國立交通大學

資訊科學系

碩 士 論 文

利用核醣核酸的二級結構及基因規劃法
找尋其共同結構元

Using RNA secondary structure information to predict common structural motif by Genetic Programming

研究生：林勁伍

指導教授：胡毓志　博士

中 華 民 國 九 十 四 年 六 月

利用核醣核酸的二級結構及基因規劃法找尋其共同結構元

Using RNA secondary structure information to predict common structural motif by Genetic Programming

研 究 生：林勁伍　　Student：Jing-Wu Lin

指導教授：胡毓志　　Advisor：Yuh-Jyh Hu

國 立 交 通 大 學

資 訊 科 學 系

碩 士 論 文

A Thesis

Submitted to Department of Computer and Information Science

College of Electrical Engineering and Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer and Information Science

June 2005

Hsinchu, Taiwan, Republic of China

中華民國九十四年六月

利用核醣核酸的二級結構及基因規劃法找尋其共同結構元

研究生：林勁伍　　指導教授：胡毓志博士

國立交通大學資訊科學系碩士班

摘　　要

功能性核醣核酸參與了許多生命中的重要反應，使得人們想要了解各核醣核酸的功能。由於核醣核酸的功能與二級結構有密切關係，使得預測核醣核酸二級結構成為生物資訊中一個發展中的課題。本論文中，我們提出一個融合多重策略的方法，自相關核醣核酸序列的二級結構，尋找具特徵的共同結構元。我們首先使用 Mfold 做為前處理器，預測單一核醣核酸序列的完整二級結構。之後，將 Mfold 的結果轉換為我們設計的核醣核酸表示法，使用基因規劃法找出其共同的結構元。在測試了數個真實的核醣核酸家族之後，我們驗證了本方法能夠有不錯的表現。

Using RNA secondary structure information to predict common structural motif by Genetic Programming

student：Jing-Wu Lin　　Advisors：Dr. Yuh-Jyh Hu

Department of Computer and Information Science
National Chiao Tung University

ABSTRACT

Functional RNA molecules play an important role in many biological activities such that people want to understand the function of them. As there is high correlation between the function and the secondary structure of RNA, RNA secondary structure prediction has become an active topic of bioinformatics. In this thesis, we propose a multi-strategy approach to find common structural motif within functionally related RNAs. We first use Mfold as a preprocessor, to predict the global secondary structure of single RNAs. After we transform the outputs of Mfold into our RNA representation, we apply genetic programming to identify common structural motifs. The new method has been tested on several real RNA families, and demonstrated promising performance.

致　　謝

終於了解為什麼台上得獎的人總是先死板的感謝父母。感謝他們能忍受一個孩子消失這麼久的時間，全力支持我取得碩士學位，使我能無後顧之憂的專注在學業上。

感謝一同奮戰、攜手走過艱苦歲月的實驗室同學－秉蔚、昀君、音璇、秀琴、世彥，沒有你們，本研究一定研究不出來。感謝有姊姊氣息的宛嫻學姊，研究相同領域的美華學姊，以及經驗豐富的萬田學長；後起之星阿貴、豐茂、貫中，還有來自八方好友們的鼓勵或詛咒，一種是支持我咬緊牙關撐下去的原因，另一種則是燃起我不服輸的鬥志。

最後，指導教授胡毓志老師在這兩年的辛苦的指導與督促，讓我這麼樣的一個生物資訊新手，能夠完成這份碩士論文。此外，口試委員葉佳炫老師、荊宇泰老師的指教，使本論文能更為完善。

目　　錄

第一章 前言

1.1 研究動機

一九九零年代所開始的人類基因體計畫完成了絕大部分人類基因的定序，爾後的重點就是要探討這些基因所代表的功能。從微觀的角度來看，維持生理機能有兩種最重要的分子：核酸及蛋白質，其中蛋白質直接影響生物體的運作，核酸則是保存蛋白質結構的藍圖。

依分子生物學的中心教條－ "去氧核醣核酸(DNA, deoxyribonucleic acid)轉錄成核醣核酸（RNA, ribonucleic acid)，核醣核酸轉譯為蛋白質" 所示，核醣核酸似乎僅是一個中間產物，攜帶去氧核醣核酸的資訊到核糖體，合成所需要的蛋白質。然而，這只是眾所皆知的一類核醣核酸－信使核醣核酸(messager RNA, mRNA）而已。其他常見的還有轉移核醣核酸，（transfer RNA, tRNA）, 核糖體核醣核酸（ribosomal RNA, rRNA)。尤其是核酶（ribozyme)、微核醣核酸（microRNA)被發現後，更是顛覆了核醣核酸原本的地位－原來它們也像蛋白質一樣的重要。這些 RNA 會折疊成特定的形狀來輔助生命機制，包含催化化學反應及調控基因表現等等。並且在少數情況下，mRNA 本身所形成的結構也會影響生物體。

在發現核醣核酸在生命中是佔有很重要的地位後，人們也積極的想要掌握核醣核酸的功能。若能由方便得知的 RNA 序列來預測其結構，進而猜測功能，未來就能透過分子層級的方法來控制生命運作。

然而，核醣核酸的結構不固定、運作時的結構不見得是擁有最小能量、以及不像 DNA 很單純的僅有兩種配對等種種特性，使其結構的預測極為不易。發展有效率、高正確性的結構預測工具是必要的。本研究試圖從一群功能相同的核醣核

酸，加入利用能量預測二級結構的資訊，以預測其共同的結構元，希望能直接點出核糖核酸行使功能的區域及結構。

1.2 研究假設

本研究設定了兩個合理的基本假設：

一、具有相同功能的序列，會擁有共同的結構元 [Pley et al, 1994; Scott et al, 1995; Lewis, 2003] 。

由演化的角度來看，重要的基因序列會在演化中保存下來，但是在功能性的核醣核酸上，並非十分的強烈。由化學的角度來看，當結構有些許變化就很可能影響分子結合的能力。因此我們認為，一群功能相同的核糖核酸序列行使功能之區域，二級結構必定幾乎相同。

二、行使功能的共同結構元不容易在完全隨機的狀態下出現。

倘若一個擁有重要功能的結構元能任意產生，必然會輕易改變生物體的生命機制，此物種不會被保留下來。相同的假設也被運用在其他系統 [Hu, 2002]。

1.3 研究目的

目前，已有許多研究者使用各種方法來嘗試解決這個問題。包含圖論方法、演化式計算、隱式馬可夫模型、序列排比等等。其中使用基因規劃法(GP, Genetic Programming)的 GPRM 是一個彈性大、正確率良好、不需要依賴序列排比的系統。但可惜對於突起結構及非對稱環狀結構的容忍性不佳，完全不使用任何領域知識（domain knowledge）也使得搜尋空間過大，需要更多時間來彌補。

在此我們同樣使用基因規劃法，嘗試改進 GPRM 對於突起結構及非對稱環狀結構處理不佳的地方，加入能量的資訊以縮小搜尋空間，以期能處理數量更多、結構更複雜的核糖核酸家族。

1.4 論文架構

本論文分為六個章節及附錄，簡述如下：

第一章前言，介紹本研究的動機及背景，以及所使用的方法與面臨的問題。

第二章文獻探討，將會介紹所需的背景知識，以及該議題過去的發展。

第三章研究方法，是本篇論文的核心，詳細介紹本研究所設計的方法。

第四章實驗結果，所有的實驗內容及結果將在此做個整理。

第五章結論與討論，分析本研究的優缺點。

第六章參考文獻，列出本研究的參考資料。

第二章 文獻探討

2.1 核醣核酸簡介

1953 年，生物學有個驚人的大突破。Watson 與 Crick 發現了去氧核醣核酸會以美麗的雙股螺旋結構存在生物體中[Watson, Crick, 1953]，並著手建立了生命基本原理的中心教條，開啟了遺傳學與分子生物學的大門。

核醣核酸在中心教條中，僅是將去氧核醣核酸的訊息帶到核糖體，轉譯成蛋白質。然而，這是一個過度簡化的流程。隨著越來越多的研究發現，維持複雜生命機制絕不是如此一條直線流程所能代表。

2.1.1 核醣核酸的重要性及資訊學扮演的角色

除了早期的功能性核醣核酸外，能作為催化劑的核酶也陸續被發現，而能夠調控許多關鍵基因表現的微核醣核酸更是令人驚豔。

微核醣核酸是一群非常短，長度約二十多個鹼基的核醣核酸，最明顯的特徵就是所有微核醣核酸的先質（precursor）都具有一個類似髮夾的構造（將在下一小節介紹），而這些構造在基因體裡是相當穩定的。微核醣核酸在後轉錄時期（post-transcription）參與調控。其影響包含控制細胞凋亡、組織成長、肥胖代謝，以及決定某些基因的表現時間。[Lee et al, 1993; Reinhart et al, 2000; Ambros, 2001]

微核醣核酸很小，而且不會轉錄成蛋白質，研究上很難從細胞中直接分離出來。在此，電腦的輔助就顯得十分重要了。Burge 及 Bartel 等人透過電腦預測基因體中的微核醣核酸的數目與位置。他們首先利用微核醣核酸在結構上的特性，尋找線蟲 C. elegans 基因體中所有會形成髮夾構造的序列。並且利用微核

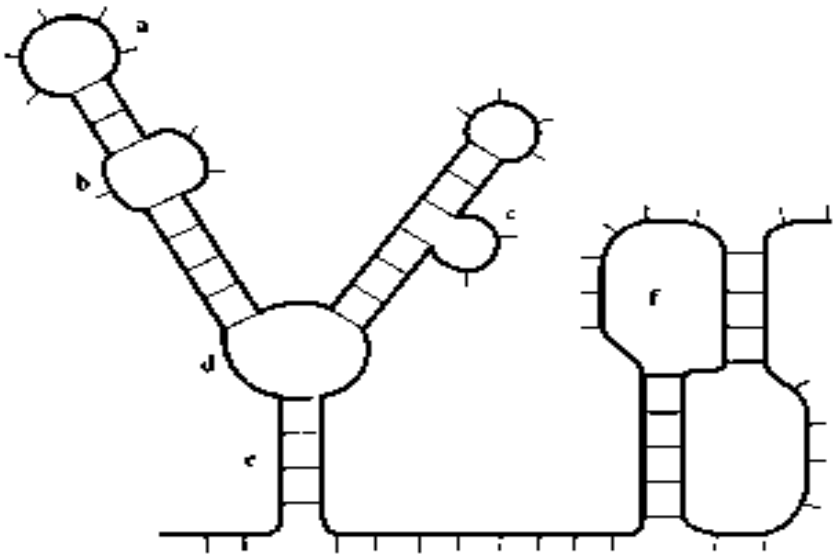
醣核酸在演化上保留的特性，與另外一個相近的線蟲種 C. briggsae 的基因體是否有相同序列。之後，再將這些微核醣核酸的可能序列進一步與目前已知的微核醣核酸序列進行比對，以找出在結構上與已知微核醣核酸更為相似者，最後進行分子試驗以確認是否真的是微核醣核酸[Rhoades et al, 2002]。除此之外，相同的特性也被運用在尋找果蠅上的微核醣核酸，同樣的透過電腦的協助 [Lai et al, 2003]。

