# 國立交通大學

## 資訊科學系

## 碩 士 論 文

# 利用共同結構元實作核醣核酸分群

# RNA clustering based on common structure elements


研究生：王美華

指導教授：胡毓志 博士


中 華 民 國 九 十 三 年 六 月

# 利用共同結構元實作核醣核酸分群

研究生：王美華　　指導教授：胡毓志博士

國立交通大學資訊科學研究所

## 摘要

本研究提出一個有關核醣核酸分群研究的新議題，針對一群未排比的核醣核酸序列，同時進行核醣核酸的分群與其結構元的預測。屬於相同家族的核醣核酸必定存在某些共同的特徵，而本研究選用共同的二級結構元作為分群的依據。我們的方法是一套反覆式的分群程序。首先，採用監督式學習(supervised learning)為訓練核心，預測出某家族的共同結構元後，再利用此結構元作為鑑定同源關係的準則，擁有此結構元的核醣核酸會被歸為同一家族。將這些序列分離出去，再重複整個程序，直到分完所有的群集。透過此系統的分析檢測，最後可以了解此組資料中包括幾群不同的核醣核酸家族、每一個家族的成員以及代表每個家族的共同結構元。此系統能準確地預測出一致性程度較高的結構元，Matthews 相關係數值可高達 0.85。雖然容易受到在非結構元區域的配對結構所影響，但依然可挑到大部份的成員，可有 0.83 擷取率。我們由實驗結果分析得知，結構元的長相一致性程度與非結構元區域是否會形成配對結構，是影響整個分群與結構元預測結果的主要兩個因素。

# RNA clustering based on common structure elements

Student：Mei-Hua Wang　　Advisor：Yuh-Jyh Hu

Institute of Computer and Information Science

National Chiao Tung University

Hsinchu, Taiwan, Republic of China

## ABSTRACT

In this study, we introduce a novel topic about RNA clustering. For a set of unaligned RNA sequences, simultaneously cluster the related RNAs together and predict the common structure element of each group. Correlated RNAs must contain some common features, and we choose the common structure element as the clustering basis. Our method is an iterative clustering procedure. At first we apply supervised learning to predict a common structure element, and then identify the homologous RNAs using this signature. After separating out those RNAs covered by this structure element, repeat the overall procedure. Finally, we can know the number of the clusters, all members and the common structure element of each cluster. Our system can correctly predict the highly conserved structure elements, the Matthews correlation coefficient is up to 0.85. Although it is sensitive to the structures in non-motif region, it can identify most of the members and has a 0.83 recall. The results suggest that structure element consensus level and the sequence similarity in non-motif region are two important factors in RNA clustering and common structure elements prediction.

# 致謝

能夠完成這篇論文研究，首先，必須感謝教授在這兩年來的指導與鼓勵。由於老師平時的督促與專業知識的傳授，讓我對生物資訊這領域有更深刻的了解。

另外，之所以能夠完成這麼多耗時的實驗，得謝謝實驗室的林宛嫻同學與可愛的陳音璇學妹、超搞笑的賴昀君與吳秉蔚以及正直的林勁伍學弟出借了他們的電腦，得以讓我及時完成所有的實驗。也非常感謝江萬田學長給予精神上的鼓勵與寶貴的建議。

最後，很感激家人的支持，使我能在無顧之憂的情況下，完成我的學業。除此之外，還有男朋友施並格平時不厭其煩的加油打氣，甚至是陪我一起找出那煩人的 bug 並且給予珍貴的技術指導。

謝謝大家!!

# 目錄


摘要.....i

Abstract.....ii

致謝.....iii

目錄.....iv

第一章　前言.....1

1.1 研究動機.....1

1.2 研究假設.....3

1.3 研究目的.....4

1.4 系統功能.....5

1.5 論文架構.....6

第二章　核醣核酸結構介紹.....7

2.1 核醣核酸的重要性.....7

2.2 核醣核酸結構基本單位元.....8

2.3 核醣核酸結構基本組成與種類.....10

第三章　文獻探討.....12

3.1 為何需要電腦的輔助.....12

3.2 為何尋找核醣核酸結構元.....14

3.3 核醣核酸共同結構尋找的相關方法.....15

3.4 核醣核酸資料庫.....18

第四章　研究方法.....20

4.1 系統設計目的與概念.....20

4.2 GPRM.....23

4.2.1 核醣核酸結構描述語言.....23

4.2.2 GPRM 模型架構.....26

4.2.3 細部修改.....27
4.3 系統主架構.....28
4.4 架構說明.....29
4.4.1 設定程式執行參數.....29
4.4.2 前置處理.....29
4.4.3 尋找代表性的結構元.....31
4.4.4 挑選家族成員.....40
4.4.5 後置處理.....42
第五章 實驗結果.....50
5.1 實驗項目.....50
5.2 測試資料.....51
5.3 執行介面與參數設定.....53
5.4 實驗流程.....55
5.5 評估方式.....58
5.6 實驗結果.....60
5.6.1 非結構區域序列相似度影響.....60
5.6.2 同源核醣核酸其二級結構相似度影響.....62
5.6.3 不同家族成員個數比例影響.....64
第六章 結論與未來研究方向.....65
6.1 結論.....65
6.2 未來研究方向.....67
6.2.1 參數設定太過寬鬆.....67
6.2.2 拉普拉斯門檻值的決定依據.....68
6.2.3 執行時間太過冗長.....69
6.2.4 負面背景序列的產生方式.....69
第七章 參考文獻.....70

# 第一章　前言

## 1.1 研究動機

人類基因定序是生物科技發展上的一大躍進，伴隨而來的難題是，該如何分析這麼大量的資料。計算式演算法陸續地提出來，希望藉由電腦的補助，發掘隱藏在這巨量資料中的重要訊息，如演化發展史中親嗣關係、蛋白質功能的確認、基因與蛋白質之間的相互影響，在生物體中所引起的一連串效應，而最終的目標還是期望能找出遺傳疾病產生的主因。

生命體內複雜的生化反應或生物的特徵等主要是由蛋白質來控制、決定，而蛋白質的產生則是依據去氧核醣核酸(DNA)序列中所隱含的遺傳密碼經由轉錄到核醣核酸(RNA)序列，再透過轉譯的過程形成蛋白質產物。藉著分析去氧核醣核酸、核醣核酸與蛋白質序列，生物學家便能更了解蛋白質的特性，甚至於整個生命系統運作的機制。

將生物物質分門別類收集，是一種常用來減低研究複雜度的方式。例如，蛋白質的分類是生物資訊領域中一個重要的議題，其主要目的在於明瞭蛋白質的功能、縮短藥物發明的時間。相同的，若希望能更快速地知道核醣核酸的功能特性，亦可先從分群的工作著手。而大部份蛋白質的分類方法，是根據序列間的相似度或是相似片段。儘管標記出蛋白質序列中共同片段的方法已發展成熟，這些演算法卻不適用於核醣核酸序列的分析，因為核醣核酸的生物功能是由它所形成的結構來決定(Pley *et al.*, 1994; Scott *et al.*, 1995; Ekland *et al.*, 1996)，因此，核醣核酸的分類必須架構在擁有相同二級結構的條件上。

由於核?(ribozyme)的發現，顛覆了生物學家對核醣核酸的刻板印象，也颳起了一股研究熱潮。然而，那些議題大多著眼於核醣核酸結構的預測。本研究以不同的角度切入此領域，研究目的主要定位於核醣核酸的分群。我們認為，每個家族的核醣核酸都具備其代表性的二級結構元，故可利用此結構元來搜尋屬於此家族的核醣核酸。以此觀點為基礎，我們希望能針對一群核醣核酸序列進行分群，並預測每一個家族共同的二級結構。

## 1.2 研究假設

在設計整體架構，作業流程時，我們以兩個基本假設為出發點，來設計演算來解決我們所定義的問題：

【假設一】 相同家族的核醣核酸之間，存在某些共同的特徵，其中二級結構元是一項重要的指標。

既然決定核醣核酸功能的主要因素為其結構（Pley *et al.,* 1994; Scott *et al.,* 1995; Ekland *et al.,* 1996），則擁有相似功能的核醣核酸應具備相同的二級結構元；因此，我們可以利用此特點來判別哪些核醣核酸是屬於相同家族的，達到分群的目的。

【假設二】 決定核醣核酸功能的結構元，不會出現在隨機產生的序列中。

這是一個合理的假設，因為在演化過程中，雖然可能產生結構的突變，但為了確保核醣核酸其功能與價值，這種關鍵性的位置與形狀，會一直被保留下來。因此，若是具有演化意義的結構應當不會出現在任意的序列中。此假設也被應用在 GPRM 的研究上(Hu 2002)。

## 1.3 研究目的

2002 年，本實驗室提出 GPRM 系統(Hu 2002)，主要是利用基因規畫方法來尋找核醣核酸的共同結構。GPRM 的特色是，不必具備專業領域的知識(domain knowledge)，並且能夠直接以二級結構當作演化的目標，同時省去了基因演算法中編碼、轉碼等瑣碎工作。這種方式不但簡單亦能達到良好的預測結果。再加上使用的結構表示語言十分具有彈性，除了基本的核醣核酸二級結構外，亦可以偵測出擬節結構。而本研究欲針對一群未排比的核醣核酸序列，但序列之間的關係及本身的二級結構皆是未知的，能取得的資訊只有序列的內容，根據上述的假設，設計一套有系統的流程，同時進行核醣核酸的分群與其結構元的預測。

## 1.4 系統功能

藉由此篇論文，我們希望能提供下列功能：

(1) 提出新的研究方向

核醣核酸相關研究是一個熱門的領域，但至今，尚未有核醣核酸分群的相關討論。本篇論文希望能由不同的方向切入此領域，提出新的研究議題，使得核醣核酸的實驗研究能夠更加充足與豐富。

(2) 整理核醣核酸資料

大量的生物資料無法單純以人力整理與分析，這樣的做法亦不夠精確與客觀，因此，發展一套自動化的分析系統是刻不容緩的。我們以資訊科學的方法，設計一套處理程序，以核醣核酸的二級結構為分群依歸，將混雜的核醣核酸資料分門別類後，能夠更清楚地看出資料中隱含的相同與相異處。

(3) 助於生物學家進行相關實驗

若能取得分析整理過後的資料，生物學家便能專注於主要的實驗探討，而且對這些統整後的資料，生物學家可以只取感興趣的部份，不必受其它無關的資訊所影響，產生預期外的結果，故能快速且正確地進行核醣核酸相關的實驗。

(4) 提供方便使用的介面

除了核心系統外，我們還設計一個方便使用的網頁介面，其它研究者可利用此工具分析他們輸入的資料，以助其實驗研究的進行。

## 1.5 論文架構

此篇論文包含六大章節，第一章為前言，主要介紹此議題的研究動機、此流程方法的基礎假設，以及主要的研究目的。第二章則是簡單地介紹核醣核酸的結構、種類。在第三章，整理了目前預測核醣核酸二級結構常用的方法，針對不同的演算法，點出其優缺點。第四章是本篇主要重點，詳細介紹本研究針對核醣核酸分群與結構元預測同時進行所提出的解決方案。在第五章，則整理本研究所進行的實驗結果。最後，第七章是本研究所參考的相關文獻。

# 第二章　核醣核酸結構介紹

## 2.1 核醣核酸的重要性

“生命的起源”，長久以來是大家探討的重要議題之一。許多理論陸續被提出來，其中最為生物學家所接受的是『RNA 世界』假說(Gilbert，1986)的觀點。在最初的世界，只存在具有催化作用的核醣核酸分子，即 T. Cech 在 1986 年發現的核? (Ribozyme)(Cech *et al.*, 1986)。Schwartz 更指出，核醣核酸是第一個顯示出自我復製(self-replication) 及演化等生命現象的生物物質(Schwartz，1995)。

從遺傳學觀點來看，核醣核酸所發揮的功能非常簡單。在合成蛋白質的過程中，攜帶由去氧核醣核酸所傳遞過來的遺傳指令，在遺傳的程序裡，扮演著”遺傳信使”的角色。但，自從 T. Cech 發現核醣核酸亦有生物催化的功能後，打破了傳統生物學”？的本質就是蛋白質”的論點。自此之後，生物學家認為，核醣核酸的作用並非只是單純地傳達遺傳訊息而已。

在某些生化反應中，核醣核酸也表現出”控制”的功能。例如，控制生物體內蛋白質的生物合成、開放或關閉某些基因、以及增加或減少某個去氧核醣核酸片段，使一種基因能夠合成出多種的蛋白質；若合成出異常的蛋白質，則會引發各種不同的疾病。

## 2.2 核醣核酸結構基本單位元

在細胞中，核醣核酸不僅擁有調控、轉錄、轉譯等重要功能，有些甚至具有酵素的功能。而我們已知，核醣核酸的功能與其結構息息相關，結構的多樣性讓核醣核酸具備多重的生物功能。因此，相較於序列的分析，核醣核酸的結構是生物學家渴望了解的範疇。以下，先簡單介紹核醣核酸的結構單位元：

核醣核酸是由四種核甘酸小分子組成的聚合物，分別是腺飄呤(adenine)，胞嘧啶(cytosine)，鳥飄呤(guanine)，連嘧啶(uracil)，常以 A、C、G、U 來代表這四種核甘酸。

最基本的核醣核酸結構可簡單分為成對與不成對，G≡C 和 A＝U 會形成標準的鹼基對(Canonical base pair)，亦稱為互補作用。在核醣核酸二級結構中出現的大多數是這種華特生-克立克(Watson-Crick)配對。另一種是 G－U 擺動對(wobble pair)，這是非標準的鹼基對(non-canonical base pair)，也常出現在核醣核酸二級結構中。G－U 配對有它獨特的化學、結構及蛋白上受體結合(ligand-binding)的特性，因此，擁有這些 G－U 配對的結合處，是蛋白質或其它核醣核酸辨識的目標，在許多生化反應中亦發揮極重要的功能(Varani *et al.*, 2000)。

G≡C 鹼基對是以三個氫鍵所鍵結而成，A＝U 則是形成兩個氫鍵，G－U 之間只有一個氫鍵，這些鍵結力便是形成核醣核酸二級結構的主要因素。在核醣核酸結構模型中，G≡C 鹼基對最為穩定，A＝U 次之，最小的便是 G－U 擺動對。

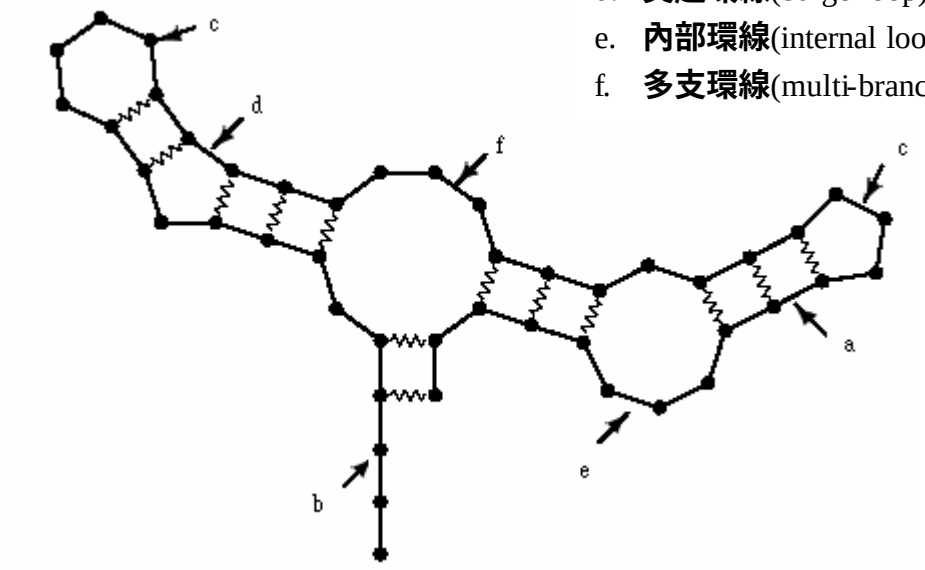
不同於去氧核醣核酸，核醣核酸是單股分子，它的二級結構是由雙股螺旋片段與單股區域交插組合而成。雙股螺旋區則是因為單股分子會互相摺疊所造成的，也

