國 立 交 通 大 學

生物資訊及系統生物研究所

碩士論文

利用 RNA 二級與三級結構資訊
識別 RNA 結構模體

Identifying RNA Structural Motifs
Using RNA 2D and 3D Structural Information

研 究 生：陳昱全

指導教授：邱顯泰 博士

盧錦隆 博士

中華民國 一〇〇 年 七 月

利用 RNA 二級與三級結構資訊
識別 RNA 結構模體

Identifying RNA Structural Motifs
Using RNA 2D and 3D Structural Information

研 究 生：陳昱全　　Student：Yi-Chiuan Chen

指導教授：邱顯泰 博士　Advisor：Dr. Hsien-Tai Chiu

盧錦隆 博士　　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

July 2011, Hsinchu, Taiwan

中文摘要

近年來，ncRNA 這種不會轉譯成蛋白質的 RNA 分子在生物領域越來越受重視。他們在細胞內扮演許多重要的角色，包括基因調控、RNA 的修改與參與染色體的複製等等。生物學家在 PDB 資料庫的 RNA 結構中發現有許多反覆出現的相似子結構，這些相似的子結構就被稱為 RNA 結構模體。許多研究也已經證明了這些 RNA 結構模體通常具有特定的功能。然而，把結構模體從 PDB 資料庫中的 RNA 結構準確地辨識出來仍是一件具有挑戰性的工作。RNAMotifScan 這個工具是利用一維序列與二級結構資訊進行序列比對以找出在一個給定的 RNA 結構中的一個特定 RNA 結構模體。但是，RNA 的一維序列在演化上比 RNA 二級與三級結構較不具有保守性，這意味著一維序列在辨識 RNA 結構模體的用處比不上二級與三級結構。在本論文中，我們使用先前所發展的結構字元將 RNA 三級結構編碼成一維結構序列，並且進一步修改 RNAMotifScan 程式使得修改後的 RNAMotifScan 能以 RNA 的鹼基對(即二級結構資訊)與一維結構序列(即三級結構資訊)來辨識 RNA 結構模體。最後，我們的實驗結果顯示上述的方法確實

能夠進一步地改善原本的 RNAMotifScan 在辨識 RNA 結構模體的執行效能。

Abstract

In recent years, the non-coding RNAs (ncRNAs) whose transcripts are not translated into proteins are becoming more and more important in the biology. They play essential roles in many cellular processes, such as gene regulation, RNA modification and chromosome replication. Biologists found that there were many recurrent conserved substructures, called RNA structural motifs, in the RNA structures currently deposited in the PDB database. Many studies have also shown that the RNA structural motifs usually have specific functions. However, it remains a challenging task to accurately identify them from the RNA structures in the PDB database. RNAMotifScan is an alignment-based tool for identifying a specified RNA structural motif from a given RNA structure by considering both the primary sequence (1D) and base pairs (2D) of the RNA structure. However, the 1D sequences of RNAs are less evolutionarily conserved than their 2D and 3D structures, suggesting that the 1D sequences are less useful than 2D and 3D structures when identifying the RNA structural motifs. In this thesis, we utilize the structure alphabet that was previously developed by our lab to transform RNA 3D structures into 1D structural sequences and further modify the program of RNAMotifScan such that it can identify the RNA structural motifs based on the base pairs of the given RNAs (i.e., 2D information) and their structural sequences (i.e., 3D information). Finally, our experimental results have shown that the above

method indeed have further improved the performance of RNAMotifScan for identifying the RNA structural motifs.

Acknowledgement

時間過得很快，兩年的碩士班生涯轉眼就到了尾聲。感謝我的指導老師盧錦隆教授的指導，讓我學到做研究以及做事情應有的態度與方法。同時也要感謝邱顯泰教授在我碩二下學期擔任我的另一位指導老師。此外，也要感謝李家同教授每個禮拜都撥空指導我們的英文作文。

謝謝彥菱學姊，在我碩一時幫我解答許多研究上的疑問還有提供許多日常生活的資訊。感謝忠翰學長總是在我肚子餓的時候適時出現，分我一小塊饅頭。感謝演富學長總是在星期六日邀我們去吃肉，一解周末的苦悶。感謝昆澤學長教我許多程式以及解答研究題目上的問題。感謝芸蓁學姊與晟宸學長讓我覺得待在實驗室是充滿活力的。感謝互亘的義氣，不管是在修課、研究總是我的最佳戰友。感謝仁駿與學妹在最後接近口試這段日子幫我跟互亘處理許多事情，也給我不少關於口試投影片的意見。

最後要謝謝我的家人的支持，讓我能無後顧之憂地完成學業。還有許多一直關心我的人，謝謝你們。

Contents

List of tables

List of figures

Chapter 1

Introduction

近年來，隨著越來越多的 RNA 結構被解出，人們越來越清楚 RNA 的功能不僅僅只是攜帶遺傳訊息而已。令人驚奇的是，RNA 更在許多的生理機制調控過程中扮演至關重要的角色。例如參與 RNA 轉錄的調控、蛋白質轉譯與轉譯後修飾、染色體複製以及協助穩定 mRNA 的結構等等。因此了解這些日益增加的 RNAs 的功能就變成生物領域上很重要的一個課題。通常這些 RNAs 大都不轉譯出蛋白質，因此這些不轉譯出蛋白質的 RNAs 就被稱為非編碼 RNA（non-coding RNAs，簡稱 ncRNAs）[24]。與蛋白質類似，ncRNAs 的功能主要取決於其本身的三級結構（tertiary structure）[2]，因此可以透過比較 RNA 三級結構之間的相似程度來了解 ncRNAs 的功能與演化的關係。目前已有許多比對 RNA 三級結構間相似程度的研究，例如 DIAL [7]與 iPARTS [29]。其中 DIAL 是利用一個時間複雜度為二次方的動態規劃演算法（dynamic programming algorithm）所設計

出來的 RNA 三級結構比對工具，此演算法根據 RNA 的一維序列、鹼基對（base-pairs）、扭轉角（torsion angles）與假扭轉角（pseudo-torsion angles）來計算兩個 RNA 三級結構之間的相似程度[7]。iPARTS 是將兩個 RNA 結構分別以結構字元（structural alphabet）編碼成一維結構序列（1D structural sequences），然後再這兩個一維結構序列用傳統的序列比對（sequence alignment）演算法進行比對以決定出這兩個 RNA 結構之間的相似程度[29]。

除此之外，結構生物學家進一步分析發現在 RNA 三級結構中有許多重複出現的相似子結構，這些子結構後來被稱為 RNA 結構模體（RNA structural motifs）。目前 RNA 結構模體的定義有許多不同的版本被提出[10, 15, 31]，在本論文中，我們把它簡化為在 RNA 結構中反覆出現的子結構（substructure）即稱為 RNA 結構模體。RNA 結構模體依骨架的組成可分為局部型（local）與複合型（composite）。所謂的局部型結構模體是指組成該結構的核苷酸皆落在同一個 hairpin loop （Figure 1-1a）或 internal loop（Figure 1-1b）上；反之則被稱為複合型的結構模體，換句話說，複合型結構模體是由三個或三個以上的 RNA 片段所組成的[20, 22] （Figure 1-1c）。

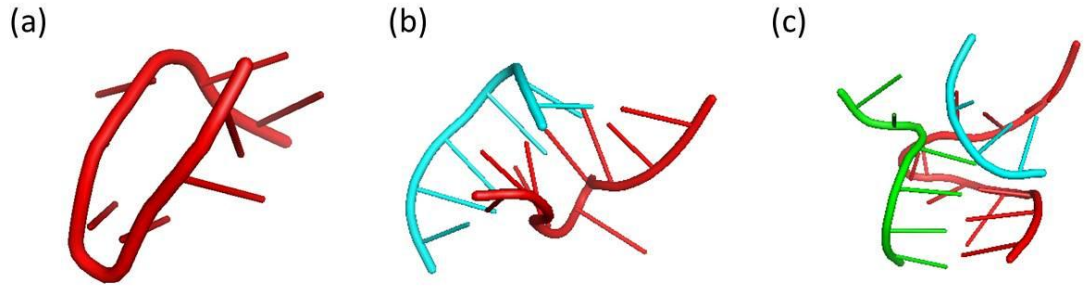


Figure 1-1: 不同類型的 RNA 結構模體。(a) α-loop：由單股 hairpin loop 所形成的局部型結構模體。 (b) Kink-turn：由兩股 internal loops 所形成的局部型結構模體。(c) Kink-turn：由三股 RNA 片段所形成的複合型結構模體。

許多研究發現 RNA 結構模體通常具有特定的功能，例如 Kink-turn 是重要的蛋白質辨識位置[14]、Sarcin-ricin 在蛋白質轉譯過程中會與延長因子（elongation factor）作用[25]。因此，把一個 RNA 結構中的特定結構模體辨識出來將有助於 RNA 功能的了解。目前已有以下兩類的工具可以用來辨識出 RNA 結構中的結構模體[19]:（1）利用 RNA 骨架的三級結構資訊所設計出來的工具，例如 PRIMOS [6]、COMPADRES [28]與 R3D-BLAST [21]。（2）利用鹼基（一維序列）與鹼基對（二級結構）的資訊所設計出來的工具，例如 FR3D [22] 與 RNAMotifScan [31]。

第一類工具的特點是先計算出 RNA 骨架上的扭轉角角度，再依據扭轉角角度來決定 RNA 骨架構型的相似程度。每個 RNA 序列上

的核苷酸都具有六個標準的扭轉角（α、β、γ、δ、ε 與 ζ）（Figure 1-2a）。若同時將此六個標準扭轉角一起納入考慮的話， 那麼核苷酸構型的分析將會是高維度的計算（high-dimensional computation），並且也不容易將分析的結果給予視覺化。為了要解決考慮全部標準扭轉角的缺點，Duarte 與 Pyle 等人提出以假扭轉角 η 與 θ 來表示核苷酸的骨架（Figure 1-2b）[5]。

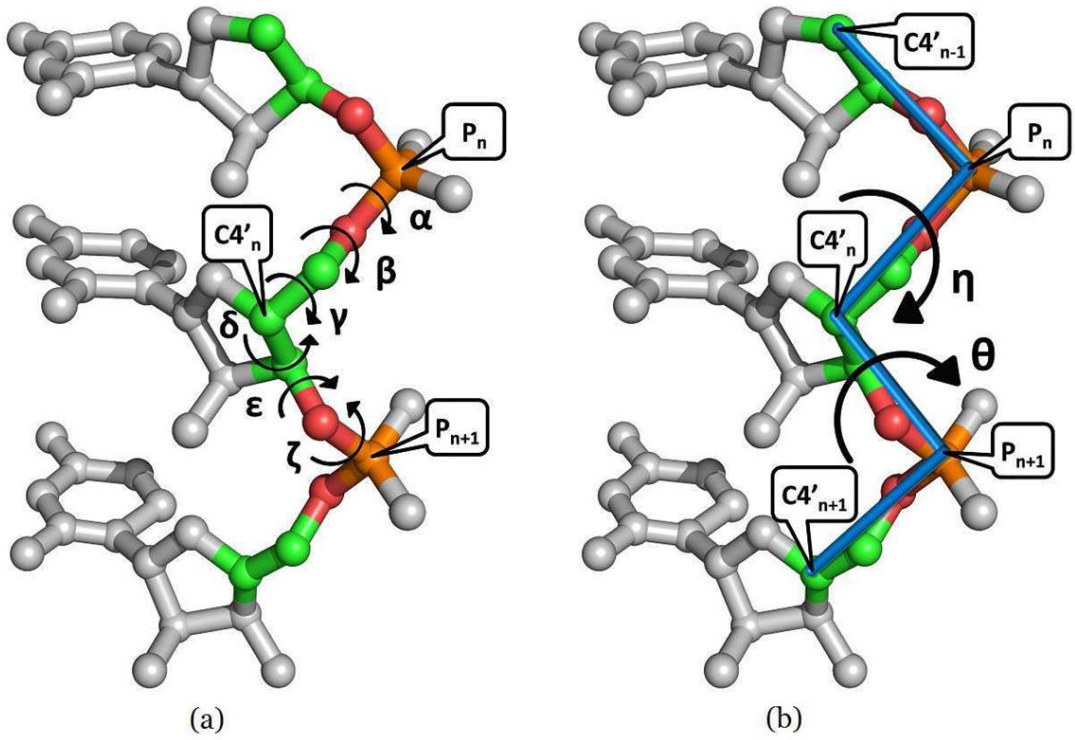


Figure 1-2: (a) 六個標準扭轉角，α、β、γ、δ、ε 與 ζ。(b) 以兩個假扭轉角 η 與 θ 表示一個核苷酸（標示為 n），η 是由 $C4'_{n-1}$, P_n, $C4'_n$ 與 P_{n+1} 四個原子所形成的假扭轉角度，θ 是由另四個原子 P_n, $C4'_n$, P_{n+1} 與 $C4'_{n+1}$ 所形成的假扭轉角度。

利用 η 與 θ 兩個假扭轉角表示核苷酸的骨架後，便可以把高維度的核苷酸分析降為二維的計算，同時也可以把核苷酸與其分析的結果視覺化地表示在一個二維的平面上，此有助於進一步地把核苷酸的骨架構型進行分類[5, 27]。利用此概念的 Duarte 與 Pyle 等人發展了 PRIMOS [6]與 COMPADRES [28]兩個工具來分析特定結構模體與 PDB 資料庫中 RNA 結構之間的骨架構型相似程度以搜尋出相似的 RNA 結構模體。我們實驗室亦利用此概念發展 R3D-BLAST [21]可以快速且準確地搜尋出單股的結構模體。在 R3D-BLAST 的方法中，首先將所有 PDB 中的 RNA 結構分別以結構字元編碼成一維結構序列，然後再利用這些一維結構序列與 BLAST 去搜尋出在 PDB 資料庫中與特定結構模體結構相似的 RNA 子結構。

