第一章　序論

1.1　生物體內遺傳訊息的運作

生物學家已經驗證基因爲製造特定蛋白質的領航員，而遺傳訊息由 DNA 傳遞至蛋白質 **(protein)** 的過程中須藉由**傳訊 RNA (messenger RNA; mRNA)** 作爲仲介。因此生物體內蛋白質產生的過程主要含括**轉錄 (transcription)** 和**轉譯 (translation)** 兩個部份。在轉錄的過程中，DNA 上的遺傳訊息會使得與其相對應的 mRNA 產生；緊接著在轉譯過程中，隱含在此 mRNA 中的遺傳訊息亦會促使與其相對應的蛋白質產生。（**圖 1.1**）

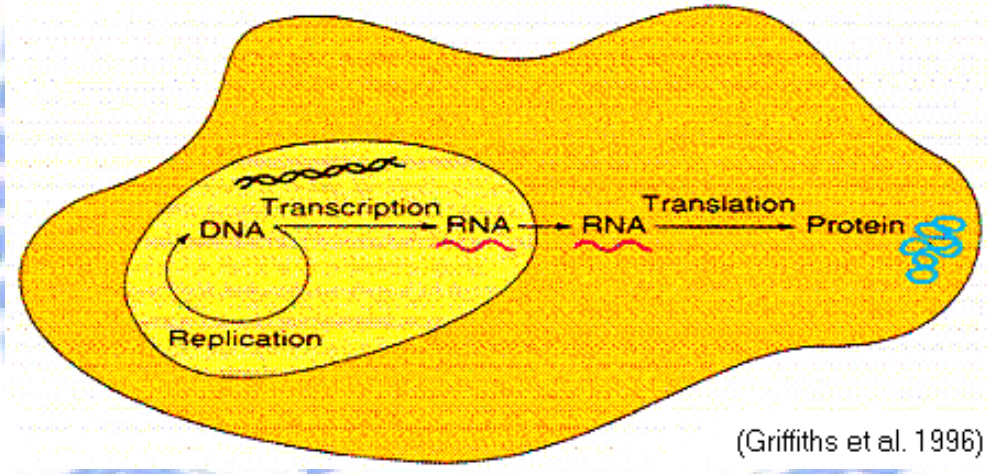


圖 1.1　轉錄和轉譯作用在遺傳訊息傳遞上所扮演的角色

在轉錄過程中，稱爲 **RNA 聚合酶 (RNA polymerase)** 的酵素將會附著至 DNA 上，沿著基因的**鑄模股 (template strand)** 將對應之 RNA 核甘酸串連，而此 RNA 核甘酸序列即爲 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的特定核甘酸序列以及起始**上游區間 (upstream region)** 的核甘酸。因此，啓動子的存在對 RNA 聚合酶的辨識而言是相當重要的。（**圖 1.2**）

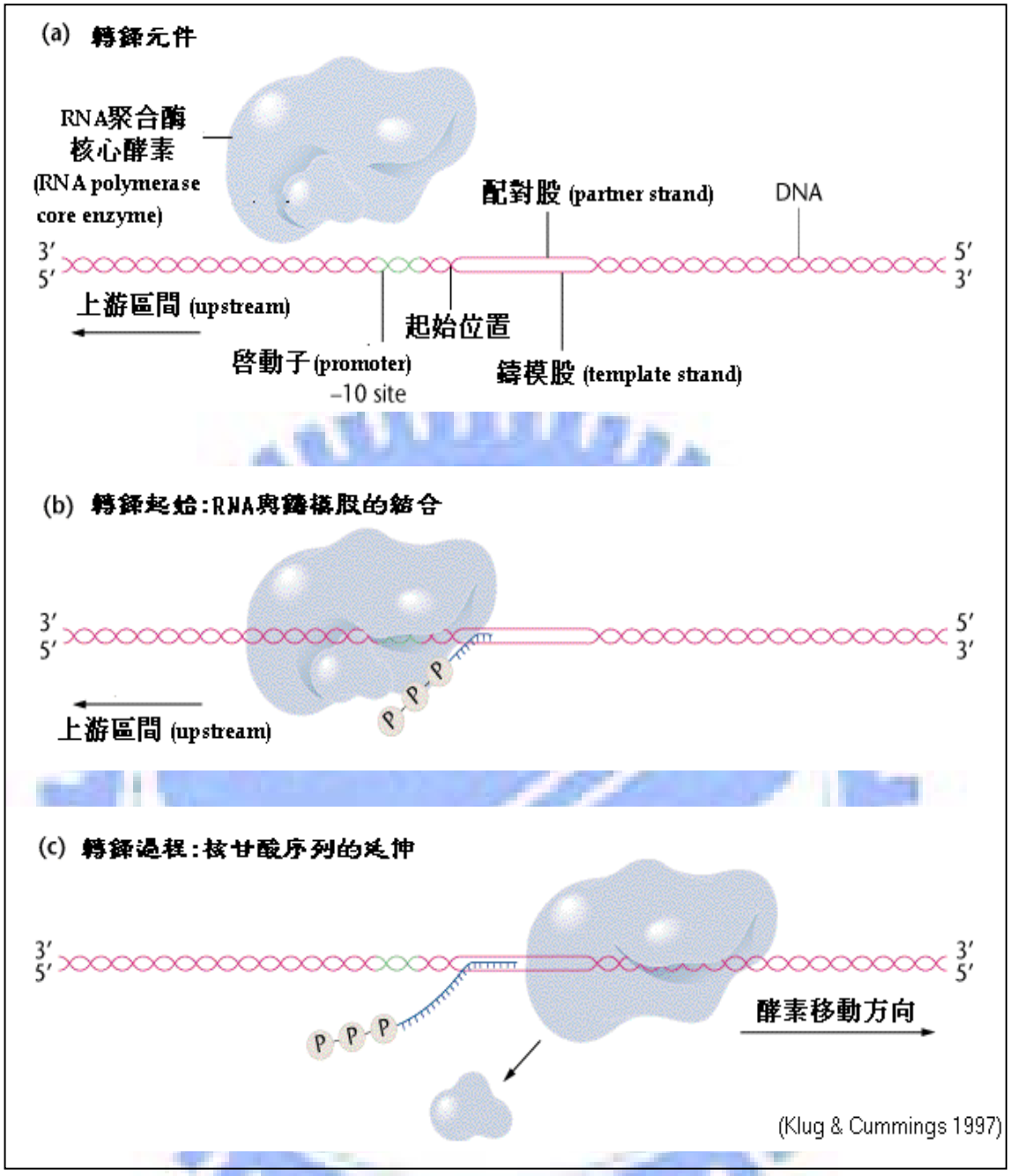


圖 1.2　轉錄作用

生物學家亦已驗證某些 RNA 聚合酶是無法自己辨認出 DNA 上啓動子的位置並加以附著的，而這些 RNA 聚合酶須依賴**轉錄因子 (transcription factor)／調控蛋白 (regulatory protein)** 才能沿著 DNA 探測啓動子的位置。不過，每個啓動子的核甘酸序列都不盡相同，其中可能含括一些重要的**調控區間 (regulatory site)** 致使得不同的轉錄因子與其結合。（**圖 1.3**）

值得一提的是，在同一生物體內的任一細胞皆擁有相同的**基因體 (genome)**。不同細胞間的差異之處僅在於其含括的**調控蛋白 (regulatory protein)** 的不同，藉由讓不同的基因轉錄來產生不同的表現。

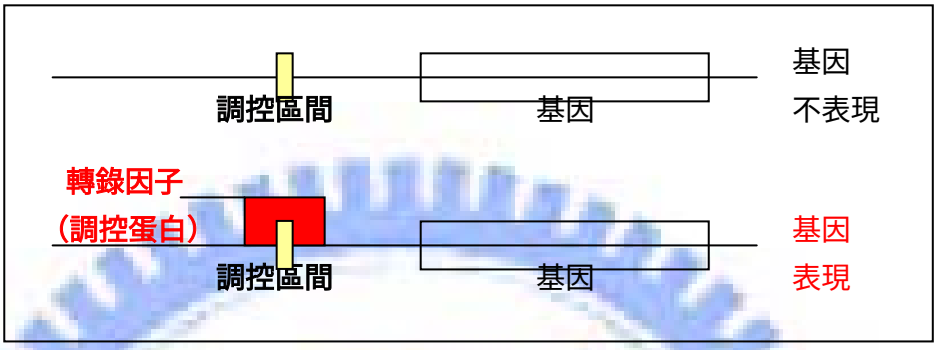


圖 1.3　調控蛋白的作用

1.2　研究動機

經由上節所介紹的遺傳訊息運作過程後和 **圖 1.3** 所示，我們可以得知藉由不同的**轉錄因子**與**重要基因片段 (motif)／調控區間**相互結合所產生的調控作用可讓基因有不同的表現。而重要基因片段主要可依據其構成子片段的個數被劃分成一僅由一**子片段 (component)** 所構成的**單元重要基因片段 (monad)** 以及由兩個以上彼此相近且具關聯度的子片段所組成的**多元重要基因片段 (polyad)**。

因此，倘若我們能探測出 DNA 上所有型態的**調控基因片段 (regulatory motif)**，我們就能更進一步去預測基因所有可能的表現。不過，由於基因定序技術發展的快速以及眾多基因計劃的進行，使得基因序列資訊量的增加以倍數方式在成長。面對如此急速增加的資料量，單純使用人工實驗由生物序列中分析出重要基因片段已不再具有可行性。因此在探測重要基因片段的議題上，勢必得使用電腦輔助的方法。

然而在許多真實調控基因片段皆爲多元重要基因片段的情況下，（例如：***S. cerevisiae*** 基因表現會由 **URS1** 和 **UASH** 兩處調控區間與其對應的調控因子結合所共同調控），現今大多使用電腦輔助探測重要基因片段的方法皆將其研究領域侷限在探測單元重要基因片段上。除此之外，其皆需使用者明定其欲搜尋重要基因片段的相關資訊，像是重要基因片段的長度等等。因而當生物學家一旦接獲到新的其中隱含某些重要基因片段的序列群時，這些需明定重要基因片段資訊的探測重要基因片段的方法即無法發揮其功效。

因此在此論文中，我們發展出由 **WEEDER** 演算法的容錯度概念 (Pavesi et al., 2001)、**Apriori** 中擷取強度相關概念 (Agrawed R. and Srikant R., 1994) 以及以 **MERMAID** 演算法 (Hu, 2003) 爲基礎所延伸結合的 **APPA 演算法 (Apriori Pruning Algorithm)** 以期符合使用者需求，進而探測所有可能的單元及多元重要基因片段。

1.3 研究目標

首先，我們將重要基因片段的探測問題定義如下：

假設有一重要基因片段 $M = (m_1,[l_1,u_1],m_2,.....,[l_{p-1},u_{p-1}],m_p)$，其由 $p \geq 1$ 個未知長度子片段 m_i 和 $(p-1)$ 個未知間距 $[l_i, u_i]$ 交錯組合而成，且 M 隱藏在有 t 條基因序列的序列群 $S=(s_1,s_2,.....,s_t)$ 中，則每條序列 s_i 中應該都有一段與此重要基因片段 M 相似的片段，$M'_i = (m'_1,g'_1,m'_2,.....,g'_{p-1},m'_p)$。（**圖 1.4**）

而 M 與 M'_i 兩片段相似的條件即爲其中的相互對應子片段 m_i 和 m'_i 間最多只能相差 e 個核甘酸，以及座落在 m'_i 和 $m'_{(i+1)}$ 間的區間長度 g'_i 須座落在 $[l_i, u_i]$ 此數值區間中，意即 $l_i \leq g'_i \leq u_i$。

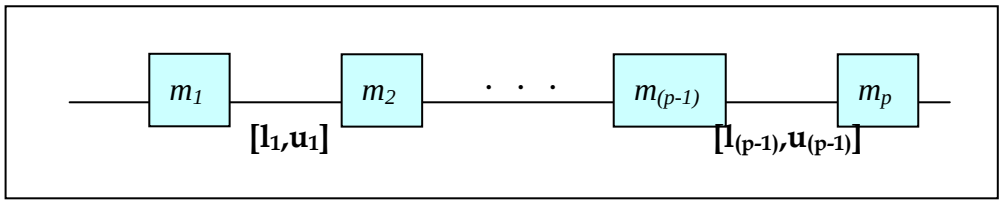


圖 1.4　重要基因片段的型式 $\boldsymbol{M}=(m_1,[l_1,u_1],m_2,.....,[l_{p-1},u_{p-1}],m_p)$

因此本論文的研究目標即在於－

在僅需使用者設定**容錯度 (error ratio)** ε 和對照基因序列群 ***B*** 的情況下，能從一群具有相同調控機制的基因序列中探測出所有可能的重要基因片段所象徵的調控區間。

例如當使用者設定容錯度 $\varepsilon=0.125$ 時，則意謂著－與重要基因片段 ***M*** 相似片段中的各個子片段只允許與其對應子片段 $\boldsymbol{m_i}$ 至多有 $0.125|m_i|$ 個位置的核甘酸相異。

在 **圖 1.5** 中有四條長度皆爲 40 且隱藏有重要基因片段的基因序列，而 **APPA** 的目的即在於在使用者設定容錯度 $\varepsilon=0.125$、含括此四條基因序列的序列群 ***S*** 以及作爲對照序列群 ***B*** 的情形下，能確切探測到重要基因片段 ***M*=ATTTCATG**，進而得知此重要基因片段爲長度 8 的單元重要基因片段。

TTAGAC**ATTTCATG**TTGTGTTGTCTCAATGCTTTGCTCAA

GTGCTCCTGTCG**AGTTCATG**CCAACCCTTTCCTAGGATAT

CACGGCATATAATCTTCCGGTAATG**AGTTAATG**GTGCCAA

TACGAATAAGCTTCTT**AGTTCATC**TAAAACTTATCCTGCT

圖 1.5　四條皆隱含有單元重要基因片段 AGTTCATG 的序列

($\varepsilon=0.125$ **代表在長度為 8 的子片段中至多只能有 1 個相異處)**

關於多元重要基因片段的實例，則在 **圖 1.6** 中。由 **圖 1.6 (a)** 中可得知 13 條表示 E. coli 啓動子的序列群，我們希望在使用者僅輸入容錯度 $\varepsilon = 0.125$、含括此 13 條基因序列的序列群 ***S*** 以及作爲對照序列群 ***B*** 的情形下， **APPA** 最終能探測到重要基因序列 ***M*** **=** **TTGACAT**－**[15, 17]**－**TATAAT**，進而得知此重要基因序列是由兩個相互間隔 15 至 17 個核甘酸的子片段所組成。（**圖 1.6(b)**）