這僅是核醣核酸參與重要調控的其中一例，而且我們不難發現，為了處理龐大的資料，藉由電腦輔助分析，已是生物研究上一個非常重要的環節。

2.1.2 核醣核酸序列及二級結構

核醣核酸由 A、G、C、U 四種含氮鹼基組成，四個字母分別代表腺嘌呤（Adenine）、鳥糞嘌呤（Guanine）、胞嘧啶（Cytosine）、尿嘧啶（Uracil）。為了版面清晰及方便，本論文以後將會使用這四個字母來表示核醣核酸的四種成分。

核醣核酸二級結構的概念開始於 1959 年，Doty 與 Fresco 揭開了這段序幕[Doty et al, 1959; Fresco et al, 1960]。不同於去氧核醣核酸，核醣核酸常以單股存在生物體中，透過分子間的作用力，自我折疊成特定結構。這些作用力主要包含標準鹼基對(canonical base pair)－C、G 間形成三個氫鍵，A、U 間形成兩個氫鍵，如此兩兩配對，構成核醣核酸的基本結構。還有一對被稱為擺動鹼基對(wobble base pair)的 G、U 配對， G、U 間只會形成一個氫鍵，所以此配對不太穩定，需透過周圍的鹼基對輔助。此外，U、U，A、G 等等都可能在特定的情況下配對。這些配對，形成了核醣核酸的基本結構，被稱為二級結構。



圖一. 核醣核酸常見的二級結構[本圖出自 http://ludwig-sun2.unil.ch/~bsondere/nussinov]

圖一列出基本的核醣核酸二級結構，說明如下：

1. 莖幹結構（stem）

   若核醣核酸序列中，連續的鹼基配對成一個長條的形狀，稱為莖幹結構（如圖中 e）。

2. 髮夾環狀結構(hairpin loop)

   當莖幹一端的鹼基完全沒有互相配對，該區域稱為髮夾環狀結構。或者可以如此定義：當一個連續的非配對區域不是在序列的終端，而且僅與一個莖幹相鄰的話，該區域就是一個髮夾環狀結構（如圖中 a）。此外，此環狀結構和相鄰的莖幹合稱一個髮夾結構。

3. 內部環狀結構(internal loop)

   一個連續的非配對區域恰與兩個莖幹相鄰，並且兩側都有未配對的鹼基，該區域就是一個內部環狀結構。該環狀結構看起來就像把一個長的莖幹結構從中間截斷，使其分為兩半。內部環狀結構分為對稱及非對稱結構。當兩旁未配對的鹼基個數相同時，稱為對稱的內部環狀結構；反之則稱為非內部環狀結構（如圖中 b）。

4. 突起結構(bulge)

在莖幹中僅一邊有未配對的鹼基，而另一邊都是連續的鹼基對，則稱這些沒配對的區域為突起結構（如圖中 c）。

5. 多分支環狀結構(multi-branched loop)

類似內部環狀結構，但當該環狀結構與三個以上的莖幹接觸時，稱為多分支環狀結構（如圖中 a）。

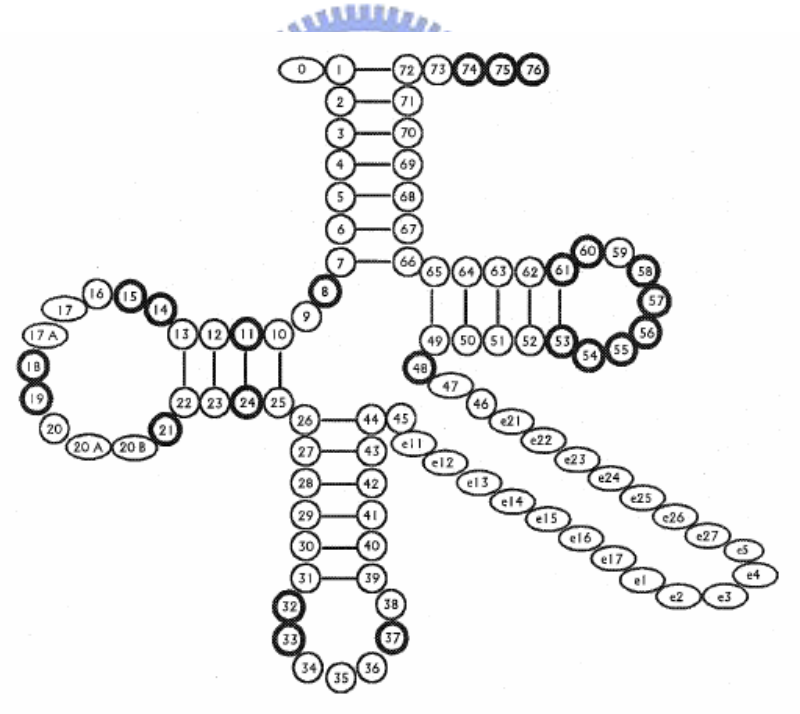
6. 擬結結構(pseudo-knot)

擬結結構是一種比較特別的結構，形成的主因是莖幹交錯配對。當莖幹間的鹼基會與莖幹外的鹼基形成配對時，該結構看似就像打結一樣，但實際上並沒有，因此稱為擬結結構。

2.2 預測核醣核酸結構的相關方法

物種經過長時間的演化後，原本擁有相同生化功能的分子會有些微變異。生物學家們會將這些分子收集成一個家族（family），以方便分析物種間的關係，並降低資料的複雜性。

目前生物學家認為，分子的結構是影響其功能的關鍵。例如常見的功能性核醣核酸之一－轉移核醣核酸，負責在轉譯作用時攜帶對應的胺基酸，長度大概介於七十到八十個鹼基。其結構都是很穩定的苜蓿葉（cloverleaf）結構：包含四個莖幹結構，形狀類似十字架或四瓣的苜蓿葉，如圖二所示[Sprinzl et al, 1998]。



圖二. 傳遞者核糖核酸的二級結構

因此，如果能有一個方便、迅速的方法，可以立即發現家族成員中的共同結構元，對生物學是很重要的。這不只能協助他們馬上了解該家族功能表現的區域，簡化盲目的嘗試，更甚者，利用已知的共同結構，檢驗功能未知的序列，以

推論其功能。

關於核醣核酸二級結構的預測可以追溯到將近四十年前，Tinoco 所提出使用最鄰近能量參數最小化能量的方法[Tinoco, 1971]，之後各類預測法如雨後春筍般的冒出，期望能一舉解決所有問題，但仍然沒有一個殺手級的系統能預測得又快又準。

底下簡述與本研究密切相關的兩種核醣核酸結構預測系統，以及近年預測家族核醣核酸共同結構元所使用的方法：

2.2.1 GPRM[Hu, 2002]

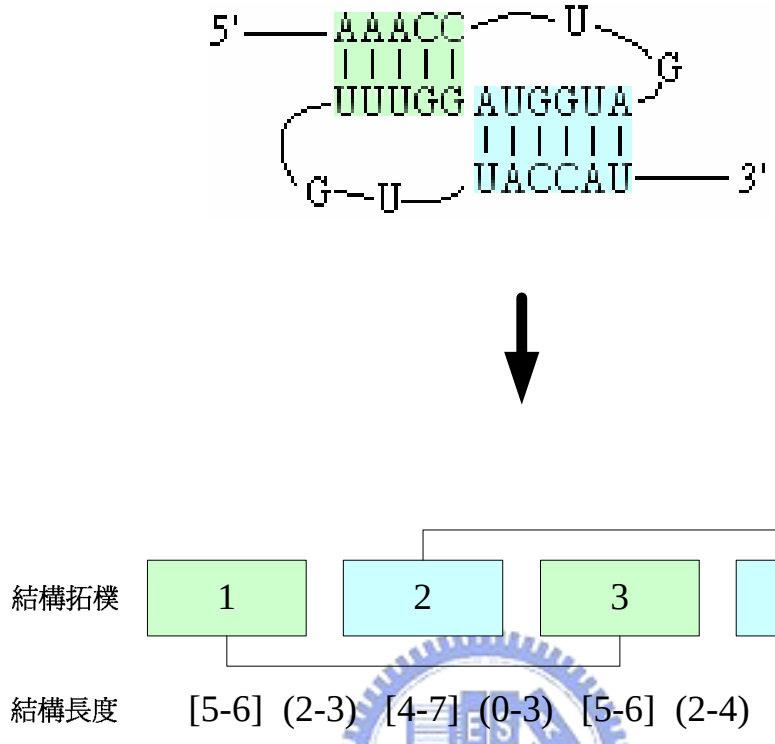
GPRM 為本實驗室於 2002 年所提出的核醣核酸二級結構的預測系統，其特點是不需考慮能量、不依賴序列排比、針對一個功能相同的家族尋找其共同結構元。採用基因規劃法，尋找相同功能核醣核酸的共同二級結構元。GPRM 將二級結構元轉為特定語言，視為基因規劃法中的一個個體（individual），透過包含突變（mutate）、互交（crossover）等機制產生變異，以演化出最適合的個體。

其關鍵在於，描述語言將二級結構元分成兩大部分：結構拓璞與結構長度的範圍。結構拓璞描述了莖幹各股（strand）的相對位置，而結構長度範圍限制莖幹與環狀結構的大小。如此一來，便可以描述一個形狀一樣，大小類似的結構，符合分子行使功能的限制。

描述語言的意義如下：

1. 允許 G、C，A、U，G、U 三種鹼基配對。
2. 一個莖幹至少要有三個鹼基對。
3. 允許長度大或等於四的莖幹包含些許的未配對鹼基（通常為一到二個）。也

就是說允許莖幹中包含小的對稱內部環線。而莖幹的長度會包含這些未配對鹼基。



圖三. 擬結結構在 GRPM 描述語言中的表示法

圖三展示了 GPRM 語言表達擬節結構的方式。結構拓樸中，矩形表示莖幹的一股，線條所連接兩個矩形表示此兩股配對成為一個莖幹結構。圖中包含四股兩個莖幹，其中第一個莖幹為第一股與第三股配對而成；第二個莖幹是第二與第四股。下方的結構長度表示出各莖幹結構與環狀結構的大小：方括號代表莖幹結構的長度範圍；小括號則代表環狀結構的長度範圍。如圖中，第一莖幹的長度範圍為五到六個鹼基，而第一股與第二股兩者間的未配對區域型成一個長度二到三的環狀結構。