RNAMotifScan 是利用鹼基（一維序列）與鹼基對（二級結構）的資訊所設計出來的工具。首先它把一個 RNA 的鹼基與鹼基對表示成一個弧線註解序列（arc-annotated sequence，即在鹼基序列中利用弧線表示鹼基對），然後再設計一個兩條弧線註解序列的比對（pairwise arc-annotated sequence alignment）演算法以尋找出特定的 RNA 結構模體。在 RNAMotifScan 發表之前，Jiang 等人[13]在 2002 年已發表過類似的方法，但是在鹼基對的部分，Jiang 等人僅考慮了傳統的 Watson-Crick 鹼基對 A/U、C/G 與 Wobble 鹼基對 U/G。然而，

隨著結構生物學的發達，生物學家發現在 RNA 裡的鹼基對配對方式除了傳統型之外，還有許多不同的配對方式。Leontis 等人將這些不同於傳統型的鹼基對歸類為非傳統型鹼基對（non-Watson-Crick 鹼基對）[17]。與 Jiang 等人的方法[13]相比，RNAMotifScan 除了考慮傳統型鹼基對之外也把非傳統型鹼基對納入考慮的範圍，但是在分數的計算上，RNAMotifScan 僅考慮鹼基對（弧線）的 match 與 mismatch，然而 Jiang 等人的方法除了考慮 arc-match 與 arc-mismatch 之外還考慮了另外三種鹼基對的比對(即 Figure 1-3 所示的 arc-removing、arc-alterning 和 arc-breaking)與給分。

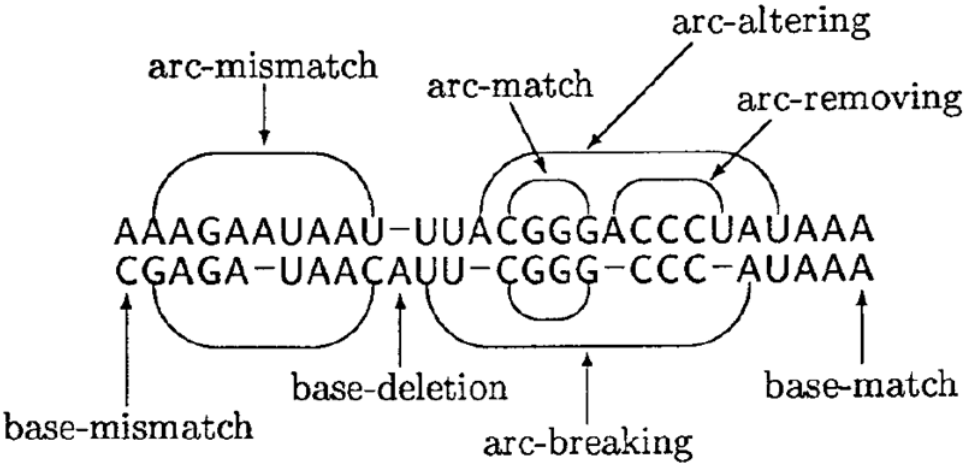


Figure 1-3: 兩條弧線註解序列的比對與編輯操作[13]。

前文提到我們實驗室在 iPARTS 與 R3D-BLAST 的方法中已經設計了一個比對兩個 RNA 三級結構的啟發式（heuristic）演算法。我們當時收集了一些具有代表性的 RNA 三級結構，然後再把這些 RNA

三級結構中核苷酸（但不包含第一個與最後一個核苷酸）的 η 與 θ 角度計算出來並將這些核苷酸依其計算出來的 η 與 θ 角度繪製在一個 η-θ 平面圖上。結果顯示在平面圖上的點是呈離散的分佈，於是我們利用分群演算法將這些點分群成 23 群，並且將此 23 群各個中心點（center）以一個字母表示，最後我們把這 23 個字母稱為 RNA 的結構字元集（structural alphabet，簡稱 SA）。對於每一個 RNA 結構，我們逐一地計算出每個核苷酸在 η-θ 圖上的位置與各個結構字元之間的距離，然後再依據最近鄰居的規則（nearest neighbor rule）將核苷酸編碼成最近距離的結構字元，這些編碼後的結構字元就形成了所謂的 RNA 一維結構序列。最後我們再利用傳統兩兩序列比對的演算法去比對兩條一維結構序列以決定出相對應的二個 RNA 結構之間的相似程度。

從演化的角度來說，RNA 的二級與三級結構會比其一維序列更具有保留性，許多研究也顯示比較 RNA 三級結構之間的相似程度會比只比較一維序列獲得更多功能及演化關係上的資訊[2]。因此我們認為使用 RNA 二級與三級的結構資訊來辨識 RNA 結構模體應該會比使用一維序列來得更為精準。前文提到 RNAMotifScan 是一個使用 RNA 一維序列與二級結構的資訊來辨識 RNA 結構模體的工具，我們認為 RNAMotifScan 使用一維序列資訊而不使用三級結構資訊是

由於序列比對演算法方法的限制，使得他們無法直接使用三級結構資訊來辨識 RNA 結構模體。然而，我們在先前的研究中已有發展出把 RNA 三級結構編碼成一維結構序列的方法，於是我們想到若能把 RNAMotifScan 方法中原本的一維序列置換成一維結構序列的話，我們就能將 RNAMotifScan 修改成為使用二級結構與三級結構資訊來辨識 RNA 結構模體的工具。我們也相信我們這樣的修改會比原本 RNAMotifScan 的方法更能提升其在辨識 RNA 結構模體的效能。

在本論文中，我們首先將 RNA 三級結構編碼成一維結構序列，並且利用二級結構註解工具找出該 RNA 的鹼基對。接著我們進一步去修改 RNAMotifScan 的程式使得修改後的 RNAMotifScan 能以 RNA 的鹼基對（即二級結構資訊）與一維結構序列（即三級結構資訊）來辨識 RNA 結構模體。除此之外，我們也修正了 RNAMotifScan 演算法遞迴函式中未敘詳盡的部分。最後我們利用五個不同的 RNA 結構模體在二個目標 RNA 結構中尋找相似的結構模體，並且將搜尋出來的結果與原本的 RNAMotifScan 所搜尋出來的結果做比對。實驗結果顯示我們的方法確實能夠進一步地改善原本的 RNAMotifScan 在辨識 RNA 結構模體的執行效能。

Chapter 2

Materials and Methods

本研究方法首先將 query 與目標 RNA 三級結構個別依照結構字元式方法編碼成一維結構序列，再利用 RNAView[30] 或 MC-Annotate[9]兩種 RNA 鹼基對註解工具之一來註解 RNA 三級結構中的 Watson-Crick 鹼基對與 Non-Watson-Crick 鹼基對關係，如此便可將 query 與目標 RNA 三級結構轉換成「弧線註解的一維結構序列 (arc-annotated 1D structural sequences) 」。接著我們修改 RNAMotifScan 的方法，使其可以將以弧線註解的一維結構序列當作輸入，並且將在目標 RNA 中所有與 query 結構模體相似的候選結構輸出。

2.1 以結構字元式方法將三級結構編碼成一維結構序列

結構字元式方法的最主要精神就是將三級結構轉換成由結構字

元所組成的序列。如前段所敘述，2D η-θ 圖可以幫助我們以圖形表示法的方式量化不同結構間的差異性，以進行 RNA 三級結構的分析。我們準備了一組不重複（non-redundant）且解構解析度最低到 3.0Å 的 PDB 資料集來繪製 η-θ 圖。這組資料包含 117 個 RNA 結構，一共有 9527 個核苷酸。接下來我們使用由 Duarte 與 Pyle 發展的 AMIGOS 工具[5]，將這些 RNA 結構中各核苷酸的 η 與 θ 角度計算出來（不包含每個結構的開頭與結尾的核苷酸，一共 9267 個）。並且將這些計算出來角度繪製在以 η 角度為 X 軸、以 θ 角度為 Y 軸的 η-θ 圖上（Fig 2-1）。接著使用 AP 分群演算法（AP clustering algorithm）[8]將 η-θ 圖上的所有點分成 23 群（Figure 2-2），並且將此 23 群各個中心點結構（center）以一個字母代表，於是我們將這 23 個字母稱為結構字元集（structural alphabet，SA），如 Figure 2-3。對於每個 RNA 結構，我們逐一計算每個核苷酸在 η-θ 圖上的位置與各中心點位置之間的距離，依據 nearest neighbor rule 將此核苷酸的結構逐一編碼成與其距離最近的中心點的結構字元，最後將整個 RNA 三級結構編碼成一維結構字元序列。接著為了要使結構字元能夠精確地比對，我們運用 Henikoff 與 Henikoff[11]提出的統計方法製作 23×23 的結構字元置換分數表，如此一來便可以應用傳統字串比對演算法比對兩結構之間的相似程度。

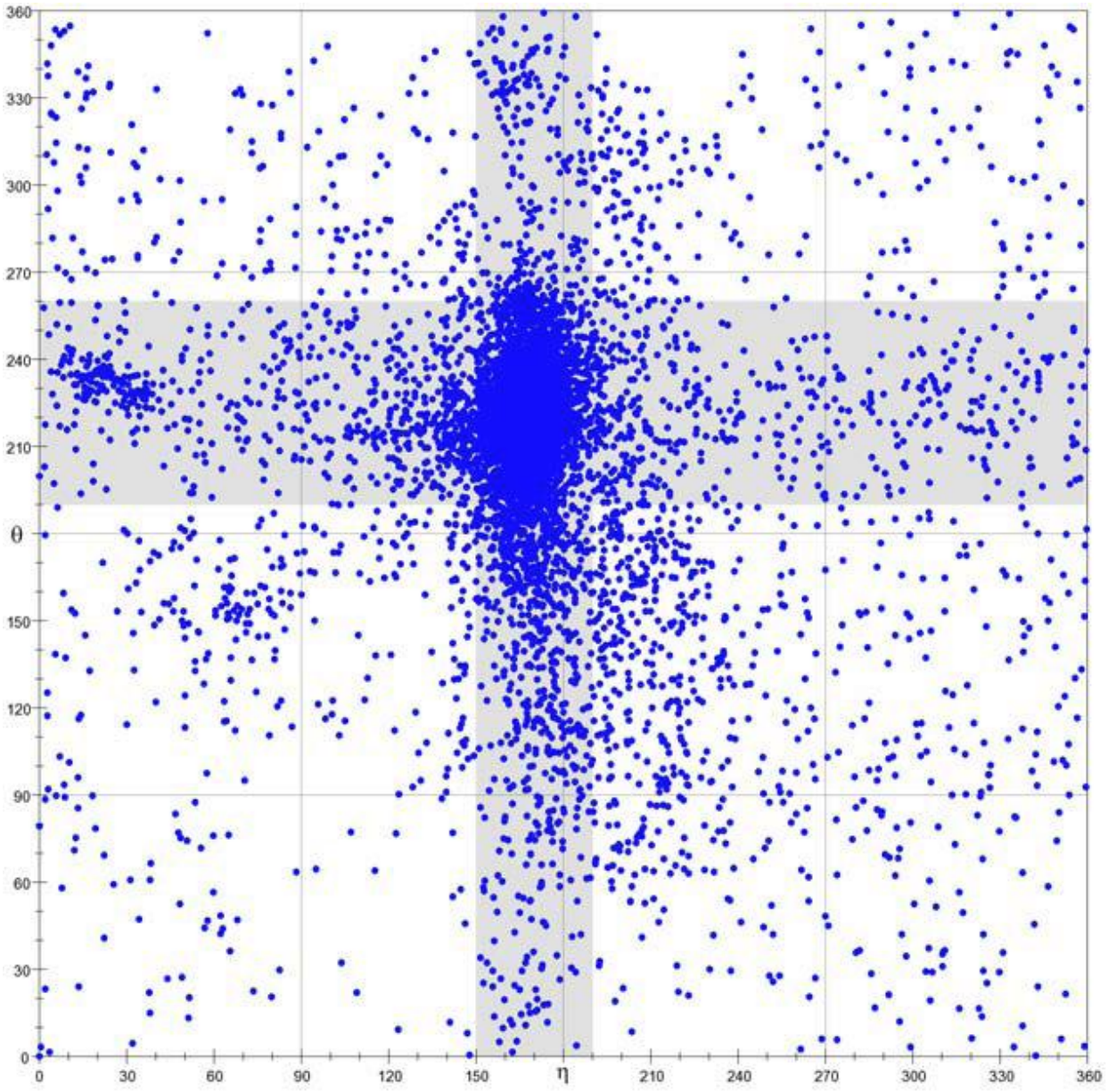


Figure 2-1: 所有在 PDB 結構資料集內且非末端的核苷酸 η-θ 圖。每個點代表一個核苷酸。中間兩灰底且垂直重疊的區域（150° ≤ η ≤ 190° 與 190° ≤ θ ≤ 260°）為 RNA 結構中的螺旋結構區域（helical region）。