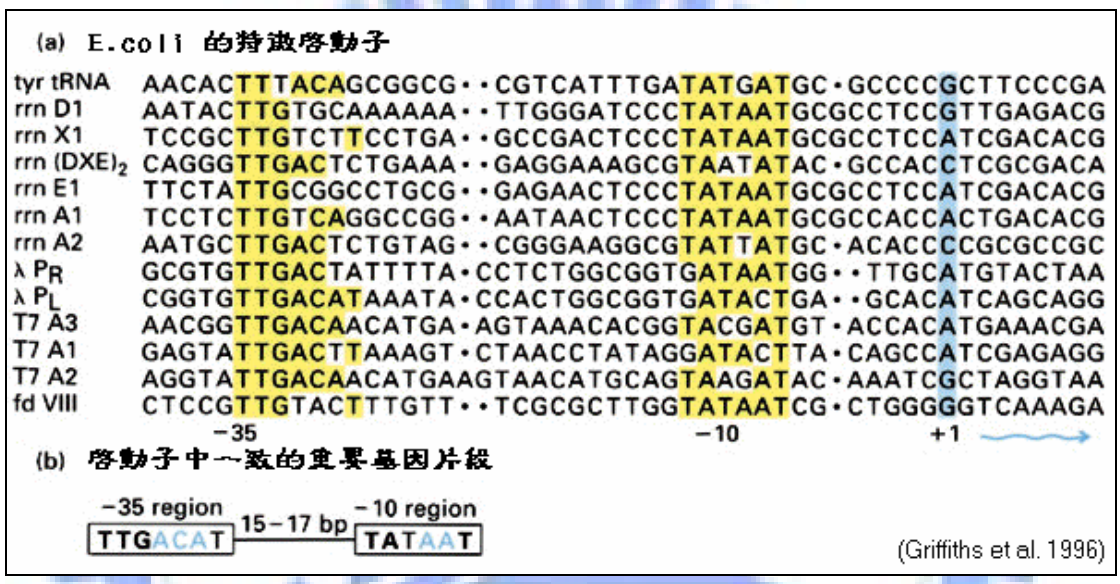


圖 1.6　E. coli 的特徵啟動子中的重要基因片段

1.4 論文架構

在此單元中，我們將述此論文的整體架構。

此論文主要分爲五章。在第一章中，我們將首先介紹生物體內遺傳訊息的運作，進而闡述探測重要基因片段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 **APPA (Apriori Pruning Algorithm)** 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

在本章中，我們會對整個探測重要基因片段的研究代表方法作精要的介紹。

在第一單元單元重要基因片段探測法概述中，我們將會首先簡單扼要地一一介紹三種探測單元重要基因片段的方法：分別是以統計方法爲基底的機器學習探測法 **MEME (Bailey and Elkan, 1995)**、以圖論爲延伸的 **WINNOWER (Pevzner and Sze, 2000)**、以及無需讓使用者設定重要基因片段資訊的 **WEEDER (Pavesi et al., 2001)**。

緊接著在第二單元多元重要基因片段探測法概述中，介紹兩種多元重要基因片段的探測法：分別是固定雙元間距的探測法 **Dyad Analysis (van Helden et al., 2000)**、以及改良 **WINNOWER** 至探測雙元可變間距重要基因片段的 **MITRA** 演算法。

最後將在第三單元中介紹 **APPA** 所採用的 **Apriori** 演算法的完整概念。

2.1 單元重要基因片段探測法概述

2.1.1　MEME (Bailey and Elkan, 1995)

MEME 是近來最常被於探測單元重要基因片段的方法之一。其運用了期望值與最佳化 **(Expectation－Maximization; EM)** 的概念，進而計算出具代表性的位置比重矩陣。在位置比重矩陣上的數值能表現出在不同位置上不同 DNA 核甘酸所出現的機率，因而可以探測到最具代表性的重要基因片段。（**圖 2.1**）

letter	*position in motif*					
	1	**2**	**3**	**4**	**5**	**6**
A	0.1	0.8	0.1	0.5	0.6	0.1
C	0.1	0.1	0.1	0.3	0.2	0.1
G	0.2	0.0	0.1	0.1	0.1	0.1
T	0.6	0.1	0.7	0.1	0.1	0.7

圖 2.1　MEME 計算出表現重要基因片段 TATAAT 的位置比重矩陣 (Bailey and Elkan, 1995)

MEME 主要藉由下述兩大步驟的反覆運作進而得到較具代表性的重要基因片段：

- **Expectation**

  一開始時針對每條序列上每一位置，計算出每一基因片段可能出現在序列群中的機率。而後每次運算都須根據前次所計算出的機率加以修改現值，最後得到一個位置比重矩陣。

- **Maximization**

  根據在 Expectation 步驟中所得的位置比重矩陣，可得知每一 DNA 核甘酸在每一基因片段上每個位置出現的機率值，進而計算出每一基因片段可能出現的機率，建立起新的機率分佈。

MEME 的優點在於它能自動搜尋最適當長度的重要基因片段，意即在使用者僅給定重要基因片段的長度範圍而非特定長度的情況下，它依舊能找到具代表性的重要基因片段。除此之外，**MEME** 亦能運用統計法估量此重要基因片段的重要性。然而，由於 **MEME** 採用 **EM** 的概念，其很有可能因一開始選定的初始位置比重矩陣而導致最後的結果陷入**局部最佳化 (local optimal)**。

2.1.2 WINNOWER (Pevzner and Sze, 2000)

在 2000 年，Pevzner 和 Sze 兩位學者發展出了 **WINNOWER**，此種融入圖論概念、將單一重要基因片段視爲圖形上的單一 **clique** 的探測演算法。

在使用者給定欲搜尋重要基因片段之特定長度 ***l*** 和容許的突變數目 ***d*** 後，**WINNOWER** 會將在任一基因序列上出現且長度爲 ***l*** 的片段 **(*l*-mers)** 先各立爲王成爲一特徵點，隨即再將相似的片段用線段連結起。而對其要尋找型態爲 ***(l, d)*** 的重要基因片段而言，要將兩片段視爲相似得要此兩片段中的差異位元數不能超過 ***2d***。

最後，交錯的線段若能形成一**完全子圖形 (complete sub-graph)**，將能構成一 clique 進而呈現出探測而得的重要基因片段。如 **圖 2.2** 中所示，當使用者想要找尋型態爲 (15, 4) 的重要基因片段時，**WINNOWER** 將會把差異在 8 個位元內的相似特徵點相互用線段兩兩連結起，最後探測到單元重要基因片段 ***M* = AAAAAAAAGGGGGGG**。

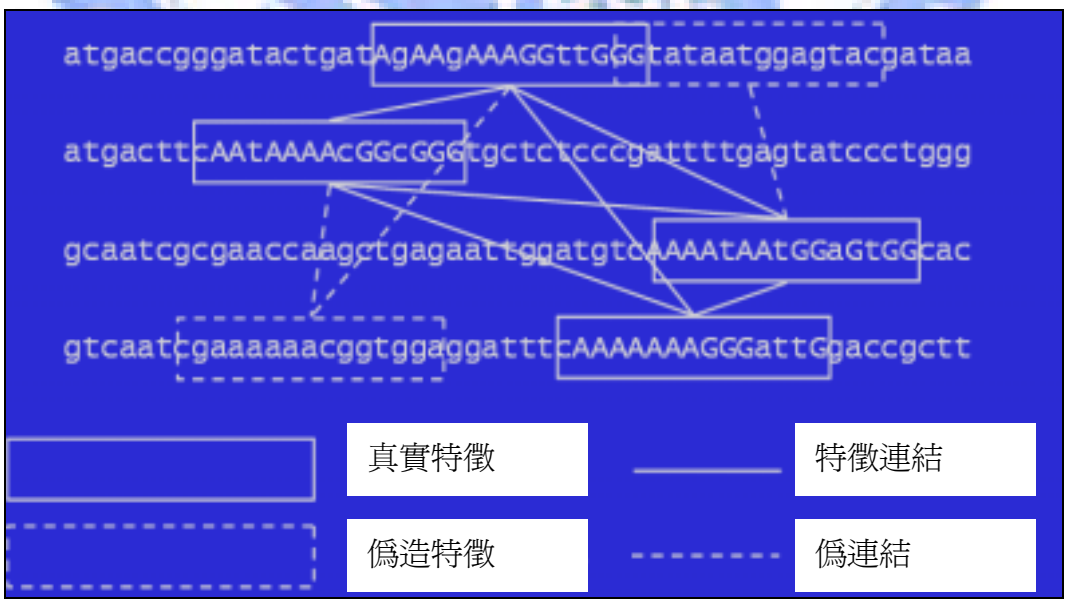


圖 2.2 WINNOWER 依據基因群及使用者輸入資訊所建構的圖形。

僞造特徵點表示並未能與其他特徵點形成一完成子圖形的特徵點。

然而由圖形中探測出 **cliques** 的存在在圖論演算法中本來就屬於 NP 的問題，意即其複雜度相當之高。因此 **WINNOWER** 得運用重覆擷取的方式將僞連結移除進而得到幾近正確的 **cliques**。

2.1.3 WEEDER (Pavesi et al., 2001)

Pavesi 等學者在 2001 年提出的 **WEEDER** 演算法的主要貢獻在於其能探測到未知長度的單元重要基因片段。這對生物學家來說是相當有助益的，畢竟在新的 DNA 序列定序完成後，未經由實驗驗證前，生物學家亦無法確定重要基因片段的長度。

使用者在使用 **WEEDER** 演算法探測單元重要基因片段時，僅須設定容錯度 ε **($\varepsilon < 1$)**。隨後 **WEEDER** 將會找尋所有在合理數量的基因序列上皆有相似片段的單元重要基因片段。

WEEDER 的核心概念在於－倘若單元重要基因片段的長度爲 l ，則其與相似片段至多僅能有 $|\varepsilon l|$ 個位置突變。而 **WEEDER** 演算法爲了降低合理相似片段的搜尋空間，其訂定一個特殊限制：單元重要基因片段與其相似片段中，第 i 個核甘酸前，僅能允許 $|\varepsilon i|$ 個核甘酸相異。（**圖 2.3**）

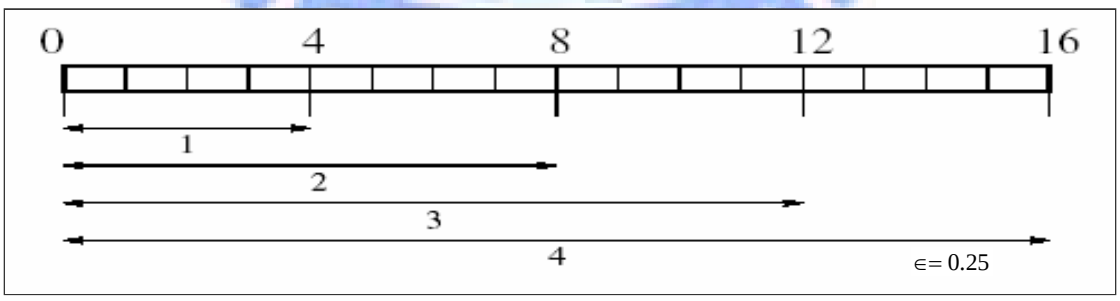


圖 2.3 WEEDER 所使用的區塊分解法能降低合理相似片段的搜尋空間。如當 ε=0.25 時，單元重要基因片段其其相似片段中，第 4 個核甘酸前僅允許 1 個核甘酸相異。 (Pavesi et al., 2001)

不過，倘若使用者允許的 DNA 核甘酸突變個數增多，意即 ε 的值變大，亦或是單元重要基因片段的長度增長，都將使得 **WEEDER** 整個搜尋流程的複雜程度以倍數成長。因此我們可以得知，探測未知長度的重要基因片段雖然較爲符合使用者的需求，不過其所需面對的難題也較具挑戰性。

2.2 多元重要基因片段探測法概述

2.2.1 Dyad Analysis (van Helden et al., 2000)

van Helden 等學家在 2000 年首度提出方法探測雙元重要基因片段，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，進而探測出雙元重要基因片段。（**圖 2.4**）

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 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。 (van Helden et al., 2000)

然而 **Dyad Analysis** 只能探測固定間距的雙元重要基因片段且間距的設定只能介於 0 到 16 間，因此其所能探測到的固定間距之雙元重要基因片段仍是有限的。

2.2.2 MITRA (Eskin and Pevzner, 2003)

爲了探測由數個特徵微弱的單元重要基因片段所組成的多元重要基因片段（**圖 2.5**），Eskin 和 Pevner 兩位學者在 2003 年所提出了 **MITRA (MIsmatch TRee Algorithm)** 演算法。

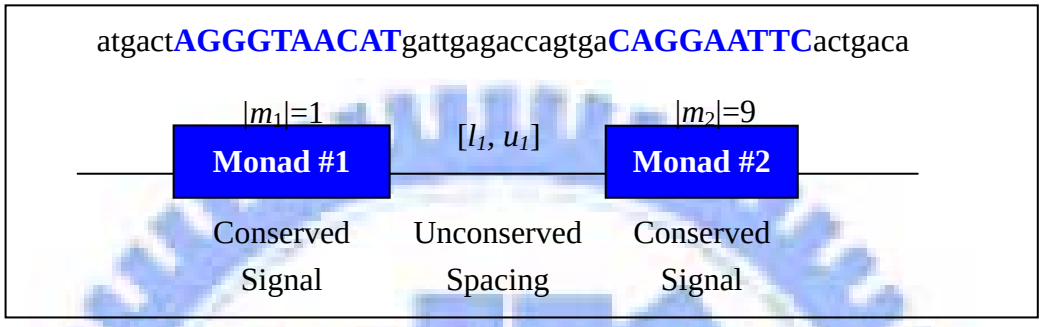


圖 2.5　由 m_1 和 m_2 兩個單元重要基因片段以及座落在其間的間距範圍 $[l_1, u_1]$ 組合而成的雙元重要基因片段

MITRA 的核心概念在於藉由基因序列群的前置處理，將多元重要基因片段轉化成較長的單元重要基因片段（我們在此稱爲仿單元重要基因片段）進而運用探測單元重要基因片段的方法探測而得。

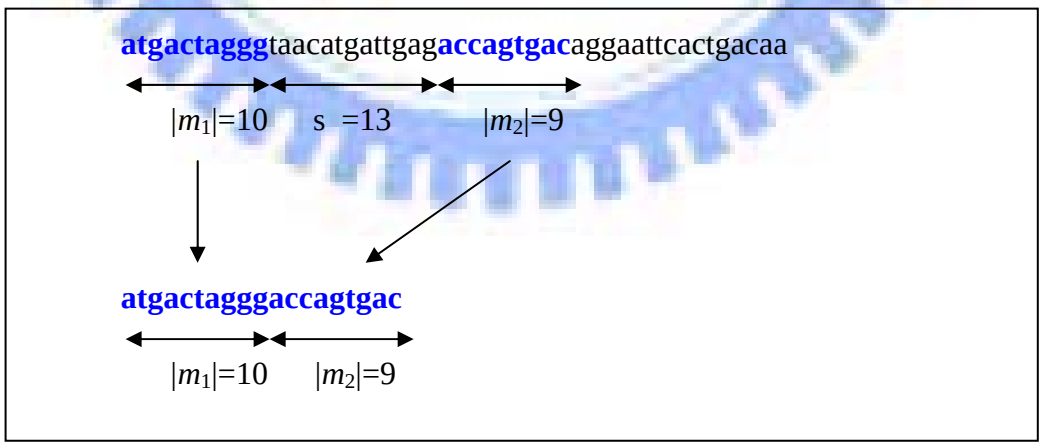


圖 2.6　MITRA 將雙元重要基因片段對應至相對的單元重要基因片段的流程

由 **圖 2.6** 中可得知－在將雙元重要基因片段對應至單元重要基因片段的過程中，**MITRA** 會將每一 $|m_1|$**-mer** 藉由連結間隔 $s \in [l_1, u_1]$ 位元的 $|m_1|$**-mer** 和 $|m_1|$**-mer** 的方式擴充成 $(|m_1|+|m_2|)$**-mer**。因此，此步驟不僅會使欲搜尋的單元重要基因片段長度變長，亦使得整個樣本空間增大爲原先的 $(u_1 - l_1 + 1)$ 倍。

