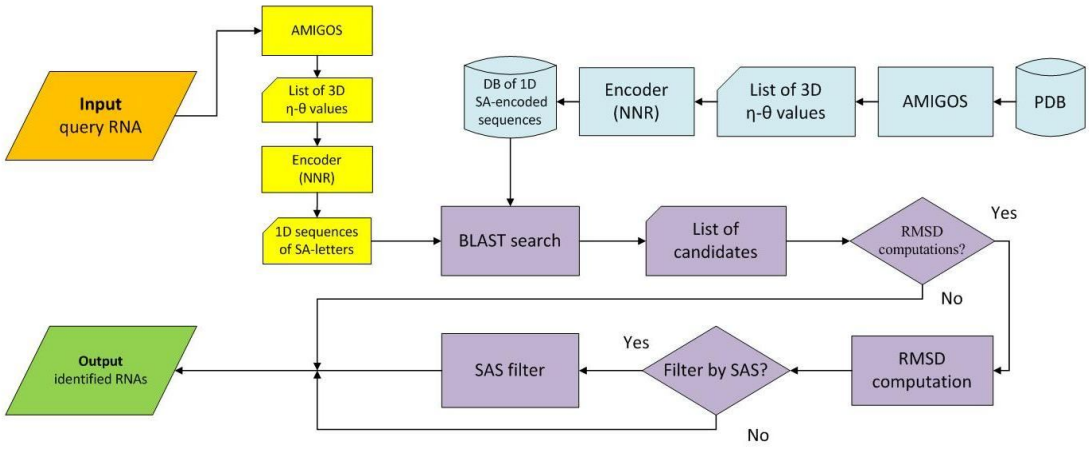



**Figure 2-6.** The procedure flowchart of R3D-BLAST

**Table 2-2.** The λ and *K* values for different sets of gap open and extension penalties.

| Open: Extension Penalty | λ | K |
|---|---|---|
| 4:1 | 0.236 | 0.009 |
| 5:1 | 0.326 | 0.041 |
| 6:1 | 0.372 | 0.079 |

| | | |
|---|---|---|
| 4:2 | 0.379 | 0.086 |
| 5:2 | 0.402 | 0.125 |
| 6:2 | 0.414 | 0.145 |

# Chapter 3

# Usage of Software Tool

Based on the structural-alphabet approach described in the previous chapter, we have developed an easy-to-operate web server, named R3D-BLAST that allows biologists to fast and accurately search the PDB database for those RNA molecules whose 3D substructures are locally similar to that of the query RNA. In the following, we shall describe the details of how to use R3D-BLAST.

## 3.1 Input of R3D-BLAST

R3D-BLAST provides an intuitive and easy-to-operate interface (see Figure 3-1) that can be freely accessed at http://bioalgorithm.life.nctu.edu.tw/R3D-BLAST/. allows user to search similar RNA 3D substructures in the database by BLAST. Below, we describe the details of its input usage step by step.

1. Enter the PDB/NDB ID (4-/6-character code) of an RNA molecule (or upload its file in the PDB format), as well as its chain ID and starting and ending residue numbers in sequence. Note that PDB/NDB ID or uploading the file is mandatory, and others are optional but the user has to specify a chain ID, if the input RNA molecule has multiple chains.

2. If user would like run the R3D-BLAST with default parameters, just have to click the "Run R3D-BLAST" button; otherwise, the user continues with the following steps of

modifying default parameter settings.

3. Choose a different set of gap open and extension penalties, where R3D-BLAST penalizes the gaps using the affine gap penalty function. Currently, R3D-BLAST provides the user six different sets of gap open and extension penalties that are 4-1, 4-2, 5-1, 5-2, 6-1, and 6-2 (default).

4. Decide on whether or not to calculate the RMSD between the query RNA and each of identified RNAs (default no). If yes, also decide on whether or not to perform SAS filter and, if needed, also enter a positive integer to serve as the cutoff of SAS score (whose default is 25).