帶有彈性的結構大小可以包容形狀一樣但大小有差異的共同結構元，但可惜突起及非對稱的環狀結構會將莖幹分割成數個小莖幹，而莖幹數目的增多則同時增加了 GPRM 的搜尋空間（search space)降低其使用性與準確度。

本系統將試圖彌補這個不足之處。

2.2.2 Mfold[Zuker, 2003; Mathews et al, 1999]

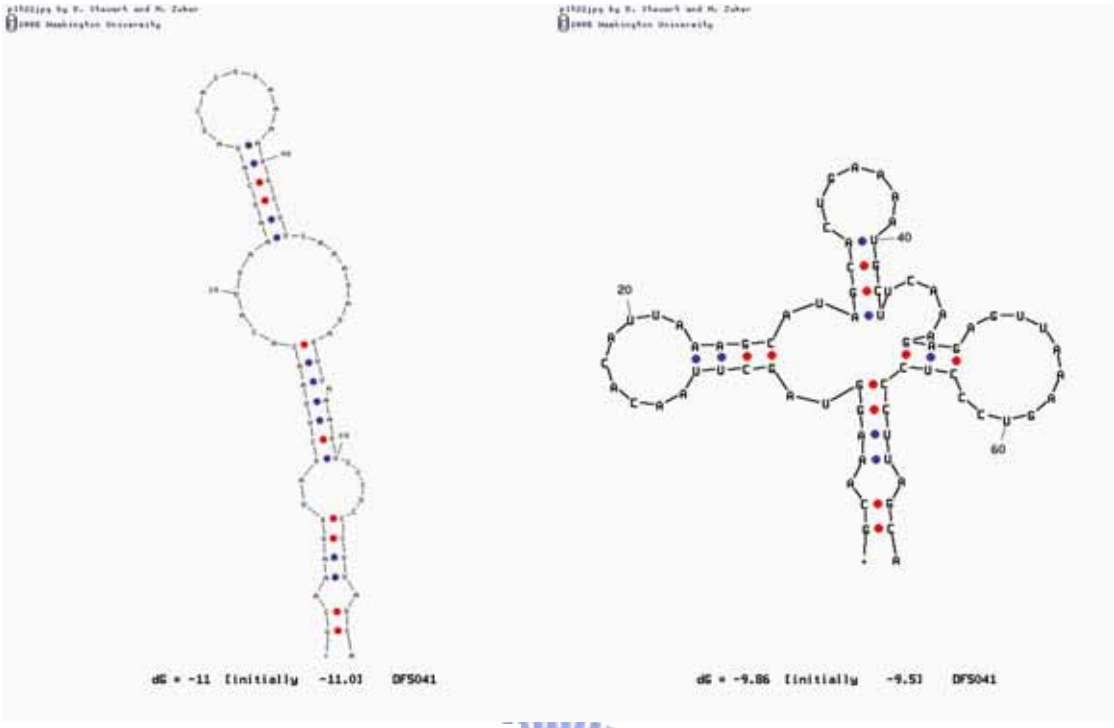
Mfold 是一套單一核醣核酸序列的二級結構預測系統，在 1989 年 Zuker 所提出[Zuker, 1989]，利用動態規劃法（Dynamic programming）計算核醣核酸序列各種摺疊方式的自由能量，以預測出能量最小、最穩定的結構。

以能量辨別二級結構有一定的正確率，不過，核醣核酸在摺疊過程可能因為某些因素甚至受其他分子影響，使得理論的上最穩定結構無法形成。因此單純以最小能量來斷定結構會有許多錯誤產生。除此之外，無法處理擬結結構也是 Mfold 的一項缺點。

然而，若能提供部分已知條件給 Mfold，例如限制拓樸結構或是強制某些鹼基配對，正確率將能大大的提升。此外 Mfold 會提供數個次佳結果，供有背景知識的使用者選擇。

因此，Mfold 仍提供了相當程度的資訊，被廣泛的運用在相關研究上[Lai et al, 2003 ;Fera et al, 2004; Punginelli et al, 2004; Thiviyanathan et al, 2004]。

下頁圖四為 Mfold 所預測 Phe-tRNA 中的 DF5041 序列。(a)圖為 Mfold 所計算出自由能最小最穩定的結構，但是(b)圖才是 DF5041 序列正確的結構，它是 Mfold 預測結果中，自由能第七小的結構。



(a)第一順位是完全錯誤的結　　(b) 正確埋藏在第七順位。

圖四. 一個 Mfold 的預測例子（使用預測的參數）:

2.2.3 其他方法

首先同時研究核醣核酸家族序列的是 Sankoff，他使用動態規劃法同時地排必、預測二級結構及建立物種的親源關係[Sankoff, 1985]。這聽起來很夢幻，可惜複雜度高達 *O(L^3N)*。之後 Gorodlin 等人改用了貪婪演算法（greedy algorithm）降低了不少複雜度[Gorodkin et al, 1997]，但仍然需要依賴序列排比。這很可能受到協同變異(comutate)的干擾。全域的序列排比工具 CLUSTAL W 也有相同的問題[Thompson et al, 1994]。

隨機前後文無關文法（stochastic context-free grammar, SCFG）也是被用來嚐試的方法[Eddy and Durbin, 1994; Holmes and Rubin, 2002]，同樣的也受到序列層級的牽制。運用到圖論演算法中的最大權重配合法（maximum weighted matching algorithm）與最大集團搜尋法（maximum clique finding algorithm）的兩類系統，十分依賴序列或是結構排比的結果[Cary and Stormo, 1995; Ji et al, 2004]。

2.3 核醣核酸資料庫

預測二級結構的原料－基因序列或核醣核酸序列及其已知結構也是相關研究的一大重點。這裡介紹一些本研究所使用到的測試資料來源，以及其他常用的資料庫：

1. Rfam[Griffiths-Jones et al, 2005]

   (http://www.sanger.ac.uk/Software/Rfam/)

   Rfam 為少數包含多數家族的核醣核酸資料庫之一，並且它也提供核醣核酸折疊的資訊，是以隨機前後文無關文法為主建立的資料庫。至 2005 年三月，已包含了 503 個核醣核酸家族，超過四十萬條核醣核酸序列。

2. 5S ribosomal RNA database[Szymanski, 2002]

   (http://rose.man.poznan.pl/5SData/)

   此資料庫是專為 5S 的核糖體核醣核酸所建置，包含目前這些序列的排比資訊及二級結構。除此之外，也提供與這些核醣核酸結合的蛋白質，資訊十分完整。

3. tRNA Compilation 2000[Sprinzl et al, 1996]

   (http://www.staff.uni-bayreuth.de/~btc914/search/)

   提供轉錄者核醣核酸的序列，及其結構資訊。該資料庫還提供搜尋功能，並依照物種及所攜帶的胺基酸分類。

4. The miRNA Registry[Griffiths-Jones S, 2004]

   (http://www.sanger.ac.uk/Software/Rfam/mirna/index.shtml)

   至 2005 四月，共收集了 1650 條微核醣核酸序列，可依物種分類瀏覽。此外，此資料庫也包含各微核醣核酸的先質，也有提供搜尋介面。使用者可以根據序列

片段、編號或名稱進行搜尋。

5. 其他常見資料庫

PseudoBase (http://wwwbio.leidenuniv.nl/~Batenburg/PKB.html) [Batenburg et al, 2000] 收集了許多結構包含擬結的序列。

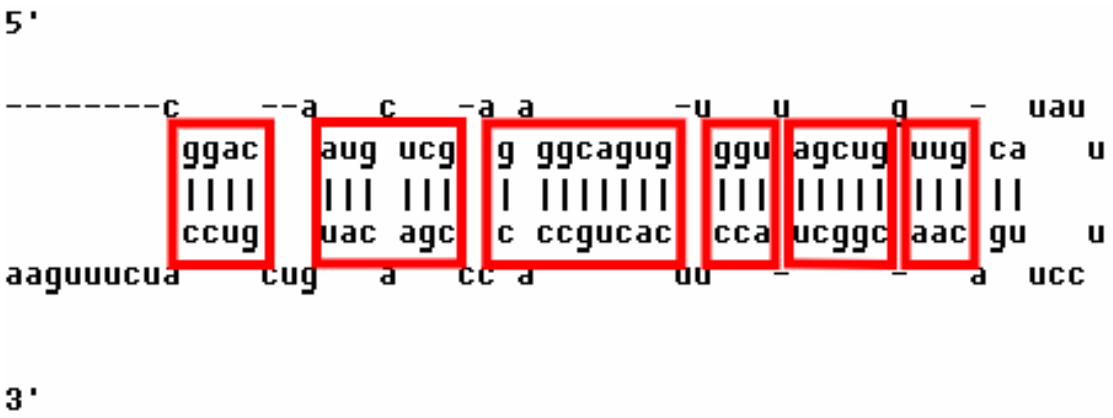
SCOR (http://scor.lbl.gov/) [Tamura et al, 2004] 提供核醣核酸的三維結構、功能以及分子的相互作用，偏向區域性（local）的資料庫。

The RNase P Database (http://jwbrown.mbio.ncsu.edu/RNaseP/home.html) [Brown, 1999] 則包含了 Ribonuclease P 家族序列的資訊。