因而我們可以推斷 **MITRA** 雖然能夠探測可變間距之多元重要基因片段，但是其搜尋的空間與時間都深受可變間距的長度及子片段的個數所影響，此意謂著 **MITRA** 演算法並不具可擴充性。換句話說，假設有一由 3 個或多個單元重要基因片段所組成的多元重要基因片段，則 **MITRA** 在探測此多元重要基因片段上便會遇到了阻礙。

Name	Dyads found by MITRA-Dyad
purC	TTTGCCAGATATATGTCTAA – –(23)– – TTTTACATAAACATGGTGAA
purF	TTCACCATGTTTATGTAAAA – –(23)– – TTAGACATATATCTGGCAAA
purT	TTAAACATATTTATGTTAAA – –(23)– – TTTAACATTTATACGTCAAT
purE	ATTAGCACATATATGTTAAA – –(23)– – ATTGACATTAAATTGCTAGG
purD	GTTAACACGTTTATGTAAAC – –(23)– – TTTGACTTAAATATGGTGAT
purA	ATTAACATAGCCCTGTCAAA – –(23)– – CTTTACTTACCCTTTGGTAA
purB	ATTTCTACAAATATGTCAAA – –(23)– – TTTACCGTGAAAATGGTGAT
purL-11	ATTGACATTTCTTTGTCAAA – –(22)– – TTTTACATTTTTCTGGCAAA
cons.	ATTAACATATATATGTACAA –(22, 23)– TTTTACATATATATGGTAAA

圖 2.7　MITRA 由 P. horikoshii 資料群中所探測而得的結果。

(Eskin and Pevzner, 2003)

2.3 Apriori 演算法

由於我們所提出的 **APPA** 演算法是延伸 **Apriori** 的概念進而藉由單元基因片段間的關聯來探測出多元重要基因片段，因此在此章節中，我們將會仔細講解整個 **Apriori** 演算法的運作流程。

Apriori 演算法主要是運用 ***k*** 個項目組成的項目群組 **(k-itemset)** 的出現頻率進而建構起 **(k+1)** 個項目所組成的項目群組 **((k+1)-itemset)**。其中的基本概念即爲：只有在 **{A}** 和 **{B}** 兩項目群組具有常頻的特色時，**{A, B}** 才會被建構成爲常頻項目群組。因此在這樣的前題下，具有高度相關的關聯性即能被探測出來。

讓我們藉由 **圖 2.8** 來了解整個 **Apriori** 運作的過程以及其核心概念。圖中的資料庫 **D** 中已知有 5 個項目出現在其中，分別爲 **{1}~{5}**。爲了擷取出項目與項目間的相關性，設定頻率門檻爲 2，而後計算每一項目的出現頻率進而作篩選。由於項目 **{4}** 的頻率值 **fre({4})=1** 低於門檻值 2，因此將其排除在擴充項目之外。隨後再針對留下的候選者，兩兩結合擴充成爲含括兩項目的項目群組 **(2-itemset)**。一直重複－計算項目群組的出現頻率、排除低於門檻值的項目群組、項目群組的擴充－這樣的流程，最後可得項目間的最強關聯爲 **{2 3 5}**。

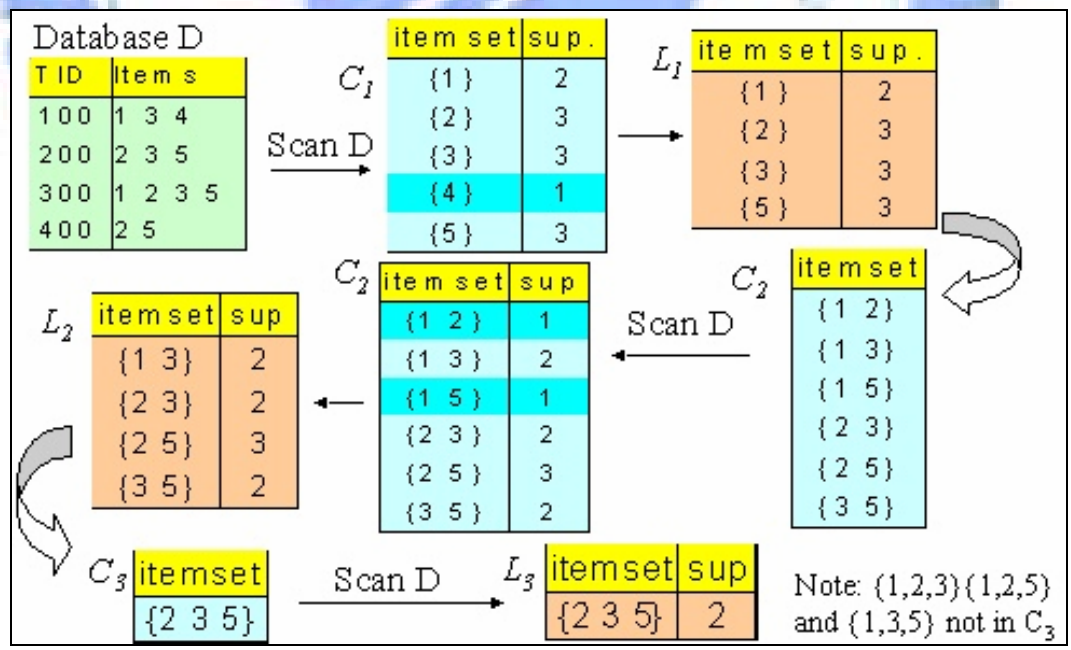


圖 2.8　運用 Apriori 從資料庫 D 中找出{2 3 5}的高度關聯性

(Agrawal and Srikant, 1994)

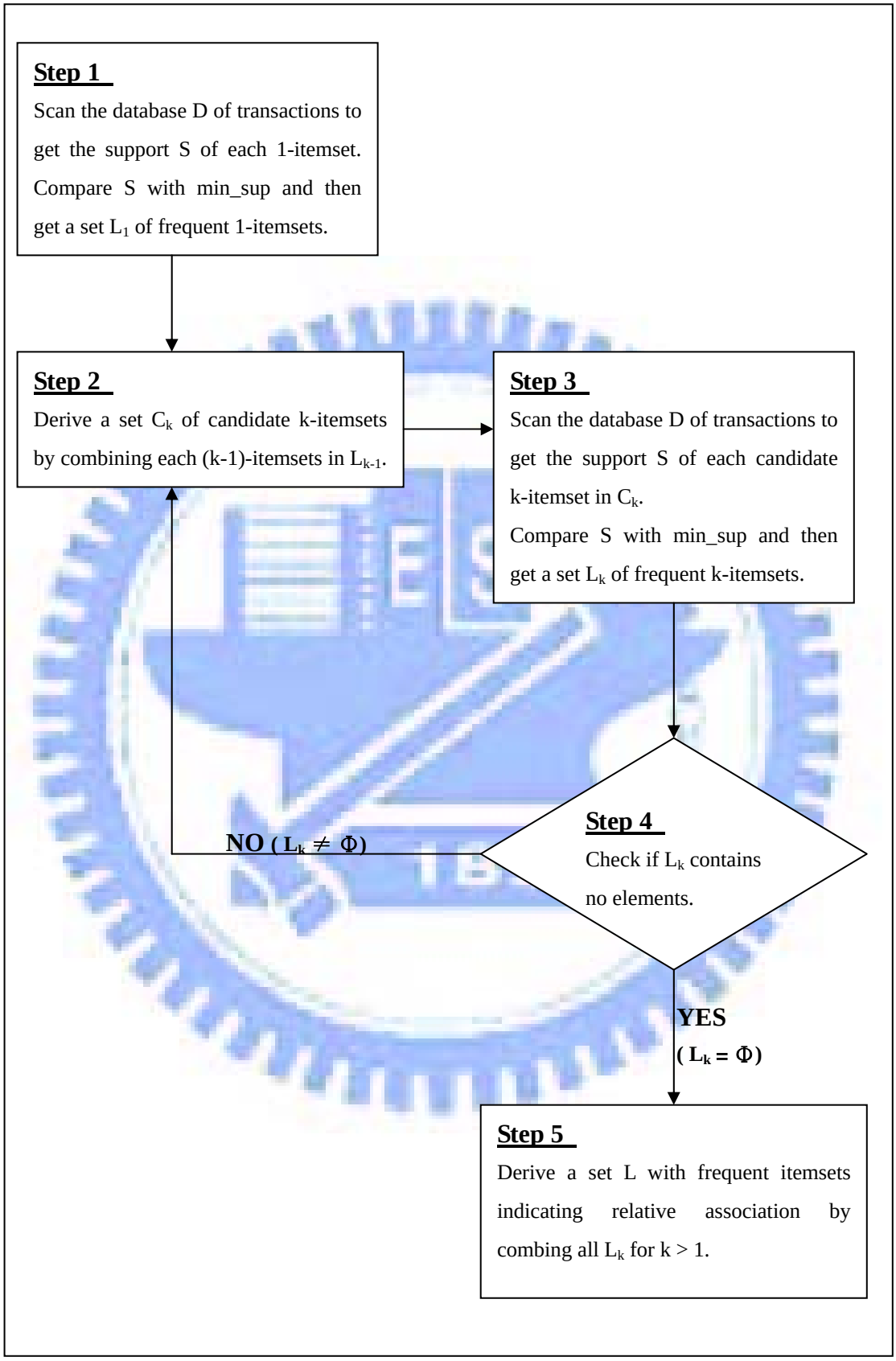


圖 2.9　Apriori 的運作流程（min_sup 表示頻率門檻）

Input: A database D of transactions and min_sup

Output: A set L containing all frequent itemsets of D

Algorithm:

(1) $L_1 \leftarrow$ {frequent 1-itemsets};

(2) for ($k = 2$; $L_{k-1} \neq \phi$; $k++$) do begin

(3) $C_k \leftarrow$ Apriori-Gen(L_{k-1}); //new candidates

(4) forall transactions $t \in D$ do begin

(5) $C_t \leftarrow$ Subset(C_k, t); //candidates contained in t

(6) forall candidates $c \in C_t$ do

(7) $c.count++$;

(8) end

(9) $L_k \leftarrow \{c \in C_k \mid c.count \geq \min\_sup\}$;

(10) end

(11) $L = \bigcup_{k>1} L_k$;

圖 2.10 Apriori 的 pesudocode。其中的 min_sup 表示頻率門檻。

(Agrawal and Srikant, 1994)

第三章　系統架構與研究方法

在此章的第一單元中，我們會先介紹 **APPA** 的核心概念；隨後在第二單元中加以詳述整個 **APPA** 演算法的循序流程，進而詳細介紹每一步驟的細節。最後簡介 **APPA** 系統的使用者介面。

3.1 核心概念

經由第二章的文獻探討後，我們希望能在讓使用者僅需設定容錯度 ε 和對應資料群 ***B*** 的情況下就能由給定的基因序列群 ***S*** 中探測到多元重要基因片段，因而我們提出以 **MERMAID** 擴充演算法作爲單元重要基因片段探測法、且以 **WEEDER** 和 **Apriori** 的概念作爲延伸的 **APPA (APriori Pruning Algorithm)** 演算法來探測由多個單元重要基因片段所組成的多元重要基因片段。（**圖 3.1**）

APPA 系統的輸入與輸出

《輸入》 基因序列群 ***S***、選定對應資料群 ***B*** 及容錯度 ε <1

《輸出》 所有可能的重要基因片段

圖 3.1　APPA 系統的輸出與輸入

而 **APPA** 的核心概念主要建立在一組成多元重要基因片段 ***M*** $=\left(m_1,(l_1,u_1),m_2,.....,(l_{p-1},u_{p-1}),m_p\right)$ 的 ***p*** 個單元重要基因片段彼此間的關係是相當密切的－這樣的假設上。因此，當 $\mathbf{m}_i$**-mer** 在某序列上存在時，$\mathbf{m}_j$**-mer** 極有可能也存在此序列上 $(i \neq j,\ 1 \leq i, j \leq p)$。

也因爲這樣的概念，我們選擇用與 **Apriori** 方法雷同的擷取方式由單元重要基因片段群集中，探測出其中的重要關聯。

圖 3.2 結合 Apriori 與 WEEDER 概念和 MERMIAD 的 APPA 演算法流程

(紅色區塊中為核心流程)

3.2 方法詳述

如同 **圖 3.2** 中所呈現的，**APPA** 演算法總共含括四大步驟：(1) 探測單元重要基因片段、(2) 標記基因序列群、(3) 探測固定間距的多元重要基因片段，以及 (4) 多元重要基因片段的整合。

3.2.1 單元重要基因片段的探測 (Monad Generation)

這一步驟的主要是運用 **MERMAID** 的擴充方法從使用者輸入的基因序列群 ***S*** 中探測出所有可能長度由 3 至 8 的單元重要基因片段。

Method	significant scale
CONSENSUS	0.06
Gibbs sampler	0.11
MEME	0.02
MEME (w/iterative)	0.09
Oligonucleotide analysis (van Helden)	0.00
WINNOWER	0.88
SP-STAR	0.23
MERMAID	0.75

圖 3.3　各演算法針對 Challenge Problem 探測(15, 4)之重要基因片段的表現 (Hu, 2003)

在 **圖 3.3** 所示爲各演算法針對 Challenge Problem 的探測表現，而我們可以從中觀察出 **MERMAID** 在效能上遠比其他演算法來得傑出。雖然亦可從中觀察出 **WINNOWER** 的效能較 **MERMAID** 好，不過由於以圖論爲延伸的 **WINNOWER** 在探測重要基因片段上的複雜度遠比 **MERMAID** 要來得高許多，所以我們選擇用 **MERMAID** 的擴充演算法作爲 **APPA** 探測單元重要基因片段的模組。

MERMAID 演算法選擇用**權重矩陣 (weight matrix)** 作爲重要基因片段的表現型態，按照實作重複取樣最佳化的程序，進而探測到最具代表性的重要基因片段。其使用了大量多樣的客觀性函數作爲重要基因片段評分的考量，如重要基因片段的**一致性 (consensus quality)**、**重複多樣性 (multiplicity significance)** 以及**覆蓋範圍 (motif coverage)**。

(a)《一致性評分函數》

$$\mathrm{con}(m)=\frac{1}{w}\sum_{1}^{w}C(n),\ \text{where}\ \ C(n)=2-E(n)\ \ \text{and}\ \ E(n)=-\sum_{i=b1}^{b4}Pm_i\cdot\log_2 Pm_i$$

《重複多樣性評分函數》 $\mathrm{mul}(m)=\dfrac{occ_S(m)}{occ_B(m)}$

《覆蓋範圍評分函數》 $\mathrm{cov}(m)=\dfrac{cont_S(m)}{|S|}$

(b) $\mathrm{Con}_{\mathrm{norm}}(m)=\dfrac{\mathrm{con}(m)}{\mathrm{MAX(Con)}}$, $\mathrm{Mul}_{\mathrm{norm}}(m)=\dfrac{\mathrm{mul}(m)}{\mathrm{MAX(Mul)}}$, $\mathrm{Cov}_{\mathrm{norm}}(m)=\dfrac{\mathrm{cov}(m)}{\mathrm{MAX(Cov)}}$

(c) 《**MERMAID 總評分**》 $\mathrm{Merit}(m)=\dfrac{1}{\frac{1}{3}\left(\frac{1}{\mathrm{Con}_{\mathrm{norm}}(m)}+\frac{1}{\mathrm{Mul}_{\mathrm{norm}}(m)}+\frac{1}{\mathrm{Cov}_{\mathrm{norm}}(m)}\right)}$