5. Modify the predefined threshold of *E*-value (whose default is 5) if needed.

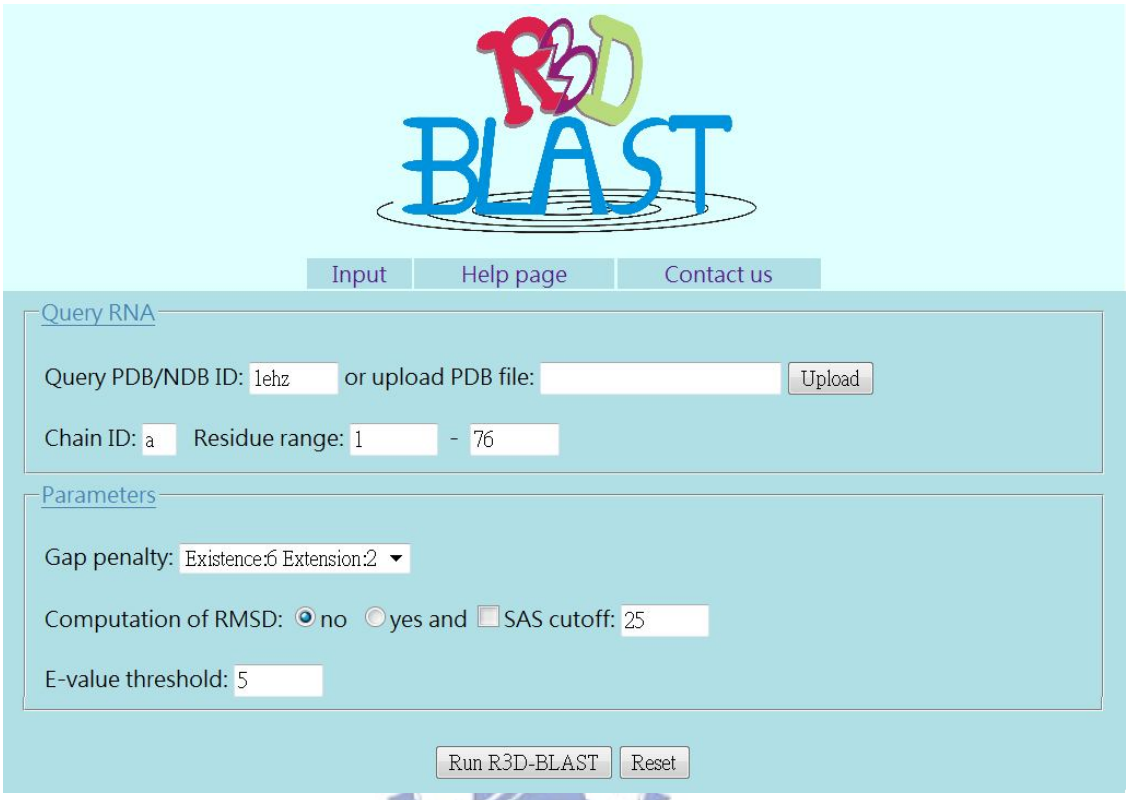



**Figure 3-1.** Interface of R3D-BLAST

## 3.2 Output of R3D-BLAST

In the output page, R3D-BLAST will first display the information about input RNA molecule and user-specified parameters. Next, it will show a detailed list of identified structurally similar RNAs (see Figure 3-2), including corresponding PDB ID, chain ID, starting and ending residue numbers of aligned region and its length, title of the RNA in the PDB ID file, classification of the RNA (based on function, metabolic role, molecule type, cellular location, and so on), experimental method used to determine the structure of the RNA, the resolution of the solved RNA, and released data of the RNA in the PDB file, *E*-value, and RMSD (Root Mean Square Deviation) measured with respect to the query RNA and its corresponding SAS score and structural superposition. Particularly, in the display of the structural superposition, the user can visually view, rotate and enlarge the 3D structures of both the query and target RNA molecules and their structural superposition in a Jmol window (see Figure 3-3). Notice that in the top panel of the Jmol window, R3D-BLAST provides the user some useful functions for displaying RNA molecules. For example, the user can choose either black or white (default) as window background color, spin RNA molecules or not (default), display RNA molecules in a scheme of either ribbon, cartoon (default), wireframe or trace, determine whether to display nucleotide IDs or not (default), and download the PDB files for the query and target RNAs and their superimposed structure. In addition, R3D-BLAST allows the user to save the search result in the CSV or text format for later process.

