國立交通大學

生物資訊及系統生物研究所

碩士論文

RNA結構的功能設定

Function assignment of RNA structures

研 究 生：陳昆澤
指導教授：盧錦隆 博士

中華民國 九十九 年 六 月

RNA結構的功能設定

Function assignment of RNA structures

研究生：陳昆澤　　Student：Kun-Tze Chen

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，生物學家對 ncRNA 這種不會轉譯成蛋白質的 RNA 分子是愈來愈有興趣了，因為他們在細胞內扮演著許多重要的角色，包括基因的調控、RNA 的修改、與染色體的複製等等。一般而言，RNA 分子的結構在演化上通常比其序列還來得保守，因此分析 RNA 分子的結構相似程度將有助於生物學家對其功能的了解。然而比較兩個 RNA 三級結構的相似度是一件困難的工作，因為它已被證明是 NP-hard 的問題。先前我們的實驗室已經利用一個啟發式的方法發展了一名為 iPARTS 的結構比對工具可以讓生物學家快速且準確地比較出兩個 RNA 三級結構的相似程度。這個啟發式方法的主要精神如下：首先我們利用 RNA 核苷酸的兩個假扭轉角η及θ畫出一張類似 Ramachandran 的平面圖表。接著我們利用所謂的親合性互動(Affinity Propagation)分群演算法來對在η-θ平面圖上的 RNA 核苷酸進行分群並得到一組含有 23 個核苷酸結構的字元集。最後我們用這個結構字元集將兩個輸入的 RNA 三級結構編碼成兩條由結構字元所組成的一級序列，然後再利用傳統的序列比對演算法去比對這兩條結構字元編碼的一級序列以決定出原 RNA 三級結構之間的相似程度。在這本論文中，我們進一步地在 iPARTS 身上添加一個新的功能來幫助生物學家準確地找出一個 RNA 三級結構的功能。為了這個目的，我們首先利用上述的結構字元集將要查詢的 RNA 三級結構與一個事先準備好的資料庫中所有已知功能的 RNA 三級結構編碼成一級的結構字元序列，然後再利用 iPARTS 去比較出要查詢的 RNA 三級結構與資料庫中每一個已知功能的 RNA 三級結構的整體相似程度，最後根據結構最相似的 RNA 來設定要查詢 RNA 三級結構的功能。最後的實驗結果顯示出我們的 iPARTS 在設定 RNA 三級結構的功能上確實比另外一個類似的工具 SARA 來得優秀，其中 SARA 是利用所謂的單位向量來比較出兩個 RNA 三級結構的相似程度。

Abstract

In recent years, there is a fast growing interest in noncoding RNAs (ncRNAs) whose transcripts are not translated into proteins, because they play essential roles in many cellular processes, such as gene regulation, RNA modification and chromosome replication. Typically, structures of RNA molecules are more evolutionarily conserved than their sequences and, therefore, the analysis of the RNAs on the structure level can be helpful for biologists to understand their functions. However, detecting structural similarities in two RNA molecules at tertiary structure level is a difficult job, because it has been shown to be an NP-hard problem. Previously, our laboratory have used a heuristic approach to develop a useful tool, called iPARTS, which allows biologists to fast and accurately compare the structural similarity of two RNA tertiary structures. The basic idea of this heuristic approach is as follows. First, we derived a Ramachandran-like diagram of RNAs by plotting the pseudo-torsion angles η and θ of their nucleotides on a two-dimensional (2D) axis. Next, we applied the so-called affinity propagation clustering algorithm to this η-θ plot to obtain a structural alphabet (SA) of 23 nucleotide conformations. Finally, we used this SA to encode RNA three-dimensional (3D) structures into one-dimensional (1D) sequences of SA letters and then applied traditional algorithms of sequence alignments to these 1D SA-encoded sequences for determining the structural similarities between two given RNA 3D structures. In this study, we have further equipped our iPARTS with a new function that is able to help biologists to accurately find the function of a given RNA 3D structure. For this purpose, we first utilize the above SA to encode the query and all the RNA 3D structures with known function in a pre-prepared database into 1D SA-encoded sequences, then use iPARTS to compare the globally structural similarity between the query RNA and each of the RNAs with known functions in the database, and finally assign the annotated function of the most structurally similar RNA to the query RNA 3D structure. Consequently, our experimental results have shown that our iPARTS indeed is superior to a similar tool, named SARA that uses the so-called unit-vector approach to align two RNA 3D structures, when assigning the functions of the RNA 3D structures.

Acknowledgement

來交大這短短的兩年讓我從自以為是的大學生變成了對一切事物都應保有嚴謹態度的研究生，這都得感謝於我的指導教授，盧錦隆教授。能將一件深奧的事物以簡單卻嚴謹並且生活化地描述出來原來是一大學問，更是一種藝術，這樣的態度必能讓我受用無窮。

感謝我的家人讓我衣食無缺的地在這樣的環境下學習，以及他們時常的關心。謝謝拖鞋學姐能不厭其煩地讓我問她問題，讓我知道報帳的各種技巧。謝謝明原學長，常常問我正在寫什麼程式幫我解救當下的窘境。謝謝之前畢業的勞學長，教導我伺服器的相關知識尤其是架設 FTP 的好用之處。更謝謝他在聚餐時常常會幫我多叫很多碗飯，或者是幫我偷加料，增長了我很多體脂肪。謝謝忠翰學長時常餵我草食動物才能吃的食物以及讓我把他可愛的玩偶擺出各種奇怪的姿勢。謝謝志偉學長留了很多可以讓我受益良多的程式給我，以及資料夾名稱不要亂取。謝謝慶恩學長讓我知道人是可以很樂觀的，但千萬不要把論文初稿揉爛丟垃圾桶。謝謝智先學長時常對我的關心及問候，讓我知道怎麼樣才是愛台灣。謝謝拖弟讓我知道他真的是為了部落的好。謝謝芸蓁讓我知道原來勳章有多重要，還有不要挖洞給自己跳。接著謝謝學弟旦旦，讓我知道原來查某人不是嘴頰甩兩下就能安撫。謝謝學弟昱全時常幫我抓龍以及幫我處理雜務，並且讓我知道食物不能亂吃，還有讓我知道烏烏茲拉確實很大聲。謝謝仁駿學弟讓我知道錢不好賺，身體確實要顧。謝謝 Thomas 教了我很多廣東話，讓我知道香港人都是很 nice 的。謝謝我的女朋友意靜，謝謝妳對我的關心及體諒。最後謝謝過去曾幫助過我的朋友。

Contents

List of tables

List of figures

Chapter 1

Introduction

These years, there is a fast growing interest in noncoding RNAs (ncRNAs) whose transcripts are not translated into proteins, because they play essential roles in many cellular processes, such as gene regulation, RNA modification and chromosome replication [9,13,25,30]. 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 [17], RNAdb [27], miRBase [16], fRNAdb [21] 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.

As both the number and the size of RNA tertiary 3D structures deposited in the database continue to grow, the techniques ofRNA structure comparison have become an increasingly crucial bioinformatics tool because structures of molecules evolve more slowly than their sequences and, therefore, their structural comparison can bring more significant insights into their functions and even evolutionary relationships that would not be detected by analyzing sequence information alone. Basically, detecting structural similarities in two RNA 3D molecules is not an easy problem 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 [24]. Due to this reason, currently available software tools for comparing two RNA 3D structures, such as ARTS [10,11], DIAL [14], PARTS [7], iPARTS [35], SARA [5,6] and LaJolla [3] are all based on some heuristic approaches.