(d) 《**APPA 評估函數**》 $\mathrm{merit}(m)=\dfrac{1}{\frac{1}{2}\left(\frac{1}{\mathrm{Mul}_{\mathrm{norm}}(m)}+\frac{1}{\mathrm{Cov}_{\mathrm{norm}}(m)}\right)}$

圖 3.4 MERMAID 評量重要基因片段的多樣客觀性函數 (Hu, 2003) 及所衍生 APPA 對單元基因片段的評估函數

圖 3.4 中爲 **MERMAID** 對於各個重要基因片段的給分方式，詳述如下：

定義特定重要基因片段的權重矩陣 **m** 中一致性的值是由**波動量 (entropy)** 來衡量的。波動量的值則由每一 DNA 核甘酸發生在此重要基因片段中的機率計算而得，在此以 Pm_{base} 作爲標識。 ***E(n)*** 表示對特定矩陣的特定行 ***n*** 計算的波動量。由於波動量的最大值爲 2，所以設定對特定矩陣的特定行的一致性爲 ***C(n)=2 −E(n)***。因而表示長度爲 **W** 的重要基因片段的特定矩陣 ***m*** 的一致性數值即爲 ***Con(m)***。

多樣重複性 ***Mul(m)*** 則是由資料擷取中所定義的精確值而得。構成 ***Mul(m)*** 的 $occ_f(m)$ 表示在基因集合 ***f*** 中，重要基因片段 ***m*** 出現的次數。此特徵能呈現重要基因片段出現在給定基因序列群 ***S*** 和對應基因序列群 ***B*** 中次數的比例。而覆蓋範圍 ***Cov(m)*** 可被定義成含有重要基因片段 ***m*** 的序列數量 $cont_S(m)$ 與給定基因序列群中序列數量 $|S|$ 的比例。

在給定基因序列群 ***S*** 後，**MERMAID** 會依據重要基因片段的一致性、多樣重複性以及覆蓋範圍的整合作爲評量方式，即爲 ***Merit(m)***，然後將探測而得的單元重要基因片段排名分等。（**圖 3.5**）

重要基因片段 *m*	評估函數				權重矩陣 weight_matrix(m)
	一般性 *con(m)*	重複性 *mul(m)*	覆蓋範圍 *cov(m)*	總分 *Merit(m)*	
CGCTG	1.00	0.87	1.00	0.954	A 0.0 0.0 0.0 0.0 0.0 G 0.0 1.0 0.0 0.0 1.0 C 1.0 0.0 1.0 0.0 0.0 T 0.0 0.0 0.0 1.0 0.0

圖 3.5　原 MERMAID 演算法所求得長度為 5 單元重要基因片段之格式

APPA 則經由此步驟將會得到一群長度由 3 至 8 不等的單元重要基因片段。而後 **APPA** 會採用類似 **MERMAID** 的計算方式 ***merit(m)*** 來給予單元重要基因片段的評分。（**圖 3.6**）

重要基因片段 *m*	MERMAID 評估函數				APPA 評估函數 *merit(m)*
	一般性 *con(m)*	重複性 *mul(m)*	覆蓋範圍 *cov(m)*	總分 *Merit(m)*	
GCG	1.00	1.00	1.00	1.00	1.00
CGCG	1.00	1.00	1.00	1.00	1.00
CGCTG	1.00	0.87	1.00	0.954	0.93
GCATTC	1.00	1.00	1.00	1.00	1.00
TGCATTC	1.00	1.00	1.00	1.00	1.00
CTGCATTC	1.00	1.00	1.00	1.00	1.00

圖 3.6　APPA 經由擴充 MERMAID 所得的單元重要基因片段及其評分

```
Given: a set of biosequences, S
Return: a set of ranked motif candidates, C={ Ci | 3 <= i <= 8}

(1) For W =3 to W=8 ( W：the length of a monad motif)
(2)     For each substring s in S Do
(3)         Set s to ss as seed

(4)         Repeat
(5)             Translate ss into candidate probability matrix m via:
                    m(i, base) = 0.50 if base occurs in position i
                               = 0.17 otherwise

(6)             Find highest match scoring substring in each sequence in S
(7)             Compute the mean of the highest match scores in S
(8)             For each sequence in S Do
(9)                 Set Potential Positions to those with match score >= mean

(10)            Randomly choose a Potential Position in each sequence to initialize matrix M
(11)            Repeat
(12)                Randomly pick a sequence s in S
(13)                Check if M's quality can be improved by using a different Potential Position in s
(14)                Update matrix M
(15)            Until (no improvement in M's quality) or (reach the cycle limit 5)

(16)            Compute the mean of match scores of substrings contributing to M
(17)            For each sequence s in S Do
(18)                Isolate motif repeats to those with match score >= mean
(19)            Form the final matrix FM with all repeats in S
(20)            Convert matrix FM into string ss as a new seed
(21)        Until (no improvement in FM's quality) or (reach the cycle limit 5)
(22)        Put FM in Cw
(23)    End For

(24)    Sort all motif candidates in Cw according to significance
(25)    C=C∪Cw
(26)End For
(27)Return C
```

圖 3.7　MERMAID 之擴充演算法的 pseudo-code

3.2.2 基因序列群的標記 (Sequence Notation)

MERMAID 所產生的單元重要基因片段在此步驟裡都將被給予唯一的辨視值 **(id)**，用以對基因序列群中的所有序列作標記的動作，進而將所有單元重要基因片段與基因序列群相關的資訊一次建構起來。（**圖 3.8**）

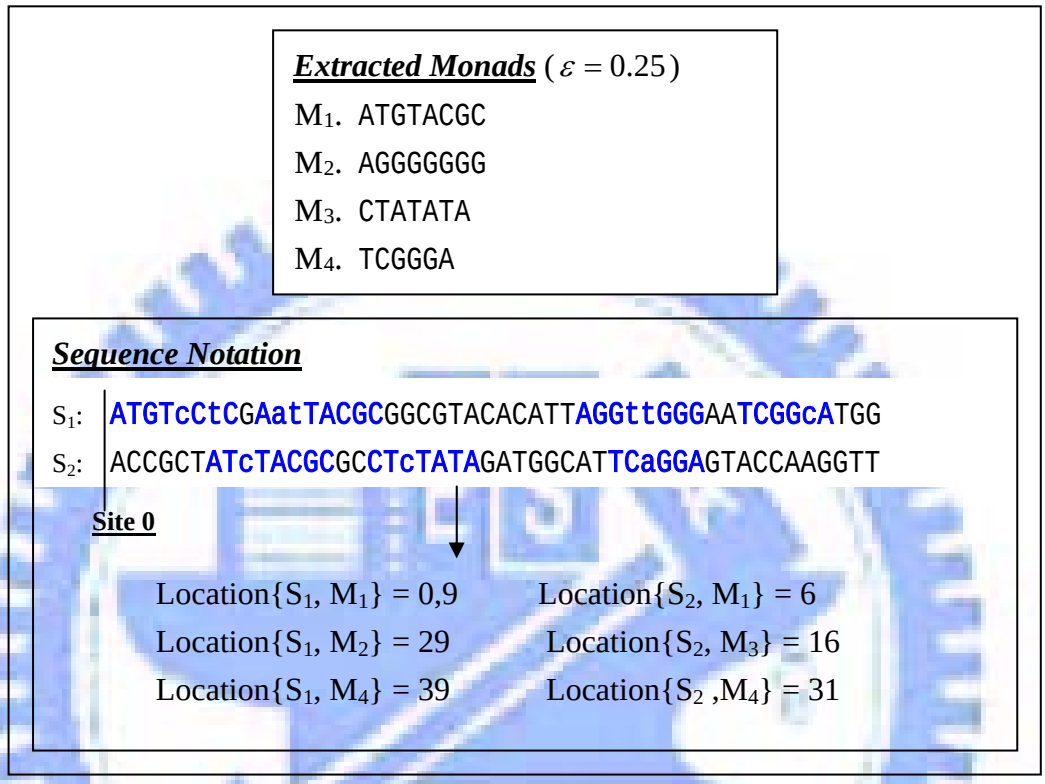


圖 3.8　APPA 運用單元重要基因片段對每一基因序列加以標記

在這步驟中，使用者所設定的容錯度 ε 將會產生效用，用以判別某一特定重要基因片段 m_j 在基因序列 S_i 上是否有相似片段 m'_j 的存在。倘若有，則 $Location\{i, j\}$ 將會記錄 m_j 在基因序列 S_i 所有出現的位置。而 m_j 和 m'_j 相似的條件在於其差異的核甘酸個數不超過 $\varepsilon \cdot |m_j|$，其中 $|m_j|$ 爲重要基因片段 m_j 的長度。

除此之外，在此步驟中，我們會將單元重要基因片段依據其 **MERMAID** 所給定的矩陣資訊作些許的調整，使得單元重要基因片段的格式可由 {A,T,G,C} 擴展爲{A,T,G,C,R,Y,K,M,S,W}。

$SG\_score\ (m) = real\_SG\_score\ (m) = \frac{cont_S(m)+1}{re\_cont_B(m)}$ $re\_cont_B\ (m) = cont_B(m) + \|S\| + 1$ $sig(m) = (\ Merit(m) + SG\_score(m)\ ) * \|m\|$
$\|S\|$－給定基因序列群中 S 的序列數目，$\|m\|$－重要基因片段 m 的長度 $cont_F\ (m)$－在序列群 F 中含有重要基因片段 m 的序列數量

圖 3.9　APPA 評量重要基因片段 m 的多樣客觀性函數

除此之外，**APPA** 在此過程中，還會給定每一單元重要基因片段 ***m*** 一個評比分數，***SG_score(m)***，以作爲之後強度關聯的篩選值。因爲 ***SG_score(m)*** 的用意在於呈現給定基因群 ***S*** 和對應資料群 ***B*** 中含有 ***m*** 序列數的比例。當 ***SG_score(m)*** 的值愈大，即表示此單元重要基因片段 ***m*** 愈顯重要。因此，***SG_score(m)*** 將會取代在 ***Apriori*** 演算法中的常頻次數，作爲 **APPA** 演算法中強度關聯的篩選值。（**圖 3.9**）

而在 **APPA** 中重要基因片段 ***m*** 的綜合優勢評估函數，***sig(m)***，則整合 **MERMAID** 所給定的優勢評估函數值 ***Merit(m)*** 和 ***SG_score(m)***。（**圖 3.9**）

Monad id	pattern	MERMAID	APPA		sig.
id{m}	*m*	*merit(m)*|m|*	*SG_score(m)*|m|*	*sig(m)*	
1	CGC	3.00	0.00900	3.0090	No
9	CGGC	4.00	0.01223	4.0122	No
17	CCGCG	5.00	0.02417	5.0242	No
18	CGCGG	5.00	0.02417	5.0079	No
23	GCCTCC	6.00	0.01485	6.0883	No
24	TCGCGG	5.99	0.00775	5.9978	No
39	GCCTCCC	7.00	0.38583	7.3858	**Yes**
40	GGGAGGC	6.69	0.01624	6.7062	No

圖 3.10　APPA 評量單元重要基因片段的格式

經由這樣的相關資訊建構法，不僅僅讓之後固定間距之多元重要基因片段的探測過程簡易許多，更可以處理單元重要基因片段相互交錯存在 **(overlapping)** 的問題以及保留各種單元重要基因片段組合而成的多元重要基因片段。

3.2.3 固定間距之多元重要基因片段的探測 (Relation Generation)

經由在第二步驟中建構所有相關的完整資訊後，**APPA** 在此過程中將會運用 **Apriori** 演算法且以 ***SG_score*** 作爲篩選高度關聯的門檻，而過濾出所有具高度相關性的單元重要基因片段組合，此任一組合皆代表著一多元重要基因片段。(**圖 3.11**)

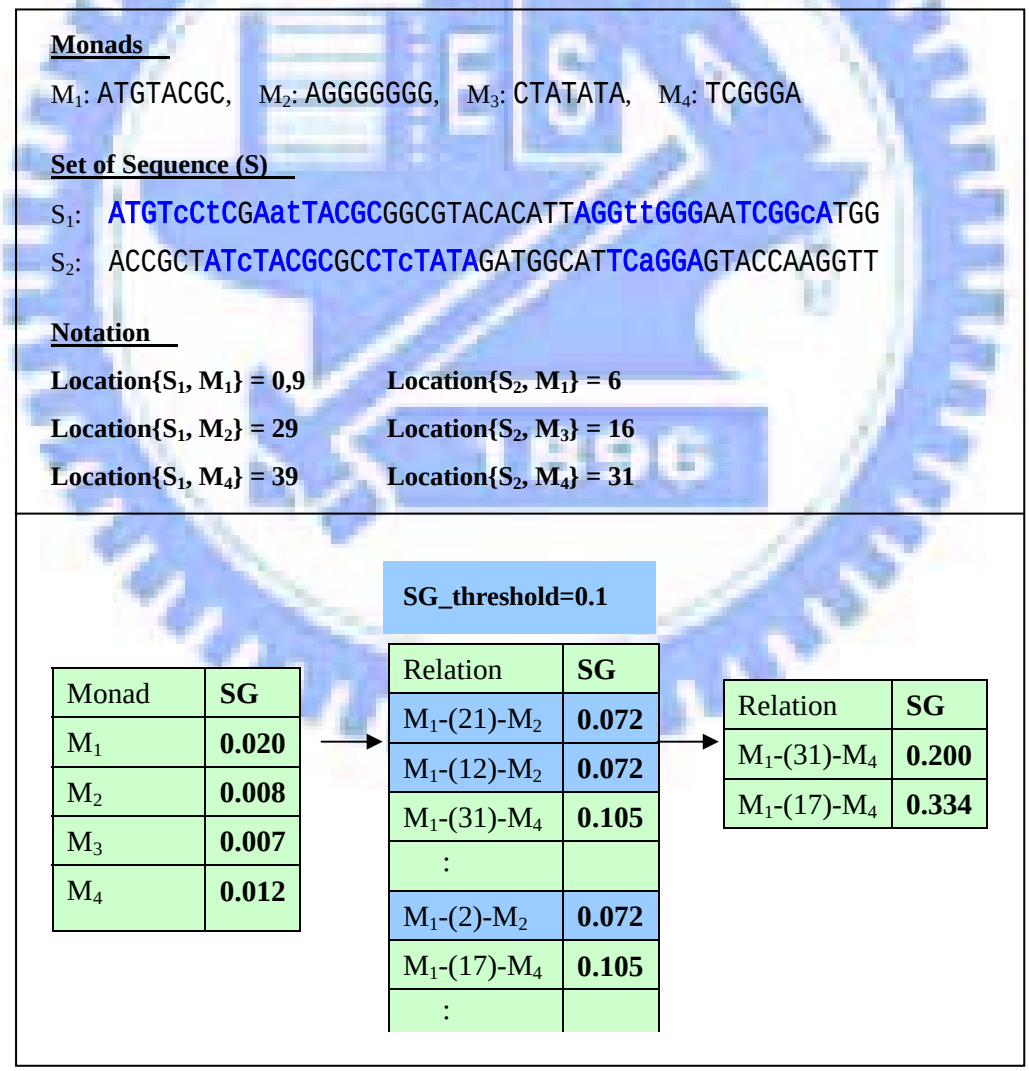


圖 3.11　APPA 運用單元重要基因片段及基因序列群標記的資訊擷取固定間距之多元重要基因片段

然而每一次重複－計算由 ***k*** 個單元重要基因片段組成的多元基因片段的 ***SG_score***、排除低於門檻值的多元基因片段、將多元重要基因片段子片段個數的擴充爲 ***(k+1)***－這樣的流程，倘若真的去計算每一多元基因片段的 ***SG_score***，會大大地增加探測時間和空間的複雜度，因爲 ***SG_score*** 需藉由此多元基因片段在對應基因序列群 ***B*** 中出現的次數加以運算而得。

