國立交通大學

資訊科學與工程研究所

碩 士 論 文

探 索 擁 有 多 調 控 序 列 的 調 控 模 組

Discover Regulatory Modules Consisting of Multiple Binding Sites

研 究 生：賴昀君

指導教授：胡毓志　教授

中 華 民 國　九 十 五　年 六 月

Discover Regulatory Modules Consisting of Multiple Binding Sites

Stident: YunJun Lai Advisors: Dr. Yuh-Jyh Hu

Institites of Computer Science and Engineering

National Chiao Tung University

Abstract

One of the keys to deciphering the secrets of life is to understand transcriptional regulation mechanisms. Such mechanisms are typically mediated by the binding of transcription factors to specific upstream or downstream regions of genes, which leads to most recent studies focused on the conservation of binding sites. Unlike current research, we address the importance of the distance between binding sites for the prediction of *Regulatory Modules*. Based on the Bayesian framework, we present an algorithm, SAMLA (***S**imulated **A**nnealing for **M**ultiple **L**ocal Sequences **A**lignment*), which takes into account the consensus levels of biding sites as well as the relative distance among them. To demonstrate the performance of our new approach, we conducted a comparative study with several current methods. We tested them on the datasets derived from *E.coli*, and the experimental results show that our method significantly outperforms the others.

探索擁有多調控序列的調控模組

研究生: 賴昀君　　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

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 2006

Hsinchu, Taiwan, Republic of China

中華民國九十五年六月

誌　　謝

不免俗套的，我要先感謝家人這幾年來的支持，讓我能夠在這個年紀總結先前的學問，這本厚厚的論文便是一個見證。我將這本論文以及學問完成的喜悅呈現給你們，希望你們能引以爲榮。

對於三個十分特別的戰友：秉蔚、勁伍、音璇，感謝你們的陪伴，一起奮鬥的過程是這輩子都不會忘記的。宛嫻學姊（塔兒）、莊破破（米兒）、ㄗㄤ小孩，沒有你們在 WOW 中的鼓勵以及陪伴，論文要完成是十分的不易。噢！也不會忘記美華學姊跟姊夫，你們的關心也讓我能夠忍住這段難熬的日子。還有亦師亦友的萬田學長，感謝你的指導以及督促。另外，實驗室的貫中、豐茂、阿貴、繼養、子緯跟巽昌，在我最煩悶的時候，你們的歡愉讓我心情舒坦了不少。此外，感謝指導老師胡毓志老師的指導。

特別要提到的是幾位「良師益友」：資四 C 的那群人渣，YY、顧大頭一號、曾大頭二號、胖阿智、蜥蜴、博宇，你們的犯賤跟對我的遵遵教誨還有 AOK 跟 D2 奮戰是我這前半段研究生活的重心。還有小緯，游泳、打球、吃飯承蒙你的陪伴。你們不愧是我的好朋友，哈哈哈。

要感謝的人太多了，我今天能夠走到這邊都是眾人的幫忙，我今天的這點小成就是你們對我的協助而累積出來的，希望我能夠在之後的日子中不負你們的協助繼續向前邁進。

第一章　前言

1.1 生物背景

爲了窮究人類生命的奧秘，在西元 1953 年，Watson 與 Crick 發現去氧核醣核酸（Deoxyribonucleic acid；簡稱 DNA）以雙股螺旋（Double Helix）結構存在於生物體中【Watson, Crick, 1953】，這不僅解開了基因的化學結構之謎，同時也揭示了 DNA 所攜帶的遺傳訊息如何完成遺傳訊息的複製與表現的機制。西元 2003 年隨著人類基因體計畫（Human Genome Project）的提前完成，DNA 序列數量以倍數的方式增加。在這麼大量的基因體序列（Genomic Sequence）中，如何快速且有效率地探尋出基因（Gene）序列所隱藏的資訊儼然已經成爲生物資訊的重要課題。在這些課題中有兩個基礎的問題：一，基因在遺傳中扮演的角色爲何？其二，基因是如何發揮其功能？

目前我們了解生物體的遺傳特性以及生理機能皆由基因來控制，而基因的表現取決於當下生物體內各項環境因子的調控，通常基因表現的調控是受到特殊的結合蛋白質（轉錄因子，Transcription Factor，TF）與基因轉錄作用起始位置（Transcription Start Site，TSS）鄰近的去氧核醣核酸序列（調控序列，Transcription Factor Binding Site，TFBS，Regulation Site）進行交互作用，對基因的轉錄做出正向或是負向的調節。所謂正向調節，即當轉錄因子與基因轉錄作用起始位置鄰近的去氧核醣核酸序列作用之後，被誘導的基因便開始發生轉錄（Transcription）的動作，產生信使核醣核酸（mRNA），並且進一步轉譯（Translation）產生蛋白質，對生物體產生調節及作用。而負向調節，則是抑制此基因的表現，使之無法產生蛋白質來與其他分子作用。這個由基因序列經過轉錄作用形成 mRNA，再透過轉譯作用產生蛋白質的過程稱爲分子生物學的中心法則（Central Dogma of Molecular Biology）。轉錄作用與轉譯作用涉及了十分複雜的生理調控過程，在本

研究中，我們關注的重點是「轉錄作用」。

在轉錄過程中，稱爲 RNA 聚合酶（RNA Polymerase）的酵素將會附著至 DNA 序列上，沿著基因的鑄模股（Template Strand）將對應的 RNA 核苷酸串連成爲一長序列，而此 RNA 核苷酸序列即爲信使核醣核酸（Messenger RNA，mRNA），而在 DNA 上讓 RNA 聚合酶附著的區域即爲啓動子（Promoter）。啓動子區域含括了標示轉錄起點（TSS）的特定核苷酸序列以及轉錄起點的上游區段（Upstream）。許多基因上游區段的鹼基序列已經被確定，而且包含許多共同的序列，稱之爲共有序列（Consensus Sequence）。這些共有序列，主要是用來幫助 RNA 聚合酶辨識轉錄的起點位置以及轉錄作用的進行。

轉錄作用發生有三個要件：

一、 RNA 聚合酶的σ子單元（σ subunit）。σ子單元爲 RNA 聚合酶中辨識啓動子的主要單元，當σ子單元發現基因序列的啓動子（Promoter）並且與之結合之後，RNA 聚合酶才會開始轉錄作用合成 mRNA。

二、 RNA 聚合酶與啓動子的結合能力。RNA 聚合酶與啓動子的結合能力會影響基因表現的程度。若結合能力低，則基因的表現程度低；反之則基因容易發揮作用。

三、 DNA 基因序列中的調控序列。這些位於 DNA 中高度保留的共有序列（Consensus Sequence）在轉錄作用中扮演著極爲重要的生物意義。

在高等真核生物（Eukaryotes）中，調控序列可略分四類：

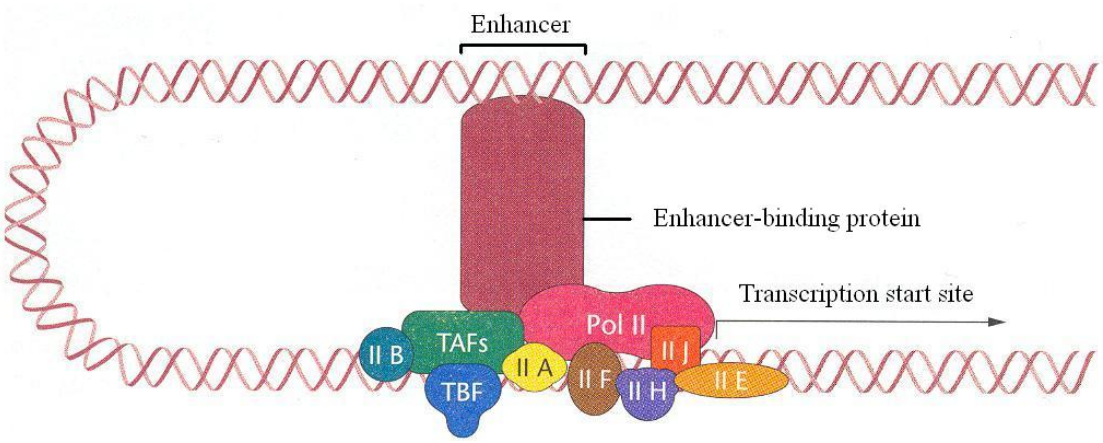
第一類，TATA box（或稱 Goldberg-Hogness）。此調控序列位於啓動子中，通常坐落於距離轉錄起點位置前 30 個核甘酸的位置。其序列以 A、T 爲主，共有序列的型式（Pattern）爲 TATAAAA，故稱之 TATA box。此序列的主要功能爲幫助 RNA 聚合酶將 DNA 展開雙股結構，以便 RNA 聚合酶沿著鑄模股合成

mRNA。

第二類，CAAT box。此調控序列位於啓動子中，通常坐落於距離轉錄起點位置前 80 個核甘酸的位置。共有序列的型式爲 GGCCAATCT，又 CAAT 爲此調控序列中高度保留的部份，故稱 CAAT box。

第三類，GC box。此調控序列位於啓動子中，通常坐落於距離轉錄起點位置前 110 個核甘酸的位置。共有序列的型式爲 GGGCGG。CAAT box 與 GC box 的功能與 TATA box 和促進子相比，其所扮演的功能比較像促進子。

第四類，促進子（Enhancer）。此區域可能涵蓋了許多不同的調控序列，不同基因的促進子中包含的調控序列也因基因的功能而有所變異。促進子相對於基因序列的位置隨著不同的基因而變化，可能與轉錄起點位置相距超過上萬個鹼基位置，也可能位於基因的上游區域（Upstream region），下游區域（Downstream region）或是位於其他的基因序列中。雖然促進子可能與轉錄起點位置相距甚遠，然而他們亦可透過 DNA 序列的結構以及不同的轉錄因子來遙控基因的轉錄作用（圖 1.1.1）。



（**圖 1.1.1**）促進子在基因轉錄作用中扮演的角色。促進子連結蛋白（Enhancer-binding protein；由許多轉錄因子聚合而成的大型複合體）（棕色）與促進子相結合之後，使得 RNA 聚合酶可以更穩定地與啓動子相結合，進而加速基因轉錄作用的發生。

在原核生物（Prokaryotes）中，調控序列較爲簡單可略分三類：

一、促進子。

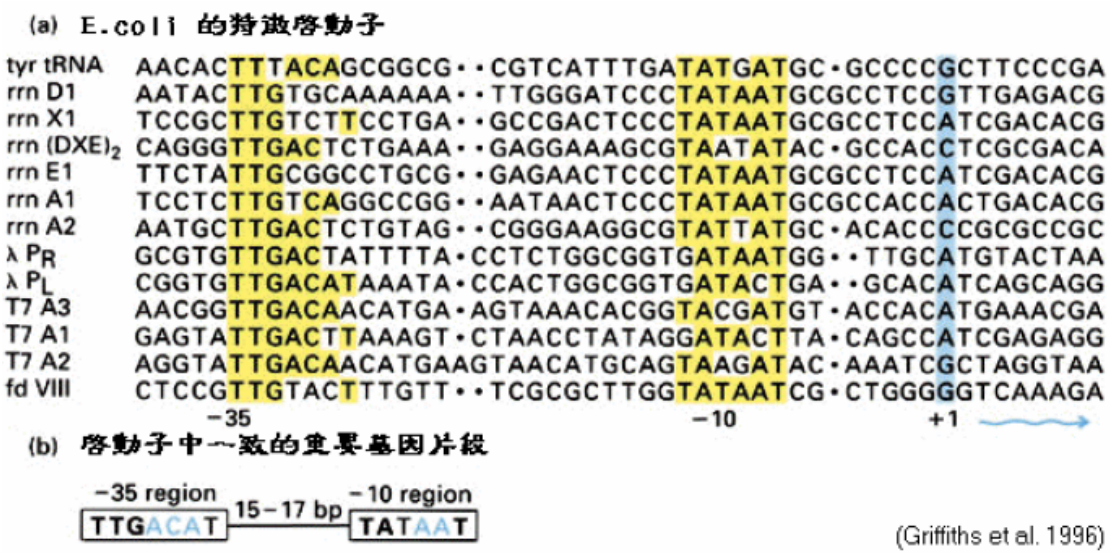
二、-10 Region（或 Pribnow box），此調控序列位於啓動子中，通常坐落於距離轉錄起點位置前 10 個鹼基的位置。其序列以 A、T 爲主，型式（Pattern）爲 TATAAT。

三、-35 Region，此調控序列位於啓動子中，通常坐落於距離轉錄起點位置前 35 個鹼基的位置。共有序列的型式爲 TTGACA。

經由生物學家驗證，在同一生物體內的任一細胞皆擁有相同的基因體（Genome）。不同細胞間功能的差異之處僅在於其含括的調控蛋白（Regulatory Protein）的不同，藉此讓相同的基因體來產生不同的基因轉錄表現。由於蛋白質擁有專一性的特色，因此不同的調控蛋白會辨識 DNA 序列中不同的調控序列，以此來控制基因的活動。

1.2 研究動機

經由上一小節介紹遺傳訊息運作的過程後，我們可以得知基因藉透過不同的轉錄因子與調控序列相互結合而有不同的表現。通常基因開始轉錄作用時需要不只一種轉錄因子（調控序列）的協助，可能會需要許多不同的轉錄因子（調控序列）共同參與反應。當RNA聚合酶與這些必需的轉錄蛋白複合體結合之後，才會沿著DNA的鑄模股來探測啓動子的位置，並且開始轉錄作用。在本研究中，我們稱一個基因發揮作用所需要的調控序列群爲調控模組（Regulatory Module）。一個基因的調控模組可能會包含一個或多種的調控序列。



（**圖1.2.1**）*E. coli* 的特徵啓動子中的調控模組。

如（圖1.2.1）(a) 顯示大腸桿菌（*E.coli*）中的 tyr tRNA，rrn D1，rrn X1，rrn $(DXE)_2$，rrn E1，rrn A1，rrn A2，λ P_R，λ P_L，T7 A3，T7 A1，T7 A2，fd VIII 等基因的啓動子中皆擁有相似的調控模組。此模組包含了兩個不同的調控序列，並且此兩序列以 15 至 17 個核甘酸的長度爲相對距離。舉例來說，於 tyr tRNA 基因上的調控模組便爲（TTTACAG - [15] - TATGAT）。而 rrn D1 的調控模組爲（TTGTGCA - [15] - TATAAT）。rrn X1 爲（TTGTCTT, [15],

TATAAT）。rrn $(DXE)_2$ 爲（TTGACTC, [15], TAAAAT）。rrn E1 爲（TTGCGGC, [15], TATAAT）。rrn A1 爲（TTGTCAG, [15], TATAAT）。rrn A2 爲（TTGACTC, [15], TATTAT）。λ PR 爲（TTGACTA, [16], GATAAT）。λ PL 爲（TTGACAT, [16], GATACT）。T7 A3 爲（TTGACAA, [16], TACGCT）。T7 A1 爲（TTGACTT, [16], GATACT）。T7 A2 爲（TTGACTA, [17], TAAGAT）。fd VIII 爲（TTGTACT, [15], TATAAT）。在這些基因的啓動子中，調控模組是由兩種調控序列所組成，分別爲 TTGACAT 與 TATAAT（以共有序列來做表示）。

隨著定序技術的進步，大量的基因體序列草圖已經隨處可得。從這麼大量的序列中探測基因以及其功能將是關鍵性的工程。若我們能了解、並且發現調控模組在基因的DNA序列中的位置以及型態，便可以加速預測基因的功能和表現。然而，現行預測調控模組的演算法大部分都將其研究領域侷限於探測其中調控序列的一致性，而忽略大部調控模組中調控序列之間的間距也必須納入考慮。因此在這篇論文中，我們將已成熟發展的模擬退火法（Simulated Annealing Algorithm），以及統計學上推論方法爲基礎所延伸結合的 SAMLA (**S**imulated **A**nnealing for **M**ultiple **L**ocal Sequences **A**lignment)，來尋找擁有相似表現基因的序列（Co-Expressed Gene Sequences）中可能的調控模組。

1.3 研究假設與目標

每條基因序列上的調控模組均不相同，然而在某種生物環境中擁有相似表現的基因（Co-Expressed Genes）必定受到類似的調控蛋白群調控。我們認爲擁有相似表現的基因其「基本的調控模組」必定擁有特定的特徵：「基本的調控模組」中的調控序列會有相似的結構以及分布。在此，所謂調控序列的結構指的是不同的基因的DNA上調控序列會有相似的序列樣式（Pattern）。例如（圖1.2.1）中的 TATA box，並非每一個基因的 TATA box 都是以 TATAAT 的模樣存在，但 TATAAT 爲這些調控序列最爲一致的樣式（Consensus Pattern）。調控序列的分布則是指不同的調控序列之間有相同的順序與以及有相似甚至是完全相同的相對應間距。如圖1.2.1，若以5'端往3'端的順序而言，TTGACAT 在 DNA 的位置會比 TATAAT 先出現，而且 TTGACAT 以及 TATAAT 之間的相對距離最大不超過17個核甘酸（17bp），最小則不少於15個核甘酸（15bp）。

因此本研究中最重要的基本假設以及主要研究目標爲，

研究假設：

相似表現基因的序列群其兩兩之間的最基本調控模組其結構以及分布十分相似。

研究目標：

給定一群擁有相似表現基因的序列，並且限制調控序列的個數以及長度，我們將從這一群基因序列中探測出每條基因序列上最可能的基本調控模組。

爲了簡化之後文章中論述的方便以及統一，我們將我們欲探測的「基本調控模組」簡稱爲「調控模組」。

此外在本篇論文中，我們亦有些額外的假設來協助系統的推論與實作：

1. 對任何的調控序列而言，其任兩位置核甘酸之間的出現我們視爲機率上的獨立事件。
2. 對於調控模組的結構以及分布，我們亦視爲機率上的獨立事件。

有了上述兩個大前提的假設，我們在第三章便可以輕易地推導出系統的理論基礎以及實做方向。

1.4 論文架構

本文主要分爲五章。在第一章中，我們將首先介紹生物體內遺傳訊息的運作，進而闡述預測調控模組此研究領域的重要性以及我們的想法與目標。緊接在第二章文獻探討的部份，我們將會著重在與探測重要基因片段相關的背景知識以及相關研究。接下來，我們會在第三章中詳細介紹演算法的流程與細節。而整個實驗方法與結果，以及和其他相關研究方法的比較都將會記錄在第四章。最後，我們會在第五章總結整個論文，並提出討論以及未來展望。