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 quadrats (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 [14], the structural alignments produced by ARTS may be incorrect sometimes.

Later on, to overcome the inaccurate problems caused by ARTS, Ferré *et al.* [14] 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 (or pseudo-dihedral) and torsion (or dihedral) 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 [i.e. 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 end 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 [26], semiglobal [28], local [29] and normalized local [2] alignments) to compare these 1D SA-encoded sequences for determining their structural similarities. As was demonstrated in [7], the structural alignments by PARTS were comparable to those by DIAL, but the running time of PARTS was generally faster than that of DIAL. 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.

More recently, Bauer et al. [3] have used a hashing algorithm to develop a tool, called LaJolla, which can perform structural alignment of two RNA 3D structures. LaJolla first translates each of input RNA 3D structures into a 1D sequence of characters according to backbone pseudo-torsion angles of constituent residues, with one of these two 1D sequences being considered as query RNA and the other as target RNA. Next, it stores all *n*-grams (i.e., substrings of length *n*) of the target RNA in a hash table and searches each of all *n*-grams of the query RNA against the hash table for its occurrences in the target RNA. Finally, all corresponding *n*-grams between the query and target RNAs are aligned to determine their anchors and a superposition of these anchors are then performed.

According to the previous study of our iPARTS, we observed that the structural similarity between two given RNA 3D structures obtained by our iPARTS were comparable with those returned by SARA. In this thesis, therefore, we try to further study whether or not our iPARTS can be used to assign the functions of RNA 3D structures that are comparable with or more accurate than those predicted by SARA. For this purpose, we first utilize the above SA to encode the query and all the RNA

3D structures with known function in a pre-prepared database into 1D SA-encoded sequences, then use iPARTS to compare the globally structural similarity between the query RNA and each of the RNAs with known functions in the database, and finally assign the annotated function of the most structurally similar RNA to the query RNA 3D structure. Consequently, our experimental results have shown that our iPARTS indeed is superior to a similar tool, named SARA that uses the so-called unit-vector approach to align two RNA 3D structures, when assigning the functions of the RNA 3D structures.

Chapter 2

Materials and Methods

The basic idea of our iPARTS algorithm is to reduce input RNA 3D structures to 1D sequences of SA letters. First we will derive a Ramachandran-like diagram of RNAs by plotting nucleotides on a 2D axis using their two pseudo-torsion angles η and θ. Then, we will apply the affinity propagation (AP) clustering algorithm to this η-θ plot to obtain an SA of 23-nt conformations. Next, we will use this SA to transform RNA 3D structures into 1D sequences of SA letters and continue to use algorithms of classical sequence alignments added with a substitution matrix to compare these 1D SA-encoded sequences. Finally, we will assign the annotated function of the most structurally similar RNA to the query RNA 3D structure.

2.1 Pseudo-torsion angles and Ramachandran-like η-θ plot

For proteins, two torsion angles (ϕ and ψ) are sufficient to describe the backbone conformation of each amino acid. In contrast, RNA molecules have much higher dimensionality, since six standard torsion angles (α, β, γ, δ, ε and ζ as shown in Figure 2-1a) are needed to specify the backbone conformation of a single nucleotide. 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 important nucleotide conformations in RNA structural motifs, because the so-called ‘crankshaft effect’, in which large changes in individual torsion angles are compensated by changes in other torsion angles, usually leads to a result that different combinations of standard torsion angles can describe identical nucleotide conformations [34]. In fact, as was suggested by Duarte and Pyle [12], the pseudo-torsion angles (η and θ 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.

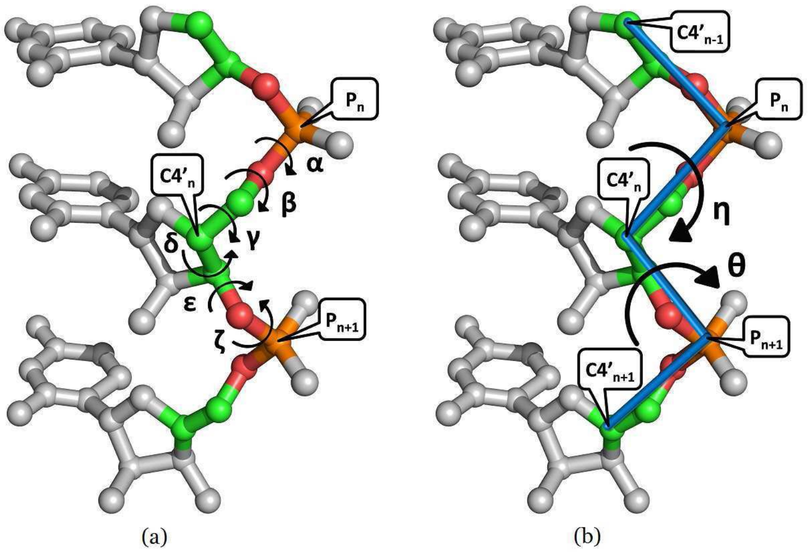


Figure 2-1. (a) Six standard backbone torsion angles of α, β, γ, δ, ε and ζ and (b) two backbone pseudo-torsion angles of η and θ for a nucleotide (denoted by n), where η is defined by the atoms C4′$_{n-1}$, P$_n$, C4′$_n$ and P$_{n+1}$, while θ is defined by P$_n$, C4′$_n$, P$_{n+1}$ and C4′$_{n+1}$.

Particularly, by representing the η and θ 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 [12,34]. Therefore, in this study, we aim to develop a novel SA for RNA 3D structures using their η-θ plot of pseudo-torsion angles, rather than using four standard torsion angles (α, γ, δ and ζ) as done in our previous work of PARTS [7] that was motivated from the works by Hershkovitz *et al.*[19,20].

As mentioned, the 2D η-θ plot is a Ramachandran-like diagram that can provide us a graphic representation of quantitatively distinct structural features for analyzing and modeling RNA 3D structures [12,34]. To depict this η-θ plot, we prepared a data set that includes non-redundant crystal structures with minimum resolution of 3.0 Å from the PDB database [4]. This data set finally contains 117 crystal RNA structures with 9527 nt in total. Next, we used the AMIGOS program developed by Duarte and Pyle [12] to calculate the η and θ pseudo-torsion angles for all non-terminal nucleotides (9267 nt in total) from all RNA molecules in the above data set and plotted these calculated pseudo-torsion angles on the axes of a 2D plot as illustrated in Figure 2-2. We then continued to use the AP clustering algorithm [15] to classify all the non-terminal nucleotides in this η-θ plot into 23 conformation clusters, each of which was further assigned a letter.