就是自身互補(self-complementary)區域。要產生這樣的雙股區域，必須在核醣核酸序列下游(downstream)的連續鹼基與上游(upstream)的連續鹼基互補形成華特生-克立克配對或者 G – U 擺動對。

## 2.3 核醣核酸結構基本組成與種類

儘管核醣核酸是單股序列，由於氫鍵鍵結，使得分子會摺疊回來形成 G≡C、A＝U 與 G－U 的配對，更組成多樣化的核醣核酸二級結構。圖表 1 顯示的便是核醣核酸結構的基本部份。

a. **莖幹**(stem)
b. **單股區域**(single stranded region)
c. **髮夾環線**(hairpin loop)
d. **突起環線**(bulge loop)
e. **內部環線**(internal loop)
f. **多支環線**(multi-branched loop)




**圖表 1．核醣核酸結構基本組成**

**莖幹(stem)**

在核醣核酸序列中，若有鹼基能夠與相同序列中反向互補的鹼基利用氫鍵鍵結，並且堆疊在它周圍的鹼基對上，形成穩定的 A-型(A-form)雙股螺旋(Dock-Bregeon *et al.,* 1989)，這一個連續區域稱之為莖幹。

**單股區域(single stranded region)**

單股區域是由未能形成配對的鹼基所組成的，在核醣核酸序列的兩端中未能形成結構。

**髮夾環線(hairpin loop)**

由莖幹夾擠起來的單股區域稱為環線(loop)，其中常出現於核醣核酸二級結構是一種髮夾環線，它是由一個莖幹與一個未形成配對的環線所組成，又稱為 U 型轉彎(U-turn)，若環線中只有兩個鹼基，則稱做急劇 U 型轉彎(sharp U-turn)。

**突起環線(bulge loop)**

在莖幹中若有一邊出現未配對的鹼基，另一邊是連續的鹼基對，則此結構稱為突起環線。

**內部環線(internal loop)**

莖幹兩邊若都有未配對的鹼基出現，則稱為內部環線。內部環線又可再細分為對稱性(symmetrical)與非對稱性(asymmetrical)，兩邊未配對的鹼基數目相同者是對稱性內部環線；反之，則是非對稱性內部環線。

**多支環線(multi-branched loop)**

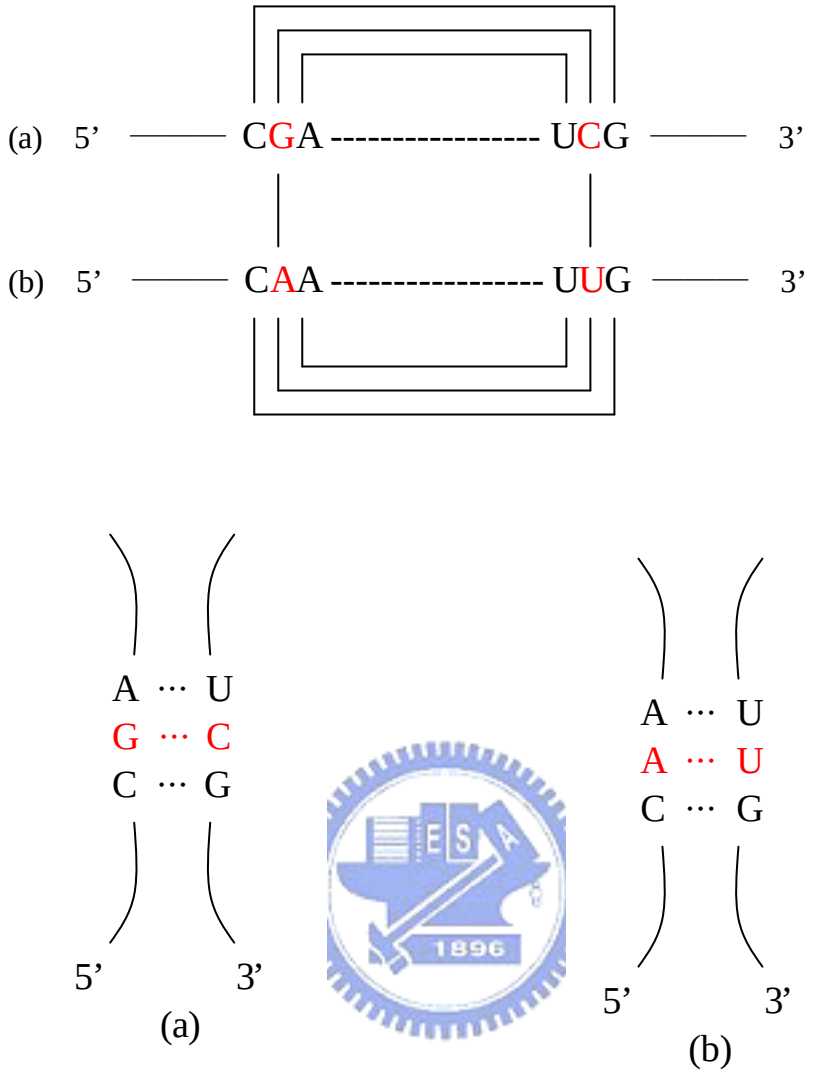
多支環線擁有三個以上的莖幹，這些莖幹又被長度不等的單股區域所分隔開來，形成放射狀的結構。

# 第三章　文獻探討

## 3.1 為何需要電腦的輔助

利用 X-ray 繞射對核醣核酸結構的決定有極高的解析度，但此種方法的最大的問題在於，很難獲得穩定的核醣核酸結晶。目前最常用來探測核醣核酸結構的方法是利用酵素對核醣核酸做部份水解，辨別雙股與單股區域，當判斷核甘酸具有單股或雙股的摺疊後，便能推測其二級結構(Ehresmann et al.，1987)，但因為酵素之間會互相干擾或有互相矛盾的結果，提高了實驗的困難度。近來，更利用一系列的刪除式突變來獲得配對資訊，但某些點突變無法造成結構的改變，且大量的點突變是不符合經濟效益的，也就無法加速核醣核酸結構的確認。

另一方面，親緣分析目前被認為預測核醣核酸結構最準確的方法，因為是在活體內進行核醣核酸分子的演化，故推測出的結構是較有可能在細胞內的核醣核酸結構。此種分析方法是假設，具有生物功能的核醣核酸不易有結構上的變化，在演化過程中便被保留下來。倘若在雙股區域的鹼基產生突變，則相配對的鹼基須發生補償性的突變(compensatory mutation)，否則可能會失去其功能而無法流傳到子代，這種突變稱之為協同變異(covariation)。圖表 2 便是協同變異的一個例子，在 5’端的鹼基 G 突變成 A 時，3’端的鹼基 C 便要突變成 U，如此才能維持相同的二級結構。




**圖表 2．協同變異**

**當電腦程式能夠快速並準確地找出共同的結構元，便能加快核醣核酸結構的研究，減少人力、物力的浪費。**

## 3.2 為何尋找核醣核酸結構元

生物學家會將某些核醣核酸序列歸類成相同的家族，是因為它們有相同的表現型，也就是有相同的生化功能。而就目前生物學家研究了解，核醣核酸的二級結構是控制它們表現結果的關鍵。例如，蛋白質的合成須要轉錄者核醣核酸(tRNA)的參與，作為連接分子，把信使核醣核酸(mRNA)中的密碼編譯成蛋白質的氨基酸，在基因表現中發揮重要作用。這類的核醣核酸分子生物特性十分明確，它的多樣特性與它的結構密切相關。因此，在同一家族的核醣核酸序列中，找出重覆出現的二級結構有助於生物學家進一步了解核醣核酸，推測出核醣核酸結構與功能之間的關係。

轉錄者核醣核酸是一種結構已經確定的分子，在目前已知的轉錄者核醣核酸中，皆可發現一致的苜蓿芽結構。除了轉錄者核醣核酸，在各個界(kingdom)中，核糖體核醣核酸(ribosomal RNA)也有非常一致的結構元，因此常用來建立親緣關係圖(Antoinette et al., 2002)。

## 3.3 核醣核酸共同結構尋找的相關方法

目前已知，具有相同生物功能的核醣核酸，並非在序列上擁有相似的片段，而是在結構具有一致性。然而，核醣核酸共同結構的尋找卻比去氧核醣核酸分析來得困難，因為核醣核酸整體結構相同並不代表序列內容相同，因為共變(covariation)的現象使得序列的突變也不至於造成結構的改變。這種自然的特性，使得鑑定核醣核酸二級結構的發展較為遲緩，儘管如此，依舊吸引許多學者紛紛加入此研究領域。

過去預測核醣核酸二級結構的方法有很多，大致可分為熱力學模型(thermodynamic)與序列比較分析(comparative sequence analysis)兩大類。若能取得同源的核醣核酸序列，則序列比較分析方法所預測出來的結構會比能量最佳化的結果來得可靠。因此，常用來建立已知是同家族的核醣核酸序列的結構。傳統的做法是，在一組已排比完成同源的序列中，偵測出有代表性的雙股區域，有些區域甚至可能會出現互補性突變。主要關鍵是，利用 $c^2$ – 統計量(Chi-square statistics)(Chiu *et al.*, 1991)或互見訊息(mutual information)的多寡(Gutell *et al.*, 1992) 作為每個鹼基對共變的證據，量測發生共變的位置，再找出共同的二級結構。

此類方法雖然可以辨認出擬節結構，但最大的困難在於，如何獲得良好的序列排比結果。有些核醣核酸序列只有些微相似，但結構上卻是一致的，因此，良好的序列比較分析方法，應該容許序列的差異卻保有結構上的相同，故序列排比時，應該考慮結構的資訊。然而，諸如 CLUSTALW(Thompson *et al.*, 1994)這類做排比的工具，並不能執行結構的排比。現在許多演算法做排比時會一起考慮序

列與結構資訊(Kim *et al.*, 1996; Notredame *et al.*, 1997)，或是排比與結構的預測同時進行(Eddy *et al.*, 1994; Gorodkin *et al.*, 1997; Hofacker *et al.*, 1998; Gorodkin *et al.*, 2001)。但這些方法受限於序列的長度，只能處理較短的序列或者要求其中某一個序列的結構是已知的。共變現象的可靠度必須參照親屬遠近關係，故亦一併考慮序列的親緣關係(Gulko *et al.*, 1996; Akmaev *et al.*, 1999; Parsch *et al.*, 2000)。

基因演算法是一種隨機最佳化的方法，運用的是達爾文的『適者生存』的概念，已廣泛地應用在各個領域中。在核醣核酸二級結構預測的議題中，亦可利用基因演算法搭配熱力學模型(Chen *et al.*, 2000)來尋找共同結構元。熱力學的觀點是以結構的自由能(free energy)當作適應度評估標準(fitness criterion)。雖然不須要做序列的排比，但自由能的計算也有它不足之處。因為這須要參考核醣核酸結構的相關知識，即使考慮了細部的條件，有些結構，利用這些能量模型的計算規則與參數(Mathews *et al.*, 1999)計算出來的能量並非是最小的(最穩定)。且能量計算公式與參數值只是實驗估計值，無法完整考慮細節部份，更不能代表生物體內的實際狀態，例如熱力學模型假設核醣核酸的結構處在熱力平衡的狀態，與環境條件無關。然而，許多核醣核酸會與蛋白質結合，而造成自由能的改變。

第三類用來尋找核醣核酸二級結構元的方法是從正規語言的角度出發(Sakakibara *et al.*, 1994; Kundsen *et al.*, 1999; Kundsen *et al.*, 2003)。不同於蛋白質序列，核醣核酸每個鹼基之間並非完全獨立，兩個位置若能形成配對，則這兩個鹼基便有很高的相依性，而隨機式前後文無關性文法(Stochastic context-free grammars ，SCFG)可以描述序列上遠距離的相依關係，恰能表示核醣核酸序列獨有的共現特性。這種機率模型在訓練完成後即代表此核醣核酸家族

的共同結構，故可再用於資料庫的搜尋，但訓練過程中須要序列排比的資訊，因此它的成效受制於排比的結果；除此之外，它最令人詬病的是，此種方法不能偵測出擬節結構(此於 4.2.1 核醣核酸結構描述語言將有詳細介紹)。

## 3.4 核醣核酸資料庫

由於技術的進步，已知結構的核醣核酸數量快速地成長，為了有系統地整理這些分散在各個文獻的資料，提高核醣核酸研究的便利性，資料庫建構的成了另一個熱門的研究方向。目前公開的資料庫已相當的多，以下簡單介紹幾個核醣核酸相關的資料庫。

(1) SCOR (Klosterman *et al.,* 2002; Tamura *et al.,* 2004)

核醣核酸多樣化的摺疊型態反應了在生物體內多功的特性，SCOR 資料庫的建立，提供了一個研究核醣核酸功能、二級結構元與三級結構之間關係的管道。截至 2003 年 5 月，SCOR 收集了 497 筆核醣核酸結構，為了不同的使用目的，以生物功能、二級結構元與立體結構作分類依據。生物功能有如轉錄者核醣核酸(tRNA)、核糖體核醣核酸(ribosomal RNA)與核酸代? (ribozyme)等等；簡單的二級結構元分類成髮夾型環線與內部環線；三級結構則有擬節結構、環線與環線間的作用(loop-loop interaction)等等。

(2) The RNA Structure Database (Murthy *et al.,* 2003)

RNABase 資料庫整合了 Protein Data Bank(PDB)(Berman *et al.,* 2000)與 Nucleic Acid Data Base(NDB) (Berman *et al.,* 1992)兩者的核醣核酸資料，再依功能與結構的不同來作分類。每一筆結構資料都包括了簡短的總結、描述立體結構的參數值、完整的摺疊構象圖示(Ramachandran-style conformational map)等等，除了提供相關資料外，還可以執行結構的分析與檢測。

**(3) PseudoBase (Batenburg *et al.,* 2000)**

**在各種核醣核酸二級結構中，擬節結構是最不易預測的，因為這種結構難以描述定義，預測擬節結構的自動化程式工具還是很少，相關的演算法雖相繼提出(Riva *et al.,* 1999; Ruan *et al.,* 2004)，但計算複雜度依舊偏高，預測正確率不高或是只限定特殊的擬節結構；但，這種結構又具有相當重要的生物功能(Dam *et al.,* 1992; David *et al.,* 2000)，因此，促使了許多相關的資料庫的建立。**

**PseudoBase 資料庫收集了擬節結構的核醣核酸相關資料，包括了序列、結構與生物功能三類資訊。每一筆資料再細分成 12 個子項目，例如擬節結構在序列的位置、EMBL 存取序號等等，提供研究擬節結構的資料來源。**

**(4) 其它相關資料庫**

**其它另有收集特定核醣核酸相關資訊的資料庫，如 Nucleic Acid Database(NDB) (Berman *et al.,* 1992)、 tRNA Compilation 2000 (Sprinzl *et al.,* 1998)、 SRPDB (Signal Recognition Particle Database) (Alm Rosenblad *et al.,* 2003)與 RNase P Database (Brown, 1999)等等。**