第二章 文獻探討

探測基因之調控模組在生物資訊的研究領域中是一項基本且重要的問題。當前大部分的研究都視調控模組中只包含單獨一條的調控序列，因此預測調控模組問題便可視爲搜尋單一調控序列問題。基本上，經過這樣的問題簡化之後，探測調控模組這個問題即可轉換成爲多序列區域性排比（Local Multiple Sequence Alignment）問題。然而，這看似簡單的多序列區域性排比問題卻已被證明爲 NP-Complete 的公開難題【Day,W.H. and McMorris,F.R, 1993】【Gusfield, 1997】。因此許多各式各樣的逼近演算法、工具紛紛地提出來解決這個難題。

接下來的討論中我們將簡單的介紹最爲常被使用的調控序列的模型－位置比重矩陣（PWM）。稍後，則簡述現行尋找調控序列以及模組的工具。

2.1 位置比重矩陣（Position Weight Matrix；PWM）

目前在大部分調控序列的研究中，位置比重矩陣爲廣泛被使用的調控序列模型。位置比重矩陣，M，爲一個 $4\times l$ 大小的陣列，l 表示調控序列的長度。M 中的每個元素 m_{ij} 表示觀察到的調控序列的第 j 個位置爲核甘酸 i 的出現頻率，其中$1 \leq j \leq l$，$i \in \{A,C,G,T\}$。如（表2.1.1）(a) 共有 11 條調控序列，長度皆爲 10。我們可以藉由計算陣列每個位置 m_{ij} 爲此 11 條調控序列的第 j 個位置爲核甘酸 i 的頻率來得到此 11 條調控序列的位置比重矩陣，如（表2.1.1）(c)。例如，（表2.1.1）(c)中的 $m_{A5} = 1/11 \cong 0.09$，正表示此 11 條調控序列中第 5 個位置爲核甘酸 A 的頻率是 1/11。

(a)

A A A A C T G T G T
A A A A A T G T G G
A A A A C T G T G G
A A A A C T G T G G
A A A A C T G T G G
A A A A T T G T G G
A A A A C T G T G G
C A A A T T G T G G
A A A A C T G T G G
C A A A C T G T G G
C A A A C T G T G G

(b)

	1	2	3	4	5	6	7	8	9	10
A	8	11	11	11	1	0	0	0	0	0
C	3	0	0	0	8	0	0	0	0	0
G	0	0	0	0	0	0	11	0	11	1
T	0	0	0	0	2	11	0	11	0	10

(c)

	1	2	3	4	5	6	7	8	9	10
A	.73	1.0	1.0	1.0	.09	.00	.00	.00	.00	.00
C	.27	.00	.00	.00	.73	.00	.00	.00	.00	.00
G	.00	.00	.00 .	00	.00	.00	1.0	.00	1.0	.09
T	.00	.00	.00	.00	.18	1.0	.00	1.0	.00	.91

(**表 2.1.1**)

(a) 我們取得一群調控序列樣式的排比。

(b) 給定一群調控序列排比樣式(a)，我們計算此排比樣式每個位置每個核甘酸出現的個數。以 4×10 的矩陣表示之。矩陣中的每個元素 m_{ij} 表示：第 j 個位置出現核甘酸 i 的個數（在已知調控序列群）。$1 \leqq j \leqq 10$，$i \in \{A,C,G,T\}$。

(c) 透過(b)，我們可以計算出此調控序列群的位置比重矩陣（PWM）。矩陣中的每個元素 m_{ij} 表示：第 j 個位置出現核甘酸 i 的頻率（在已知調控序列群）。$1 \leqq j \leqq 10$，$i \in \{A,C,G,T\}$。

使用位置比重矩陣作爲一群調控序列的模型，其優點是：比起以共同字串（Consensus String）搭配 IUPAC 碼表示調控序列來得更加有彈性。而缺點則爲，由於視調控序列當中任兩位置核甘酸的出現爲獨立事件，因此很難從位置比重矩陣中描述調控序列中不同位置核甘酸之間共變（Co-variance）的關係。

2.2 MEME【Bailey and Elkan, 1995】

MEME（Multiple Expectation-maximization for Motif Elicitation）爲一個非監督式學習的演算法（Unsupervised Learning Algorithm）。給定一群未排比的序列，MEME 可以尋找出這群序列中最具代表性的一群子序列，這些子序列代表著輸入序列的特徵。因此若給定的序列爲共同被調控基因的 DNA 序列，則 MEME 可以預測基因序列中統計意義上顯著的調控序列。

MEME 的運作主要是運用了期望與最佳化（Expectation-maximization）的概念，反覆地計算出最具代表性的位置比重矩陣，而此代表性的位置比重矩陣便是輸入序列特徵的另一種呈現方式。MEME 首先隨機選擇輸入序列中任一位置的子序列，來形成初始的位置比重矩陣。經過隨機出始化之後，MEME 藉由以下兩大步驟的反覆運作進而得到較具代表性的調控序列：

- 期望（Expectation）

  透過上一最佳化階段得到的位置比重矩陣，$M^{(t)}$。MEME 對所有的子序列計算此子序列爲調控序列的機率，並且建立其機率分布。

- 最佳化（Maximization）

  依照期望階段中所所有子序列的機率分布，MEME 選擇出最可能的子序列，並且形成下一階段的位置比重矩陣，$M^{(t+1)}$。

在設計上，MEME 可以測試不同長度的調控序列。亦即，在使用者給定長度範圍之內，MEME 會於此長度限制之下找尋最爲可能的代表性序列。然而，由於 MEME 採用 EM 的概念運行，因此程式很有可能因爲初始化時位置比重矩陣建構的偏差，導致最後的結果陷入局部最佳化（Local Optimal）。

2.3 Gibbs Sampler【Lawrence *et. al.* 1993】

Gibbs Sampling 爲馬可夫鏈蒙地卡羅（Markov-chain Monte-Carlo）演算法中最容易了解的技巧，多應用於條件概率（Conditional Distribution）容易計算之問題。在【Lawrence *et. al.* 1993】這篇論文中，首先將 Gibbs Sampling 的技巧應用於多序列區域性排比。給定一群基因序列 G 而調控序列以 PWM 的形式呈現，Gibbs Sampler 主要是尋找調控序列模型與背景之間差距的最大值，此差距即以所謂的相對亂度（Relative Entropy 或 Kullback-Leibler distance）來呈現：

$$\sum_{j=1}^{l} \sum_{i \in \{A,C,T,G\}} m_{i,j} \log \frac{m_{i,j}}{b_i}$$

。調控序列模型，M，爲一個 $4 \times l$ 大小的陣列，l 表示調控序列的長度。m_{ij} 表示觀察到的調控序列的第 j 個位置爲核甘酸 i 的出現頻率，其中$1 \leq j \leq l$，$i \in$ {A,C,G,T}。b_i 表示背景（非調控序列的基因序列）中核甘酸 i 的出現頻率。

Gibbs Sampler 搜尋之流程可簡述如下：

首先假設 a_1，…，a_N 爲基因序列 s_1，…，s_N 中調控序列出現的起始位置。$M^{(t)}$ 表示於第 t 時間點時 Gibbs Sampler 所發現的調控序列模型。程式之初，隨機從序列 s_1，…，s_N 選擇 a_1，…，a_N 等位置；並且根據 a_1，…，a_N 計算調控序列模型 $M^{(0)}$。接下來演算法不斷的重複執行以下兩步驟直到調控序列模型不再變動：

(一) 更新（Update Phase）

針對所有的序列 s_1，…，s_N，任意或依序選擇其中一條 s_i，$1 \leq i \leq N$。移除 a_i，根據 a_1，…，a_{i-1}，a_{i+1}，…，a_N 計算調控序列模型 $M^{(t)}_{-i}$。$M^{(t)}_{-i}$ 表示從此時間點的調控序列模型 $M^{(t)}$ 中移除位於第 i 條序列上的調控序列而

得到的調控序列模型。

(二) 取樣（Sampling Phase）

若 s_i 長度爲 k。針對 s_i 上所有長度爲 l 的子序列 x_1，…，x_{k-l+1}，分別依據 $M^{(t)}_{-i}$ 與背景 B 來計評估其分數，Q_{x_j} 與 P_{x_j}。$1 \leq j \leq k-l+1$。根據 Q_{x_j} 與 P_{x_j} 的分數高低，我們可以計算子序列 x_j 的機率分布爲：

$$P(x_j) = \frac{Q_{x_j} / P_{x_j}}{\sum_{m=1}^{k-l+1} Q_{x_m} / P_{x_m}}$$

演算法便依據 $P(x_j)$ 來隨機選擇新的子序列以及其於 s_i 上的位置 z。根據 a_1，…，a_{i-1}，z，a_{i+1}，…，a_N 產生下一步驟新的調控序列模型 $M^{(t+1)}$。

由上述之介紹不難看出 Gibbs Sampling 與 EM 十分相似，然而 Gibbs Sampling 的概念比起 EM 更容易實做，其原因爲條件機率 $P(z| a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_N)$ 的機率分布較容易計算與及實作。雖然 Gibbs Sampling 實際上也和 EM 的技巧都存在局部最佳解的問題存在（初始化時位置比重矩陣建構的小偏差，導致最後的結果陷入局部最佳化的困境），且序列長度必須事先給定爲系統參數爲另一遺憾之處，然而這些缺點與其貢獻比較之依舊是瑕不掩瑜。

2.4 Dyad Analysis【van Helden *et al.* 2000】

van Helden 等學者在 2000 年首度提出方法探測雙元調控模組，此方法即爲 Dyad Analysis。Dyad Analysis 是藉由統計所有可能雙元基因片段的出現頻率將其重要性數值化，進而探測出雙元調控序列。（圖2.4）。

Dyad Analysis 中提出調控模組的核心應該是雙元基因片段 D 所形成，而雙元基因片段 D 應該是兩對較爲短小的調控序列，且以一固定的距離作爲分隔，以符號表示之爲：

$$D = w_1.n_s.w_2$$

w_1 與 w_2 代表兩對較爲短小的調控核心，n_s 則爲核心之間的序列，以 s 代表其長度。在分析中，w_1 與 w_2 的長度都設定爲 3，而間距爲 0 到 16 之間。簡單的來說，倘若 D 爲輸入序列群的調控模組的核心部份，則雙元基因片段 D 出現於輸入序列群的機率必定遠遠的超過出現於背景中的機率。（這裡指的背景爲輸入序列群所屬物種的所有非基因序列，Non-coding Region）。在此前提之下，作者首先計算出現於背景（Non-coding Region）中所有可能的雙元基因片段 D 的機率，以符號 $f_{exp}(D)$ 表示之。接下來套用二項分布的概念來計算：在輸入序列中觀察到雙元基因片段 D 至少出現次數大於 n 次的機率，

$$P(D \geqq n) = \Sigma^{T}_{i=n} C^{T}_{i} [f_{exp}(D)]^{i} \times [1 - f_{exp}(D)]^{T-i}$$

。T 則爲輸入序列中所有可能發現雙元基因片段 D 的位置，$T = \Sigma_j (l_j - 2\times k - s + 1)$。$l_j$ 爲輸入序列 j 的長度，k 則表示調控核心的長度。因此，雙元基因片段 D 的重要性便可以數值化如下式，

$$sig = - \log_{10}[P(D \geqq n) \times N_p]$$

，其中 N_p 爲所有可能的雙元模組的總數。雙元基因片段 *D* 的 sig 代表著此片段 *D* 的重要性，數值越高，越爲真實模組的可能性越高。如（圖2.4）中所顯現，CGG.n_{16}.CCG的 sig 爲最高，因此 CGG.n_{16}.CCG 爲雙元模組核心的機率最高。分析的最後便會將 sig 最高的片段 *D* 爲核心主體，以 *D* 來重組（還原）調控模組。見（圖2.4）。

然而 Dyad Analysis 只能探測固定間距的雙元調控序列，而且調控序列之間距的設定只能介於 0 到 16 之間，因此其所能探測到包含雙元調控序列仍是有所限制的。

pattern	**total occurrences**	**overlaps**	**non-overlapping occurrences**	**expected occurrences**	**proba**	**sig**
.CGG . . . . . .CCG	22	2	20	0.59	$1.9e^{-12}$	7.2
.CGG . . . . . .CGa	12	2	10	0.50	$2.1e^{-10}$	5.1
tCG . . . . . .CCG	12	2	10	0.50	$2.1e^{-10}$	5.1
.CGG . . . . . .tCC	12	3	9	0.91	$6.7e^{-07}$	1.6
. .GGa . . . . . .CCG	12	3	9	0.91	$6.7e^{-07}$	1.6
tCGGa . . . . . .tCCGa	**Assembly**					

（**圖2.4**）Dyad Analysis【van Helden *et al.* 2000】由 GAL 資料群中探測的結果。其中 sig 表示各個基因片段重要性的數值。

2.5 Bioprospector【Liu *et. al.* 2001】

Bioprospector 延伸 Gibbs Sampler【Lawrence *et. al.* 1993】的概念來預測單一調控序列或是雙重調控序列之模組。此外 Bioprospector 利用三階的馬可夫模型作爲背景模型以其改進原先於 Gibbs Sampler 中使用零階馬可夫背景模型的預測能力。Bioprospector 在預測單一調控序列時的程序與 Gibbs Sampler 相似（見2.3節）。預測雙重調控序列之模組時，Bioprospector 針對原先於 Gibbs Sampler 中的「更新」與「取樣」兩步驟做了以下之修改：首先假設 $(a_1, a_2)_1$，…，$(a_1, a_2)_N$ 分別代表基因序列 s_1，…，s_N 中雙重調控序列出現的起始位置，而 $M^{(t)}_{(j)}$ 表示於第 t 時間點時 Gibbs Sampler 所發現的第 j 條調控序列模型。

(一)更新

我們對所有的序列 s_1，…，s_N，任意或依序選擇其中一條 s_i，$1 \leq i \leq N$。移除 $(a_1, a_2)_i$，依據 $(a_1, a_2)_1$，…，$(a_1, a_2)_{i-1}$，$(a_1, a_2)_{i+1}$，…，$(a_1, a_2)_N$ 計算調控序列模型 $M^{(t)}_{-i(1)}$ 以及 $M^{(t)}_{-i(2)}$。其中 $M^{(t)}_{-i(1)}$ 表示於第 t 時間點調控序列模型 $M^{(t)}_{(1)}$ 中移除位於第 i 條序列上的調控序列而得到的調控序列模型；$M^{(t)}_{-i(2)}$ 表示於第 t 時間點調控序列模型 $M^{(t)}_{(2)}$ 中移除位於第 i 條序列上的調控序列而得到的調控序列模型。

(二)取樣

針對序列 s_i 上的兩子序列 x_1 和 x_2，我們依據調控序列模型 $M^{(t)}_{-i(1)}$ 以及 $M^{(t)}_{-i(2)}$，和背景模型可以計算其的機率分布爲：

$$A(x_1, x_2) = \frac{Q_{x_1}}{P_{x_1}} \times \frac{Q_{x_2}}{P_{x_2}}$$

，Q_{x_k} 表示由 $M^t_{-i(k)}$ 評估 x_k 出現的機率；P_{x_k} 表示由背景評估 x_k 出現的機率，$1 \leqq k \leqq 2$。然而演算法卻並非計算所有可能的字串組 (x_1, x_2)，而是先依照 $A(x_1,*) = \sum_{x_2} A(x_1, x_2)$ 分布來取得第一個序列的新位置 z_1，再利用此新位置根據 $A(x_1, x_2) / A(x_1,*)$ 選出第二條序列的新位置 z_2。最後演算法便依據 $(a_1, a_2)_1, \ldots, (a_1, a_2)_{i-1}, (z_1, z_2)_i, (a_1, a_2)_{i+1}, \ldots, (a_1, a_2)_N$ 計算調控序列模型 $M^{(t+1)}{}_{(1)}$ 以及 $M^{(t+1)}{}_{(2)}$。

此外，爲了要能夠評定其探測出模組的優劣，Bioprospector 訂定了評分的標準，如下式：

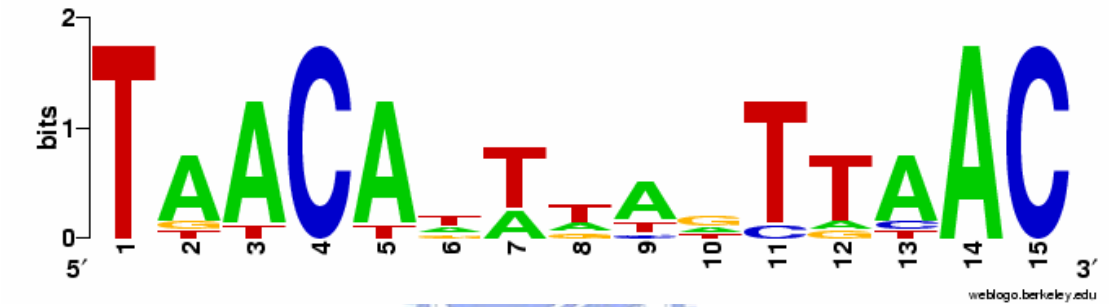
$$Motif\ Score = \#seg \times \exp\{ [\Sigma_{all\ positions\ i} \Sigma_{all\ nucleotides\ j} q_{i,j} \times \log(q_{i,j} / p_j)] / w \}$$

#seg 代表模組出現於給定序列中的次數；w 表示模組的大小，爲第一個調控序列的長度加上第二個調控序列長度之總和；$q_{i,j}$ 則是模組（$M_{(1)}$ 以及 $M_{(2)}$）中第 i 個位置出現核甘酸 j 的機率；p_j 則是從背景模型中出現核甘酸 j 的機率。

Bioprospector 最主要的貢獻在於依靠較爲高階的背景模型增進了 Gibbs Sampling 的預測精確度，並且提供了預測雙元調控序列模組的功能。這啓發了我們針對調控模組的探測初步認識。然而，模組中相同調控序列之間有相當程度的一致性之外，不同調控序列之間的間距也有著高度的關連性。因此，指針對調控序列的一致性來進行探測仍然顯得不足。

2.6 SeSiMCMC【Favorov *et. al.* 2004】

SeSiMCMC 亦延伸 Gibbs Sampler【Lawrence *et. al.* 1993】的概念來預測單一調控序列或是雙元的調控序列模組。最大不同的地方在於 SeSiMCMC 可以自動調整調控序列本身的長度並且預測兩條控序列之間的間距。(圖2.6.1)。其演算法主要是利用修改後的 Gibbs Sampler 作爲核心，等待核心程式收斂之後，再針對調控模組的大小以及兩條控序列之間的間距做調整，如此重複執行此兩大步驟，直到程式整體抵達收斂的狀態。