2.2 Affinity Propagation and Structural Alphabet

Instead of using the vector quantization (VQ) approach as done in our previous work [7], we here applied the so-called *affinity propagation* (AP) clustering algorithm, introduced by Frey and Dueck recently [15], to classify all the non-terminal nucleotides in our prepared dataset according to their η and θ pseudo-torsion angles. Like *k*-means clustering algorithms, the VQ approaches usually find locally optimum clusters and are sensitive to outliers and noise [36],

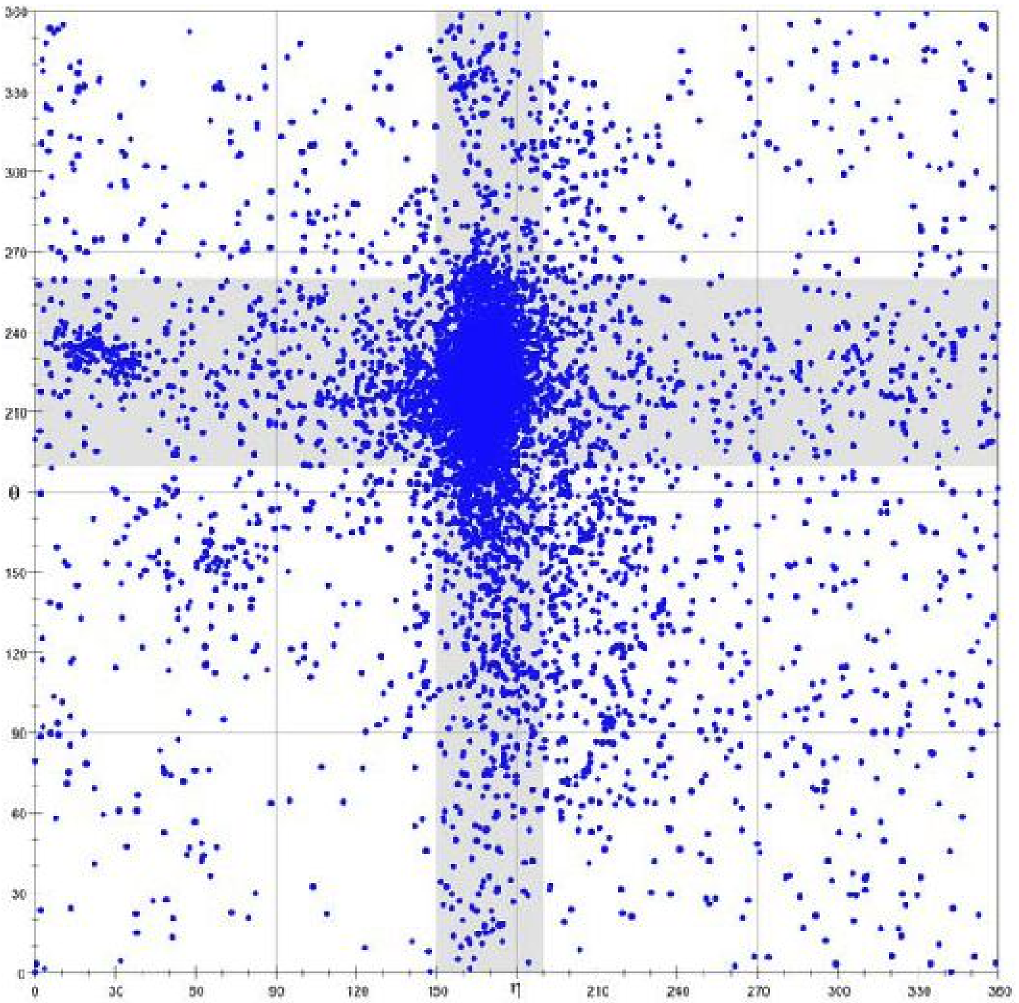


Figure 2-2. An η-θ plot of all non-terminal nucleotides from all RNA molecules in the dataset, where the intersection of the perpendicular gray regions (150° ≤ η ≤ 190° and 190° ≤ θ ≤ 260°) is designated the helical region.

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 the fitness function includes the term $s(x_i, c_i)$.

Basically, the AP algorithm operates by simultaneously considering all input data points $x_1, x_2, \ldots, x_n$ as potential exemplars and exchanging messages between data points until a good set of exemplars and clusters emerges. For simplicity, the similarity $s(x_i, x_j)$ 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* $r(i, k)$ that 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:

$$r(i,k) = s(i,k) - \max_{k':k'\neq k}\{\mathrm{a}(i,k') + \mathrm{s}(i,k')\}$$

The *availability* $a(i, k)$ that 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. For $i \neq k$ the value of $a(i, k)$ is updated as follows:

$$a(i,k) = \min\left\{0, r(k,k) + \sum_{i':i'\notin\{i,k\}} \max\{0, r(i',k\ )\}\right\}$$

; otherwise,

$$a(k,k) = \sum_{i':i'\notin\ k} \max\{0, r(i',k)\}$$

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 λ times its value from the previous iteration plus 1 − λ times its currently prescribed updated value, where λ is a *damping factor* whose value is between 0 and 1. In this study, we used a default damping factor of λ = 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 η-θ 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 [15], 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 η-θ 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. The reasons for choosing the certain number of clusters (i.e., $N = 23$) will be described in Chapter 4.

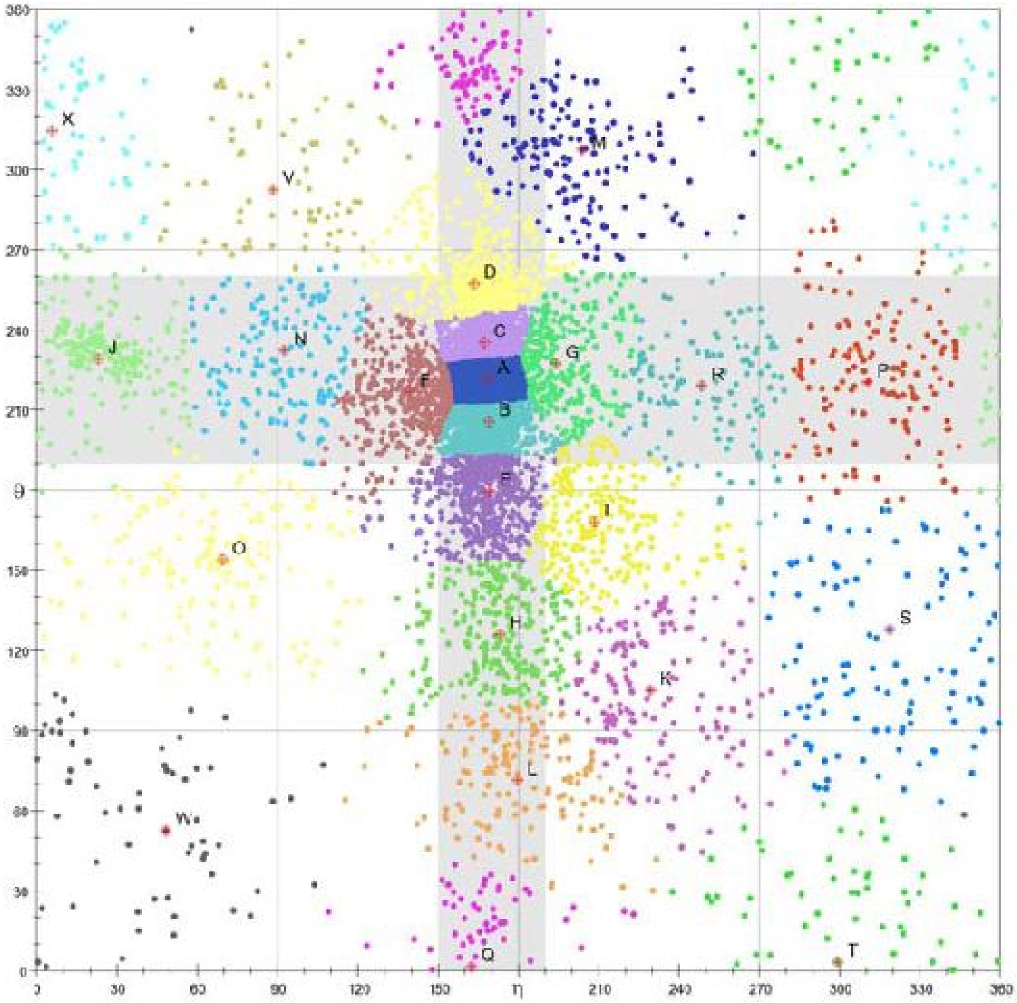


Figure 2-3. Twenty-three clusters classified by the AP algorithm.