因此，我們將用 ***pseudo_SG_score*** 來衡量雙元基因片段中的 ***SG_score***，意即雙元基因片段的 ***SG_score*** 將由爲其構成要素的單元重要基因片段的資訊建構而得。

假設有一雙元基因片段 $d = m_1 - (g) - m_2$，其由單元重要基因片段 m_1 和距離 m_1 後方間隔長度 g 的 m_2 所組成。而 $cont_S(m_i) = c_i$ 表示 m_i 在給定基因序列群 S 中共出現在 c_i 條序列上。

$$pseudo\_cont_S(d) = \min(\, cont_S(m_1), cont_S(m_2)\,)$$

$$pseudo\_re\_cont_B(d) = \frac{(cont_B(m_1) + cont_S(m_1) + 1) \times (cont_B(m_2) + cont_S(m_2) + 1)}{(|B| + |S| + 1)}$$

$$SG\_score(d) = pseudo\_SG\_score(d) = \frac{pseudo\_cont_S(d) + 1}{pseudo\_re\_cont_B(d)}$$

$|F|$－基因序列群 F 中的序列數目

$cont_F(m)$－在序列群 F 中含有重要基因片段 m 的序列數量

當 ***SG_score(d)*** 的分數高過於分檻時，雙元基因片段才被視爲具有代表性的雙元重要基因片段，且能作爲擴充成三元基因片段的候選者。爲什麼我們能確定 ***pseudo_SG_score(d)*** 能模擬真實的 ***SG_score(d)*** 呢？我們將依據 **(a) *d* 確爲應探測到的雙元重要基因片段** 以及 **(b) *d* 並非爲探測結果** 此兩種假設作討論，進而闡述 ***pseudo_SG_score(d)*** 的可行性。

(a) 假設雙元基因片段 ***d*** 爲應被探測到的結果，則 ***d*** 理應在對應基因序列群 ***B*** 中不具任何特徵性，意即組成 ***d*** 的單元重要基因片段 m_1 和 m_2 間在 ***B*** 中並不具有高度關聯。在此種假設下，我們可以將 m_1 和 m_2 在對應基因序列群 ***B*** 中出現機率視爲獨立，則 m_1 和 m_2 在對應基因序列群 ***B*** 中同時出現的機率等同於 m_1 和 m_2 分別在對應基因序列群 ***B*** 中出現機率的乘積，即 $P_B(m_1 \cap m_2)=P_B(m_1) \times P_B(m_2)$。因此 $pseudo\_re\_cont_B(d)$ 的值將會大於或等於 $re\_cont_B(d)$ 的數值，因爲 $pseudo\_re\_cont_B(d)$ 中並無考量到 m_1 和 m_2 間的間距長度 g。由於 ***d*** 爲須被探測到的雙元重要基因片段，則 ***pseudo_SG_score(d)*** 理應大於門檻值 *SG_threshold* 。最後可推得－

$$real\_SG\_score(d) \geqq pseudo\_SG\_score(d) \geqq SG\_threshold$$

意即運用 ***pseudo_SG_score*** 而被選取的雙元重要基因序列在使用一開始定欺的 ***real_SG_score*** 的方式下也定會被選取。

(b) 倘若雙元基因片段 ***d*** 在給定的基因序列群 ***S*** 和對應基因序列群 ***B*** 中都是具有特徵性的，由於 m_1 和 m_2 具有高度關聯，所以 $re\_cont_B(d)$ 的數值將會比 $pseudo\_re\_cont_B(d)$ 的值來的大。又因爲 ***d*** 並非我們所要探測的結果，所以我們希望在過程中能將雙元基因片段 ***d*** 從擴充候選人中篩選掉，因此我們會得到－

$$SG\_threshold \geqq pseudo\_SG\_score(d) \geqq real\_SG\_score(d)$$

換句話說，在利用 ***pseudo_SG_score*** 篩選掉的雙元基因片段在運用 ***real_SG_score*** 的方式下也定會被篩選掉。

同理，多元重要基因片段的建構篩選方式也與雙元重要基因片段的方式同。且藉由此種擬真的 ***pseudo_SG_score*** 評分方式能使得多元重要基因片段的擴充快速很多。而此快速評分方式的可行性也將在第四章的核心模組實驗中被驗證。

在此步驟雖用 ***pseudo_SG_score*** 過濾多元重要基因片段，不過之後具代表性的多元重要基因片段的資訊會再用 ***SG_score*** 將其正確資訊建構起來，以正確建構以此多元重要基因片段擴充的多元重要基因片段。

3.2.4 多元重要基因片段的整合 (Motif Derivation)

經由固定間距多元重要基因片段的探測過程後，**APPA** 會得到所有具代表性的多元重要基因片段。而 **APPA** 在最後一步驟中會由固定間距的多元重要基因片段盡可能地推導出可變間距的多元重要基因片段。

而決定兩個恆定間距的多元重要基因片段 d_1 和 d_2 可合而爲一可變間距多元重要基因片段 d 的條件如下－

(a) 構成 d_1 和 d_2 的任一單元重要基因片段 $m^1_i = m^2_i$ 皆需成立，d_1 和 d_2 的差異僅在座落於兩兩單元重要基因片段間距的長度不同。

(b) $SG\_score(d) \geqq \dfrac{SG\_score(d_1) + SG\_score(d_2)}{2}$

type	relation	string representation	SG_score
fixed	**M_1-(31)-M_4**	ATGTACGC -(31)-TCGGGA	0.200
fixed	**M_1-(17)-M_4**	ATGTACGC-(17)-TCGGGA	0.334
cons./variant	**M_1-(17, 31)-M_4**	ATGTACGC-(17, 31)-TCGGGA	0.286

圖 3.12 APPA 由固定間距多元重要基因片段對應至可變間距的推導

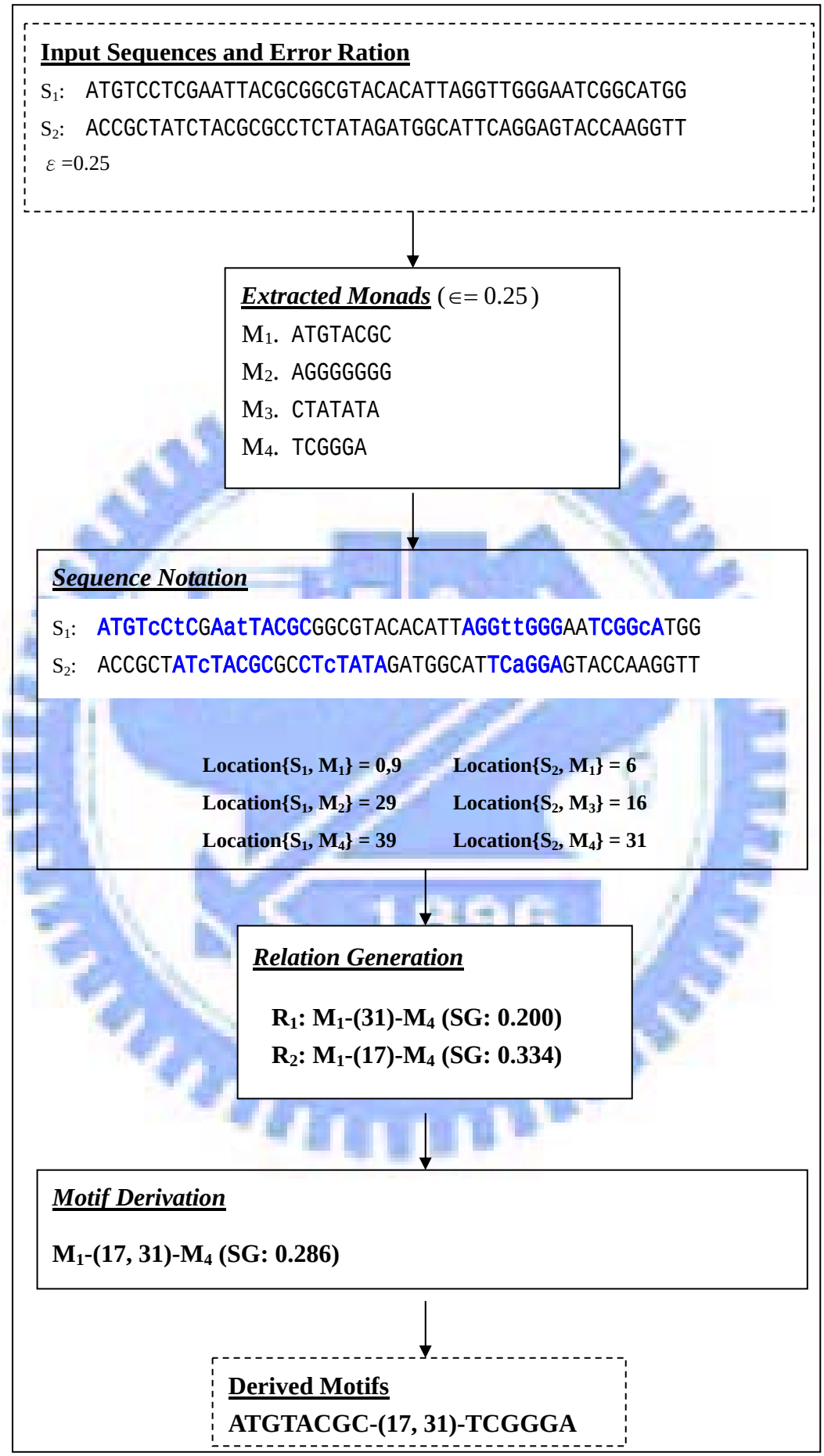


圖 3.13　APPA 完整流程的舉例

3.3 系統介面

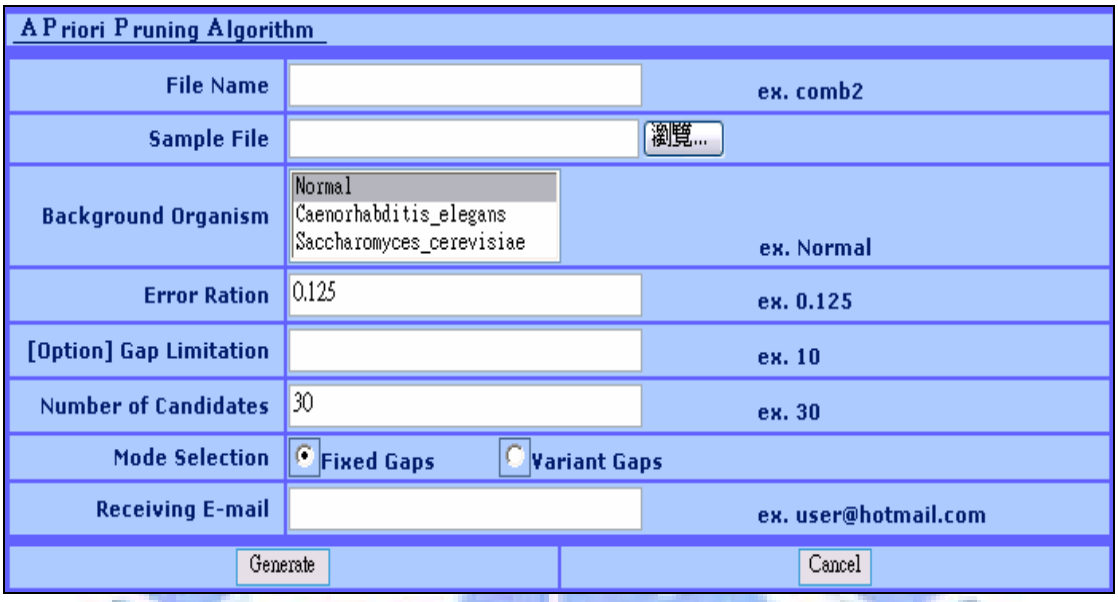


圖 3.14　APPA 系統之使用者介面

圖 3.14 中呈現 **APPA** 系統之使用者介面，使用者僅需輸入隱含重要基因片段的基因序列群 ***S* (sample file)**、選定特定對應基因序列群 ***B* (background organism)**、以及設定所允許的容錯值 **ε (error ration)** 、欲探測之重要基因片段爲固定間距 **(fixed gaps)** 或是可變間距 **(variant gaps)** 和使用者信箱等資訊，**APPA** 系統會自動將探測重要基因片段的結果寄到使用者的信箱中。

而系統的實驗結果與分析將會在下一章節中有詳細的描述。

第四章　實驗結果與分析

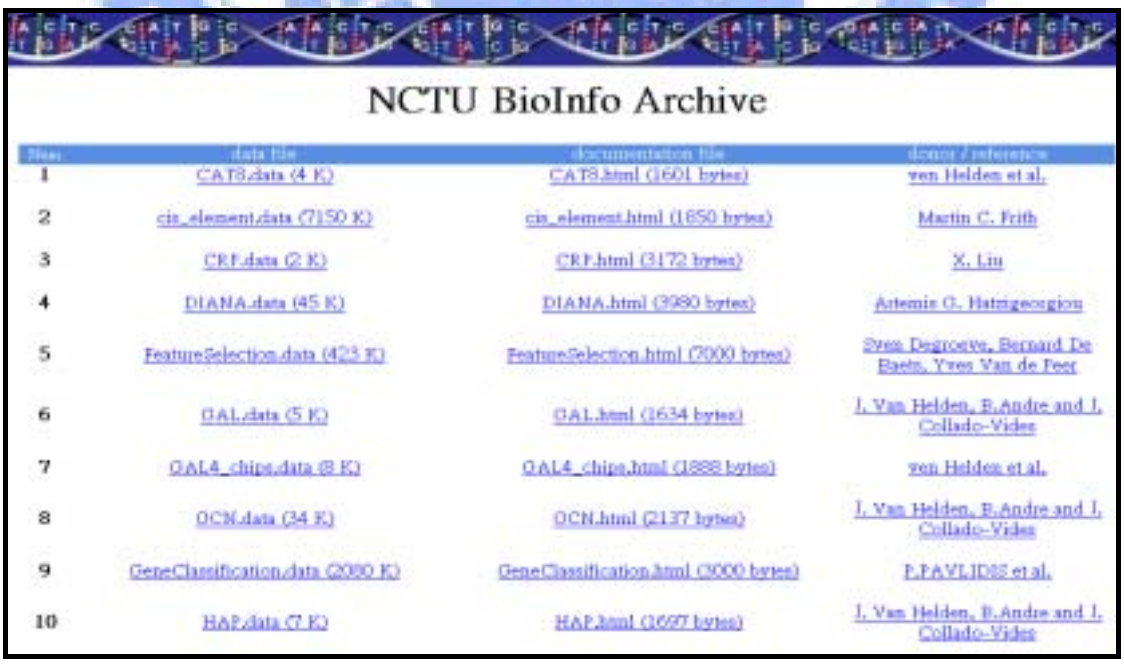
在詳盡介紹完 **APPA** 的運作流程之後，針對 **APPA** 在探測重要基因片段的實作以及效能作驗證動作是必然的。因此在此章節的一開始會詳述實驗資料的來源，緊接著利用人工基因序列群測試核心模組的正確性。最後經由實驗結果所得的數據，進而對 **APPA** 的效能作全面性的評估。