（**圖2.6.1**）SeSiMCMC 預測 ArcA 調控模組的 Logo【Crooks *et. al.* 2004】。此模組主要由兩個相同的調控序列（TAAC）以間距爲7個鹼基所組成。【Favorov *et. al.* 2004】。

SeSiMCMC 與 Gibbs Sampler 不甚相同的部份在於：SeSiMCMC 修改 Gibbs Sampler 在取樣階段計算方式，且 SeSiMCMC 提出了另一套的評分標準來分辨模組的優劣。

在取樣階段 SeSiMCMC 利用下式來建立新位置的機率分布，

$$P([k] \mid s_i, M_{-i}^{(t)}, B) = \frac{P(s_i \mid [k], M_{-i}^{(t)}, B) \cdot P([k] \mid M_{-i}^{(t)}, B)}{P(s_i \mid M_{-i}^{(t)}, B)}$$

，其中機率 $P([k] \mid M_{-i}^{(t)}, B)$ 的計算可以簡化爲機率 $P([k] \mid M_{-i}^{(t)}, B) = P([k])$ ，這是

因爲目前的模組 $M_{-i}^{(t)}$ 以及背景與模組出現位置等事件假設爲機率上的獨立事件。 $P([k]\mid s_i, M_{-i}^{(t)}, B)$ 可解釋爲，以目前所見的模組模型 $M_{-i}^{(t)}$ ，背景模型 B 以及欲探知調控序列位置的序列 s_i，新模組位置出現在序列 s_i 上的第 k 位置的機率。而 $P(s_i \mid [k], M_{-i}^{(t)}, B)$（在觀察到模組模型 $M_{-i}^{(t)}$，背景模型 B 以及在序列 s_i 調控序列位置爲 k 的前提下，序列 s_i 的出現機率）的計算方式如下，

$$P(s_i \mid [k], M_{-i}^{(t)}, B) = \prod_{i=1}^{k-1} b(r_i) \prod_{i=k}^{k+w-1} m_{-i}^{(t)}(i-k+1, r_i) \prod_{i=k+w}^{L(S_i)-w+1} b(r_i) \text{ ，} k \neq 0$$

。其中 $b(r_i)$ 表示於背景中觀察序列第 i 位置核甘酸的出現機率； $m_{-i}^{(t)}(j, r)$ 則是於目前的模型 $M_{-i}^{(t)}$ 中觀察到第 j 個位置的核甘酸爲 r 的機率；w 表示目前的模型的長度。 $P(s_i \mid M_{-i}^{(t)}, B)$ 則可以利用 $P(s_i \mid [k], M_{-i}^{(t)}, B)$ 來計算得知：

$$P(s_i \mid M_{-i}^{(t)}, B) = \sum_{k=0}^{L(S_i)-w+1} P(s_i \mid [k], M_{-i}^{(t)}, B) \cdot P([k])$$

。在此定義 k=0 代表序列 s_i 上無調控序列， $P(s_i \mid [0], M_{-i}^{(t)}, B) = \prod_{i=1}^{L(S_i)-w+1} b(r_i)$ 。值得一提的是關於先驗機率 $P([k])$ 的計算方式：

$$P([k] = \frac{1}{L(s_i) - w + 1}(1 - P([0])) \text{ ，} k \neq 0$$

。$P([0])$ 則需要使用者來設定，表示輸入序列中不存在模組的機率。透過先驗機率 $P([k])$ 的定義，便可以完整的估計 $P([k] \mid s_i, M_{-i}^{(t)}, B)$ 機率分布，並且以此來選擇於序列 s_i 上新的調控序列位置。

在評定模組的優劣部份，SeSiMCMC 不只使用了原先 Gibbs Sampler 的評分方式，另外地創立新的評分公式，（G 表示輸入的序列群）

$$I_{spatial} = \sum_{s_i \in G} \sum_{k=0}^{L(s_i)-w+1} P([k] \mid s_i, M, B) \log_2 \left(\frac{P([k] \mid s_i, M, B)}{P([k])} \right)$$

，此式子可視爲另一形式的相對亂度（Relative Entropy 或 Kullback-Leibler distance），主要爲希望最後模組出現位置的機率分布必定要與原先所預定的 $P([k])$ 分布有一定程度上的差異。除了上式與眾不同的評分方式之外，原先於 Gibbs Sampler 中所使用的評分公式也會納入考慮，

$$I_{struct} = \sum_{i=1}^{w} \sum_{j \in \{A,C,G,T\}} c(i,j) \log_2 \left(\frac{m(i,j)}{b(j)} \right)$$

，$c(i,j)$ 代表目前模組中第 i 位置上核甘酸 j 出現的個數。而最後將 $I_{spatial}$ 與 I_{struct} 最整合來形成模組的評分 ICP（the highest information content per site position）：

$$ICP = \left(\frac{I_{struct} + I_{spatial}}{w + C} \right)$$

，C 爲模組出現於整個輸入序列中的總次數。

SeSiMCMC 透過上述的修改 Gibbs Sampler 取樣步驟以及評分公式，便依照 Gibbs Sampler 的流程來逼近 ICP 的最佳值。而針對雙核心的調控模組，SeSiMCMC 會在每一核心間距設定下執行修改後的 Gibbs Sampler，等待收斂後

再針對調控模組的大小以及兩個核心的控序列之間的間距做調整，重複地執行修改的 Gibbs Sampler 以及長度與間距的調整兩大步驟，直到程式整體抵達收斂。

2.7 總結

在本章節的最後，我們將之前介紹的工具或演算法以表格的方式做個簡單的總結。（表 2.7.1）。

預測工具	考慮調控序列之間的間距	自動調整調控序列長度	描述序列所使用的模型	使用的技巧
MEME	✘	✔	PWM	EM
Dyad Analysis	✔	✔	Consensus word	Words Counting
Gibbs Sampling	✘	✘	PWM	Gibbs Sampling
Bioprospector	✔	✘	PWM	Gibbs Sampling
SeSiMCMC	✔	✔	PWM	Gibbs Sampling

（**表 2.7.1**）簡介各項工具的特徵

此外，我們也將這些工具所需要的參數詳細的列表（表 2.7.2）。

預測工具	描述序列所使用的模型	必須使用的參數
MEME	PWM	1. 調控序列的長度範圍（MEME 會針對此長度範圍之內的搜尋空間來探測實際的調控序列）。 2. 每一條輸入序列群中所能包含的調控序列之總數。有三種選擇：每條序列恰好包含唯一一個調控序列；每條序列包含零個或是一個調控序列；每條序列恰好包含多個調控序列。
Dyad Analysis	Consensus word	1. 調控模組中兩個核心的大小（網頁上只允許長度爲 3 的核心大小）。 2. 調控模組中兩個核心之間的距離（網頁上只允許長度變化於 0 至 20）。 3. 背景模型的選擇。
Gibbs Sampling	PWM	1. 調控序列的長度。
Bioprospector	PWM	1. 調控序列的長度。 2. 對於兩個不同調控序列之間間距範圍的設定，Bioprospector 會於這範圍之內選擇出一致性最高的兩個調控序列。 3. 背景模型的選擇。 4. 每一條輸入序列群中所能包含的調控序列之總數。有兩種選擇：每條序列包含零個或是多個調控序列；每條序列恰好包含多個調控序列。
SeSiMCMC	PWM	1. 調控模組的長度範圍。 2. 模組中是否包含兩個不同調控序列。 3. *P*([0]) 數值的設定。*P*([0]) 表示輸入序列中不存在模組的機率。

（**表 2.7.2**）簡介各項工具所需要的參數設定

第三章 演算法與系統架構

3.1 系統流程

系統整體流程見（圖3.1）。本系統中最爲重要的是爲模擬退火元件。它是被使用來尋求調控模組的核心程式。然而，模擬退火法爲一種隨機程序的演算法，因此，初始化的狀態，隨機變數的取得，以及使用者輸入的參數都會影響最後成果的品質良宥。本章，我們討論的重點爲本研究的系統核心演算法。

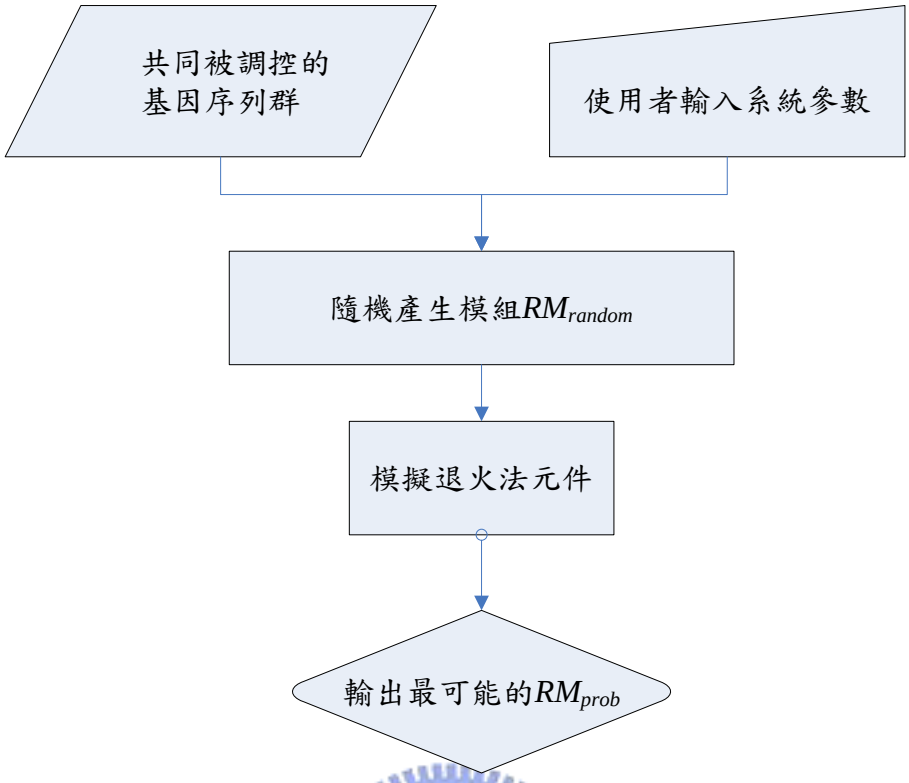
使用者可以控制的參數包括，調控模組所包含的調控序列數目，調控序列的長度，每條序列包含最多的組態個數，以及退火策略。我們的系統會針對這些輸入的參數，從輸入序列群中探測出可能的調控模組。

A. 調控序列種類：定義調控模組包含調控序列的種類。
   - 參數，-c，預設爲 2，表示調控模組包含 2 種不同的調控序列。

B. 調控序列的長度：定義不同的調控序列之長度。
   - 參數，-w，預設爲10,10，兩種不同的調控序列之長度分別爲10。

C. 調控模組個數的上限：定義一條基因序列中調控模組出現的最多次數。
   - 參數，-m，預設爲 2，表示每條序列最多能提供 2 組調控模組。
   - 此參數是爲了避免系統於初始化時過度地包羅每條基因序列中可能的調控模組。

D. 退火策略：
   - 由兩種參數代表。
     1. -T，表示系統的初始溫度設定。預設爲100，表示系統的初始溫度設定爲100。
     2. -d，表示退火時降溫的比例。預設爲0.98。


共同被調控的
基因序列群
使用者輸入系統參數
隨機產生模組RM_{random}
模擬退火法元件
輸出最可能的RM_{prob}

(圖3.1) 系統流程圖

3.2 調控序列與調控模組模型

假設調控模組中有 m 種不同的調控序列，代號分別爲1，…，m。我們首先介紹需要使用的基本符號：（圖3.2.1）。

- ◆ G：使用者輸入基因序列之集合，$G = \{ g_i \mid g_i$表示每一條基因序列，$1 \leq i \leq n \}$。
  - $|G| = n$，表示總共的序列數目。
  - $|g_i| = l_i$，表示序列 g_i 的長度。
- ◆ $G.RM$ 表示基因序列 G 的調控模組之集合。

$$G.RM = \bigcup_{g_i \in G} g_i.RM$$

- $g_i.RM$ 表示基因序列 g_i 所擁有的調控模組。

  $g_i.RM = \{ g_i.RM_j \mid 1 \leq j \leq h$，若 g_i 擁有 h 個調控模組 $\}$

- $g_i.RM_j$ 表示基因序列 g_i 的第 j 個調控模組。

$$g_i.RM_j = (d_S, d_{12}, d_{23}, \ldots\ldots, d_{m\text{-}1,m})$$

  - 模組中擁有 m 種不同的調控序列。
  - d_S，調控模組中第一個調控序列出現的位置。
  - $d_{i,j}$ 則爲第 i 個調控序列與第 j 個調控序列之間的相對距離。$1 \leq i \neq j \leq m$。

- ◆ $G.C$：所有可能的調控模組（Possible Regulatory Module，簡稱 PRM）的集合。

$$G.C = \bigcup_{g_i \in G} g_i.C$$

$$G.C = G.C^{[-G.RM]} \cup G.RM$$

- 基因序列 g_i 中所有 PRM 的集合以 $g_i.C = \{ g_i.c_j \mid 1 \leq j \leq p \}$，$p$ 表

示基因序列 g_i 中所有 PRM 的個數。

- 稱 $g_i.c_j = (d_S, d_{12}, d_{23}, \ldots\ldots, d_{m-1,m})$ 爲基因序列 g_i 某一可能的調控模組（PRM） j，若 $0 \leqq (d_S + \Sigma d_{i,j}) \leqq l_i$。

- $g_i.C^{[-G.RM]}$：表示在序列 g_i 中不屬於 $g_i.RM$ 的所有 PRM 的集合。

$$g_i.C_i^{[-RM]} = g_i.C - g_i.RM$$

- $G.C^{[-G.RM]}$：不包含於 $G.RM$ 中所有 PRM 的集合。

$$G.C^{[-G.RM]} = \bigcup_{g_i \in G} g_i.C^{[-G.RM]}$$

◆ $G.B$：背景（Background）。基因序列群 G 中非調控模組的序列。

◆ $\Theta = (\Theta_0, \Theta_1, ..., \Theta_i, ..., \Theta_m)$。$\Theta_i$ 表示第 i 個調控序列的位置比重矩陣，而 Θ_0 則代表序列群 G 中非調控序列的部份（背景）所形成的機率分布。

- $\Theta_i = (\theta_1, ..., \theta_j, ..., \theta_{w_i})$，$\theta_j$ 可視爲第 i 個調控序列第 j 個位置四種核甘酸的機率分布。$i \neq 0$。

- $\theta_{i,j}^r$ 則爲第 i 個調控序列第 j 個位置核甘酸 r 的出現機率。$r \in \{A, C, G, T\}$。$i \neq 0$。

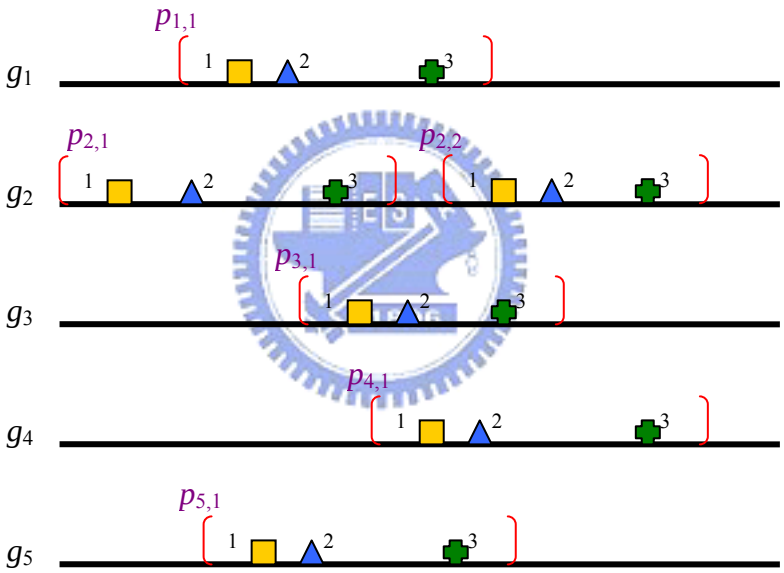
- 理論上 Θ_0 可以利用 k-階的馬可夫鏈模型（k^{th}-order Markov Chain Model）來描述。在本研究的實做中，我們使用 0 階的馬可夫鏈模型（Zero-order Markov Chain Model）來表示背景字元出現的機率。

◆ $\Gamma = \{ \Gamma_{1,2}, \ldots, \Gamma_{i,i+1}, \ldots, \Gamma_{m-1,m} \}$。$\Gamma_{i,i+1}$ 表示第 i 個調控序列與第 i+1 個調控序列相對距離的機率模型。

- 在本系統中，我們以較爲簡單的方式來實做，亦即觀察目前調控模組中

第 i 個調控序列與第 i+1 個調控序列在各條序列中的相對距離來建立機率分布。（圖 3.2.1）。

- A：表示調控模組出現的起始位置，爲一個矩陣。
  - $A = \{ a_{ij} \mid a_{ij} = 1$，表示在序列 g_i 上第 j 個位置爲模組的起始位置。$a_{ij} = 0$，表示在 g_i 上第 j 個位置爲非調控模組的起始位置。$\}$
- $D = \{ D_{1,2}, \ldots, D_{i,i+1}, \ldots, D_{m-1,m} \}$。$D_{i,i+1}$紀錄目前模組中第 i 個調控序列與第 i+1 個調控序列在各條序列中的相對距離。



（**圖 3.2.1**）輸入的基因序列，$G = \{ g_i \mid 1 \leq i \leq 5 \}$。圖中的調控模組包含三種調控序列，調控序列 p_1(黃正方)，調控序列 p_2(藍三角)與調控序列 p_3(綠十字)。長度各別爲 w_1，w_2 以及 w_3。每條序列的調控模組均由粉紅色括弧所顯示。$p_{i,j}$ 則記載著在第 i 條序列上第 j 個模組出現的位置。舉例來說，$g_2.RM = \{ g_2.RM_1, g_2.RM_2 \}$，$g_2.RM_1 = \{ p_{2,1}, d_{12}, d_{23} \}$，$g_2.RM_1 = \{ p_{2,2}, d'_{12}, d'_{23} \}$。倘若此時我們觀察到 p_1 與 p_2 的相對距離爲 $g_1.RM_1.\ d_{1,2}$=2, $g_2.RM_1.\ d_{1,2}$=4, $g_2.RM_2.\ d_{1,2}$=2, $g_3.RM_1.\ d_{1,2}$=2, $g_4.RM_1.\ d_{1,2}$=2, $g_5.RM_1.\ d_{1,2}$=2，則我們可以計算其出現頻率分布成爲 $\Gamma_{1,2} = \{\ p(d=2)=5/6, p(d=4)=1/6 \}$。