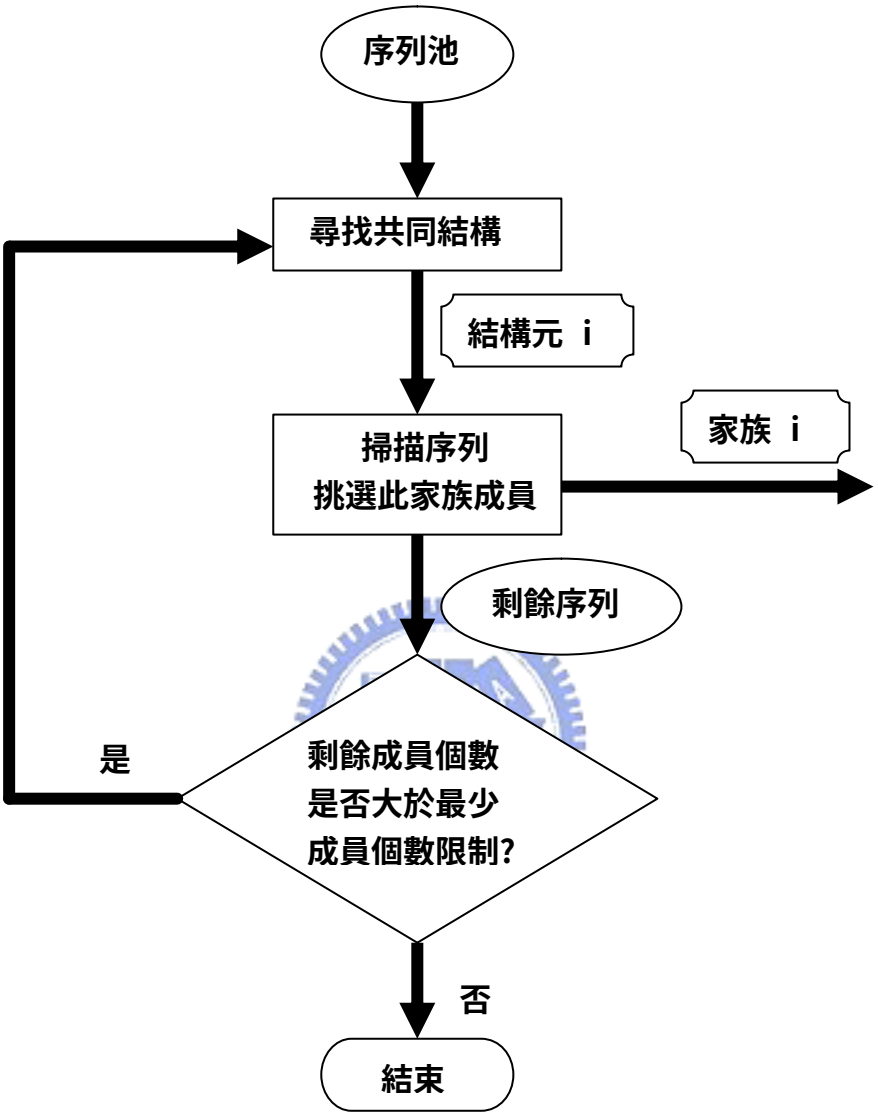
# 第四章　研究方法

## 4.1 系統設計目的與概念

核醣核酸在生物體內有許多種不同的功能，而這些功能大多取決於核醣核酸的二級結構。若能知曉每一條核醣核酸之間的關係，便能獲得許多有價值的資訊。例如，由二級結構已確定的核醣核酸，透過親緣關係的分析判斷，預測另一條核醣核酸的摺疊結構。甚至是藉由已知生化功能與摺疊結構的核醣核酸，利用結構相似度的比較，推測另一條核醣核酸在生物體上可能發揮的功用。然而，每兩條核醣核酸都執行結構比對，是一種沒有效率的作法。最佳的方法是將這些序列先進行分群，分群的依據便是核醣核酸的二級結構。

本研究希望能提出一套有系統的程序，同時進行核醣核酸的分群與其結構元的預測。透過這樣的分析，整理大量的核醣核酸序列資料，助於生物學家進行核醣核酸實驗研究。鑑於 GPRM 在尋找核醣核酸共同結構的成功，本研究將 GPRM 稍做修改後，與我們的系統整合在一起，藉由 GPRM 來尋找每一群核醣核酸共同的二級結構元。由於本研究的重點，並非提出一套全新的結構描述語言，故仍沿用 GPRM 系統所採用的表示方式。

在本篇論文研究中，我們將此系統定位為一套核醣核酸分群的工具，以監督式學習(supervised learning)為訓練核心，預測共同的結構元。再利用此結構元作為鑑定同源關係的準則，達到分群的目的。圖表 3 即為本系統簡略的流程圖。


序列池
尋找共同結構
結構元 i
掃描序列
挑選此家族成員
家族 i
剩餘序列
剩餘成員個數
是否大於最少
成員個數限制?
是
否
結束

**圖表 3．系統流程圖**

此種分群方式類似 CN2 演算法(Clark and Niblett, 1989)，這是機器學習(machine learning)領域中常用的規則庫建構法。當尋找出一條規則法後，便將符合此規則描述的範例挑出去，此後便不再參與其它規則的推導。

本篇研究的基本假設是，相同家族的核醣核酸會擁有相似的二級結構。我們希望能先找出某個有意義的二級結構元，再以此結構元進行序列的掃描，辨識相同家族的核醣核酸。擁有此結構的核醣核酸便歸為同一家族，並將這些核醣核酸從原本的資料中分離出來，故不會影響之後的結構元預測與家族分群的作業。重複這樣的程序直到分完所有的群集。

以下便先簡單介紹 GPRM 系統以及針對本研究的須求而修改的部份，再概述本研究的整體流程，最後，再針對各步驟做詳細的解說。

## 4.2 GPRM

GPRM 為本實驗室在 2002 年所提出的一套系統，此研究提出一套十分有彈性的核醣核酸結構表示語言，並能針對一個核醣核酸家族，利用基因規畫方法尋找其共同結構。

### 4.2.1 核醣核酸結構描述語言

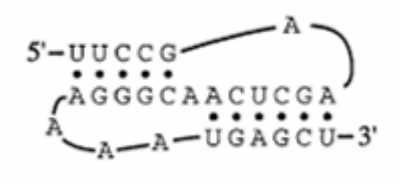
GPRM 使用三種參數來定義核醣核酸的二級結構，即莖幹的個數、莖幹在結構中的相對位置以及每個莖幹及環線的長度範圍。

(1) 莖幹個數

表示在此結構中含有幾段雙股區域。

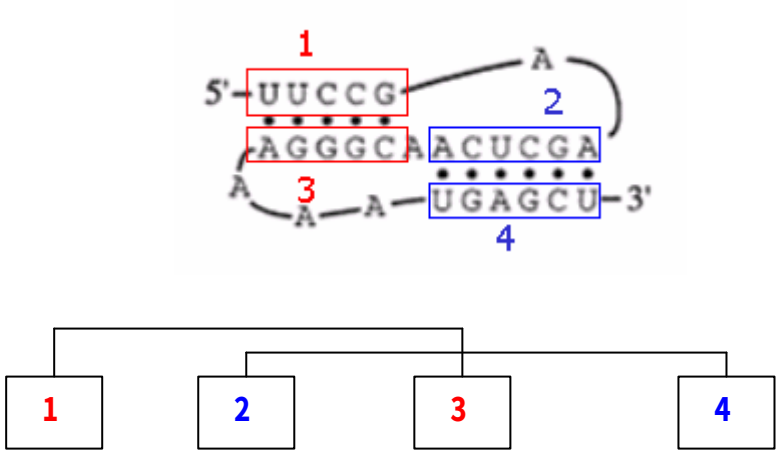
(2) 莖幹在結構中的相對位置

若將所有的雙股區段由 5’端到 3’端的順序加以編號，如 1,2,3,…,2N，其中 N 為莖幹個數，則各種區段間的配對組合便形成了各種不同的結構。如圖表 4 即為過去許多語言所無法描述的擬節結構(pseudoknot)。這種簡單的結構是由兩段單股環線連接兩段雙股區域所組成，即所謂的 H-式/(H-type)髮夾型(hairpin)擬節，由圖可看出，每一個髮夾環線緊接著一段單股區域。




圖表 4．擬節結構

以此種結構為例，由 5’端到 3’端將所有莖幹編為 1 號，2 號，3 號及 4 號區段。擬節結構則是 1 號與 3 號，2 號與 4 號區段形成雙股結構，如圖表 5 所示。此種配對方式即是 soil-borne mosaic virus 家族共同的二級結構。

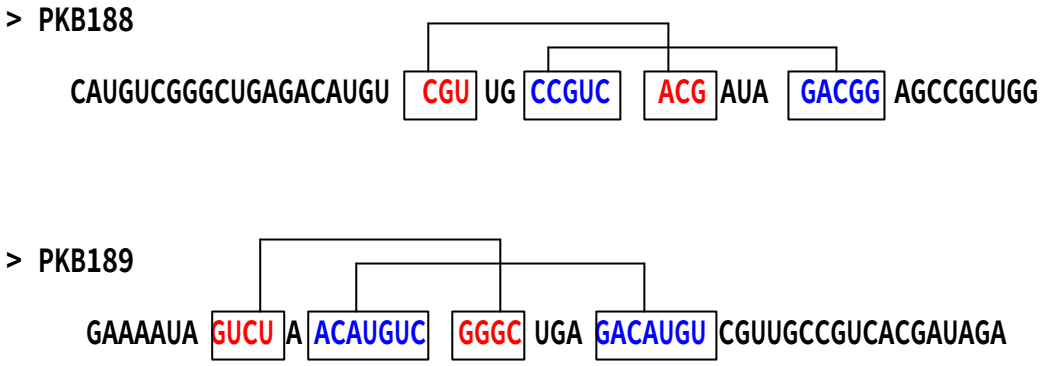



圖表 5．擬節結構中莖幹的相對位置

(3) 每個莖幹及環線的長度範圍

GPRM 並不考慮末端的單股區域，因此，每個結構的開始與結束皆是莖幹，且兩個莖幹之間會被一個環線所隔開。由於不同的核醣核酸雖擁有相同的二級結構，但所形成的雙股區域長度可能有些微的差異，因此，使用長度上限與下限來指定莖幹與環線在此家族中可能出現的長度，只要在長度範圍[Min, Max]內者都算是合法的莖幹與環線。Min 是指此莖幹最短的長度，Max 則是最長可出現的長度。

圖表 6 為 soil-borne mosaic virus 中的兩條核醣核酸序列，由此圖可看出，不同的核醣核酸雖然擁有相同的二級結構，但莖幹與環線的長度卻不同。因此，若以 GPRM 對 soil-borne mosaic virus 家族預測它的共同二級結構，所得到的長相為 [3,6] (0,3) [4,10] (0,1) [3,6] (0,6) [4,10]。



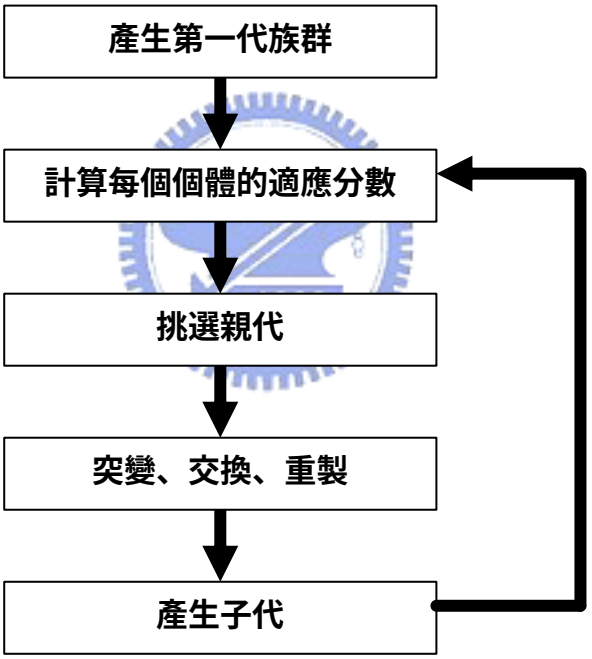

**圖表 6．相同結構，但莖幹、環線長度不同**

**下述為一個完整描述結構的範例。**

**莖幹個數 ： 2**
**莖幹相對位置：(1, 3)(2, 4)**
**莖幹與環線長度範圍：[3,6] (0,3) [4,10] (0,1) [3,6] (0,6) [4,10]**

### 4.2.2 GPRM 模型架構

GPRM 利用基因規畫的方法來尋找某核醣核酸家族的共同結構，主要概念是先隨機產生許多可能的結構，再透過突變、交換與複製的機制來改變這些結構。此外，再設計一套評分方式，比較每個結構的優劣。藉由分數直接反應出此結構便是解答的可能性，得分愈高者表示愈可能是此家族的共同結構。下圖即為 GPRM 流程圖。

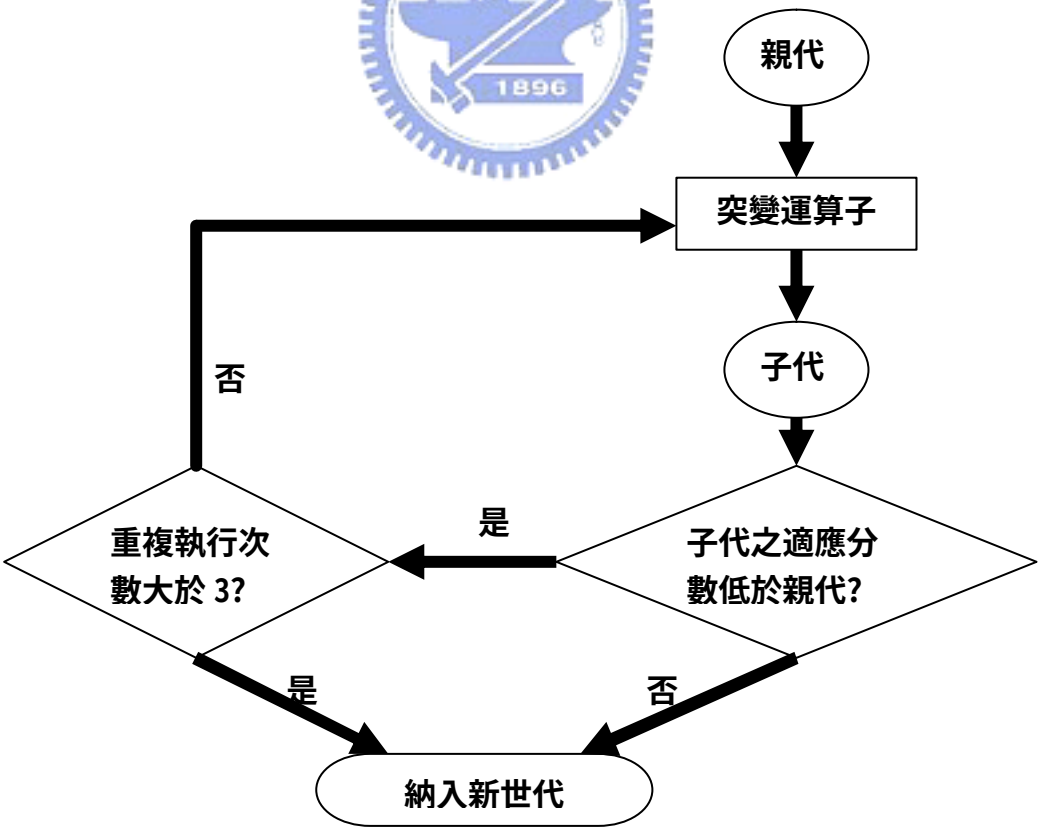



圖表 7．GPRM 系統流程圖

### 4.2.3 細部修改

本系統整合了 GPRM，但為了符合此研究的須求，我們在某些步驟做了修改，例如前置處理、適應函數與產生下一個子代等等，在之後的幾個章節會陸續提及。在此，我們先說明在產生子代時所做的變更。

