國立交通大學

資訊科學與工程研究所

碩 士 論 文

利 用 結 構 分 解 與 抽 象 處 理 搜 尋 核 糖 核 酸 結 構

Utilizing Structure Decomposition and Abstraction
to find RNA Structure

研 究 生：黃繼養

指導教授：胡毓志　教授

中 華 民 國　九 十 五　年 七 月

利用結構分解與抽象處理搜尋核糖核酸結構
Utilizing Structure Decomposition and Abstraction to find RNA Structure

研 究 生：黃繼養　　Student：Chi-Yang Huang

指導教授：胡毓志　　Advisor：Yuh-Jyh Hu

國 立 交 通 大 學

資 訊 科 學 與 工 程 研 究 所

碩 士 論 文

A Thesis

Submitted to Institute of Computer Science and Engineering

College of Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer Science

June 2007

Hsinchu, Taiwan, Republic of China

中華民國九十五年六月

利用結構分解與抽象處理

搜尋核糖核酸結構

研究生：黃繼養　　指導教授：胡毓志博士

國立交通大學資訊科學與工程研究所

摘要

近年來，在核醣核酸的相關研究上有許多新發現，研究發現核醣核酸有許多我們以往不清楚的功能，而這些功能在生物學上皆扮演著重要的角色，2006/10/02 公佈的諾貝爾醫學獎得獎的主題「RNAi (RNA interference, RNA 干擾現象)」就是最好的例子。由於核醣核酸的功能與其二級結構有密切的關係，因此若能提供核醣核酸二級結構的相關資訊給生物學家，則能協助他們加快檢驗出核醣核酸的功能。在本研究中，我們提供一個生物資訊的方法，自一群相關的核醣核酸序列中，找出其二級結構的共同結構元。我們以現有的單一核醣核酸二級結構預測系統作爲前處理器，預測出每條序列數個可能的二級結構，將其結果使用我們所設計的演算法做分解並抽象化之後，使用 Gibbs-like 流程方法來找出其共同結構元，並且爲了能提升系統的效能，設計了一些新的對核醣核酸結構做分析的演算法，它們的時間復雜度都比現今提出的演自法來得更有效率，使得我們的系統能在記憶體和執行時間上遠勝於其它的系統。爲了驗證我們系統的準確定與效能，我們從 Rfam 中下載數個 RNA 家族的資料來做測試，實驗結果也顯示出本方法有不錯的表現。

Utilizing Structure Decomposition and Abstraction to find RNA Structure

Student：Chi-yang Huang　　Advisor：Dr. Yuh-Jyh Hu

Institute of Computer and Information Science
National Chiao Tung University
Hsinchu, Taiwan, Republic of China

Abstract

Motivation: RNA molecules are the key players in the biochemistry of the cell, playing many important roles in regulation, catalysis and structural support. Many functional RNAs have evolutionarily conserved secondary structures in order to fulfill their roles in a cell. Although current approaches can identify common structure motifs from a set of RNAs, they typically rely on the assumption that the given sequences are from a single family, which is not necessarily true in practice.

Results: We develop a new method based on structure decomposition and Gibbs sampling to predict consensus structure motifs in unaligned RNA sequences. Unlike most current approaches, our method is applicable to a set of mixed sequences from different families, and is able to predict multiple motifs for multiple families. Furthermore, as we separate motif finding from sequence folding in our system, new folding algorithms other than Mfold or RNAfold, etc. can be easily integrated with our motif finding process. Extensive testing on 17 families from Rfam shows that our method competes well with other current tools in single family predictions. As for multi-family predictions, experiments also demonstrate that our new approach outperforms recent alternative methods.

國 立 交 通 大 學

博碩士論文全文電子檔著作權授權書

（提供授權人裝訂於紙本論文書名頁之次頁用）

本授權書所授權之學位論文，為本人於國立交通大學資訊科學與工程研究所 資訊 組， 96 學年度第 2 學期取得碩士學位之論文。

論文題目：利用結構分解與抽象處理搜尋核糖核酸結構
指導教授：胡毓志

■ 同意

本人茲將本著作，以非專屬、無償授權國立交通大學與台灣聯合大學系統圖書館：基於推動讀者間「資源共享、互惠合作」之理念，與回饋社會與學術研究之目的，國立交通大學及台灣聯合大學系統圖書館得不限地域、時間與次數，以紙本、光碟或數位化等各種方法收錄、重製與利用；於著作權法合理使用範圍內，讀者得進行線上檢索、閱覽、下載或列印。

論文全文上載網路公開之範圍及時間：

本校及台灣聯合大學系統區域網路	■ 立即公開
校外網際網路	■ 中華民國 97 年 7 月 19 日公開

■ **全文電子檔送交國家圖書館**

授 權 人：黃繼養

親筆簽名： 黃繼養

中華民國 96 **年** 7 **月** 19 **日**

國 立 交 通 大 學

博碩士紙本論文著作權授權書

（提供授權人裝訂於全文電子檔授權書之次頁用）

本授權書所授權之學位論文，為本人於國立交通大學資訊科學與工程研究所 資訊 組， 96 學年度第 二 學期取得碩士學位之論文。

論文題目：利用結構分解與抽象處理搜尋核糖核酸結構
指導教授：胡毓志

■ 同意

本人茲將本著作，以非專屬、無償授權國立交通大學，基於推動讀者間「資源共享、互惠合作」之理念，與回饋社會與學術研究之目的，國立交通大學圖書館得以紙本收錄、重製與利用；於著作權法合理使用範圍內，讀者得進行閱覽或列印。

本論文為本人向經濟部智慧局申請專利(未申請者本條款請不予理會)的附件之一，申請文號為：＿＿＿＿＿＿＿＿＿＿，請將論文延至＿＿＿年＿＿＿月＿＿＿日再公開。

授 權 人：黃繼養

親筆簽名： 黃繼養

中華民國 96 **年** 7 **月** 19 **日**

國家圖書館
博碩士論文電子檔案上網授權書

（提供授權人裝訂於紙本論文本校授權書之後）

ID:GT009455560

本授權書所授權之論文為授權人在國立交通大學資訊科學與工程研究所 96 學年度第 2 學期取得碩士學位之論文。

論文題目：利用結構分解與抽象處理搜尋核糖核酸結構
指導教授：胡毓志

茲同意將授權人擁有著作權之上列論文全文（含摘要），非專屬、無償授權國家圖書館，不限地域、時間與次數，以微縮、光碟或其他各種數位化方式將上列論文重製，並得將數位化之上列論文及論文電子檔以上載網路方式，提供讀者基於個人非營利性質之線上檢索、閱覽、下載或列印。

※ 讀者基於非營利性質之線上檢索、閱覽、下載或列印上列論文，應依著作權法相關規定辦理。

授權人：黃繼養

親筆簽名： 黃繼養

民國 96 **年** 7 **月** 19 **日**

國家圖書館
博碩士論文電子檔案上網授權書

（請於辦理離校程序時繳至系所助理用）

ID: GT009455560

本授權書所授權之論文為授權人在國立交通大學資訊科學與工程研究所 96 學年度第 2 學期取得碩士學位之論文。

論文題目：利用結構分解與抽象處理搜尋核糖核酸結構
指導教授：胡毓志

茲同意將授權人擁有著作權之上列論文全文（含摘要），非專屬、無償授權國家圖書館，不限地域、時間與次數，以微縮、光碟或其他各種數位化方式將上列論文重製，並得將數位化之上列論文及論文電子檔以上載網路方式，提供讀者基於個人非營利性質之線上檢索、閱覽、下載或列印。

☐ 論文全文上載網路公開之範圍及時間：2008.7.19 公開。

授權人：黃繼養

親筆簽名： 黃繼養

民國 96 **年** 7 **月** 19 **日**

國立交通大學

研究所碩士班

論文口試委員會審定書

本校 資訊科學與工程 研究所 黃繼養 君

所提論文:

利用結構分解與抽象處理搜尋核糖核酸結構元（Utilizing Structure Decomposition and Abstraction to find RNA Structure）

合於碩士資格水準、業經本委員會評審認可。

口試委員：黃崇源 荊宇泰

胡毓志

指導教授：胡毓志

所 長：曾煜棋

中 華 民 國 九十六 年 7 月 12 日

致謝

大學讀完後，發現自己對演算法的設計比較有興趣，但只解一些理論的題目又有些無聊，再加上對生物有些興趣，所以考上研究所後就選擇了生物資訊做爲研究的領域，回想起來，兩年過去了，能有不錯的研究成果，就覺得當時的選擇做對了，現在寫致謝文時的感覺是開心、是欣慰。

有人說做研究的日子是苦悶的，但很可惜，這句話無法做爲我這兩年生活的寫照，因爲我的身邊有一群閃亮亮的學長姐們，博班中有孔子接班人釗民、神之子子緯、實驗室常駐程式異昌、與美食專家均木，碩班中有積哥豐茂、熱哥勁伍、海哥昀君、帥哥秉蔚、海軍閃電貫中、憲兵之光登貴、生物專家世彥、以及最近才得到港姐稱號的音璇，有你們在生活中就是快樂與歡笑，謝謝你們。

而在研究實驗上最重要的，最要感謝的人當然是我的指導教授胡毓志老師，在這兩年中給我指導、傳授我知識，在實驗室人少的情況下，老師您還是努力的在研究上付出，您對實驗結果的執著與堅持，一直是我在偷偷效法的榜樣，感謝您這兩年來給我的教誨與幫助，謝謝您。此外，還有荊宇泰老師與黃崇源老師在在百之中抽空來幫我口試，並爲我的研究提出寶貴的意見，謝謝你們。

謝謝。

目錄

第一章、前言

1.1 研究動機

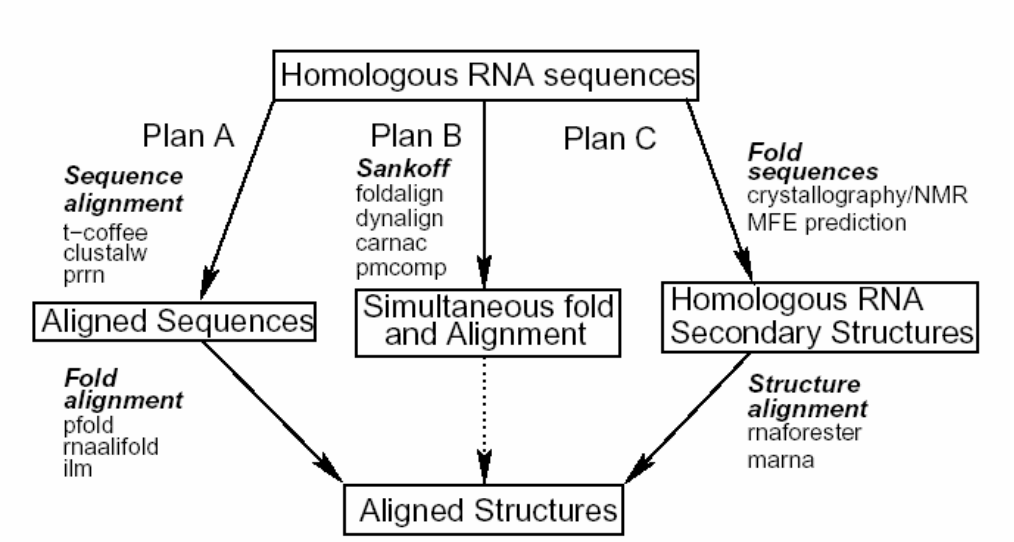
核醣核酸(RNA)在生命體中扮演很重要的角色，其中最爲人知的信使核醣核酸(mRNA)傳遞核醣核酸的資訊到核糖體，合成所需要的蛋白質。其他常見的還有轉錄核醣核酸(tRNA)、核糖體核醣核酸(rRNA)、微核醣核酸(microRNA)等。這些 RNA 會褶疊成特定的形狀來輔助生命機制，如催化化學反應及調控基因表現等等。

從已知的生物知識可知，褶疊形狀相似的核醣核酸很有可能也會有相似的功能。因此，若能由已知的核醣核酸序列來預測其褶疊而成的二級結構，進而猜測其功能，將能更迅速的瞭解生命運作的機制。

然而，在生物實驗室裡進行實驗來決定一個核醣核酸的結構是很費時的，單用人工的方式實驗非常沒有效率。因此，我們希望利用已知序列上的資訊，加入能量預測二級結構的資訊，藉由電腦的輔助以提供一個快速的方法，希望能預測出核醣核酸的結構，更進一步的從一個家族的核醣核酸序列中，預測出他們的共同結構元(motif)，因爲這些共同的結構在生物演化上可能是有意義的，他們可能控制著一種重要的生物機能，所以在經過長時間的演化之後，這些結構仍然保留至今。

研究核醣核酸二級結構預測(RNA secondary structure prediction)的方法有很多，例如使用動態程式規劃(dynamic programming)的方法尋找化學上能量最爲穩定的結構；或是以排比(alignment)的方式，利用一條已知二級結構核醣核酸序列上的資訊，去預測另外一條結構未知的相關核醣核酸序列；以及用基因演算法(genetic algorithm)的方式尋找二級結構和摺疊路徑(folding pathway)等。以上的方法都是只針對單一條核醣核酸序列提供唯一的最佳二級結構預測結果，或是包含多個次佳解的結果。

近年來，對於核醣核酸二級結構的研究主題多在預測同一核醣核酸家族的共同結構元，目前常見的方法有三大類：



(A) 先對所有核醣核酸序列做多重排比(multiple sequence alignment)，再將排比好的序列利用單一核醣核酸序列的二級結構預測系統進行摺疊(folding)，最後所得的摺疊結構即爲該家族的預測共同結構元。

(B) 以 Sankoff algorithm 爲基礎，使用動態程式規劃同時考慮序列排比與摺疊的資訊來預測同一家族序列的共同結構元。

(C) 利用單一核醣核酸序列的二級結構預測系統，對此家族的每一條核醣核酸序列各自進行單一序列的摺疊，再對所有產生的結構進行結構排比(structure alignment)。

本研究與上述的 Plan C 方法有點相似，在前半段使用單一核醣核酸序列的二級結構預測系統作爲前處理器，來預測單一核醣核酸序列的完整二級結構。然而在後半段，也是最主要的核心部分，我們並非只是對其產生的結構進行排比，而是將其預測的結構做分解後，再利用 Gibbs-like 流程來預測出此家族序列的共同結構元。

1.2 研究假設

關於核醣核酸二級結構的共同結構元預測，本研究設定了兩個合理的基本假設：

[假設一] RNA 的功能大部份是由其二級結構所決定

這點假設在 2002 Science magazine 中已經有 Couzin 等人所發表出來了，所以現在大家也大多認定這一點，因此我們的系統在判斷序列之間是否爲同家族時，大多會從結構的觀點出發，而不是從序列的觀點出發。

[假設二] 同一家族的核醣核酸序列在二級結構上有相似的子結構

一群核醣核酸序列之所以會被視爲同一家族，就是因爲他們有類似的功能。由化學的角度來看，當結構有些許改變就很有可能影響分子結合的能力，因而影響其功能，所以我們認爲，一群功能相同的核醣核酸序列行使功能之區域，其二級結構必定極爲相似。

本研究假設一群被歸類爲同一家族的相關核醣核酸序列中，從在某些共同的結構，而這些共同的結構則是決定此家族核醣核酸所行使的功能。

1.3 研究目的

在過去的研究中，預測核醣核酸二級結構的共同結構元用到許多不同的方法，包含動態程式規劃、隱藏式馬可夫模型(Hidden Markov Model)、序列排比、圖論方法以及演化式計算等等。每一個研究所切入的角度都不太一樣，對於不同的家族的共同結構元預測能力也不太相同，但目前的系統大多只能預測出長度較短的共同結構元。