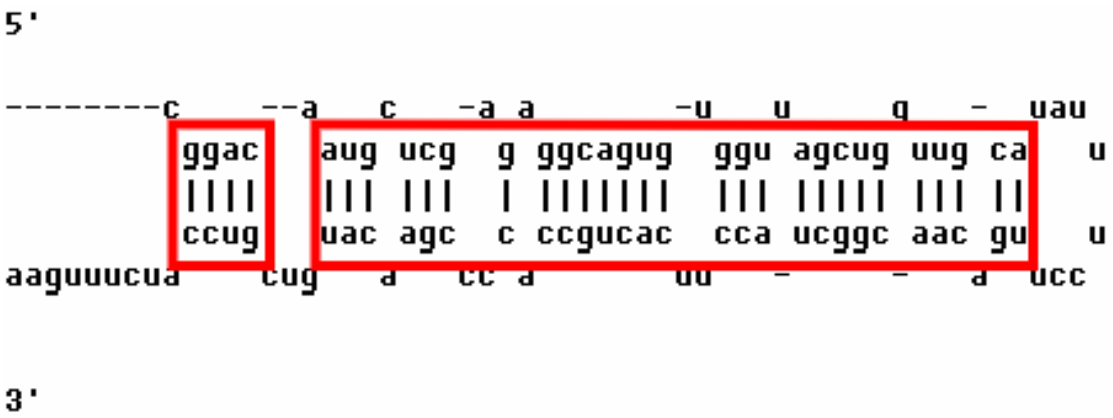
第三章 研究方法

3.1 系統設計目的

不使用任何領域知識的 GPRM 可以尋找結構穩定的區域性共同結構元，但對於一個包含長度較小的非對稱環狀結構的莖幹，無法將其視為一個莖幹，而會分割成兩個、甚至更多莖幹(如圖五所示)。如此一來，將大大增加問題的複雜度，需要花費許多額外的時間來處理。為了解決此問題，本研究設計一個更具有包容性的描述語言，並將能量的資訊納入考慮，以減少搜尋空間，加速系統的運作以處理更長更多的序列。



(a)舊的 GRPM 二級結構元表述法會因非對稱環狀結構而分割莖幹



(b)若能不受小的內部環狀結構影響，可以大幅減低複雜度

圖五 C.elegans miR-34 的髮夾結構

本研究設計一套新的系統Fold^{2}GP（Fold to Genetic Programming），採用眾所公認的Mfold系統為前處理器，自單一核醣核酸二級結構的觀點，來尋找共同的二級結構元。但鑑於Mfold仍有許多不足之處，我們提高Mfold中允許自由能提升的容忍量，採用能容許大量雜訊（nosie）的基因規劃法，以期能在廣大的搜尋空間中，尋找出家族核醣核酸的共同結構元。

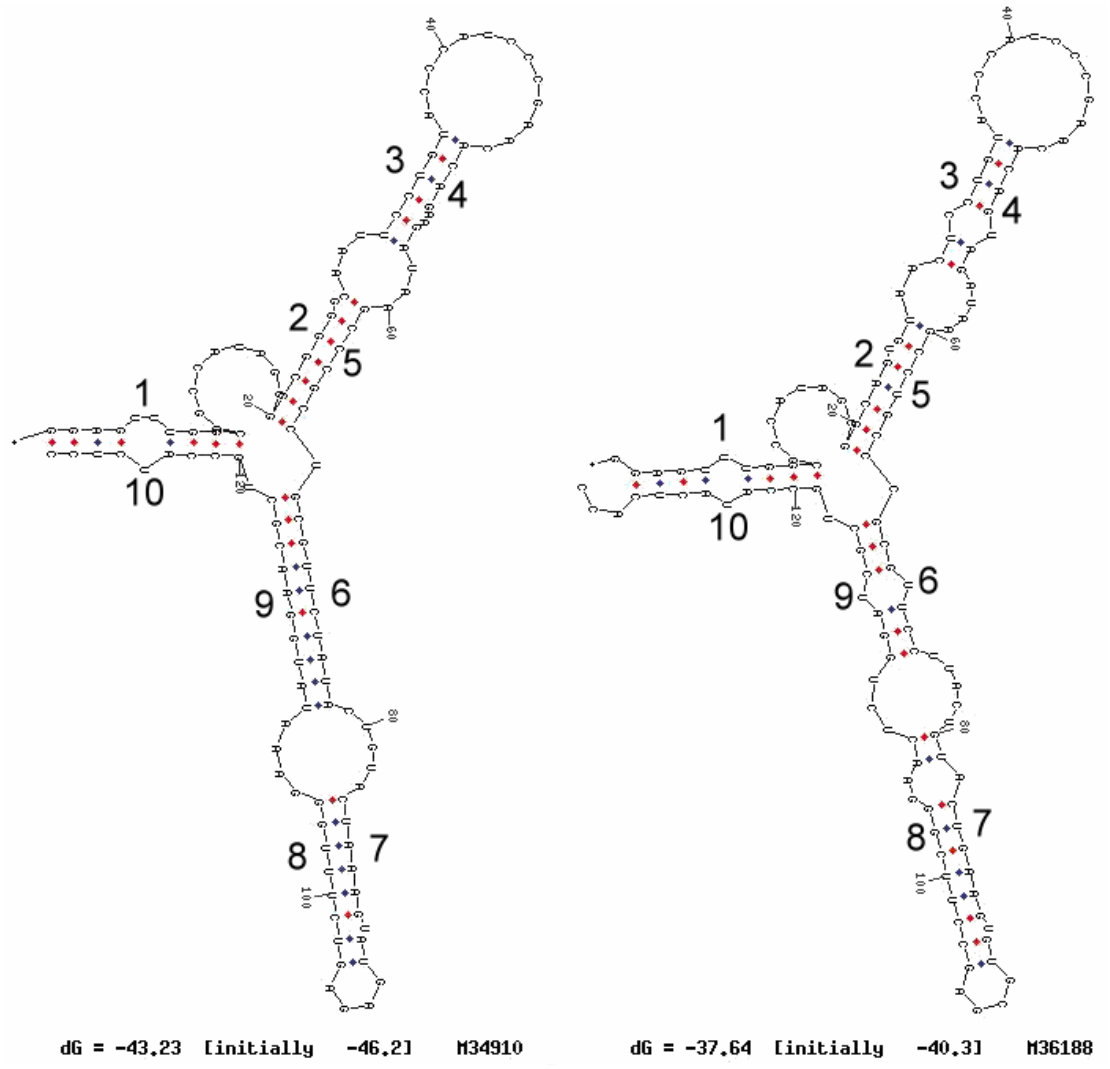
3.2 核醣核酸描述語言

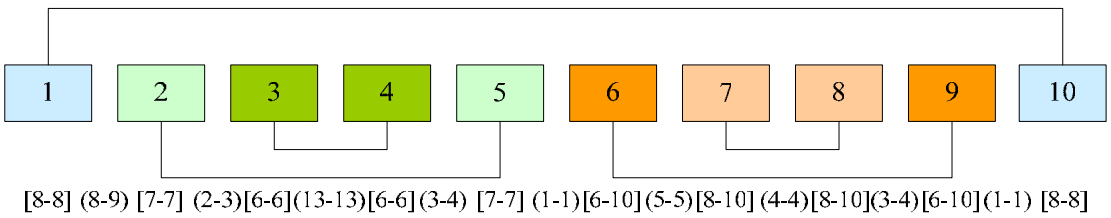
本研究針對二級結構設計一套方便基因規劃法演化的描述語言，原則上承襲GPRM的型態，但所包含的意義略有不同。語言的定義如下：

1. 分為結構拓樸及結構長度範圍。
2. 結構拓樸中的莖幹辨別是根據Mfold計算而得到的，不單純只是G≡C、A＝U、G－U三種鹼基配對。
3. 一個莖幹中可包含特定長度以下的內部環狀結構或突起結構。這個數目可以讓使用者自由定義，本實驗預設值為三個鹼基。
4. 莖幹結構的每股與每個環狀結構分別給予一個長度範圍，表示該結構元所涵蓋的結構大小。
5. 內部環狀結構的長度不包含在莖幹長度中。

透過上述定義，任何的二級結構能輕易的轉換為描述語言。下頁圖六中我們展示了以描述語言表示兩個序列的共同結構。描述語言的上半部為結構拓樸的示意圖，下半部括號裡表示上述定義中的第四點，結構長度的範圍，其中方括號代表莖幹結構的長度範圍、小括號則代表環狀結構的長度範圍。根據上述定義的第三點，當設定為允許莖幹包含三個以下的內部環狀結構，第一股與第十股（如圖中編號）形成長度為八個鹼基的莖幹，忽略內部環狀結構。

當序列所形成的結構中，有子結構與某結構元有相同的拓樸結構，並且每個莖幹與環狀結構的長度都在結構元允許的結構長度範圍內，就稱此序列擁有該結構元。



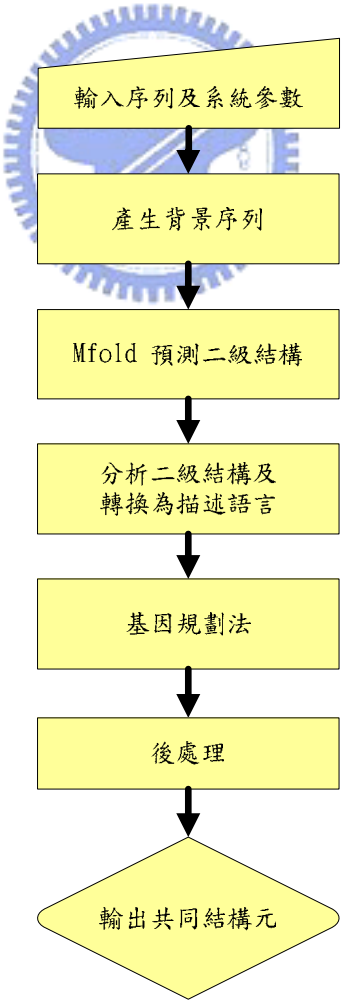


圖六. 以描述語言表示兩個富含非對稱內部環狀結構的 5SrRNA。

3.3 系統架構

本系統主要包含三大部分：系統前處理、預測及轉換二級結構，與主要的核心基因規劃法。

系統前處理包含兩個步驟，首先是處理系統參數。本系統可供使用者自行定義系統參數，包含設定莖幹及環狀結構的範圍，以及莖幹內所允許的內部環狀結構大小等等。完全使用系統預設的參數，也能達到一定效果。若使用者對於欲預測的資料有所了解，能提供更嚴謹的系統參數，還可大幅增進預測的正確性。此外，基因規劃法有本身系統執行時所需的參數，例如族群數量（population size）、與各運算子發生的機會。



圖七. 系統流程圖

第二步驟則是產生負面的背景資料（negative data set）。本研究根據研究假設，採用監督學習（supervised learning）來加強系統的效能。背景資料的數量為輸入序列數的倍數，使用者可以自由調整。一般而言，當輸入資料的序列數很少時，我們會建議使用較大的倍數來加強背景資料的效果。背景資料的序列長度根據各個輸入序列的長度而定。比方說，當背景資訊的序列倍數為三時，第一到三條的長度會取輸入序列第一條的長度；第四到六條則取輸入序列第二條的長度，依此類推。另外，本系統採用一級（first-order）序列產生法，也就是每個鹼基被挑選的機會是根據前一個鹼基來決定。