3.3 核心評分公式

爲了引導出最後的核心評分公式，我們藉助上一小節介紹的符號來協助推論。我們知道某個特定的調控序列模型，調控模組出現的位置，調控序列相對距離的機率模型以及實際出現於基因序列中調控序列的相對距離可以代表某個相對應的模組 $G.RM$。因此給基因序列群 G，我們希望能夠發現模組 $G.RM$，而其代表參數 (Θ, A, Γ, D) 並定會使得條件機率 $P(\Theta, A, \Gamma, D \mid G)$ 值達到最大。

後驗機率（Posterior Probability）$P(\Theta, A, \Gamma, D \mid G)$ 可表示成：

$$P(\Theta, A, \Gamma, D \mid G) = \frac{P(G, A, D \mid \Theta, \Gamma) P(\Theta, \Gamma)}{P(G)} \tag{1}$$

，在這樣的模型中我們可以假設各個調控序列的位置比重矩陣（Θ）與其相鄰的調控序列之間的相對距離模型（Γ）爲相互獨立（Independent），因此，整個式子可以簡化成，

$$P(\Theta, A, \Gamma, D \mid G) = \frac{P(G, A, D \mid \Theta, \Gamma) P(\Theta) P(\Gamma)}{P(G)} \tag{2}$$

。若將調控序列的位置比重矩陣（Θ）以及調控序列相對距離機率分布（Γ）視爲已知的資料，則輸入序列群 G 與 A，D 整體的條件機率便可寫成：

$$P(G, A, D \mid \Theta, \Gamma) = P(G, D \mid \Theta, A, \Gamma) P(A \mid \Theta, \Gamma) \tag{3}$$

。其中 $P(G, D \mid \Theta, A, \Gamma)$ 的計算方式：

$$P(G, D \mid \Theta, A, \Gamma) = P(G \mid \Theta, A, \Gamma, D) P(D \mid \Theta, A, \Gamma)$$

。而機率 $P(G \mid \Theta, A, \Gamma, D)$ 的推導則可以參考【Jensen *et. al.* 2004】【Lawrence *et. al.* 1993】【Liu *et. al.* 1995】【Favorov *et. al.* 2004】等人的立論而得：

$$P(G \mid \Theta, A, \Gamma, D) = \prod_{i=1}^{|G|} \prod_{p=1}^{|l_i|} \theta_{j,k}^{g_i(p)} \quad , \quad 0 \le j \le m \quad , \quad 1 \le k \le w_j$$

，其中 $\theta_{j,k}^{g_i(p)}$ 爲序列 g_i 上第 p 位置核甘酸（以 $g_i(p)$表示之）的出現機率，其機率可以依據背景模型或是由第 j 種調控序列上第 k 個位置的機率決定。判定序列 g_i 第 p 個位置上的核甘酸屬於背景或是第 j 種調控序列，可以依靠模組起始矩陣 A 以及其序列相對距離矩陣 D 加以輔助計算而得知。實際上想要確定 $P(D \mid \Theta, A, \Gamma)$ 的計算方式是相當困難的，然而由於之前的假設，「各個調控序列的位置比重矩陣（Θ）與其相鄰的調控序列之間的相對距離模型（Γ）爲相互獨立」，因此我們可以簡化 $P(D \mid \Theta, A, \Gamma) = P(D \mid \Gamma)$。若單純地想要直接透過調控序列模型（$\Theta$）以及調控序列相對距離的機率模型（$\Gamma$）來得知基因序列中模組出現的位置實爲天方夜譚，因此我們假設調控序列模型（Θ）以及調控序列相對距離的機率模型（Γ）與基因序列中模組出現的位置的出現（A）爲機率上的獨立事件，則式(3)中的 $P(A \mid \Theta, \Gamma)$ 便可簡化成 $P(A \mid \Theta, \Gamma) = P(A)$。基於以上的推論，我們便可以將式(3)改寫成，

$$P(G, A, D \mid \Theta, \Gamma) = \prod_{i=1}^{|G|} \prod_{p=1}^{|l_i|} \theta_{j,k}^{g_i(p)} \cdot P(D \mid \Gamma) \cdot P(A) \quad , \quad 0 \le j \le m \quad , \quad 1 \le k \le w_j \qquad (4)$$

將上述式子(2)(4)加以整理，再套用【Mitchell, 1997】中推論 *MAP*（maximum a posteriori）的方式，可得：

$$(\Theta, A, \Gamma, D)_{MAP} \equiv \arg\max_{(\Theta, A, \Gamma, D) \in S} P(G, A, D \mid \Theta, \Gamma) P(\Theta) P(\Gamma) \tag{5}$$

，其中 $(\Theta, A, \Gamma, D)_{MAP}$ 表示最可能的模組參數；S 則爲所有模組參數所組成的集合。然而在實際情況中，計算 $P(G, D \mid \Theta, A, \Gamma) P(A) P(\Theta) P(\Gamma)$ 爲十分困難的一件事。原因出自於先驗機率 $P(A)$，$P(\Theta)$ 以及 $P(\Gamma)$ 是未知的機率分布，我們無法得知。因此爲了推論與計算上的便利，我們假設 A，Θ 以及 Γ 各自出現的先驗機率（Prior Probability）分布爲隨機分布（Uniform Distribution），即 $\forall i \neq j, P(A_i) = P(A_j)$，$\forall k \neq l, P(\Theta_k) = P(\Theta_l)$ 與 $\forall m \neq m, P(\Gamma_m) = P(\Gamma_n)$。則式(6)便可以輕易推得。

$$(\Theta, A, \Gamma, D)_{MAP} \equiv \arg\max_{(\Theta, A, \Gamma, D) \in S} P(G, D \mid \Theta, A, \Gamma) \tag{6}$$

因此，給定任兩種調控模組參數：調控模組位置矩陣 A_1，調控序列相對距離 D_1 以及分布調控序列相對距離 Γ_1 和調控序列模型 Θ_1；調控模組位置矩陣 A_2，調控序列相對距離 D_2 以及分布調控序列相對距離 Γ_2 和調控序列模型 Θ_2，我們可以利用式(6)來判斷何者較可能爲輸入序列 G 的調控模組。亦即，我們只要比較 $P(G, D_1 \mid \Theta_1, A_1, \Gamma_1)$ 與 $P(G, D_2 \mid \Theta_2, A_2, \Gamma_2)$，其值越大者，爲輸入序列 G 的調控模組的可信度越高。

由於參數 (Θ, A, Γ, D) 可以代表某個相對應的模組 $G.RM$，爲了簡化稍後的論述，我們定義，

$$Score(G.RM) = (\Theta, A, \Gamma, D)_{MAP} \tag{7}$$

。

3.4 系統核心推演方法

經過3.2節的討論，我們可以將問題思考成爲尋找一組調控模組參數，(A, D, Θ, Γ)使得$P(\Theta, A, \Gamma, D \mid G)$值爲最大。針對所有可能的調控模組參數做評估，而後選擇分數最高者是一種最爲簡單的方式，然而卻由於搜尋空間（Search Space）過於龐大而是爲最不實際可行的方法。爲此我們必須參考 Gibbs Sampling 的推論技巧，搭以模擬退火法的退火過程來加速搜尋所需要的時間以及減少搜尋所需要的空間。

3.4.1 模擬退火法

模擬退火法（Simulated Annealing）【Kirkpatrick *et. al.* 1983】與 Gibbs Sampler 同屬於馬可夫鏈蒙地卡羅（MCMC；Markov Chain Monte Carlo）演算法的推廣，是爲一種隨機程序。其主要是模擬自然界物質從高溫至冷卻結晶的過程。在大自然中，物質分子由於高溫狀態時擁有較多的能量而能夠在狀態空間中自由移動。隨著溫度適當地漸漸降溫，這些分子的能量會慢慢散失而使分子逐漸停留於某些狀態。在溫度最低時，分子重新以最穩定的結構排列。因此分子結晶的完美與否取決要素則爲降溫的過程。

模擬退火法是 Kirkpatrick 等人在1983年提出並且成功地應用於組合最佳化問題中。其概念沿襲自然界的退火步驟，但在不同溫度時分子狀態的分布則是依循波茲曼概率（Boltzmann）分布：$e^{-E(S)/T}$；T ＞0表示溫度，S 代表溫度爲 T 時的狀態，$E(S)$ 則是狀態 S 時物質擁有的能量函數。模擬退火法中最爲關鍵的一環便是退火溫度的控制與排程（退火策略；Annealing Schedule）。當溫度趨近於0時，程式可能會落入函數 E 的局部最小值（Local Minimum）。然而，當退火策略的制定夠精細時，模擬退火法將會發現函數 E 的全局最小值（Global Minimum）。（圖3.4.1）爲模擬退火法的虛擬碼（pseudo-code）。

Simulated Annealing

```
E(*) ← energy function;
T ← annealing temperature schedule;
S ← randomly initialize a state;
For n-th iteration
{
    If( S no changed || System "freezes" )
    {
        Break;
    }
    S_next ← randomly select one state;
    ΔE = E(S_next) - E(S);
    t ← T_n;
    Draw a Uniform(0,1) random variable U;
    If( U ≦ min{1, exp(-ΔE/t)} ), then S ← S_next;
}
Return S;
```

（**圖3.4.1**）模擬退火法的虛擬碼。

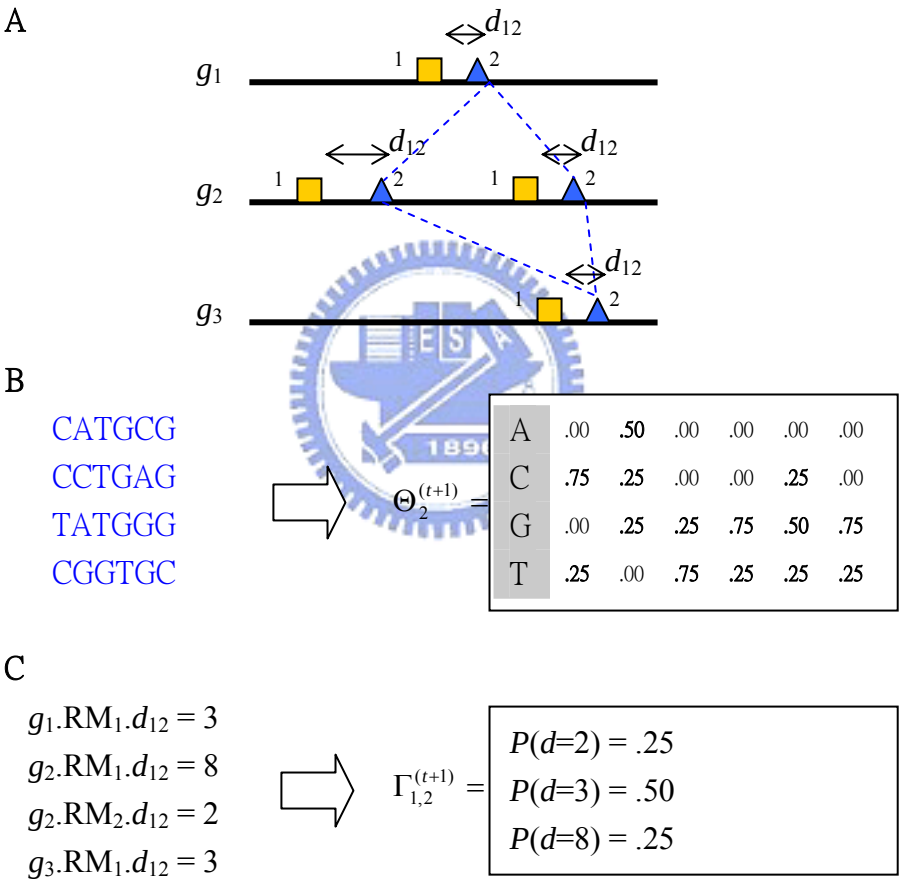
3.4.2 核心概觀

首先，我們將模組參數中的 A 與 D 視爲未知的資訊（Missing Data），利用模擬退火法的程序來執行推導的過程。程式啓動之初，會從序列群 G 中隨機地選擇模組的起始點以及調控序列之間的距離，之後，系統便反覆地執行以下步驟：（圖3.4）。

藉著片面地改變已知的調控模組位置矩陣 $A^{(t)}$ 以及調控序列相對距離 $D^{(t)}$ 而得到的 $A^{(t+1)}$ 與 $D^{(t+1)}$，我們將參數 $(\Theta^{(t)},\Gamma^{(t)})$ 更新。並且根據新參數 $(\Theta^{(t+1)},\Gamma^{(t+1)})$，我們將計算 $P(G,D\,|\,\Theta,A,\Gamma)$ 之值來測試這些新的參數 $(A^{(t+1)},D^{(t+1)},\Theta^{(t+1)},\Gamma^{(t+1)})$ 是否可以取代前一時間的舊參數 $(A^{(t)},D^{(t)},\Theta^{(t)},\Gamma^{(t)})$。

I. 若 $P(G, D^{(t+1)} \mid A^{(t+1)}, \Theta^{(t+1)}, \Gamma^{(t+1)})$ 大於 $P(G, D^{(t)} \mid A^{(t)}, \Theta^{(t)}, \Gamma^{(t)})$，選擇接受新的參數。

II. 若 $P(G, D^{(t+1)} \mid A^{(t+1)}, \Theta^{(t+1)}, \Gamma^{(t+1)})$ 小於 $P(G, D^{(t)} \mid A^{(t)}, \Theta^{(t)}, \Gamma^{(t)})$，則依照 $\exp(\Delta p / T)$ 分布來決定是否選擇接受新的參數。其中 T 表示目前退火的溫度，而 $\Delta p = P(G, D^{(t+1)} \mid A^{(t+1)}, \Theta^{(t+1)}, \Gamma^{(t+1)}) - P(G, D^{(t)} \mid A^{(t)}, \Theta^{(t)}, \Gamma^{(t)})$ 。



（**圖3.4**）(A)根據已知的 $A^{(t+1)}$ 與 $D^{(t+1)}$，我們可以對照輸入序列 G 找出相對應的子序列。(B)將這些代表各序列中的調控序列字串做排比，我們便可推導出此調控序列的位置比重矩陣。圖中是以調控序列編號2爲例，透過 $A^{(t+1)}$ 與 $D^{(t+1)}$，我們可以取得其此時刻的位置比重矩陣 $\Theta_2^{(t+1)}$ 。(C)依據相對位置記錄者 $D^{(t+1)}$，我們可以進而計算其距離的機率分布函數 $\Gamma_{1,2}^{(t+1)}$ 。

3.5 系統核心實做與架構

本研究主要的目標是探測基因序列中的調控模組，亦即搜尋出一組調控模組 *G.RM* 使得 *Score*(*G.RM*) 爲最大值。我們利用模擬退火法【Kirkpatrick *et. al.* 1983】爲核心並且參考 Gibbs Sampler 【Lawrence *et. al.* 1993】的流程，撰構出我們的主程式核心程式，SAMLA（**S**imulated **A**nnealing for **M**ultiple **L**ocal Sequences **A**lignment）。（圖3.5.1）爲虛擬碼。原始的模擬退火法中，每一次迭代只需要隨機的選取新的狀態，並且以波茲曼概率分布來決定是否改變目前狀態，進而逼近最佳解。系統若以此方式來進行預測，勢必需要花費較長的時間來收尋，方能收斂於最佳解。因此我們參考 Gibbs Sampler 的流程【Lawrence *et. al.* 1993】【Liu,J.S. 1994】【Liu *et. al.* 1995】，改進原先隨機選取新狀態的過程，亦即，我們將會針對所有可能的調控序列位置進行比較，選取分數高者與目前的狀態做比較，來決定下一迭代的新狀態。

程式之初，我們從 *G.C*（所有 PRM 的集合）中隨意選出起始的模組 $RM_{current}$。每一次的迴圈中，我們針對每條輸入之基因序列 g_i 執行三種不同的操作來嘗試改變 $RM_{current}$ 的狀態，分別爲：加入新的 PRM（ADD-Operator），移除一組已存在的 PRM（DELETE-Operator），將一組已存在的 PRM 與其他代換（SWAP-Operator）。除此之外，由於模擬退火法的原始設計爲尋找能量函數 *E*(*) 的最小值，但是我們現在的問題卻是必須尋找 *Score*(*) 的最大值，因此我們修改了狀態變動時的機率計算方式：原先是計算 $exp(-\Delta E/t)$ 值，我們改成計算 $exp(\Delta E/t)$ 來符合尋求最大值的要求。

(1) $Score(*) \leftarrow$ energy function (definition in 3.2);
(2) $T \leftarrow$ annealing temperature schedule;
(3) $G.RM_{current} \leftarrow$ randomly initialize from $G.C$;
(4) While(!($Score(G.RM_{current})$ no changed || System "freezes" || $iteration > 200$))
{
(a) For each $g_i \in G$, DO
{
i. $G.RM_{next} \leftarrow \max\{$ ADD-Operator(g_i,$G.RM_{current}$),
DELETE-Operator(g_i,$G.RM_{current}$),
SWAP-Operator(g_i,$G.RM_{current}$) $\}$;
ii. $\Delta E = Score(G.RM_{next}) - Score(G.RM_{current})$;
iii. Draw a Uniform(0,1) random variable U;
iv. If($U \leqq \min\{1, exp(\Delta E/t)\}$), then
$G.RM_{current} \leftarrow G.RM_{next}$;
}
(b) $G.RM_{current} \leftarrow$ SHIFT-Operator($G.RM_{current}$);
(c) $t \leftarrow T_n$;
}
(5) Return $G.RM_{current}$ as $G.RM$;

（**圖3.5.1**）SAMLA的虛擬碼

- ADD-Operator：走訪基因序列 g_i 中所有的 PRM，$c \in g_i.C^{[-G.RM]}$，嘗試加入新的 PRM 於現有的模組。計算 $Score(G.RM_{current} \cup c)$。最後回傳分數最高的新調控模組 $G.RM^{+c}$。時間複雜度爲 $O(l_i^m)$，m 表示調控序列的各數，l_i 則是目前正在操作的序列 g_i 的長度。

ADD-Operator
For all $c \in g_i.C^{[-G.RM]}$, DO
{
Calculate $Score(G.RM_{current} \cup c)$;
}
Return the best scored $G.RM^{+c}$;