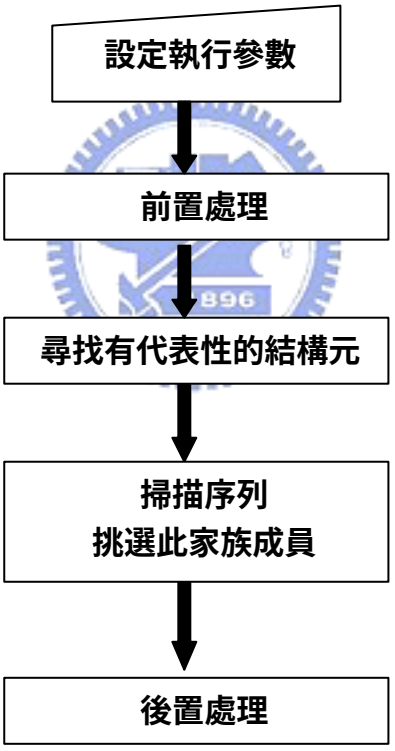
GPRM 使用競賽(tournament)的方式來挑選親代，勝出者會依照使用者所設定的突變率、交換率來決定執行突變或交換運算子，甚至是直接保留到下一代。執行突變者會產生一個新子代，交換運算子則有兩個子代產生。本系統維持與 GPRM 相同的突變與交換機制，但為了加快收斂的速度，我們規定唯有適應分數高於其親代的子代才會納入新的世代，否則便捨棄此子代，重新執行原本的運算子。然而，受限於程式執行時間，重覆執行運算子的次數限定在三次內，若超過三次，則一律將此子代納入新的世代中。圖表 8 為突變機制的流程。




圖表 8．突變運算子

## 4.3 系統主架構

下圖為本系統的主要流程圖，整體架構大約可分成五個部份。一開始使用者可以透過網頁介面設定程式執行參數、輸入核醣核酸序列。之後的前置處理部份，會根據這些設定值，利用特定的資料結構來表示這些序列資訊；分析完這些序列內容後，便可利用修改後的 GPRM 來尋找具有家族代表性的二級結構元；此共同結構元便可挑出此家族的成員。在後置處理時回報此組序列資料的分群結果，以及所預測的結構元，除此之外，視情況所須，簡單地說明分群結果。




圖表 9．系統架構圖

## 4.4 架構說明

在此章節中，則針對系統架構的每一部份做詳盡的說明。

### 4.4.1 設定程式執行參數

透過網頁介面，使用者可設定一些環境變數，除了一些原本 GPRM 系統預測共同結構所須的項目外，如突變率(mutation rate)、交換率(crossover rate)、族群大小(population size)等等，另新增一些在分群時須要的指標變數，如群集大小的最小限制(minimum cluster size)、拉普拉斯門檻值(Laplace threshold)等等。關於此筆核醣核酸序列的結構設定，保留莖幹與環線長度範圍，與 GPRM 不同的是，莖幹個數須指定最小值及最大值。這些參數的用處，會在之後的相關章節中詳細介紹。

### 4.4.2 前置處理

(1) 分析核醣核酸序列

本系統所能接受的核醣核酸序列格式與 GPRM 相同，皆為 FASTA 格式(FASTA format)，但在檢查輸入格式及合法字元的程序時再做補強，使得系統偵錯能力更加健全，並且給予更詳盡的錯誤訊息，以便使用者更快速地找出錯誤之處。在分析核醣核酸序列同時，會將四種鹼基及十六種相鄰鹼基對出現的頻率記錄下來，提供產生負面背景序列時所須的相關資訊。

(2) 產生負面背景序列

GPRM 利用基因規畫法尋找共同核醣核酸二級結構，在此研究中假設，具有生物意義的二級結構元不會任意出現在隨機產生的序列中, 因此須產生一組對照的負面背景資料(negative set)當做錯誤範例。為了行文方便，我

們稱使用者輸入的序列為正例，負面背景序列為反例，在之後的章節內容中這兩種用詞會交替使用。

依照 GPRM 的設計，所有的負面背景序列長度皆相同，雖然 A、C、G、U 四種鹼基出現的頻率與使用者輸入的序列相同，但相鄰的鹼基對之間是獨立的，互不影響。然而，已知自然界中的核甘酸序列，相鄰的鹼基對之間是有相關性的。因此，為了產生接近真實的生物序列，在準備負面背景序列時，則須考慮這一個特性。

本系統可產生不同長度的負例，每一條序列的長度則是根據正例而定，例如，第一條負例的長度會取正例中的第一條序列長度，餘此類推。若負例個數超過正例個數，一個循環後，再取正例的第一條序列長度。而每一條序列的第一個鹼基乃由四種鹼基個別出現的機率來決定，從第二個鹼基之後，便須考慮前一個鹼基的種類，由條件機率決定出現的鹼基。這樣的負例產生方式，我們稱之為一級(first order)序列產生方法。

(3) 取出所有合法莖幹

為了方便 GPRM 計算每個個體的分數，在前置處理的階段便先從使用者輸入的序列資料中取出所有合法的莖幹。而所謂合法的莖幹，便是長度符合使用者設定的範圍內的所有莖幹。這些莖幹將作為之後 GPRM 演化過程的材料。

### 4.4.3 尋找代表性的結構元

前置處理只是一般的準備程序，之後才會進行結構預測與分群的工作。此小節我們介紹如何利用 GPRM 來預測核酸核酸二級結構。

**適應函數的誤導**

GPRM 考慮正確率(precession)與擷取率(recall)來評估演化群體中二級結構的好壞，其定義為：

正確率(precession) ： $P = \frac{M}{M+N}$

擷取率(recall) ： $R = \frac{M}{C}$

其中 C 為使用者輸入的序列總數；M 表示在使用者輸入的序列中擁有此結構的序列總數；N 則是在負面背景資料中擁有此結構的序列總數。

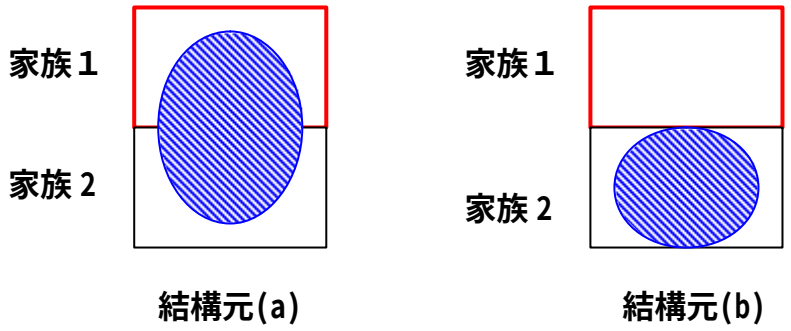
但為了比較的方便，透過下列公式的轉換將兩種數值結合起來，

$$f(M,N,C) = \begin{cases} \dfrac{1}{\dfrac{1}{2}\left[\dfrac{1}{\left(\frac{M}{C}\right)} + \dfrac{1}{\left(\frac{M}{M+N}\right)}\right]} & ,M \neq 0 \\ 0 & ,M = 0 \end{cases} \qquad (1)$$

$$= \begin{cases} \dfrac{2M}{C+M+N} & ,M \neq 0 \\ 0 & ,M = 0 \end{cases}$$

此式即是 F 分數(F-score)(Lewis and Gale, 1994)的定義，GPRM 便是以此當做適應函數。由上式來看，只有在正確率與擷取率都高的情況下，分數才會較高，故 GPRM 在尋找答案時會偏好愈多序列擁有的共同結構。但若考慮以下例子(圖表 10)，便會發現這樣的搜尋方向與實際答案背道而馳：

假設使用者輸入的序列分屬兩個不同的家族，藍色斜線區塊為擁有此二級結構的序列數目。




結構元(a)　　結構元(b)

圖表 10.

若以分群結果的角度來看，很明顯可看出，結構元(a)不能區分出不同家族的核醣核酸，反倒是有結構元(b)出現的核醣核酸比較單純，皆是家族 2 的成員。顯然，結構元(b)才是代表家族的結構元，但 GPRM 卻會偏愛結構元(a)。這是由於 GPRM 系統的前提假設是，所有的核醣核酸序列皆為同一家族，因此在計算擷取率時，錯把全部序列數目當作答案總數。由此例可知，真實的共同結構元出現次數只要達到某定程度即可，並非愈多愈好，最好是只在其家族的核醣核酸中出現。若要知道某結構的出現是否恰如其分，或只是偶然，則可利用統計方法來檢定。為了去除 F-分數的迷思，我們必須先預測家族成員個數，亦即目標結構元在正例中最佳的出現次數。

本研究運用拉普拉斯估計量(Laplace-estimate)(Kruskal and Tanur, 1978)來調整家族大小的猜測方向。以此作為臨界值來判斷結構元出現的次數是否達到可接受的程度。本研究依然使用 F-分數做為選擇結構的依據，只是 F-分數計算的方式須再加以修正。以下我們先介紹拉普拉斯估計量及其使用時機，再說明如何調整 F-分數計算公式。

**拉普拉斯估計量及其用處**

**在規則庫的建構領域中，拉普拉斯數值可用來估量符合此規則的範例個數是否夠多。它的原始定義如下：**

$$Laplace\_value = \frac{n_c + 1}{N + k} \qquad (2)$$

**其中 k 為此資料中共有多少種類(classes)；**

$n_c$ **則是群集 c 中符合此規則的範例個數；**

**N 是此資料中符合此規則的範例總數；**

**實際上，核醣核酸二級結構預測的問題亦可轉換成分類問題，分類規則便是演化得到的二級結構；所有的序列只分為正例(使用者所輸入的序列)與反例(隨機產生的序列)。利用結構元分類時，將擁有此結構的序列歸類為一類(稱為 A 類)，沒有出現者為另一類(稱為 B 類)。當 A 類中皆為正例，B 類中皆為反例時，此種分類結果最佳。在計算拉普拉斯值時，k 即為 2，因為處理之序列分為正反二例。而 N 便是所有序列中擁有此結構的總數。由於欲衡量的是『目標結構元在正例中出現的次數是否夠頻繁』，故 $n_c$ 便是在正例中擁有此結構的數目。於是，拉普拉斯公式便可以修改為**

$$Laplace\_value = \frac{n_p + 1}{n_p + n_n + 2} \qquad (3)$$

**其中 $n_p$ 為使用者所輸入的序列中擁有此結構的數目；**

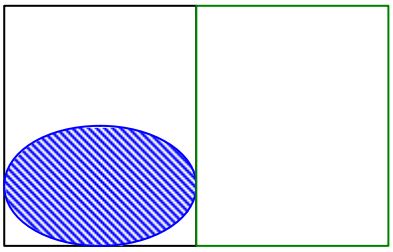
$n_n$ **為負面背景序列中擁有此結構的數目；**

在尋找具有代表性結構的過程中，GPRM 採用 F-分數來顯示二級結構的好壞。透過演化的運算子來改變二級結構以期提高適應分數，最後 GPRM 會預測出一個最好的二級結構，而其拉普拉斯值是否通過拉普拉斯門檻，則透露出不同的訊息。接下來我們仔細探討這兩種狀況。

【情況一】通過拉普拉斯門檻

GPRM 是一個採用監督式學習(supervised learning)來獲得最佳解的系統，因此會有一組負面背景序列來當做學習時的錯誤範例。根據假設所述，具有代表性的共同結構不會出現在這些序列中。換言之，我們希望結構元儘可能只在正例中出現，負面背景序列擁有此目標結構的數目愈少愈好。

對於相同家族的序列，利用 GPRM 可以尋找出它們共同的二級結構。這種結構在負面背景序列理當不常出現。如圖表 11 所示，此結構元並沒有出現在任何一條負例中，且在正例中，擁有此結構的數目接近一半。會造成此結果的原因極可能是因 GPRM 所找到的共通結構太嚴格，因此，雖然沒有任何負例包含此結構，但也僅出現在少數正例中。換言之，這可能是(overfitting)的結果，但真正的家族成員不只這些核醣核酸，因而，必須調高家族的大小(family size)。



圖表 11.

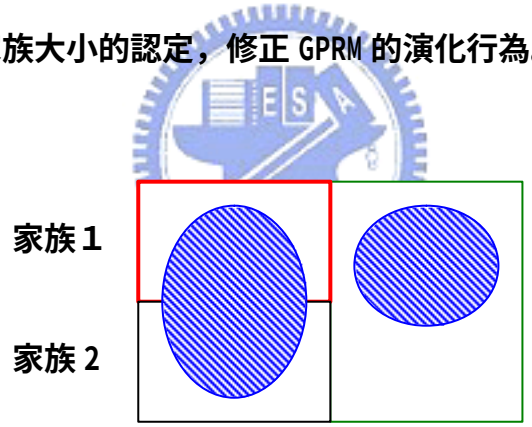
左邊區塊為正例個數；右邊區塊為反例總數。

負面背景序列數目與使用者輸入的序列數目相等。

【情況二】不及拉普拉斯門檻值

當有一組核醣核酸序列，其屬於不同的家族，例如，兩個大小相同的家族，則不管是哪一個家族，其真正的結構元平均出現的次數應該都不會超過一半，然而因 GPRM 優先選擇大量出現的結構，故演化終止時，可能會收斂到一個不具代表性的結構，不僅這兩個家族的核醣核酸大多會擁有此種結構，甚至在負面背景序列中亦會常出現。其拉普拉斯數值應當較低。若能透過適當的門檻刪選，應可以過濾這些不具代表性的共通結構。

如圖表 12 所示，此二級結構元的拉普拉斯值不及門檻值，表示 GPRM 所預測的二級結構為無意義的共同結構，其出現的核醣核酸數目太多，不但包含了另一個家族的核醣核酸，也包含了許多序列負例，因此，我們能依此校正 GPRM 對於家族大小的認定，修正 GPRM 的演化行為。




圖表 12.