完成背景資料後，連同欲預測共同結構元的輸入序列，輸入至 Mfold 來預測二級結構。Mfold 對於每條序列通常會有多個預測結果（大部分為十幾個），這些結果稱為本系統的“候選結構”。每個候選結構包含以下資訊：根據能量排序的排名、估計的能量、以及各鹼基的配對資訊。

取得候選結構後，根據下列三點將配對資訊轉為結構描述語言：

1. 將連續的鹼基配對視為一個莖幹。
2. 允許特定大小的內部環狀結構及突起結構(預設為三個鹼基)。
3. 根據系統參數，過濾掉過小的莖幹。

至此，我們得到許多以描述語言表示的結構。這些結構包含輸入序列及背景序列的所有候選結構。

3.4 基因規劃法

基因規劃法源自基因演算法（genetic algorithm, GA），為 Koza 在 1992 年所提出的方法[Koza, 1992]。同樣都是模擬達爾文所提出“物競天擇；適者生存”的概念，自一個隨機產生的初代個體中，透過突變、互交、及複製等演化運算，逐步演化出適應度高的個體。

與基因演算法最大的不同在於，基因規劃法將個體直接轉換為樹狀結構，稱為分析樹（parse tree），而不需要編碼為基因演算法所使用的固定長度二元字串。不過，鑑於使用樹狀結構會耗費的大量運算時間，以及考量核醣核酸共同結構元的複雜度，本研究並不採用傳統的樹狀結構，而直接使用結構描述語言來模擬樹狀結構。

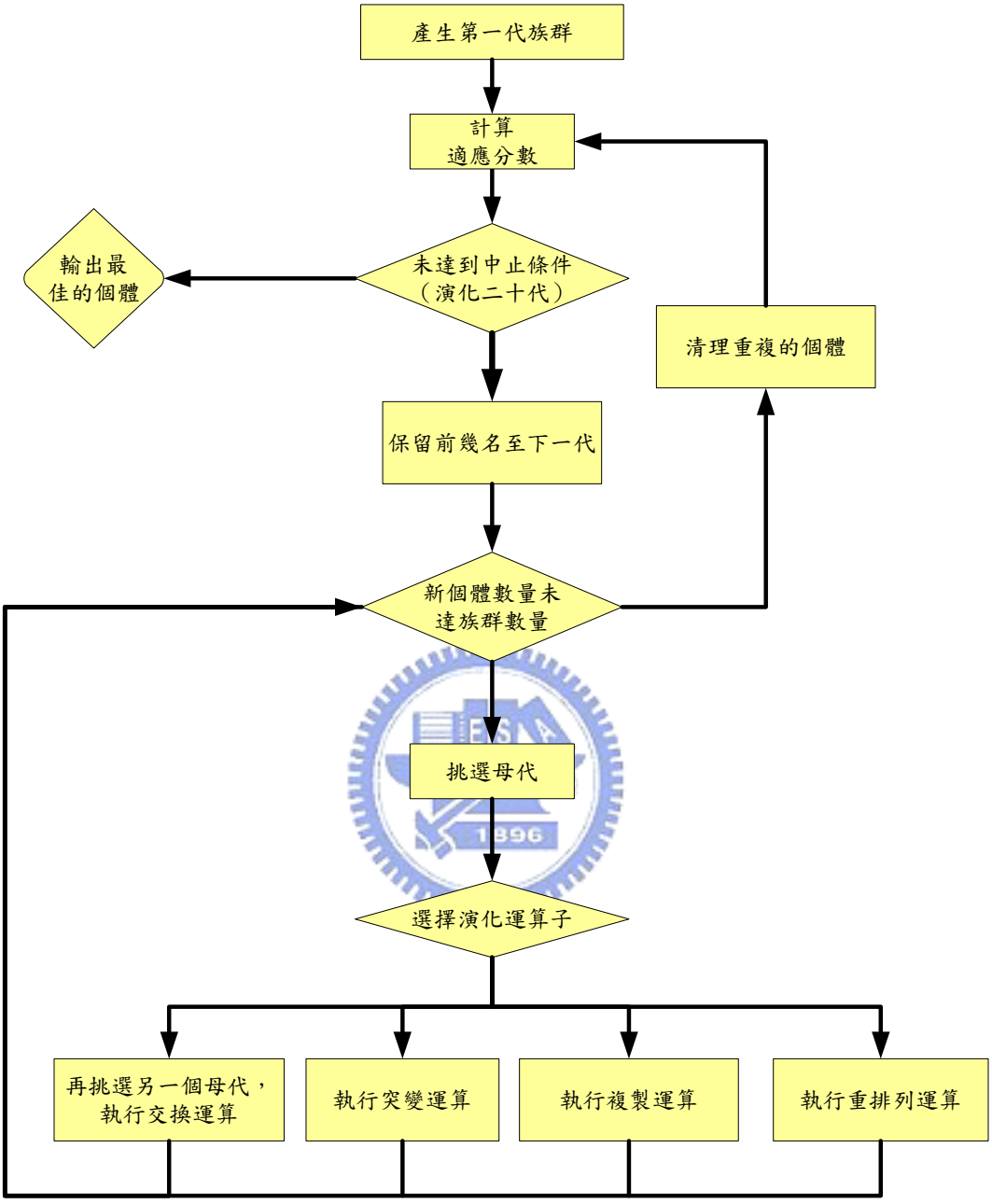
基因規劃法分為五個部份：

1. 產生初代個體（initial population）
2. 適應函數(fitness function)
3. 母代挑選機制(selection)
4. 演化運算子(genetic operators)
5. 終止條件(termination criterion)

以及最後的後處理(post process)。

系統將二級結構資訊轉為結構描述語言後，根據這些資訊產生初代個體，並一一計算其適應分數。當全部個體適應分數計算完畢後，再透過挑選機制選出適當的母代，而後由演化運算產生新的個體，直到新個體數目到達族群數量。如此反覆製造新一代的個體，直到滿足終止條件。

以下各小節將詳述各部份的實作策略。

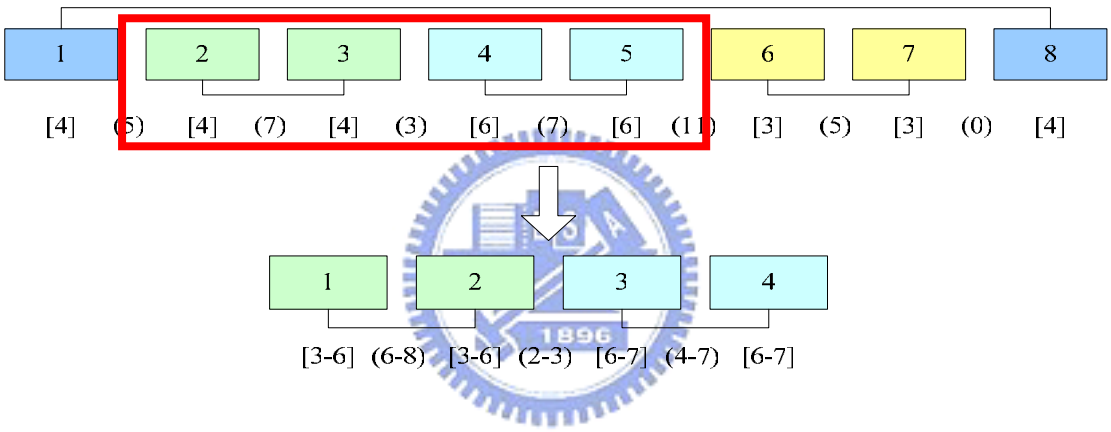

產生第一代族群
計算
適應分數
輸出最
佳的個體
未達到中止條件
（演化二十代）
清理重複的個體
保留前幾名至下一代
新個體數量未
達族群數量
挑選母代
選擇演化運算子
再挑選另一個母代，
執行交換運算
執行突變運算
執行複製運算
執行重排列運算

圖八 基因規劃法流程圖。

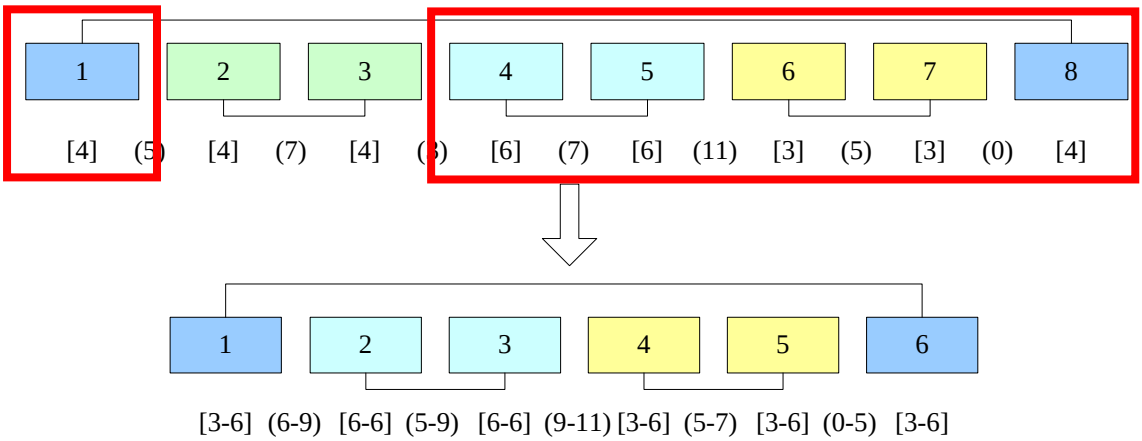
3.4.1 產生初代個體

共同結構元意指會出現在各序列上的相似結構，因此，本研究根據下面三個要點，隨機擷取輸入序列的子結構來當作第一代。

1. 每個輸入序列會產生相同數量的子結構，順位越高（意即結構越穩定）的候選結構會產生越多的初代個體。
2. 自每個候選結構中，擷取給定數量的不同子結構。該子結構的莖幹數目不會超過所有候選結構的平均數。
3. 將子結構的每個長度隨機往外拓展，形成一長度區間。



(a)對於原本四個莖幹的候選結構，擷取其第二、三個莖幹，並將結構長度往外拓展，當作一個初代個體



(b)同一候選結構，但選取到的莖幹為第一、三、四個，形成另一個新個體

圖九 一個候選結構可能會產生數個子結構，成為初代個體。

上頁圖九表達了兩個初代個體自同一候選結構產生的可能狀況。圖(a)的個體擷取了第二與第三個莖幹，並將各個結構的長度範圍往外拓展一些，形成了一個新的個體。圖(b)中則取到了第一、三、四個莖幹。