- DELETE-Operator：此操作會移除目前 $G.RM_{current}$ 中屬於基因序列 g_i 的一個 PRM，$c \in g_i.RM_{current}$。計算 $Score(G.RM_{current}/c)$。最後回傳 $Score(G.RM_{current}/c)$ 分數最高的新調控模組 $G.RM^{c}$。時間複雜度爲 $O(max\_sites\_per\_seq)$，$max\_sites\_per\_seq$ 爲使用者輸入的參數，代表每個輸入的基因序列中最多允許模組的個數。

DELETE-Operator

For all $c \in g_i.RM_{current}$, DO
{
    Calculate $Score(G.RM_{current} / c)$;
}
Return the best scored $G.RM^{c}$;

- SWAP-Operator：將目前 $g_i.RM_{current}$ 中的所有 PRM 與 $g_i.C^{[-G.RM]}$ 中的 PRM 作置換動作。此操作會針對所有的 $c_{RM} \in g_i.RM_{current}$ 以及 $c_{SWAP} \in G.C^{[-G.RM]}$ 來嘗試置換的結果，時間複雜度爲 $O(l_i^m)$，m 表示調控序列的各數，l_i 則是目前正在操作的序列 g_i 的長度。此操作回傳置換之後分數最高的新調控模組 $G.RM^{SWAP}$。

SWAP-Operator

For all $c_{RM} \in g_i.RM_{current}$, DO
{
    $S \leftarrow G.RM_{current} / c_{RM}$;
    For all $c_{SWAP} \in g_i.C^{[-G.RM]}$, DO
    {
        Calculate $Score(S \cup c_{SWAP})$;
    }
}
Return the best scored $G.RM^{SWAP}$;

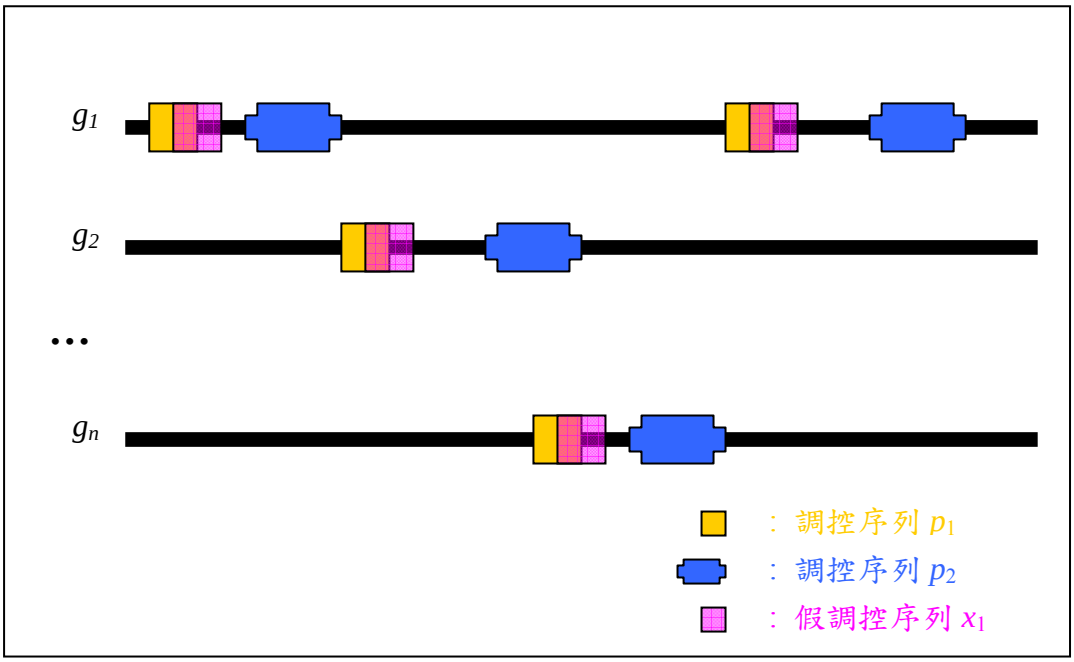
最後，爲了避免程式落入區域最佳位移解（圖3.5.2），我們也設計

SHIFT-Operator 來避開此種困境【Lawrence *et. al.* 1993】。

- SHIFT-Operator：爲了避免程式落入全域最佳位移解，我們在每一次迭代之後，將目前 $G.RM_{current}$ 中每一個調控序列 p_i 向左以及向右移動一個核甘酸位置。最後回傳分數最高的移動方式。SHIFT-Operator 的時間複雜度爲 $O(m)$ 。m 表示調控序列的個數。

SHIFT-Operator

$RM_{shift} \leftarrow$ NULL;
For all $p_k \in G.RM_{current}$, DO
{
(1) $A_{left} \leftarrow A_{current}$;
(2) If($a_{i,j} == k$), then set $a_{i,j} = 0$ and $a_{i,j-1} = 1$, for $1 \leqq i \leqq n$, $1 \leqq j \leqq l_i$, and $a_{i,j} \in A_{left}$;
(3) Form one new RM_{left} by A_{left} and D;
(4) $A_{right} \leftarrow A_{current}$;
(5) If($a_{i,j} == k$), then set $a_{i,j} = 0$ and $a_{i,j+1} = 1$, for $1 \leqq i \leqq n$, $1 \leqq j \leqq l_i$, and $a_{i,j} \in A_{left}$;
(6) Form one new RM_{right} by A_{right} and D;
(7) If($Score(RM_{right}) > Score(RM_{left})$), then $RM_{shift} \leftarrow RM_{right}$;
Otherwise, $RM_{shift} \leftarrow RM_{left}$;
}
If($Score(G.RM_{current}) > Score(RM_{shift})$), then return $G.RM_{current}$;
Otherwise, return RM_{shift};



（**圖3.5.2**）黑色粗線爲輸入的序列，$G = \{ g_i \mid 1 \leq i \leq n \}$。假設真正的調控模組 *G.RM* 由 p_1，p_2 所組成。而又有另一調控模組 *G.RM'* 由 x_1，p_2 所組成。程式在迭代搜尋 *G.RM* 時，會因爲 x_1 與 p_1 重疊而無法找出比 *Score*(*G.RM'*) 更高分數的 *G.RM''* 而停止程式。因此在每次迭代搜尋時必須測試位移調控序列 p_i，來解決落入區域最佳位移解的困境。

第四章 實驗結果與分析

4.1 Zn群組調控因子【van Helden *et. al.* 2000】

在【van Helden *et. al.* 2000】中提到了所謂的「雙核心」的調控模組。由於調控蛋白以二聚化合物（Dimer）的複合體樣式而形成雙核心的調控蛋白，這使得此二核心在與 DNA 序列的結合區域呈現出以較小、較爲相似的樣式。在此型態下的調控模組其兩調控序列（調控蛋白雙核心部份與 DNA 序列的結合區域）之間必定有著固定的間距。我們收集【van Helden *et. al.* 2000】中提到與酵母菌 Zn 群組調控因子所調控的生物基因序列來測試系統的能力。基因序列以及調控模組的資訊詳見（表4.1.1），所有的序列均取自於ORF前 800 個長度的核甘酸序列。在此實驗中系統預設的參數爲：收尋兩個相同長度的調控序列，長度分別爲 3，其間距爲可變動等參數進行收尋。

SAMLA 收尋的結果，詳見（表4.1.1）與（表4.1.2）。（表4.1.2）中詳列了系統預測樣式的 Logo【Crooks *et. al.* 2004】。SAMLA 準確地預測了 6 個家族的調控模組，分別爲 GAL4，CAT8，LEU3，LYS，PPR1，PUT3，UGA3，UME6 等家族。其中由於 PUT3 家族中包含的基因序列個數太少（只有 2 條），因此長度爲 3，間距固定的兩個調控序列核心隨處可見，例如 CGGN[10]GCC（真正的模組樣式）與 CGGN[3]GCC，在我們的評分中這兩種樣式的分數是完全一樣，因此我們需要更多被 PUT3 調控的基因序列來幫助判定更精確的調控模組樣式。

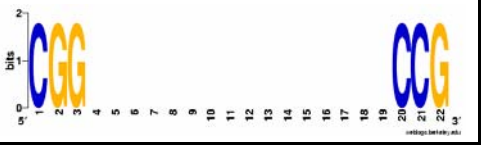
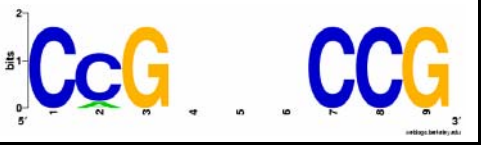
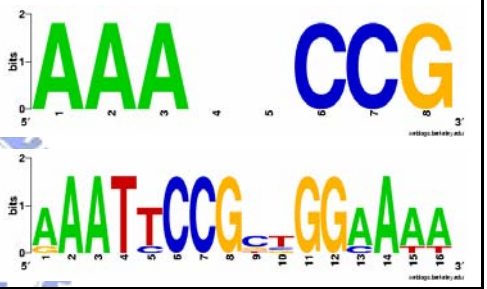
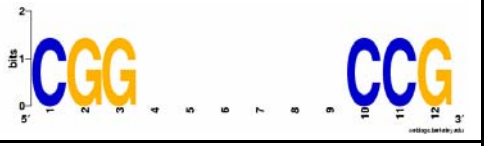
然而在長度爲 3 的兩調控序列與可變間距的參數設定下，SAMLA 並不能發現 HAP1 與 PDR 家族的正確樣式。其中 HAP1 家族，系統探測的最高分數的樣式爲 GGGn{3,12}GGC（n{3,12}代表系統探尋出GGG與GGC兩樣式之間的距離爲 3 或 12），與已知的樣式（CGGnnnTAnCGG）明顯的不同。爲此，我

家族	基因序列	已知的樣式	SAMLA 預測樣式	dyad analysis 預測樣式
GAL4	GAL1 GAL2 GAL7 GAL80 MEL1 GCY1	CGGRnnRCYnYnCnCCG	CGGnnnnnnnnnnnCCG	TCGGAnnnnnnnnnTCCGA
CAT8	ACR1 ICL1 MLS1 PCK1 FBP1	CGGnnnnnnGGA	CGGnnnnnnGGA	CGGnnnnATGGAA
*HAP1	CYB2 CYC1 CYC7 CTT1 CYT1 ERG11 HEM13 HMG1 ROX1	CGGnnnTAnCGG CGGnnnTAnCGGnnnTA	GGAnnnnnCGG GGAnnnnnnGGC	GGAnnnnnCGGC
LEU3	GDH1 ILV1 LEU1 LEU2 LEU4	RCCGGnnCCGGY	CGGnnnCGG	ACCGGCGCCGGT
LYS	LYS1 LYS2 LYS4 LYS9 LYS20 LYS21	WWWTCCRnYGGAWWW	AAAnnCCG AAATYCCGnnGGMAWW	AAATTCCG TCCGCTGGA
*PDR	YOR1 PDR11 PDR10 GAS1 STE6 SNQ2 PDR5	TYTCCGCGGARY TCCGCGGA TCCGTGGA	CCGYGGAA	TTCCGCGGAA
PPR1	URA1 URA3 URA4	WYCGGnnWWYKCCGAW	CGGnnnnnnCCG	CGGnnnnnnCCG
PUT3	PUT1 PUT2	YCGGnAnGCGnAnnnCCGA CGGnAnGCnAnnnCCGA	CGGnnnnnnnnnnGCC CGGnnnGCC	CGGnnnnnnnnnCCG
UGA3	UGA1 UGA4 YBR006W	AAARCCGCSGGCGGSAWT	CCGnnGGC	GCCGnCGGCGGC
UME6	BAR1 CAR1 CAR2 DMC1 GAL1 HOP1 HSF1 ILV2 IME1 IME2 INO1 MEI4 MER1 REC102 REC114 RED1 RME1 SPO11 SPO13 SPO16 TOP1 ZIP1	TAGCCGCCGA	GCCGCC	TAGCCGCCGA

（**表4.1.1**）序列資訊【van Helden *et. al.* 2000】。

[A] 收錄【van Helden *et. al.* 2000】所使用的雙核心調控模組資訊，欄位“家族”代表調控蛋白的名稱；“基因序列”則是被此蛋白所調控的基因序列群；經由生物實驗驗證出的模組樣式則置於欄位“已知的樣式”中；“SAMLA”則顯示我們系統所預測的樣式；最後欄位“dyad analysis”則是取自於【van Helden *et. al.* 2000】中的分析結果。所有的樣式均以 $W_1.N_S.W_2$ 表示。$W_i = x_1\ x_2\ x_3$，$x_j \in \{A,C,G,T\}$，$i = 1$或2，$j = 1,2$或3。N_S 代表兩個調控序列之間的間距，以“n”的個數表示距離長度，例如：CAT8的樣式CGGnnnnnnGGA，其中CGG與GGA之間的間距固定爲6，因此以六個“n”來表示。若間距爲0，則中間沒有“n”表示。

[B] *：表示 SAMLA 無法利用長度爲 3 的兩調控序列與可變間距的參數來正確的探測出實驗已知的樣式。因此在內文說明中我們有詳細的說明如何利用不同的參數來探測出正確的答案

家族	已知的樣式	預測 Logo
GAL4	**CGG**RnnRCYnYnCn**CCG**	
CAT8	**CGG**nnnnnn**GGA**	
HAP1	C**GG**nnnTAn**CGG** CGGnnnTAnCGGnnnTA	
LEU3	R**CCG**Gnn**CCG**GY	
LYS	**WWWTCCRnYGGAWWW**	
PDR	TYT**CCGCGGA**RY T**CCGCGGA** T**CCGTGGA**	
PPR1	WY**CGG**nnWWYK**CCG**AW	
PUT3	YCGGnAnGCGnAnnnCCGA **CGG**nAnGCnAnnnC**CGA**	
UGA3	AAAR**CCG**CS**GGC**GGSAWT	
UME6	TA**GCCGCC**GA	

（**表4.1.2**）模組樣式的 Logo。

們跟隨【van Helden *et. al.* 2000】中提到 dyad analysis 的方式來作分析：我們讓系統搜尋兩個調控序列長度爲 3 且固定間距的模組。間距設定由 0 至 20，每個間距設定中，我們讓系統執行五次，選取分數最高者作爲此間距最爲可能的模組。經過總共 105 次的收尋，我們發現間距爲 5，12 以及 6 的樣式分數各爲第一、二、三名，樣式分別爲 GGAnnnnnCGG，GGGnnnnnnnnnnnnGGC 和 GGAnnnnnnGGC。實驗證實的樣式與這兩個樣式之間以及 dyad analysis 所預測樣式的排比如下：

已知樣式	CGGnnnTAnCGG
dyad analysis預測樣式	GGAnnnnnCGGC
系統預測樣式	GGAnnnnnCGG GGAnnnnnn GGC
	* * * * *

我們發現，一、我們將間距爲五以及六所預測的兩樣式作重組，便與 dyad analysis 所預測樣式一致（GGAnnnnnCGGC）。二、不論是 dyad analysis 或是 SAMLA 所預測的樣式與實驗證實的樣式之間的相似度都非常的高。由於高度的相似性，我們不諱言的認定我們的系統亦可以利用固定間距的方式發現 HAP1 家族的正確樣式。然而最讓我們訝異的是 PDR 家族的探索失敗。在兩長度爲 3 且可變間距的調控序列參數設定之下，系統預測 PDR 家族的樣式爲 GGTGCC，此樣式與已知的樣式十分不一致。接下來，我們重新設定系統使用參數，只搜尋長度爲 8 的單一調控序列模組，並且讓 SAMAL 以此設定執行五次，選擇分數最高的樣式。SAMAL 探測得到的樣式爲 CCGYGGAA，與實驗證實的樣式排比如下：

已知樣式	TCCGCGGA TCCGTGGA
系統預測樣式	CCGYGGAA

* * * * * * *

4.2 大腸桿菌中的雙核心模組

在這部份的實驗中，我們自現有的大腸桿菌基因體分析資料中取得 6 種與雙核心模組相關的基因序列來測試我們的系統。這些調控蛋白分別是 ArcA-P、CRP、TyrR、cpxR、narL、rpoH 等。接下來我們將逐一介紹 SAMLA 對於這些調控蛋白的測試結果。

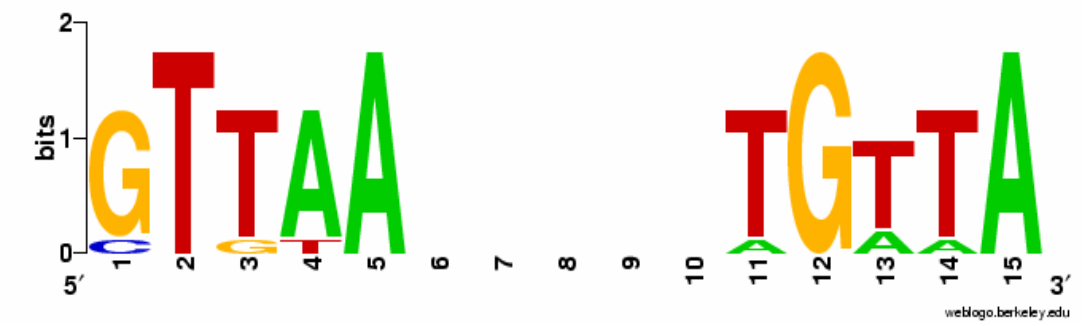
4.2.1 Phospho-ArcA【Favorov *et. al.* 2005】

Arc 系列調控蛋白主要是存在於大腸桿菌中負責調控細菌體於有氧與無氧環境度過時的生理變化。當處於無氧環境時， Phospho-ArcA（ArcA-P）調控蛋白會開始作用，一方面抑制某些蛋白質的活動（e.g. *icd*, *lld*, *glt*, *glc*, *sdh* and *soda*），另一方面也會刺激某些調控因子（e.g. *cyd* and *pfl*）來協助細菌體適應無氧環境。我們從 http://favorov.imb.ac.ru/SeSiMCMC/ 取得相關基因序列資料。共有9條序列，長度最長的爲844，最短爲192，平均長度爲375.2。由 SAMLA 探測出此調控模組主要由兩個十分相似的調控序列所組成，兩條控序列之間的相對距離恰巧均爲 7 個核甘酸，此兩條控序列樣式爲 GTTA。（圖4.2.1 a）。