| No. | Details of RNAs identified by R3D-BLAST | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | PDB ID: 1EHZ | Chain ID: A | Subject:Query Range (Length): 1-76:1-76 (76:76) | | Title: THE CRYSTAL STRUCTURE OF YEAST PHENYLALANINE TRNA AT 1 93 A RESOLUTION | | | |
| | E-value: 5E-48 | RMSD: 0.000 | SAS: 0 | Jmol: Jmol3D | Class: RNA | Method: X-RAY DIFFRACTION | Å: 1.93 | Release Date: 02-OCT-00 |
| 2 | PDB ID: 1EVV | Chain ID: A | Subject:Query Range (Length): 3-76:3-76 (74:74) | | Title: CRYSTAL STRUCTURE OF YEAST PHENYLALANINE TRANSFER RNA AT 2 0 A RESOLUTION | | | |
| | E-value: 2E-35 | RMSD: 1.191 | SAS: 1.6 | Jmol: Jmol3D | Class: RNA | Method: X-RAY DIFFRACTION | Å: 2.00 | Release Date: 01-MAY-00 |
| 3 | PDB ID: 1PNS | Chain ID: V | Subject:Query Range (Length): 3-72:3-72 (70:70) | | Title: CRYSTAL STRUCTURE OF A STREPTOMYCIN DEPENDENT RIBOSOME FROM E COLI, 30S SUBUNIT OF 70S RIBOSOME THIS FILE, 1PNS, CONTAINS THE 30S SUBUNIT, TWO TRNAS, AND ONE MRNA MOLECULE THE 50S RIBOSOMAL SUBUNIT IS IN FILE 1PNU | | | |
| | E-value: 8E-34 | RMSD: 1.549 | SAS: 2.2 | Jmol: Jmol3D | Class: RIBOSOME | Method: X-RAY DIFFRACTION | Å: 8.70 | Release Date: 15-JUL-03 |
| 4 | PDB ID: 1GIX | Chain ID: C | Subject:Query Range (Length): 3-72:3-72 (70:70) | | Title: CRYSTAL STRUCTURE OF THE RIBOSOME AT 5 5 A RESOLUTION THIS FILE, 1GIX, CONTAINS THE 30S RIBOSOME SUBUNIT, THREE TRNA, AND MRNA MOLECULES 50S RIBOSOME SUBUNIT IS IN THE FILE 1GIY | | | |
| | E-value: 8E-34 | RMSD: 1.562 | SAS: 2.2 | Jmol: Jmol3D | Class: RIBOSOME | Method: X-RAY DIFFRACTION | Å: 5.50 | Release Date: 04-MAY-01 |

**Figure 3-2.** Partial display of the R3D-BLAST result when queried with a tRNA (PDB ID: 1EHZ, chain ID: A, residue range: 1-76).