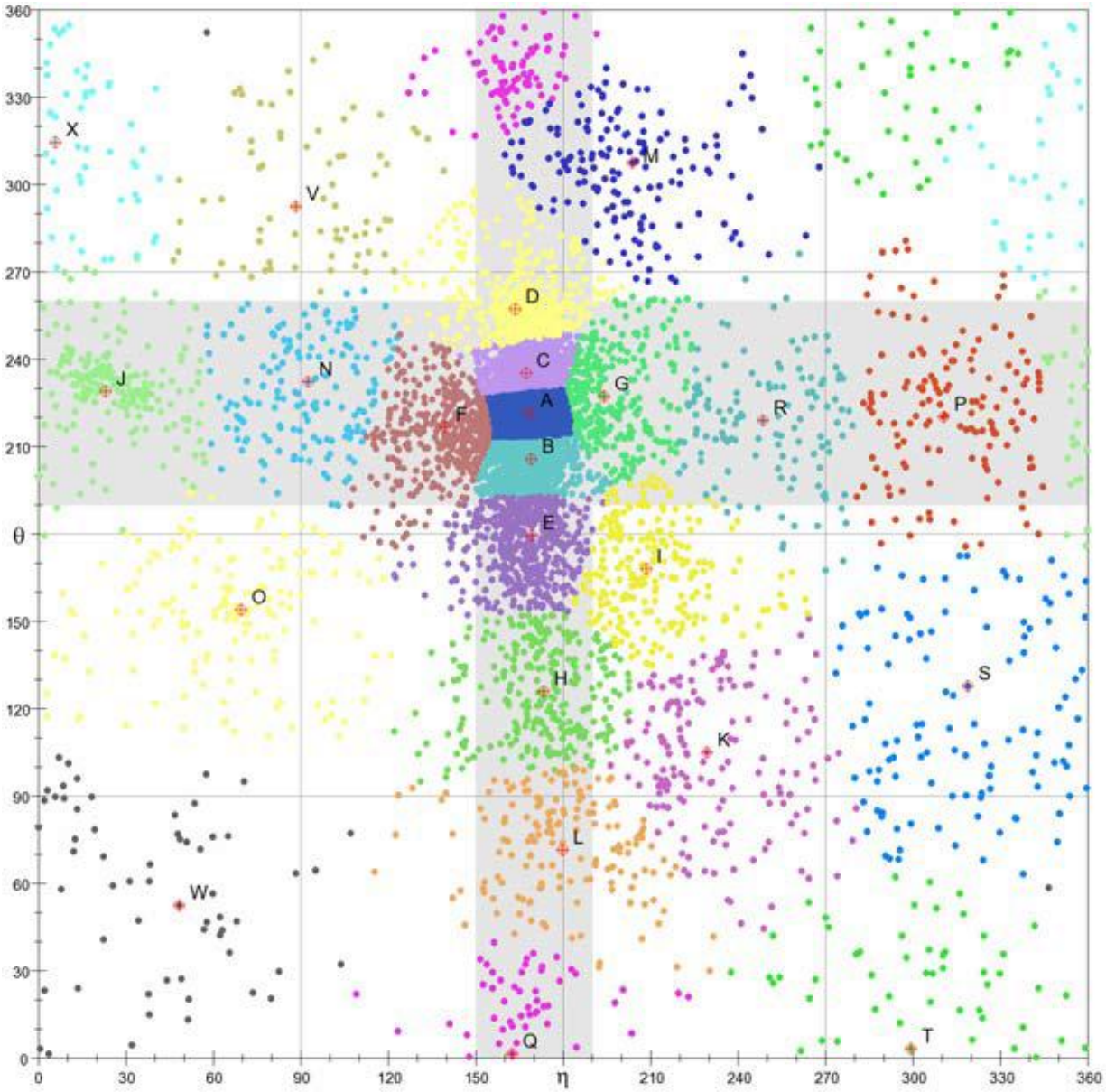


Figure 2-2: 以 AP 演算法所分出來的 23 個群組。每群分別以不同顏色表示，並以字母標記出該群的中心點。

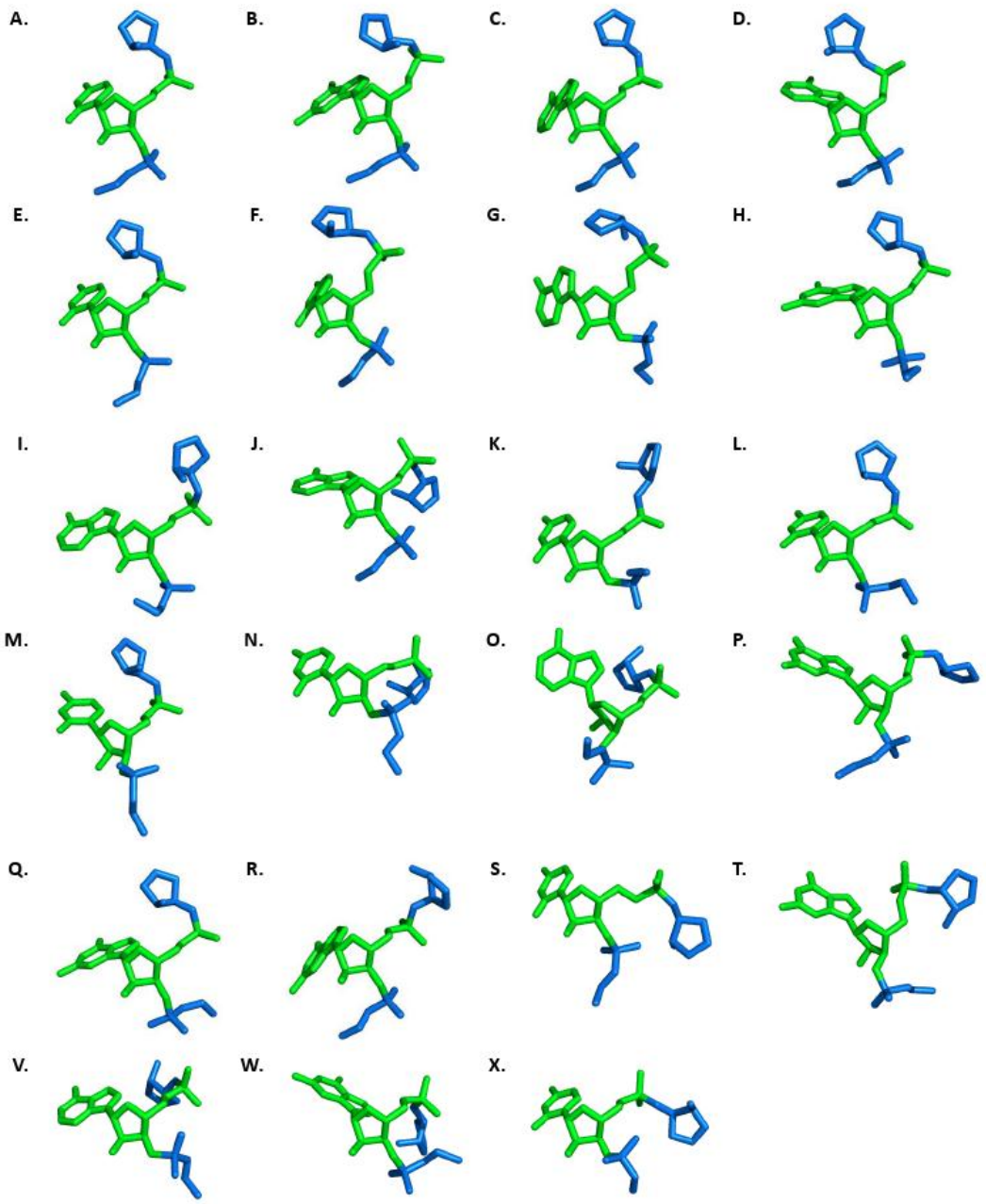


Figure 2-3: 23 個分群中心點核酸的三級結構圖。中心點核酸以綠色表示；中心點前後各一個會影響中心點核酸假扭轉角度的核酸以藍色表示。

2.2 以鹼基對註解工具註解 Watson-Crick 鹼基對與 Non-Watson-Crick 鹼基對

RNA 鹼基對是由鹼基上的極性原子在兩鹼基之間形成氫鍵鍵結的邊對邊關係（edge-to-edge relationship）[18]。RNA 的鹼基對關係最早被發現的是大量存在於 A-型雙螺旋（A-form double helix）中的傳統型 Watson-Crick 鹼基對。但是在結構生物學迅速發展之下，大量的 RNA 結構被解析出來。生物學家從這些 RNA 的結構中發現鹼基對的結合有非常多種的形式，Watson-Crick 鹼基對僅是眾多形式中的其中一種。Leontis 與 Westhof 於 2001 年針對 RNA 鹼基對提出了一套分類方法[18]。在這套方法中，他們將鹼基上可形成氫鍵的位置分成三個邊，分別是 Watson-Crick 邊（命名依據為傳統 Watson-Crick 鹼基對形成氫鍵的邊）、Hoogsteen 邊（或在嘧啶中稱 C-H 邊）與 Sugar 邊（Figure 2-4）。另外依據鹼基對兩醣苷鍵方向可分為順式（*cis*）與反式（*trans*）（Figure 2-5）。依據邊對邊關係與醣苷鍵方向把鹼基對分為十二個家族（Table 2-1）。Leontis 等人發現在各家族中，C1’-C1’ 原子間距離相近的鹼基對在結構上是相似的，且在 RNA 結構中可以互相抽換而不會劇烈影響到整體的結構。這些相似的鹼基對就稱為

同構鹼基對（isosteric pairs）。這個分類方法，可以幫助辨識在同源的 RNA 分子結構中，可以互相置換而不會影響結構的鹼基對，並且也可幫助觀察等量鹼基對的共同變化（co-vary）關係[17]。

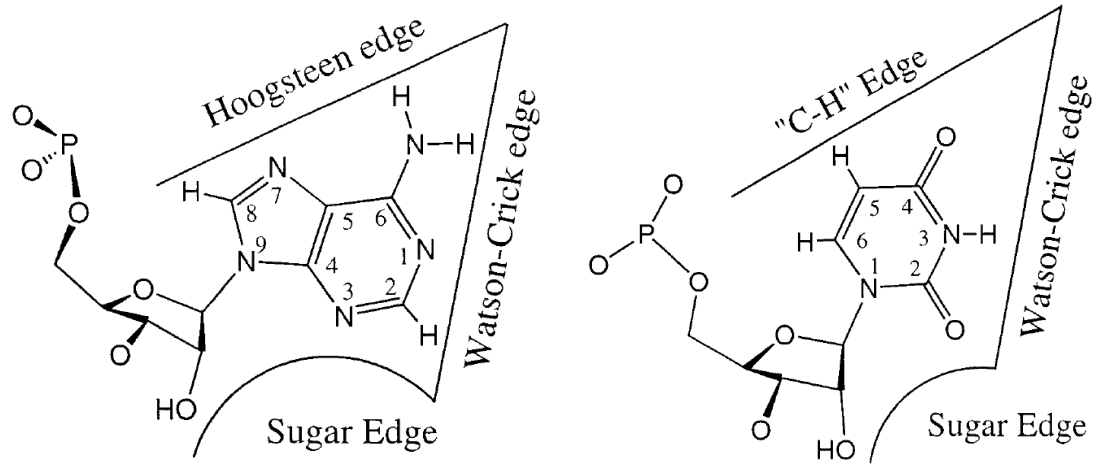


Figure 2-4: RNA 鹼基上的三個可形成氫鍵的邊。（左）嘌呤上的邊與其對應名稱。（右）嘧啶上的邊與其對應名稱[17]。

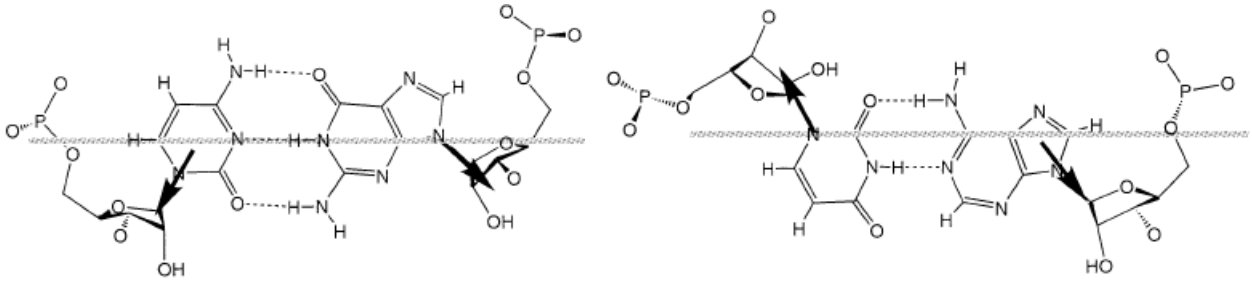
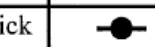
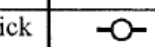
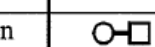
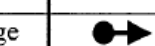
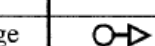
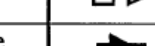


Figure 2-5: 順式（*cis*）與反式（*trans*）醣苷鍵方向。（左）順式醣苷鍵方向。（右）反式醣苷鍵方向[17]。

Table 2-1: 以醣苷鍵方向（第二行）與邊對邊關係（第三行）為依據將 RNA 鹼基對分為 12 個家族。第四行表示以圖形代表各個家族，

圓形代表 Watson-Crick 邊、三角形代表 Sugar 邊、四方形代表 Hoogsteen 邊、圖形填滿代表順式（*cis*）、圖形中空代表反式（*trans*）[17]。其中傳統型 Watson-Crick 鹼基對與 Wobble 鹼基對屬於第一個家族，除了這兩種鹼基對以外，其他的鹼基對都稱為非-Watson-Crick 鹼基對。

No.	Glycosidic Bond Orientation	Interacting Edges	Symbol	Default Local Strand Orientation
1	*Cis*	Watson-Crick / Watson-Crick		Anti-parallel
2	*Trans*	Watson-Crick / Watson-Crick		Parallel
3	*Cis*	Watson-Crick / Hoogsteen		Parallel
4	*Trans*	Watson-Crick / Hoogsteen		Anti-parallel
5	*Cis*	Watson-Crick / Sugar Edge		Anti-parallel
6	*Trans*	Watson-Crick / Sugar Edge		Parallel
7	*Cis*	Hoogsteen / Hoogsteen		Anti-parallel
8	*Trans*	Hoogsteen / Hoogsteen		Parallel
9	*Cis*	Hoogsteen / Sugar Edge		Parallel
10	*Trans*	Hoogsteen / Sugar Edge		Anti-parallel
11	*Cis*	Sugar Edge / Sugar Edge		Anti-parallel
12	*Trans*	Sugar Edge / Sugar Edge		Parallel

為了註解 RNA 結構中的鹼基對資訊，我們使用了 RNAView[30] 或 MC-Annotate[9]兩種 RNA 鹼基對註解工具，其註解包含了配對關係、鹼基、醣苷鍵方向與邊對邊關係。隨後將註解工具的結果輸出，並與第一個步驟得到的一維結構序列整理成有弧線註解的一維結構序列，以此來當作比對演算法的輸入檔案。