而在本研究中，我們同樣使用 Gibbs-like 的法，試圖找出同一家族的共同結構元，加入能量的資訊縮小搜尋空間以節省搜尋時間，而資料結構的表示法我們也自己設計了 SCC(Self-Closed Component)和 R-Grammar 來表示，希望可以藉此找出較長或者是更複雜的共同結構元。

1.4 論文架構

本篇論文包含六個章節：

第一章爲前言，介紹本研究的動機、背景、此研究所使用的方法及其基本假設，以及主要的研究目的。

第二章爲文獻探討，將介紹核醣核酸的背景知識，以及此研究過去的發展。

第三章爲研究方法，是本篇論文的核心，詳細介紹本研究設計的方法流程與細節。

第四章爲實驗結果，整理所有實驗的內容與實驗的結果。

第五章爲結論與討論，分析本實驗的優缺點。

第六章參考文獻，則列出本研究參考的相關文獻。

第二章、文獻探討

2.1 核醣核酸簡介

長期以來，人們對於核醣核酸(ribonucleic acid, RNA)的瞭解不多，僅知道 RNA 在合成蛋白質的過程中扮演著”遺傳信使”的角色：將去氧核醣核酸(deoxyribonucleic acid, DNA)所攜帶的訊息帶到核醣體，作爲轉譯(translation)蛋白質使用。最近幾年隨著對 RNA 的研究發現愈來愈多，RNA 在生物學上的地位也愈來愈爲重要。

2.1.1 核醣核酸的重要性

除了早期所知的信使核醣核酸(messenger RNA, mRNA)外，有其他重要功能的核醣核酸也陸續被發現，如許多未編碼的核醣核酸(non-coding RNAs, ncRNAs)，其中有些甚至可以促進生化反應，控制細胞內蛋白質(酶)的合成，這類的核醣核酸包括轉錄核醣核酸(transfer RNA, tRNA)和核醣體核醣核酸(ribosomal RNA, rRNA)等。

還有能夠調控基因表現的核醣核酸，如微核醣核酸(microRNAs)。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，而這些構造在基因體裡是非常穩定的。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、組織生長、肥胖代謝，以及決定某些基因的表現時間。

在科學(Science)雜誌所刊載的 2002 年研究表明，一些長度較短的核醣核酸，即所謂的小分子核醣核酸(Small RNA)，能夠對細胞和基因的很多行爲進行控制，比如打開和關閉多種基因，刪除一些不需要的 DNA 片段等。它們在細胞分裂過程中更是發揮了至關重要的控制作用，可指導染色體中的物質形成正確的結構，防止 DNA 片段位移出錯。若 DNA 功能的產生錯亂，可能是引發癌症的一個重要原因。

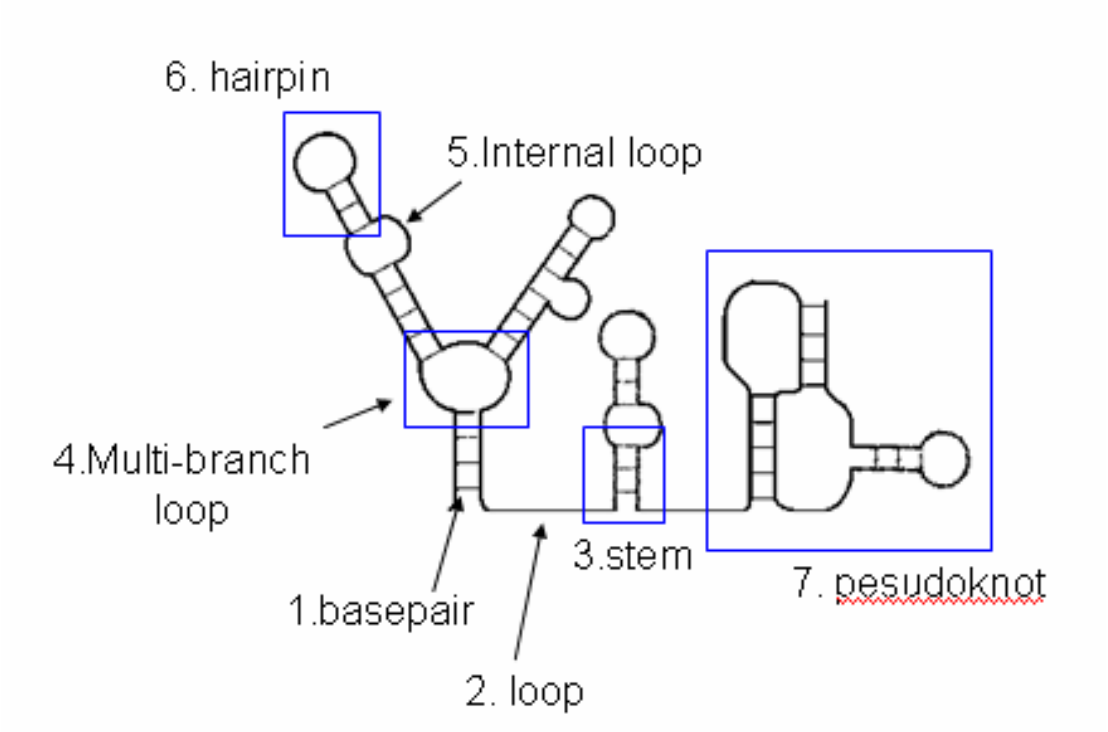
2.1.2 核醣核酸結構基本單位元

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，在核醣核酸的相關研究上，我們對於核醣核酸的結構所產生的興趣，遠大於對於序列的分析。談結構之前，還是必須先對序列組成有基本的了解。

核醣核酸由四種含氮鹼基組成，分別是腺嘌呤(Adenine) 、胞嘧啶(Cytosine)、鳥糞嘌呤(Guanine)、尿嘧啶(Uracil)，習慣上常分別以 A、C、G、U 來代表這四種含氮鹼基。

核醣核酸常以單股存在於生物體中，透過分子間的作用力，會自己摺疊成特定的結構，產生摺疊的作用力主要來自於 C≡G 三個氫鍵的鏈結以及 A＝U 兩個氫鍵的鏈結，此兩組鏈結的個別配對稱之爲標準鹼基對(canonical base pair)。此外還有一個搖擺鹼基對(wobble base pair)G－U，爲一個氫鍵的鏈結，此結構較不穩定，需要週圍的鹼基對輔助。由 A、G、C、U 各自配對所產生的摺疊形成了 RNA 的基本結構，稱之爲 RNA 的二級結構。

2.1.3 核醣核酸二級結構



核醣核酸基本的二級結構如下：

1. 莖幹結構(stem)

   核醣核酸序列中，連續鹼基配對所形成的一個長狀形狀，稱之爲莖幹結構。

2. 髮夾環狀結構(hairpin loop)

   當一個連續的非配對區域不是出現在序列的終端，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。而此環狀結構與相鄰的莖幹則合稱一個髮夾結構。

3. 內部環狀結構(inner loop)

   一個連續的非配對區域恰與兩個莖幹相連，而且兩側都有未配對鹼基，則該區域即爲內部環狀結構。內部環狀結構又可分爲對稱性(symmetrical)與非對稱性(asymmetrical)，當兩側未配對鹼基個數相同時，則稱爲對稱性內部環狀結構，反之則稱爲非對稱性內部環狀結構

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，則稱這些未配對的區域爲突起結構。

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，則稱爲多分支環狀結構。

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，該結構就稱之爲擬結結構。

2.2 預測核醣核酸結構的相關方法

研究核醣核酸的目的是希望能夠了解核醣核酸在生物體裡所擁有的功能，而讓這些核醣核酸有其功能的原因不在於它的一級結構(序列)，而是它所折疊而成的二級結構。目前生物學家認爲，分子的結構是影響核醣核酸功能的關鍵，例如常見的轉移核醣核酸，其結構都是很穩定的苜蓿葉(cloverleaf)結構：包含四個莖幹而形狀類似四瓣的苜蓿葉。另外，擁有相同生化功能的同一家族成員，也常會擁有相似的二級結構片段。

因此，若能利用計算機的輔助，能夠迅速的發現同一家族成員中的共同結構元，這對生物學是很有幫助的。這不僅能協助生物學家快速找出該家族行使其功能的結構片段，亦能利用已知的共同結構，檢驗未知功能的序列，來推論出其功能，進而找到所屬家族。

過去研究核醣核酸二級結構預測的方法很多，本節簡述過去幾個比較具代表性的方法：

2.2.1 單一核醣核酸序列二級結構預測

給定一條核醣核酸序列，我們希望預測出它摺疊而成的二級結構，最常見的方法則是使用熱力學(thermodynamics)的知識來推論此序列可能折疊而成的形狀，以下爲幾個代表的系統：

2.2.1.1 Mfold

Mfold是一套單一核醣核酸序列的二級結構預測系統，實作Zuker與Stiegler所提供的演算法，利用動態程式規劃法(Dynamic Programming)計算出核醣核酸序列擁有最小自由能量(minimal free energy, MFE)的摺疊結構，以預測最爲穩定的結構。當序列的長度爲n時，系統所需的時間複雜度爲$O(n^3)$，所需要的空間複雜度爲$O(n^2)$。此系統可以依照自由能量的大小，由小到大輸出數個可能的二級結構供使用者參考。

然而核醣核酸在摺疊的過程中可能受到某些因素或是受到其他分子的影響，使得理論上最穩定的結構無法形成，單純依靠最小能量來斷定結構形狀仍會有很大的不足。另外，Mfold 無法摺疊成擬結結構，這也是其缺點之一。但在許多相關的研究上，Mfold 仍被廣泛應用，其確實提供了相當程度的資訊。

.

2.2.1.2 RNAfold

RNAfold 是維也納 RNA 研究團隊(Vienna RNA package)所實作的單一核醣核酸序列的二級結構預測系統，其運作原理與 Mfold 一樣皆是建立在 Zuker 與 Stiegler 所提供的演算法上，以動態程式規劃法找出能量最小而最爲穩定的二級結構。

RNAfold 與 Mfold 的運作原理相同，差別只在於實作的方式不同，兩者所預測出來的核醣核酸二級結構結果差異性很小，從兩者的比較研究顯示，這兩個不同的單一核醣核酸序列二級結構預測系統，從準確性上看來沒有重大的差別存在。

2.2.1.3 Sfold

Sfold 實作了另一種以能量爲基礎的單一序列折疊演算法，給定一條核醣核酸序列，利用統計的方法取出其二級結構的樣本，接著依據給予的熱力學參數產生核醣核酸二級結構的相稱分割函數(equilibrium partition functions)，根據分割函數使用條件機率對所有可能的結構進行遞迴取樣，而後產生二級結構的統計上典型樣本，最後使用分群(clustering)的技術獲得可能的結構。可根據最小自由能取出前幾名可能的結構以供使用者參考。

根據先前的研究分析，從準確度上看來，Sfold 的結果與 Mfold 和 RNAfold 產生的結果非常相似，但 Sfold 相較於其他兩者而言，它的結果的變異性(variance)有稍微高出了一些。

2.2.2 根據多重序列排比結果進行摺疊來預測核醣核酸共同結構元

預測核醣核酸共同結構元的一類逼近方法，先同時對所有核醣核酸序列進行多重排比，再將排比結果的序列摺疊成二級結構。而進行多重排比的方法，最常見的爲 ClustalW，其不僅擁有長久的歷史，且其結果也優於許多其他類似的工具。而摺疊的方法則是各有明顯的差異。

2.2.2.1 RNAalifold

RNAfold是維也納RNA研究團隊(Vienna RNA package)所開發的系統之一，可預測出多條已排比好序列的一致結構，其原理爲Zuker-Stiegler演算法的延伸，摺疊結構時同時考慮最小自由能(MFE)和共變(covariation)關係。當資料有N條序列，而最長序列長度爲n時，本系統的時間複雜度爲$O(N \cdot n^2+n^3)$，空間複雜度爲$O(n^2)$。

2.2.2.2 Pfold

Pfold 使用隨機前後無關文法(stochastic context free grammar, SCFG)，產生核醣核酸結構的先前機率分配(prior probability distribution)，針對輸入的已排比核醣核酸序列和系統發生的樹狀結構(phylogenetic tree)，計算出此結構的後端機率(posterior probabilities)，而後進行行(column)的排比或行的配對。最後在 SCFG model 中找到最大可能發生樹(maximum-likelihood tree)，產生最有可能的核醣核酸二級結構。

2.2.2.3 ILM

ILM(iterated loop matching)使用熱力學和相互資訊(mutual information)的結合產生一個二級結構，接著產生所有可能的莖幹，根據熱力學和相互資訊的結合分數對莖幹進行排序。選擇分數最高的莖幹，更新分數，然後將與被選上的莖幹有衝突的莖幹移除，之後再選擇分數第二高的莖幹，接著一直重複此動作直到沒有其他莖幹剩下，最後的所有莖幹則決定了結構。

2.2.3 同時考慮序列排比與摺疊的資訊來預測核醣核酸的共同結構元

Sankoff algorithm是一種合併了做序列排比與做結構摺疊的動態程式規劃方法，它可以被用來獲得排比結果和一致性的共同結構。而最原始的Sankoff algorithm實作雖然可以同時做結構摺疊與序列排比，但其負擔卻是相當的大，當資料有N條序列而最長序列長度爲n時，其運作所需的時間複雜度爲$O(n^{3N})$，空間複雜度爲$O(n^{2N})$。因此，爲了減少系統運作的負擔，則有了一些新的實作方法，針對原始的Sankoff algorithm加了一些限制，而能在預測核醣核酸的共同結構元時仍有不錯的表現。

2.2.3.1 Foldalign

Foldalign 可被視爲一個區域性排比(local alignment)與鹼基配對數最大化(maximum number of base-pairs)演算法的混合體，它使用了與 CLUSTAL 和 CONSENSUS 相似的啓發式方法(heuristics)，由兩條序列的排比與鹼基配對的關係建立了分數矩陣(scoring matrix)，使用由 Sankoff algorithm 延伸的動態程式規劃法求出兩條最佳配對排比結果(pairwise alignment)。而系統將所有序列兩兩成對個別求出其排比結果，從中取出分數最高的排比結果，再個別與其他序列進行排比，從中再取出最好的配對排比結果，此時的配對排比結果即爲三條序列的最佳配對排比結果。之後再依此方法持續循環下去，最後所得即爲所有序列的最佳配對排比結果。

Foldalign將Sankoff algorithm延伸實作，但限制了尋找的共同結構元最大長度，而且禁止了多分支環狀結構(multi-loops)的產生，因此可以降低系統運作的負擔。當資料有N條序列而最長序列長度爲n時，其運作所需的時間複雜度爲$O(n^4N)$。

Foldalign 被專門設計來預測短區域的調控共同結構元，例如 IREs(iron response element)中的髮夾結構(hairpin structures)，因此在找尋全域性(global)的結構與多分支環狀結構上的表現不佳。

2.2.3.2 Dynalign

Dynalign 結合了自由能最小化(free energy minimization)與比較序列分析(comparative sequence analysis)，依此找出兩條序列低自由能的共同結構。系統先對兩條序列進行排比，再分別對兩條序列進行摺疊，而摺疊的結構其鹼基可以產生配對的條件爲：必須兩條序列在排比結果的同個位置上皆能產生標準鹼基對，亦即使兩條序列可以摺疊成相同的結構。