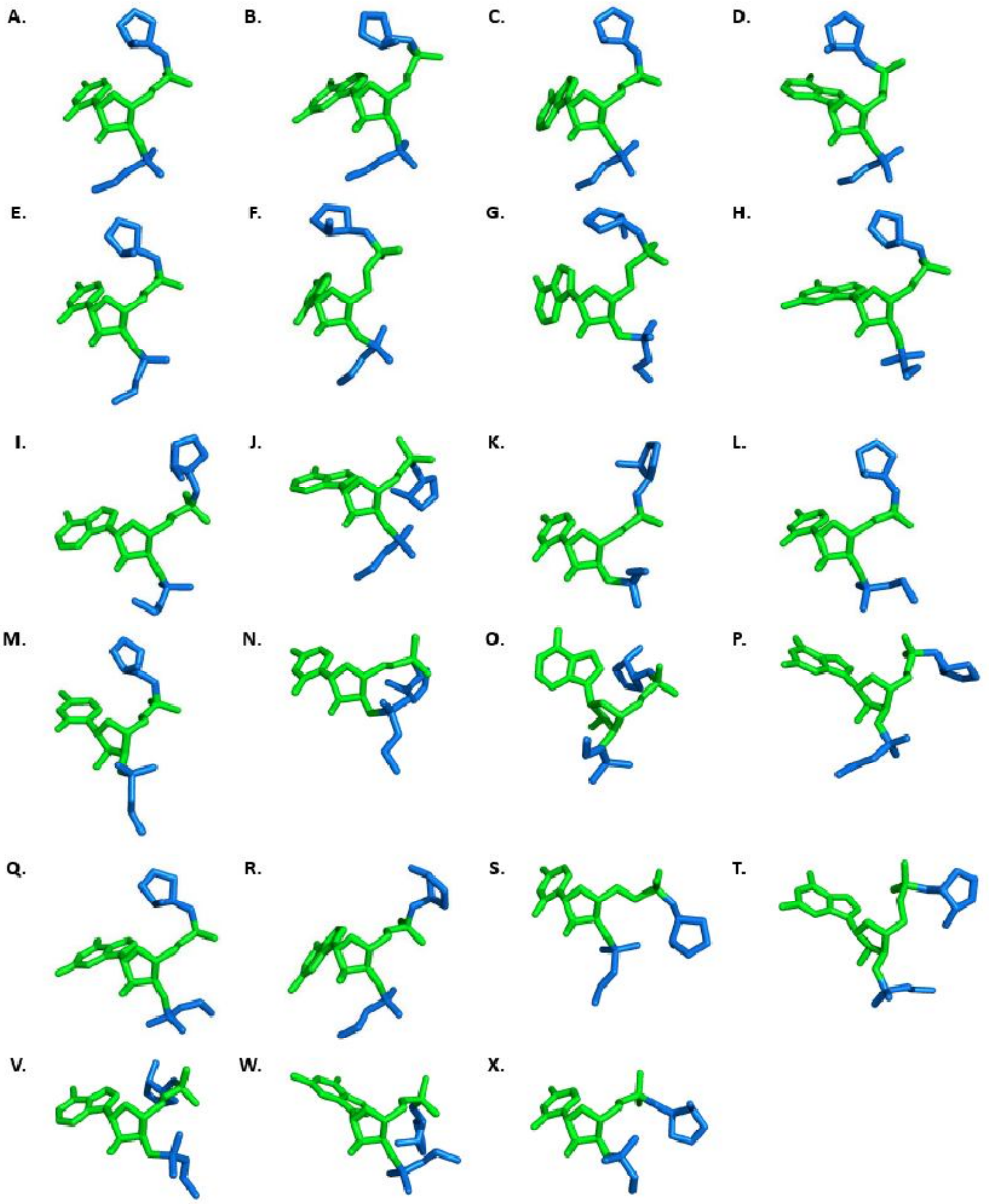


Figure 2-4. Three-dimensional conformations of 23 exemplar nucleotides, where the exemplar nucleotides are shown in green, whereas the portions of the previous and next nucleotides that affect the pseudo-torsions are shown in blue.

Table 2-1. The structural alphabet of 23 conformational clusters classified by the AP algorithm with their associated letters and the η and θ pseudo-torsion angles of their exemplars.

No.	Letter	(η, θ)	No.	Letter	(η, θ)	No.	Letter	(η, θ)
1	A	(168.7, 221.4)	9	I	(208.5, 167.9)	17	Q	(162.5, 1.4)
2	B	(169.1, 205.7)	10	J	(23.1, 228.9)	18	R	(248.7, 218.9)
3	C	(167.3, 235.1)	11	K	(229.4, 104.9)	19	S	(318.9, 127.7)
4	D	(163.7, 257.1)	12	L	(179.8, 71.4)	20	T	(299.4, 3.2)
5	E	(169.4, 179.5)	13	M	(203.8, 307.5)	21	V	(88.3, 292.5)
6	F	(139.7, 216.6)	14	N	(92.5, 232.2)	22	W	(48.3, 52.5)
7	G	(194.1, 227.2)	15	O	(69.6, 153.8)	23	X	(5.9, 314.3)
8	H	(173.3, 125.9)	16	P	(310.6, 220.1)			

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 $\{a_1, a_2, ..., a_{23}\}$ 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 a_i in an SA-letter pair (a_i, a_j) is:

$$p_i = q_{ii} + \frac{\sum_{k=1,k\neq i}^{23} q_{ik}}{2}$$

The expected frequency e_{ij} for a substitution between two SA-letter (a_i, a_j) is then $p_i p_j$ for $i = j$ and $p_i p_j + p_j p_i = 2p_i p_j$ for $i \neq j$. The logarithm of the odds matrix is finally calculated by:

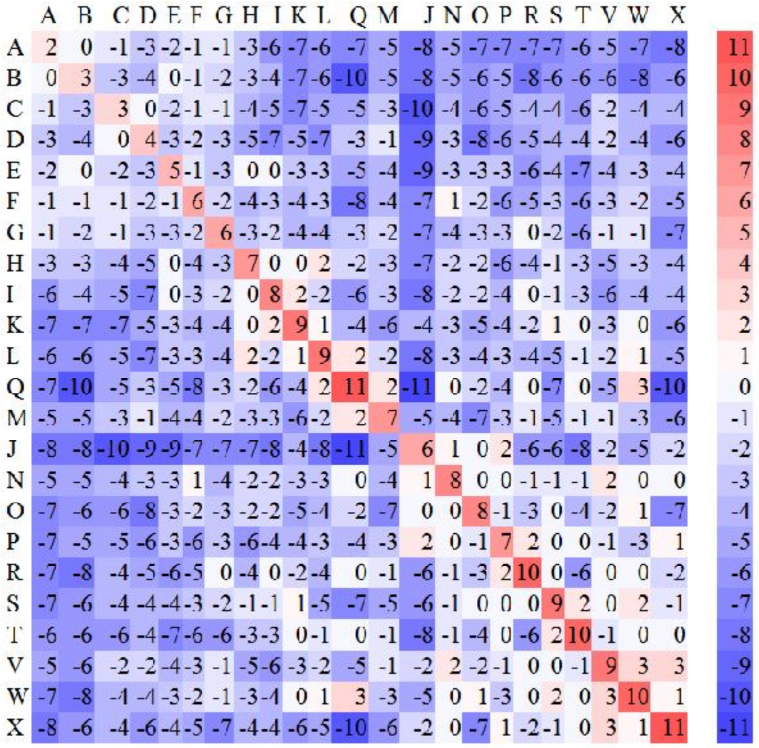
$$score(a_i, a_j) = \lambda log_2\left(\frac{q_{ij}}{e_{ij}}\right)$$

where λ 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 [22,33], 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[1] 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