2.3 應用 RNAMotifScan 的弧線註解序列比對演算法比較兩 RNA 結構之間的相似程度

RNAMotifScan 的核心是一個動態規劃（dynamic programming）的序列比對演算法，其輸入為一對弧線註解序列。在方法中，一維序列的計分方法與傳統序列比對方法相同，鹼基對部分仿照一維序列設計一個操作的分數計算方法，加總各鹼基對操作的分數，最後再將一維序列與鹼基對的分數加總，即為比對的分數結果。RNAMotifScan 便是尋找最高分比對結果的演算法。在搜尋方面，RNAMotifScan 先將目標 RNA 結構依照 query 長度與鹼基對個數切成片段，再將各片段逐一與 query 比對，同時計算每個比對結果的 p-value 與偽陽性比率（false positive rate，FPR）。在這個步驟，我們修正 RNAMotifScan 演算法遞迴函式中三個未敘詳盡的部分並且修改 RNAMotifScan 的程式碼，使其可以將兩弧線註解的一維結構序列當作輸入，並且將可能的結果依照比對分數由高至低排序輸出。

對於每個鹼基對 P，將弧線左端與右端的鹼基分別命名為 P_l 與 P_r。假設 $A = A[1]A[2]\ldots A[\alpha]$ 且 $B = B[1]B[2]\ldots B[\beta]$代表含有 m 與 n 個鹼基對的弧線註解一維結構序列。此外，在進入遞迴函式前，兩

個序列個別在頭尾加上由 $A[0]$與 $A[\alpha+1]$、$B[0]$與 $B[\beta+1]$構成的虛擬鹼基對（dummy base pairs）。加上虛擬鹼基對可以減少遞迴函式起始事件的討論件數。在演算法一開始先強迫兩個虛擬鹼基對對在一起，接著在被這兩個鹼基對包含的序列區間做序列比對，比對結束後再將對在一起的虛擬鹼基對去除，就能得到原本問題的解。令 $\mathcal{P}^A = P_1^A, P_2^A, \ldots, P_{m+1}^A$ 與 $\mathcal{P}^B = P_1^B, P_2^B, \ldots, P_{m+1}^B$ 分別代表在 A 與 B 中的鹼基對集合，鹼基對序號根據 P_r 右端鹼基位置由左至右遞增排序，若兩鹼基對右端鹼基位置相同，則依據 P_l 左端鹼基位置排序。

假設在同一序列中有兩鹼基對P' 與 P，此兩鹼基對之間的關係可分為下列三種：（1）並列（juxtapose）：P' 與 P 個別包含的序列區間無重疊，且P'在 P 左側，記為$P' <_p P$ （Figure 2-6a）。（2）巢狀（nested）：P'所包含的序列區間被 P 所包含的序列區間完全包含，記為$P' <_I P$ （Figure 2-6b）。（3）交叉（crossing）：兩鹼基對個別包含的序列區間部分重疊，又名假結（pseudoknot），但在此不考慮此種情形（Figure 2-6c）。

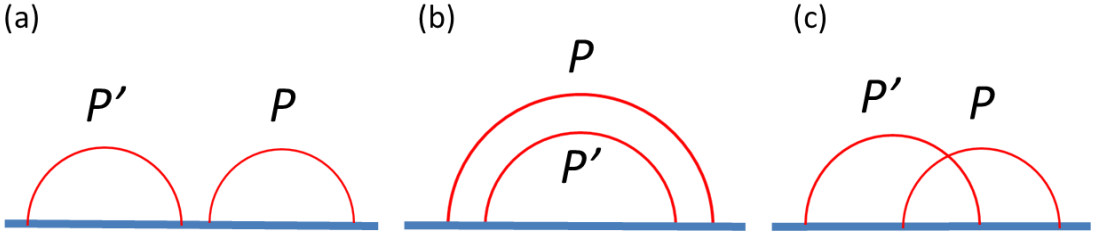


Figure 2-6: 序列中之兩鹼基對關係。(a) 並列。(b) 巢狀。(c) 交叉。

在 RNAMotifScan 的方法中以（1）$Loop(P^A)$表示位於 $A[P_l^A]$ 與 $A[P_r^A]$之間的子序列，但不包含 $A[P_l^A]$ 與 $A[P_r^A]$。（2）$Loop(P^{A'}, P^A)$ 表示介於兩個並列的鹼基對$P^{A'}$與P^A之間的子序列 $A[P_r^{A'}+1]$ … $A[P_l^A-1]$，如 Figure 2-7。在遞迴函式中所使用的分數符號定義如下：（1）$I(P^A, P^B)$：兩鹼基對P^A 與 P^B對起來的分數。如果P^A 與 P^B是同構鹼基對則為 match，分數較高；反之為 mismatch，分數較低。（2）$S(A[i…j],B[k…l])$：在 $A[i…j]$ 與 $B[k…l]$兩子序列間比對的分數。（3）$Gap(k)$：在序列中插入或刪除長度為 k 的子序列的扣分分數。（4）$M[P^A, P^B]$：令P^A 與 P^B對在一起，被P^A 與 P^B包含的序列區間的最佳比對分數，若計算 $M[P^A_{dummy}, P^B_{dummy}]$則是完整 A 與 B 序列的最佳比對分數。

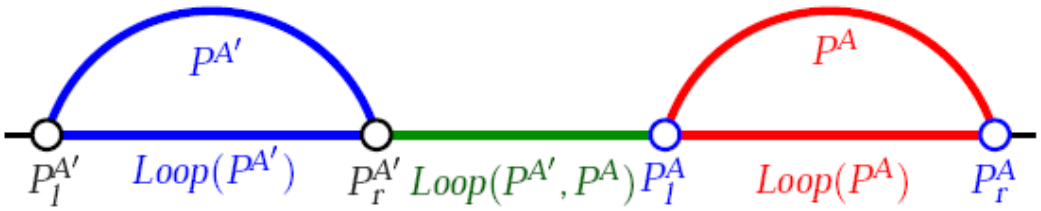


Figure 2-7: 被鹼基對包含的子序列與介於兩並列鹼基對之間的子序列。

在 RNAMotifScan 的方法中，必須計算所有 P^A 對 P^B 的 $M[P^A, P^B]$，且 $P^A \in \mathcal{P}^A$、$P^B \in \mathcal{P}^B$。$M[P^A, P^B]$的計算方式如下：

$M[P^A, P^B] = M_s[P^A, P^B] + \max\{M_h[P^A, P^B], M_l[P^A, P^B]\}$. (Figure 2-8)

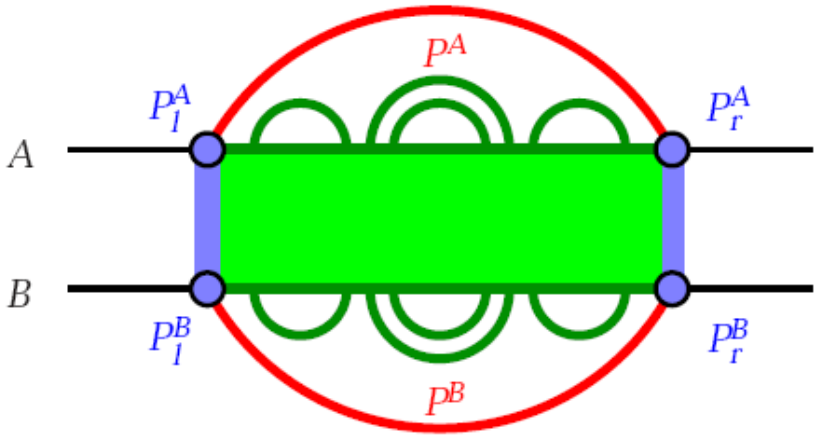


Figure 2-8: $M[P^A, P^B]$的示意圖。P^A與P^B以紅色弧線表示，P^A與P^B的左右鹼基以藍紫色表示，中間綠色表示被P^A與P^B包含的子序列區間。

其中 $M_s[P^A, P^B]$是鹼基對P^A與P^B比對的分數，包含鹼基對是否為同構鹼基對與組成P^A與P^B的左右端鹼基比對的分數，算法如下：

$M_s[P^A, P^B] =$

$w_1 \times I(P^A, P^B) + w_2 \times (S(A[P_l^A], B[P_l^B]) + S(A[P_r^A], B[P_r^B]))$. (Figure2-9)

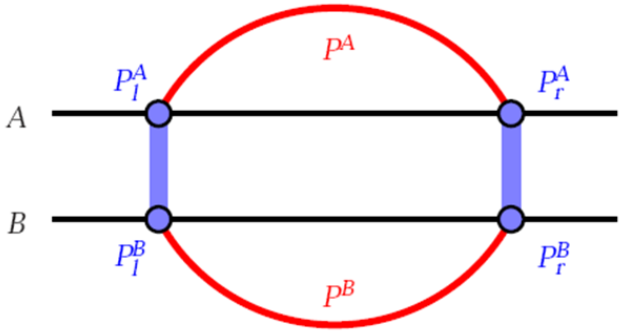


Figure 2-9: $M_s[P^A, P^B]$的示意圖。P^A與P^B以紅色弧線表示，計算鹼

基對比對分數 $I(P^A, P^B)$。P^A與P^B的左右鹼基以藍紫色表示，計算P_l^A對P_l^B、P_r^A對P_r^B序列比對分數。

被P^A與P^B包含的子序列區間的分數計算分成兩種情況討論，並取兩者之間最大值納入計算。第一種情況為 $M_h[P^A, P^B]$，假設在此區間沒有任何對在一起的鹼基對存在，$M_h[P^A, P^B]$的計算方式如下：

$M_h[P^A, P^B] = w_3 \times S(Loop(P^A), Loop(P^B))$. (Figure 2-10)

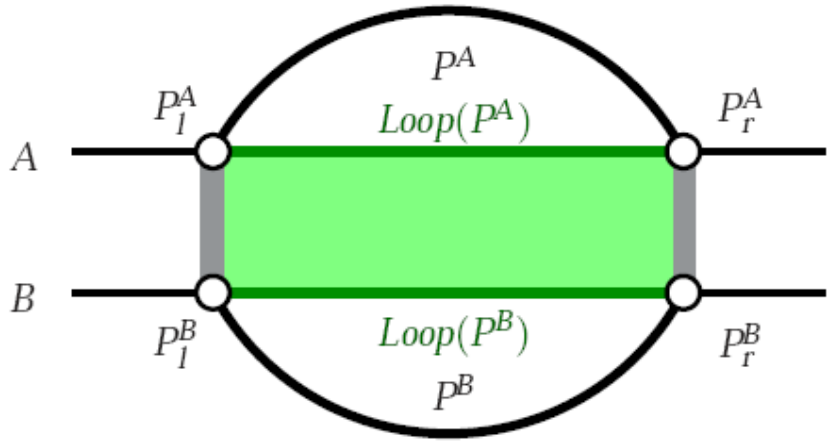


Figure 2-10: $M_h[P^A, P^B]$的示意圖。假設在被P^A與P^B包含的子序列區間沒有對在一起的鹼基對，因為不考慮鹼基對 insertion/deletion 的操作，所以直接計算此區間子序列比對的分數 $S(Loop(P^A),Loop(P^B))$。

第二種情況為 $M_l[P^A, P^B]$，假設在此區間有對在一起的鹼基對存在。假設有兩個鹼基對P^A與$P^{A'}$且$P^{A'} <_I P^A$，則 $Loop(P^A)$可被分成三個部分（Figure 2-11）：（1）$LoopL(P^{A'}, P^A) = A[P_l^A+1 \ldots P_l^{A'}-1]$。

（2）$Loop(P^{A'}) = A[P_l^{A'}+1 \ldots P_r^{A'}-1]$。（3）$LoopR(P^{A'}, P^A) = A[P_r^{A'}+1 \ldots P_r^A-1]$。$M_l[P^A, P^B]$的計算方法如下：$M_l[P^A, P^B] = \max_{i,j}\{M_c[P_i^A, P_j^B] + w_3 \times S(LoopR(P_i^A, P^A), LoopR(P_j^B, P^B))\}$. (Figure 2-12)

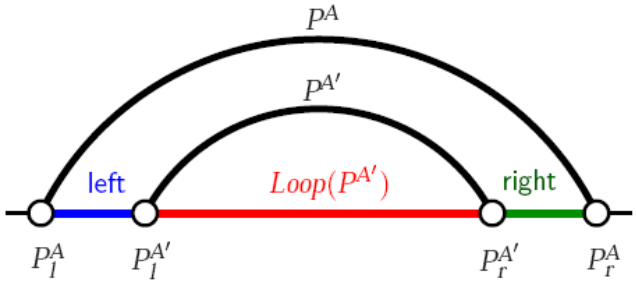


Figure 2-11: P^A與$P^{A'}$為巢狀關係，原本 $Loop(P^A)$ 被分成三個部分。右邊綠色部分為 $LoopR(P^{A'}, P^A)$；左邊藍色部分為 $LoopL(P^{A'}, P^A)$；中間紅色部分為 $Loop(P^{A'})$。

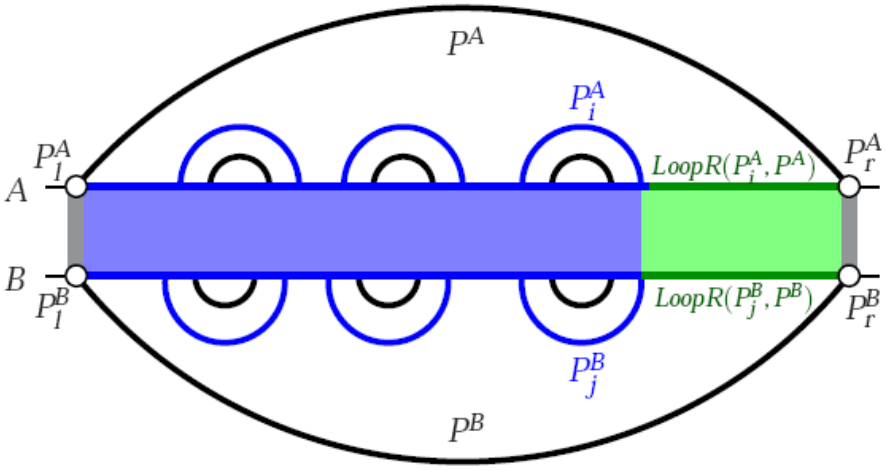


Figure 2-12: $M_l[P^A, P^B]$的示意圖。分數由兩部分組成：（1）綠色 $LoopR$ 區域的序列比對分數。（2）藍色 $M_c[P_i^A, P_j^B]$部分。