Dynalign 的目的將整個系統的總自由能做最小化，總自由能的求法爲：

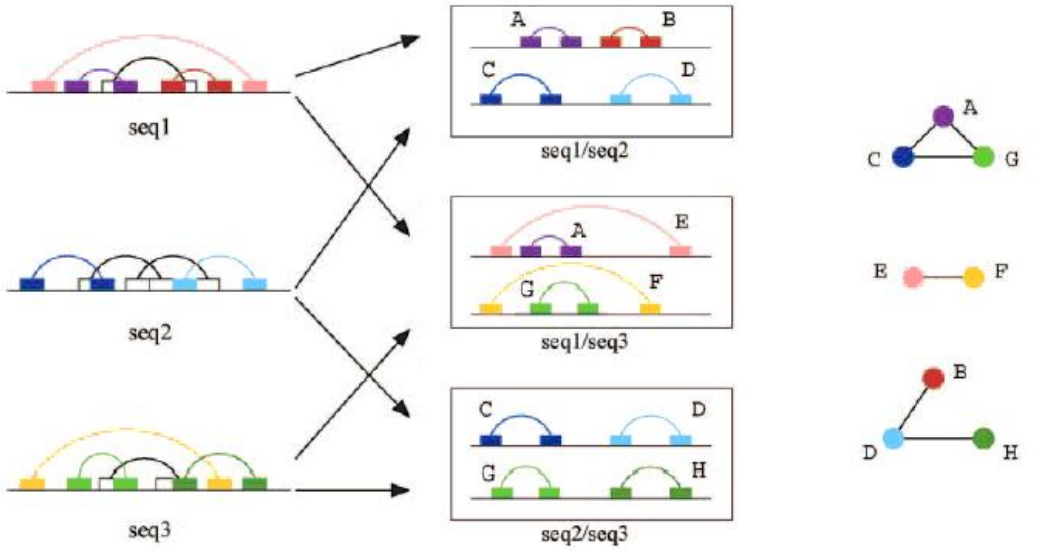
$$\Delta G^{\circ}_{\text{total}} = \Delta G^{\circ}_{\text{sequence 1}} + \Delta G^{\circ}_{\text{sequence 2}} + (\Delta G^{\circ}_{\text{gap}})\,(\text{number of gaps})$$

ΔG°_{total} 表示整個系統的總自由能，$\Delta G^{\circ}_{sequence\ 1}$ 與 $\Delta G^{\circ}_{sequence\ 2}$ 分別爲序列 1 與序列 2 的構造自由能(conformational free energy)，ΔG°_{gap} 爲兩條序列排比產生的缺口(gap)造成的處罰值(penalty)，此值根據經驗設置。

Dynalign使用全能量模型(full energy model)，進行Sankoff algorithm的動態程式規劃法對系統的總自由能做最小化，但在進行演算時則限制了兩條序列在進行排比時的最大距離，即當序列 1 的第i個鹼基要與序列 2 第j個鹼基排比在一起，則i與j的差值必須小於由使用者設定的M值。使用這樣的限制可以使系統的時間複雜度降爲$O(n^3M^3)$，而空間複雜度則爲$O(n^2M^2)$，其中n爲較短序列的序列長度。

Dynalign只能同時找兩條序列的共同結構元，儘管可以擴展至多條序列，但會造成系統的嚴重負擔，例如當序列數爲三條時，系統的時間複雜度會增至$O(n^3M^6)$，而空間複雜度則增爲$O(n^2M^4)$。

由實驗的測試結果顯示，Dynalign 在較短的且較多樣性的 tRNA 預測上有比較好的表現。



2.2.3.3 Carnac

Carnac 同時考慮區域相似性(local similarity)、莖幹能量(stem energy)和共變關係(covariation)，產生序列的共同摺疊二級結構。此系統採用啓發式的演算法，概略圖如上圖，演算法步驟如下(設有 N 條序列)：

Step1: 對所有的序列分別找出每條序列所有可能的莖幹，再使用熱力學的知識，利用動態程式規劃法計算出每條莖幹的自由能，留下能量低於預設門檻值的所有莖幹。

Step2: 將所有的序列兩兩成對，分別建立所有可能的N*(N-1)/2 個序列對成對摺疊(pairwise foldings)。方法爲先找出兩條序列鹼基高度相似的區域，考慮區域相似性與共變關係找出成對的莖幹(pairwise stems)，然後根據所選到的所有莖幹，考慮能量最小化使用類似Sankoff algorithm的動態程式規劃法找出最佳的共同摺疊。而此動態程式規劃法與Sankoff algorithm的差異點在於Carnac將序列的莖幹視爲基本單位元去運作，而不是像一般皆以含氮鹼基視爲基本單位元。因此找兩條序列的共同結構元的時間複雜度只需要$O(n^2)$，所需要的空間複雜度亦爲$O(n^2)$。

Step3: 此步驟將 Carnac 擴展至可以同時找多條序列的共同結構元。在經過

step 2 之後，每條序列皆得到 N-1 個預測結構，爲了得到最有可靠度(reliable)的莖幹，於是建立了一套新的資料結構，稱之爲莖幹圖(stem graph)。在莖幹圖中的所有點(vertices)的集合表示所有序列預測出來的所有莖幹的集合，觀察序列 1 中的任一莖幹 A 與序列 2 中的任一莖幹 C，若配對(A,C)出現在 step 2 產生的成對摺疊中，則在莖幹圖中的點 A 與點 C 建立一致邊(identity edge)。觀察完每條莖幹，建立完整的莖幹圖，再對莖幹圖中的每個連通單元(connected component)進行排序，排序的依據則考慮各個莖幹圖的幾個拓撲特徵(topological features)：(i)莖幹圖中節點的數目，(ii)每條序列的莖幹數目，(iii)所有邊的總數，以及(iv)各個一致邊的數目。最理想的情形是連通單元可以構成一個頂點數爲 N 的完全圖(complete graph)，而每個頂點皆來自於不同的序列。最後則根據排序好的最佳連通單元完成其二級結構。

Carnac 在鹼基對的預測測試中發現其結果有很高的選擇性(selectivity)，然而它的敏感性(sensitivity)一般而言卻偏低，儘管可以藉由限制最小自由能的摺疊來提高其敏感性，但如此一來則相對的會因此降低其選擇性，而相關性(correlation)則有非常些微的提高。

2.2.4 對核醣核酸摺疊結構進行排比來預測核醣核酸共同結構元

當已知序列有可信賴的二級結構時，我們可以考慮藉由結構提供的資訊進行多重結構排比，由此預測核醣核酸的共同結構元。而每一條序列的二級結構，可以使用 2.2.1 節裡介紹的單一核醣核酸序列二級結構預測工具來取得，其中的 Mfold 與 RNAfold 皆常被各相關研究引進使用。

2.2.4.1 RNAforester

RNAforester 建立樹狀排比模型(tree alignment model)，依此推論核醣核酸二級結構的多重排比，只考慮核醣核酸分子的二級結構而不需要知道其序列的相似性。系統使用其他單一核醣核酸序列二級結構預測工具將序列轉爲二級結構，再將預測的二級結構轉換成樹狀結構(tree)或森林結構(forest)的輪廓圖(profile)，之後將 ClustalW 多重序列排比的演算法延伸爲多重結構排比，以此演算法對所有序列轉換成的輪廓圖進行多重結構排比，由排比結果可得預測的核醣核酸共同結構元。

當實驗的資料序列有N條，其平均的長度爲n，設d值爲輪廓圖中的樹狀結構節點的最大分支度(degree)，則此系統運作的時間複雜度爲$O(n^2d^2N^2)$，空間複雜度爲$O(Nn+N^2+n^2d)$。

2.2.4.2 MARNA

MARNA 同時考慮核醣核酸一級序列與二級結構產生 RNA 的多重排比，它建立了權重排比邊(weighted alignment edges)的集合，而這些邊的權重則反映了序列的和結構的共通性(conservation)，其計算方法須考慮到序列與結構兩部分，而結構部份則參考由單一核醣核酸序列二級結構預測工具產生的預測結構。之後將這些邊的集合輸入 T-coffee 系統，產生多重排比的結果，最後可從此結果擷取出一致性的序列與一致性的結構。

當實驗的資料序列有N條，假設每條序列長度皆接近爲n，設E值爲每條序列所產生的預測結構個數(此值通常極小)，則此系統運作的時間複雜度爲$O(E^2N^2n^4)+O(N^3n^2)$。

2.3 核醣核酸資料庫

由於核醣核酸的相關研究蓬勃發展，已知的核醣核酸序列及結構資料量快速地成長，於是有許多相關的生物資料庫收集了分散在各個文獻的資料，以各自設計的方法系統化地將核醣核酸的資料分門別類整理，公開提供給所有相關的研究人員使用。目前核醣核酸相關的資料庫有許多，以下則簡單介紹幾個常用的資料庫。

2.3.1 Rfam

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

Rfam(RNA families database of alignments and CMs)，是由多重序列排比與共變模型所建立的資料庫，其中儲存了多數家族的核醣核酸資料，包含家族成員各自的鹼基資料、家族的多重序列排比結果、以及家族二級結構的共同結構元等，是一個廣泛被使用的資料庫。

2.3.2 tRNA Compilation 2000

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

此資料庫收集了大量的轉錄核醣核酸序列，亦包含了明確的結構資訊。資料庫中提供了查詢的功能，可在 11 個界(kingdom)中選擇適當的分類，再從界之下的有機體(organism)分立中做選擇，最後再查詢是攜帶哪一種胺基酸的轉錄核醣核酸，這比較適合有生物背景的使用者使用。

2.3.3 RNABase

(http://www.rnabase.org/)

RNABase(The RNA Structure Database)資料庫整合了 Protein Data Bank(PDB)與 Nucleic Acid Data Base(NDB)兩者的核醣核酸資料，再依功能與結構的不同來做分類。此資料庫的主要特色是能提供核醣核酸的 3D(three-dimensional)結構圖，另外還能執行結構的分析與檢測。

2.3.4 SCOR

(http://scor.lbl.gov/)

SCOR(Structural Classification or RNA)提供了核醣核酸共同結構元的階級分類，分別以生物功能、二級結構元和三級立體結構爲依據，提供了三種不同的分類方法。而生物功能類別則分別以分子功能、結構元功能與結構模型向下細分；二級結構元類別則分類成髮夾結構和內部環狀結構，而各類別底下再依據結構的形狀做更小的細分；三級立體結構類別則以各種形狀不同的相互作用作爲細分的依據。

2.3.5 RAG

(http://monod.biomath.nyu.edu/rna/rna.php)

RAG(RNA-As-Graphs web resource)是一個存放 RNA 二級結構的資料庫，利用圖學理論(Graph Theory)的結果，提供了一個量化的方法可以對 RNA 二級結構的拓撲(topology)進行分類，相較於其他 RNA 的資料庫，RAG 容易用於比較相異二級結構的相似與相異處。

RAG 提供了兩種二級結構拓撲的表示法：RNA tree graphs 及 RNA dual graphs，此兩種表示法可以列舉出所有可能的 RNA 二級結構元。

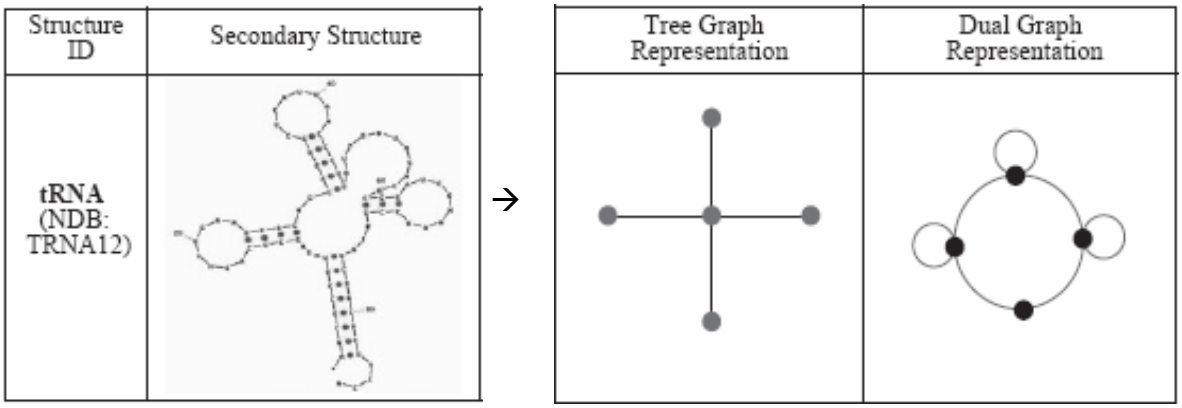
RNA tree graphs：

將突起結構與所有環狀結構都視爲一個點(vertex)，而莖幹結構則視爲一個邊(edge)，如此便能將一個 RNA 二級結構表示成一個 RNA tree graph。但此表示法無法表示擬結結構。

RNA dual graphs：

將莖幹結構視爲一個點，而突起結構或環狀結構的單股(single strand)則視爲一個邊，如此便能將一個 RNA 二級結構表示成一個 RNA dual graph。此表示法可以表示所有可能的 RNA 二級結構，包含了擬結結構。而 RAG 的接續下來的研究也都著墨在 RNA dual graphs 的特性。

RNA tree graph 與 RNA dual graph 示意圖：



RAG 提供了圖學中圖形拓撲的表示法，然而擁有同一種拓撲的相異 RNA，其二級結構還是很有可能會有很大的不同，因爲缺少了每個點跟邊的長度資訊，即使在每個點跟邊都只有些微差異的情況下，累積起來的差異依舊不小，這對尋找 RNA 二級結構的共同結構元影響頗大。這是目前 RAG 的圖形拓撲表示法較爲不足的地方。

2.3.5 其他常見資料庫

PseudoBase(http://wwwbio.leidenuniv.nl/~Batenburg/PKB.html)收集了擬節結構的核醣核酸相關資料，包含了序列、結構與生物功能三類資訊。

5S ribosomal RNA database(http://rose.man.poznan.pl/5SData/)專門爲 5S 的核醣體核醣核酸所建置，提供了這些序列的排比資訊與二級結構。另外也提供了與這些核醣核酸結合蛋白質資訊。

miRBase(http://microrna.sanger.ac.uk/)收集了微核醣核酸序列，可依物種分類瀏覽，此資料庫亦包含了各微核醣核酸的先質(precursor)，也有提供搜尋介面，使用者可根據序列片段、編號或名稱進行搜尋。

第三章、研究方法

3.0 系統設計目的與概念

在我的研究中，主要分成 2 大部份，第一部份是提出一個描敘核醣核酸結構的語言，第二部份是利用 2 次 Gibbs-like 的流程尋找屬於同家族的共同結構元。在第一部份中的描敘語言是由 Content Free Grammar 表示，並爲了第二部們中的 Gibbs-like 流程延伸出 4 個演算法：

(1) 分解演算法(Decomposition Algorithm)：列舉出一個核醣核酸結構之中，所有可能爲共同結構元的候選者。因爲第二部份中是用 Gibbs-like 的方式尋找共同結構元，因此列舉出所有的候選者可以方便 Gibbs-like 搜尋。

(2) 計算抽象形狀演算法(Abstraction)：標示出核醣核酸結構的形狀(Shape)和抽象形狀(Abstract Shape)。在第二部份的系統中，我們提出一個假設「相同家族之中的結構們有非常高的機率其抽象形狀會相同」，因此當要在一大群候選者中找出共同結構元之前，可以先用抽象形狀做簡單的分群。

(3) 計算相對長度差異演算法(Related Length Difference，簡稱爲RLD)：當 2 個核醣核酸結構擁有相同的抽象形狀時，可以計算彼此之間的相對長度差異。在許多核醣核酸結構研究中，若要判斷 2 個結構之間的相關性，往往都需要做結構排比(Structure Alignment)，雖然結構排比的復雜度($O(n^2)$)不算高，但核醣核酸實驗中往往需要上百萬次的結構排比，這是很花時間的數量，因此設計出RLD這個復雜度爲O(stem個數)的演算法來在做結構排比之前的一個門檻，因爲計算RLD幾乎是常數時間的復雜度，因此可以大大的減少實驗時間。

(4) 新的排比演算法(new Structure Alignment Algorithm)：因爲第二部份中，有先用抽象形狀做初步分群，因此會被 Gibbs-like 流程取來做結構排比的 2 個結構必定抽象形狀相同，所以我們加了這個條件後，設計出新的結構排比演算法。