[1] The size ratio here is defined as the number of nucleotides of the bigger structure divided by the number of nucleotides of the smaller structure

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 λ value used in this study was set to 1.6 for the best performance, by testing various values ranging from 1 to 2. Figure 2-6 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	O	P	R	S	T	V	W	X
A	2	0	-1	-3	-2	-1	-1	-3	-6	-7	-6	-7	-5	-8	-5	-7	-7	-7	-7	-6	-5	-7	-8
B	0	3	-3	-4	0	-1	-2	-3	-4	-7	-6	-10	-5	-8	-5	-6	-5	-8	-6	-6	-6	-8	-6
C	-1	-3	3	0	-2	-1	-1	-4	-5	-7	-5	-5	-3	-10	-4	-6	-5	-4	-4	-6	-2	-4	-4
D	-3	-4	0	4	-3	-2	-3	-5	-7	-5	-7	-3	-1	-9	-3	-8	-6	-5	-4	-4	-2	-4	-6
E	-2	0	-2	-3	5	-1	-3	0	0	-3	-3	-5	-4	-9	-3	-3	-3	-6	-4	-7	-4	-3	-4
F	-1	-1	-1	-2	-1	6	-2	-4	-3	-4	-3	-8	-4	-7	1	-2	-6	-5	-3	-6	-3	-2	-5
G	-1	-2	-1	-3	-3	-2	6	-3	-2	-4	-4	-3	-2	-7	-4	-3	-3	0	-2	-6	-1	-1	-7
H	-3	-3	-4	-5	0	-4	-3	7	0	0	2	-2	-3	-7	-2	-2	-6	-4	-1	-3	-5	-3	-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
K	-7	-7	-7	-5	-3	-4	-4	0	2	9	1	-4	-6	-4	-3	-5	-4	-2	1	0	-3	0	-6
L	-6	-6	-5	-7	-3	-3	-4	2	-2	1	9	2	-2	-8	-3	-4	-3	-4	-5	-1	-2	1	-5
Q	-7	-10	-5	-3	-5	-8	-3	-2	-6	-4	2	11	2	-11	0	-2	-4	0	-7	0	-5	3	-10
M	-5	-5	-3	-1	-4	-4	-2	-3	-3	-6	-2	2	7	-5	-4	-7	-3	-1	-5	-1	-1	-3	-6
J	-8	-8	-10	-9	-9	-7	-7	-7	-8	-4	-8	-11	-5	6	1	0	2	-6	-6	-8	-2	-5	-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
O	-7	-6	-6	-8	-3	-2	-3	-2	-2	-5	-4	-2	-7	0	0	8	-1	-3	0	-4	-2	1	-7
P	-7	-5	-5	-6	-3	-6	-3	-6	-4	-4	-3	-4	-3	2	0	-1	7	2	0	0	-1	-3	1
R	-7	-8	-4	-5	-6	-5	0	-4	0	-2	-4	0	-1	-6	-1	-3	2	10	0	-6	0	0	-2
S	-7	-6	-4	-4	-4	-3	-2	-1	-1	1	-5	-7	-5	-6	-1	0	0	0	9	2	0	2	-1
T	-6	-6	-6	-4	-7	-6	-6	-3	-3	0	-1	0	-1	-8	-1	-4	0	-6	2	10	-1	0	0
V	-5	-6	-2	-2	-4	-3	-1	-5	-6	-3	-2	-5	-1	-2	2	-2	-1	0	0	-1	9	3	3
W	-7	-8	-4	-4	-3	-2	-1	-3	-4	0	1	3	-3	-5	0	1	-3	0	2	0	3	10	1
X	-8	-6	-4	-6	-4	-5	-7	-4	-4	-6	-5	-10	-6	-2	0	-7	1	-2	-1	0	3	1	11

Figure 2-5. BLOSUM-like scoring matrix for the 23 SA-letters.

2.4 Alignment for SA-encoded sequences and function assignment

We first utilize the above SA to encode the query and all the RNA 3D structures with known function in a pre-prepared database into 1D SA-encoded sequences, then use iPARTS to compare the globally structural similarity between the query RNA and each of the RNAs with known functions in the database, and finally assign the annotated function of the most structurally similar RNA to the query RNA 3D structure. Notice that a grid-like search procedure was performed to optimize the default parameters of open and extension gap penalties, which are −6 and −1, respectively, used in iPARTS by varying the open gap penalty from −15 to −1 in steps of 1 and the extension gap penalty from −3 to −0.5 in steps of 0.5.

Chapter 3

Implementation of Software Tool

Based on the SA-based approach described in the previous chapter, we have further equipped iPARTS with a function that can perform function assignment of RNA tertiary structures. In the following, we will describe the details of how to use this new function of iPARTS.

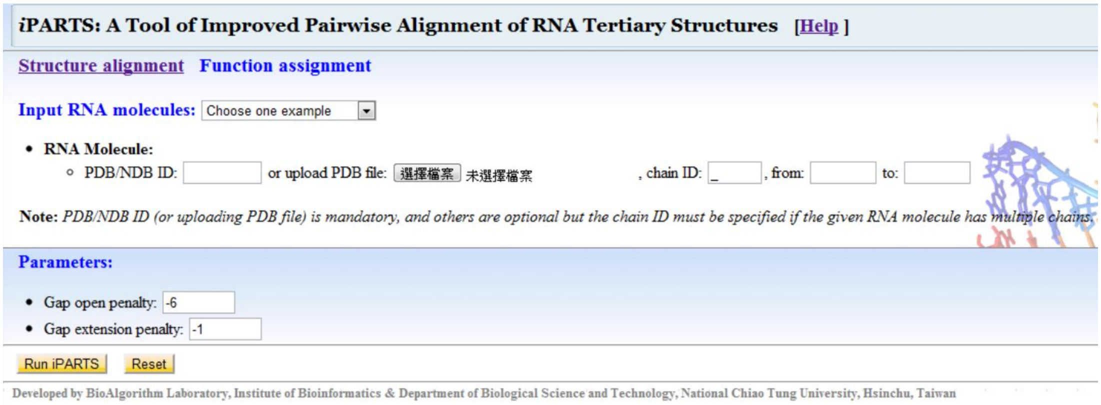
3.1 Input of iPARTS

As illustrated in Figure 3-1, our iPARTS provides an interface that is intuitive and easy to operate. The user can choose one of the examples we prepared in advance for testing iPARTS, or submit a job according to the procedures described as follows. Notice that the user can reset all the input values, as well as parameter settings, by pressing the "Reset" button.