4.1 資料來源

4.1.1 隱藏有重要基因片段的生物基因序列群 (biological data)

由於現今提及藏有多元重要基因片段的生物基因序列群的文獻仍在少數，因此我們主要選取數個由 van Helden 等學者在其研究中所搜集匯整而得且已證實被 **Zn 群組因子 (zinc cluster factor)** 所調控的生物基因序列群作爲測試 **APPA** 的生物基因序列群。。（**圖 4.2**）

而此些生物基因序列的主要由 **NCTU BioInfo Archive** (http://bioinfo.cis.nctu.edu.tw) 網站上所取得。（**圖 4.1**）



NCTU BioInfo Archive

Num	data file	documentation file	donor / reference
1	CAT8.data (4 K)	CAT8.html (1601 bytes)	van Helden et al.
2	cis_element.data (7150 K)	cis_element.html (1650 bytes)	Martin C. Frith
3	CRF.data (2 K)	CRF.html (3172 bytes)	X. Liu
4	DIANA.data (45 K)	DIANA.html (3980 bytes)	Artemis G. Hatzigeorgiou
5	FeatureSelection.data (423 K)	FeatureSelection.html (7000 bytes)	Yves Degroeve, Bernard De Baets, Yves Van de Peer
6	GAL.data (5 K)	GAL.html (1634 bytes)	J. Van Helden, B.Andre and J. Collado-Vides
7	GAL4_chips.data (8 K)	GAL4_chips.html (1888 bytes)	van Helden et al.
8	OCN.data (34 K)	OCN.html (2137 bytes)	J. Van Helden, B.Andre and J. Collado-Vides
9	GeneClassification.data (2080 K)	GeneClassification.html (3000 bytes)	P.PAVLIDIS et al.
10	HAP.data (7 K)	HAP.html (2697 bytes)	J. Van Helden, B.Andre and J. Collado-Vides

圖 4.1　NCTU BioInfo Archive (http://bioinfo.cis.nctu.edu.tw)

family	genes	known motif
GAL4	GAL1 GAL2 GAL7 GAL80 MEL1 GCY1	**CGGRn$_2$RCYnYnCnCCG**
CAT8	ACR1 ICL1 MLS1 PCK1 FBP1	**CGGn$_6$GGA**
LEU3	GDH1 ILV1 LEU1 LEU2 LEU4	**RCCGGn$_2$CCGGY**
LYS	LYS1 LYS2 LYS4 LYS9 LYS20 LYS21	**WWWTCCRnYGGAWWW**
PDR	YOR1 PDR11 PDR10 GAS1 STE6 SNQ2 PDR5	**TYTCCGCGGARY TCCGGGA TCCGTGGA**
PPR1	URA1 URA3 URA4	**WYCGGn$_2$WWYKCCGAW**
UGA3	UGA1 UGA4 YBR006W	**AAARCCGCSGGCGGSAWT**
UME6	BAR1 CAR1 CAR2 DMC2 DMC1 GAL1 HOP1 HSF1 ILV2 IME1 IME2 INO1 MEI4 MER1 REC102 REC114 RED1 RME1 SPO11 SPO13 SPO16 TOP1 ZIP1	**TAGCCGCCGA**

圖 4.2　由 *Zn* 群組因子所調控的基因序列群 (van Helden, 2000)

4.1.2 隱藏有重要基因片段的人工基因序列群 (artificial data)

我們除了選取上述已知含有重要基因片段的生物基因序列群作爲測試 **APPA** 的資料之外，我們將會依據與 Pevner 和 Sze 所定義 **Challenge Problem** 類似的衍生資料法產生隱含有可變間距之多元重要基因片段的人工基因序列群，藉其更進一步來測試 **APPA** 的效能以及全面性。

假設我們欲產生含有一多元重要基因片段和數個單元重要基因片段的人工基因序列群 ***S***，我們需設定此人工基因序列群的相關資訊，如：基因序列群中的序列數量爲 ***seq_num***、基因序列的統一長度爲 ***seq_len***、共有 ***c*** 個單元重要基因片段、其中 ***p*** 個單元重要基因片段共同組成一多元重要基因片段 ***M***、而各個單元重要基因片段的長度限制爲 ***len_limit***、間距長度限制爲 ***inter_limit***、以及突變機率爲 ***mut_rate***。(**圖 4.3**)

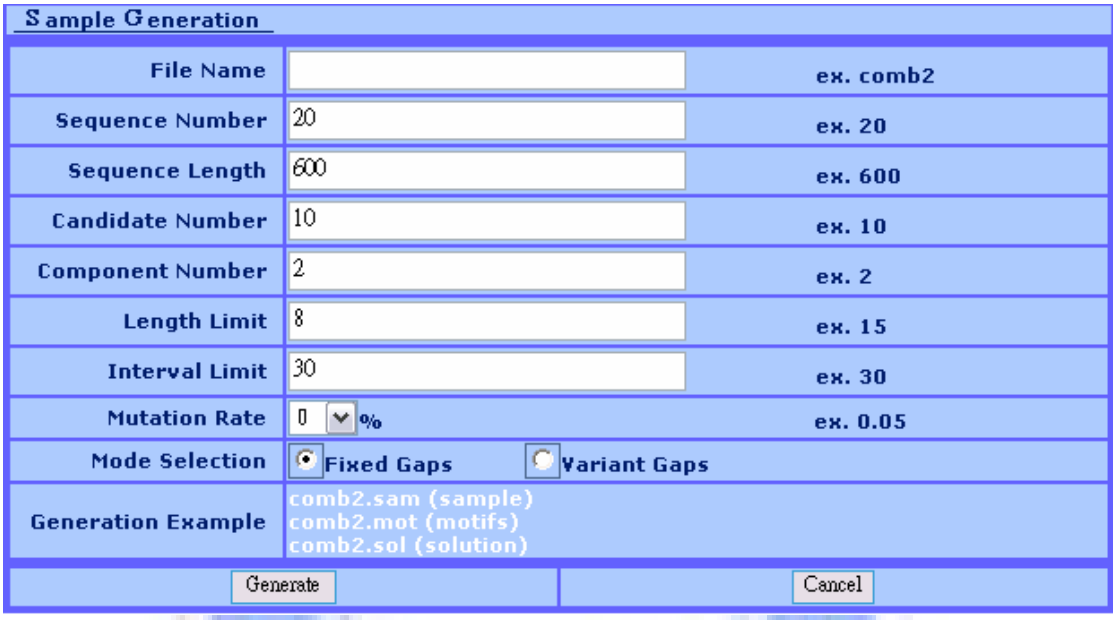


圖 4.3　隱含可變間距多元重要基因片段之基因序列群產生器

在使用者設定好所有參數後，此基因序列群產生器所運作的第一步驟就是亂數產生 c 個長度介於 3 至 ***len_limit*** 的單元重要基因片段。緊接著選取前 ***p*** 個單元重要基因片段和 ***(p-1)*** 個由亂數產生且長度介於 0 至 ***inter_limit*** 的間距區間進而組成多元重要基因片段 ***M***。最後亂數產生 ***n*** 條基因序列，將多元重要基因片段 ***M*** 和剩餘 ***(c-p)*** 個單元重要基因片段安插入每一序列中，然後再針對在每一序列上每一基因片段的每一位置作是否突變的處置。（**圖 4.4**）

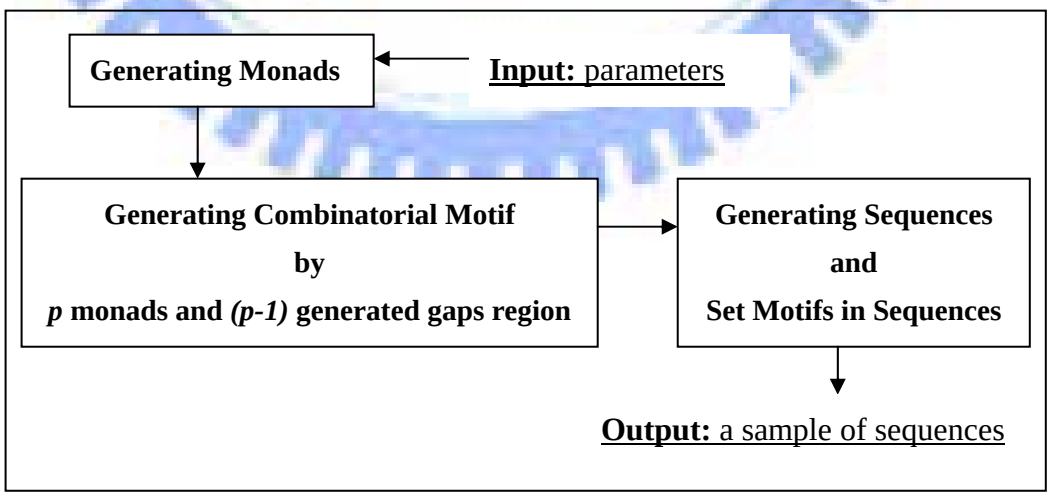


圖 4.4　基因序列群產生器的概略流程

4.2 核心模組測試

首先，我們會藉由人工基因序列群的實驗驗證 **APPA** 演算法中核心模組流程的正確性，因爲我們希望在單元重要基因片段探測結果皆爲正確的情況下，**APPA** 中的核心模組能將出正確的多元重要基因片段推導出來。

(a)

No.	input monad (m)	merit(m)	No.	input monad(m)	merit(m)
1	TTT	0.015	6	GGT	0.011
2	CCCTATA	0.686	7	CGTTGGCC	2.521
3	ACTGCT	0.155	8	CTAT	0.026
4	GTAA	0.015	9	CATC	0.018
5	ACTCG	0.045	10	AGTTCAT	0.683

(b)

know motif (mutation rate = 0)	APPA (ε = 0.125 , background= *Normal*)			
	Derived Motif	***MERIT***	***SG***	***SIG***
TTT-(1)-CCCTATA-(10)-ACTGCT	TTT-(1)-CCCTATA-(10)-ACTGCT [1]	16.00	15.24	31.24
GTAA	CCCTATA-(10)-ACTGCT	13.00	12.38	25.38
ACTCG	TTTCCCTATA-(10)-ACTGCT	8.00	13.72	21.72
GGT	TTT-(2)-CCCTATA-(10)-ACTGCT	7.11	13.34	20.45
CGTTGGCC	TTT-(1)-CCCTATA	7.92	6.86	14.18
CTAT	TTTCCCTATA	3.89	4.00	7.89
CATC	TTT-(2)-CCCTATA	3.53	3.85	7.38
AGTTCAT	CGTTGGCC	2.52	2.08	4.60
	CCCTATA	0.69	0.63	1.32
	AGTTCAT	0.68	0.62	1.30

[$rank_m$]－表重要基因片段 *m* 在探測結果中的排名爲 $rank_m$。

圖 4.5　核心模組之驗證 (MERIT= ***merit(m)*|m|***, SG= ***SG_score(m) *|m|***)
(a) 輸入之正確單元重要基因片段 (b) APPA 核心模組所產生的結果

在 **圖 4.5** 中，我們將藉由輸入正確單元重要基因片段，觀察 **APPA** 核心模組是否能精確地探測到正確的多元重要基因片段，進而評比 **APPA** 核心模組的精確度。

在驗證 **APPA** 核心模組的過程中，我們採用**平均排名倒數法則 (mean reciprocal rank; MRR)** 作爲探測多元重要基因片段之相似精確度的評比。由於在此過程中，輸入的單元重要基因片段皆爲正確單元重要基因片段，因此影響固定間距之多元重要基因片段的評估因素即爲此多元重要基因片段中子片段與正確多元重要基因片段中子片段相符的個數－

《固定間距之重要基因片段的相似評估函數》

$$MRR = \frac{1}{n}\sum_{i=1}^{n} RR_i \text{, where } RR_i = \begin{cases} \frac{1}{rank_i} \times \frac{c_{r\cap d}}{\max(c_r, c_d)} & rank_i \le 10 \\ 0 & rank_i > 10 \end{cases}$$

而影響可變間距之多元重要基因片段的評估因素除了上述相符子片段的個數之外，還加入了可變間距精準度的考量－

《可變間距之重要基因片段的相似評估函數》

$$MRR = \frac{1}{n}\sum_{i=1}^{n} RR_i, \quad RR_i = \begin{cases} \frac{1}{rank_i} \times \frac{C_{r\cap d} + \sum_{j=1}^{(C_{r\cap d}-1)} \frac{\left|g_{r\cap d,j} \cap g_{d\cap r,j}\right|}{\left|g_{r\cap d,j} \cup g_{d\cap r,j}\right|}}{2\times\max(c_r, c_d)-1} & rank_i \le 10 \\ 0 & rank_i > 10 \end{cases}$$

此處的 n 表示實驗的次數，$rank_i$ 則表示第 i 次實驗中的正解多元重要基因片段在傳回數據中的排名，c_r 爲正確多元重要基因片段 r 的子片段個數，c_d 爲探測所得多元重要基因片段 d 的子片段個數，而 $c_{r\cap d}$ 則爲 r 和 d 所共有之子片段個數。$g_{r\cap d,j}$ 表示 r 上第 j 個和 d 共同擁有的可變區間，而 $g_{d\cap r,j}$ 則表示 d 上第 j 個和 r 共同擁有的可變區間。因此，$|g_{r\cap d,j} \cup g_{d\cap r,j}|$ 表示兩區間聯合後的長度，而 $|g_{r\cap d,j} \cap g_{d\cap r,j}|$ 則表示兩區間交集後的長度。

除了採用重要基因片段的相似程度作爲評估之外，我們另外還採用了前學者慣用的 **Mean Performance Coefficient (MPC)** 作爲評估考量。

$$MPC = \frac{1}{n}\sum_{i=1}^{n} PC_i \qquad PC_i = \begin{cases} \frac{1}{rank_i} \times \frac{|K \cap P|}{|K \cup P|} & rank_i \le 10 \\ 0 & rank_i > 10 \end{cases}$$

(a)

Exp. *i*	know motif (mutation rate = 0%)	APPA (ε = 0.125, background= *Normal*, threshold=*low*)					
		Derived Motif	***MERIT***	***SG***	***SIG***	RR_i	PC_i
1	GCGGT-(3)-GCCCCGAA	GCGGT-(3)-GCCCCGAA [1]	13.00	12.38	25.38	1	1
2	GCCTCTG-(27)-GCAGGCTT	GCCTCTG-(26)-TGCAGGCTT [1]	16.00	15.24	31.24	0.96	1
3	CGGAT-(14)-TTAGA	CGGAT-(14)-TTAGA [1]	8.93	7.70	16.63	1	1
4	GCATAG-(16)-GCAGCATC	GCATAG-(16)-GCAGCATC [1]	14.00	13.34	27.34	1	1
5	TCCT-(3)-AAGGTCT	TCCT-(3)-AAGGTCT [1]	11.00	10.48	21.48	1	1
6	ATTC-(29)-GTC	[x]	×	×	×	0	0
7	GTTAC-(2)-CGGTC	GTTAC-(2)-CGGTC [1]	8.75	7.41	16.16	1	1
8	TCTAAAG-(3)-CGTG	TCTAAAG-(3)-CGTG [1]	10.50	9.57	20.07	1	1
9	CAA-(29)-TTAGA	CAA-(29)-TTAGA [1]	5.44	4.83	10.27	1	1
10	TGCCGAA-(26)-GGTTCCAC	GTGCCGAA-(26)-GGTTCCAC [1]	16.00	15.24	31.24	0.96	1