第二部份則是延伸原先用來找去氧核醣核酸序列共同結構元的 Gibbs Sampling，使其擴展到核醣核酸二級結構上的共同結構元，其中提出一個進階的權重矩陣(Modified Weight Matrix)使它能表示成員長度不一的共同結構元集合，並對 Gibbs 的流程做一些簡易的修改，使其能更快、更適合用於核醣核酸共同結構元上的搜尋，其內容會在隨後的文章中詳細介紹。

3.1 核醣核酸結構描述語言

這節中我們將設計一套用來描敘核醣核酸結構的語言，以方便我們對核醣核酸的結構有一套規則去解釋它，再利用這個規則，設計或修改一些已知的方法來對敘核醣核酸結構做進一步的分析，下面爲詳細的介紹。

3.1.0 核醣核酸結構的形成

(1) 自身封閉單元(Self-closed Component)：

SCC 是核醣核酸二級結構中的子結構，並符合以下幾個條件：

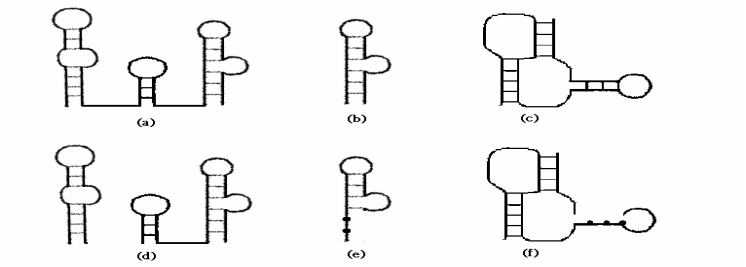
1. 由連續的核醣核酸序列組成。
2. 其起點與終點必爲莖幹結構(子結構中無基對時可忽略這個條件)。
3. 由起點的鹼基到最後的鹼基之間所有的基對都必需是完整的(基對的 5’和 3’都要包含在內)，

則此核醣核酸子結構稱之爲自身封閉單元(Self- closed Component，簡稱爲 SCC)，如下圖中(a)(b)(c)爲合格的 SCC，而(d)(e)(f)則不是)。

爲了方便而後的研究討論，先定義以下 2 種特定的 SCC：

*. Max-SCC：若某 SCC 不包含在任何其它的 SCC 之中，則稱它爲 Max-SCC(如下圖中(a)爲 Max-SCC，但(b)包含在(a)中所以不是)。

*. Min-SCC：若某 SCC 不包含其它 SCC，則稱它爲 Min-SCC

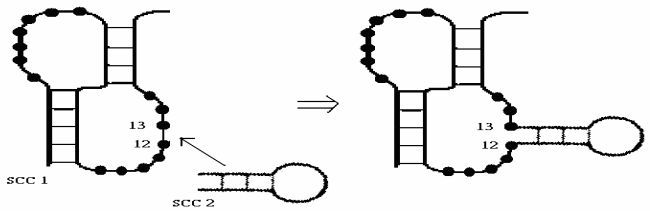


(2) 核醣核酸結構的結合：

核醣核酸結構的結合方式有 2 種：

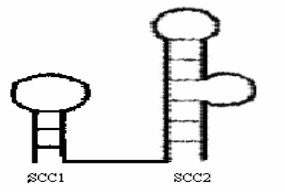
1. 插入

將SCC_2插入SCC_1上第i和第i+1個鹼基之間計作SCC_1.add(SCC_2, i)。如下圖爲SCC_2加到SCC_1的第12和第13個loop之間、即爲SCC_1.add(SCC_2, 12)。



2. 並排

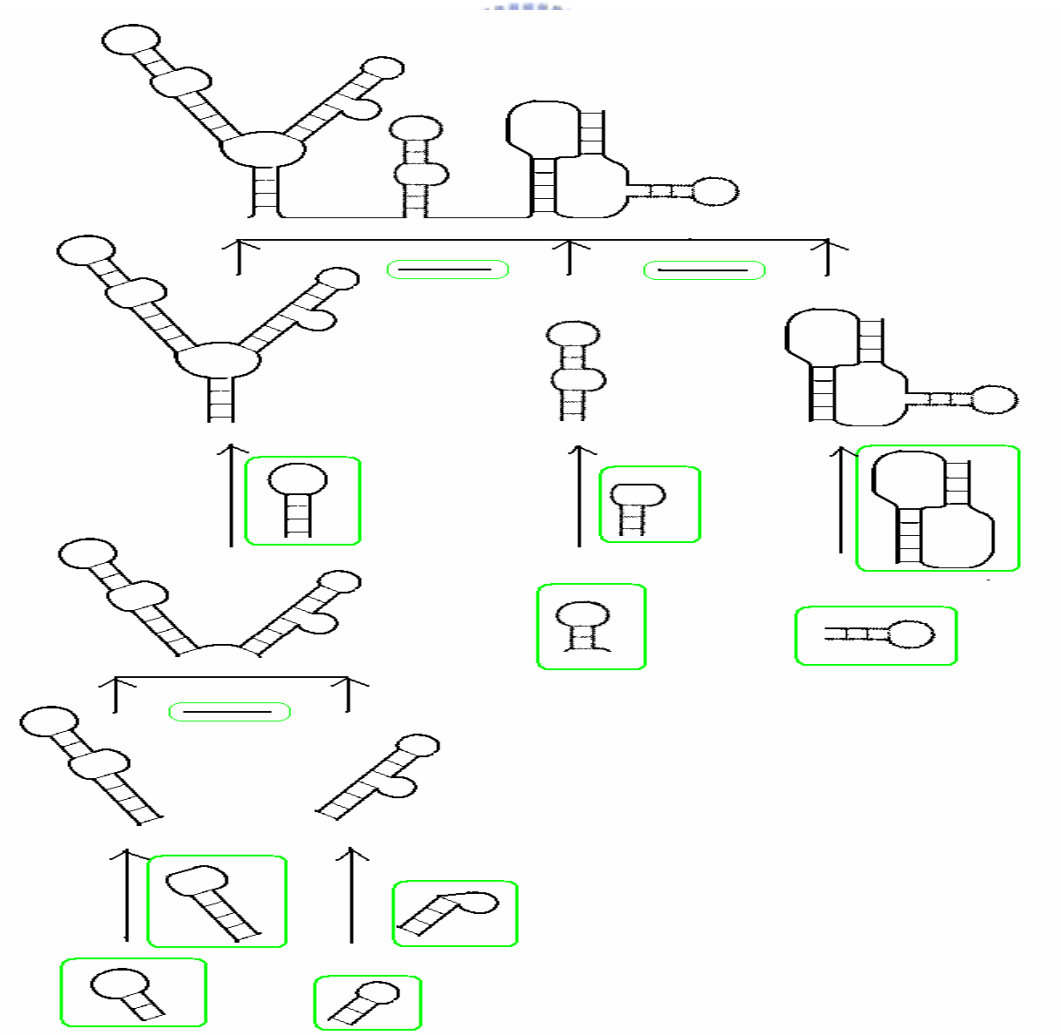
將SCC_1和SCC_2並排在一起，之間用若干個單股鹼基連接而結合在一起，記做SCC_1-SCC_2，如右圖所示。



在(1)和(2)的討論就可以感覺的出來，其實再復雜的核醣核酸結構都是由若干個min-SCC再利用(2)中所說的2個動作結合而成的，說起來有點抽象，但觀察下圖即能得到一個詮釋，其中有些被框起來的小圖即爲構成整個核醣核酸結構的min-SCCs(下圖中的樹稱爲「生成樹」，其性質會在下一頁中介紹，在此不贅述)。因此我們選擇用Content Free Grammar的方式來表示核醣核酸結構遞迴結合的過程(之後簡稱之爲R-Grammar)，如下所示：

S -> S-S | M.add(S, i) | M

M -> min-SCC



(3) 核醣核酸結構的生成樹(RNA Generating Tree)：

顧名思義此樹就是用來表示被分析的核醣核酸結構的生成情況，此生成樹有以下幾個特性：

<1> 樹根爲被分析的核醣核酸結構。

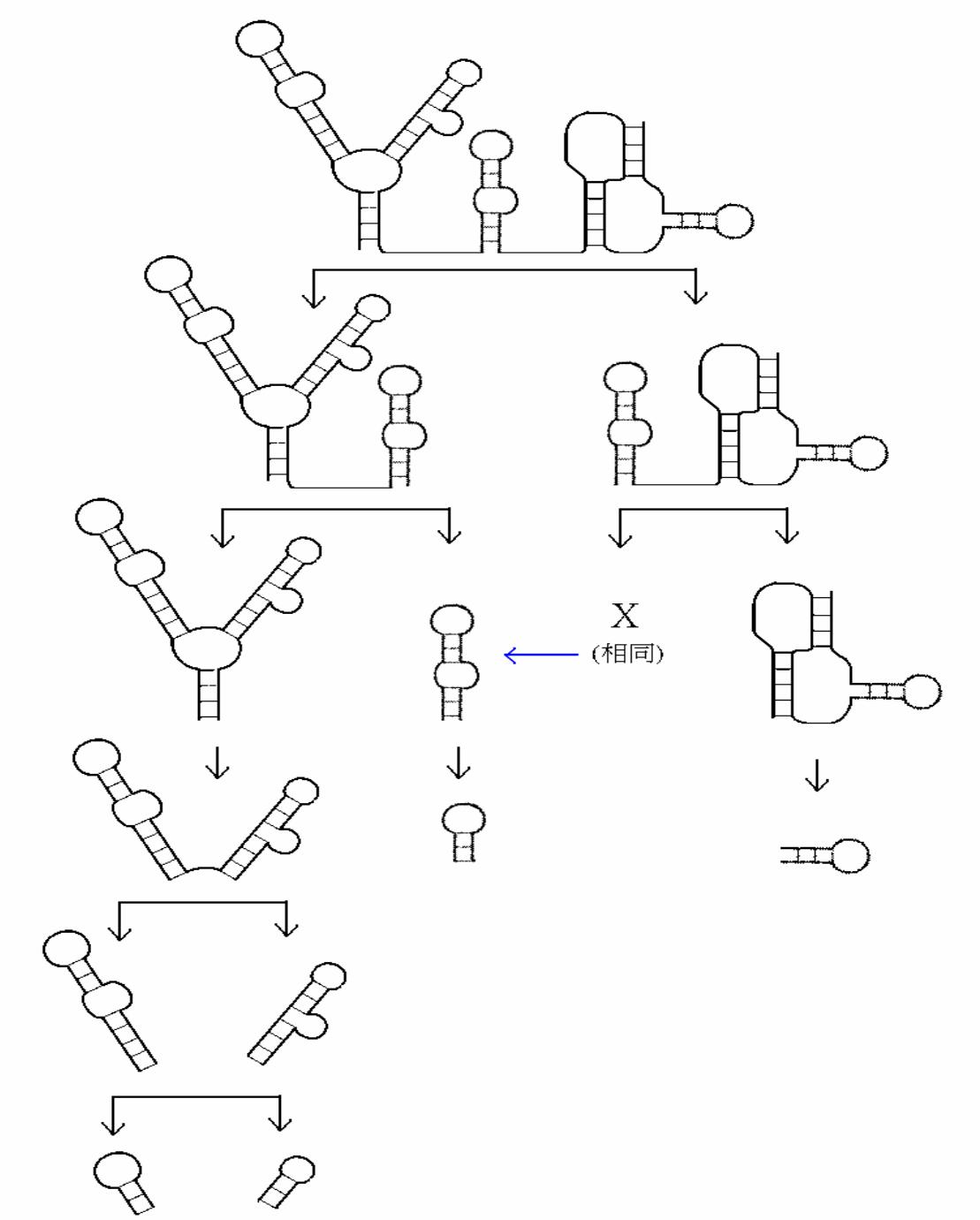
<2> 所使用到的min-SCC們會出現在樹葉和分支數爲一的邊上。

<3> 分支數爲一的邊上會放著此時加入的min-SCC，也就是由第一種核醣核酸結構的結合方式「插入」所構成的；而分支數大於或等於二的則爲「並排」所構成的，其中會加入（子結點個數-1）個內部環結來將子結點中的SCC們連接起來。

有了生成樹之後，可以由它是待知原核醣核酸結構是由那些min-SCC們組合的，依何種順序方式組成的，這些資訊對核醣核酸結構的分類或是設計分析核醣核酸結構的演算法都會有很大的幫助。

3.1.1 分解演算法(Decomposition Algorithm)

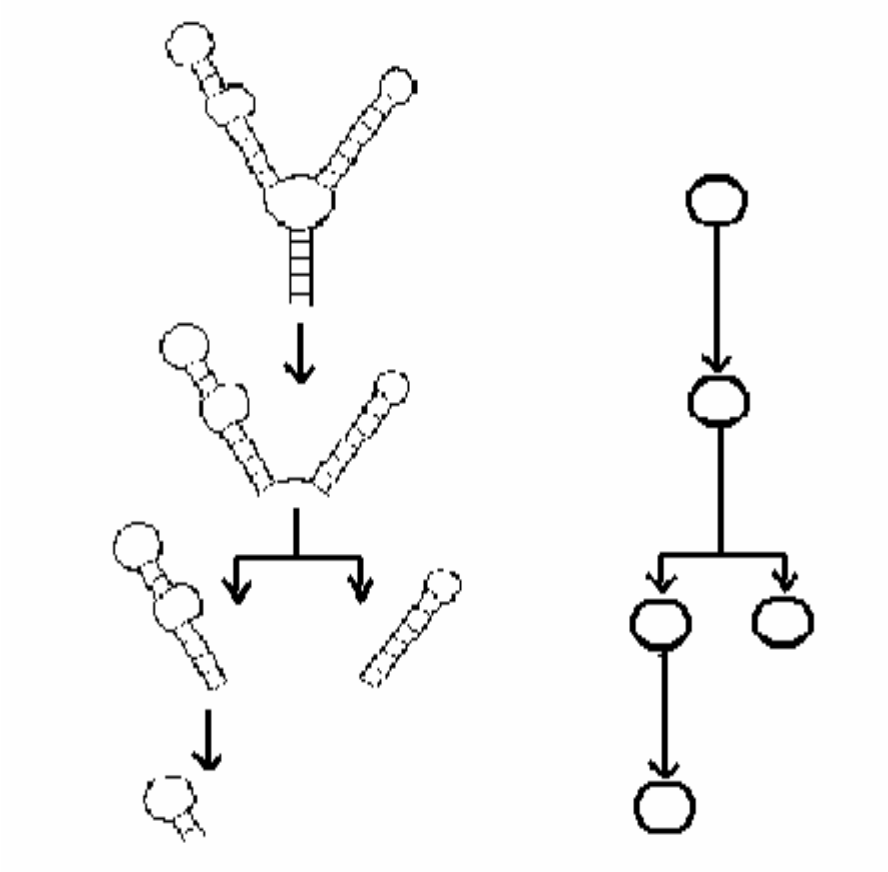
爲了方便 Gibbs-like 搜尋，我們必需要能列舉出一個核醣核酸結構之中所有可能爲共同結構元的候選者，因此需要分解演算法，幸運的是 R-Grammar 本身就能爲輸入的核醣核酸結構做分解了，其中只要注意不要分解出相同的子結構即可。如下圖則是使用 Left-most 分解的結果，而右子樹中的 X 表示有分解出之前已經分解過的子結構，此時就會停止不繼續分解下去。這個分解演算法非常的有效率，如下圖樹中的每個結點都是一個候選者，並且不會有重復的情況，因此時間復雜度和結構中的莖幹個數成正比，爲 O(s)，s 爲莖幹個數。