左上紅色區塊為第一群正例；左下黑色區塊為第二群正例；

右邊綠色區塊代表反例總數。

在此範例中，負面背景序列數目與使用者輸入的序列數目相等。

**預估某家族的序列數目**

本研究欲處理的資料是一組未經排比(unaligned)的核醣核酸序列。我們希望利用共通結構元為基礎，對這些核醣核酸序列作分群。為了計算正確的 F-分數，必須事先知曉每個家族的核醣核酸數目。我們採取二分逼近法來預測其中某家族可能的成員個數，希望由上下夾擠的方式找出最適當的家族大小。

一開始先假設全部的核醣核酸為同一家族，序列總數即為答案總數。此時利用 GPRM 預測共同的二級結構，若此結構的拉普拉斯值通過門檻，雖符合上述情況一的條件(請參照**拉普拉斯估計量及其用處**)，但因為家族大小已無法再擴大，則表示這群核醣核酸皆屬相同家族，且這答案便是正確的結構元；反之，這些核醣核酸便分屬兩個以上的家族。因此，須要先估計出其中某一家族的核醣核酸數目，有了較正確的答案總數，GPRM 才能找出正確的二級結構元。

當結構元的拉普拉斯值不及門檻值時，則為上述情況二的條件(請參照**拉普拉斯估計量及其用處**)，故須將猜測數值縮小。本研究的做法是將家族大小減為一半，再利用 GPRM 重新尋找共同的二級結構。當決定增加家族成員數目時，便取前一次拉普拉斯值未通過門檻時的大小，與此次猜測的家族大小的中間值；反之，須降低家族個數時，便取前一次通過門檻時的大小，與此次猜測的家族大小的中間值。重複上述的程序，直到找到最適當的大小。而所謂的”最適當的大小”，本研究定義為，當序列數少 1 則可通過拉普拉斯門檻，多 1 則拉普拉斯值又嫌不足。

我們希望在預測序列數目的過程中，會如下述的劇本發展：

以下例子中，N 表示猜測的成員個數。此例中核醣核酸總數為 50。


圖表 13．例一。

紅色數字表示拉普拉斯值不足；

藍色數字則是超過拉普拉斯門檻。

1．N=50，拉普拉斯值不足，故 N 降為 25，重新尋找共同結構。

2．N=25，超過拉普拉斯門檻，取 25 與 50 之中間值 37，重新尋找共同結構。

3．N=37，拉普拉斯值不足，故取 25 與 37 之中間值 31，重新尋找共同結構。

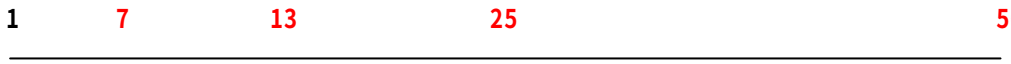
4．N=31，拉普拉斯值不足，故取 25 與 31 之中間值 28，重新尋找共同結構。

5．N=28，超過拉普拉斯門檻，取 28 與 31 之中間值 29，重新尋找共同結構。

6．N=29，超過拉普拉斯門檻，取 29 與 31 之中間值 30，重新尋找共同結構。

與 30 最接近且拉普拉斯值不及門檻值的成員數目為 31, 因之間的差距為 1, 且成員數目為 29 的拉普拉斯值有超過門檻值，符合我們所定義的終止條件，故此時可宣稱 30 為某家族的成員數目。

由於本研究認為成員數目須達到一定程度才足以構成一個家族, 但多大才算有意義的核醣核酸家族，我們預留設定彈性給有不同認定的使用者，故使用者須先指定最小的家族大小(minimum cluster size)。圖表 14 所顯示的例子是，當拉普拉斯值一直無法通過門檻值時，成員數目會一直減半，直到估計的數量小於最小的家族大小。




圖表 14．例二。

假設一個家族至少須要 10 條序列。

1．N=50，拉普拉斯值不足，故 N 降為 25，重新尋找共同結構。

2．N=25，拉普拉斯值不足，故 N 再降一半，變成 13，重新尋找共同結構。

3．N=13，拉普拉斯值不足，故 N 再降一半，變成 7。

因為最小的家族大小為 10，此時猜測的成員數目已小於最小限制，達到終止條件。對於這種情況，因為找不到一個適當的家族成員個數，於是便猜測是結構參數設定的問題。

修改結構參數

2002 年本實驗室發展的 GPRM 系統，提供使用者輸入結構參數的部份，其中之一便是莖幹個數。而在本研究中，此部份修改為輸入最少的莖幹個數與最多的莖幹個數。正是因為使用者可能輸入不同家族的序列，這些家族的共同結構可能莖幹個數不同。

由於 GPRM 使用 ramped half-and-half(Koza 1992)的概念產生第一代的族群，若允許個體可以擁有不同的莖幹個數，則須要很大的族群才能囊括所有可能的二級結構相對位置，而且須要夠長的演化時間才能收斂到正確的答案。因此，在預測可能的共同結構時，整個族群中的個體會固定相同的莖幹個數。

本研究先以最多的莖幹數目開始嘗試，在此結構參數的設定下，預測可能的家族成員個數。若未發現任何合適的家族大小，則將莖幹個數減 1，再重新推測其家族成員數目。倘若目前的結構設定已是最小的莖幹數目，我們則認為此組核醣核酸為同一家族。

**適應函數修改**

當猜測新的家族成員個數時，GPRM 會再重新尋找另一個共同的二級結構。本研究所使用的 GPRM 仍舊沿用 F-分數充當適應分數，由於家族成員總數預設不同，故計算它的適應分數時，擷取率的計算應作適當修正。當在正例中，出現此結構的核醣核酸數目超過所猜測的家族成員個數，便將擷取率(recall)設為 1。因此，適應函數便修改如下：

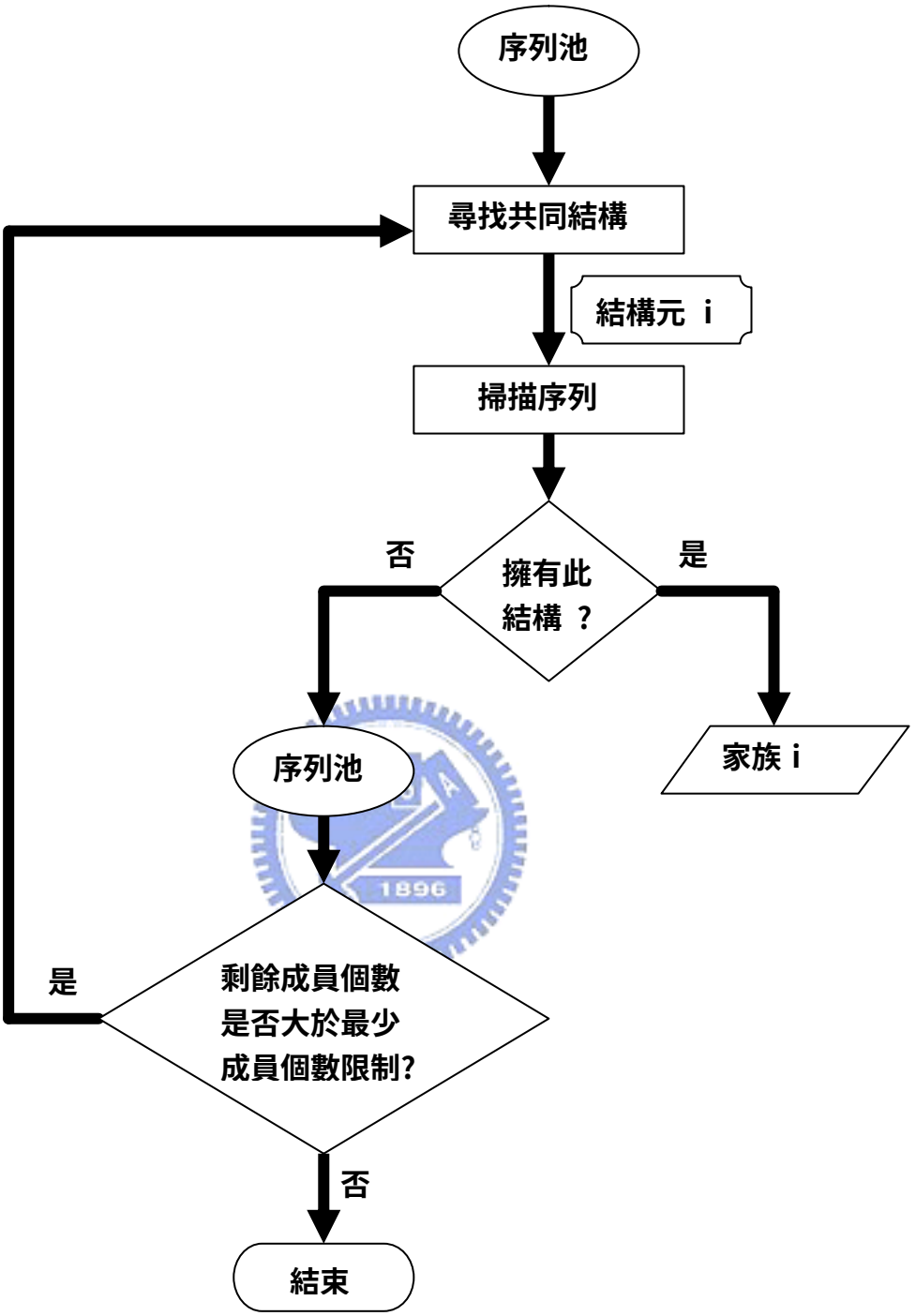
$$f(M,N,C)=\begin{cases}\dfrac{1}{\dfrac{1}{2}\left[\dfrac{1}{\left(\dfrac{M}{C}\right)}+\dfrac{1}{\left(\dfrac{M}{M+N}\right)}\right]} & ,M<C,M\neq 0\\[2ex] \dfrac{1}{\dfrac{1}{2}\left[1+\dfrac{1}{\left(\dfrac{M}{M+N}\right)}\right]} & ,M\geq C,M\neq 0\\[2ex] 0 & ,M=0\end{cases} \qquad (4)$$

### 4.4.4 挑選家族成員

在整個分群流程中，最後步驟是將擁有此共同結構的序列分離出來。本研究分類的準則是，將擁有相同二級結構的核醣核酸歸為同一個家族，這是建立在『同一家族的核醣核酸，其二級結構會非常相似』的假設上。因此，須要設計一個有系統、有依據的工作流程，來預測這個代表一整個家族的共同結構元(詳細流程，可參照 4.4.3 尋找代表性的結構元)。

下圖為分群的流程圖，經過一套有系統的預測程序，先預測出家族大小，之後利用 GPRM 獲得一個良好的結構答案，而這就是某一家族的共同結構元。將這個二級結構對所有序列進行掃描的動作，若發現某序列擁有此結構，則將它從其它序列分離出來。最後我們可以收集一群具有此結構的核醣核酸，這些核醣核酸就形成一個家族。

一個家族若只有兩、三個成員，直覺上，這是一個偶然形成的群體。於是，我們設定一個數值來限定每一個家族最基本的成員數目。每當分出一個核醣核酸家族後，我們便要檢查，剩餘的核醣核酸數目是否足以構成一個家族。條件滿足時，才須要繼續尋找另一個二級結構，否則，便可視為分群工作完成，而這些剩餘的核醣核酸則視為雜訊(outlier)，不屬於任何一個核醣核酸家族。


序列池
尋找共同結構
結構元 i
掃描序列
否
擁有此
結構 ?
是
家族 i
序列池
是
剩餘成員個數
是否大於最少
成員個數限制?
否
結束

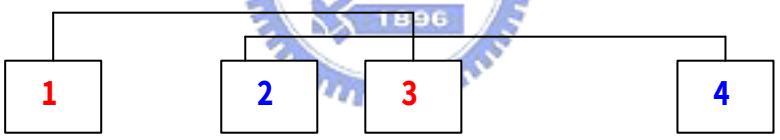
**圖表 15．系統流程圖**

### 4.4.5 後置處理

**減少錯誤的正預測**

本研究承襲了 GPRM 所提出的二級結構表示法，除了描述莖幹的相對位置外，還有每段莖幹與環線的長度範圍(請參照 4.2.1 核醣核酸結構描述語言)。根據 GPRM 的基本假設，共同結構元在每一條核醣核酸中，只會出現一次。除了正確答案外，其它符合結構限制條件的位置，則稱為錯誤的正預測(false positive)。

雖然核醣核酸的共同二級結構長相會有些微的差異，即某莖幹或某環線的長度不相同，但大致上，長度的變化量應該是很小的。在本篇論文中，將莖幹與環線的變化量總合定義為『變易度』(flexibility)。例如，以下為 GPRM 所預測出來的 soil-borne mosaic virus 家族的共同二級結構。




[3,6] (0,3) [4,10] (0,1) [3,6] (0,6) [4,10]

圖表 16.

其中第一個莖幹的長度最小是 3，最大是 6；而第一個環線的長度範圍是 0 到 3。則此二級結構的『變易度』即為

(6-3)+(3-0)+(10-4)+(1-0)+(6-3)+(6-0)+(10-4) = 28。

一個二級結構若堪稱是一個家族的識別標記，則在每一條核醣核酸上的長相應該非常一致。故它的『變易度』應該非常地小。我們希望再進一步調整二級結構元，使得它的『變易度』是所有可能的結構元中最小的。在調整

過程中須滿足的條件是，調整後的共同結構元，依然必須出現在每一個家族成員中。由於莖幹與環線長度範圍限制更嚴格，便可濾掉一些錯誤的結果。

本研究利用 Branch and Bound 演算法(Narendra and Fukunaga 1977)來搜尋『變易度』最小的二級結構元。這是一種深度優先(depth-first)的搜尋方式，逐一地整合每條核醣核酸的一個答案，直到全部家族成員都附加進來。此時便產生一個可能的二級結構元，比較此結構的『變易度』，若是小於目前為止最小變易度的結構元，則更新結構元與最小變易度的結構記錄保持者。遇到相等的情況，便取莖幹總長度最大者。例如

結構記錄保持者：[3,5](2,3)[3,5]，變易度 = 5，莖幹總長 = 10

新的二級結構元：[4,6](1,2)[4,6]，變易度 = 5，莖幹總長 = 12

以這兩個結構而言，最後會更新結構記錄。這是由於莖幹愈長者，結構愈穩定，也更有可能是家族結構代表。

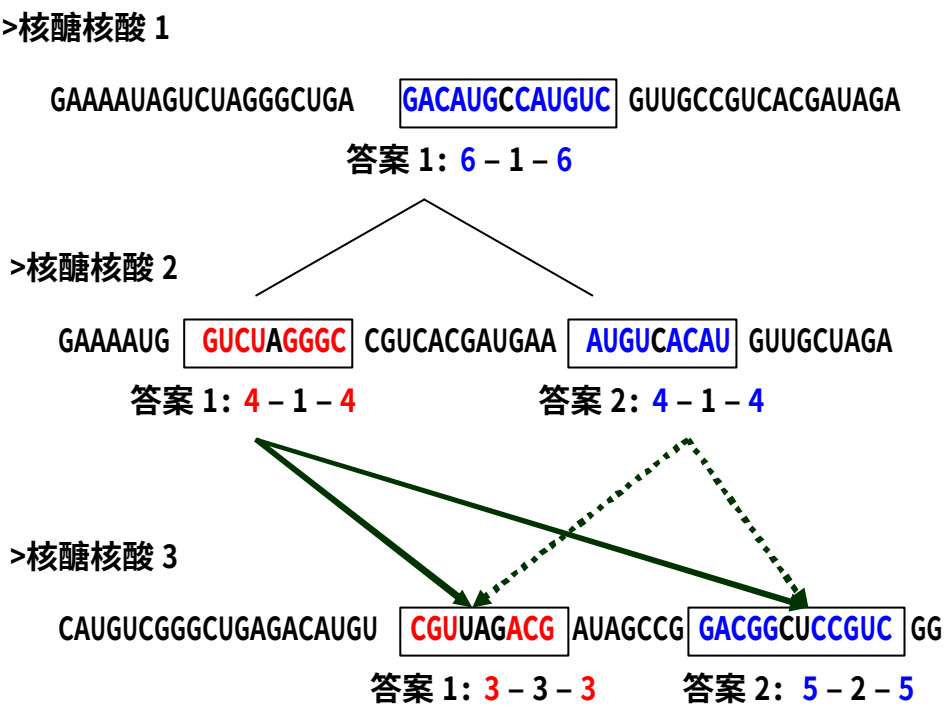
為了減短搜尋時間，以下任一種情況發生時，便可省去之後的探索動作，嘗試新的搜尋方向：

(1)目前的二級結構元變易度比結構記錄中的來得大。

每整合一條核醣核酸的答案後，便要重新計算此時的二級結構變易度。若大於結構記錄中的變易度，因為不論再怎麼結合其它成員的答案，都不可能找到一個變易度更小的結構，故可放棄這個探查方向。