[*x*] 爲第 *i* 次實驗中，正確多元重要基因片段並未出現在探測結果中。

(b) $MRR = \frac{8.92}{10} = 0.892$ (c) $MPC = \frac{9}{10} = 0.9$

圖 4.6 針對十個含有固定間距之雙元重要基因片段的人工基因序列群所得的實驗結果 (MERIT= *merit(m)*|m|*, SG= *SG_score(m) *|m|*)

(a)

Exp. *i*	know sites (mutation rate = 0%)	APPA (ε = 0.125 , background= *Normal*, threshold=*low*)					
		Derived Motif	***MER***	***SG***	***SIG***	RR_i	PC_i
1	ACTCGT-(10,10)-AAGTG	ACTCGT-(10)-AAGTG [1]	10.74	10.01	20.75	1	1
2	TCA-(7,8)-CCATGC	TCA-(7,8)-CCATGC [3]	3.59	2.77	6.36	0.33	1
3	CGGCGAT-(5,10)-CGCTCC	CGGCGAT-(5,10)-CGCTCC [1]	4.78	12.40	17.19	1	1
4	TTGATTA-(7,10)-GTGAAAG	TTGATTA-(7,10)-GTGAAAG [1]	6.14	14.00	20.14	1	1
5	CCTAGCGT-(7,9)-AGTACCTT	CCTAGCGT-(7,9)-AGTACCTT [1]	8.43	16.00	24.43	1	1
6	CGGT-(3,7)-GTTCT	[x]	×	×	×	0	0
7	GCAT-(2,4)-TTGCATC	GCAT-(2,4)-TTGCATC [1]	5.41	8.89	14.30	1	1
8	GAC-(6,9)-GGACCCG	GAC-(6,9)-GGACCCG [1]	3.73	4.89	8.62	1	1
9	CCGC-(1,10)-GTTACATC	CCGC-(4,10)-GTTACATC [1]	4.15	8.64	12.79	0.9	0.85
10	AGTCACTA-(8,10)-AAACACC	AGTCACTA-(8,10)-AAACACC [1]	7.92	15.00	22.92	1	1

(b) $MRR = \frac{8.23}{10} = 0.823$ (c) $MPC = \frac{8.85}{10} = 0.885$

圖 4.7 針對十個含有可變間距之雙元重要基因片段的人工基因序列群所得的實驗結果 (MER= *merit(m)*|m|*, SG= *SG_score(m) *|m|*)

圖 4.6 和 圖 4.7 中分別記錄 APPA 核心模組針對十筆含有固定間距和可變間距之雙元重要基因片段的人工基因序列群測試所得的實驗數據。

(a)

Exp. *i*	know sites (mutation rate = 0%)	APPA (ε = 0.125, background= *Normal*, threshold=*low*)					
		Derived Combinatorial Motif	***MER***	***SG***	***SIG***	RR_i	PC_i
1	TTT-(1)-CCCTATA-(10)-ACTGCT	TTT-(1)-CCCTATA-(10)-ACTGCT	16.00	15.24	31.24	1	1
2	CTTAC-(2)-CATTGCA-(8)-GTCAGA	CTTAC-(2)-CATTGCA-(8)-GTCAGA	18.00	17.15	35.15	1	1
3	CAG-(3)-TTTTGCA-(10)-AAGTACAT	CAG-(3)-TTTTGCAG-(8)-AAAGTACAT	20.00	19.05	39.05	0.92	1
4	CTCATC-(8)-AACCG-(4)-AGCGA	CTCATC-(8)-AACCG-(4)-AGCGA	16.00	15.12	31.12	1	1
5	CAGGAAC-(10)-TCCGT-(10)-ATT	CAGGAAC-(10)-TCCGT-(10)-ATT	15.00	14.29	29.29	1	1
6	CTAGAC-(1)-GACA-(3)-TGT	CTAGAC-(1)-GACA	9.13	8.00	17.13	0.67	1
7	TTCACACG-(1)-GACCCTAG-(3)-GCCCG	TTCACACG-(1)-GACCCTAG-(3)-GCCCG	21.00	20.00	41.00	1	1
8	CGAGGT-(8)-ATCAGT-(9)-ACAGTGG	CGAGGT-(8)-ATCAGT-(9)-ACAGTGG	19.00	18.10	37.10	1	1
9	ATAT-(3)-GTTC-(2)-GAGCCC	ATAT-(3)-GTTC-(2)-GAGCCC	14.00	13.34	27.34	1	1
10	ACTTCCCG-(10)-CCTC-(3)-CCGATTT	ACTTCCCG-(10)-CCTC-(3)-CCGATTT	19.00	18.10	37.10	1	1

[$rank_i$] 爲第 i 次實驗中，正確多元重要基因片段在探測結果中的排名爲 $rank_i$。

(b) $MRR = \frac{9.59}{10} = 0.959$

(c) $MPC = \frac{10}{10} = 1.00$

圖 4.8 APPA 核心模組針對十個含有固定間距之三元重要基因片段的人工基因序列群所得的實驗結果 (MERIT= *merit(m) | *m* |, SG= *SG_score(m)* * | *m* |)**

(component number, mutation rate)	*(2, 0)*	*(3, 0)*	*(4, 0)*	*(5, 0)*	*(6, 0)*
MRR	0.9	0.959	0.886	0.8	**0.696**
MPC	0.9	1.00	1.00	1.00	**1.00**

圖 4.9 APPA 核心模組探測子片段個數不同之多元重要基因片段的精確度。

(ε = 0.125, background= *Normal*)

當我們訂定好精確度的評比分式後，我們隨即對由不同個數單元重要基因片段所建構的多元重要基因片段進行探測實驗，因爲我們想要瞭解多元重要基因片段之子片段個數是否會對於 **APPA** 核心模組的精確度造成影響。而從 **圖 4.9** 所呈現的數據中，我們可以清楚地觀察到－當多元重要基因片段由愈多單元重要基因片段所構成時，**APPA** 對其探測的精確度也就愈低。這樣結果是必然的，因爲在 **APPA** 的核心概念中，***k*-**元重要基因片段需由 ***(k-1)*-**元重要基因片段和特定雙元重要基因片段共同建構而成。因而倘若 ***(k-1)*-**元重要基因片段或此特定雙元重要基因片段並未被探測出，則此 ***k*-**元重要基因片段的資訊也會遺失。

(a)

Exp. *i*	know sites (mutation rate = 5%)	APPA (ε = 0.125, background=*Normal*)					
		Derived Motif	***MERIT***	***SG***	***SIG***	RR_i	PC_i
1	TAC-(3)-TAGGTT	TAC-(3)-TAGGTT	4.50	3.06	7.56	1	0.80
2	GAA-(9)-TGGC		×	×	×	0	0
3	TGCGTAAC-(4)-CTCGC	TGCGTAAC-(4)-CTCGC	10.40	12.07	22.47	1	0.65
4	TTAGG-(3)-GAGTTA	TTAGG-(3)-GAGTTA	7.12	8.46	15.58	1	0.50
5	TCG-(8)-TGGCGAC	TCG-(6)-CGTGGCGAC	6.67	6.19	12.86	1	0.65
6	TCAC-(9)-TTAATATT	TCAC-(9)-TTAATATT	11.40	11.37	22.77	1	0.90
7	GCAA-(6)-CATCGCT	GCAA-(6)-CATCGCT	9.51	10.32	19.83	1	0.75
8	ACAACTTT-(8)-CGGAAAC	ACAACTTT-(8)-CGGAAAC	13.42	14.12	27.54	1	0.80
9	AATATAA-(7)-ATAAG	AATATAAG-(6)-ATAAG	8.13	10.63	18.76	0.94	0.45
10	GATCTCA-(8)-TGCGT	GATCTCA-(8)-TGCGT	8.25	10.91	19.16	1	0.50

[$rank_i$] 爲第 *i* 次實驗中，正確多元重要基因片段在探測結果中的排名爲 $rank_i$。

[×] 即表示在第 *i* 次實驗中，並未搜尋到此正確多元重要基因片段。

(b) $MRR = \frac{8.94}{10} = 0.894$

(c) $MPC = \frac{6}{10} = 0.6$

圖 4.10　APPA 核心模組針對十個含有雙元重要基因片段且突變率為 5% 的人工基因序列群所得實驗結果

(MERIT= *merit(m)*|m|*, SG= *SG_score(m) *|m|*)

(component number, mutation rate)	*(2, 0%)*	*(2, 5%)*	*(2, 10%)*	*(2, 15%)*	*(2, 20%)*	*(2, 25%)*
MRR	0.9	0.894	0.8	0.66	0.577	0.521
MPC	0.9	0.6	0.455	0.433	0.387	0.326

圖 4.11　APPA 核心模組探測受不同突變率影響之重要基因片段的精確度。

(ε = 0.125, background= *Normal*)

而由 **圖 4.11** 的表格中則可觀察出－當多元重要基因片段所受影響的突變率愈高，**APPA** 對其探測的精確度也就愈低。值得注意的是，在容錯度 ε 設定爲 0.125 的情況下，當突變率提高至 15% 時，**APPA** 仍能有 73% 的精確度。

4.3 實驗結果

在實驗結果的分析上，我們依舊採用**平均排名倒數法則 (mean reciprocal rank; MRR)** 和 **MPC** 作爲精確度的評比。不過由於在探測單元重要基因片段的步驟上，並不能保證被探測出的單元重要基因片段皆爲正確，因此在此要須加入多元重要基因片段中各子片段和各區間的精確數值作爲相關的評比標準－

$$MRR = \frac{1}{n}\sum_{i=1}^{n} RR_i$$

$$RR_i = \begin{cases} \dfrac{1}{rank_i} \times \dfrac{\sum_{j=1}^{C_{r\cap d}} \dfrac{\left|m_{r\cap d,j} \cap m_{d\cap r,j}\right|}{\left|m_{r\cap d,j} \cup m_{d\cap r,j}\right|} + \sum_{j=1}^{(C_{r\cap d}-1)} \dfrac{\left|g_{r\cap d,j} \cap g_{d\cap r,j}\right|}{\left|g_{r\cap d,j} \cup g_{d\cap r,j}\right|}}{2\times \max(c_r, c_d) - 1} & rank_i \le 10 \\ 0 & rank_i > 10 \end{cases}$$

此處的 n 表示實驗的次數，$rank_i$ 則表示第 i 次實驗中的正解多元重要基因片段在傳回數據中排名，c_r 爲正確多元重要基因片段 r 的子片段個數，c_d 爲探測所得多元重要基因片段 d 的子片段個數，而 $c_{r\cap d}$ 則爲 r 和 d 所共有之子片段個數。$m_{r\cap d,\,j}$ 表示 r 上第 j 個和 d 共同擁有的子片段，而 $m_{d\cap r,\,j}$ 則表示 d 上第 j 個和 r 共同擁有的子片段。因此，$|m_{r\cap d,\,j} \cup m_{d\cap r,\,j}|$ 表示兩子片段聯集後的長度，而 $|m_{r\cap d,\,j} \cap m_{d\cap r,\,j}|$ 則表示兩子片段交集後的長度。同理，$g_{r\cap d,\,j}$ 表示 r 上第 j 個和 d 共同擁有的可變區間，而 $g_{d\cap r,\,j}$ 則表示 d 上第 j 個和 r 共同擁有的可變區間，$|g_{r\cap d,\,j} \cup g_{d\cap r,\,j}|$ 表示聯集後的長度，而 $|g_{r\cap d,\,j} \cap g_{d\cap r,\,j}|$ 則表示兩區間交集後的長度。

舉例而言，假設我們欲探測雙元重要基因片段 r= GCAA-(4)-GCCTCCTT，而 APPA 的探測結果排名第一的重要基因片段爲 d=GGC-(6)-GCCTCCTT，則此次實驗的精確度爲－

$$RR = 1 \times \frac{\left(\frac{2}{5} + 1 + 1\right)}{3} = 0.8$$

4.3.1 探測單元重要基因片段

由於 **APPA** 在探測單元重要基因片段的步驟上使用 **MERMAID** 作爲探測方法，因此在探測單元重要基因片段的實驗上，我們僅針對 **PDR**、**UGA3** 以及 **UME6** 此三個生物基因序列群作探討。藉此三個生物基因序列群的探測結果驗證－雖然 **MERMAID** 僅幫 **APPA** 探測出長度 3 至 8 的單元重要基因片段，可是經由 **APPA** 核心模組的運作，**APPA** 仍可正確探測出長度大於 8 的單元重要基因片段。

【parameter】 ε = 0.125, background organism=*S. cerevisiae*, gap limit=10

Family	Known motif	Dyad Analysis	APPA *Derived Motif*		*MERIT*	*SG*	*SIG*
PDR°	TYTCCGCGGARY	TCCGTGGAA [1]	TTTCCGCGGA	[1]	13.29	1.82	15.11
	TCCGCGGA	TCCGCGGA [2]	TCCGCGGAAA	[2]	8.37	1.54	9.91
	TCCGTGGA		TTCCGTGGA	[5]	6.90	0.69	7.59
UGA3°	AAARCCGCSGGCGGSAWT	CCTn{14}CCG [x]	AAGCCGCGGGCGGGAY	[2]	12.88	5.34	18.22
UME6	TAGCCGCCGA	TAGCCGCCGA [1]	AGCCGCCG	[4]	6.12	1.02	7.14
			GCCGCCGA	[6]	5.67	0.73	6.40
			TAGCCGCC	[7]	5.80	0.51	6.31

圖 4.12　PDR、UGA3 以及 UME6 此三個生物基因序列群所得實驗結果

(MERIT= $merit(m)*|m|$**, SG=** $SG\_score(m)*|m|$**,** $[rank_i]$ **)**

而由 **圖 4.12** 所記錄的實驗結果中，我們可以察覺在使用者僅輸入容錯度和對應基因序列群的情況下，**APPA** 皆能在此三個生物基因序列群中探測到正確且精確的結果，其中其中尤以探測 **UGA3** 所隱含的單元重要基因片段最爲精準。

4.3.2 探測多元重要基因片段

在驗證探測固定間距之多元重要基因片段的實驗中，我們將會分別對依據上單元所述方法所產生的十筆人工基因序列群以及 **GAL4**、**CAT8**、**LEU3** 和 **LYS** 此四個已知重要序列片段爲何的生物基因序列群作驗證實驗。（**圖 4.13**、**圖 4.14**）進而估量 **APPA** 在探測固定間距之多元重要基因片段上的效能。

而由 **圖 4.13** 中，針對第三筆人工基因序列群所作的探測結果作分析，我們發現－由於 **MERMAID** 所探測到的單元重要基因片段中並未包括 **ATGTA** 和 **GTA** 此兩個單元重要基因片段，因此 **APPA** 無法將正確的雙元重要基因片段探測出。

(a)