3.1.2 計算形狀與抽象形狀演算法

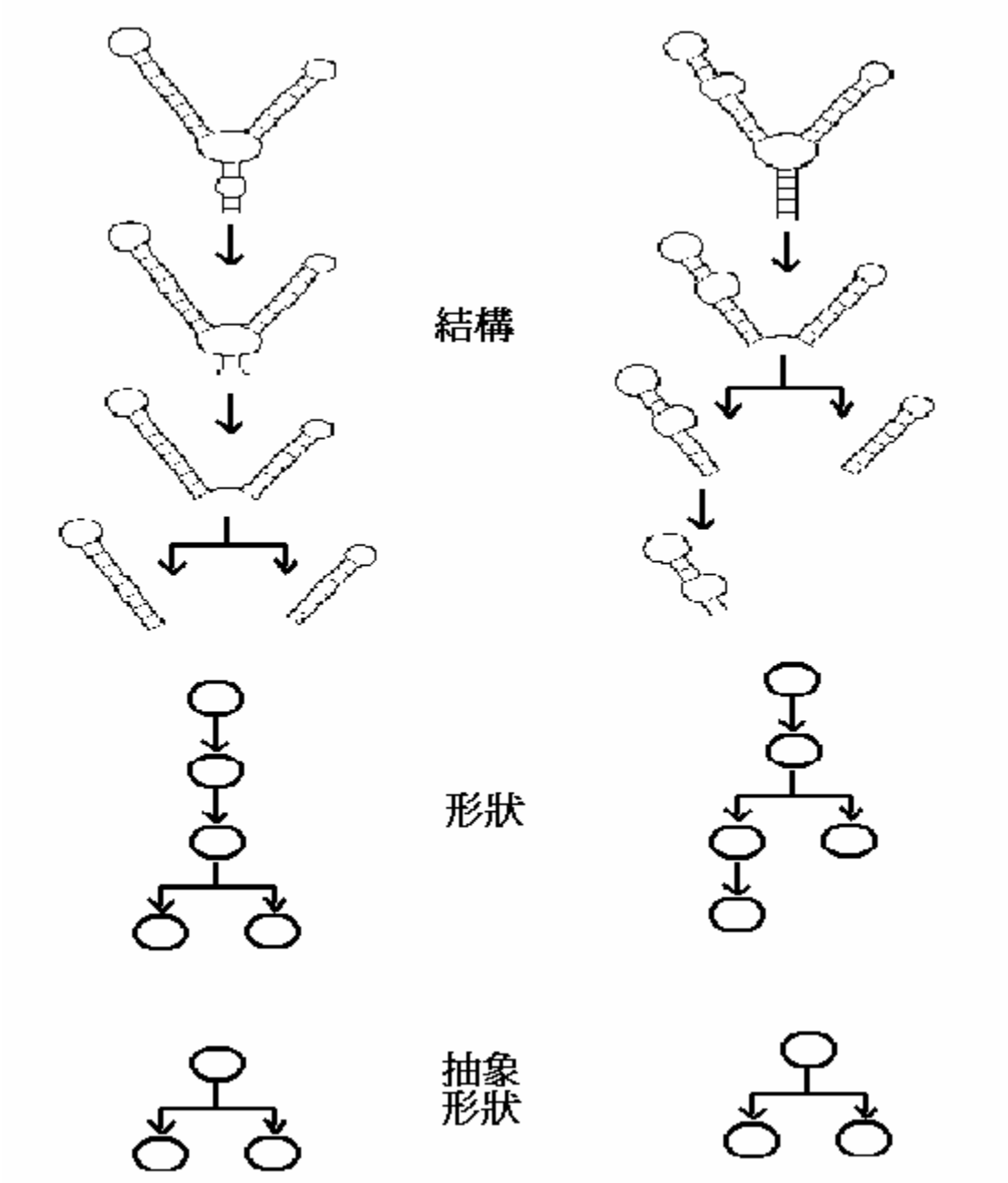
[1] 核醣核酸結構的形狀(Shape)：

「形狀」顧名思義就是一個物體忽略其尺寸大小後所留下的外形，而核醣核酸結構的形狀也是，忽略每個莖幹部位與非配對部份的大小，只考慮莖幹們之間的相對位置。換而言之，用 3.1.0 節中我們用來定義核醣核酸結構的理論來表示的話，就是只考慮一個核醣核酸結構中 min-SCC 子元件們彼此之間組合成整個核醣核酸結構的順序，而不考慮它們的大小。如下圖中左邊爲一個核醣核酸結構用 R-Grammar 分析出來的生成樹，而我們對形狀(Shape)的定義是以樹的資料結構表示，內容就是生成樹本身的簡化，忽略每個結點中的內容、邊上的內容，只留下樹的整體外形，如圖中右半邊所示。



[2] 核醣核酸結構的抽象形狀(Abstract Shape)：

核醣核酸結構在演化的過程中經常會因爲鹼基突變使得配對的鹼基分開了，而在結構中出現一些突起結構或是內部環狀結構，使得它們的形狀變的不一樣，但如下圖的中兩個核醣核酸結構，它們的形狀雖然不同，但只要拿遠一點看起來這兩個核醣核酸結構又好像相同，爲此我們設計了抽象形狀(Abstract Shape)來表示粗略觀察下的結構形狀，使得鹼基突變造成的變異被忽略掉。用我們定義的形狀來得到抽象形狀是非常簡單的，只要將[1]中所介紹的「形狀」中只有一個子結點的父結點與其子結點合併即可，因此抽象形狀依然是以樹的資料結構來表示。



3.1.2 計算相對長度差異演算法

在搜尋核醣核酸共同結構元的過程中，往往會用排比演算法來計算兩個核醣核酸結構爲同家族之共同結構元的可能性，但是在搜尋的過程中，難免會需要多達百萬次以上的機會需要做排比演算法，這會花費相當長的時間，因此我們設計了「相對長度差異 (Related Length Difference , RLD) 」演算法，在我們眼中 RLD 是最精簡的排比演算法，它建立在要進行比較的兩個核醣核酸結構的形狀或是抽象形狀相同的這個前提之下做運算，最重要的是我們能保證當 RLD 說兩個核醣核酸結構相似時它們不一定相似，但說他們不相似時則它們一定不相似，有了這層的保證我們可以知道，RLD 的目的並非取待傳統的排比演算法，而是擔任門檻的角色，在 RLD 說兩個結構之間可能相似度高時才交給排比演算法做排比；由於 RLD 幾乎是 constant time 演算法，因此能爲系統省下大量的時間，以下將詳細介紹。

在介紹演算法之前先闡述一段在實驗過程中的發現，如下表中(1)的部份是兩個我們想要做比較的核醣核酸結構，用現有較知名的核醣核酸結構排比演算法（如：RNAforest、RNAcomp…等等）做排比的結果常會出現如(2)中所顯示的結果，但在我們的眼中，這樣的排比結果來詮釋這兩個結構之間的相似關係並不恰當，然而如同(3)中所顯示的排比結果，雖然加入的 GAP 個數多上許多，但是看起來比較合理。爲何呢？試想如果 Sequence1 要演化到 Sequence2，如果由(3)來解釋，就是演化過程中失去了若干個鹼基，但由(2)來解釋則是失去右半部的鹼基並且左半部的配對關係也發生了變化，聽起來很明顯(2)的發生機率要低上許多。用我們之前介紹的理論來翻譯這個發現可解釋爲【任兩個形狀或抽象形狀相同的核醣核酸結構，它們生成樹上相同位置的 SCC 應該是相對應的】

(1) 原本的核醣核酸結構

((((.....)))).((((..)))) >Sequence 1

((((...)))).....((((....)))) >Sequence 2

(2) 一般的排比結果

((((.....)))).((((..))))

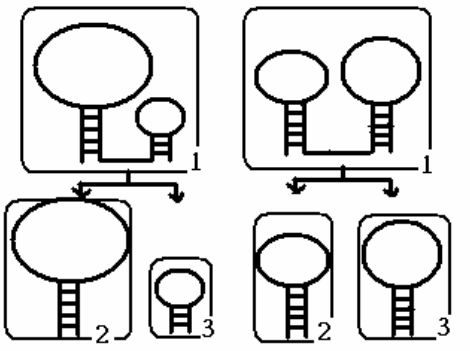
((((...)))).....((((....))))----------

(3) 比較符合生物意義的排比結果

((((.....)))).---(((.----.)))

((((..----------.)))).....((((....))))

，意思就如右圖所示，它是前面表格中兩個核醣核酸結構的生成樹，因爲它們的形狀樹是相同的，因此在相對應的結點中的結構就應該要互相對應(如右圖就是結點 1、2、3 內的 SCC 都應該要互相對應)。這是一個很重要的發現，我們也將會利用這個發現來設計 RLD 的計算，與下一節中將介紹的新的核醣核酸結構排比演算法。

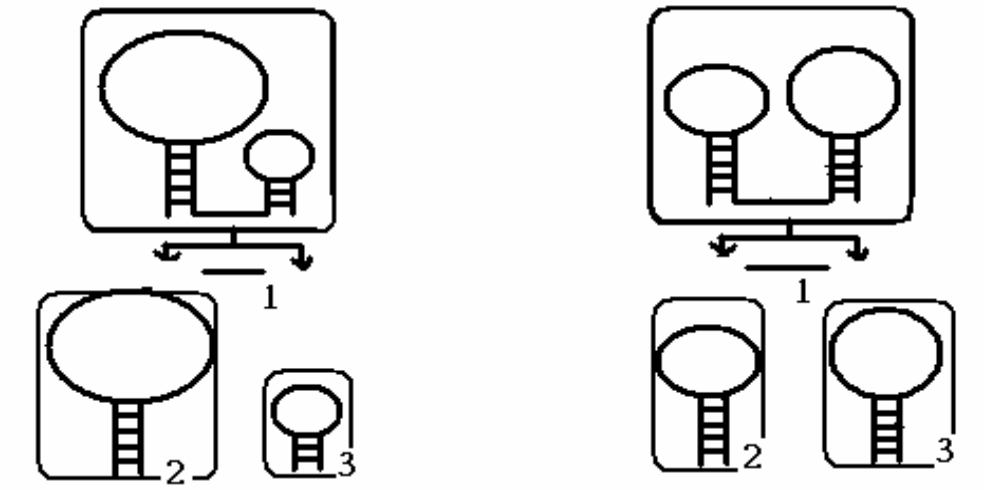


上圖中的生成樹爲了避免太雜亂而省去了一些邊上的應該要畫出來的 min-SCC，在下圖中我們畫出完整的生成樹，並且將編號放在構成核醣核酸結構的 min-SCC 們之上，當兩個結構的形狀樹相同時，則構成的 min-SCC 們必定也會互相對應，因此我們定義相對長度差異 (Related Length Difference，RLD)爲每個互相對應之 min-SCC 長度差平方之合再開根號，再除以兩個結構中最大的長度，式子爲：

$$RLD = （\Sigma(\text{SCC長度差})^2）^{1/2} / \text{Max-Length(結構 1、結構 2)}$$

如上頁中的例子就有三組 min-SCC 長度差值，分別爲 17、5、4，而最長的結構長度是 Sequence1 的 38，因此 RLD 爲 0.477。在此例中是討論形狀樹相同時的情況，當抽象形狀樹相同時，也是依此類推。

時間復雜度看的出來就與形狀樹的結點個數成正比，而不是與結構的長度相關，因此非常的快速，是個近乎於 constant time 的演算法。

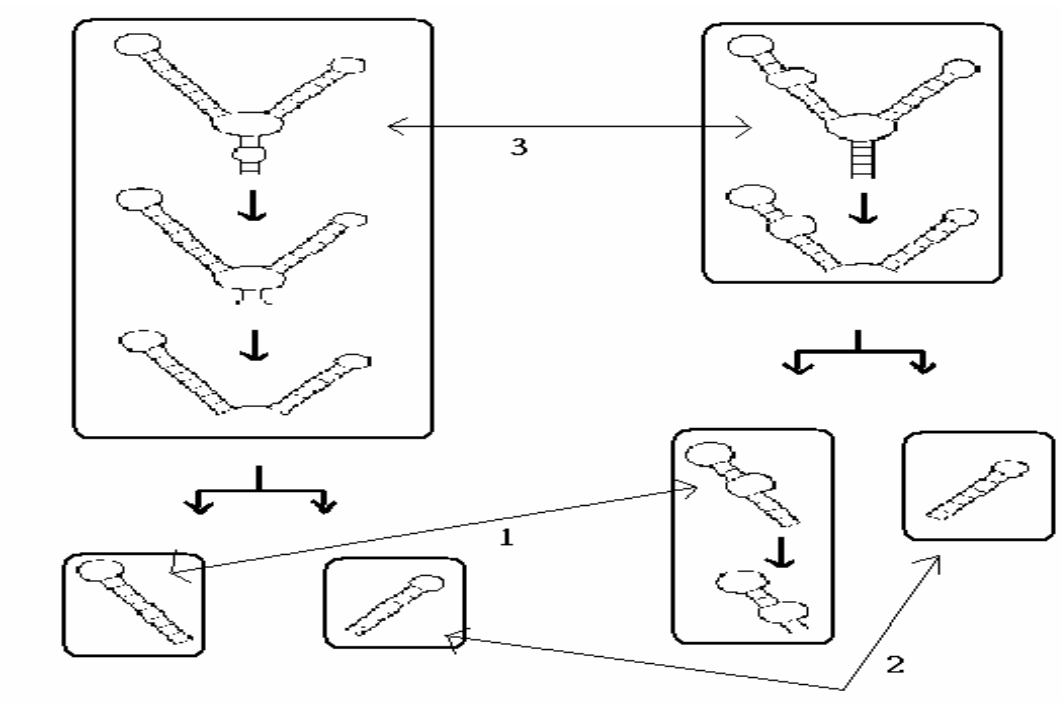


3.1.2 新的核醣核酸結構排比演算法

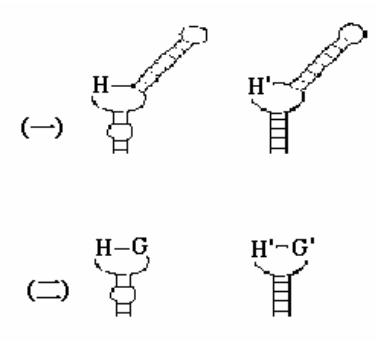
在上一節中敘述了一段現今大多排比演算法會有的不合理處，因此提出了【任兩個形狀或抽象形狀相同的核醣核酸結構，它們生成樹上相同位置的 SCC 應該是相對應的】這個想法來矯正這個不合理的現象，在這一節中我們將會把這個想法加入設計出新的排比演算法。這個條件的加入不但會避免掉上述的不合理處，而且還發現排比演算法可以只對結構中的 min-SCC 做排比，使得問題的難度下降，也使時間復雜度降低了。所以接下來會介紹兩大部份，一是如何決定 min- SCC 之間的排比順序，另一部份就是 min-SCC 之間要如何做排比；最後我們爲了 Gibbs-like 的流程設計了一個排比模型，因爲在做 Gibbs-like 的過程中，常會問一個結構 A 與另一組結構集合中的結構們的相似度有多高，則需要拿 A 去與集合中的每個結構一一做排比，但有了這個排比模型後，就只需做一次排比即可。接下來將詳細敘述之。

[1] 如何決定排比的順序

當兩個核醣核酸結構的抽象形狀樹相同時，我們會對每個相對應結點中的結構做排比，但是抽象形狀樹不像形狀樹，它的每個結點中可能會有一個以上的結構（因爲抽象形狀樹是由形狀樹對某些結點做合併而產生的），但我們永遠都只拿最上面的結構來比較就可以了，因爲最上面的結構一定是最大的。而在決定順序方面，排比演算法是由 Dynamic Programming 所設計出來的，因此在排比一個結構時，它的子結構一定要先被排比完，所以我們所設計的排比順序是由樹的樹葉部份開始，所以決定出來的順序會如下圖的編號所示。在每一個結點被排比