3.4.2 適應函數

本研究所定義適應函數包含兩個部份： GPRM 所使用的 F-score 以及結構長度。

F-sorce 包含兩個部份－正確率（presicion）與擷取率（recall），取其調和平均數，以確保兼顧到兩個數值。不單純僅使用傳統的算術平均數而改用 F-sorce，是因為我們預期適應分數高的個體其正確率與擷取率分數也應該都是高的。而傳統的算術平均數很容易因為某一個值過高而拉高整體的分數，產生高估或錯估的情況。F-sorce 定義如下：

$$precision\ (I) := \frac{M}{M+N}\ ,\ recall(I) := \frac{M}{C} \qquad (1)$$

$$F(I) = \frac{1}{\frac{1}{2}\left[\frac{1}{recall(I)} + \frac{1}{precision(I)}\right]} = \frac{2M}{C+M+N} \qquad (2)$$

其中 M 代表輸入序列中，包含結構元 I 的個數； N 代表背景序列包含結構元 I 的個數； C 為輸入序列的總個數。

然而，背景資料對於大的共同結構元，影響力稍嫌不足。在 F-score 相近的情況下，我們會傾向選取結構較大且較完整的個體。因此本研究加入輔助的結構長度，將適應函數定義為：

$$f(I)=\alpha\times F(I)+(1-\alpha)\times S(I)$$

其中 $S(I)$代表結構元 I 符合各序列的平均鹼基數除以該代所有個體的最大平均鹼基數。如此一來，$S(I)$ 的值會被正規化在零到一之間。α 為一個介於零到一的權重，當 α 越大時，系統會偏好被更多輸入序列包含的結構元，但可能受雜訊影響，只能找到共同結構元的子結構。一般而言，我們設定 α 為 0.6，可以滿足絕大部份的資料。

3.4.3 母代挑選機制

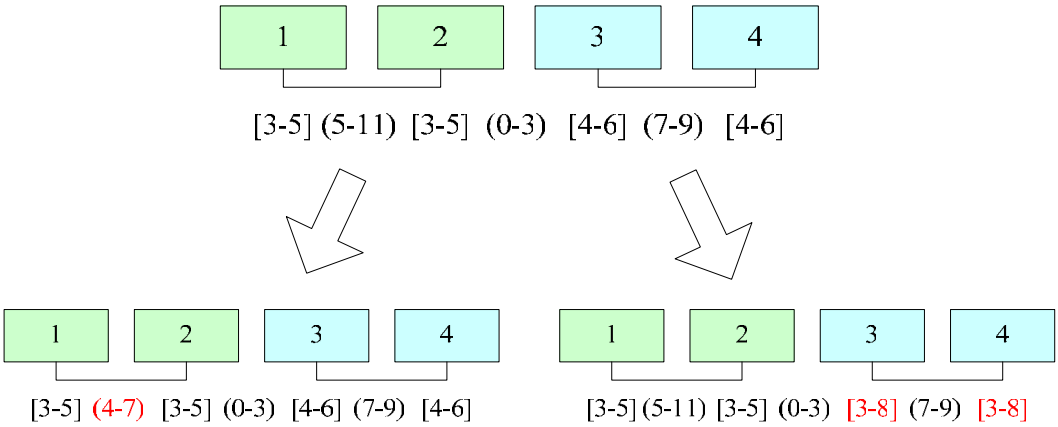
產生新一代個體時，我們先保留一定比例適應分數高的個體至下一代。剩餘的子代個體則透過挑選機制選出母代，執行演化運算後產生新一代的個體。本系統採用的挑選機制為競賽法(tournament)：自母代群體中挑選一定數量的個體，個體之間彼此比較適應分數，其中分數最高者被挑選出來產生子代。

3.4.4 演化運算子

以下介紹本研究所使用到的各演化運算子：

點突變

選擇一個莖幹或環狀結構，給予一個全新的、符合系統參數限制的範圍。下頁圖十(a)為挑選到第一個環狀結構的突變結果。若選取到莖幹結構，則該莖幹的兩股都要同時改變，如圖十(b)所示，挑選到的是第二個莖幹，改變其結構長度。點突變運算為本系統最基本的演化運算。

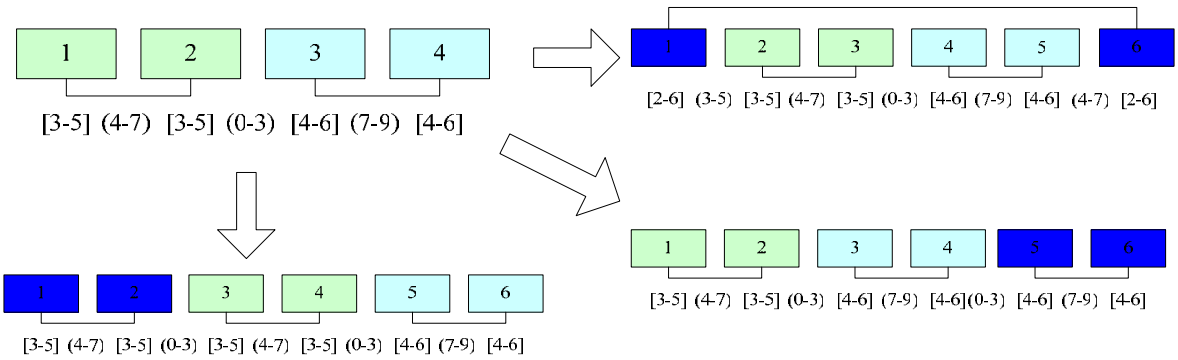


(b)突變環狀結構的長度。　　(b)突變莖幹的長度。

圖十 點突變運算示意圖：紅色表示突變點

結構突變

隨機製造一個新的莖幹結構，連接在被挑選的結構前、後、或是將其夾在中間。這樣做的目的在於，原本透過挑選機制選到的個體，其表示的結構很可能僅是共同結構元的子結構。在此我們並不刪除莖幹，因為刪除莖幹會使大型結構化為小結構，而小結構本身已經很容易存在，不需要多此一舉。

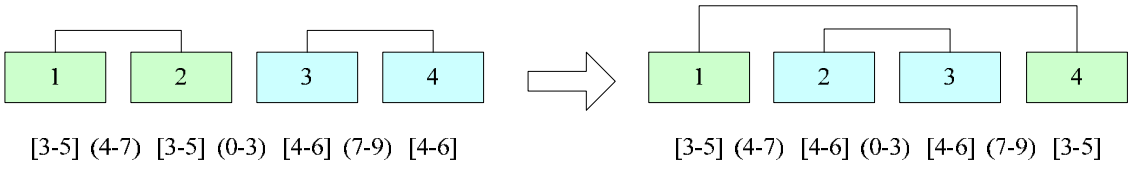


圖十一 結構突變運算示意圖：因突變新增的結構以深色表示

結構重排

將莖幹的相對位置全部打散重排，結構長度也隨之移動，環狀結構的長度則保持原來的位置不變。結構重排的目的在創造出新的結構。當莖幹數量大時，有

限的族群數量很難包含所有的結構拓樸。透過結構的重新排列，將有機會產生出全新的個體。



圖十二 結構重排運算示意圖：原本並排的兩個莖幹被重排為一個包含大內部環線的莖幹

互交

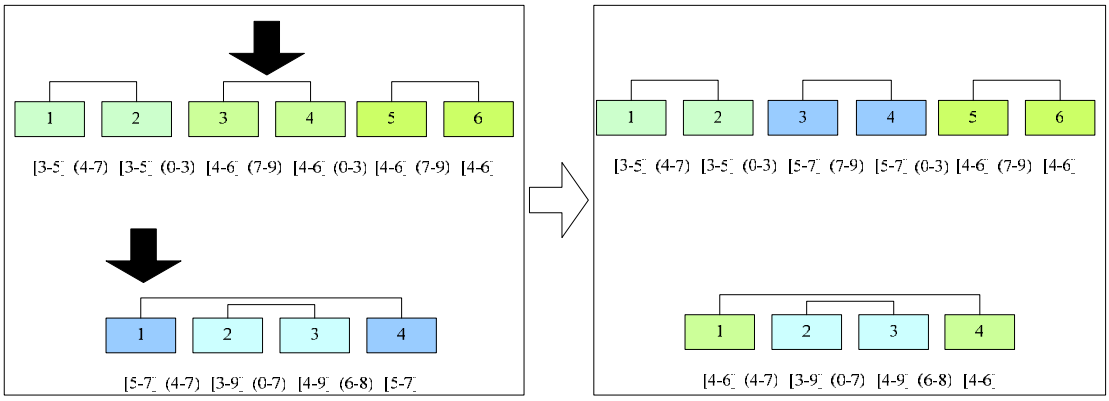
互交運算不同於其它運算僅挑選一個母代個體，而是挑選兩個母代個體出來。然後，隨機選取其中一個母代個體中的一個莖幹結構、與由另一母代個體隨機選出的一至三個莖幹結構進行交換。

互交不僅讓結構拓樸改變，莖幹數也隨之增減，當被挑到莖幹恰好是共同結構元上的不同子結構時，透過互交運算將兩者結合起來，使得跳脫區域最佳解（local optimum）的機會大增。

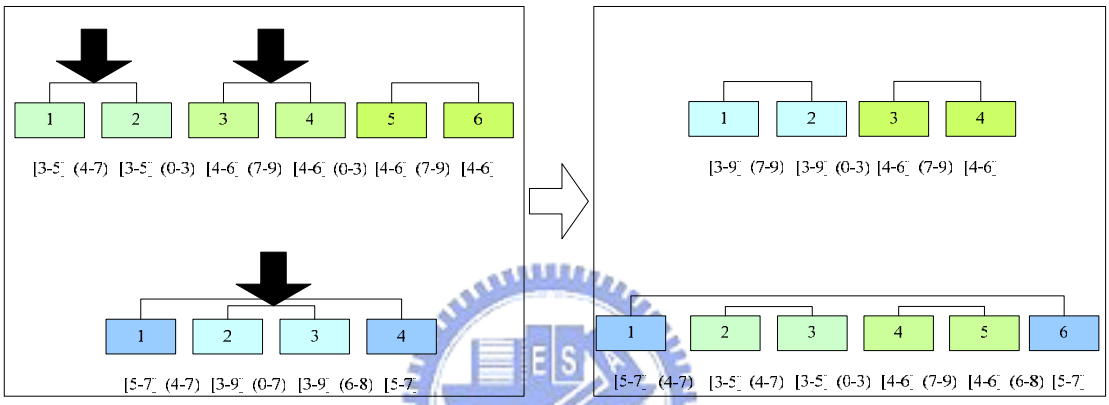
下頁圖十三說明了兩種不同的互交結果。圖(a)中，選取了上方個體的第二個莖幹，與下方個體的第一個莖幹，將其互換，分別成為兩個結構長度被改變的新個體。圖(b)則選取了上方個體的第一、二個莖幹，與下方個體的第一個莖幹，交換後不僅結構長度改變，結構拓撲甚至莖幹數目也都改變了。