$M_l[P^A,\ P^B]$以最右側對起來的鹼基對P_i^A與P_j^B為準，將此區間分成兩個部分，第一個部分為P_i^A與P_j^B起以左計算 $M_c[P_i^A,\ P_j^B]$，P_i^A與P_j^B右側計算序列比對的分數。$M_c[P_i^A,\ P_j^B]$需再分成四種情況討論，取四種情況中最大值納入分數計算。介紹 $M_c[P_i^A,\ P_j^B]$前，需先定義$P_{i_1}^A \in F(P_{i_2}^A)$，$P_{i_1}^A \in F(P_{i_2}^A)$必須滿足$P_{i_1}^A <_p P_{i_2}^A$且在$P_{i_1}^A$與$P_{i_2}^A$之間沒有存在任何鹼基對$P_{i_3}^A$使得$P_{i_1}^A <_p P_{i_3}^A <_p P_{i_2}^A$。$M_c[P_i^A,\ P_j^B]$計算方法如下：

$M_c[P_i^A,\ P_j^B]=$

$$\max_{\substack{P_x^A \in F(P_i^A),\\ P_y^B \in F(P_j^B)}} \begin{cases} w_3 \times S(LoopL(P_i^A, P^A), LoopL(P_j^B, P^B)) + M(P_i^A,\ P_j^B), \\ M_c[P_x^A,\ P_y^B] + w_3 \times S(Loop(P_x^A, P_i^A), Loop(P_y^B, P_j^B)) + M[P_i^A,\ P_j^B], \\ M_c[P_i^A,\ P_y^B] + w_3 \times Gap(|Loop(P_y^B\ ,\ P_j^B)| + |Loop(P_j^B)| + 2), \\ M_c[P_x^A,\ P_j^B] + w_3 \times Gap(|Loop(P_x^A\ ,\ P_i^A)| + |Loop(P_i^A)| + 2). \end{cases}$$ (Figure 2-13)

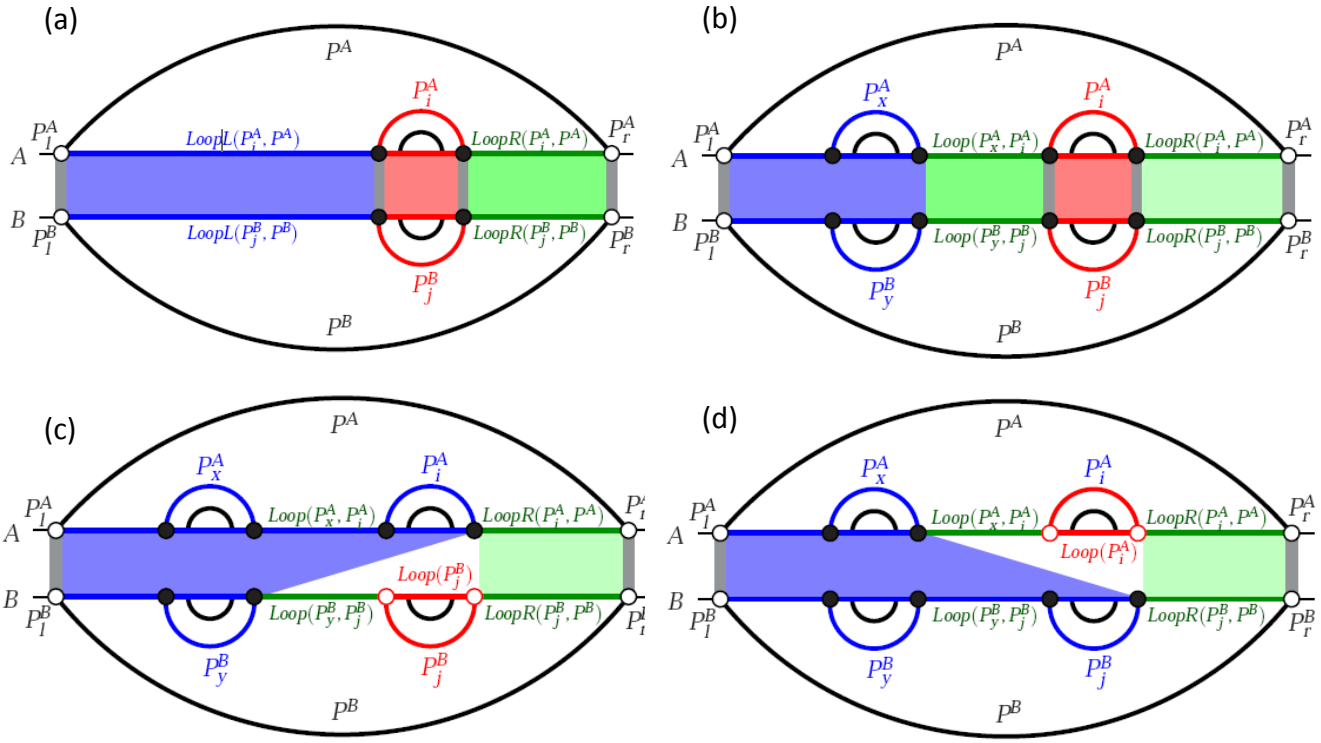


Figure 2-13: $M_c[P_i^A, P_j^B]$的示意圖。(a)~(d)依序為$M_c[P_i^A, P_j^B]$的第 1~4 種情況。

$M_c[P_i^A, P_j^B]$的第一種狀況討論在P^A與P^B包含的區間只有P_i^A與P_j^B有對在一起，於是在P_i^A與P_j^B左側（不包含P_i^A與P_j^B）計算序列比對的分數。另外，原本遞迴函式缺少P_i^A與P_j^B這個部分的最佳比對分數，我們在此將它修正。第二種情況討論在P^A與P^B包含的區間除P_i^A與P_j^B有對在一起之外還有其他有對在一起的鹼基對。於是分數的計算可分成三個部分，第一個部分是P_x^A與P_y^B起以左的部分繼續往左討論，第二個部分是介於P_x^A與P_y^B、P_i^A與P_j^B之間的序列比對分數；第三個部份是P_i^A與P_j^B這個部分的最佳比對分數；第三種情況討論P_i^A對到P_j^B左邊的P_y^B的情況，在分數的計算上可以分成兩部分，第一個部分是P_i^A與P_y^B起以左的部分繼續往左討論，第二個部分是 $Loop(P_y^B, P_j^B)$與從P_j^B左端鹼基到右端鹼基這段序列對到另外一股的空格的分數，在原本遞迴函式中空格長度少考慮了P_j^B的左端與右端鹼基，我們在此將其修正補上長度 2 的空格；第四種情況跟第三種情況相反，是討論P_j^B對到P_i^A左邊的P_x^A的情況，其中空格長度有錯誤的部分，我們亦將其修正。

Chapter 3

Results and Discussions

為了驗證將 RNAMotifScan 從使用一維序列修改成使用一維結構序列在辨識 RNA 結構模體上是否如我們預期的比原本的方法在執行效能上更有效率，我們實行兩組實驗來比較原本的方法與修改後的方法之間的差異。第一組實驗我們使用 RNAMotifScan 的論文中所記載的四個 RNA 結構模體：Kink-turn、Sarcin-ricin、C-loop 與 E-loop 當作 query，以及一個嗜鹽性古細菌（*H. marismortui*）的 23S rRNA（1S72）當作目標結構[31]。第二個實驗中，我們使用了一個來自 *H. m* . rRNA 的 RNA 結構模體－α-loop 當作 query 在另一個嗜熱性葛蘭氏陰性菌（*T. thermophilus*）的 30S rRNA（1N32）目標結構中尋找相似的結構模體。我們將這些 query 與目標結構當作兩方法的輸入，並且將兩方法的輸出結果加以比較與分析。實驗結果討論以 FPR 0.01 為基準，FPR 數值小於 0.01 的結果為高分結果並討論之，或 FPR 雖大於 0.01 但 RNAMotifScan 文獻有記載的 RNAMotifScan 結果亦

列入討論範圍。

3.1 在 *H. marismortui* 23S rRNA 中辨識多股 RNA 結構模體

Kink-turn

Kink-turn 是一個非對稱 internal loop，（Figure 3-1a），結構的特點與命名由來為 Kink-turn 的骨架片段其中一股有大角度的彎曲（Figure 3-1b）。Kink-turn 在 RNA 中是很重要的蛋白質辨識位置（Figure 3-1c）[14]。在這個實驗中當作 query 的 Kink-turn 結構模體來自 *H. marismortui* 的 23S rRNA [20]。

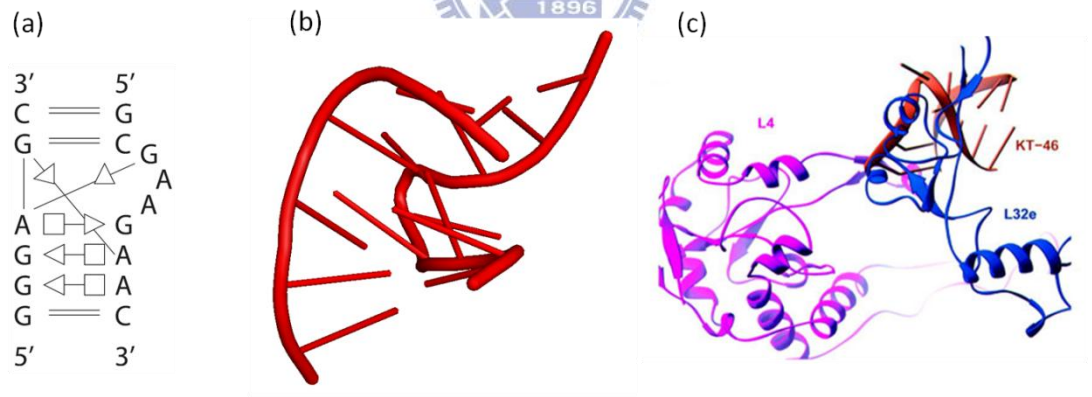


Figure 3-1: Kink-turn 結構模體。(a)query 的鹼基對示意圖。(b) query 的三級結構。(c) Kink-turn-蛋白質交互作用示意圖[14]。

我們的方法與 RNAMotifScan 個別都找到了其他文獻也記載的在 1S72 中的六個局部型 Kink-turn [14, 20, 22]。但是我們有四個結果

的 RMSD 比 RNAMotifScan 的結果小，比對的結果比較好（Table 3-1）。

Table 3-1: 我們的方法與 RNAMotifScan 分別以 Kink-turn 結構模體在 1S72 中搜尋相似結構模體高分的結果比較。

No.	Chain	Location	Our method		RNAMotifScan	
			RMSD	FPR	RMSD	FPR
1	0	77-82/92-100	**0.000**	**0.000**	3.240	0.000
2	0	936-941/1025-1034	2.466	0.000	2.466	0.000
3	0	1338-1343/1311-1319	1.996	0.000	1.996	0.000
4	0	1212-1217/1146-1155	**2.230**	**0.000**	4.030	0.000
5	0	1587-1594/1600-1608	**2.984**	**0.000**	5.699	0.000
6	0	244-250/259-267	**3.195**	**0.000**	4.198	0.000

由於 Kink-turn query 的出處即是 1S72，所以第一個搜尋結果便是 query 本身，RMSD 數值應該為 0。我們的結果是完全吻合的，但是 RNAMotifScan 結果的 RMSD 卻不為 0。我們分析發現，我們所做的實驗與 RNAMotifScan 文獻上所使用的 query 的鹼基對 pattern 是根據 Lescoute 於 2005 年所發表的文獻[20]， Lescoute 文獻中 Kink-turn 鹼基對 pattern 與 RNAVIEW 所註解的 pattern 一樣（Figure 3-2a），在我們所做的實驗中，目標結構的鹼基對是由 MC-Annotate 所註解（Figure 3-2b）。結果不同的鹼基對註解在這個區域出現一個鹼基對註解差異，由於 RNAMotifScan 的鹼基對在分數的計算上比重遠大於一維序列，所以 RNAMotifScan 選擇在 loop 區域插空格，使鹼基對能夠對起來（Figure 3-2c）。然而，我們的方法有三級結構資

訊的輔助，且三級結構資訊在結構的辨識比一維序列可靠，所以在分數計算上我們將三級結構的比重提高，使得縱使鹼基對的 pattern 有差異，但是我們三級結構資訊還是能夠反映出這個區域的結構是很相似的（Figure 3-2d）。

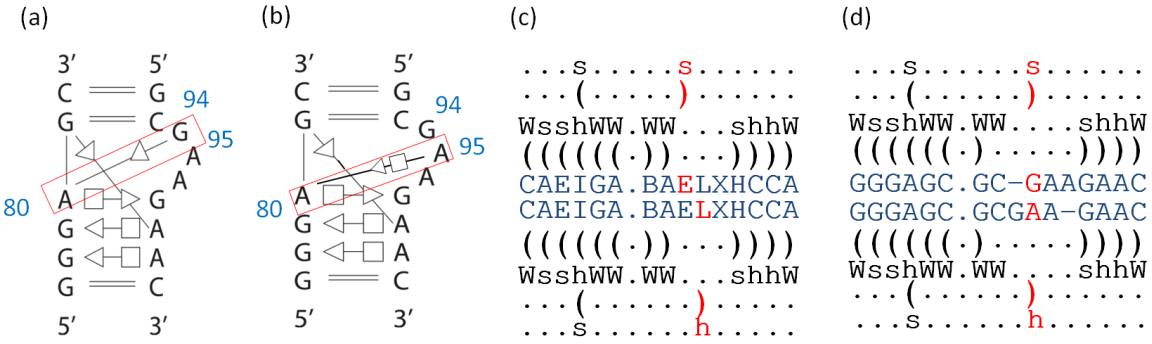


Figure 3-2: Kink-turn 第一個結果的比較。(a) query 的鹼基對。(b)目標結構的鹼基對。(a)與(b)鹼基對的差異為(a)在 80-94 有一個 *trans* S-S 的鹼基對，(b)是在 80-95 有一個 *trans* S-H 的鹼基對。(c)我們的比對結果。(d)RNAMotifScan 的比對結果。