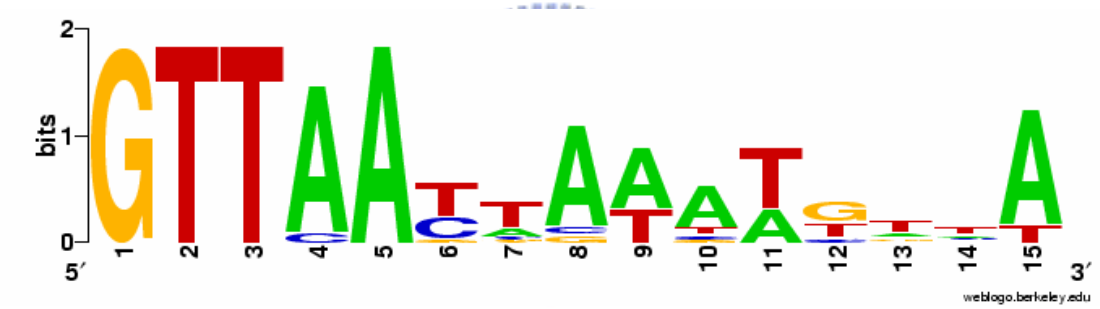
（圖4.2.1 a）顯示系統探測出顯著的雙重調控序列 GTTAA 與 TGTTA，且雙重調控序列的間距爲 5。由此可知，ArcA-P 調控因子於 DNA 序列上的調控序列是由兩個長度爲 5，間隔 5 的較小的調控序列（GTTAA 與 TGTTA）所組合而成。而上述之結果也與【Favorov *et. al.* 2005】中所預測的結果相同（亦即所發現模組的位置均爲一樣）。針對爲何預測的樣式與實驗數據不相一致，在【Favorov *et. al.* 2005】當中提到，雖然所預測出的模組樣式與實際上生物實驗【McGuire *et al.* 1999】發現的樣式（圖4.2.1 b）不同，但爲了驗證找出的調控模組樣式的可信度，他們重新針對大腸桿菌家族的基因體做調控模組的搜尋。最後從這些基因體中辨識出許多在【McGuire *et al.* 1999】所沒介紹的調控模組位置，而這些證據也

在最近幾年被驗證實爲 ArcA-P 蛋白所調控的區域（更爲詳細的內容請參照【Favorov *et. al.* 2005】）。因此，我們相信 ArcA-P 調控蛋白的樣式應該有著兩個長度爲 5 的雙重調控序列且其間距爲 5 個核甘酸的特徵。

(**a**)



(**b**)



（**圖4.2.1**）受 ArcA-P 調控之序列的Logo【Crooks *et. al.* 2004】。

(a) 系統以雙重調控序列的參數而取得結果。

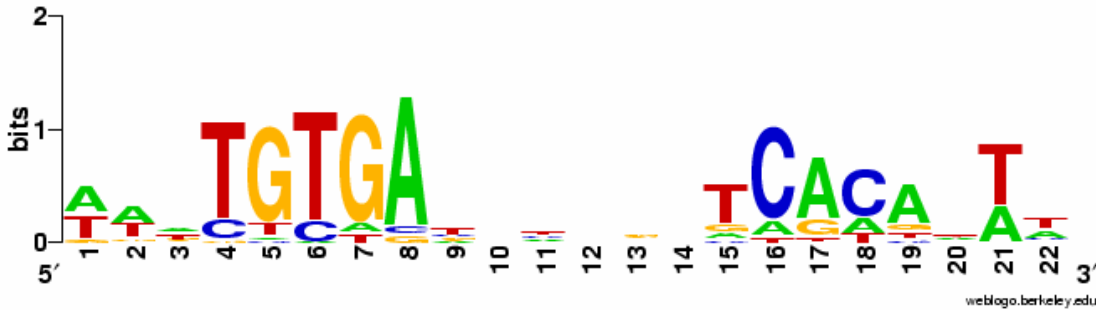
(b) 實驗已經證實的調控序列。資料取自 DPInteract 資料庫【Robison *et. al.* 1998】。

4.2.2 Cyclic AMP Receptor Protein

接下來，我們介紹另一種調控蛋白：Cyclic AMP Receptor Protein（CRP，環腺苷受體蛋白）。存在於大腸桿菌中的 CRP 蛋白其主要功能在於調控醣類的代謝。當細菌體內的葡萄糖濃度低時，CRP 蛋白與 cAMP（環腺苷酸）會相互結合成爲複合體進而加速基因轉錄作用，產生能夠合成、吸收和分解乳糖的酵素；一旦細菌體中葡萄糖的濃度充裕時，cAMP 便會減少使得 CRP 蛋白不能夠活化控制乳糖酵素的基因。

(a)

(b)



（**圖4.2.2**）大腸桿菌調控模組 Crp 的Logo【Crooks *et. al.* 2004】。

(a) 系統以雙重調控序列的參數而取得結果。

(b) 實驗已經證實的模組的Logo。資料取自 DPInteract 資料庫【Robison *et. al.* 1998】。

我們從大腸桿菌的基因體中收集 33 個受到 CRP 蛋白所調控的基因序列。我們測試 SAMAL 的是否能夠找到其調控序列。我們已知 CRP 基因結合區的樣式爲回文型態（Palindromic Motif），爲兩個十分明顯的調控序列間以 6 個

核甘酸爲距離所形成，透過實驗得知 CRP 蛋白所調控的共有字串爲 WWnTGTGAnnnnnnTCACAnWW，調控模組的 Logo 見（圖4.2.2 b）。我們以長度爲 8 的兩個調控序列，而此兩個調控序列之間距爲可變動的設定來讓 SAMLA 探測模組，結果如（圖 4.2.2 a），共有字串爲 DWWNTGTGAnnnnnnnTCACAWWW。

與實驗中發現的位置（圖4.2.2 a）相比，我們所發現的模組（圖4.2.2 b）的第一個序列恰與實驗發現的第一個序列位置重疊了 7 個核甘酸，而且我們所發現的模組其間距也收斂爲 7 個核甘酸。

已知樣式	WWnTGTGAnnnnnnTCACAn WW
系統預測樣式	DWWnTGTGnnnnnnn TCACAWWW
	* * * * * * * * * * * * * *

在預測的結果當中，其中有兩基因條序列不旦預測的道的間距錯誤，且其預測的位置也錯誤。第一條爲位於 ansB 基因序列上，我們預測的序列爲 AAATTGTTtaacgTCAAATTT，以 AAATTGTT 與 TCAAATTT 爲兩核心，其中小寫字體部份代表相對距離（5 個核甘酸）。實驗驗證的模組則爲 TTTTGTTAcctgccTCTAACTT，以 TTTTGTTA 與 TCTAACTT 爲兩核心，其間以 6 個核甘酸爲距離。

已知樣式	WWnTGTGAnnnnnnTCACAnWW
實驗驗證位置	T TTTGTTAnnnnnnTCT AACTT
系統預測位置	AAATTGTTn n n n n TCAAA TTT

第二條預測錯誤則發生於位在 aldB 基因序列，我們系統預測得到的序列爲 AAATTGTTagccgcttttTCAACTAT，以 AAATTGTT 與 TCAACTAT 爲兩核心，其中間隔了 10 個核甘酸。然而實驗驗證的調控模組則爲 ATTCGTGAtagctgTCGTAAAG，以 ATTCGTGA 與 TCGTAAAG 爲兩核心，其中間隔了 6 個核甘酸。

已知樣式	WWnT GTGAn n n n nnTCACAnWW
實驗驗證位置	ATTCGTGAn n n n nnTCGT AAAG
系統預測位置	AAATTGTTnnnnnnnnnnTCAACT AT

接下來，我們收集了 Bioprospector【Liu *et. al.* 2001】、SeSiMCMC【Favorov *et. al.* 2005】等能夠預測雙核心模組或是雙調控序列模組的系統。我們針對 Bioprospector 做了兩次不同設定的實驗來評估其效能。在參數設定上 Bioprospector$^{(A)}$ 與 Bioprospector$^{(B)}$ 的兩核心均設定長度爲 8；不同的部份在於 Bioprospector$^{(A)}$ 以 0 至 10 的間距爲設定，而Bioprospector$^{(B)}$ 則是設定 5 至 8 的間距；亦即，我們測試不同間距設定對 Bioprospector 預測能力的影響。SeSiMCMC 則均以預設的參數來執行。爲了統一起見，我們以每個程式所回報的最高分者爲依歸，針對這些預測結果，只要兩核心序列有與實驗驗證的位置相護重疊超過百分之六十，我們便認爲正確的預測。然而對於 SAMLA 的預測結果中，由於實際模組的間距爲固定的距離，因此每條基因上所預測的模組的間距也必須要一致，我們才認定爲正確的預測。此外，我們使用預測單一調控序列的工具 AlignACE【Hughes *et. al.* 2000】來測試是否 CRP 模組的任一核心可以透過預測單一調控序列的 AlignACE 而探尋。若可以找尋出其中的核心部份，則表示使用能夠預測多核心（多調控序列）模組的系統，SAMAL，是不需要的，亦即，我們只要依靠預測單一調控序列的工具便可以發現多核心模組。AlignACE 則各利用兩種不同的參數設定來執行：一爲設定能夠搜尋較長的調控序列，另一

則爲較短，並且各自在最高分數的輸出結果中只要預測的調控序列有與實驗驗證的位置相護重疊超過百分之六十，我們才認爲正確的預測。在希望能夠搜尋較長的調控序列方面，我們設定「預期多少位置需要高度一致」爲 10（測試AlignACE 是否可以準確的將 CRP 模組中兩個長度各爲五的核心找出來），另一爲 5（測試AlignACE 是否可以準確的將 CRP 模組中長度爲五的核心單獨地找出來）。

Tools	Number of Correctly Predicted	Total Number of Predicted Sites	Precision	Sensitivity	F-Score
SAMLA	27	35	77.14%	55.10%	0.6429
Bioprospector$^{(A)}$	28	63	44.44%	57.14%	0.5000
Bioprospector$^{(B)}$	35	50	70.00%	71.43%	0.7071
SeSiMCMC	28	44	63.64%	57.14%	0.6022
AlignACE(S)	0	55	0.00%	0.00%	0.0000
AlignACE(L)	19	48	39.58%	38.78%	0.3918

（**表 4.2.2**）與各種不同的預測工具的比較。已知在 33 條基因序列總共包含了 49 個實驗證實的模組位置。其中 Bioprospector$^{(A)}$ 與 Bioprospector$^{(B)}$ 之間的差距在於所使用的核心間距參數不相同，Bioprospector$^{(A)}$ 以 0 至 10 的間距爲設定，而 Bioprospector$^{(B)}$ 則是設定 5 至 8 的間距。而精確率（Precision），涵蓋率（Sensitivity）以及 F-Score 的計算方式詳見附錄 B。AlignACE(S) 代表設定較短的調控序列長度（「預期多少位置需要高度一致」爲 5）所得之結果，而 AlignACE(L) 代表設定較長的調控序列長度（「預期多少位置需要高度一致」爲 10）所得之結果。

我們透過生物實驗得知在 33 條基因序列中一共包含了 49 個受到 CRP 蛋白所調控的位置。粗略地來檢視（表4.2.2）的實驗結果統計數據，我們可以發現 Bioprospector$^{(B)}$ 的效能遠遠的超過其他預測工具，當然也超過了 SAMLA。但此結果是可預期的，這是由於 Bioprospector$^{(B)}$ 間距參數的設定（5~8個核甘酸）十分接近真實調控模組中兩核心的相對距離（6個核甘酸）。更加深入地檢視 Bioprospector$^{(B)}$ 所發現的模組，我們發現在這 35 個正確的預測結果中，有 1 個模組其兩核心的間距爲 5 個核甘酸；10 個模組其兩核心的間距爲 8個核甘酸；而符合正確間距（6 個核甘酸）的預測則只有 24 個。若將核心之間的間距做爲評定正確答案的要件，則 Bioprospector$^{(B)}$ 不論是在精確率或是涵蓋率的評比中均不及五成。若不將 Bioprospector$^{(B)}$ 的評比納入考量，我們可以發現 SAMLA

或是 Bioprospector(A) 亦或是 SeSiMCMC 所能探測到已證實位置的模組數目幾乎相同，SAMLA 正確的預測了 27 個位置，Bioprospector(A) 正確的預測了 28 個位置，SeSiMCMC 正確的預測了 28 個位置。能夠比 Bioprospector 的效能更加突出說明了模組核心之間的間距真正會影響著預測的準確。只能收尋單一調控序列的 AlignACE 則不能夠有效的尋找出正確的位置。

4.2.3 TyrR調控蛋白

大腸桿菌中的 TyrR 調控蛋白（Tyrosine Transcriptional Repressor）調控了 aroF、aroG、aroL、aroP、tyrB、tyrP、TyrR、mtr 等基因的表現作用。TyrR 調控蛋白的於這些被調控基因上的結合區域有著共同的樣式爲 RTGTAAWnnnnnnTTTACAnM。是爲一個雙核心的模組，以 RTGTAAW 與 TTTACAnM 作爲兩核心，其間相隔了 6 個核甘酸的距離，見（表 4.2.3.1）。TyrR 調控蛋白在轉錄時期大部分扮演著抑制的腳色，主要抑制 aroP、aroL、tyrP、tyrB、TyrR、aroG、aroF 等基因，特別的是活化 mtr 基因使其發生作用。

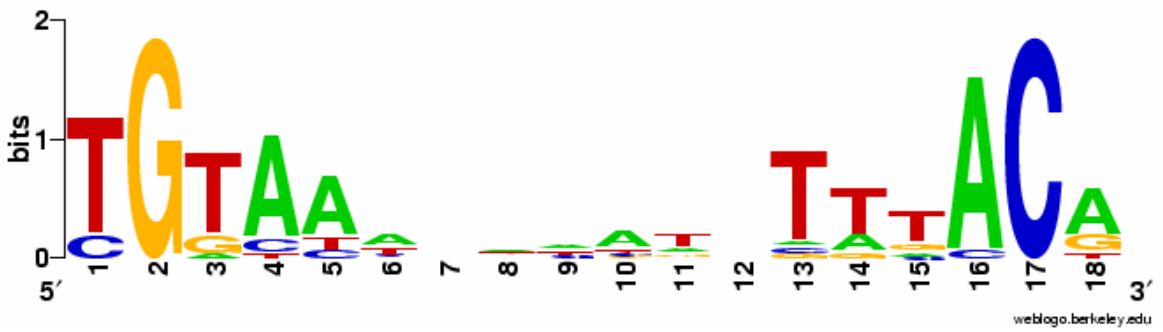
Regulon	**被調控基因**	**實驗得知樣式**
TyrR	tyrB tyrP mtr aroG aroL aroF aroP TyrR	RTGTAAWnnnnnnTTTACAnM

(表 4.2.3.1) TyrR 調控蛋白的資訊。資料取自 DPInteract 資料庫【Robison *et. al.* 1998】。

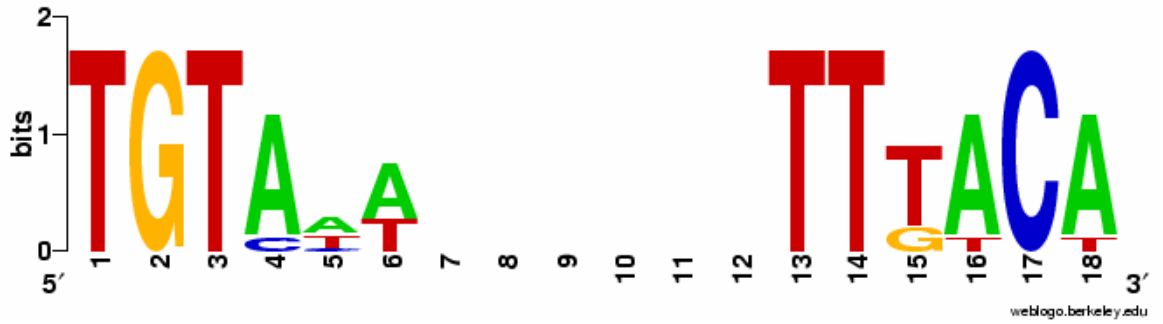
我們從大腸桿菌的基因體中取出這 8 條基因序列轉錄起始位置前的上游區域來測試我們的系統是否能發現生物實驗所證實的位置。這 8 條序列的長度均爲 300。SAMLA 的參數設定爲：兩核心長度均設定爲 6 ，其他則設以預設值。由於 SAMAL 爲隨機演算法，因此我們讓 SAMAL 以上述的設定執行五次，並且以這五次中分數最高的結果來做爲預測 TyrR 調控蛋白的作用區域。

我們比較 SAMAL 的預測結果與生物實驗所預測的結果，見（表 4.2.3.2），從實驗數據中發現 TyrR 蛋白的雙核心調控模組在 8 條基因序列中出現了 16 個位置，亦即在這 8 條基因序列上有 16 個位置將會與 TyrR 蛋白的雙核心模組相結合使得 TyrR 蛋白發揮調控的作用。透過 SAMLA 的收尋，我們卻只能在這 8 條基因序列中預測出 8 個會與 TyrR 蛋白發生作用的位置，這 8 個預測的位置中只有一個基因序列上爲預測錯誤。此錯誤發生在 tyrB 基因，此基因

(a)



(b)



（**表 4.2.3.2**）(a) 實驗已證實【Robison *et. al.* 1998】的樣式 Logo 與 (b)我們所設計的 SAMLA 預測樣式的 Logo 【Crooks *et. al.* 2004】。

上的 TyrR 調控模組爲 CGTAAAcctggaGAACCA，以 CGTAAA 和 GAACCA 爲核心。然而 SAMLA 所發現的模組爲 TGTAATatttgaTTGTCT，是以更爲一致的 TGTAAT 與 TTGTCT 爲核心所形成的模組。

已知樣式	TGTAAWnnnnnnTTTACA
實驗驗證位置	CGTAAAnnnnnnGAACCA
系統預測位置	TGTAATnnnnnnTTGTCT

接下來我們也比較其他工具與 SAMLA 的預測能力，見（表 4.2.3.3）。與上一小節雷同，我們收集了 Bioprospector【Liu *et. al.* 2001】、SeSiMCMC【Favorov *et. al.* 2005】、以及 AlignACE【Hughes *et. al.* 2000】等工具來評估。我們針對 Bioprospector 將兩核心均設定長度爲 6 並且以 0 至 10 的間距爲設定。SeSiMCMC 則均以預設的參數來執行。對於各個程式的最高分預測結果，只

Tools	Number of Correctly Predicted	Total Number of Predicted Sites	Precision	Sensitivity	F-Score
SAMLA	7	8	87.50%	43.75%	0.5833
Bioprospector	6	26	23.08%	37.50%	0.2857
SeSiMCMC	9	10	90.00%	56.25%	0.6923
AlignACE(S)	0	34	0.00%	0.00%	0.0000
AlignACE(L)	5	9	55.56%	31.25%	0.4000