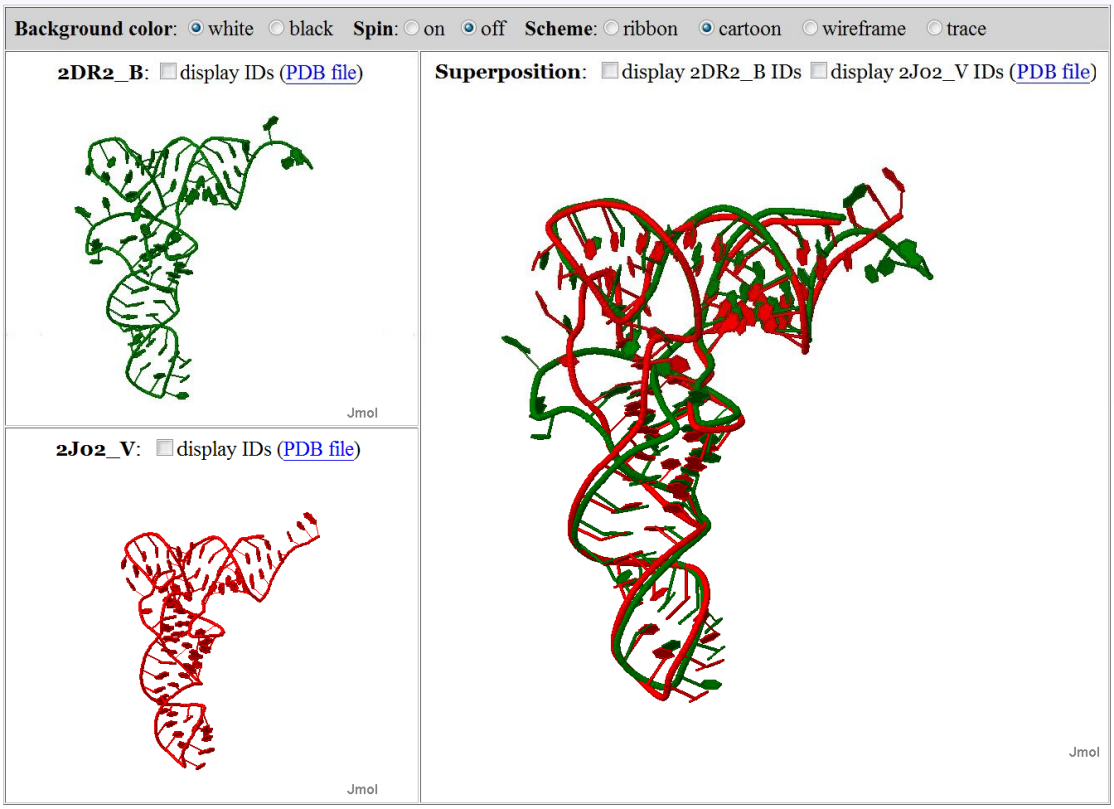



**Figure 3-3.** Visual display of the tertiary structures of two input tRNA molecules and their superposition.

# Chapter 4

# Results and Discussions

In this chapter, we shall test our R3D-BLAST on some RNA 3D structures and also compare its experimental results to those obtained by other available tools, such as BLAST and FASTR3D. Unless specified, all the experiments were carried out using these three tools with their default parameters.

## 4.1 Comparison with BLAST

In fact, the sequences of two RNA sub-molecules can diverge greatly, even while they have a similar tertiary structure. For example, the RNA substructures as shown in Figures 4-1a and 4-1b are highly similar, because the RMSD of their superimposition, as shown in Figure 4-1c, is 1.535 angstrom. However, these two RNA substructures belong to two different RNA molecules and their sequence percentage identity is 36% only, as illustrated in Figure 4-2. This is because that the 3D structures of RNA molecules are more evolutionarily conserved than their sequences. Therefore, it can be expected that our R3D-BLAST can search the PDB database for more RNAs whose whole structures or substructure are similar to those of the query RNA, as compared to BLAST. To demonstrate this, we tested our R3D-BLAST, as well as BLAST on some tRNAs, whose results are shown in Table 4-1. Consequently, the average number of RNA

(a) (b) (c)

**Figure 4-1.** Tertiary structures of two RNA sub-molecules: (a) PDB ID: 1HR2, chain ID: A, residue range: 103-260 and (b) PDB ID: 1U6P, chain ID: B, residue range: 280-301, and (c) their superimposition

```
1HR2  103
      GUCUCAGGGGAAACUUUGAGAU 260
            ** *    **** * *** *   *
1U6P  280
      GUACUAGUUGAGAAACUAGCUC 301
```

**Figure 4-2.** Sequence alignment of two RNA sub-molecules whose percentage identity is about 36%.

sub-molecules, whose 3D structures are similar to those of the query RNA, identified by R3D-BLAST are much larger than that returned by BLAST. Particularly, when queried with PDB IDs 1WZ2 and 2DU6, our R3D-BLAST searched for 37 and 16, respectively, structurally similar RNA sub-molecules with at least 50% query coverage, while BLAST found only 3 and 4, respectively, homologous RNA sub-sequences with at least 50% query coverage, where the *query coverage* is defined as the percent of the query length that is included in the alignment.

**Table 4-1.** Comparison of experimental results between R3D-BLAST and BLAST for some tRNA molecules.

| PDB:Chain:Residue range | R3D-BLAST | | BLAST | |
|---|---|---|---|---|
| | 50-100% | 0-49% | 50-100% | 0-49% |
| 1EHZ:A:1-76 | 194 | 149 | 77 | 14 |
| 1WZ2:D:902-988 | 37 | 246 | 3 | 37 |
| 2DU6:D:902-971 | 16 | 111 | 4 | 2 |

## 4.2 Comparison with FASTR3D

In this section, we further evaluated our R3D-BLAST by testing it on some RNA molecules and also comparing its results with those obtained by FASTR3D. FASTR3D is a tool also developed by our laboratory based on the information of RNA secondary structures for identifying structural similarities for a query of RNA molecule in the PDB database. As demonstrated in [23], FASTR3D can serve as a useful tool that allows biologists to fast and accurately search the PDB database for structurally similar RNAs. However, FASTR3D can find only those RNAs whose secondary structures exactly match with that of the query RNA and, therefore, it cannot search for those structurally similar RNAs whose secondary structures are just approximately equal to that of the query RNA, or even those RNAs that have only substructures similar to those of the query RNA.