(2)目前的二級結構元會引導搜尋動作至重覆的探查路線。

以下我們舉例說明此種情況。為了行文方便，估且將核醣核酸編號為 1，2，3。在每條核醣核酸中，亦標示出所有符合結構限制條件的答案。




圖表 17. Branch and Bound 搜尋路線

以此例子來說，我們尋找目標結構元的第一步驟會整合核醣核酸 1 的答案 1 與核醣核酸 2 的答案 1，之後可得到一個二級結構元 [4,6](1,1)[4,6]。再往下一層的搜尋路徑是，整合此結構元與核醣核酸 3 的答案 1，最後得到結構元[3,6](1,3)[3,6]，其變易度為 8。另一個的搜尋路徑是整合核醣核酸 3 的答案 2，最後的結構元長相為[4,6](1,2)[4,6]，變易度是 5。可發現變易度最小的二級結構元是[4,6](1,2)[4,6]，變易度最小記錄為 5。而往上推一層，核醣核酸 1 會整合核醣核酸 2 的答案 2，得到的二級結構元依然是[4,6](1,1)[4,6]。若再繼續往下一層搜尋，其路徑是整合核醣核酸 3 的答案 1，與整合核醣核酸 3 的答案 2(圖中綠色虛線部份)，這會得到完全相同的結構元長相。也就是說，這是重覆的探究的路線(所得結果與圖中綠色實線部份相同)，故可省去圖中綠色虛線的比對動作，即使目前結構元的變

易度尚未超過記錄中的變易度。

總而言之，由 GPRM 所預測出來的共同結構，先記錄它在此家族的所有核醣核酸上出現的所有位置。根據下述的五個步驟來尋找『變易度』最小的二級結構元。

步驟一：取出此核醣核酸中一個答案，若答案已全部檢視完畢，則回到前一條核醣核酸。

步驟二：與目前的二級結構元結合。

步驟三：計算新的二級結構元的『變易度』，若比記錄中最小的『變易度』來得大，則回到步驟一。

步驟四：溯及以往整合至本條核醣核酸中，曾出現過的二級結構元，若重覆，則回到步驟一。

步驟五：若此家族中，尚有核醣核酸未被檢查，則任取其中一條，再回到步驟一。反之，則表示此二級結構元在所有核醣核酸中皆有出現。計算它的『變易度』，若它是目前最小的『變易度』，則更新記錄中的二級結構元及『變易度』。否則便捨棄它。若變易度相等，則取莖幹總長度最大者。

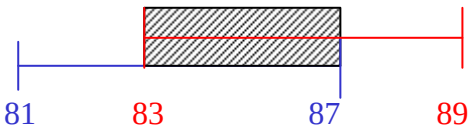
**過濾相似的答案**

本研究輸出分群結果時，每一個家族的序列，除了顯示序列名稱外，還會標示出結構元出現的位置。而在相同序列上，可能會出現好幾個答案。若要將這些結果逐一顯示，則會造成使用者的負擔，而且重要的信息往往會被一堆無用的資訊給掩蓋住。因此須要過濾一些太相似的答案。

本研究重新定義兩個答案的相似程度，根據使用者設定的門檻值(Basepairing overlap allowance rate)來決定兩個答案是否相似。本研究中相似度的定義是, 將兩個莖幹結合後, 重疊部份的長度除以整體的總長度。

$$\text{相似度} = \frac{\text{重疊區域長}}{\text{重疊後整體總長}} \qquad (5)$$

舉例來說，若有兩個莖幹分別出現在(81-87)與(83-89)的位置上，則兩者結合後如下所示，灰色斜線區域便是重疊的部份。

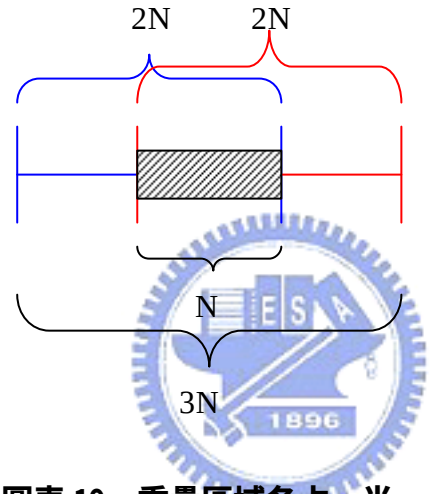



**圖表 18．重疊後的兩個莖幹**

根據定義，其相似度為 0.55。

$$\text{相似度} = \frac{(87 - 83 + 1)}{(89 - 81 + 1)} = 0.55$$

下圖的例子是，兩個長度為 2N 的莖幹，當有一半的部份重疊在一起時，重疊區域的長度為 N，結合後的總長度是 3N，因此相似度為 $\frac{N}{3N} = 0.33$。




圖表 19．重疊區域各占一半

若相似度通過門檻值，則認為這是兩個相同的莖幹。對兩個不同的結構答案而言，當所有的莖幹都被視為相同時，才代表這兩個是完全相同的答案。最後，會保留莖幹總長度較長的答案。

**分群結果解釋說明**

**當最終分群結果將所有序列視為相同的家族時，本系統會根據結果發生原因提供三種可能的建議：**

**【情況一】**

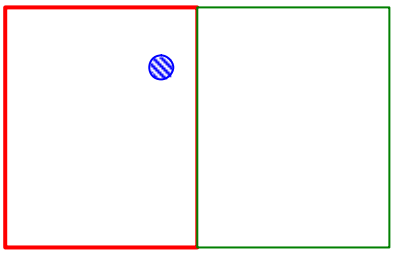
**在預測家族大小時，一開始會將所有的序列視為同一家族。若所得到的二級結構元，其拉普拉斯值通過門檻值，因個數無法再增加，故認為全部的核醣核酸屬於相同的家族。或許使用者輸入的便是同源的核醣核酸序列，但亦可試著提高拉普拉斯門檻值，再重新分群以驗證此想法。**

**【情況二】**

**另一種可能是結構參數設定太過寬鬆，例如合法的莖幹、環線長度範圍太大。因此預測出一個無生物意義的二級結構元。由於此結構出現在所有的核醣核酸中，而被錯認為家族代表結構。因此，可將莖幹或環線合法長度範圍減小後，再重新進行分析實驗。**

**【情況三】**

**預測家族成員數目時，若嘗試所有可能的莖幹個數，依然無法找到合適的大小，最後會將所有的核醣核酸視為相同的家族。這可能是此組資料中的核醣核酸之間，其二級結構長相並不是非常一致，擁有非常相似結構的核醣核酸數量不夠多。如同圖表 20 所顯示的一個成員個數非常少的群集。**



**圖表 20.**

**左邊紅色區塊為正例；右邊綠色區塊代表反例總數。**

**負面背景序列數目與使用者輸入的序列數目相等。**

**由於一開始設定的門檻值太高，當無法通過門檻值時，這一群核醣核酸的共同結構元會被視為無意義的結構。如上圖所示，此結構元並沒有出現在任何一條負例中，但在正例中，擁有此結構的數目過低，故認為這個結構的出現只是一個偶然的情況。可試著降低拉普拉斯門檻值，放鬆群集大小的限制，再重新分群。**

# 第五章　實驗結果

為了測試此演算法的正確性，我們需要實作一些實驗來驗證我們的想法，以下便介紹實驗主題、測試資料的準備以及實驗的結果。

## 5.1 實驗項目

在此篇論文研究中，我們想要探討的三種主題分別為：

(1) 非結構區域序列的相似度：若不同家族的核醣核酸在非結構元的區域擁有相似的序列內容，是否會預測出一個無意義的共同結構，而無法分離出這些不同家族的核醣核酸。

(2) 同源核醣核酸二級結構長相的一致性：即使屬於相同的家族，二級結構的長相仍有所差異。若不同核醣核酸的莖幹長度差距太大，則在隨機產生的序列中，亦有可能出現此種結構。那麼此家族的核醣核酸是否依然能與其它家族的成員區隔開來。

(3) 家族成員個數的比例：若是家族成員個數不同，差距懸殊，是否成員稀少的家族容易被忽略。

## 5.2 測試資料

**真實序列資料**

目前我們可以收集到三組已經確定共同結構的核醣核酸家族，包括 IRE like、soil-borne mosaic virus 與 archaea 16S rRNA。本篇論文研究的實驗資料便由這組所組合而成的。以下為這三個家族的基本資料整理。

| | 莖幹個數 | 家族成員個數 | 序列平均長度 |
|---|---|---|---|
| IRE like | 2 | 56 | 206 |
| soil-borne mosaic virus | 2 | 18 | 62 |
| archaea 16S rRNA | 3 | 34 | 100 |

**表格 1 三個核醣核酸家族基本資料**

**人造測試資料**

由於，目前真實的核醣核酸序列資料量較少，甚至可能無法能夠完全切合本研究所提出的實驗主題，因此，須要自行準備實驗的材料。產生測試資料的方式大約可以分成三個步驟：

**步驟一 設定參數**

在產生資料時，為了符合不同實驗項目的需求，故利用一些參數的設定，來指導資料產生器輸出何種類型的資料。

(1) 家族個數：此測試資料中總共含有幾個不同的家族。

(2) 非結構區序列相似度：決定不同家族間，非結構區序列相似的程度。

(3) 家族成員個數：個別設定所有家族的成員個數，此變數可控制不同家族間成員個數的差距。

(4) 四種鹼基出現的機率：個別設定所有家族中，鹼基出現的分佈。

**步驟二 設定共同結構元**

為了達到高彈性的目的，每個家族的共同結構元依然由使用者自行設定，即須指定莖幹個數、配對結構以及莖幹與環線的長度。透過這些參數，可以控制不同家族間共同結構元的相似度，或是同一家族內，每一條核醣核酸二級結構的相似程度。

**步驟三 產生核醣核酸序列**

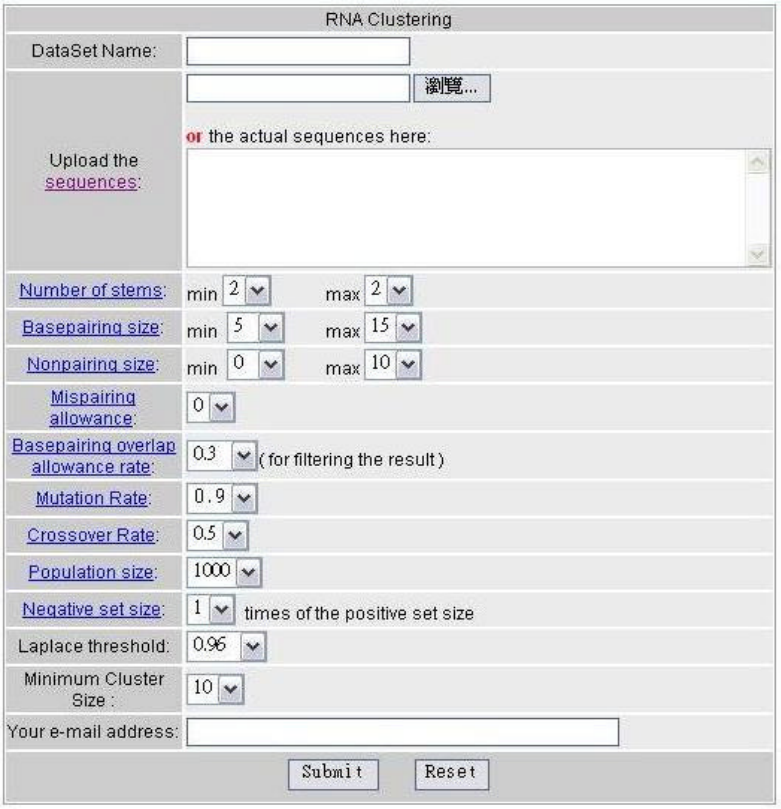
核醣核酸序列大致可切割成兩種片段，答案(motif region)與非答案區域(non-motif region)。這兩種不同片段會有不同的產生方式，當答案與非答案區域的內容分別決定後，再由非答案區中隨機選一個位置將答案插入。舉例來說，一條序列產生的步驟如下：

假設二級結構為 [3,5] (4,6) [3,5]

1. 決定莖幹內容的長度。例如 4-5-4，則表示兩邊莖幹長度為 4，中間的環線長度為 5。
2. 決定雙股區內容。左邊的莖幹，是由四種鹼基出現的機率來決定每一個位置上的內容。由於本研究允許的配對方式為 A-U、G-C 與 G-U，因此，左邊的莖幹出現 A 時，右邊的莖幹就必須是 U；若左邊出現 G，則右邊可出現 C 或 U，兩者機率相同。
3. 決定環線內容。此片段的序列亦是根據四種鹼基出現的機率來決定每一個位置上的內容。
4. 產生非答案區域之序列。假設四種鹼基呈現獨立分佈，則根據鹼基出現的機率來決定每一個位置上的內容。
5. 隨機挑選非答案區中的一個位置，將答案序列插入。
6. 重覆上述 1-5 的步驟，最後將此家族的序列內容全部產生。

## 5.3 執行介面與參數設定

下圖為本研究所提供的網頁介面，方便使用者進行相關的分析實驗。




圖表 21. 使用介面

在此介面中，有幾類的參數需要事先設定：

(1)分析資料

資料名稱(Data SetName) – 用來分辨不同的實驗。

核醣核酸序列(sequence) – 使用者欲分析的資料。可直接上傳檔案或輸入所有序列內容。

(2)演化環境

突變率(Mutation Rate) – 演化過程中執行突變運算子機率。

交換率(Crossover Rate) – 演化過程中執行交換運算子機率。

群體大小(population size) – 演化時群體的大小。

背景序列數目(Negative Size) – 計算分數時，反例數目。總數為正例的整數倍。

(3)結構參數

莖幹個數範圍(Number of stems) – 最小與最大的莖幹數目。

莖幹長度範圍(Basepairing size) – 雙股區域中鹼基對數範圍。

環線長度範圍(Nonpairing size) – 單股區域中鹼基個數範圍。

錯誤配對容忍度(Mispairing allowance) – 雙股區域中，允許幾對錯誤的配對，但這些錯誤配對不能出現在莖幹兩端。

答案相似度(Basepairing overlap allowance rate) – 分辨兩個結構是否相似，用以刪除重覆出現的答案。

(4)分群參數

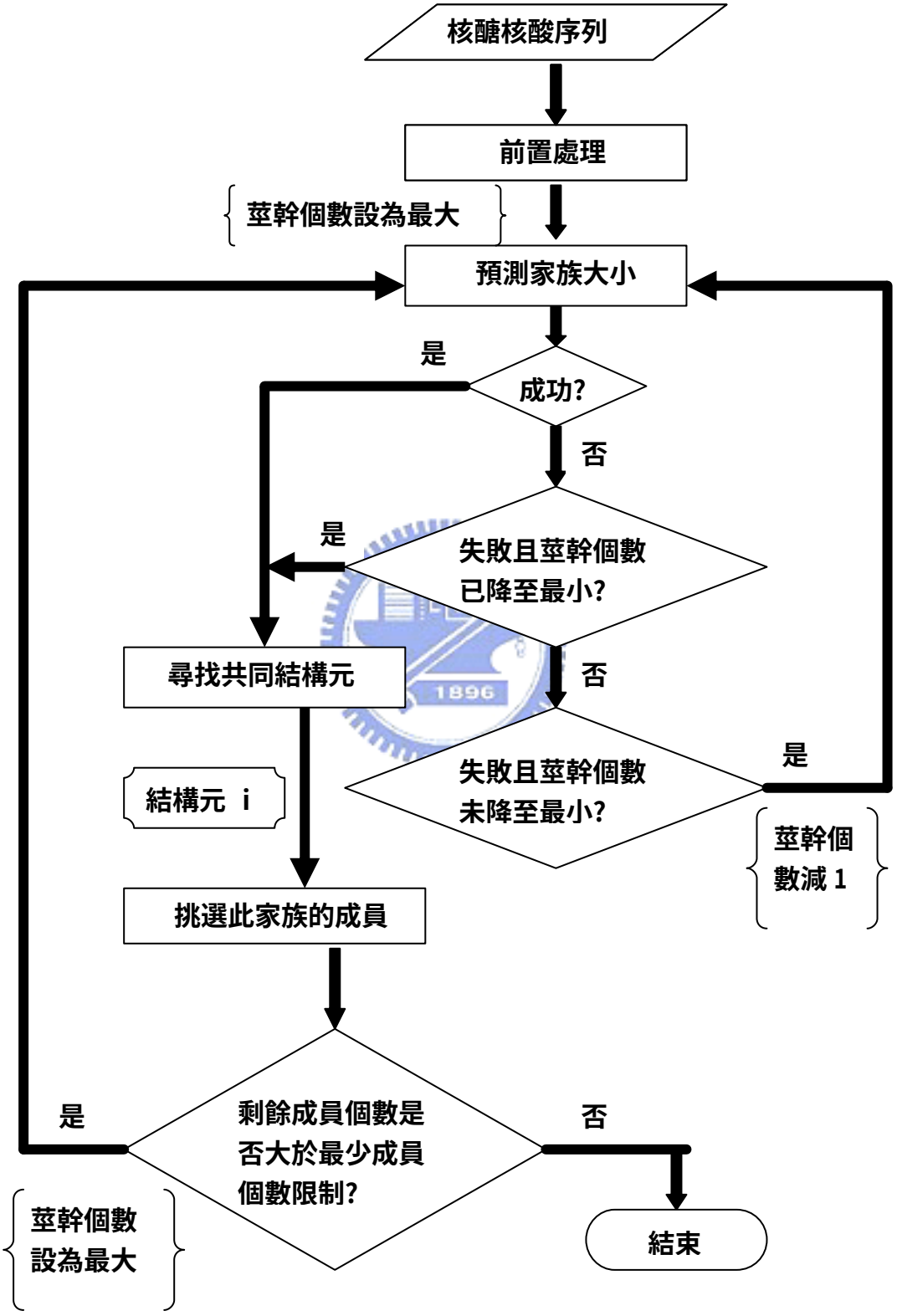
拉普拉斯門檻值(Laplace threshold) – 調整家族大小的猜測值。