在第四個結果中，目標結構有部分位於 stem 的鹼基對與 query 不像，RNAMotifScan 選擇在 query 插入較多空格，使 query stem 上的鹼基對能與 Figure 3-3b 上較外側的鹼基對對在一起。然而，我們的方法因為在此區域部分三級結構也是很像的，所以我們的方法沒有插入很多空格，捨棄掉部分鹼基對，但是比對的結果卻更好（Figure 3-3）。

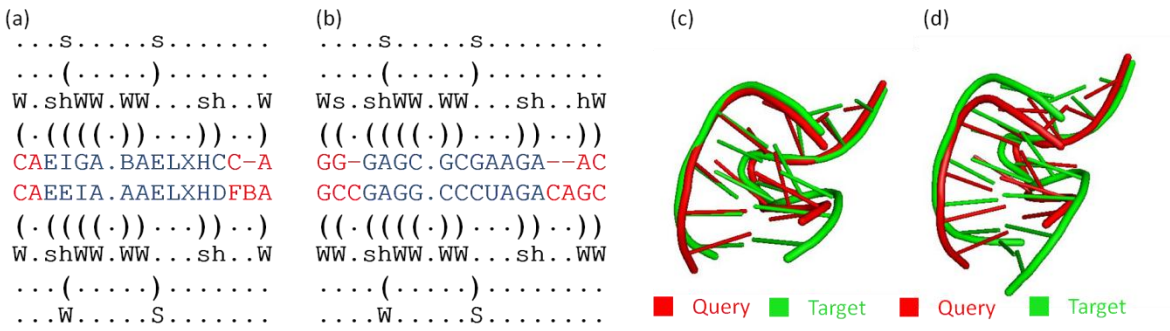


Figure 3-3: Kink-turn 第四個結果的比較。(a)我們的方法比對結果。(b) RNAMotifScan 的比對結果。(c)我們的方法三級結構比對結果。(d) RNAMotifScan 三級結構比對結果。

第五與第六個結果，我們的方法比對的結果能比 RNAMotifScan 好的原因，也是因為目標結構與 query 的鹼基對有部分不像，RNAMotifScan 選擇在 loop 區域插入許多空格。然而我們的方法，有辦法反映出在 loop 區域三級結構的相似性，所以在整體比對上，我們方法的表現較好（Figure 3-4 與 Figure 3-5）。在搜尋 Kink-turn 的結果顯示，我們的方法在因為註解上的差異或二級結構不像但三級結構像的例子能夠展現出其優勢。

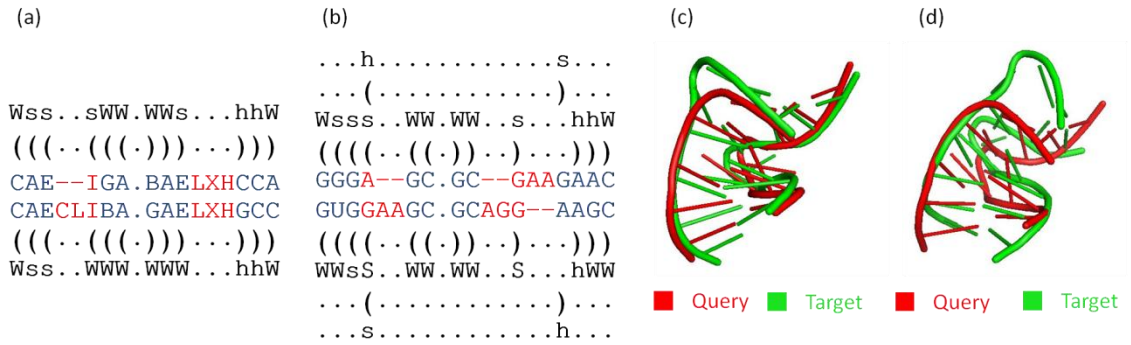


Figure 3-4: Kink-turn 第五個結果的比較。(a)我們的方法比對結果。

(b) RNAMotifScan 的比對結果。(c)我們的方法三級結構比對結果。

(d) RNAMotifScan 三級結構比對結果。

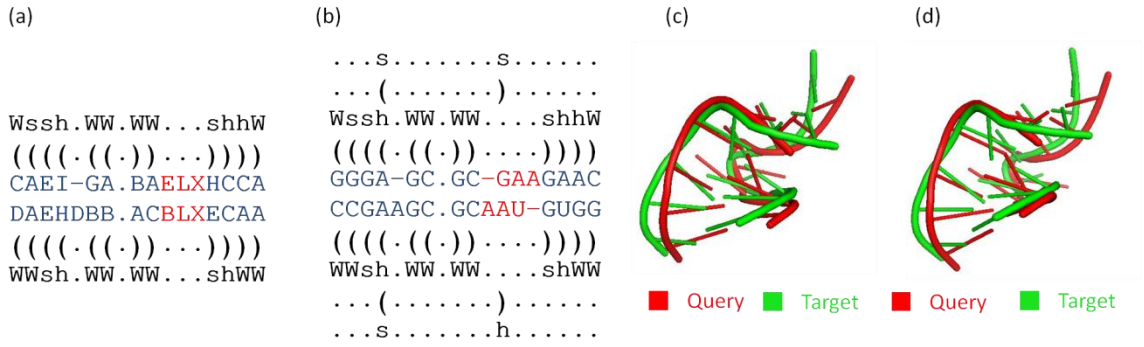


Figure 3-5: Kink-turn 第六個結果的比較。(a)我們的方法比對結果。

(b) RNAMotifScan 的比對結果。(c)我們的方法三級結構比對結果。

(d) RNAMotifScan 三級結構比對結果。

Sarcin-ricin

Sarcin-ricin 結構模體在核醣體的大次單位中是具有高度保留性的，尤其是在其中一股骨架片段有一個「S」型的彎曲為特點（Figure 3-6）。Sarcin-ricin 結構模體在蛋白質轉譯過程中會與延長因子（elongation factor）作用而使轉譯持續進行[25]。但是 Sarcin-ricin 結構模體名稱的由來是因為它的功能會受到 α-sarcin 與 ricin 兩個酵素的抑制[23, 31]。在這個實驗中當作 query 的 Sarcin-ricin 結構模體來自 *H. marismortui* 的 23S rRNA（PDB id: 1JJ2）（Figure 3-6）[16]。

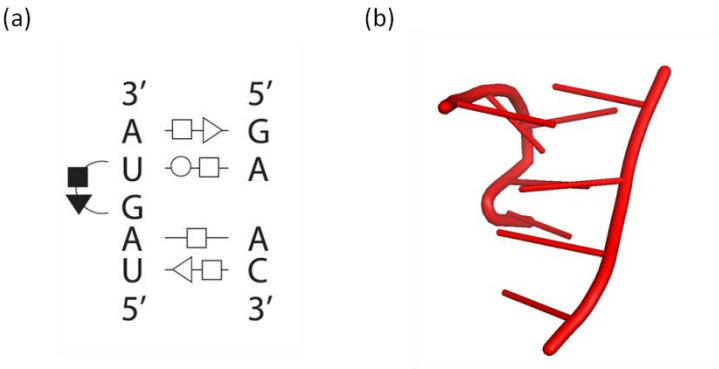


Figure 3-6: Sarcin-ricin 結構模體。(a) query 結構模體的鹼基對示意圖。(b) query 結構模體的三級結構。

我們的方法一共搜尋到 10 個 FPR 數值小於 0.01 的結果，且在這十個結果中除第十個結果 RMSD 略高之外，其它的結果 RMSD 皆小於 2，三級結構的相似程度也很高。RNAMotifScan 則找到了七個 FPR 小於 0.01 的結果。此外我們對照 RNAMotifScan 的文獻，發現 RNAMotifScan 亦另外列舉了六個 FPR 大於 0.01 的結果，在此我們將這六個結果也納入結果討論的範圍（以†記號註記），因此 RNAMotifScan 一共有 13 個結果（Table 3-2）。

在 Table 3-2 中第 11、12、13 個結果是 RNAMotifScan 有搜尋到，但是我們的方法沒有搜尋到的結果。這三個結果 RMSD 的數值皆較高，我們對照這三個三級結構比對的結果發現這三個目標結構模體與 query 並不相似。RNAMotifScan 會認為這三個結果是高分結果的原因是其鹼基對的 pattern 與 query 很像（Figure 3-7），但是一維序列比對卻無法反映三級結構相似與否。然而，我們的方法在使用一維

結構序列的結果之下，確實能夠排除掉這類鹼基對 pattern 相似，三級結構卻不像的結果。

Table 3-2: 我們的方法與 RNAMotifScan 分別以 Sarcin-ricin 結構模體在 1S72 中搜尋相似結構模體高分的結果比較。

No.	Chain	Location	Our method		RNAMotifScan	
			RMSD	FPR	RMSD	FPR
1	0	2690-2694/2701-2704	0.029	0.000	0.029	0.000
2	0	1368-1372/2053-2056	0.571	0.000	0.571	0.000
3	0	211-215/225-228	0.594	0.000	0.594	0.000
4	9	76-80/102-105	0.886	0.000	0.886	0.000
5	0	173-177/159-162	0.760	0.000	0.760	0.022 (†)
6	0	461-466/475-478	1.706	0.000	1.706	0.002
7	0	586-590/568-571	**0.938**	**0.000**	2.879	0.037 (†)
8	0	356-360/292-295	**1.078**	**0.002**	2.814	0.0369 (†)
9	0	380-383/406-408	0.838	0.002	0.838	0.004
10	0	1775-1779/1765-1768	3.595	0.007	3.595	0.020 (†)
11	0	951-955/1012-1016	–	–	4.099	0.006
12	0	2090-2094/2651-2654	–	–	4.897	0.018 (†)
13	0	1542-1545/1640-1643	–	–	3.392	0.037 (†)

符號註記：「-」代表未有結果。「†」表示 RNAMotifScan 文獻有記載但 FPR 數值大於 0.01 的結果。

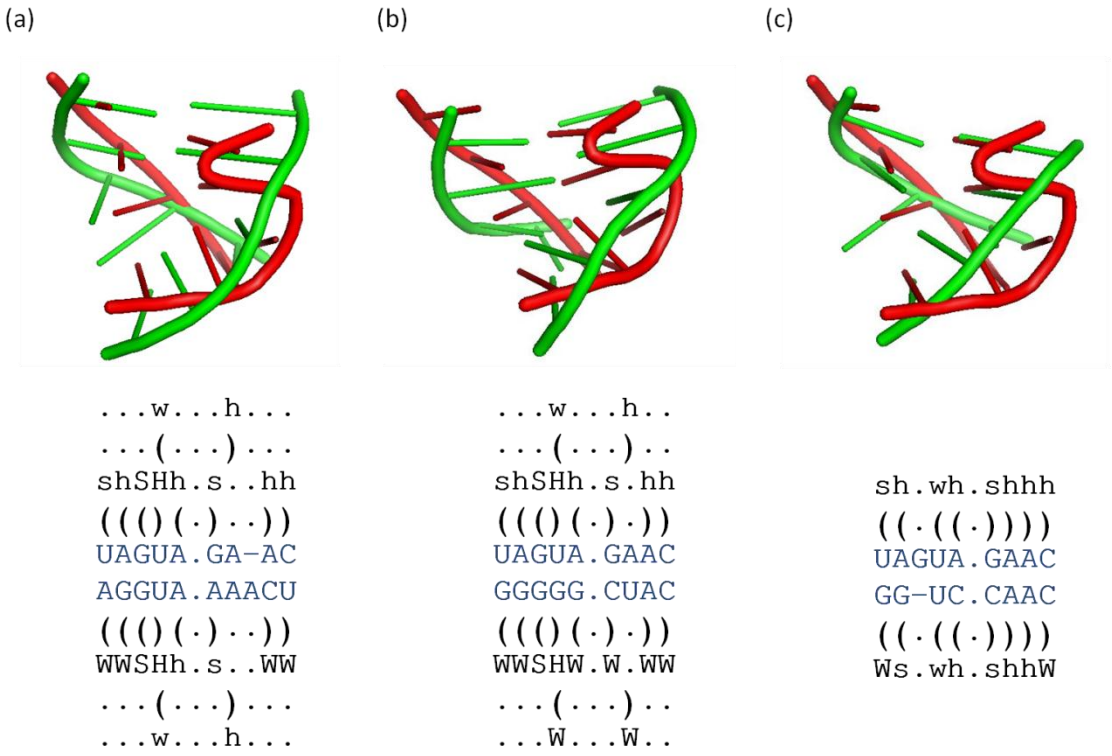


Figure 3-7: Sarcin-ricin 第 11、12、13 個比對結果與三級結構圖。(a)~(c)依序為第 11、12、13 個結果的三級結構圖（上）與比對結果（下）。

在第七與第八個結果，我們的方法比對的結果 RMSD 數值都遠比 RNAMotifScan 小，分別是 0.938 與 1.078。我們比對的結果比 RNAMotifScan 好的原因也是因為在目標結構模體中有部分的鹼基對不相似，RNAMotifScan 在 loop 區域插入了較多的空格，使得能有較多鹼基對對在一起。我們的方法能在鹼基對不相似的情況下，利用三級結構的資訊凸顯出整體結構的相似性（Figure 3-8 與 Figure 3-9）。

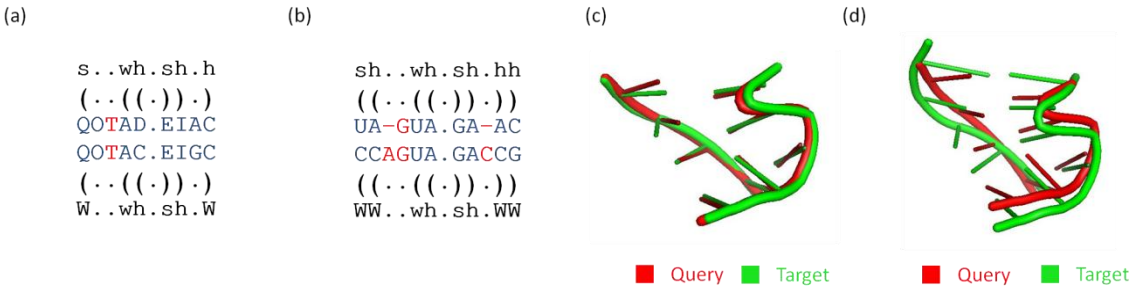


Figure 3-8: Sarcin-ricin 第七個結果的比較。(a)我們的方法比對結果。(b) RNAMotifScan 的比對結果。(c)我們的方法三級結構比對結果。(d) RNAMotifScan 三級結構比對結果。