The experimental results we obtained by testing R3D-BLAST and FASTR3D on five different kinds of RNA molecules with different length are shown in Table 4-2. For some RNAs, such as riboswitch and tRNA, FASTR3D found more candidates, whose tertiary structures entirely similar to those of the query RNAs For some RNAs, such as riboswitch and tRNA, FASTR3D found more candidates, whose tertiary structures entirely similar to those of the query RNAs (i.e., the query coverage is 100%), than R3D-BLAST. The main reason is that FASTR3D was designed to search for those RNAs whose whole structures are globally similar to that of the query, but R3D-BLAST was designed to search for those RNAs whose substructures are locally similar to that of the query. Actually, those RNAs that were found by FASTR3D but not by R3D-BLAST were still can be identified by R3D-BLAST, when the query coverage is set to at

least 90%. On the other hand, there are a lot of structurally similar RNA sub-molecules (i.e., with at least 90% query coverage) that still were able to identified by R3D-BLAST, but not by FASTR3D, as illustrated in Table 4-2.

**Table 4-2.** Comparison of experimental results between R3D-BLAST and FASTR3D for five RNA molecules with length from 46 to 1530 bp**.**

| RNA | PDB:Chain:Length | R3D-BLAST | | FASTR3D |
|---|---|---|---|---|
| | | 100% | At least 90% | 100% |
| Riboswitch | 1Y27:X:46 | 6 | 20 | 10 |
| Pseudoknot | 1YMO:A:47 | 1 | 2 | 1 |
| tRNA | 1EHZ:A:76 | 10 | 133 | 17 |
| Ribozyme | 1HR2:A:157 | 1 | 7 | 1 |
| 16S rRNA | 1J5E:A:1530 | 4 | 74 | 0 |

# Chapter 5

# Conclusion

In this chapter, we have developed a bioinformatics tool R3D-BLAST that allows biologists to fast and accurately search the PDB databases for those RNAs that have substructures similar to that of the query RNA. The basic idea behind our R3D-BLAST is as follows. We first encoded all the RNA 3D structures deposited in the PDB database as 1D sequences using the structural alphabet of 23 letters, which was obtained by using the two pseudo-torsion angles of RNA nucleotide backbones and the affinity propagation clustering approach. We then applied BLAST to searching for RNA sub-molecules whose 3D structures are similar to that of the query. Our experimental results have also demonstrated that our R3D-BLAST indeed has better performance than BLAST f for identifying those RNA molecules whose tertiary substructures are locally similar to that of the query RNA, as well as FASTR3D for finding those RNAs whose structures are entirely similar to that of the query RNA. Therefore, we believe that our R3D-BLAST can serve as a useful bioinformatics tool in the study of structural biology.

# References


[1] Abraham,M., Dror,O., Nussinov,R., Wolfson,H.J. (2008) Analysis and classification of RNA tertiary structures. *RNA*, 14, 2274-2289.

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

[3] Berman,H.M., Westbrook,J., Feng,Z., Iype,L., Schneider,B. and Zardecki,C. (2002) The nucleic acid database. *Acta Crystallogr. Biol. Crystallogr.*, 58, 889–898.

[4] Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Research*, 28, 235–242.

[5] Capriotti,E. and Marti-Renom,M.A. (2008) RNA structure alignment by a unit-vector approach. *Bioinformatics*, 24, i112–i118.

[6] Capriotti,E. and Marti-Renom,M.A. (2009) SARA: a server for function annotation of RNA structures. *Nucleic Acids Research*, 37, W260–W265.

[7] Chang,Y.F., Huang,Y.L. and Lu,C.L. (2008) SARSA: a web tool for structural alignment of RNA using a structural alphabet. *Nucleic Acids Research*, 36, W19–W24.

[8] Charikar,M., Guha,S., Tardos,E. and Shmoys,D.B. (2002) A constant-factor approximation algorithm for the *k*-median problem. *Journal of Computer and System Sciences*, 65,

129–149.

[9] Doudna,J.A. (2000) Structural genomics of RNA. *Nat. Struct. Biol.,* 7, 954–956.

[10] Dror,O., Nussinov,R. and Wolfson,H.J. (2005) ARTS: alignment of RNA tertiary structures. *Bioinformatics*, 21, 47–53.

[11] Dror,O., Nussinov,R. and Wolfson,H.J. (2006) The ARTS web server for aligning RNA tertiary structures. *Nucleic Acids Research*, 34, W412–W415.