完之後，就會用一個字元（如：H 和 H’ 或 G 和 G’ 等等）來取待其父結點中的結構中自己的那一個部份，如右圖就是前一頁例子的取待順序，在 1 的部份排比完之後就會用 H 和 H’ 來取待，在 2 的部份排比完後，就會用 G 和 G’ 來取待，並且把 H、H’、G、G’ 這 4 個字元加到排比演算法要使用的分數表中供排比演算法使用，其中 H<->H’ 的分數就是 H 和 H’ 排比的分數，G<->G’ 也是，而和其它的字元對應分數都爲-∞，這樣子就能保證在做第 3 部份排比時，排比演算法會把 H 和 H’ 對在一起、G 和 G’ 對在一起。由於一個結構在排比前，其子結構一定會排比完並用一個字元取待之，所以在排比每一個結構時，它一定會是一個 min-SCC（因爲它不可能含其它的 SCC，都被取待了），因此我們只需設計出能排比 min-SCC 的演算法即可。



[2] 比較兩個 min-SCC 的排比演算法

現有的核醣核酸排比演算法除了有3.1.2節中提到的不合理外，其實還有另一個不合理的情況如右表所示，(一)上半部中的兩個結構其實不需要再加入任何的GAP就已經排比好了，但常會被排比成(一)下半部的情況，原因出在它們都不准許「.」和「（或）」排在一起，因爲如果想要可以排在一起的話，問題的復雜度會升高，但如果簡單的用Needleman-Wunsch的排比演算法去排比的話又會有(二)中的問題，上半部的結構被排比成像下半部的結果，這是不可能出現的配對方式，基對的配對完全跑掉了。但我們還是希望能在准許「.」和「（或）」排在一起的情況下，依然設計出時間復雜度相同的排比方法，且排比出來的結果是合理的，爲了這個目的，我們設計了一個逼近演算法來達成這件事，其策略是由Needleman-Wunsch的排比方式出發做一些修改，對於它排比出的不全理結果，再設計一個平衡（Balance）演算法做調整，雖然我們無法保證結果會是最佳解，但能達到之前提出的三點要求。所以這個演算法對輸入的兩個核醣核酸結構分成三個階段分別處理：

（一）	(((.....))) ((((....))))
	-(((.....)))- ((((-....-))))
（二）	((...)) (...)
	((...)) (-...)-

<1>使用三次 Needleman-Wunsch 做初步排比

<2>對前面的結果做平衡(Balance)

<3>對平衡後的結果再做分斷排比

<1> 使用三次 Needleman-Wunsch 做初步排比：

Needleman-Wunsch爲DP演算法，所以它會需要使用到評分表，我們所使用的評分表如右表所示，這個評分表有個特點是在於它會動態變大，如[1]中所說，在用字元取代某個SCC之後會把字元加到評分表中，如右圖中的H和H’在一開始時是不存在的，是有取代動作出現時才加進來的。第二個特點是，排比的過程並非只考慮核醣核酸序列或只考慮核醣核酸結構，而是一起考慮，所以會需要這兩個評分表M_{seq}和M_{str}。

M_{seq}	A	C	G	U	-	H	…
A	1.0	.25	.25	.25	-1	-∞	
C	.25	1.0	.25	.25	-1	-∞	
G	.25	.25	1.0	.25	-1	-∞	
U	.25	.25	.25	1.0	-1	-∞	
-	-1	-1	-1	-1	1	-∞	
H’	-∞	-∞	-∞	-∞	-∞		
:							

M_{str}	(	.	)	-	H	…
(	1.0	.50	-1	-1	-∞	
.	.50	1.0	.50	-1	-∞	
)	-1	.50	1.0	-1	-∞	
-	-1	-1	-1	-1	-∞	
H’	-∞	-∞	-∞	-∞		
:						

爲了讓排比演算法能同時考慮核醣核酸的序列和結構，我們將它的遞迴式設計如下，

$$Sim(i,j)=\begin{cases} Sim(i,j-1)+gap \\ Sim(i-1,j-1)+pair(i,j) \\ Sim(i-1,j)+gap \end{cases}$$

$$gap = w_1 \cdot gap_{struct} + w_2 \cdot gap_{seq}$$

$$pair(i,j) = w_1 \cdot match_{struct}(i,j) + w_2 \cdot match_{seq}(i,j)$$

其中w_1和w_2分別爲使用者對系統下的參數，分別表示對序列和結構所指定的權重，而$match_{seq}(i, j)$表示第一個結構的第i個鹼基和第二個結構的第j個鹼基在M_{seq}中對應的分數。

<2> 平衡(Balance)

對於由<1>排比好的結構，可以用 3 個動作把它做好平衡：

一. 加入 GAP 時得 2 個左括對齊，再找到它們的右括，也一樣加入 GAP 使他們的右括對齊。

二. 當有其中一條已經對完時，但另一條還有鹼基未對齊時，一樣加入 GAP 使它們長度相同。

三. 去掉結構中同時爲 GAP 的位置。

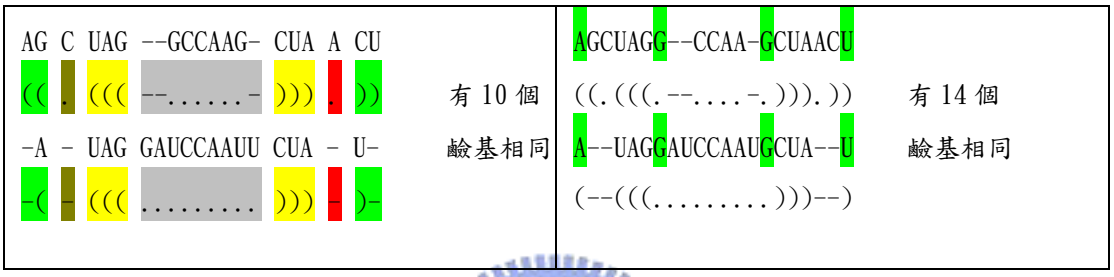
下表爲一個例子：

((.(((-.....))).)) ((-((.....-.))))		
一	((.(((--.....))).)) ((-((.....-.))))	((.(((--.....--))).)) ((-((.....-.)))))
一	((.(-((--.....--))).)) ((-((.....-.))))	((.(-((--.....--)))).)) ((-((.....-.)))))
一	((.(-((--.....--))).)) (-(-((.....-.))))	((.(-((--.....--))).)) (-(-((.....-.)))-)
二	((.(-((--.....--))).)) (-(-((.....-.)))-)	((.(-((--.....--))).)) -(-(-((.....-.)))-)-
三	((.(-((--.....--))).)) -(-(-((.....-.)))-)-	((.(((--.....-))).)) -(-(((.....)))-)-

<3> 對平衡後的結果再做分斷排比：

平衡後的結果雖然合理，但我們的目的並非求得合理的排比即可，還要儘可能是相似分數最高的排比，但平衡的動作並不保證這一點，因此才有第<3>個動作，要提高平衡後的排比分數。

因爲之前平衡動作已經把結構對好了，那就表示兩個本來形狀不同的核醣核酸結構，因爲加入 GAP 的關係而使得它們的形狀也想同了，而後只要將每塊莖幹和環結中的鹼基再配對排比就可以了，如下圖左中，用顏色分開要排比的區塊，對每個區塊做排比後，就能得到下圖左中更高的相似度，本來對應到相同的鹼基有 10 個，而做後就有 14 個。



在做完以上所說的三個步驟之後，雖然不能保證最後能得到最佳的排比結果，但是整個排比演算法所花的時間復雜度能保持在$O(n^2)$，排比出來的結構也是合理的核醣核酸二級結構，而且最重要的是我們用這個排比演算法幫已知家族的共同二級結構元做排比，能得到相似度非常高的結果（如下表爲其中一個例子），這個就表示這個排比演算法對同家族的共同二級結構元的辨識度是非常高的。

```
C-CUCAUAUAAUCUUGGGAAUAUGGCCCAAAAGUUUCUACCUGCUGACCGUAAAUCGGCGGACUAUGGG-G
(-((((((...(.(((((.......)))))).).........(((((((.......)))))))..)))))-)
UACUUGUAUAAUAGUUGCGAU-UGGGCGACGAGUUUCUACCUGGUUACCGUAAAUAACCGGACUAUGAGUA
(((((((...(.(((((...-...)))))).).........(((((((.......)))))))..)))))))
UACUUAUAUAAUAGUUGCGAU-UGGGCGACGAGUUUCUACCUGGUUACCGUAAAUAACCGGACUAUGAGUA
(((((((...(.(((((...-...)))))).).........(((((((.......)))))))..)))))))
CACUCAUAUAAUCGCGUGGAUAUGGCACGCAAGUUUCUACCGGGC-ACCGUAAAU-GUCCGACUAUGGGUG
(((((((...(.(((((.......)))))).).........(((((-(.......)-)))))..)))))))
UUUUCAUAUAAUCGCGGGGAUAUGGCCUGCAAGUUUCUACCGGUUUACCGUAAAUGAACCGACUAUGGAAA
(((((((...(.(((((.......)))))).).........(((((((.......)))))))..)))))))
CACUCAUAUAAUCGCAAGAAUACGGCUUGCGAGUCUCUACCAGCCGACCGUAAAUCGGCUGACUAUGAGUG
(((((((...(((((((.......)))))))........(((((((.......)))))))..)))))))
CACUCGUAUAAUCGCGGGAAUAGGGCCCGCAAGUUUCUACCAGGCUGCCGUAAACAGCCUGACUACGAGUG
(((((((...(.(((((.......)))))).).........(((((((.......)))))))..)))))))

uacUCgUAUAAuc###ggaAUAUGG#c####aGUuUCUACC#gg#uaCCGUAAAu##cC#GACUAuGag#g
(((((((...(((((((.......)))))))........(((((((.......)))))))..)))))))
```

[3] 排比模型：

在 Gibbs-like 的流程中，經常會需要拿一個結構去與種子中的每個結構做排比，這是很花時間的，因此我們設計了一個「排比模型」來表示種子中的結構們，而後有需要和此種子內全部的成員做排比的時候，就只需要和這個「排比模型」做一次排比就可以了。

「排比模型」的內容是用兩個矩陣來表示種子中成員們在做完多重排比後的內容，一個矩陣是存序列內容，另一個是存結構內容，用統計的方式計算每個位置中｛ A , C , G , U ｝或是｛ (, . ,) ｝所占的百分比，而後如果有結構要跟這個種子排比時，就只需要與這兩個矩陣做排比就可以了，下表爲一例。

```
GCAUCCCUUUUGGGAGC
((..(((....))).))
GGA--CCUUUUGGCACC
((.--((....))..))
GC-UGCC-UUUGGCAGC
((-((((-...))))))
```

M_S=

	(	)	.	-
1	1.00	0.00	0.00	0.00
2	1.00	0.00	0.00	0.00
3	0.00	0.00	0.67	0.33
4	0.33	0.00	0.33	0.33
5	0.67	0.00	0.00	0.33
6	1.00	0.00	0.00	0.00
7	1.00	0.00	0.00	0.00
8	0.00	0.00	0.67	0.33
9	0.00	0.00	1.00	0.00
10	0.00	0.00	1.00	0.00
11	0.00	0.00	1.00	0.00
12	0.00	1.00	0.00	0.00
13	0.00	1.00	0.00	0.00
14	0.00	0.67	0.33	0.00
15	0.00	0.00	1.00	0.00
16	0.00	1.00	0.00	0.00
17	0.00	1.00	0.00	0.00

M_B=

	A	G	C	U	-
1	0.00	1.00	0.00	0.00	0.00
2	0.00	0.33	0.67	0.00	0.00
3	0.67	0.00	0.00	0.00	0.33
4	0.00	0.00	0.00	0.33	0.67
5	0.00	0.33	0.33	0.00	0.33
6	0.00	0.00	1.00	0.00	0.00
7	0.00	0.00	1.00	0.00	0.00
8	0.00	0.00	0.00	0.67	0.33
9	0.00	0.00	0.00	1.00	0.00
10	0.00	0.00	0.00	1.00	0.00
11	0.00	0.00	0.00	1.00	0.00
12	0.00	1.00	0.00	0.00	0.00
13	0.00	1.00	0.00	0.00	0.00
14	0.00	0.33	0.67	0.00	0.00
15	1.00	0.00	0.00	0.00	0.00
16	0.00	0.67	0.33	0.00	0.00
17	0.00	0.00	1.00	0.00	0.00

「排比模型」是有品質優劣的，評分依據是矩陣中每個位置相似度的平均值，再去乘上排比演算法中的評分表，其式子如下。

$$Score\ (M) = 1/L \cdot \sum_{k=1}^{L} w_1 \cdot StructSim\ (k) + w_2 \cdot SeqSim\ (k)$$

$$StructSim\ (k) = \sum_{i \in \{(,),.,\ -\}} \sum_{j \in \{(,),.,\ -\}} M_S(k,i) \cdot M_S(k,j) \cdot StructMatch\ (i,j)$$

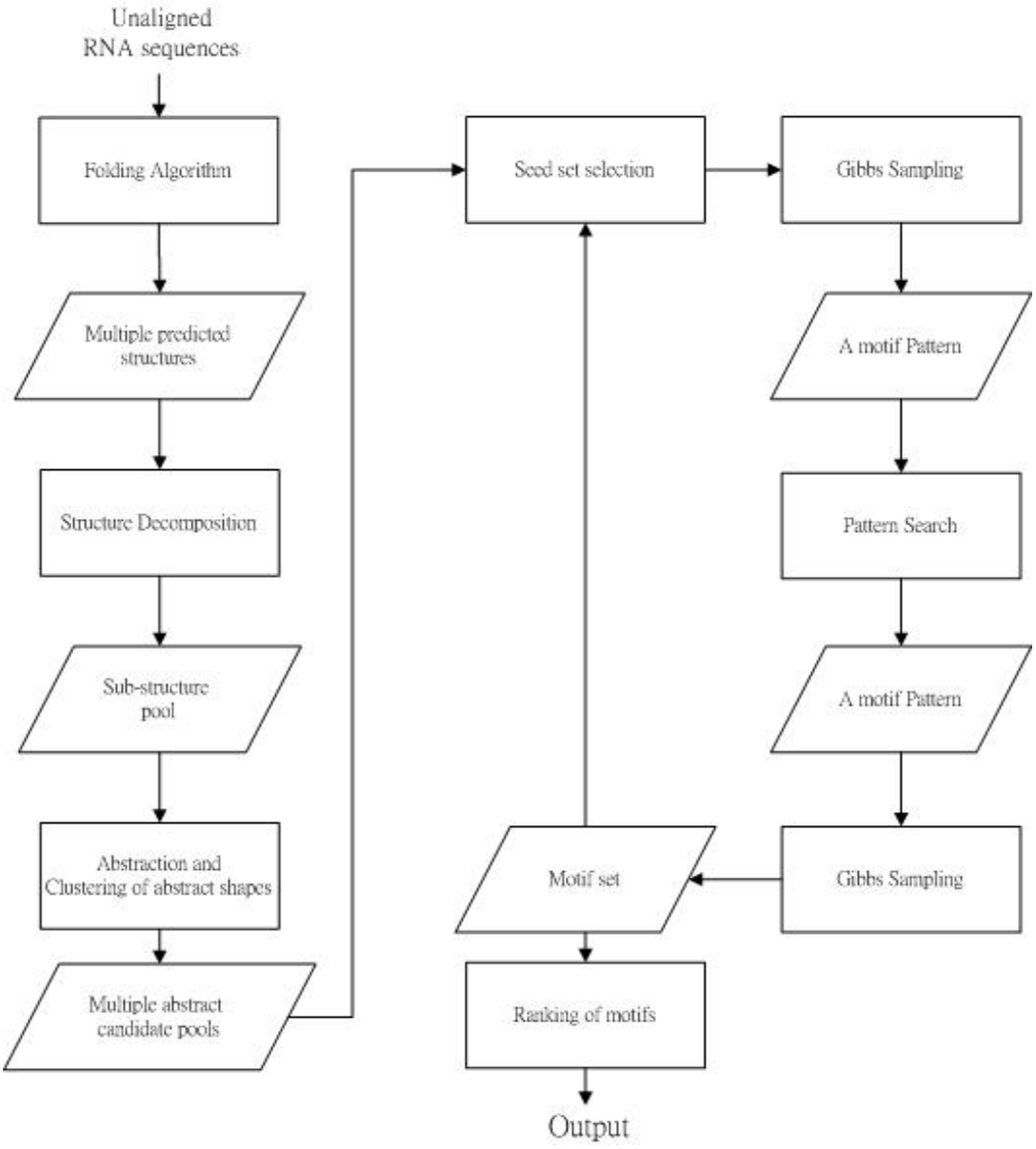
$$SeqSim\ (k) = \sum_{i \in \{A,G,C,U,-\}} \sum_{j \in \{A,G,C,U,-\}} M_E(k,i) \cdot M_E(k,j) \cdot SeqMatch\ (i,j)$$