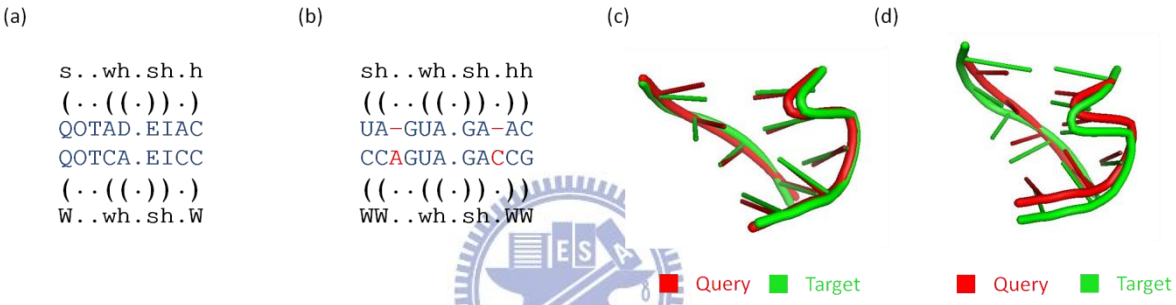


Figure 3-9: Sarcin-ricin 第八個結果的比較。(a)我們的方法比對結果。(b) RNAMotifScan 的比對結果。(c)我們的方法三級結構比對結果。(d) RNAMotifScan 三級結構比對結果。

C-loop

C-loop 是非對稱的 internal loop。在生物體內，C-loop 是重要的 RNA-蛋白質結合位，在 16S rRNA[3]、23S rRNA[1]與大腸桿菌的蘇胺醯基-tRNA 合成酶（threonyl-tRNA synthetase）的 mRNA 5 端-不轉譯區域（5’-UTR）[26]中皆可發現其蹤影。C-loop 的特色在較長的一

股骨架片段的 loop 區域有兩個胞嘧啶分別形成交叉的鹼基對[19]（Figure 3-10）。在這個實驗中當作 query 的 C-loop 結構模體來自 *E. coli* 的蘇胺醯基-tRNA 合成酶 mRNA（PDB id: 1KOG）（Figure 3-10）[16]。

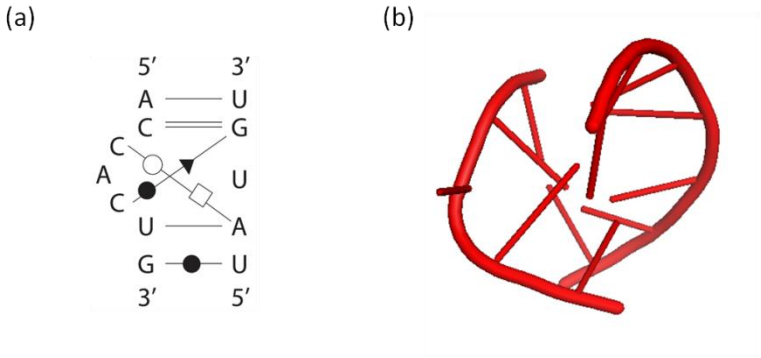


Figure 3-10: C-loop 結構模體。(a) query 結構模體的鹼基對示意圖。(b) query 結構模體的三級結構。

在我們的方法中，我們一共找到四個 FPR 低於 0.01 的結果；RNAMotifScan 找到五個（Table 3-3）。其中第一到第三個結果，兩方法的結果都是一致的，第四個以我們的方法比對的結果，RMSD 數值比 RNAMotifScan 比對的結果小，比對的結果較好（Figure 3-11）。第五個結果，以 RNAMotifScan 比對是高分的結果，但是以我們的方法比對的則非高分結果。

Table 3-3: 我們的方法與 RNAMotifScan 分別以 C-loop 結構模體在 1S72 中搜尋相似結構模體高分的結果比較。

No.	Chain	Location	Our method		RNAMotifScan	
			RMSD	FPR	RMSD	FPR
1	0	02760-02764/02716-02722	1.288	0.000	1.288	0.000
2	0	01939-01945/01892-01898	2.813	0.000	2.813	0.000
3	0	01436-01440/01424-01430	1.330	0.000	1.330	0.000
4	9	28-31/49-54	2.582	0.000	4.725	0.008
5	0	01004-01009/0957-0964	2.914	0.152 (‡)	1.957	0.000

符號註記:「‡」表示我們的方法有找到但 FPR 數值大於 0.01 的結果。

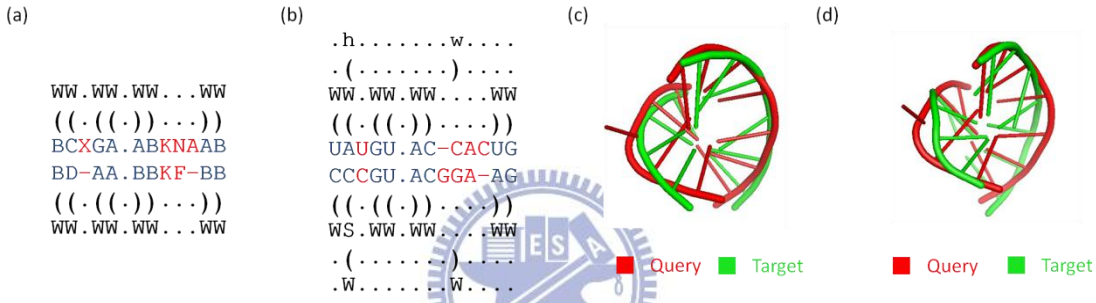


Figure 3-11: C-loop 第四個結果的比較。(a)我們的方法比對結果。(b) RNAMotifScan 的比對結果。(c)我們的方法三級結構比對結果。(d) RNAMotifScan 三級結構比對結果。

第五個結果所搜尋到的目標結構模體是 Leontis 於 2003 年也有發表過的 C-loop 一種變型[19]。在這個變型的 C-loop 中（Figure 3-12b），於兩股骨架片段的 loop 區域都各比傳統型 C-loop（即本實驗所使用的 query；Figure 3-10a 與 Figure 3-12a）多了一個鹼基。在鹼基對的部分，變型的 C-loop 比傳統型在 958-961 多了一個 *trans* -WC 鹼基對，其他鹼基對則與傳統型一致，因此 RNAMotifScan 認

為這一個目標結構模體與 query 結構模體很相似（Figure 3-13b）。但是目標結構模體在 loop 區域多出來的兩個鹼基，使得 loop 區域變長並且使兩股骨架在此有很大的彎曲角度（Figure 3-13c 與 Figure 3-13c 綠色）。我們的方法認為這兩個彎曲角度與 query 並不相似，在比對分數的計算上扣分較多，於是比對分數較低，FPR 評估的數值變高。然而，我們的結果 RMSD 比 RNAMotifScan 高的原因是在較長股 loop 區域插空格位置的差異所造成，在我們的一維結構序列中，N 對 W 與 N 對 P 的置換分數計算是一樣的（Figure 3-13a），換言之，空格如果插到左邊分數不會改變，這樣比對的結果便與 RNAMotifScan 一樣，RMSD 數值也會一樣。

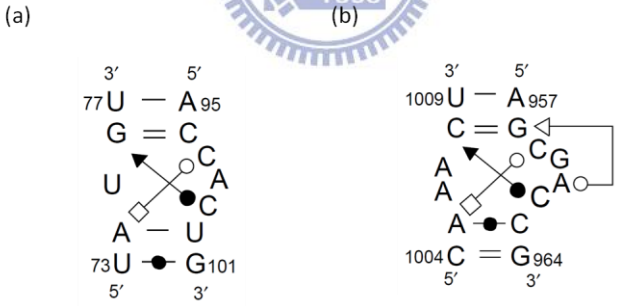


Figure 3-12: 傳統型與變型 C-loop[19]。(a)傳統型 C-loop。(b)變型 C-loop，在 loop 區域兩股骨架片段都各比傳統型多一個鹼基。

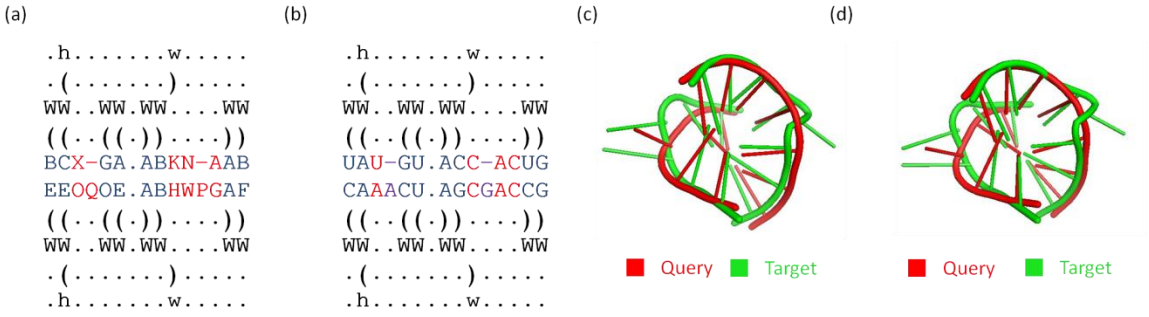


Figure 3-13: C-loop 第五個結果的比較。(a)我們的方法比對結果。(b) RNAMotifScan 的比對結果。(c)我們的方法三級結構比對結果。(d) RNAMotifScan 三級結構比對結果。

E-loop

E-loop 結構模體以 2D 的角度去看是一個對稱 internal loop。在大腸桿菌與葉綠體的 5S rRNA 中可發現。在這個實驗中當作 query 的 E-loop 結構模體來自 *E. coli* 的 5S rRNA（PDB id: 1PNX）（Figure 3-14）[19]。

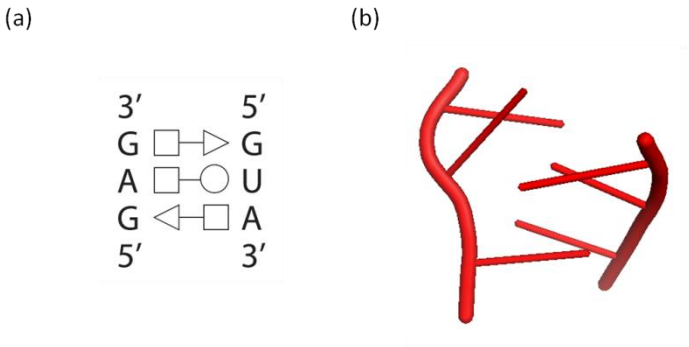


Figure 3-14: E-loop 結構模體。(a) query 結構模體的鹼基對示意圖。(b) query 結構模體的三級結構。

我們的方法一共找到六個 FPR 數值低於 0.01 的結果，RNAMotifScan 可找到十個（FPR 數值低於 0.012）。如 Table 3-4 所示，兩方法一樣的結果只有第一與第二個。其餘搜尋到的結果差異很大。在這些結果中，我們的方法搜尋到的結果的 RMSD 皆比 RNAMotifScan 的結果低。我們分析兩方法的三級結構比對結果發現，RNAMotifScan 的結果除第八個外，其他的目標結構模體都與 query 不甚相似（Figure 3-15）。然而，我們的結果，目標結構模體的骨架比對結果都與 query 很相似（Figure 3-16）。依實驗結果所示，我們的方法在處理像 E-loop 這類序列長度較短，鹼基對個數較少的結構模體辨識比 RNAMotifScan 更精準。

Table 3-4: 我們的方法與 RNAMotifScan 分別以 E-loop 結構模體在 1S72 中搜尋相似結構模體高分的結果比較。

No.	Chain	Location	Our method		RNAMotifScan	
			RMSD	FPR	RMSD	FPR
1	0	1543-1545/1640-1642	1.071	0	1.071	0.005
2	0	706-708/720-722	0.673	0.001	0.673	0.006
3	0	816-818/795-797	2.241	0.001	–	–
4	0	23-25/518-520	2.392	0.001	–	–
5	0	1339-1341/1316-1318	2.230	0.004	–	–
6	0	2502-2504/2516-2518	2.567	0.009	–	–
7	0	174-177/159-161	–	–	2.734	0.01
8	0	663-666/680-683	–	–	2.919	0.012
9	0	586-590/568-571	–	–	3.568	0.012
10	0	356-360/292-295	–	–	3.591	0.012
11	0	2691-2694/2701-2703	–	–	2.787	0.012
12	0	1369-1372/2053-2055	–	–	2.723	0.012
13	0	463-466/475-477	–	–	2.633	0.012
14	0	380-383/406-408	–	–	2.720	0.012