1. Enter the PDB/NDB ID (4-/6-character code) of the input RNA molecule (or upload its file in the PDB format), its chain ID, and the starting and ending residue numbers of a chain fragment to be aligned. Note that PDB/NDB ID or uploading the file is mandatory, and others are optional but the chain ID must be specified if the given RNA molecule has multiple chains.

2. Just click "Run iPARTS" button, if the user would like to run iPARTS with default parameters; otherwise, the user continues with the following steps of parameter settings.

3. Key in two real values for gap open penalty and gap extension penalty, respectively, because iPARTS uses the so-called affine gap penalty function to charge the gaps.



iPARTS: A Tool of Improved Pairwise Alignment of RNA Tertiary Structures [Help]

Structure alignment Function assignment

Input RNA molecules: Choose one example

- RNA Molecule:
  - PDB/NDB ID: or upload PDB file: 選擇檔案 未選擇檔案 , chain ID: _ , from: to:

Note: *PDB/NDB ID (or uploading PDB file) is mandatory, and others are optional but the chain ID must be specified if the given RNA molecule has multiple chains.*

Parameters:

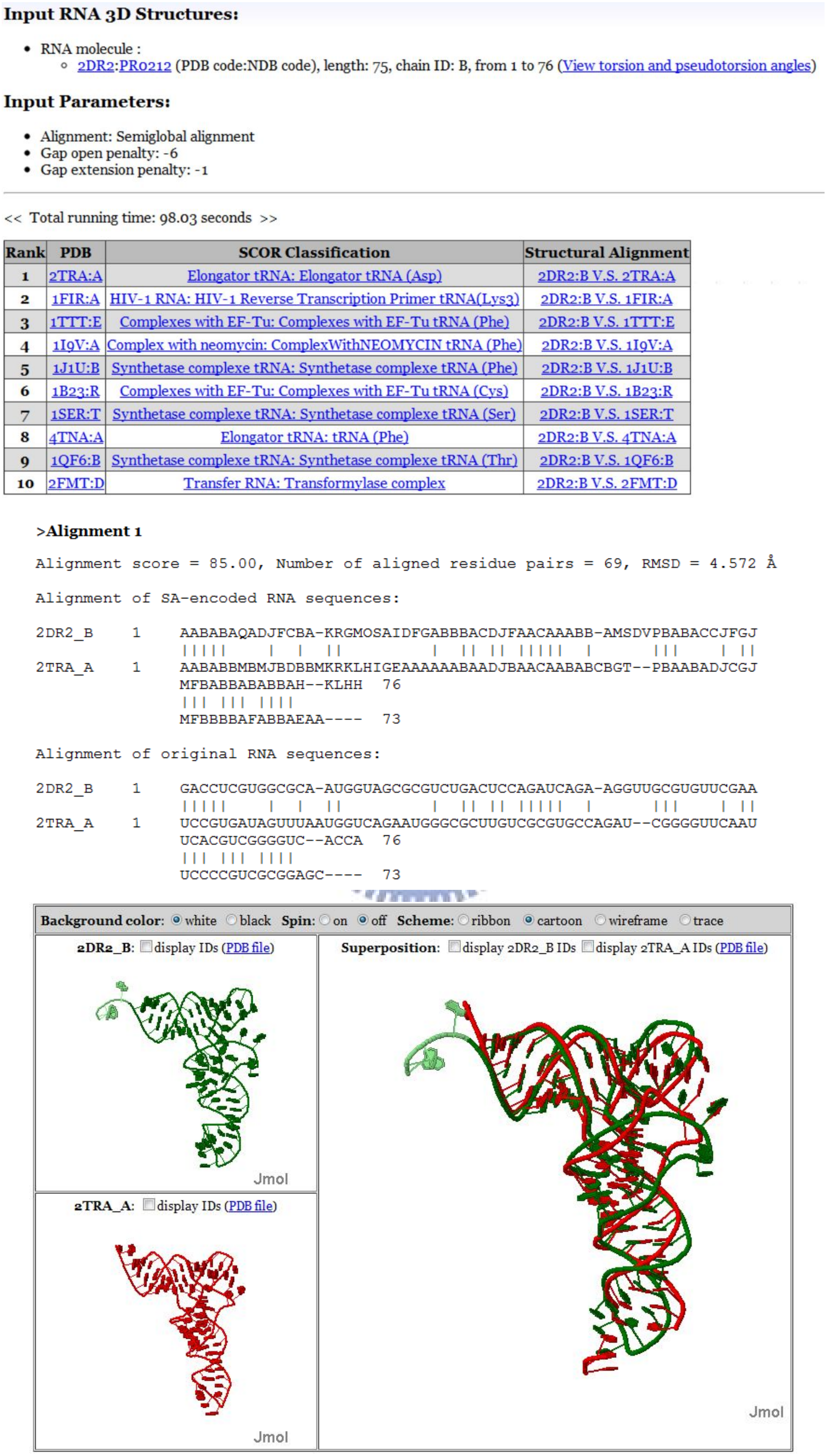
- Gap open penalty: -6
- Gap extension penalty: -1

Developed by BioAlgorithm Laboratory, Institute of Bioinformatics & Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, Taiwan

Figure 3-1. The interface of iPARTS.

3.2 Output of iPARTS

In the output page as shown in Figure 3-2, iPARTS will first show the details of input RNA molecules, user-specified parameters and total running time. Next, iPARTS will list the details of the top ten structurally similar target RNAs, including their PDB IDs, their SCOR classification, and their structural alignments with respect to the query. In addition, the user can click the hyperlink of the PDB ID to show the details of the target RNA annotated in the PDB database; click the hyperlink of the SCOR classification to show the details of the target RNA annotated in the SCOR database; click the hyperlink of the structural alignment to show the details of the structural alignment between the target and query RNAs. Particularly, the display of the structural alignment will show its alignment score, RMSD (root mean square deviation), and detailed alignment of SA-encoded sequences and its corresponding alignment for original RNA sequences. In addition, the user can click the "hyperlink to display structural superposition" link to visually view, rotate and enlarge the 3D structures of the query and target RNAs and their superposition in a Jmol window.



Input RNA 3D Structures:

- RNA molecule :
  - 2DR2:PR0212 (PDB code:NDB code), length: 75, chain ID: B, from 1 to 76 (View torsion and pseudotorsion angles)

Input Parameters:

- Alignment: Semiglobal alignment
- Gap open penalty: -6
- Gap extension penalty: -1

<< Total running time: 98.03 seconds >>

Rank	PDB	SCOR Classification	Structural Alignment
1	2TRA:A	Elongator tRNA: Elongator tRNA (Asp)	2DR2:B V.S. 2TRA:A
2	1FIR:A	HIV-1 RNA: HIV-1 Reverse Transcription Primer tRNA(Lys3)	2DR2:B V.S. 1FIR:A
3	1TTT:E	Complexes with EF-Tu: Complexes with EF-Tu tRNA (Phe)	2DR2:B V.S. 1TTT:E
4	1I9V:A	Complex with neomycin: ComplexWithNEOMYCIN tRNA (Phe)	2DR2:B V.S. 1I9V:A
5	1J1U:B	Synthetase complex tRNA: Synthetase complex tRNA (Phe)	2DR2:B V.S. 1J1U:B
6	1B23:R	Complexes with EF-Tu: Complexes with EF-Tu tRNA (Cys)	2DR2:B V.S. 1B23:R
7	1SER:T	Synthetase complex tRNA: Synthetase complex tRNA (Ser)	2DR2:B V.S. 1SER:T
8	4TNA:A	Elongator tRNA: tRNA (Phe)	2DR2:B V.S. 4TNA:A
9	1QF6:B	Synthetase complex tRNA: Synthetase complex tRNA (Thr)	2DR2:B V.S. 1QF6:B
10	2FMT:D	Transfer RNA: Transformylase complex	2DR2:B V.S. 2FMT:D