重製

除了適應分數高的個體，我們也保留了一些機會讓不是那麼優秀的母代能夠完整保留下來，以增加族群的多樣性。



（a）各選擇一個莖幹互換，不會改變莖幹數量



（b）上面的序列選取其中兩個莖幹，與下面的一個莖幹交換，會大大改變結構拓樸

圖十三 互交運算示意圖

清除重複個體

當演化到一定程度時，可能會往某幾個區域最佳解收斂，導致整個族群充斥著特定的個體。當族群變異度太小、缺乏多樣性時，很容易導致演化結果圍繞在這些區域最佳解。因此，我們在每代演化完成後，把幾乎類似的個體刪除，並根據初代個體產生的方式產生新個體，彌補因個體刪除後族群大小缺少的部份。

3.4.5 終止條件

通常在二十代內，系統就能找到不錯的答案。因此本系統為確保能在一定時間完成，將終止條件設定為演化二十代。

3.4.6 後處理

我們透過基因規劃法找到輸入序列的共同結構元後，將每條序列上所發生的位置標示出來後輸出給使用者。

然而，我們觀察到，當內部環狀結構的大小遊走在門檻值時，類似的結構可能會被解析成不同的拓樸結構。例如，當輸入序列的共同結構元為一個包含長度三到四的內部環狀結構的莖幹，倘若我們將內部環狀結構的容忍值設定為三個鹼基，長度為四的序列將會被分成兩個莖幹；長度為三的則被合併為一個，目前我們的描述語言無法明確的表述此種狀況。

為此，我們以找到的共同結構元，再次檢查那些沒有發現結構的序列，並允許其中一個莖幹被合併或是被分割，找出因上述原因遺失的序列。

第四章 實驗結果

4.1 實驗評估方式

對於預測核醣核酸二級結構的評估有兩大重點：莖幹層級與鹼基層級的效果。莖幹層級意指二級結構的外貌，如轉移核醣核酸的苜蓿葉結構，成十字架狀的型態。由於本系統放寬了許多限制，十分具有彈性，對於莖幹層級都能確實的預測出來，因此本系統將使用鹼基層級的預測評估方式。

目前許多二級結構預測軟體使用化簡後的 Matthews 相關係數（Matthews correlation coefficient)來評估一個二級結構預測系統的效能[Ji et al, 2004; Hu, 2002; Gorodkin, 2001]，因此，本研究也使用 Matthews 相關係數來作為評估的標準。其原始定義如下：

$$C = \frac{P_t N_t - P_f N_f}{\sqrt{(N_t + N_f)(N_t + P_f)(P_t + N_f)(P_t + P_f)}} \qquad (3)$$

其中 P_t 為正確正預測的總鹼基對數（true positive)；P_f 為錯誤正預測的總鹼基對數（false positive)；N_t 為正確負預測的總鹼基對數（true negative)。N_f 為錯誤的負預測的總鹼基對數（false negative)。

化簡後的式子如下：

$$C \approx \sqrt{\frac{P_t}{P_t + N_f}\frac{P_t}{P_t + P_f}} \qquad (4)$$

在本系統中，正確的正預測代表預測結構中所包含的鹼基對確實是正確的鹼基對；預測結構中所包含的鹼基對不存在正確結構中的話，稱為錯誤的正預測；存在於正確結構，卻沒有被系統發現的鹼基對，稱為錯誤的負預測。

4.2 測試資料

本研究使用五個核醣核酸家族作為實驗的測試資料，包含 GPRM 所使用過的三個測試資料，與另兩個結構元包含許多內部環狀結構的核醣核酸家族。簡介如下：

Archaea 16S rRNA

這組資料僅有 34 條序列，是本研究測試資料中最小的一組。其序列的平均長度為 97 個鹼基，共同結構元包含三個莖幹結構，都不包含非對稱的環狀結構，屬於比較單純的二級結構。

Phe-tRNA

Phe-tRNA 為攜帶胺基酸 Phe 的轉移核醣核酸，包含 74 條序列。其共同結構元幾乎涵蓋序列，形狀看起來像四瓣的苜蓿芽，包含四個莖幹結構。這個資料的特點是並非所有的序列都擁有完整的結構元，也就是說資料本身就包含些許雜訊，可以來測試本系統對於雜訊的處理情況。

IRE-like

IRE 的全名是 iron response element，其特色為其共同結構元在序列上與結構上都十分一致，為長度十到十三個鹼基，包含一個突起結構的莖幹結構，使得序列層級的分析也能很快的找到。因此，在 GPRM 的研究中，將序列做些許改變，以增加序列排比時的複雜度。此資料的平均長度為兩百個鹼基，用來觀察本系統對於共同結構元遠比序列小時的處理情況

C.elegans microRNA precursor

這份資料是從“The miRNA Registry”取得，屬於 Caenorhabditis elegans 所有的微核醣核酸，共 116 條序列。微核醣核酸先質的共同結構元是一個莖幹長

度很長的髮夾結構，內部包含大量對稱及非對稱環狀結構與突起結構，十分適合來檢驗本系統處理這些結構的能力。

5SrRNA

5SrRNA 完整的資料來自“Rfam”資料庫，本研究取其“seed”的部份一共 602 條序列。這組測試資料的特色是結構元包含五個莖幹結構，並且一樣含有凌亂的內部環狀以及突起結構。這份資料除了能測試本系統對於內部環狀結構外，大量的序列也是一項新的嘗試。不過，由於我們無法找到所有序列的正確答案，在評估實驗結果時，只會取其中有正確答案的 269 條序列來評估。

資料名稱	序列數量	平均長度	結構元莖幹數目	Annotation
16SrRNA	34	97	3	• 結構單純 • 無非對稱內部環線
tRNA	74	68	4	• 結構元莖幹數多 • 含少量非對稱內部環線 • 並非所有序列都擁有結構元
IRE-like	56	202	1	• 結構元相對於序列較小
C.ele. microRNA	116	98	1	• 僅有一個莖幹 • 含大量內部環線
5SrRNA	602 (269)	119	5	• 序列數量多 • 結構元莖幹數多 • 含大量內部環線

表一 測試資料一覽表

4.3 實驗結果

本系統是以C語言實作，測試環境的作業系統是Mandrake Linux 10.1，電腦配備為Pentium IV 3.2G Hz的中央處理器與2 Giga-bytes的記憶體。

所有實驗基因規劃法的參數：族群數量10000、長度突變率60%、結構突變率20%、結構重排率10%、互交率5%以及重製率5%。此外，Mfold所使用的次最佳解的比例設為35%、背景資料的個數為輸入序列的兩倍。結構限制部份，若沒有特別說明，採取十分寬鬆的限制：所允許的莖幹結構長度介於三到二十之間，並且限制單一莖幹的最大與最小值差距在十五以內；允許環狀結構的長度介於零到二十，限制單一環狀結構的最大最小值差距在十五以內，並容許莖幹內包含三個鹼基以下的內部環狀結構或突起結構。

實驗數據是根據上百次的實驗結果來計算，每次的實驗都會設定不同的隨機種子（random seed）以確保每次實驗不會重複。

4.3.1 與GPRM的結果比較

表二整理了本系統對於IRE-like, archaea 16S rRNA, tRNA三種資料的Matthews相關係數及所花費的平均時間，以及GPRM的實驗結果。

Dataset	Runtime	Fold^{2}GP	GPRM
16SrRNA	3 min	.93	.94
tRNA	12 min	.84	.64
IRE-like	25 min	.44	.91

表二 實驗結果一

對於共同結構元單純的 16S rRNA，本系統與 GPRM 的表現都不錯。對於較共同結構元複雜、包含雜訊的 tRNA，GPRM 就處理的比較差些，而本系統則有不錯的表現。在此我們發現到系統所找到結構元，僅有其中的六十三條序列所擁有，另外十一條的序列，除了有三條序列本身就沒有符合的結構元外，其他八條都是 Mfold 無法預測出正確結構所致。

然而，對於序列長度的平均長度為兩百個鹼基，而其共同結構元卻僅有十幾個鹼基的 IRE-like，本系統使用預設的寬鬆參數，無法找到正確的結構元。而在我們參考 GPRM 所使用的參數後，雖然能夠找到正確的結構元，但因為 Mfold 有超過二十條序列無法將該區域正確的配對，導致 Matthews 相關係數僅有四成。

4.3.2 富含突起及內部環狀結構資料的實驗結果

表三列舉 C. elegans microRNA 與 5S rRNA 兩組資料的實驗結果。

Dataset	Runtime	Fold^{2}GP	GPRM
C.ele. microRNA	17 min	.84	< .6
5SrRNA	110 min for 602 seqs 45 min for 269 seqs	.74 (269 seqs)	< .3 (269 seqs)

表三 實驗結果二

C.elegans microRNA 的共同結構元僅有一個莖幹，不過其長度約為三十多個鹼基，並包含許多大小不一、位置不定的突起與內部環狀結構，屬於 GPRM 處理不佳的資料。由於的結構元的莖幹較大，因此我們將莖幹結構的限制放大到三到三十五個鹼基，並且容許單一莖幹的最大最小值差距為二十，將莖幹內容許為配對的鹼基調整為四。對於如此大的莖幹結構，本系統依然能找出正確的結構元，並且在絕大部分的序列上標示出正確的位置。

最後，對於結構元複雜、包含突起與內部環狀結構以及序列數量很大的 5S rRNA，本系統依然能找到五個莖幹的正確結構元，而且花費時間在兩個小時內。我們無法找到所有序列的正確結構，僅自“5S Ribosomal RNA database”收集其中 269 條擁有結構資訊的序列，因此這裡的 Matthews c.c. 僅計算此 269 條序列。

第五章 結論與未來研究方向

5.1 結論

本研究嘗試利用 Mfold 提供能量資訊，使用基因規劃法來尋找一組具有相同功能核醣核酸的共同結構元，而且不依賴序列排比，直接以二級結構來作為搜尋的目標。

相較於 GRPM，本系統加入能量資訊，改良結構描述語言後，確實能解決突起結構與非對稱內部環狀結構所產生的問題，並且因為搜尋空間的縮減，可以處理複雜更結構、數量更多的核醣核酸序列。此外，本系統將基因規劃法的族群數量拉大到一萬個個體，因此不需定義莖幹數量、可以使用十分寬鬆的結構參數來處理大部分的資料。