（**表 4.2.3.3**）與各種不同的預測工具的比較。已知在 8 條基因序列總共包含了 16 個實驗證實的模組位置。AlignACE(S) 代表設定較短的調控序列長度（「預期多少位置需要高度一致」爲 6）所得之結果，而 AlignACE(L) 代表設定較長的調控序列長度（「預期多少位置需要高度一致」爲 12）所得之結果。

要兩核心序列有與實驗驗證的位置相護重疊超過百分之六十，我們便認爲正確的預測。AlignACE 設定兩種不同的參數，我們設定「預期多少位置需要高度一致」爲 12（測試AlignACE 是否可以準確的將模組中兩個長度各爲 6 的核心找出來），另一爲 6（測試AlignACE 是否可以準確的將模組中長度爲 6 的核心單獨地找出來），對於較短的設定，則只要與實驗已知的任一個核心重疊 60 %認定爲正確的預測。此外對於 SAMLA 的預測結果中，由於實際模組的間距爲固定的距離，因此每條基因上所預測的模組的間距也必須要一致，我們才認定爲正確的預測。

我們透過生物實驗得知在 8 條基因序列中一共包含了 16 個受到 tyrR 蛋白所調控的位置。檢視（表4.2.3.3）的實驗結果統計數據，我們可以發現 SeSiMCMC 的效能遠遠的超過其他預測工具，當然也超過了 SAMLA。這項結果讓我們十分納悶，因此我們詳細的比較了 SeSiMCMC 與 SAMLA 的比較結果，發現 SAMLA 錯誤預測的 tyrB 基因模組位置是由於真正實驗驗證的位置比 SAMLA 發現的位置還要不一致，也就是說，在 tyrB 基因序列中存在一個模組比生物實驗發現的模組還要來得更加與其他基因序列的模組一致。SeSiMCMC 能夠發現此正確的位置則是在預設參數中有一個「於程式收斂後，選入比目前結果更加一致的位置」（雖然程式能夠發現"每條"序列中最一致的位

置，仍然有可能於其他序列中存在比此序列更爲一致的位置，而此位置不一定是程式所選入的最高分者）。至少在此部分的評比 SeSiMCMC 的效能是爲最高的。而只能收尋單一調控序列的 AlignACE 則不能夠有效的在此資料中尋找出正確的模組位置。

4.2.4 cpxR調控蛋白

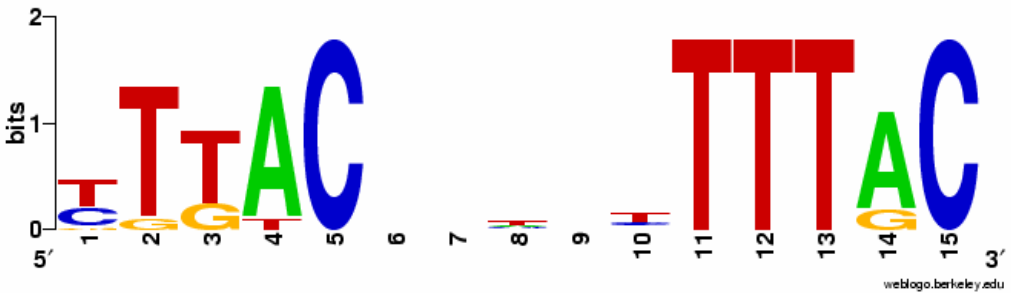
cpxR 調控蛋白存在於大腸桿菌中，主要調控 ppiA、ppiD、tsr、csgB、csgD、motA、cpxR、cpxP、alx、degP、yihE 等 11 種基因的表現。cpxR 調控蛋白的於這些被調控基因上的結合區域有著共同的樣式爲 GYAAAnnnnnGTAAR。是爲一個雙核心的模組，以 GYAAA 與 GTAAR 作爲兩核心，其間相隔了 5 個核甘酸的距離，見（表 4.2.4.1）。

Regulon	被調控基因	實驗得知樣式
cpxR	ppiA ppiD tsr csgB csgD motA cpxR cpxP alx degP yihE	GYAAAnnnnnGTAAR

(**表 4.2.4.1**) cpxR 調控蛋白的資訊。資料取自 DPInteract 資料庫【Robison *et. al.* 1998】

我們從大腸桿菌的基因體中取出這 11 條基因序列轉錄起始位置前的上游區域來測試我們的系統是否能發現生物實驗所證實的位置。這 11 條序列的長度

(a)



(b)

（**表 4.2.4.2**）(a) 實驗已證實【Robison *et. al.* 1998】的樣式 Logo 與 (b)我們所設計的 SAMLA 預測樣式的 Logo【Crooks *et. al.* 2004】。

均爲 300。SAMLA 的參數設定爲：兩核心長度均設定爲 5 ，其他則設以預設值。我們比較 SAMAL 的預測結果與生物實驗所預測的結果，見（表 4.2.4.2），從實驗數據中發現 cpxR 蛋白的雙核心調控模組在 11 條基因序列中出現了 12 個位置。透過 SAMLA 的收尋，我們在這 11 條基因序列中預測出 11 個會與 cpxR 蛋白發生作用的位置，這 11 個預測的位置中有兩個基因序列上爲預測錯誤，錯誤發生在 ppiD以及 cpxP 基因。系統在 cpxP 上所發現的樣式爲 TTTACnnnnnTTTAC，而非實驗中所發現的 CTGACnnnnnTTTGC。我們也可以明顯的看出 SAMLA 發現的樣式與（表4.4.2）中所呈現的 Logo 較爲相似。而針對 ppiD 基因序列上所發掘的位置則是一個有趣的現象。我們所預測的樣式總長爲 25，樣式與位置爲 (81)GTAAAnnnnnnnnnnnnnnGTAAA(105)，我們預測爲序列上的第八十一個位置開始一直到第一百零五位置爲模組出現位置。而實驗所驗證的樣式與位置則爲 (81)GTAAAnnnnnGAAAG(95)。其相對位置如下表示。即系統在預測 ppiD 序列上的模組時捨棄了距離上的統一而以核心部份的一致性爲優先考量。這是可預期的，原因在於預測大部分的模組均有著相同的間距時，捨棄了 ppiD 基因序列上模組的間距統一，而置換更加一致的調控序列樣式（將 GAAAG 換成更爲一致的樣式 GTAAA ），其於間距上的放棄能夠換取更大的統一。

位置	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05
SAMLA	**G**	**T**	**A**	**A**	**A**	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	**G**	**T**	**A**	**A**	**A**
Known	**G**	**T**	**A**	**A**	**A**	N	N	N	N	N	**G**	**A**	**A**	**A**	**G**	N	N	N	N	N	N	N	N	N	N

(**表**) ppiD 基因序列上實驗已知的模組位置與系統所預測的位置分佈示意表。

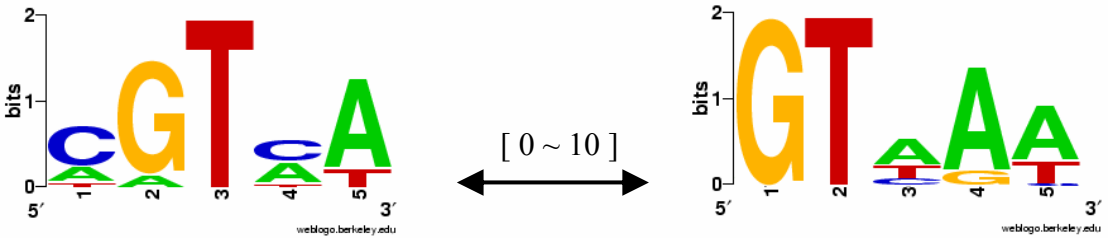
[a] 位置表示 ppiD 基因序列的實際位置。

[b] 灰體大寫字"N"，代表著非模組部分。列"SAMLA"則示意著系統發現的模組，以 GTAAA 與 GTAAA 爲核心，相距20個核甘酸。列"known"則示意著實驗證實的模組，以 GTAAA 與 GAAAG 爲核心，相距5個核甘酸。

Tools	Number of Correctly Predicted	Total Number of Predicted Sites	Precision	Sensitivity	F-Score
SAMLA	9	11	81.82%	75.00%	0.7826
Bioprospector	4	39	10.26%	33.33%	0.1569
SeSiMCMC	9	11	81.82%	75.00%	0.7826
AlignACE(S)	0	27	0.00%	0.00%	0.0000
AlignACE(L)	0	24	0.00%	0.00%	0.0000

（**表 4.2.4.3**）與各種不同的預測工具的比較。已知在 11 條基因序列總共包含了 12 個實驗證實的模組位置。AlignACE(S) 代表設定較短的調控序列長度（「預期多少位置需要高度一致」爲 5）所得之結果，而 AlignACE(L) 代表設定較長的調控序列長度（「預期多少位置需要高度一致」爲 10）所得之結果。

我們針對 Bioprospector 將兩核心均設定長度爲 5 並且以 0 至 10 的間距爲設定。SeSiMCMC 則均以預設的參數來執行。對於各個程式的最高分預測結果，只要兩核心序列有與實驗驗證的位置相護重疊超過百分之六十，我們便認爲正確的預測。AlignACE 設定兩種不同的參數，我們設定「預期多少位置需要高度一致」爲 10（測試AlignACE 是否可以準確的將模組中兩個長度各爲 5 的核心找出來），另一爲 5（測試AlignACE 是否可以準確的將模組中長度爲 5 的核心單獨地找出來），對於較短的設定，則只要與實驗已知的任一個核心重疊 60%認定爲正確的預測。然而對於 SAMLA 的預測結果中，由於實際模組的間距爲固定的距離，因此每條基因上所預測的模組的間距也必須要一致，我們才認定爲正確的預測。



(**圖**) Bioprospector 預測結果，以 MGTMW 與 FTWRW 爲核心，而其間距則於 0~10 之間變動。以另一個角度來看便是以 WYWAG 與 WKACK 爲核心。

檢視（表4.2.4.3）的實驗結果統計數據，SAMAL 與 SeSiMCMC 的效能相同，遠遠的超過其他預測工具。這項結果讓我們十分納悶的是 Bioprospector 的

結果是如此的糟糕。從 Bioprospector 的結果(上圖)中我們發現與實驗驗證的 LOGO 結果（表4.2.4.2）(a) 十分不相似，而兩核心之間的間距也在 0 與 10 之間變化，並無趨向一致的現象。而只能收尋單一調控序列的 AlignACE 依舊不能夠有效的在此資料中尋找出正確的模組位置。

4.2.5 narL調控蛋白

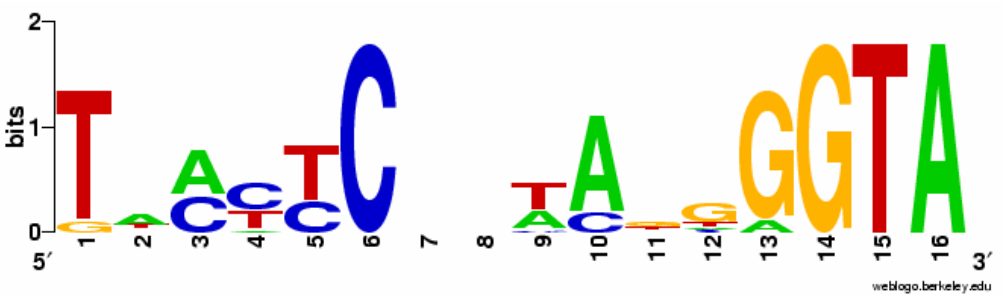
narL 調控蛋白主要調控 fdnG、narK、napF、frdA、dmsA、nrfA、narG、nirB 等 8 種大腸桿菌基因的表現。narL調控蛋白的於這些被調控基因上的結合區域有著共同的樣式爲 TWMYYCnnWAKGGGTA。看似爲一個雙核心的模組，以 TWMYYC 與 WAKGGGTA 作爲兩核心，其間相隔了 2 個核甘酸的距離，見（表 4.2.5.1）。

Regulon	被調控基因	實驗得知樣式
narL	fdnG narK napF frdA dmsA nrfA narG nirB	TWMYYCnnWAKGGGTA

(**表 4.2.5.1**) narL 調控蛋白的資訊。資料取自 DPInteract 資料庫【Robison *et. al.* 1998】

我們用 SAMLA 測試這 8 條基因序列，設定兩核心的長度均爲 7，其他參數則用預設。結果 LOGO 見(表 4.2.5.2)。基本上 SAMLA 的預測 LOGO (表

(a)



(b)

（**表 4.2.5.2**）(a) 實驗已證實【Robison *et. al.* 1998】的樣式 Logo 與 (b)我們所設計的 SAMLA 預測樣式的 Logo【Crooks *et. al.* 2004】。

4.2.5.2)(b) 已經描繪出真實模組的大致組成。不同的是我們發現的是以三個核甘酸爲核心間距，而實際上則是兩個核甘酸的間距。雖然有些許的不同，但是我們所預測 8 個基因上所得到的 8 個位置卻是真實已經過驗證的模組位置。推測 LOGO 會與實驗模組所表現的 LOGO 之間的大同小異是由於實際上的 10 個位置我們只發現了 8 個。

Tools	Number of Correctly Predicted	Total Number of Predicted Sites	Precision	Sensitivity	F-Score
SAMLA	8	8	100.00%	80.00%	0.8889
Bioprospector	3	25	12.00%	30.00%	0.1714
SeSiMCMC	7	8	87.50%	70.00%	0.7778
AlignACE(S)	8	13	61.54%	80.00%	0.6957
AlignACE(L)	6	13	46.15%	60.00%	0.5217

（**表 4.2.5.3**）與各種不同的預測工具的比較。已知在 8 條基因序列總共包含了 10 個實驗證實的模組位置。AlignACE(S) 代表設定較短的調控序列長度（「預期多少位置需要高度一致」爲 7）所得之結果，而 AlignACE(L) 代表設定較長的調控序列長度（「預期多少位置需要高度一致」爲 14）所得之結果。

Bioprospector 將兩核心均設定長度爲 6 並且以 0 至 10 的間距爲設定。SeSiMCMC 則均以預設的參數來執行。對於以上程式的最高分預測結果，只要兩核心序列有與實驗驗證的位置相護重疊超過百分之六十，我們便認爲正確的預測。AlignACE 設定兩種不同的參數，我們設定「預期多少位置需要高度一致」爲 14（測試 AlignACE 是否可以準確的將模組中兩個長度各爲 7 的核心找出來），另一爲 7（測試 AlignACE 是否可以準確的將模組中長度爲 7 的核心單獨地找出來），對於較短的設定，則只要與實驗已知的任一個核心重疊 60 %認定爲正確的預測。與其他預測工具所比較（表4.2.5.3），有考慮間距對模組影響的 SAMLA 效能爲最高。讓人較爲訝異的是 AlignACE 的表現也絲毫不遜色，這是因爲 narL 模組的兩核心擁有較多的 C、G 核甘酸。這樣的調控模組(調控序列)反而能夠在以 A、T 居多的背景中特別突顯，因此 AlignACE 也能夠預測出部分的核心位置。

4.2.6 rpoH調控蛋白

在這部分，我們要介紹一個與先前十分不同的調控蛋白，rpoH 調控蛋白。目前已知 rpoH 蛋白調控了 gapA、dnaK、grpE、htpG、hslV、rrmJ、clpB、ibpA、clpP、lon、groS、rpoD 等 12 個基因。(表4.2.6.1)。其模組序列以 MTTGWMW 和 CCCCATWW 爲核心部份，此兩個核心之間的相對距離則存在於兩種變動的數值，13 或 14 個核甘酸位置。我們選擇此與先前介紹的模組（固定間距），來突顯 SAMLA 的強大。

Regulon	**被調控基因**	**實驗得知樣式**
rpoH	gapA dnaK grpE htpG hslV rrmJ clpB ibpA clpP lon groS rpoD	MTTGWMWn$^{\{13,14\}}$CCCCATWW

(**表 4.2.6.1**) 調控蛋白的資訊。資料取自 DPInteract 資料庫【Robison *et. al.* 1998】

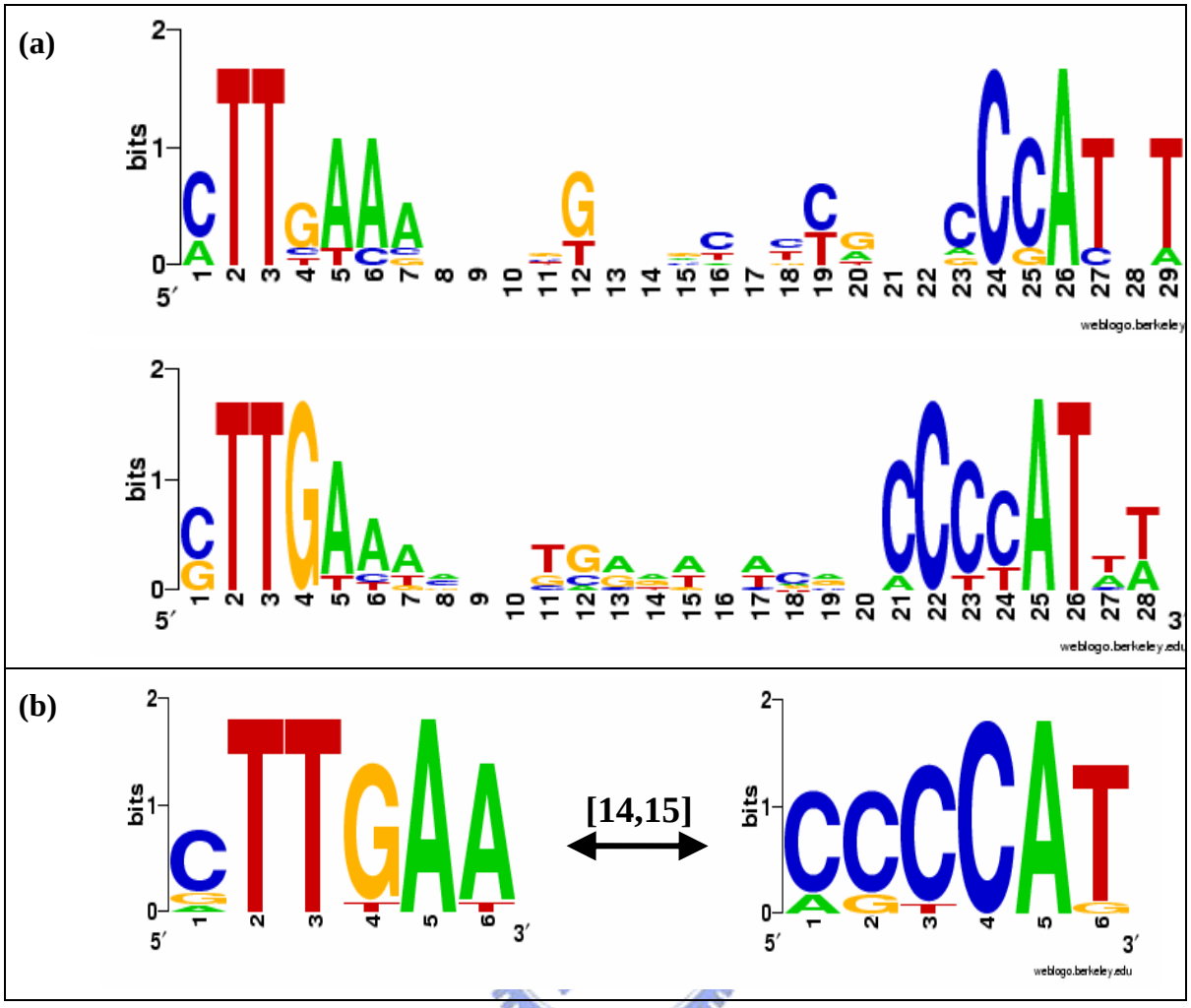
由於 rpoH 調控蛋白所形成的調控模組區域則爲一個十分有趣的現象，調控模組中兩個調控序列之間的相對距離並不固定，但分布仍然有限制，即兩個調控序列之間的相對距離只能間隔 13 或 14 個核甘酸的距離。在兩個長度爲 6 的核心，可變間距的設定之下 SAMLA 探尋出 rpoH 模組結果見（表4.2.6.2）。SAMLA 預測的樣式爲 MTTGAAn$^{\{14,15\}}$CCCCAT。與已知的樣式之間的排比如下：