基本成員個數(Minimum Cluster Size) – 每個家族最少須擁有多少成員。

(5)使用者聯絡方式

電子郵件信箱(e-mail address) – 用以回報分群結果。

## 5.4 實驗流程




圖表 22. 實驗流程圖

上圖為本研究之實驗流程圖，模型設計概念是，先預測某家族的大小，有了正確的成員個數，GPRM 才能尋找出代表此家族的共同結構元。以此結構去挑出屬於此家族的核醣核酸，移除這些序列後，再尋找另外家族的結構元，以避免不相關序列誤導了 GPRM 搜尋答案的窘況。

在前置處理中進行鹼基字元檢查、計算四種鹼基與十六種連續鹼基出現的頻率，以及列出所有合法的莖幹，這一連串準備的作業結束後，才進入核心的結構預測與分群程序。

本研究修改 GPRM 後，將它應用在本系統中尋找代表家族的共同結構。本實驗室發展的 GPRM，是利用基因規畫尋找共同的二級結構。由於受限於族群的大小，故 GPRM 在尋找共同結構元時，演化的個體會固定相同的莖幹個數。本研究先以最多的莖幹數目開始嘗試。

在利用 GPRM 尋找代表性的二級結構元前，需要先預測家族的大小。使用二分逼近法尋找成員個數時，會有以下三種可能的情況：

(1) 成功：正確地預測出某家族的大小。

(2) 失敗且莖幹個數已降至最小值：所有莖幹個數都猜測過後，依然無法找出合適的家族大小，表示此組核醣核酸的二級結構差異性頗大。若只將非常一致的核醣核酸分為同一群，成員個數又無法滿足最小的限制，因此，勉強視此組序列為相同家族。

(3) 失敗且莖幹個數大於最小值：以此結構設定下，無法找出合適的家族大小，這意味著目標結構並不存在這麼多的莖幹個數，故將莖幹個數減 1 後，再重新預測可能的家族個數。

有了正確的家族大小與莖幹個數，才能使用 GPRM 來預測共同的二級結構元。尋找出家族結構代表後，接下來的步驟便要掃描序列，將擁有此結構的核醣核酸分離出來，自成一個家族。剩餘的核醣核酸若足以形成另一個家族，便再重覆整個預測與分群的程序。由於無法確知另一家族共同結構的莖幹個數，故再從最大的數目開始搜尋。

## 5.5 評估方式

**分群結果評估**

本篇論文研究主要是提出一套核醣核酸分群的系統，在評估分群結果時，選擇常用的擷取率(recall)與正確率(precision) 來衡量挑選到正確家族成員的能力，其定義分別如下所示：

$$\text{正確率} = \frac{\text{真正屬於此群集的成員數目}}{\text{此群集目前的成員總數}} \qquad (6)$$

$$\text{擷取率} = \frac{\text{真正屬於此群集的成員數目}}{\text{此群集真正的成員總數}} \qquad (7)$$

**整體結果(分群與結構元預測)評估**

本研究分群的依據主要是核醣核酸共同的二級結構。唯有預測出正確的結構元，分群結果才會是可靠的。因此在評估系統的優劣時，亦須要考慮二級結構預測的結果。本篇論文沿襲了 GPRM 評估核醣核酸二級結構預測的方式，依舊採用 Matthews 的相關係數評分法(Matthews correlation coefficient)(Matthews 1975)。

經過推導後，相關係數的計算公式為

$$C \approx \sqrt{\frac{P_t}{P_t + N_f}\frac{P_t}{P_t + P_f}} \qquad (8)$$

其中 $P_t$：正確的正預測　$P_f$：錯誤的正預測　$N_f$：錯誤的反預測

在本研究議題中，正確的正預測(true positives)是指，確實屬於此家族的核醣核酸，其預測的結構元中，所包含的鹼基對亦出現在正確結構元中的鹼基

對個數；錯誤的正預測(false positives)則是，不屬於此家族的核醣核酸鹼基對個數，以及雖被歸類在正確家族的核醣核酸，在其預測的結構元中，所包含的鹼基對並沒有出現在正確結構元中的鹼基對個數；另外，屬於此家族之核醣核酸，其正確結構元中的鹼基對，並沒有出現在預測的結構元中，則稱為錯誤的反預測(false negatives)。

## 5.6 實驗結果

本研究以 C 語言來撰寫整個系統，執行環境為 Mandrake Linux 9.0 作業系統，兩顆 PIV 2.4G Hz 的中央處理器以及 1G 記憶體。在本篇論文所進行之實驗，其群體大小設為 1000，突變與交換率分別為 0.9 與 0.2，而反例數目為正例的 1 倍。拉普拉斯門檻值設為 0.96。

### 5.6.1 非結構區域序列相似度影響

在生命體中無生化功能的核醣核酸二級結構元，在不同的家族中都有出現的可能。為了探討本系統是否會被這些結構的誤導，而預測出錯誤的共同結構元，我們設計了此類實驗。由於在非結構區域中，若有非常相似的序列內容，且能摺疊成二級結構，則會有較高的機會在此區域中出現共同的二級結構元，而此結構便是無代表性的結構，但有可能誤導系統運作。我們希望藉此類資料測試本系統的容錯能力。

此組實驗資料為 IRE like 與 soil-borne mosaic virus 混合而成的。在這次實驗中，我們故意放鬆結構參數的設定，莖幹個數設為 2-3。表格 2 為其實驗結果整理。

| | 擷取率 | 正確率 | Matthews 相關係數 |
|---|---|---|---|
| IRE | 0.83 | 0.98 | 0.78 |
| VIRUS | 0.77 | 0.76 | 0.62 |

表格 2 IRE + VIRUS (2-3 個莖幹)

由表可看出，IRE 的正確率(Precision)很高，但擷取率(Recall)卻很低。這是因為 GPRM 預測出來的共同結構擁有三個莖幹，如下所示。

[4,6] (2,2) [5,5] (7,7) [5,5] (1,1) [4,6] (0,6) [3,4] (3,9) [3,4]

實際上，IRE 共通結構元只有兩個莖幹，GPRM 雖能正確地預測出來，但也找出另一個無意義的莖幹(上例中的綠色莖幹)。這個莖幹的長度很短，而中間所夾擠的環線範圍很寬鬆，這對 IRE 這種序列很長的家族來說，在非結構區域有很高的機率可以找到這種莖幹。但以這種結構元去掃描家族成員時，只挑出其中一部份的成員，故在 IRE 這群中有很高的正確率，但擷取率卻很低。而因為遺漏掉太多的序列，使得錯誤的反預測(false negative)太高，而拉低了 Matthews 相關係數值。

雖然 GPRM 預測出 VIRUS 家族的結構元為擬節結構，但受限於此組資料天生上共同的結構元長相差異性比較大的關係，擷取率依然偏低。由於 IRE 序列比較長，有些序列在非結構區域出現擬節結構，因此，即使以 VIRUS 的結構元來掃描家族成員，還是會挑出 IRE 的序列，而拉低了 VIRUS 的正確率。因為在搜尋共同結構元時，還是會被 IRE 序列的影響，所有預測出來的結構元多多少少還是會與真實的長相有所出入，再加上太多不屬於此家族的序列，而有很高的錯誤正預測(false positive)，故 VIRUS 的 Matthews 相關係數值便降低了。

若我們嚴格的限制共同結構元的莖幹個數為 2，則不僅是擷取率與正確率，連 Matthews 相關係數也可大幅改進，實驗結果如下表所示。

| | 擷取率 | 正確率 | Matthews 相關係數 |
|---|---|---|---|
| IRE | 0.97 | 0.99 | 0.97 |
| VIRUS | 0.71 | 0.95 | 0.79 |

表格 3 IRE + VIRUS (2-2 個莖幹)

### 5.6.2 同源核醣核酸其二級結構相似度影響

我們以共同結構元的變易度來表現同源的核醣核酸，其二級結構長相的一致性。當變易度愈小時，每一條核醣核酸的二級結構元會非常相似，且在隨機產生的反例中，亦很少出現。

若以 GPRM 只針對 16SRNA 家族預測其共同的二級結構元，會如下所示

[8,9] (1,3) [8,14] (1,6) [8,14] (1,6) [8,14] (1,6) [8,14] (1,4) [8,9]

此結構元擁有三個很長的莖幹，在此家族的核醣核酸二級結構元的長相都滿相似的。而 GPRM 針對 VIRUS 家族所預測出來的共同結構元是包含二個莖幹的擬節結構，如 [3,6] (0,3) [4,10] (0,1) [3,6] (0,6) [4,10]

在 VIRUS 家族中的核醣核酸，其二級結構元相對來說差異性較大，並沒有存在一個結構元是出現在所有序列中。因此，我們取 soil-borne mosaic virus 與 archaea 16S rRNA 混合而成此組實驗資料。在本次實驗中，莖幹個數設為 2-3，結果如表格 4 所示。

| | 擷取率 | 正確率 | Matthews 相關係數 |
|---|---|---|---|
| 16SRNA | 0.97 | 0.95 | 0.83 |
| VIRUS | 0.77 | 0.98 | 0.77 |

表格 4 VIRUS + 16SRNA

因為在 16SRNA 中，序列之間的結構元比較一致，而且又有三個莖幹，因此其共同的特徵很容易被發掘，預測出來的結構元較準確。故掃描家族成員時，便可挑出絕大部份的序列，而有極高的擷取率、正確率以及 Matthews 相關係數值。

而 VIRUS 家族的共同結構元，因差異比較大，而無法找到一個結構元是所有序列都有的，只能儘可能地找出一個大部份序列都出現的結構元。因此，所挑選

出來的家族成員，只是一部份的序列，故平均而言，擷取率會較低。但 16SRNA 只有少數幾條序列擁有擬節結構，故 VIRUS 還是有很高的正確率。但因為遺漏掉一些序列，錯誤的反預測(false negative)比較高，而拉低 Matthews 相關係數值。

另一組實驗資料是由 IRE 與 16SRNA 的序列所混合而成。下面所顯示的結構長相亦是利用 GPRM 針對 IRE 家族所預測出來的二級結構元。

[5,8] (1,1) [5,5] (6,7) [5,5] (0,0) [5,8]

雖然 IRE 家族共同結構元的莖幹比較短，但其長度變化量很小，因此，同源序列之間的結構元長相會很相似。所以不同於 VIRUS + 16SRNA 的測試資料，此組資料的特性是，兩個家族的結構元長相都非常一致。

| | 擷取率 | 正確率 | Matthews相關係數 |
|---|---|---|---|
| 16SRNA | 0.81 | 0.73 | 0.67 |
| IRE | 0.73 | 0.99 | 0.85 |

表格 5 IRE + 16SRNA

表格 5 即此次實驗結果。雖然兩家族的結構元長相都非常一致，但由於 IRE 家族的核醣核序列比較長，故在非結構元區域很有可能會出現其它結構，而其中某些擁有 16SRNA 結構元的核醣核酸會被歸到 16SRNA 的家族，因此拉低了 16SRNA 的正確率與 IRE 的擷取率。但為了能包括到 IRE 的序列，才會使得 GPRM 最後收斂到一個較不精確的結構元。因此，以此當成共同特徵去挑選 16SRNA 的家族成員，便會遺漏一些序列，造成較低的擷取率與較高的錯誤的反預測(false negative)。再加上多出了 IRE 的序列，故有較高的錯誤的正預測(false positive)，因此，Matthews 的相關係數值才會比較低。而預測出來的 IRE 共同結構元長相雖然是正確的，但因為一部份的序列已經被歸到 16SRNA 家族了，故有較高的錯誤的反預測(false negative)，而拉低了 Matthews 的相關係數值。

### 5.6.3 不同家族成員個數比例影響

即使 VIRUS + 16SRNA 的序列數目大約是 1:2 (18 條 VIRUS 與 34 條 16SRNA)，IRE + 16SRNA 的數目比大約是 1:3 (18 條 VIRUS 與 56 條 16SRNA)，但由於影響的因素太多，因此我們使用人造測試資料來驗證此系統的可行性。在產生資料時，我們選定兩個長相在其同源序列中較一致的結構作為兩個家族的共同結構元，但改變序列數目的比例。 表格 6、表格 7 與表格 8 分別為序列數目比 1:1、1:2 與 1:3 之實驗結果。

| | 擷取率 | 正確率 | Matthews相關係數 |
|---|---|---|---|
| 第一群 | 1.00 | 0.88 | 0.89 |
| 第二群 | 0.86 | 1.00 | 0.93 |

表格 6 序列比 1:1

| | 擷取率 | 正確率 | Matthews相關係數 |
|---|---|---|---|
| 第一群 | 1.00 | 0.90 | 0.94 |
| 第二群 | 0.95 | 1.00 | 0.96 |

表格 7 序列數目比 1:2

| | 擷取率 | 正確率 | Matthews相關係數 |
|---|---|---|---|
| 第一群 | 1.00 | 0.87 | 0.90 |
| 第二群 | 0.95 | 1.00 | 0.97 |

表格 8 序列數目比 1:3

由實驗結果可看得出來，不同的序列數目比例，對此系統分群與結構預測的正確性並不會有太大的影響。只是因為第二群的序列恰巧擁有第一群的結構元，故少許的序列會被歸類到第一群，使得第一群的正確率與第二群的擷取率無法達到 1.00。但因為兩群的 Matthews 相關係數值都滿高的，顯示出預測出來的共同結構元長相是正確的。