符號註記：「-」代表未有結果。

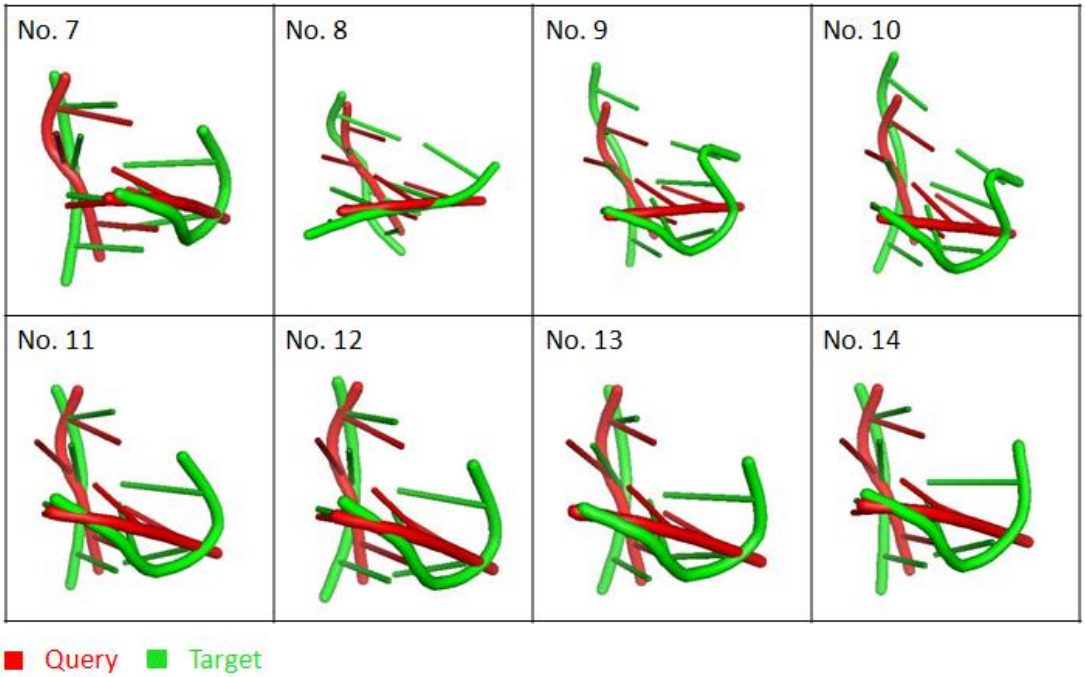


Figure 3-15: RNAMotifScan 的 E-loop No.7~14 三級結構比對結果。

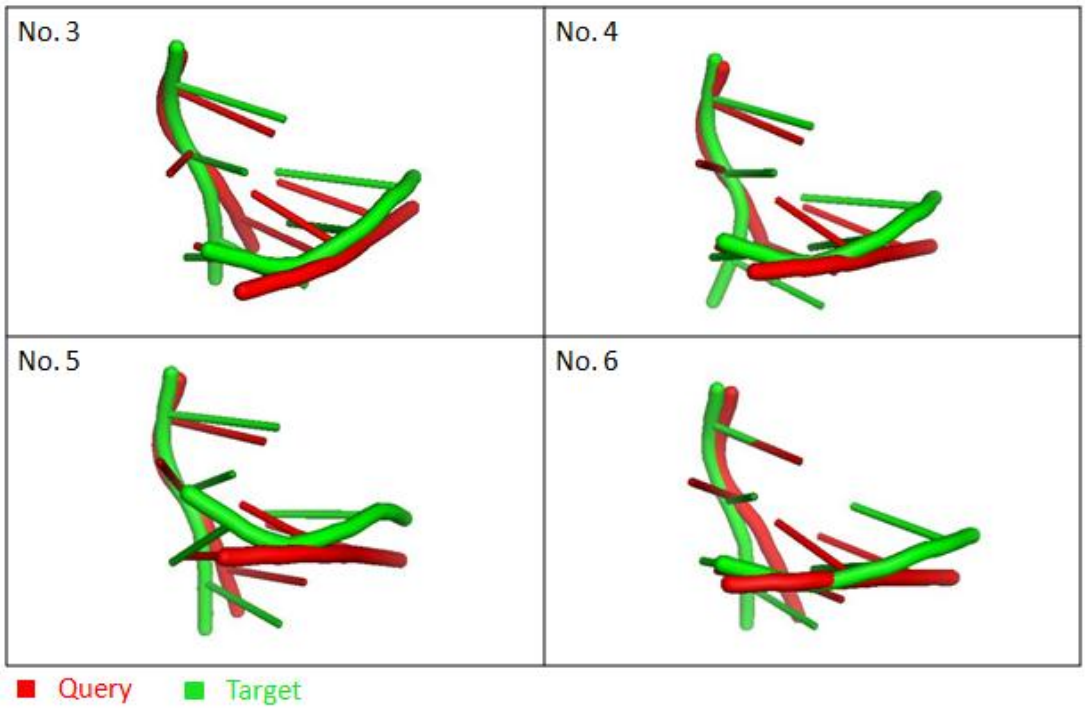


Figure 3-16: 我們的 E-loop No.3~6 三級結構比對結果。

3.2 在 *T. thermophilus* 30S rRNA 中辨識單股α-loop 結構模體

α-loop 的名稱由來是它的三級結構外型就像「α」字母。α-loop 一共有八個鹼基，其中六個鹼基構成 loop 區域（Figure 3-17）。α-loop 在 23S 與 16S rRNA 中可以發現到，α-loop 可以被提供輔助大小次單元核醣體組合的蛋白質所辨識，以幫助核醣體正確組合[4, 12]。本實驗取 COMPADRES[28]所記載的兩個α-loop 其中之一當作 query（PDB id 為 1JJ2，0 chain，位置 1100-1107）搜尋另一個在 *T. thermophilus* 30S rRNA 中的α-loop。

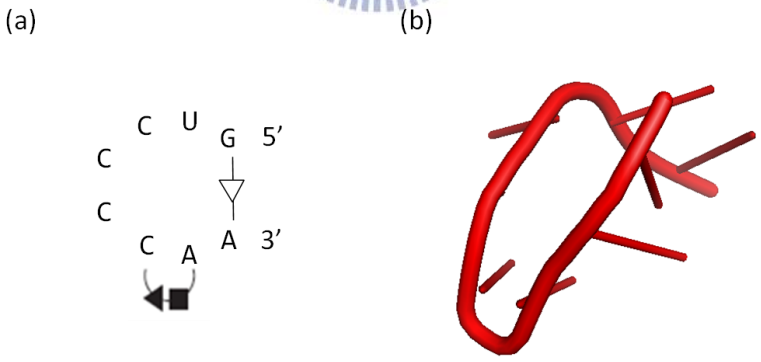


Figure 3-17: α-loop 結構模體。(a) query 結構模體的鹼基對示意圖。(b) query 結構模體的三級結構。

RNAMotifScan 一共找到兩個結果（Table 3-5）。在我們分析其第一個結果的三級結構後，我們認為第一個是錯誤結果，因為三級

結構並不相似（Figure 3-18b）。但是 RNAMotifScan 認為這一個結果是最高分的原因是因為在目標結構模體裡的兩個鹼基對與 query 的兩個鹼基對都是 match 的（Figure 3-18a）。比起第二個正確答案的目標結構模體只有一個鹼基對與 query 是 match 的分數高很多（Figure 3-19a 與 3-19b）。因為我們的方法能夠反映出 loop 區域的相似性（Figure 3-19a），所以能準確地辨認出正確答案，排除掉錯誤的答案。經過這個實驗所示，我們的方法在辨識 loop 區域較長且鹼基對數目相對較少的結構模體比 RNAMotifScan 能更精準。

Table 3-5: 我們的方法與 RNAMotifScan 分別以 α-loop 在 *T. thermophilus* 30S rRNA 中搜尋相似結構模體高分的結果比較。

No.	Chain	Location	Our method		RNAMotifScan	
			RMSD	FPR	RMSD	FPR
1	A	568-574	–	–	3.630	0.000
2	A	503-510	1.048	0.000	1.048	0.000

符號註記：「-」代表未有結果。

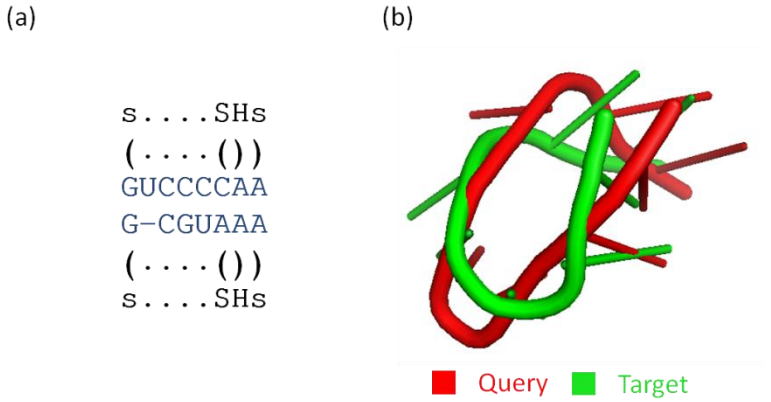


Figure 3-18: RNAMotifScan 的第一個結果。(a)比對結果，兩組對在

一起的鹼基對皆為 match。(b)比對結果的三級結構圖。

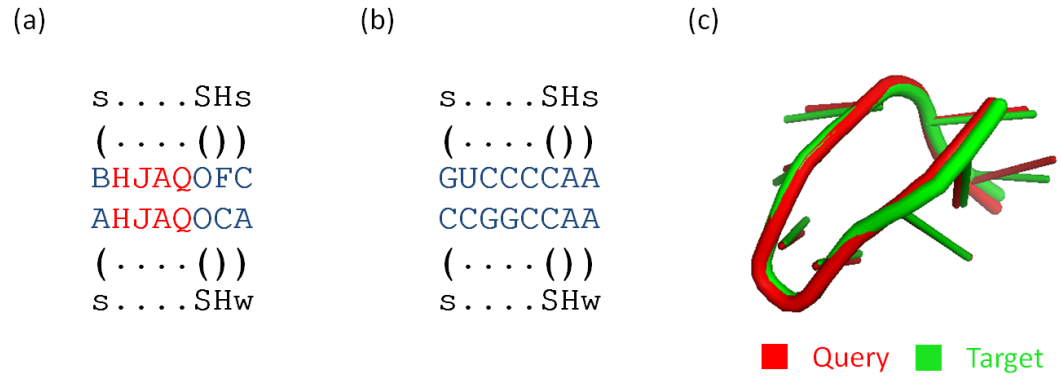


Figure 3-19: 我們的結果與 RNAMotifScan 的第二個結果。(a)我們的比對結果，loop 區域的結構序列完全 match。(b) RNAMotifScan 的比對結果。(c) 比對結果的三級結構圖。

Chapter 4

Conclusions

在這個研究中，我們首先將 RNA 的二級與三級結構資訊編碼成弧線註解的一維結構序列，接著再把 RNAMotifScan 的程式修改成利用 RNA 二級與三級結構資訊來辨識 RNA 的結構模體。最後，我們的實驗結果也顯示出我們的方法確實可以進一步地提升 RNAMotifScan 在辨識 RNA 結構模體的執行效能。然而，如本論文的 Introduction 中所述，RNAMotifScan 在做弧線註解序列的比對時，對鹼基對（弧線）而言，它只考慮了 match 與 mismatch 而已，事實上還有其它鹼基對比對方式（如 Figure 1-3 所示的 arc-breaking、arc-altering 與 arc-removing）。因此，如何在弧線註解一維結構序列兩兩比對方法中將所有的鹼基對比對方式都納入考慮將會是一個有趣且值得研究的問題。

References

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

2. Capriotti, E., Marti-Renom, M.A. (2010) Quantifying the relationship between sequence and three-dimensional structure conservation in RNA. *BMC Bioinformatics*, 11, 322.

3. Clemons, W.M. Jr, Brodersen, D.E., McCutcheon, J.P., May, J.L., Carter, A.P., Morgan-Warren, R.J., Wimberly, B.T. and Ramakrishnan, V. (2001) Crystal structure of the 30S ribosomal subunit from *Thermus thermophilus*: purification, crystallization and structure determination. *J. Mol. Biol.*, 310, 827-843.

4. Culver, G.M. (2003) Assembly of the 30S ribosomal subunit. *Biopolymers*, 68, 234-249.

5. 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.

6. 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 Res*., 31, 4755–4761.

7. Ferrè , 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 Res*., 35, W659–W668.

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

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

10. Hendrix, D.K., Brenner, S.E. and Holbrook, S.R. (2005) RNA structural motifs: building blocks of a modular biomolecule. *Q. Rev. Biophys*., 38, 221–243.

11. 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.

12. Herold, M. and Nierhaus, K.H. (1987) Incorporation of six additional proteins to complete the assembly map of the 50S subunit from Escherichia coli ribosomes. *J. Biol. Chem.*, 262, 8826-8833.

13. Jiang, T., Lin, G., Ma, B. and Zhang, K. (2002) A general edit distance between RNA structures. *J. Mol. Biol.*, 9, 371–388.

14. Klein, D.J., Schmeing, T.M., Moore, P.B. and Steitz, T.A. (2001) The kinkturn: a new RNA secondary structure motif. *EMBO J*., 20, 4214-4221.

15. Leonits, N.B., Lescoute, A. and Westhof, E. (2006) The building blocks and motifs of RNA architecture. *Curr. Opin. Struct. Biol*., 16, 279–287.

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

17. Leontis, N.B., Stombaugh, J., Westhof, E. (2002) The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.*, 30, 3497-3531.

18. Leontis, N.B. and Westhof, E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, 7, 499-512.

19. Leontis,N.B. and Westhof,E. (2003) Analysis of RNA motifs. *Curr. Opin. Struct. Biol*., 13, 300–308.

20. Lescoute, A., Leontis, N., Massire, C. and Westhof, E. (2005) Recurrent structural RNA motifs, isostericity matrices and sequence alignments. *Nucleic Acids Res*., 33, 2395–2409.

21. Liu, Y.C., Yang, C.H., Chen, K.T., Wang, J.R., Cheng, M.L., Chung, J.C., Chiu, H.T. and Lu, C.L. (2011) R3D-BLAST: a search tool for similar RNA 3D substructures. *Nucleic Acids Res*., 39, W45-W49.

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

23. Spackova, N. and Sponer, J. (2006) Molecular dynamics simulations of sarcin-ricin rRNA motif. Nucleic Acids Res., 34, 697-708.

24. Storz, G. (2002) An Expanding Universe of Noncoding RNAs. *Science*, 296, 1260−1263.

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

26. Torres-Larios, A., Dock-Bregeon, A.C., Romby, P., Rees, B., Sankaranarayanan, R., Caillet, J., Springer, M., Ehresmann, C., Ehresmann, B. and Moras, D. (2002) Structural basisof translational control by *Escherichia coli* threonyl tRNA synthetase. *Nat. Struct. Biol.*, 9, 343-347.

27. 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.

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

29. Wang, C.W., Chen, K.T., Lu, C.L. (2010) iPARTS: an improved tool of pairwise alignment of RNA tertiary structures. *Nucleic Acids Res*., 38, W340-W347.

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

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