3.2 核醣核酸共同結構元搜尋系統

系統的流程圖如下圖所示，整體來看會分爲 2 大部份：

(1) 因爲本系統是使用 Gibbs-like 解搜尋核醣核酸共同結構元的問題，因此第一部份就是在爲 Gibbs-like 建立多個「抽象形狀候選者池（Abstract Shape Candidate Pools）」，以供 Gibbs-like 使用。

(2) 接下來就是 Gibbs-like 流程的部份，每個抽象形狀候選者池都會選出多個初使化種子，而後每個種子再經過 2 次 Gibbs-like 而得到具有相似結構的共同結構元集合。最後對所有的共同結構元集合做排序後輸出給使用者。



系統流程圖

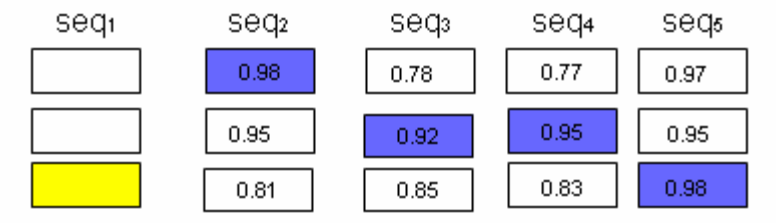
3.2.1 產生核醣核酸共同結構元候選者集合

在接收到資料序列集時，會先由 RNAFold 預測出每條核醣核酸序列的結構，要注意的是，RNAFold 對於過長的序列的預測結果不佳，而且預測時間也很久，所以我們不會直接將整條序列送給 RNAFold 做預測，而是會對序列做分段輸入，每次輸入的長度只有 1.5 倍的「max candidate length(使用者設定)」，而 sliding window size 爲 0.5 倍的「max candidate length(使用者設定)」。接下來經由分解演算法將核醣核酸結構中所有共同結構元候選者分解出來，最後將每個候選者依造抽象形狀分類得到多個「抽象形狀候選者池（Abstract Shape Candidate Pools）」，以供 Gibbs-like 使用。

讀者現在一定會好奇，爲什麼要用抽象形狀來做分類呢？這其實是一個觀察出來的現象，在我們目前所見到的同家族的共同結構元，它們的抽象形狀都是相同的，當然這也是爲何我們要設計”抽象形狀”這個核醣核酸表示法的原因，所以我們將它做爲系統的實驗假設，而且也在實驗結果中得到很不錯的結果，如此初步分群後的結果能大大的縮小系統的搜尋空間，加速搜尋的時間。

3.2.2 如何選擇種子

其實 Gibbs-like 是可以隨機選取種子的，但是爲了能讓結果快速收斂，所以我們決定先選擇一些相似的結構們來當種子，這是選取種子的目的，但難度在我們如何找”相似”的結構呢？用排比演算法來計算並不划算，因爲我們只是在選種子而已，所以決定只用 RLD 來計算兩兩結構之間的相似度，並且將相似度很高的結構們集合起來當成是一個種子。如下圖的例子是用 Seq1 最下面的 candidate 做爲中心，到其它的序列中將與自己相似度最高的 candidate 集合起來，就可以形成一個種子。在候選者池中的每個候選者都要做一次這個動作，去收集與自己高度相似的結構，因此會有很多個種子，最後以「平均相似度」做排序後，輸出給 Gibbs-like 做處理。



3.2.3 Gibbs-like 流程

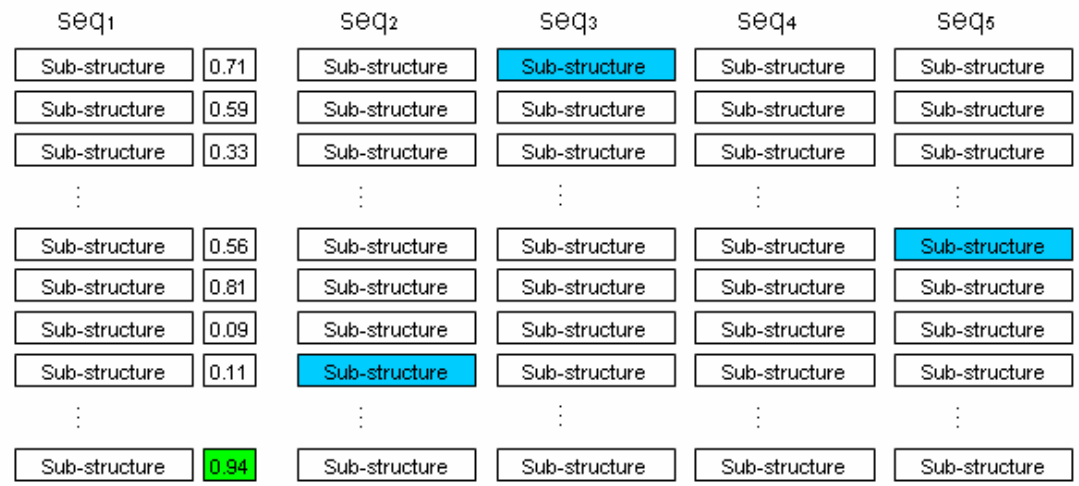
n：測試序列條數
S：$(s_i)_{1 \le i \le n}$，序列集合
T：$(t_i)_{1 \le i \le n}$，種子，其中$t_i = \phi$表示種子中沒有第i條序列的成員。
C：$(c_{ij})_{1 \le i \le n}$，候選者集合，c_{ij}表示s_i的第j個候選者
C_i表示s_i的所有候選者集合。

M_T：　由T所建立出來的排比模型。
sim(T)：　種子T的內成員們之間的相似度，定義爲Score(M_T)。
sim(c_{ij}, T)：　候選者與種子T之間的相似度，定義爲c_{ij}與M_T排比分數。

(文中會用到符號的定義)

Gibbs-like 這個方法的目的，就類似是要在 k-partite graph 中找 max-Clique，它的步驟我們由下圖來說明：

(1) 如下圖中被標記起來的Sub-structure就是種子的成員，現在我們計算C_i中每一個候選者與種子的相似度，並且選擇相似度最大者加入種子中，如下圖就是選最下面那個候選者（相似度爲 0.94）加入種子中。要注意到的是，計算sim(c_{ij}, T)時，若RLD的值小於θ_1（RLD的門檻值，由使用者設定）則sim(c_{ij}, t_k)會直接設定爲 0 而不用排比演算法計算相似度，用以限定Gibbs的搜範圍。
(2) 如上述的方法，對每一條序列都處理過一次後稱爲一個回合（Iteration）。
(3) 如果某個回合在執行前和執行後，種子的內容是不變的，即爲收斂停止。



Gibbs-like 在每個回合的過程中有以下的特性：

(1) 目的是使 sim(T)的值最大化。

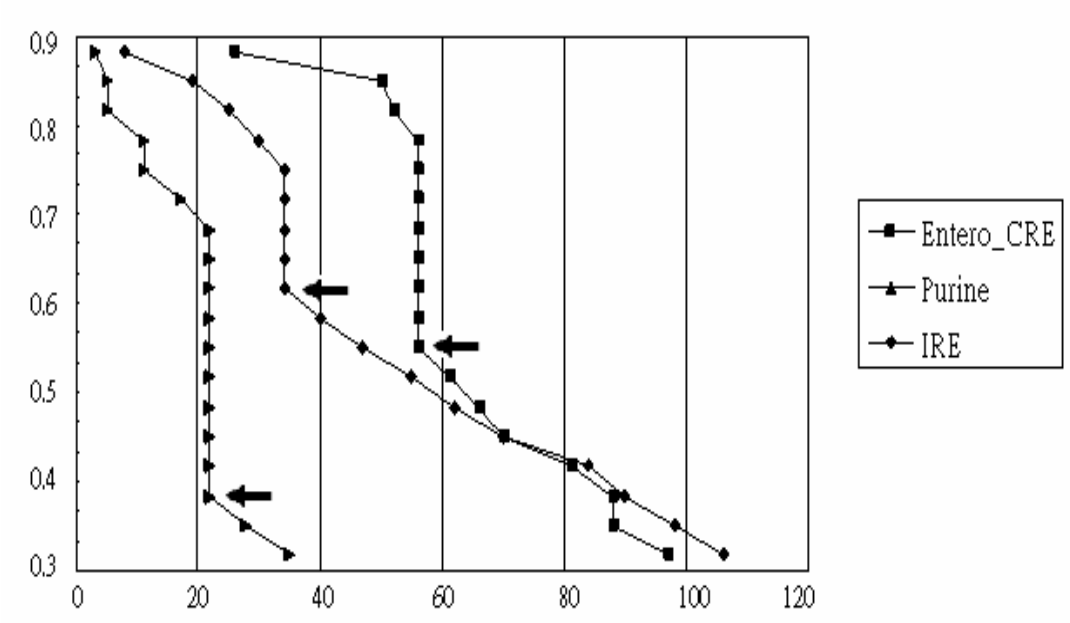
(2) 每條序列必定會選擇一個與種子相似度最高者加入種子。

如(1)中所描敘使 sim(T)值最大化表示 Gibbs-like 會從種子出發，試著找其它的候選者加入使得找到的種子集合變大，並且依造 RLD 來限制 Gibbs，在計算 sim(C_i, T)的時候，若與 T 中暫定共同結構元的 RLD 小於 θ_1 的候選者則計算相似度，否則相似度直接設定爲 0，那麼 Gibbs-like 就只會嘗試選擇可能在 T 附近的候選者來比較，以避免做不必要的搜尋。

而(2)中說出了一項 Gibbs-like 的缺點，當一組資料中，並非每一條序列都是屬於同一個家族時(就是有些它們沒有共同結構元)，則這幾條序列對 Gibbs-like 來說就是雜訊，因爲 Gibbs-like 一定會幫每條序列選擇一個候選者加入。

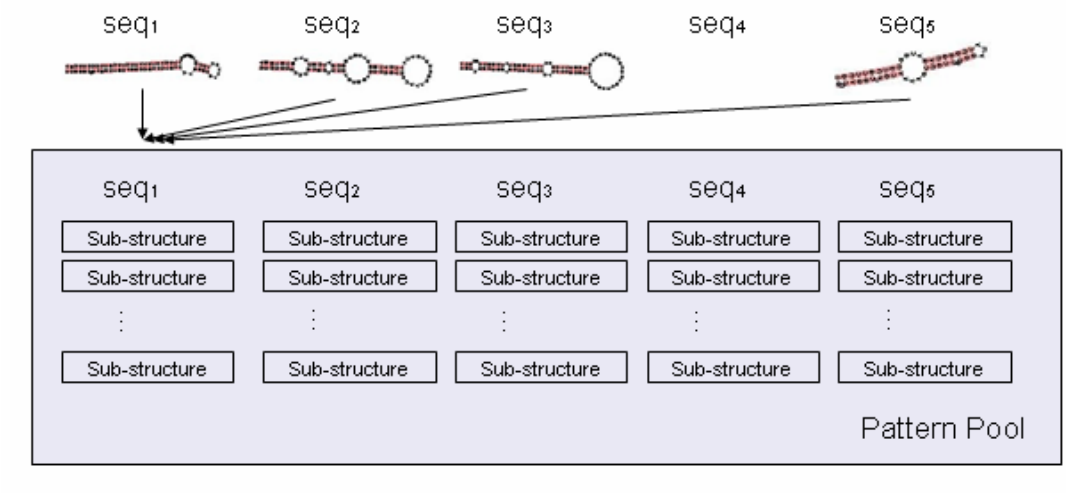
解決方法是設定相似度的門檻值(θ_s)，在爲C_i選擇加入種子的候選者時，如果其中沒有任何一個候選者與種子的相似度高於 θ_s 則不選，t_i 選定爲 ϕ。

但是由實驗的結果看來，直接設定一個 θ_s 是可以做到過濾雜訊的功用，但效果並不佳，所以後來決定 θ_s 會由一個最大值一直遞減 0.02，直到 0，再由系統去選要停在那一個值會比較好，而最大值就是用種子中最小的成員相似度，結果就如下圖所示，在遞減的過程中會有一段蠻長的”穩定區間”，系統就會選擇這時的種子爲結果，也表示後來再進來的就被視爲是雜訊。



3.2.4 樣版搜尋(Pattern Search)和第二次 Gibbs-like

樣版搜尋(Pattern Search)就是拿一個種子去搜尋測試資料，將每條序列中所有能折疊出與種子中任何一個成員相同的子序列都搜尋出來，然後構成新的候選者池，如下圖一個種子中有 4 個成員，拿來搜尋Seq1 後，所有搜尋到的Sub-structure都存在C_1中，如此搜尋完全部的序列後，就可以可以形成新的候選者池，對於新的候選者池再跑一次 3.2.3 中所敘述的Gibbs-like即可，得到新的共同結構元。



爲何要做第二次的 Gibbs-like 呢？想想看在第一次 Gibbs-like 時，爲何有些序列沒有找到共同結構元呢？有以下 2 種可能性：

(1) 這條 sequence 並非同一個家族(它沒有與其它序列相同的共同結構元)。

(2) 從結構預測程式接收到的結果中沒有正確的二級結構。

如果是(1)，則沒有找到是對的，但如果是(2)的話，樣版搜尋(Pattern Search)就能可能將答案找回來，因爲共同結構元之間的結構是很相似的，那麼用已找到的結構去掃瞄沒有共同結構元的序列，就會有可能找到，而能簡易的擺拖只依懶結構預測程式來預測序列的結構的現象。

3.2.5 結果

下表爲系統的結果範例，其中第一行是這組共同結構元排序後的排名，第二行中是它的相似度，而後每一行各代表一個共同結構元，第一個爲序列核醣核酸序列的，第二個##中的數字爲共同結構元在序列中的起點與終點，最後是共同結構元的結構。

```
# Rank 1 ####
SS_Similarity : 0.85972878335044
AB062402.1/11-40          ##(44-71):0.7718689356620393          (((.((.(((((.....)))))))))).
AB073371.1/5-34           ##(142-167):0.7668859450468646        (.(((.(((((.....))))))))).
AC073115.5/47515-47486    ##(71-96):0.7190369230599117          (((...(((((.....))))).))).
AF117958.1/132-161        ##(29-54):0.7353796703166375          ((.((.(((((.....)))))))))
AF171078.1/1416-1442      ##(33-58):0.7026030589248979          (.(((.(((((.....)))))))))
AF266195.1/14-43          ##(62-89):0.7703516913194333          (((.((.(((((.....)))))))))).
AF285177.1/3-32
AF338763.1/11-40          ##(3-31):0.71486837226548             (...(((.(((((.....))))))))).)
AJ426432.1/1593-1619      ##(100-124):0.688661836248043         (((..(((((.....))))).))).
AJ426432.1/1658-1684      ##(9-33):0.7231902495120888           (((..(((((.....))))).))).
AL034379.8/68035-68064    ##(77-104):0.7740334378265413         ((.(((.(((((.....)))))))))).
AL355837.6/87643-87614    ##(62-89):0.7502985520226898          ((.(((.(((((.....)))))))))).
AL513423.3/108544-108573 ##(19-46):0.7394387221973429           ((.(((.(((((.....)))))))))).
AP003174.2/91762-91734    ##(99-125):0.744721283592757          ((((..(((((.....))))).)))).
AY120878.1/50-76          ##(9-31):0.7108776973249286           (((.(((((.....)))))))).
BC001188.1/3791-3817      ##(59-82):0.7321221307428204          (.((.(((((.....)))))))).
```