# 第六章　結論與未來研究方向

## 6.1 結論

儘管核醣核酸的研究愈來愈受重視，但焦點議題大多關於二級結構的預測，無論是單一序列，或是針對一個家族。本研究最重要的貢獻是提出新的研究方向，我們希望能提出一套有系統的方法，輸入一群核醣核酸序列，預測所有家族的共同二級結構，再依據此結構元進行分群。

根據本篇研究所提出的方法，我們設計出一套核醣核酸分群的工具，可以減低生物學家進行實驗時的複雜度。使用者可以輸入未排比的核醣核酸序列，透過此系統的分析檢測程序，可以了解此組資料共有幾群不同的核醣核酸家族，並且自動預測每一個家族的共同結構元。

由 VIRUS + 16SRNA 的測試實驗來看，此系統針對共同結構元長相非常一致的資料，如 16SRNA 家族，可以有較好的分群結果，且預測出來的結構元長相也較精準。但對於 VIRUS 家族來說，因為共同結構元長相差異性較大，故容易遺漏了一些成員序列。

另外，透過 IRE + 16SRNA 與 IRE + VIRUS 兩組測試結果來看，此系統容易受到非真正結構元的影響。由於 IRE 中的序列比較長，故在非結構元區域容易形成配對結構，甚至是出現另一家族的配對結構，而影響其它家族結構元長相。例如 IRE + 16SRNA 的實驗中，預測 16SRNA 的結構元長相並不太精準，便是為了囊括某些 IRE 的序列所造成的。

對於 IRE + VIRUS 此組資料來看，雖然 IRE 家族真正的共同結構元只擁有二個莖幹，但由於本系統搜尋共同結構元是從較大的莖幹個數開始，若放鬆了結構參數設定，讓 GPRM 嘗試演化一個較大的結構，最後得到的結構雖有包含到真正的結構區域，但也多了一個不重要的莖幹。受限於 GPRM 的所有演化個體是相同莖幹數目的結構，這類的問題的確是本研究的限制。雖然正確的結構參數可以解決這樣的問題，但往往正確的結構參數是不可預知的。不過，卻可嘗試修改 GPRM，允許演化個體中存在不同莖幹個數的結構，讓真正的結構元與其它結構互相競爭。然而，這需要增加群體個數，那麼勢必會造成演化時間要更長才能收斂到正確的共同結構元。

而由人造的測試資料發現，不同的家族大小對本系統在分群與結構元預測上，並不會造成太大的困擾。即使序列數目相差懸殊，高達 1:3，依然能準確地預測出結構元，並挑出正確的家族成員。因為我們認為，決定性的因素有兩個，其一是每個家族的共同特徵是否夠明顯，即共同的二級結構元長相是否夠一致。而另一個關鍵在於，非結構元區域是否容易形成配對結構，這會影響分群的結果，甚至是最後預測出來的結構元長相。

## 6.2 未來研究方向

此章節將提出幾項在研究過程中遭遇到的困難，未來可針對這些問題進行改善與更進一步的相關研究。

### 6.2.1 參數設定太過寬鬆

在使用者輸入的序列中，不同家族之間的結構元可能非常不相似，但卻只用同一組參數來描述所有家族的結構元長相。而為了涵蓋這些不同結構元的特性，會傾向設定較寬鬆的結構參數值，例如莖幹個數太多、莖幹環線長度範圍太大，或者是不恰當的錯誤配對容忍度等等。這容易造成錯誤的正預測太高，正確的答案淹沒在一堆無用的資料中，增加使用者讀取結果的負擔。甚至更糟糕的情況是預測出不精確的結構元。

經過多次的實驗測試，我們發現過鬆的錯誤配對容忍度會造成共同結構元預測失敗。我們希望能透過演化的？爭機制來決定此家族的共同結構是否擁有錯誤配對，因此，若是允許存在錯誤配對的情況，在產生演化個體時，擁有與不擁有錯誤配對的二級結構數量會各占一半。互相競爭的結果，擁有正確結構設定的個體會脫穎而出。透過這樣的機制，便可預測出正確的共同結構元。

除了錯誤配對容忍度，太大的莖幹長度範圍，依然會導致 GPRM 無法預測出正確的結構元。我們認為自動調整此項結構參數值可以解決這樣的問題。既然一開始設定的莖幹長度範圍太鬆，那麼就先嘗試用較嚴格的參數值，之後再逐次拉開差距。舉例來說，一開始設定的莖幹長度最小是 3，最大是 15。再選定一個梯度 5，那麼就有 10-15、5-15、3-15 這三組莖幹長度範圍需要一一檢查。每用一組參數跑完本系統的流程，就會預測出一個結構元，若此結構元的拉普拉斯值超

過門檻，或者預測出此組資料存在兩個以上的家族(即預測出某一個家族真正的成員總數且並非最大的)，則計算此結構元的平均鹼基數。最後，再從所有可能的結構元中，選平均鹼基數最大者當作家族代表的結構元。這理由便是鹼基數愈多的結構，其自由能(free energy)會愈小，結構愈穩定。

若這三組參數值預測出來的結構元拉普拉斯值都沒有超過門檻，或者家族成員個數預測皆失敗(即認為所有序列為相同家族)，則從中選擇適應分數最高者為我們的答案。因為，根據 GPRM 設計的理念，針對同樣家族的核醣核酸序列，預測出來的結構元是適應分數最高者，即偏好那種出現次數愈多的結構元。

因此，即使為了配合其它家族代表結構的設定，而被迫採用一組不恰當的結構參數，依然可預測出正確的共同結構元。

### 6.2.2 拉普拉斯門檻值的決定依據

在分群程序開始前，須先設定的參數其中一項為拉普拉斯門檻值。此數值是在預估家族大小時，用來調整猜測的方向(詳細過程，可參照 4.4.3 尋找代表性的結構元)。然而，此門檻值的選定缺乏理論的依據，太高或太低都會誤導猜測結果，只能透過多次的實驗，來決定較好的門檻值。理論上，門檻值的大小須要根據序列內容自動調整，例如，對於那些共同結構元長相不太一致的核醣核酸，門檻值應該降低才能預測出正確的家族大小。這雖然可減少參數設定的困擾，但卻會使得系統太過複雜，且增加程式執行的時間。目前此參數可交由使用者設定，未來可嘗試設計出一套演算法自動尋找適當的門檻值。

### 6.2.3 執行時間太過冗長

由於本研究在尋找共同結構元時，並沒有家族大小(答案數目)的資訊，因此採用二分逼進法來估計可能的數目。對每一個新的數值，都必須重新尋找可能的目標結構，由此答案的拉普拉斯值來決定家族大小該如何調整。故此過程須反覆執行 GPRM，如此一來，執行時間便會拉長。在 GPRM 原本的設定中，每一次結構元的預測會固定演化三十代。然而，就本研究所進行的實驗觀察中，不必三十代，便會出現結構元的拉普拉斯值通過門檻了。故在預測家族大小時，若有此種情況發生，演化便可提早結束。雖然利用拉普拉斯值的判斷，可縮短執行時間，但往後依舊須要找出新的解決方法，能更快且依然能準確地預測出家族大小，才能再提高此系統的實用性。

### 6.2.4 負面背景序列的產生方式

GPRM 預測共同的二級結構時，採用監督式學習(supervised learning)，故另須一份對照的序列資料充當錯誤範例。在本系統中，主要使用一級(first order)方式來產生反例資料，但亦曾嘗試零級(zero order)的產生方式。零級(zero order)方式指的是，四種鹼基出現的機率是根據正例中四種鹼基的機率來決定，且在決定每個位置上的鹼基時並不會受到鄰近的鹼基類型所影響，即這些是完全獨立(i.i.d.)的序列。

當目標結構擁有兩個莖幹時，此種反例序列產生方式還算恰當。但是對於擁有三個莖幹的結構元，無論何種莖幹配對組合，在反例序列中幾乎很少出現。這浮現出的問題是，負面背景序列並沒有表現出該有的功能，亦即區分出有代表性的結構元。因此，接下來的改進目標之一是，找到一個更合適的反例序列產生方法。

# 第七章　參考文獻


Akmaev V., Kelley S. and Stormo G. (1999). A phylogenetic approach to RNA structure prediction. In *Proc Int Conf Intell Syst Mol Biol,* pp. 10-7 AAAI Press.

Alm Rosenblad M., Gorodkin J., Knudsen B., Zwieb C., and Samuelsson T. (2003). SRPDB (Signal Recognition Particle Database). *Nucleic Acids Res.,* 31, Database Issue, 363-364. (http://psyche.uthct.edu/dbs/SRPDB/SRPDB.html)

Antoinette C.K. van, Donato L.P.B., John T.D., Jolanda M., Ronald E.W. and Jaap G. (2002). Phylogenetic Relationships among the Species of the Genus Testudo (Testudines: Testudinidae) Inferred from Mitochondrial 12S rRNA Gene Sequences. *Molecular Phylogenetics and Evolution,* 174-183.

Batenburg F.H.D.van, Gultyaev A.P., Pleij C.W.A., Ng J. and Oliehoek J. (2000). Pseudobase: a database with RNA pseudoknots. *Nucleic. Acids Res.* 28,1, 201-204. (http://wwwbio.leidenuniv.nl/~batenburg/pkb.html )

Berman H.M., Olson W.K., Beveridge D.L., Westbrook J., Gelbin A., Demeny T., Hsieh S.H., Srinivasan A.R. and Schneider B. (1992). The Nucleic Acid Database: A Comprehensive Relational Database of

Three-Dimensional Structures of Nucleic Acids. *Biophys. J.,* 63, 751-759. (http://ndbserver.rutgers.edu/NDB/)

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 Res.,* 28, 235 – 242.

Brown J.W. (1999). The Ribonuclease P Database. *Nucleic Acids Res.,* 27,314. (http://www.mbio.ncsu.edu/RNaseP/)

Cech T.R. and Bass B.L. (1986). Biological catalysis by RNA. *Annual Review of Biochemistry,* 55, 599-629.

Chen J.H., Le S.Y., and Maizel J.V. (2000). Prediction of common secondary structure of RNAs: a genetic algorithm approach. *Nucleic Acids Res.,* Vol. 28, No. 4, 991-999.

Chiu D.K.Y. and Kolodziejczak T. (1991). Inferring consensus structure from nucleic acid sequence. *Comput. Applic. Biosci.,* 7,347-352.

Clark P. and Niblett T. (1989). The CN2 Induction Algorithm. *Machine Learning* 3, 261-283,

Dam E., Pleij K. and Draper D. (1992). Structural and functional aspects of RNA pseudoknots. *Biochemistry,* 31, 11665-11676.

David P.G., Carla A.T. and Paul L.N. (2000). Structure, Stability and Function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *J. Mol. Biol.,* 298, 167-185.

Dock-Bregeon A.C., Chevrier B., Podjarny A., Johnson J., de Bear J.S., Gough G.R., Gilham P.T. and Moras D. (1989). Crystallographic structure of an RNA helix: $[U(UA)_6A]_2$. *J. Mol. Biol.* 209:459-474.

Eddy S.R. and Durbin R. (1994). RNA sequence analysis using covariance models. *Nucleic Acids Res.,* 22, 2079-2088.

Ehresmann C., Baudin F., Mougel M., Romby P., Ebel J.P. and Ehresmann B. (1987). Probing the structure of RNAs in solution. *Nucleic Acids Res.,* 15, 9109-9128.

Ekland E.H. and Bartel D.P. (1996). RNA-catalysed RNA polymerization using nucleoside triphosphates. *Nature* 382:373-376.

Gilbert W. (1986). The RNA world. *Nature,* 319:618.

Gorodkin J., Heyer L.J. and Stormo G.D. (1997). Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res.,* Vol. 25, No. 18, 3724-3732.

Gorodkin J., Shawn L. and Stormo G.D. (2001). Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids Res.,* Vol. 29, No.

10, 2135-2144.

Gulko B. and Haussler D. (1996). Using multiple alignments and phylogenetic trees to detect RNA secondary structure. In *Proc Pac Symp Biocomput,* pp. 350-367.

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

Han K. and Kim H.J. (1993). Prediction of common folding structures of homologous RNAs. *Nucleic Acids Res.,* 21, 1251-1257.

Hofacker I.L., Fekete M., Flamm C., Huynen M.A., Rauscher S., Stolorz P.E. and Stadler P.F. (1998). Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucleic Acids Res.,* 26, 3825-3836.

Hu Y.J. (2002). Prediction of consensus structural motifs in a family of coregulated RNA sequences. *Nucleic Acids Res.,* 30, 3886-3893.

Kim J., Cole J.R. and Pramanik S. (1996). Alignment of possible secondary structures in multiple RNA sequences using simulated annealing. *Comput. Appl. Biosci.,* 12, 259-267.

Klosterman P.S., Tamura M., Holbrook S.R., Brenner S.E. (2002). SCOR: a structural classification of RNA database. *Nucleic Acids Res.* 30. 392-394. (http://scor.lbl.gov )

Koza J.R. (1992). *Genetic Programming: on the programming of computers by means of natural selection.* MIT Press.

Kruskal W.H. and Tanur J.M. (1978). International encyclopedia of statistics. New York, NY: Free Press.

Kundsen B. and Hein J. (1999). RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics,* Vol. 15, No. 6, 446-454.

Kundsen B. and Hein J. (2003). Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.,* Vol. 31, No. 13, 3423-3428.

Lewis, D. and Gale, W.A. (1994). A sequential algorithm for training text classifier. *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval.* 3-12.

Matthews, B.W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochem. Biophys. Acta.* 405, 442-451.

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

Murthy V.L. and Rose G.D. (2003). RNABase: an annotated database of RNA structures. *Nucleic Acids Res.,* 31,502-504. (http://www.rnabase.org/)

Narendra P. and Fukunaga K. (1977). A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers,* 26(9): 917-922.

Notredame C., O'Brien E.A. and Higgins D.G. (1997). RAGA: RNA sequence alignment by genetic algorithm. *Nucleic Acids Res.,* 25, 4570-4580.

Parsch J., Braverman J.M. and Stephan W. (2000). Comparative Sequence Analysis and Patterns of Covariation in RNA Secondary Structures. *Genetics,* 154, 909-921.

Pley H.W., Flaherty K.M. and Mckay D.B. (1994). Three-dimensional structure of a hammerhead ribozyme. *Nature,* 372:68-74.

Rivas E. and Eddy S.R. (1999). A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.,* 285, 2053-2068.

Ruan J., Stormo G.D., and Zhang W. (2004). An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics,* Vol. 20, No. 1,58-66.

Sakakibara Y., Brown M., Hughey R., Mian I.S., Sjoelander K., Underwood R. and Haussler D. (1994). Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res.,* 22, 5112-5120.

Schwartz A. W. (1995). "The RNA World and its origins". *Planet and Space Science,* 43, 1-2.

Scott W.G., Finch J.T. and Klug A. (1995). The crystal structure of an all-RNA hammerhead ribozyme: a proposed mechanism for RNA catalytic cleavage. *Cell* 81:991-1002.

Sprinzl M., Horn C., Brown M., Ioudovitch A. and Steinberg S. (1998). Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic. Acids Res.*, 26, 148-153. (http://www.staff.uni-bayreuth.de/~btc914/search/index.html)

Tamura M., Hendrix D.K., Klosterman P.S., Schimmelman N.R.B., Brenner S.E. and Holbrook S.R. (2004). SCOR: Structural Classification of RNA, Version 2.0. *Nucleic Acids Res.* 32. 182-184.

Thompson J.D., Higgines D.G. and Gibson T.J. (1994) CLUSTAL w: improving the sensitivity of progressive multiple sequence alignment through

sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22, 4673-4680.

Varani G. and McClain W.H. (2000). The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Rep.* Jul;1(1):18-23.