實驗驗證樣式　　MTTGWMW - n$^{\{13,14\}}$ - CCCCATWW

系統預測樣式　　MTTGAA - n$^{\{14,15\}}$ - CCCCAT

* ** * * *　　*** * **

已知於這 12 條基因上有 14 個以驗證的模組位置，SAMLA 發現了其中的11個位置。



（**表 4.2.6.2**）(a) 實驗已證實【Robison *et. al.* 1998】的樣式 Logo。模組核心有兩種不同的相對距離，分別爲上 LOGO 的 14 個核甘酸以及下 LOGO 的 13 個核甘酸。 (b)我們所設計的 SAMLA 預測樣式的 Logo【Crooks *et. al.* 2004】。

（表4.2.6.3）中列舉了不同的預測工具的比較。由於 rpoH 蛋白所形成的模組爲可變間距的模組，因此，我們認定所謂的正確預測核心部份不到需要與實驗所得部分相互重疊超過 60% 之外，SALMA 以及 Biopospector 工具預測每條核心間距也必須一致，我們才將其列入正確的預測。然而，在這部份 SeSiMCMC 由於只能搜尋固定間距之模組，因此我們以較爲寬鬆的標準來認定其正確的預測。只要 SeSiMCMC 所預測的結果能夠與實驗驗證的部份相互重疊百分之 70，我們便認爲正確的預測。AlignACE 設定兩種不同的參數，我們設定「預期多少位置需要高度一致」爲 12（測試 AlignACE 是否可以準確的將模組中兩個長度各

爲 6 的核心找出來），另一爲 6（測試 AlignACE 是否可以準確的將模組中長度爲 6 的核心單獨地找出來），對於較短的設定，則只要與實驗已知的任一個核心重疊 60 %認定爲正確的預測。即便是放寬了 SeSiMCMC 與 AlignACE 的評定標準，從（表4.2.6.3）依舊可見 SAMLA 的效能評比爲第一。AlignACE 則無法精準的預測實驗驗證的模組位置。

Tools	Number of Correctly Predicted	Total Number of Predicted Sites	Precision	Sensitivity	F-Score
SAMLA	11	12	91.67%	78.57%	0.8462
Bioprospector	11	24	45.83%	78.57%	0.5789
SeSiMCMC	12	15	80.00%	85.71%	0.8276
AlignACE(S)	0	28	0.00%	0.00%	0.0000
AlignACE(L)	0	28	0.00%	0.00%	0.0000

（**表 4.2.6.3**）與各種不同的預測工具的比較。已知在 12 條基因序列總共包含了 14 個實驗證實的模組位置。AlignACE(S) 代表設定較短的調控序列長度（「預期多少位置需要高度一致」爲 6）所得之結果，而 AlignACE(L) 代表設定較長的調控序列長度（「預期多少位置需要高度一致」爲 12）所得之結果。

4.2.7 總結

在以上小節中，我們評估了 ArcA-P，CRP，TyrR，cpxR，narL，以及 rpoH 等調控蛋白之模組。他們所調控的基因序列群以及從實驗已得知的樣式見（表 4.2.7.1）。

Regulon	**被調控基因**	**實驗得知樣式**
ArcA-P	aldA cydA glcC gltA icdA lldP lpdA pflA sodA	GTTAAYWMWWWKNNW
CRP	lacZ tsx nagE fur galE ompA dadA cdd glpT ptsH srlA ansB nupG tdcA crp ppiA ompR malT glpE aldB mtlA ilvB tnaL cyaA rhaB glpF cytR malE malK melR uxuA deoC	WWWTGTGAnnnnnnTCACANWW
TyrR	tyrB tyrP mtr aroG aroL aroF aroP TyrR	RTGTAAWnnnnnnTTTACAnM
cpxR	ppiA ppiD tsr csgB csgD motA cpxR cpxP alx degP yihE	GYAAAnnnnnGTAAR
narL	fdnG narK napF frdA dmsA nrfA narG nirB	TWMYYCnnWAKGGGTA
rpoH	gapA dnaK grpE htpG hslV rrmJ clpB ibpA clpP lon groS rpoD	MTTGWMWn$^{\{13,14\}}$CCCCATWW

(表 4.2.7.1) 調控蛋白的資訊總整理。

實際比較 SAMLA 的預測能力不難發現，SAMLA 的能力略比 SeSiMCMC 爲強大。反而是 Bioprospector 差強人意，深入的觀察 Bioprospector 預測的結果可以發現 Bioprospector 所預測各個基因序列上的兩核心之間之間距並非固定，是不一致的，這與實驗所驗證出的固定間距模組有所差距，這使得 Bioprospector 的預測十分不精確。透過這些評比可以觀察到 SAMLA 的預測能力並不亞於其他的工具（SeSiMCMC），而 Bioprospector 卻有可能輸出許多不同間距的答案，這使得使用者不能夠放心的相信 Bioprospector 對於多核心模組的偵測能力。因此，我們所設計的 SAMLA 不但可以預測固定間距的模組亦可

以探測可變間距之模組，這對模組預測的技巧是爲一大躍進。

從這些結果中不難發現，不論是 CRP、TyrR、cpxR、或是 rpoH 調控蛋白所形成模組，其兩核心部份均無法利用預測單一調控序列的工具（AlignACE）來準確的探測，進而將其結合還原爲雙核心模組。因此能夠同時探測多核心（調控序列）的 SAMLA 與 SeSiMCMC 反而能夠更爲精準的預測模組的位置。至於 narL 調控蛋白所形成之模組足爲特例，但是從比較結果中可以看出利用 SAMLA 同時搜尋 narL 模組的雙核心反而能夠提升更多的效能。

	Number of Real Binding Sites	Tools	Number of Correctly Predicted Sites	Total Number of Predicted Sites	Precision	Sensitivity	F-Score
CRP	49	SAMLA	27	35	77.14%	55.10%	0.6429
		Bioprospector	28	63	44.44%	57.14%	0.5000
		SeSiMCMC	28	44	63.64%	57.14%	0.6022
		AlignACE(S)	0	55	0.00%	0.00%	0.0000
		AlignACE(L)	19	48	39.58%	38.78%	0.3918
TyrR	16	SAMLA	7	8	87.50%	43.75%	0.5833
		Bioprospector	6	26	23.08%	37.50%	0.2857
		SeSiMCMC	9	10	90.00%	56.25%	0.6923
		AlignACE(S)	0	34	0.00%	0.00%	0.0000
		AlignACE(L)	5	9	55.56%	31.25%	0.4000
cpxR	12	SAMLA	9	11	81.82%	75.00%	0.7826
		Bioprospector	4	39	10.26%	33.33%	0.1569
		SeSiMCMC	9	11	81.82%	75.00%	0.7826
		AlignACE(S)	0	27	0.00%	0.00%	0.0000
		AlignACE(L)	0	24	0.00%	0.00%	0.0000
narL	10	SAMLA	8	8	100.00%	80.00%	0.8889
		Bioprospector	3	25	12.00%	30.00%	0.1714
		SeSiMCMC	7	8	87.50%	70.00%	0.7778
		AlignACE(S)	8	13	61.54%	80.00%	0.6957
		AlignACE(L)	6	13	46.15%	60.00%	0.5217
rpoH	14	SAMLA	11	12	91.67%	78.57%	0.8462
		Bioprospector	11	24	45.83%	78.57%	0.5789
		SeSiMCMC	12	15	80.00%	85.71%	0.8276
		AlignACE(S)	0	28	0.00%	0.00%	0.0000
		AlignACE(L)	0	28	0.00%	0.00%	0.0000

（**表 4.2.7.2**）與各種不同的預測工具的比較表格。AlignACE(S) 代表設定較短的調控序列長度所得之結果，而 AlignACE(L) 代表設定較長的調控序列長度所得之結果。（AlignACE 詳細的設定，請詳看章節內容中）。

第五章 結論與未來研究方向

5.1 結論與討論

過去幾年以來，針對預測調控序列問題所研發的工具多如牛毛，然而針對調控模組的研究相對比較起來顯得如鳳毛麟角。爲此，我們參考眾多以統計機率角度來分析調控序列問題的工具【Lawrence *et. al.* 1993】【Liu *et. al.* 1995】【Bailey and Elkan, 1995】【Hu *et. al.* 2000】【van Helden *et al.* 2000】【Liu *et. al.* 2001】【Jensen *et. al.* 2004】【Favorov *et. al.* 2004】，進而推導出可以對調控模組評分的函數 *Score*(*)。只要給定基因序列以及調控模組 *RM*，我們便可以對調控模組 *RM* 計算、評估其好壞。而真實的調控模組必定爲 *Score*(*) 數值最高者，因此將預測調控模組問題轉換成尋求函式 *Score*(*) 最佳值問題。

與其他大部分現行的工具比較之，SAMLA 最主要的貢獻爲，

SAMLA 提出一個用來描述模組中各調控序列之間間距擁有相似性特點的機率模型，這使得整個系統能夠在不失去偵測固定間距模組的能力之外，還能更有彈性地來針對可變間距的調控模組進行預測，而不用限定於以往的預測模式（只能探測固定間距之模組）。

其次則爲，

1. 不需要使用者預先定義背景。特別是某些生物的背景資料十分難以取得和製作時，SAMLA 依舊能夠利用目前已知的序列資料來進行運算。
2. SAMLA 能夠探尋多調控序列的模組。
3. SAMLA 能夠自動選擇模組於基因序列群中的出現次數（但需要設定每個基因序列上最多出現的次數）。
4. 利用貝式推論得到評分調控模組的方式，公式(7)，並且利用模擬退火法來取得使 *Score*(*RM*) 爲最大值的調控模組 *RM*。此外，利用公式(7)

可以針對不同的調控模組來評論其優劣。

每種分析方法皆具有其特長，不可言諱的亦有其相對的缺失之處。利用模擬退火法的概念搭配 Gibbs Sampler 的流程設計的系統，有著以下之缺點，

1. 需要使用者輸入調控序列之種類以及長度。
2. 需要人工預先定義模組中調控序列的個數。
3. 每次迭代時執行的 ADD-Operator 與 SWAP-Operator 所需要的時間複雜度隨著調控模組大小以及調控序列之種類而改變，執行 ADD-Operator 與 SWAP-Operator 將是整個系統中最花費時間的操作，時間複雜度約可略估爲 $O(\prod_{j=1}^{j=k}(l_i - w_j + 1))$。其中 l_i 爲基因 g_i 之長度，共有 k 種調控序列需要被搜尋。在調控模組大小未明確的定義之下，第 j 種調控序列出現於序列 g_i 的可能位置共有 $l_i - w_j + 1$ 種，整體而言，此操作必須嘗試所有可能的組合，因此時間複雜度爲 $O(\prod_{j=1}^{j=k}(l_i - w_j + 1)) \cong O(l_i^k)$。

5.2 未來研究方向

雖然我們已經得到了不錯的評分函式可以用來評比不同的調控模組，然而，在我們所發展的系統中卻沒有針對不同長度以及種類的調控序列來做搜尋。其中最爲主要的原因是，若加入不同長度以及種類的調控序列等變因，依照目前的 SAMLA 必定無法於合理的時間（多項式時間，Polynomial Time Complexity）中來完成搜尋。因此未來將會朝向發展能夠更全方面地探尋調控模組的程式以及演算法，並且能夠在合理的時間中來完成搜尋。

目前 SAMLA 描述模組中各調控序列之間間距的方式，是以一個十分簡單的機率模型來描繪調控序列之間的距離。隨著生物資訊上的新發現，我們可以融匯這些新訊息來調整間距模型，以期能夠更爲精確的描述調控序列之間的間距關係。

此外，改用更爲精確的背景模型亦是一種改進的方式。經由第三章的介紹之後可以發現我們所利用的背景模型是零階的馬可夫模型（Zero-order Markov Model）。在【Thijs G *et. al.* 2001, 2002】【Sinha, S. *et. al.* 2000】研究中提出了改進背景模型爲更高階的馬可夫模型可以增加系統預測的能力。使用高階馬可夫模型的背景模型亦可增進系統分辨不同調控模組之間的差距，改進系統的預測能力。

參考文獻

1. Agrawed,R. and Srikant,R. (1994) Fast algorithms for mining association rules. *Proceedings 199J International Conference VLDB*. 487-499.

2. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. 28-36.

3. Bailey,T.L. and Elkan,C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Mach. Learn.*, 21, 51.

4. Crooks,G.E., Hon,G., Chandonia,J.M., and Brenner,S.E. (2004) WebLogo: A sequence logo generator. *Genome Research*, 14, 1188-1190.

5. Day,W.H. and McMorris,F.R. (1993) The computation of consensus patterns in DNA sequence. *Math. Comput. Model.*, 17, 49-52.

6. GuhaThakurta,D. and Stormo,G.D. (2001) Identifying target sites for cooperatively binding factors. *Bioinformatics*, 17, 608-621.

7. Favorov,A.V., Gelfand,M.S., Gerasimova1,A.V., Ravcheev,D.A., Mironov, A.A., and Makeev,V.J. (2004) A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics*, 21, 2240-2245.

8. Gusfield. (1997) *Algorithms on strings, trees and sequences*. Cambridge University Press.

9. Hu,Y., Sandmeyer,S., McLaughlin,C., and Kibler,D. (2000) Combinatorial motif analysis and hypothesis generation on a genomic scale. *Bioinformatics*, 16, 222-232.

10. Hu,Y. (2003) Finding subtle motifs with variable gaps in unaligned DNA sequences. *Computer Methods and Programs in Biomedicine*, 70, 11-20.

11. Grayson,J., Bassel-Duby,R., and Williams,R.S. (1998) Collaborative Interactions Between MEF-2 and Sp1 in Muscle-Specific Gene Regulation. *Journal of Cellular Biochemistry*, 70, 366-375.

12. Jensen,S.T., Liu,X.S., Zhou,Q. and Liu,J.S. (2004). Computational discovery of gene regulatory binding motifs: a Bayesian perspective. *Statistical Science* 19:188-204.

13. Kirkpatrick,S., Gelatt,Jr., C.D., and Vecchi,M.P.. (1983) Optimization by Simulated Annealing. *Science*, 220, 671-680.

14. Lawrence,C.E., Altshul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262, 208-214.

15. Liu,J.S. (1994) The collapsed Gibbs Sampler in Bayesian computations with applications to a gene regulatory problem. *J. Amer. Statist. Assoc.* 89, 958-966.

16. Liu,J.S., Neuwald,A.F. and Lawrence,C.E. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Amer. Statist. Assoc.* 90, 1156-1170.

17. Liu X, Brutlag,D.L., and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*, 127-38.

18. Mitchell, Tom M.. (1997) Machine Learning. McGraw-Hill.

19. Robison,K., McGuire,A.M., and Church,G.M. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K12 genome. *Journal of Molecular Biology*, 284, 241-254.

20. Sinha,S. and Tompa,M. (2000) A statistical method for finding transcription factor binding sites. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, 344-354.

21. Thijs,G., Lescot,M., Marchal,K., Rombauts,S., De Moor, B., Rouze, P., and Moreau, Y. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, 17, 1113-1122.

22. Thijs G., Marchal K., Lescot M., Rombauts S., De Moor B., Rouze P., and

Moreau Y. (2002) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *Journal of Computational Biology*, Vol. 9, No. 2: 447-464.

23. van Helden, J., Andre, B, and Collado-Vides, J. (1998) Extracting Regulatory Sites from the Upstream Region of Yeast Genes by Computational Analysis of Oligonucleotide Frequencies. *Journal of Molecular Biology*, 281, 827-842.

24. van Helden, J., Rios, A. and Collado-Vides, J. (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res*., 28, 1808-1818.

25. Zhu, J. and Zhang, M.Q. (1999) SCPD: A promoter database of the yeast Saccharomyces cerevisiae. Bioinformatics, 15, 607-611.

A. IUPAC對照表

Code	Description
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
R	Purine (A or G)
Y	Pyrimidine (C, T, or U)
M	C or A
K	T or G
W	T or A
S	C or G
B	C, T or G
D	A, T or G
H	A, T or C
V	A, C or G
N	any base (A, C, G or T)

核甘酸的 IUPAC Code 對照表

B. Precision and Sensitivity

首先定義三個名詞：

A. *TP* 表示程式所預測的答案中，實際上正確的預測數目。

B. *FP* 表示程式所預測的答案中，實際上錯誤的預測數目。

C. *RA* 表示真實正確的答案數目。

，接下來精確率（Precision）以及涵蓋率（Sensitivity）便可依此來計算：

$$\text{Precision} = \frac{TP}{TP + FP} \text{。}$$

$$\text{Sensitivity} = \frac{TP}{RA} \text{。}$$

$$\text{F-Socre} = \frac{1}{\frac{1}{2}\left(\frac{1}{\text{Pr}ecision} + \frac{1}{Sensitivity}\right)}$$