>Alignment 1

```
Alignment score = 85.00, Number of aligned residue pairs = 69, RMSD = 4.572 Å

Alignment of SA-encoded RNA sequences:

2DR2_B    1    AABABAQADJFCBA-KRGMOSAIDFGABBBACDJFAACAAABB-AMSDVPBABACCJFGJ
               |||||    |  |  ||         |  || || |||||  |      |||    | ||
2TRA_A    1    AABABBMBMJBDBBMKRKLHIGEAAAAAABAADJBAACAABABCBGT--PBAABADJCGJ
               MFBABBABABBAH--KLHH  76
               ||| ||| ||||
               MFBBBBAFABBAEAA----  73

Alignment of original RNA sequences:

2DR2_B    1    GACCUCGUGGCGCA-AUGGUAGCGCGUCUGACUCCAGAUCAGA-AGGUUGCGUGUUCGAA
               |||||    |  |  ||         |  || || |||||  |      |||    | ||
2TRA_A    1    UCCGUGAUAGUUUAAUGGUCAGAAUGGGCGCUUGUCGCGUGCCAGAU--CGGGGUUCAAU
               UCACGUCGGGGUC--ACCA  76
               ||| ||| ||||
               UCCCCGUCGCGGAGC----  73
```

Background color: ● white ○ black **Spin:** ○ on ● off **Scheme:** ○ ribbon ● cartoon ○ wireframe ○ trace

2DR2_B: □ display IDs (PDB file)

2TRA_A: □ display IDs (PDB file)

Superposition: □ display 2DR2_B IDs □ display 2TRA_A IDs (PDB file)

Figure 3-2. The output page of iPARTS.

Chapter 4

Results and Discussions

In this chapter, we will demonstrate the capability of iPARTS in assigning the functions of RNA tertiary structures. Moreover, we will describe the reasons why we classify the nucleotides in the η-θ plot into 23clusters, instead of 46 clusters.

4.1 ROC analyses for RNA function assignment

The ROC curve is to depict the trade-off between true-positive rate (i.e. sensitivity) and false-positive rate (i.e. 1 – specificity). The ROC curve for each experiment in this thesis was obtained as follows. First, the alignments of all pairs of RNA structures are sorted by their native alignment or a geometric match measure, called structural alignment score (SAS) [23,31], where SAS = 100 x RMSD / (number of aligned residues). A threshold is then varied between the maximum and minimum of the sorted alignment/SAS scores for producing the points of the ROC curve. For a fixed threshold, all pairs of aligned RNA structures whose alignment/SAS scores are above the threshold are assumed positive and all below it negative. Moreover, the pairs assumed positive are counted as true positives (TP) if they belong to the same family (i.e. they are structurally similar) and false positives (FP) otherwise (i.e. they are not structurally similar); the pairs assume negative are counted as true negatives

(TN) if they do not belong to the same family and false negatives (FN) otherwise. Then a point of the ROC curve corresponding to this fixed threshold is produced by plotting its true positive rate on the y-axis and its false positive rate on the x-axis, where the ‘true positive rate’ is defined as TP/(TP+FN) and the ‘false positive rate’ as FP/(FP+TN).

To evaluate the accuracy of our iPARTS on RNA function assignment, we tested it on three data sets (called FSCOR, R-FSCOR and T-FSCOR, respectively) that were prepared by Capriotti and Marti-Renom [6] from the SCOR database on their recent study of SARA. The FSCOR data set includes 419 RNA chains that were classified into 192 classes, the R-FSCOR data set contains the representative structures of 192 classes in the FSCOR data set and the T-FSCOR data set has all structures of the FSCOR data set not present in the R-SCOR data set. In the study by Capriotti and Marti-Renom [6], two RNA structures have a ‘geodesic distance’ $d = 0$ if they were annotated with the same function in the SCOR database, and $d \leq 2$ if the number of edges between their SCOR function annotations, which are organized in a directed acyclic graph, is ≤ 2. The evaluation of structure-based function assignment was usually done by searching with a query RNA structure against a representative data set of annotated RNA structures and predicting the function of the query as the annotated function of the top hit RNA structure. For this purpose, Capriotti and Marti-Renom [6] performed two different tests using their SARA tool: (i) a leave-one-out test using the FSCOR data set and (ii) a test using each structure in the T-SCOR data set as the query and searching it against the R-FSCOR data set. As described in [6], SARA resulted in an AUC of 0.61 and 0.83 for $d = 0$ and $d \leq 2$, respectively, on the leave-one-out test and an AUC of 0.58 and 0.85 for $d = 0$ and $d \leq 2$, respectively, on the other test. Here, we repeated these two experiments using our iPARTS tool. Consequently, the AUC values obtained by iPARTS on the leave-one-out test are 0.72 and 0.92 for $d = 0$ and $d \leq 2$, respectively (see Figure 4-1a for their ROC curves) and 0.77 and 0.90 for $d = 0$ and $d \leq 2$, respectively, on the second test (Figure 4-1b), suggesting that our iPARTS performs better than SARA on the function assignment of RNA 3D structures.

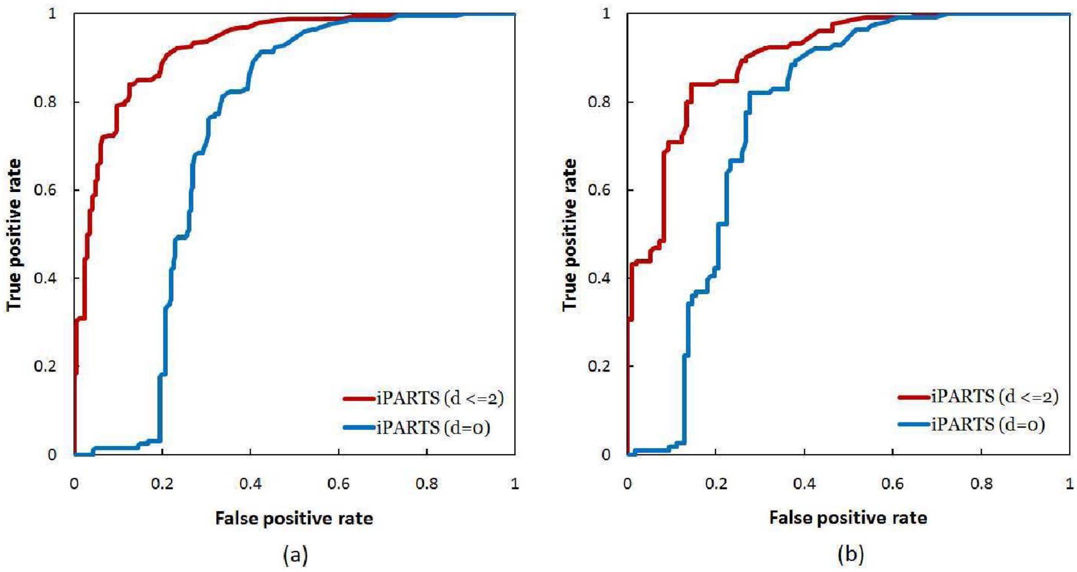


Figure 4-1. The ROC curves when testing our iPARTS for its capability of function assignment using (a) the FSCOR dataset, where the AUC values for d = 0 and d ≤ 2 are 0.72 and 0.92, respectively, and (b) the R-FSCOR and T-FSCOR datasets, where the AUC values for d = 0 and d ≤ 2 are 0.77 and 0.90, respectively.