第四章、實驗結果

4.1 實驗評估標準

與目前多數預測二級結構的研究一樣，我們採用 Matthews 的相關係數(Matthews correlation coefficient)來做爲評估的標準。其原始定義如下：

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

其中P_t爲正確正預測的總鹼基對數(true positive)，即系統預測到的鹼基對也出現在正確答案中的鹼基對數；P_f爲錯誤正預測的總鹼基對數(false positive)，即系統預測到的鹼基對沒有出現在正確答案中的鹼基對數；N_t爲正確負預測的總鹼基對數(false negative)，即沒有出現在預測結果上也沒有出現在正確答案上的鹼基對數；N_f爲錯誤負預測的總鹼基對數(false negative)，即沒有出現在預測結果上卻有在正確答案上出現的鹼基對數。

化簡後的式子如下：

$$C \approx \sqrt{\frac{P_t}{P_t + N_f}\frac{P_t}{P_t + P_f}}$$

其中 $\frac{P_t}{P_t + P_f}$ 表示系統的正確率(precession)，又稱爲選擇性(selectivity)，

而 $\frac{P_t}{P_t + N_f}$ 表示系統的擷取率(recall)，又稱爲敏感性(sensitivity)。

4.2 實驗測試資料

下表爲我們所準備的測試資料，共有 13 組，資料的來源是到 Rfam 中下載共同結構元的檔案，再隨機爲每條序列做延伸，使得共同結構元在序列中所暫的比例降低，以提高搜尋的難度。

我們的實驗是使用以下的測試資料設計出多個不同的主題，有(1)單一家族測試、(2)多個家族測試、(3) 龐大資料量測試…等等，其中我們要比較的除了 MCC 之外，還有系統的速度與可載量，並在結果中會發現，我們的系統有不輸其它系統的 MCC 準確值，而且因爲提出新的核醣核酸表示法，並延伸出多個時間復雜度低的分析程式，而使得我們的程式在執行時間上，以及可處理的資料量上，能大大的勝過其它的系統，更能勝任實際上核醣核酸分析實驗的情況，以下會一一介紹。

	Sequence Avg. Length	Motif Avg. Length	M-Len / S-Len	Sequence Number	Motif Abs-Shape
ctRNA_pGA1	300.29	62.5	0.208	15	[][]
Entero_CRE	261.89	39.0	0.148	56	[]
glmS	497.57	154.8	0.311	11	[][][][]
HepC_CRE	173.34	48.0	0.276	52	[]
Intron_gpII	285.74	69.0	0.241	114	[][]
IRE	139.09	29.1	0.209	39	[]
let-7	400.33	80.9	0.202	14	[]
lin-4	311.88	68.8	0.220	9	[]
Lysine	381.40	172.2	0.451	43	[[][][][]]
mir-10	374.09	71.2	0.190	11	[]
Purine	193.91	68.9	0.355	22	[[][]]
s2m	162.92	41.0	0.251	38	[]
SECIS	164.62	64.3	0.390	43	[]

4.2 實驗結果

我們的系統是以 Java 實作而成，測試環境的作業系統爲 Mandrake Linux 10.1，電腦配備爲 Pentium IV 3.2GHz 的中央處理器與 512MB 的記憶體。比較的對象爲目前 4 個最有效的核醣核酸分析系統：FoldAlignM 、CMfinder、RNAshape、MARNA。

[1] 單一家族 MCC 測試結果：

	Our	CMfinder	RNAshape	FoldAlign	MARNA
ctRNA_pGA1	0.942	0.933	0.866	0.936	0.871
Entero_CRE	0.940	0.951	0.872	0.922	0.728
glmS	0.723	0.805	0.415	0.747	0.444
HepC_CRE	0.985	0.998	0.907	0.970	0.663
Intron_gpII	0.821	0.792	0.731	0.804	0.542
IRE	0.851	0.902	0.625	0.827	0.504
let-7	0.788	0.841	0.647	0.765	0.611
lin-4	0.711	0.783	0.774	0.720	0.761
Lysine	0.824	0.883	0.755	0.767	0.687
mir-10	0.683	0.720	0.441	0.694	0.317
Purine	0.882	0.910	0.588	0.892	0.795
s2m	0.863	0.845	0.341	0.850	0.384
SECIS	0.657	0.708	0.540	0.719	0.588
Avg	0.821(2)	0.851(1)	0.654(4)	0.806(3)	0.607(5)

[2] 多個家族 MCC 測試：

在多家族測試中，因爲資料量會是單一家族實驗中的倍數，因此在本實驗中不只會計錄 MCC 的測試值，還會將將執行的時間一併納入。

1. Abstract Shape Different：這組資料是要測試當測試資料中家族共同結構元的抽象型態都不同時，各個系統的實驗狀況。

	Our	CMfinder	RNAshape	FoldAlign	MARNA
Entero_CRE(56)	0.864	0.866			
ctRNA_pGA1(15)	0.942	0.933			
SECIS(43)	0.611	0.708			
Lysine(43)	0.796	0.883			
CPU time	1”54	6”24			

2. Big Family Size：這組資料是要測試當測試資料量很大時（本實驗共有 203 條序列），各個系統的實驗狀況。

	Our	CMfinder	RNAshape	FoldAlign	MARNA
Intron_gpII (95)	0.807	0.776			
Entero_CRE(56)	0.907	0.930			
HepC_CRE(52)	0.985	0.998			
CPU time	3”27	12”41			

3. Abs-Shape Equal：這組資料是要測試當測試資料中家族共同結構元的抽象型態都不同時，各個系統的實驗狀況。

	Our	CMfinder	RNAshape	FoldAlign	MARNA
s2m(38)	0.844	0.813			
HepC_CRE(52)	0.985	0.998			
Entero_CRE(56)	0.881	0.922			
mir-10(11)	0.541	0.660			
CPU time	2”09	6”03			

4. Small Family Size：這組資料是要測試當測試資料中家族中序列的個數都很小時，各個系統的實驗狀況。

	Our	CMfinder	RNAshape	FoldAlign	MARNA
Lin-4(9)	X	0.751			
glmS(11)	0.723	0.805			
Mir-10(11)	0.582	0.720			
ctRNA_pGA1(15)	0.937	0.933			
Let-7(14)	0.731	0.839			
CPU time	0”21	0”47			

[3] 龐大資料量分析：

	Our	CMfinder	RNAshape	FoldAlign	MARNA
ctRNA_pGA1	0.834				
Entero_CRE	0.813				
glmS	0.625				
HepC_CRE	0.937				
Intron_gpII	0.721				
Lysine	0.672				
mir-10	0.653				
Purine	0.802				
s2m	0.813				

(共 350 條序列)

4.3 實驗結果分析

[1] 精確度分析：

在單一家族測試的實驗中，平均 MCC 我們的系統是排名第二，但其實前三名的系統 MCC 差距都是很小的，可以說結果是幾乎相同的，但 CMfinder 確實是容易在 MCC 上能有較好的結果，因爲它不只是依賴結構預測程式所預測的二級結構，還在它所用的 Covariance Model 中對結構做微調，使的找到的共同結構元的相似度更高，所以比起我們的系統是單純的依賴結構預測程式所預測的結構，它能有更高一點的預測結果，這方面也是我們系統未來能再加強的部份。

另外，現有的系統大多都是用 MFE 來計算預測核醣核酸的二級結構，而大多搜尋結構元的系統都是用它們來決定核醣核酸序列的二級結構（我們的系統、CMfinder、RNAcast、MARNA…等等都是），用 MFE 來預測核醣核酸的二級結構的結構，並非都能有 8~9 成以上的準確度，所以這些依賴它的系統所能找到最好的 MCC 當然也不可能高過結構預測程式的準確度，使得有些家族的 MCC 都不能到 8 成，這並非共同結構元搜尋系統的效能不好，而是沒有收到準確度高的二級結構的關係，而我們的系統也受到這個限制。

[2] 執行速度分析：

在多個家族測試中，因爲測試資料會倍增，使得 FoldAlignM 這三個系統會跑相當久的時間，一般都要數到到數個星期都有可能，而 RNAshpae、MARNA 也是要跑好幾天，而且出來的結果也是錯的，所以表格中只有 CMfinder 有測試結果，這個部份，我們將只和 CMfinder 做討論。

在整體的時間復雜度計算上，CMfinder的時間復雜度爲$O(M^2 L^2+k_1 N|Q| L^3)$，而我們的系統則是$O(k_2 M L^2)$，其中N爲序列條數、M爲候選者個數、L爲序列最大長度、Q爲CMfinder的CM中state個數、K_1和K_2分別是2個系統所跑的回合（Ieteration）數，由時間復雜度上的分析非常明顯的可以看的出來，爲何我們的系統在多個家族測試時，速度會是CMfinder的數倍，而且測試資料越大時差距也會越來越大。

[3] 可執行量分析：

在龐大資料量分析的實驗之中，結果更是明顯，連CMfinder都沒有結果，原因除了處理時間的問題外，其實在測試資料到達 200 條序列之後，它所需要的記憶體往往會需要數GB，因爲它的記憶體需要$O(M^2)$，其中M爲候選者個數，而我們的記憶體也只需要O(M)，這是我們的系統能處理龐大資料量而其它資料不能最大的一個重點，雖然龐大資料量時雜訊過多而使的MCC會下降，但大量資料的實驗才近於生物學家們所做的生物實驗的實際情況，能處理大量資料的系統其可使用度也才會高，這點我們做到了。

第五章、結論與建議

對於我們的系統，其實還有幾個能再加強改進的地方，我們接下來一一討論它們，與可能的改進方法。

(1) 在「4.3 實驗結果分析」的章節中有提到，現有的系統大多都是用 MFE 來計算預測核醣核酸的二級結構，而大多搜尋結構元的系統都是用它們來決定核醣核酸序列的二級結構，所以這些依賴它的系統所能找到最好的 MCC 也當然不可能高過結構預測程式的準確度，我們的系統也是有這樣的特性，雖然我們有設計第二次 Gibbs-like 和 Pattern Search 來補強這個地方，但是在搜尋上是找與種子成員”完全相同”的子結構，而不是”高度相似”的子結構，但同家族的共同結構元大多相似且未必相同，因此這個方法其實能力有限。最好的方法當然是拿種子的排比模型去和測試資料中每條序列的每個子序列做排比，將序列相似度高的部份拿出來做成新的候選者，但是很花時間的程序，所以要想好的、有效率的搜尋法後才加上去，才不會使系統爲了能有好的 MCC 而失去能處理大量測試資料的能力才行。

(2) 我們的系統目前無法找有 pesudoknot 的核醣核酸家族，很明顯的重點在於以下幾點：

<1>形狀樹和抽象形狀樹的結點只有一種，無法表示多種 min-SCC。

<2>新的「核醣核酸結構排比演算法」無法排比 pesudoknot 的結構。

在<1>中所說的問題很容易處理，只要將樹上的結點分成不同的種類即可，而<2>就是比較難改進的地方，必需要花較多心力設計新的演算法才可以。

我們系統的缺點就在於過度依賴 MFE 結構預測程式的結果，和無法處理 pesudoknot 這兩點，這也是未來可以再加強的目標。

第六章、參考文獻

Chiu, D.K. and Kolodziejczak, T. (1991) Inferring consensus structure from nucleic acid sequences. *Compu Appl Biosci*, **7**, 347-352.

Conne, B. *et al.* (2000) The 3' untranslated region of messenger RNA: A molecular 'hotspot' for pathology? *Nat. Med.*, **6**, 637-641.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) Biological sequence analysis. Cambridge University Press.

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

Furtig, B., Richter, C., Wohnert, J. and Schwalbe, H. (2003) NMR spectroscopy of RNA. *Chembiochem.*, **4**, 936-962.

Gardner, P.P. and Giegerich, R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, **5**:140.

Giegerich, R., Voβ, B. and Rehmsmeier, M. (2004) Abstract shapes of RNA. *Nucleic Acids Research*, **32**, 4843-4851.

Gorodkin, J., Heyer, L. and Stormo, G. (1997) Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Research*, **25**, 3724-3732.

Gutell, R.R. (1993) Evolutionary characteristics of RNA: inferring higher-order structure from patterns of sequence variation. *Curr Opin. Struct. Biol.*, **3**, 313-322.

Hamada, M., Tsuda, K., Kudo, T., Kin, T. and Asai, K. (2006) Mining frequent stem patterns from unaligned RNA sequences. *Bioinformatics*, **22**, 2480-2487.

Hochsmann, M., Toller, T., Giegerich, R. and Kurtz, S. (2003) Local similarity of RNA secondary structures. *Proc of the IEEE Bioinformatics Conference*. 59-68.

Hofacker, I.L., Priwitzer, B. and Stadler, P.F. (2004) "Prediction of locally stable RNA secondary structures for enome-wide surveys", *Bioinformatics*, **20**, 186-190.

Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, L., Tacker, M. And Schuster, P. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte fur Chemie*. **125**, 167-188.

Hofacker, I.L., Fekete, M. and Stadler, P. (2002) Secondary structure prediction for aligned RNA sequences. *J. Molecular Biology*, **319**, 1059-1066.

Huttenhofer, A. *et. al.* (2001) RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.*, **20**, 2943-2953.

Ji, Y., Xu, X and Stormo, G.D. (2004) A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics*, **20**, 1591-1602.

Juan, V. and Wilson, C. (1999) RNA secondary structure prediction based on free energy and phylogenetic analysis. *J. Mol. Biol.*, **289**, 935-947.

Knudsen, B. and Hein, J. (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, **15**, 446-454.

Knudsen, B. and Hein, J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.*, **31**, 3423-3428.

Lai, E.C. (2003) RNA sensors and riboswitches: self-regulating messages. *Current Biology*, **13**, R285-R291.

Mandal, M. *et al.*, (2003) Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell*, **113**, 577-586.

Mattick, J.S. (2001) Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.*, **2**, 986-991.

Moulton, V. (2005) Tracking down noncoding RNAs. *Proc. Nat. Acad. Sci. USA*, **102**, 2269-2270.

Nudler, E. and Mironov, A.X. (2004) The riboswitch control bacterial metabolism. *Trends Biol. Sci.*, **29**, 11-17.

Pesole, G., Mignoe, F., Gissi, C., Grillo, G., Licciulli, F. and Liuni, S. (2001) Structural and functional features of eukaryotic mRNA untranslated regions. *Gene*, **276**, 73-81.

Rouillard, J.M., Zuker, M. and Gulari, E. (2003) OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res.*, **31**, 3057-3062.

Ruan, J., Stormo, G. and Zhang, W. (2004) An iterative loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, **20**, 58-66.

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

Sankoff, D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Applied Math*, **45**, 810-825.

Siebert, S. and Backofen, R. (2003) MARNA: A server for multiple alignment of RNAs. Proc of the German Conference on Bioinformatics, 35-40.

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

Thompson, J., Higgins, D. and Gibson, T (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673-4680.

Touzet, H. and Perriquet, O. (2004) CARNAC: folding families of related RNAs. *Nucleic Acids Research*, **32**, W142-145.

Wolfinger, M.T., Svrcek,-Seiler, W.A., Flamm, C., Hofacker, I.L. and Stadler, P.F. (2004) Efficient computation of RNA folding dynamics. *J. Phys. A: Math Gen*, **37**, 4731-4741.

Yao, Z., Weinberg, Z and Ruzzo, W. (2006) CMfinder: a covariance model based RNA motif finding algorithm. *Bioinformatics*, **22**, 445-452.

Zuker, M and Stiegler, P. (1981) Optimal computer folding of larger RNA sequences using therdynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133-148.

Zuker, M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48-52.