[12] Duarte,C.M., Pyle,A.M. (1998) Stepping through an RNA structure: A novel approach to conformational analysis. *Journal of Molecular Biology*, 284, 1465-1478.

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

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

[15] Ferre`,F., Ponty,Y., Lorenz,W.A. and Clote,P. (2007) DIAL: a web server for the pairwise alignment of two RNA three-dimensional structures using nucleotide, dihedral angle and base-pairing similarities. *Nucleic Acids Research*, 35, W659–W668.

[16] Frey,B.J. and Dueck,D. (2007) Clustering by passing messages between data points. *Science*, 315, 972-976.

[17] Griffiths-Jones,S., Saini,H.K., van Dongen,S. and Enright,A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Research*, 36, D154–D158.

[18] Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89, 10915–10919.

[19] He,S., Liu,C., Skogerbo,G., Zhao,H., Wang,J., Liu,T., Bai,B., Zhao,Y. and Chen,R. (2008) NONCODE v2.0: decoding the non-coding. *Nucleic Acids Research,* 36, D170–D172.

[20] Kin,T., Yamada,K., Terai,G., Okida,H., Yoshinari,Y., Ono,Y., Kojima,A., Kimura,Y., Komori,T. and Asai,K. (2007) fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Research,* 35, D145–D148.

[21] Klosterman, P.S., Tamura, M., Holbrook, S.R. and Brenner, S.E. (2002) SCOR: a structural classification of RNA database. *Nucleic Acids Research*, 30, 392–394.

[22] Kolodny,R. and Linial,N. (2004) Approximate protein structural lignment in polynomial time. *Proc. Natl Acad. Sci. USA,* 101, 12201–12206.

[23] Lai,C.E., Tsai,M.Y., Liu,Y.C., Wang,C.W., Chen,K.T., Lu,C.L.,(2009) FASTR3D: a fast and accurate search tool for similar RNA 3D structures. *Nucleic Acids Research*, 37, W287-W295.

[24] Macke,T.J., Ecker,D.J., Gutell,R.R., Gautheret,D., Case,D.A. and Sampath,R. (2001)

RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Research,* 29, 4724–4735.

[25] Mattick,J.S. and Makunin,I.V. (2006) Non-coding RNA. *Hum. Mol. Genet.,* 15, R17–R29.

[26] Olsen,R., Bundschuh,R. and Hwa,T. (1999) Rapid asseseement of extremal statistics for gapped local alignment. In *Proceeding of the Seventh International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, 1999, 211-222

[27] Pang,K.C., Stephen,S., Dinger,M.E., Engstrom,P.G., Lenhard,B. and Mattick,J.S. (2007) RNAdb 2.0–an expanded database of mammalian non-coding RNAs. *Nucleic Acids Research*, 35, D178–D182.

[28] Popenda,M., Blazewicz,M., Szachniuk,M. and Adamiak,R.W. (2008) RNA FRABASE version 1.0: an engine with a database to search for the three-dimensional fragments within RNA structures. *Nucleic Acids Research*, 36, D386–D391.

[29] Popenda,M., Szachniuk,M., Blazewicz M., Wasik,S. KBurke E, Blazewicz,J. and Adamiak,R.W. (2010) RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. *BMC Bioinformatics*, 11, 231-241.

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

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

[32] Szymanski,M., Erdmann,V.A. and Barciszewski,J. (2007) Noncoding RNAs database (ncRNAdb). *Nucleic Acids Research,* 35, D162–D164.

[33] Tamura, M., Hendrix, D.K., Klosterman, P.S., Schimmelman, N.R., Brenner, S.E. and Holbrook, S.R. (2004) SCOR: structural classification of RNA, version 2.0. *Nucleic Acids Research*, 32, D182–D184.

[34] Wadley,L.M., Keating,K.S., Duarte,C.M., Pyle,A.M. (2007) Evaluating and learning from RNA pseudotorsional space: quantitative validation of a reduced representation for RNA structure. *Journal of Molecular Biology*, 372,942-957.

[35] Wang,C.W., Chen,K.T. and Lu,C.L. (2010) IPARTS: an improved tool of pairwise alignment of RNA tertiary structures. *Nucleic Acids Research,* 38, W340-W347.

[36] Xu,R. and Wunsch,D., I. (2005) Survey of clustering algorithms. *IEEE Transactions on Neural Networks,* 16, 645–678.