4.2 The number of the clusters in the η-θ plot

In this study, we chose 23 as the number of the clusters on the η-θ plot based on the following two reasons. First, over 60% of nucleotides on the η-θ 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 result 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.

Here we prepared a filtered and non-redundant testing dataset that consists of 34 families of 100 RNA structures to calculate the values of the areas under ROC curves (AUC) for *N* = 23 and *N* = 46, respectively. According to our experimental results, the value of the AUC obtained using our testing dataset with *N* = 46 is not better than that with *N* = 23 from the viewpoints of both native alignment and SAS scores (Figure 4-2). Second, choosing *N* = 23 will allow 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.

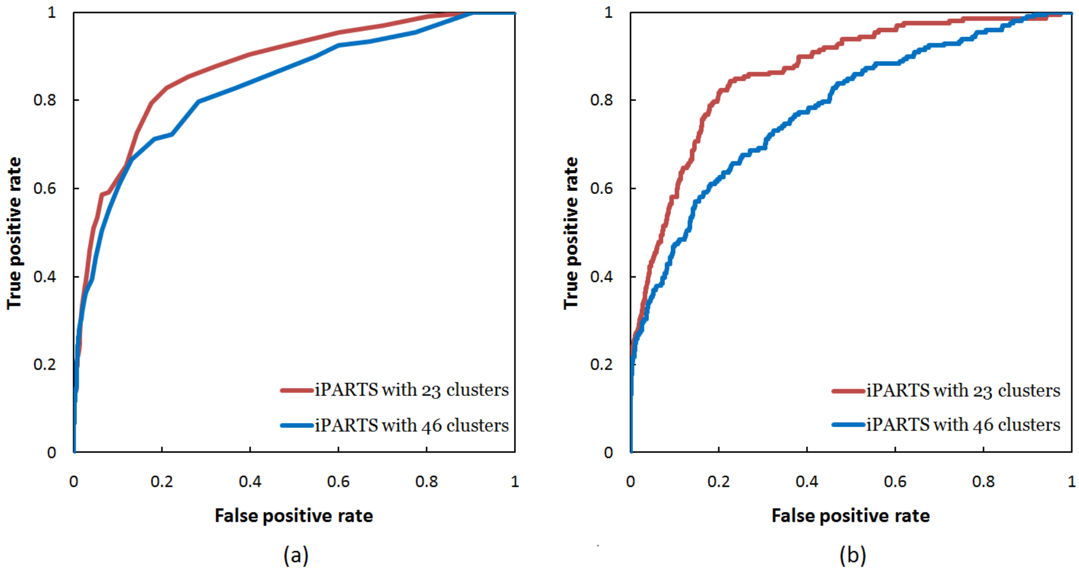


Figure 4-2. The ROC curves of iPARTS with 23 and 46 clusters (a) based on native alignment score, where the AUC value of 23 clusters is 0.87, while the AUC value of 46 clusters is 0.83, and (b) based on SAS score, where the AUC value of 23 clusters is 0.86, while the AUC value of 46 clusters is 0.78.

Chapter 5

Conclusions

In this thesis, we have equipped our iPARTS with a new function that is able to help biologists to accurately assign the function of a given RNA 3D structure. The basic idea behind this new function of iPARTS is as follows. We first obtained a structural alphabet (SA) of 23 letters by using the two pseudo-torsion angles of RNA nucleotide backbone and the affinity propagation clustering approach. Next, we utilized the SA to encode the query and all the RNA 3D structures in a pre-prepared database into 1D SA-encoded sequences, then used iPARTS to compare the globally structural similarities between the query RNA and each of the RNAs with known functions in the database, and finally assigned the annotated function of the most structurally similar RNA to the query RNA 3D structure. Our experimental results have also demonstrated that iPARTS indeed outperforms SARA on the function assignment of RNA 3D structures. Therefore, we believe that iPARTS can serve as a useful tool in the study of structural and functional biology.

References

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

2. Arslan, A.N., Egecioglu, O. and Pevzner, P.A. (2001) A new approach to sequence comparison: normalized sequence alignment. *Bioinformatics*, **17**, 327-337.

3. Bauer, R.A., Rother, K., Moor, P., Reinert, K., Steinke, T., Bujnicki, J. and Preissner, R. (2009) Fast Structural Alignment of Biomolecules Using a Hash Table, N-Grams and String Descriptors. *Algorithms*, **2**, 692-709.

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. *Nature Structural Biology*, **7**, 954-956.

10. Dror, O., Nussinov, R. and Wolfson, H. (2005) ARTS: alignment of RNA tertiary structures. *Bioinformatics*, **21 Suppl 2**, ii47-53.

11. Dror, O., Nussinov, R. and Wolfson, H.J. (2006) The ARTS web server for aligning RNA tertiary structures. *Nucleic Acids Research*, **34**, W412-415.

12. Duarte, C.M. and 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. Ferré, 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.

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

16. 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.

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

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. Hershkovitz, E., Sapiro, G., Tannenbaum, A. and Williams, L.D. (2006) Statistical analysis of RNA backbone. *Ieee-Acm Transactions on Computational Biology and Bioinformatiocs*, **3**, 33-46.

20. Hershkovitz, E., Tannenbaum, E., Howerton, S.B., Sheth, A., Tannenbaum, A. and Williams, L.D. (2003) Automated identification of RNA conformational motifs: theory and application to the HM LSU 23S rRNA. *Nucleic Acids Research*, **31**, 6249-6257.

21. 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.

22. Klosterman, P.S., Hendrix, D.K., Tamura, M., Holbrook, S.R. and Brenner, S.E. (2004) Three-dimensional motifs from the SCOR, structural classification of RNA database: extruded strands, base triples, tetraloops and U-turns. *Nucleic Acids Research*, **32**, 2342-2352.

23. Kolodny, R., Koehl, P. and Levitt, M. (2005) Comprehensive evaluation of protein structure alignment methods: Scoring by geometric measures. *Journal of Molecular Biology*, **346**, 1173-1188.

24. Kolodny, R. and Linial, N. (2004) Approximate protein structural alignment in polynomial time. *Proc Natl Acad Sci USA*, **101**, 12201-12206.

25. Mattick, J.S. and Makunin, I.V. (2006) Non-coding RNA. *Human Molecular Genetics*, **15**, R17-R29.

26. Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**, 443-453.

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. Setubal, J.C. and Meidanis, J. (1997) *Introduction to computational molecular biology*. PWS Pub., Boston.

29. Smith, T.F. and Waterman, M.S. (1981) Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, **147**, 195-197.

30. Storz, G. (2002) An expanding universe of noncoding RNAs. *Science*, **296**, 1260-1263.

31. Subbiah, S., Laurents, D.V. and Levitt, M. (1993) Structural Similarity of DNA-Binding Domains of Bacteriophage Repressors and the Globin Core. *Current Biology*, **3**, 141-148.

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**, D182-184.

34. Wadley, L.M., Keating, K.S., Duarte, C.M. and 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 Res*, **38 Suppl**, W340-347.

36. Xu, R. and Wunsch, D., 2nd. (2005) Survey of clustering algorithms. *IEEE Trans Neural Netw*, **16**, 645-678.