由實驗結果可看出，基因規劃法確實能夠避開大量雜訊的干擾，並且不會因為序列數量的增加而花費太多時間。這表示選用基因規劃法來協助我們處理 Mfold 不佳的輸出結果，確實是可行的。

5.2 未來研究方向

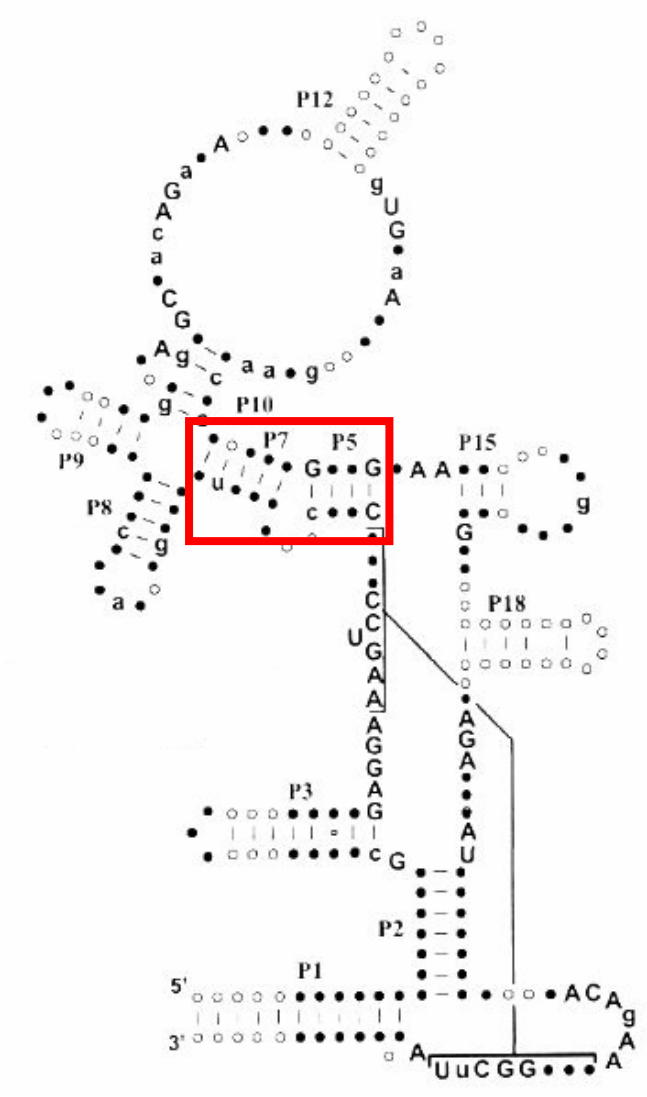
在此提出研究過程中所遭遇的問題，以供相關研究參考。

5.2.1 描述語言

改良後的描述語言雖然成功的解決解決突起結構與非對稱內部環狀結構的問題，但是仍有一種狀況無法處理。語言的特點可以連續、區域性的連續莖幹所形成的結構元，但是無法只描述莖幹的兩股在序列上相距太長的情況。

如下頁圖十四所示，這是 RNase P 家族的結構示意圖，我們假設它的共同結構元是方框所圍起來的莖幹（P5, P7）。由於該莖幹後面還有很長的一段序列，並且成疊成數個莖幹結構（P8, P9, P10, P12），以目前所定義的結構描述語言，無法單獨表示 P5, P7 所形成的莖幹，必須包含把此莖幹兩股間序列所形成的結構。

對此，本研究曾嘗試設計解決方法，但礙於能量資訊無法提供精確的結構，使得我們無法達到預期的目標。如何尋找一個方法表述該結構，又不會因為過於彈性而失去正確性，是今後研究的一大課題。



圖十四 對於方框內的莖幹結構，描述語言無法單獨表示

5.2.2 前處理器

本研究使用 Mfold 來當前處理器，希望透過能量的資訊可以過濾掉一些自然界不可能形成的結構，但任何系統都不可能完美，即使我們試著將 Mfold 輸出的候選結構數拉大，也無法包含所有正確的結構。

此外，Mfold 最大的缺點是無法預測擬結結構，這使得我們所提出的新系統必須犧牲掉所有包含擬結結構的測試資料。倘若能結合更多的二級結構預測系統，並能分辨出其中各系統所預測出候選結構的優劣，整合起來做為前處理器，

或許是解決本問題的一個方法。

5.2.3 背景資料與適應函數

對於共同結構元大的核醣核酸家族，以目前方式所產生的背景資料效果不佳。例如，預測 5S rRNA 時，我們發現共同結構元中的其中四個莖幹結構就已經很不容易出現在背景資料的序列中，使得找到的結構可能變的不完整。

為了解決背景資料對於大共同結構元效果太弱的問題，我們在適應函數中加入長度的資訊，但這產生了新的問題。對於很長的核醣核酸序列，如果 Mfold 的效果不夠好，適應分數可能會被少數幾個長相類似的序列拉高。如何在此兩者間取得真正的平衡，也是一項需要改進的地方。

第六章 參考文獻

Ambros V. “microRNAs: tiny regulators with great potential.” Cell. 2001 Dec 28;107(7):823-6. Review.

Brown JW. “The Ribonuclease P Database.” Nucleic Acids Res. 1999 Jan 1;27(1):314.

Fera D, Kim N, Shiffeldrim N, Zorn J, Laserson U, Gan HH, Schlick T. “RAG: RNA-As-Graphs web resource.” BMC Bioinformatics. 2004 Jul 6;5(1):88.

Cary RB, Stormo GD. “Graph-theoretic approach to RNA modeling using comparative data.” Proc Int Conf Intell Syst Mol Biol. 1995;3:75-80.

Gorodkin J, Heyer LJ, Stormo GD. “Finding the most significant common sequence and structure motifs in a set of RNA sequences.” Nucleic Acids Res. 1997 Sep 15;25(18):3724-32.

Gorodkin J, Stricklin SL, Stormo GD. “Discovering common stem-loop motifs in unaligned RNA sequences.” Nucleic Acids Res. 2001 May 15;29(10):2135-44.

Griffiths-Jones S. “The microRNA Registry.” Nucleic Acids Res. 2004 Jan 1;32(Database issue):D109-11.

Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. “Rfam: annotating non-coding RNAs in complete genomes.” Nucleic Acids Res. 2005 Jan

1;33(Database issue):D121-4.

Hu YJ. “Prediction of consensus structural motifs in a family of coregulated RNA sequences.” Nucleic Acids Res. 2002 Sep 1;30(17):3886-93.

Ji Y, Xu X, Stormo GD. “A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences.” Bioinformatics. 2004 Jul 10;20(10):1591-602. Epub 2004 Feb 12.

John R. Koza. “Genetic Programming: On the Programming of Computers by Means of Natural Selection.” MIT Press, 1992.

Lai EC, Tomancak P, Williams RW, Rubin GM. “Computational identification of Drosophila microRNA genes.” Genome Biol. 2003;4(7):R42. Epub 2003 Jun 30.

Lee RC, Feinbaum RL, Ambros V. “The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.” Cell. 1993 Dec 3;75(5):843-54.

Lewis, B. “Genes VIII.” Oxford, 2003.

Mathews DH, Sabina J, Zuker M, Turner DH. “Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure.” J Mol Biol. 1999 May 21;288(5):911-40.

Pley HW, Flaherty KM, McKay DB “Three-dimensional structure of a hammerhead

ribozyme.” 1994 Nov 3;372(6501):68-74

Scott WG, Finch JT, Klug A “The crystal structure of an all-RNA hammerhead ribozyme: a proposed mechanism for RNA catalytic cleavage.” Cell. 1995 Jun 30;81(7):991-1002

Punginelli C, Ize B, Stanley NR, Stewart V, Sawers G, Berks BC, Palmer T. “mRNA secondary structure modulates translation of Tat-dependent formate dehydrogenase N.” J Bacteriol. 2004 Sep;186(18):6311-5.

Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G. “The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans.” Nature. 2000 Feb 24;403(6772):901-6

Rhoades MW, Reinhart BJ, Lim LP, Burge CB, Bartel B, Bartel DP. “Prediction of plant microRNA targets.” Cell. 2002 Aug 23;110(4):513-20.

Sankoff,D. “Simultaneous solution of the RNA folding, alignment and protosequence problems.” SIAM J. Appl. Math. 1985; 45:810-25.

Sprinzl M, Steegborn C, Hubel F, Steinberg S. “Compilation of tRNA sequences and sequences of tRNA genes.” Nucleic Acids Res. 1996 Jan 1;24(1):68-72.

Sprinzl M, Horn C, Brown M, Ioudovitch A, Steinberg S. “Compilation of tRNA sequences and sequences of tRNA genes.” Nucleic Acids Res. 1998 Jan 1;26(1):148-53.

Szymanski M, Barciszewska MZ, Erdmann VA, Barciszewski J. “5S Ribosomal RNA Database.” Nucleic Acids Res. 2002 Jan 1;30(1):176-8.

Tamura M, Hendrix DK, Klosterman PS, Schimmelman NR, Brenner SE, Holbrook SR. “SCOR: Structural Classification of RNA, version 2.0.” Nucleic Acids Res. 2004 Jan 1;32(Database issue):D182-4.

Thiviyanathan V, Yang Y, Kaluarachchi K, Rijnbrand R, Gorenstein DG, Lemon SM. “High-resolution structure of a picornaviral internal cis-acting RNA replication element (cre).” Proc Natl Acad Sci U S A. 2004 Aug 24;101(34):12688-93. Epub 2004 Aug 16.

Thompson JD, Higgins DG, Gibson TJ. Related Articles, Links “CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.” Nucleic Acids Res. 1994 Nov 11;22(22):4673-80.

Tinoco I Jr, Uhlenbeck OC, Levine MD. “Estimation of secondary structure in ribonucleic acids.” Nature. 1971 Apr 9;230(5293):362-7.

van Batenburg FH, Gultyaev AP, Pleij CW, Ng J, Oliehoek J. “PseudoBase: a database with RNA pseudoknots.” Nucleic Acids Res. 2000 Jan 1;28(1):201-4.

Watson JD, Crick FH. “The structure of DNA.” Cold Spring Harb Symp Quant Biol. 1953;18:123-31

Zuker M. “On finding all suboptimal foldings of an RNA molecule.” Science. 1989 Apr 7;244(4900):48-52.

Zuker M. “Mfold web server for nucleic acid folding and hybridization prediction.” Nucleic Acids Res. 2003 Jul 1;31(13):3406-15.