Exp. *i*	know motif (mutation rate = 0%)	APPA (ε = 0.125, background= *Normal*)					
		Derived Motif	***MERIT***	***SG***	***SIG***	***RR$_i$***	***PC$_i$***
1	GCAA-(4)-GCCTCCTT	GGC-(6)-GCCTCCTT [1]	11.00	3.67	14.67	0.80	0.30
2	TTACGGC-(8)-GAGG	WTTACGGC-(10)-GGC [1]	9.69	3.50	13.19	0.76	1.00
3	ATGTA-(3)-GTA	[×]	×	×	×	0.00	0.00
4	CATAG-(5)-CGC	TCCATAGT-(4)-CGC [1]	9.38	4.05	13.88	0.90	
5	CATCGGT-(10)-CTATTCA	CCATCGGT-(9)-GCTATTCA [1]	14.13	16.00	30.13	0.96	
6	TGCGCCCT-(8)-CCGACTC	TGCGCCCT-(8)-CCGACTC [1]	15.00	15.00	30.00	1.00	1.00
7	CAATTC-(9)-ATTAACGG	CAATTC-(9)-ATTAACGG [1]	13.40	14.00	27.40	1.00	1.00
8	GCGGGTAC-(6)-TGGATT	GCGGGTAC-(6)-TGGATT [1]	12.89	14.00	26.89	1.00	1.00
9	TTCAGAGT-(6)-CGGT	TTCAGAGT-(6)-CGG [1]	10.67	9.63	20.30	0.92	1.00
10	CTGAATTT-(10)-CACCAT	GCCTGAATTT-(10)-CACCAT [1]	22.52	7.43	29.95	0.93	

[$rank_i$] 爲第 i 次實驗中，正確多元重要基因片段在探測結果中的排名爲 $rank_i$。

(b) $MRR = \frac{8.27}{10} = 0.827$

圖 4.13　APPA 針對十個含有雙元重要基因片段的人工基因序列群所得的實驗結果（MERIT= ***merit(m)*\|m\|***, SG= ***SG_score(m) *\|m\|***）

(a) 針對各個生物基因序列群，Dyad Analysis 和 MERMAID 的探測結果

Family	Known motif	Dyad Analysis		MERMAID	
GAL4	CGGRn2RCYnYnCnCCG	CGGn{10}TCC	[1]	CGG-X(11)-CCG	[1]
CAT8	CGGn6GGA	CGGn{4}ATGGA	[1]	CGG-X(6)-GGA	[1]
LEU3	RCCGGn2CCGGY	CCGn{3}CCG	[1]	GCCGG-X(2)-CCGGC	[1]
LYS	WWWTCCRnYGGAWWW	AAATTCCG	[1]	TTCCR-X(1)-YGGAA	[10]
PPR1	WYCGGn2WWYKCCGAW	CGGn{6}CCG	[1]	TTCGG-X(2)-AACCCCGAG	[4]

(b) 針對各個生物基因序列群，APPA 所得的探測結果

Family	Known motif	APPA (ε = 0.125, background organism=*S. cerevisiae*)				
		Derived Motif		MERIT	SG	SIG
GAL4°	CGGRn2RCYnYnCnCCG	TCGGRGCA-(9)----GAA	[1]	8.54	7.34	15.88
		TCGGRGC-(7)---CCG	[2]	8.55	5.50	14.05
		TCGGRGCA-(6)-TCC	[4]	8.51	5.50	14.01
CAT8	CGGn6GGA	CGTCCGGA-(5)-GGA	[1]	8.45	5.50	13.95
LEU3	RCCGGn2CCGGY	GCGCC-(1)-GAACCGGC	[1]	14.72	6.50	21.22
		GCCGGAACCGGC	[2]	10.90	6.00	16.90
LYS	WWWTCCRnYGGAWWW	TTT-(1)-CCAGCGGAA	[L]	16.91	5.20	22.11
		AATTCCG-(2)-GGAA	[L]	10.24	4.72	14.96
PPR1	WYCGGn2WWYKCCGAW	ATCTTCGGCATTCGCCGAAAAT	[1]	31.71	11.00	42.71

[$rank_i$] 爲第 i 次實驗中，正確多元重要基因片段在探測結果中的排名爲 $rank_i$。

[L] 則表示第 i 次實驗中，正確多元重要基因片段在探測結果中的排名遠在十名之外。

圖 4.14　各演算法針對各生物基因序列群所得的探測結果

(MERIT= $merit(m)*|m|$, SG= $SG\_score(m)*|m|$)

(a)

Exp. *i*	know sites (mutation rate = 0%)	APPA (ε = 0.125, background= *Normal*)				
		Derived Combinatorial Motif	***MER***	***SG***	***SIG***	***RR$_i$***
1	GTA-(5)-AGTACTT-(6)-TAT	AGTACTT [1]	6.97	0.34	7.31	0.60
2	ACCTCGT-(7)-TAT-(1)-GATGTATC	GATGTATC [1]	7.90	1.31	8.31	0.60
3	CATGGC-(1)-GCAAAT-(7)-GCTCA	CATGGC-(1)-GCAAAT-(6)-CGC [1]	20.59	6.29	26.88	0.88
4	TTCTTCAG-(6)-AGGTTA-(4)-ACG	TTCTTCA-(7)-AGGTTA-(5)-CGC [1]	21.58	12.00	33.58	0.88
5	GTT-(8)-ACGTC-(3)-ATGTAACT	ACGTC-(3)-ATGTAACT [1]	11.26	13.00	24.26	0.80
6	TGTGTG-(9)-AAGGCGAC-(8)-CTTTGCGG	TGTGTG-(9)-AAGGCGA-(10)-TTTGCGG [1]	26.23	39.00	65.23	0.95
7	TTCTCTGT-(1)-GGT-(6)-TGGGCG	TTCTCTGT-(10)-TGGGCG [1]	13.92	14.00	27.92	0.80
		TTCTCTGTCGG-(7)-TGGGCG [2]	27.82	8.00	27.82	
8	GGCT-(1)-GAT-(1)-TCC	GGC-(2)-GATTTCC [1]	8.59	2.96	11.55	0.95
9	CCACTCCG-(1)-GTTC-(6)-ACGC	CCACTCCGGG [1]	9.94	2.69	12.63	0.65
10	ACG-(6)-GTCCG-(3)-TGAC	GTCCG-(4)-GACG [1]	8.28	1.29	9.57	0.72
		CGC-(5)-GTCCGC [2]	7.99	1.10	9.09	

[$rank_i$] 爲第 i 次實驗中，正確多元重要基因片段在探測結果中的排名爲 $rank_i$。

(b) $MRR = \frac{7.83}{10} = 0.783$

圖 4.15　APPA 針對十個含有固定間距之三元重要基因片段的人工基因序列群所得的實驗結果（MERIT= *merit(m)*|m|*, SG= *SG_score(m) *|m|*）

4.4 結果分析

由上述的實驗結果可驗證－在探測重要基因片段的問題上，**APPA** 不但符合使用者需求且具有彈性以及完整性。因爲 **APPA** 不僅僅能探測到多元重要基因片段，更能辨視出長度較長的單元重要基因片段。除此之外，由上述實驗中亦不難看出 **APPA** 所辨視出的重要基因片段亦較其他演算法來得精確。

圖 4.16 爲我們將 **APPA** 和 **Dyad Analysis** 針對受不同突變率影響的人工基因序列群作探測實驗後所得的精確度數值。此數據亦驗證 **APPA** 所探測的重要基因片段遠比 **Dyad Analysis** 來得精確。

Method＼mut_rate	*0%*	*5%*	*10%*
APPA (MRR, MPC)	(0.842, 0.615)	(0.781, 0.529)	(0.718, 0.430)
Dyad Analysis (MRR, MPC)	(0.668, 0.643)	(0.620, 0.424)	(0.472, 0.310)

圖 4.16　APPA 和 Dyad Analysis 之精確度數值比較

Exp. *i*	know sites (mutation rate = 0%)	Dyad Analysis spacing range=(0, 10)	APPA ε = 0.125
1	GCAA-(4)-GCCTCCTT	GCAAGGAGGCCTCCTTGC [2] 0.47/0.00	GGC-(6)-GCCTCCTT [1] 0.80/0.30
2	TTACGGC-(8)-GAGG	ACGGCn{8}GAGG [2] 0.45/0.50	WTTACGGC-(10)-GGC [1] 0.76/0.30
3	ATGTA-(3)-GTA	ATGTAn{3}GTA [1] 1.00/1.00	[x] 0.00/0.00
4	CATAG-(5)-CGC	CATAGn{5}CGC [1] 1.00/1.00	TNCATAGT-(3)-GCGC [1] 0.93/0.30
5	CATCGGT-(10)-CTATTCA	GGTn{10}CTA [2] 0.31/0.46	CATCGGTR-(8)-KCTATTCA [1] 1.00/1.00
6	TGCGCCCT-(8)-CCGACTC	GCCCTn{8}CCGAC [1] 0.78/1.00	TGCGCCCT-(8)-CCGACTC [1] 1.00/1.00
7	CAATTC-(9)-ATTAACGG	ATTCn{9}ATTA [2] 0.36/0.50	CAATTC-(9)-ATTAACGG [1] 1.00/1.00
8	GCGGGTAC-(6)-TGGATT	GCGGGTACn{5}CTGGATT [1] 1.00/0.50	GCGGGTAC-(6)-TGGATT [1] 1.00/1.00
9	TTCAGAGT-(6)-CGGT	TTCAGAGTn{6}CGGT [1] 1.00/1.00	TTCAGAGT-(6)-CGGT [1] 1.00/1.00
10	CTGAATTT-(10)-CACCAT	TTTn{10}CAC [2] 0.31/0.47	GCCTGAATTT-(10)-CACCAT [1] 0.93/0.25

圖 4.17　APPA 和 Dyad Analysis 針對十筆含有固定間距之雙元重要基因片段的人工基因序列群的測試結果

Exp. *i*	know sites (mutation rate = 10%)	Dyad-Analysis spacing range=(0, 10)	APPA result ε = 0.125
1	CCACGC-(8)-CGGGTGG	[x] 0.00/0.00	CCACGC-(8)-CGGG-(1)-GGG [1] 0.92/0.25
2	TTGTG-(9)-AACCAT	[x] 0.00/0.00	[x] 0.00/0.00
3	AGAACCT-(7)-TGC	ACCn{8}TGC [1] 0.81/0.61	AGAACCT [1] 0.67/0.50
4	CCTAAC-(2)-GTG	[x] 0.00/0.00	MCCTAACY [1] 0.62/0.81
5	GCTGGC-(1)-GAGC	CTGGCn{1}GAG [1] 0.86/0.65	GCTGGCNG-(1)-GCG [1] 0.87/0.30
6	CCCGATA-(6)-TTCCAGGA	ATAn{6}TTC [2] 0.32/0.19	CCGATA-(7)-TCCAGGA [1] 0.95/0.25
7	CGTT-(10)-CAGCTTCC	AGCnTCC [1] 0.58/0.46	CAGCTTCC [1] 0.67/0.86
8	CTGCTA-(3)-CAATT	CTGCTAn{3}CAATT [1] 1.00/0.45	CTGCTAT [1] 0.67/0.29
9	TGTAGCA-(1)-CTTT	TGTn{5}CTT [1] 0.73/0.41	TAGCARCT [1] 0.81/0.64
10	CAAAT-(1)-ATCTACGG	CAAATn{1}ATCT [2] 0.42/0.33	CAAATAATCTACGG [1] 1.00/0.40

圖 4.18　APPA 和 Dyad Analysis 針對十筆含有固定間距之雙元重要基因片段的人工基因序列群的測試結果，其中突變率為 10%。

第五章　結論與未來研究方向

在本章將會藉由討論分析探討的論文中提出方法的優缺點和特性，對此論文作個總結。並且在之後的單元中，呈述可以改善的地方，以及未來可能的研究方向。

5.1 結論與討論

我們提出的 **APPA** 演算法意圖在最符合使用者使用需求的前提下，盡可能探測出所有正確的重要基因片段。這些重要基因片段可能爲單元重要基因片段、固定間距之多元重要基因片段或者是可變間距之多元重要基因片段。

總結 **APPA** 主要與其他方法不同點之處在於：

1. 使用者僅需輸入可容許的誤差範圍 $\varepsilon < 1$，而不必明定欲作搜尋之重要基因片段資訊。
2. 運用對應基因序列群的設定進而探測出特徵不甚明顯的重要基因片段以及過濾掉不甚重要的基因片段。
3. 運用單元重要基因片段與給定基因序列群的相關資訊，快速建構起多元重要基因片段的資訊。
4. 解決過往探測重要基因片段演算法的不可擴充性質，可探測由 $p>2$ 個子片段所組成的多元重要基因片段。

然而，**APPA** 也有其受限之處。由於 **APPA** 運用 **MERMAID** 之擴充方法探測所有可能長度由 3 至 8 的單元重要基因片段，倘若構成多元重要基因片段的某一子片段並沒有被 **MERMAID** 所探測出，則此多元重要基因片段可能會因此子片段資訊的移失而無法經由 **APPA** 演算法探測而得。

此外，由於 **APPA** 採用擬真的 **pseudo_SG_score** 方法估量雙元以上重要基因片段的 **SG_score**，會使得由多個長度皆很短的子片段組成的多元重要基因片段無法被探測出，因爲 **pseudo_SG_score** 的值遠比原先 **real_SG_score** 來得小很多。

5.2 未來研究方向

雖然 **APPA** 能探測出長度超過 8 的單元重要基因片段，不過其中有個大前題存在，即爲此單元重要基因片段的子片段所形成的子集合中至少要有一個爲 **MERMAID** 所探測單元重要基因片段集合的子集。

MERMAID	**APPA**	
pattern	**pattern**	**detected**
ATGT	ATGTCATGC	No
TCATGC	ATGTTCATGC	**Yes**

圖 5.1　MERMAID 所探測得的單元重要基因片段與 APPA 間的關聯

舉例來說，假設今天 **MERMAID** 僅探測到兩個單元重要基因片段 **ATGT** 和 **TCATGC** ，那麼 **APPA** 所能探測長度超過 8 的單元重要基因片段只單單有 **ATGTTCATGC** 而沒有 **ATGTCATGC** 。（**圖 5.1**）

因此，在未來研究方向中，能運用 van Helden 等學者所提出的 **Pattern assembly** 方式將彼此間有重疊且相關的片段整合，進而得到更精確的重要基因片段資訊。由之前所作的實驗可以發現，在 **UME6** 這個生物基因序列群中，**APPA** 僅能找到具代表性卻非絕對完整的單元重要基因片段，倘若我們能將相關片段整合，則能得到最精確的片段資訊。（**圖 5.2**）

family	known motif	derived motif by APPA	summarized
UME6	TAGCCGCCGA	AGCCGCCG [4] GCCGCCGA [6] TAGCCGCC [7]	TAGCCGCCGA
GAL4	CGGRn2RCYnYnCnCCG	TCGGGGC-(7)-TCC [1] TCGGGGC-(8)--CCG [3] TCGGGG--(8)-TCC [6]	TCGGGGC-(7)-TCCG

圖 5.2　MERMAID 所探測得的單元重要基因片段與 APPA 間的關聯

參考文獻

1. Eskin, E. and Pevzner, P. A. (2002) Finding composite regulatory patterns in DNA sequences. Bioinformatics, 18, 354-363

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

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

4. Marsan, L. and Sagot, M. (2002) Algorithms for extracting structured motifs using a suffix tree with application to promoter and regulatory site consensus identification. J. Comput. Biol., 7, 345-360

5. Pevzner, P. A. and Sze, S. (2000) Combinatorial approaches to finding subtle signals in DNA seuqneces. In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology. pp. 269-278

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

7. 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. J. Mol. Biol., 281, 827-842

8. Pavesi, G., Mauri, G. and Pesole, G. (2001) An algorithm for finding signals of unknown length in DNA sequences, *Bioinformatics*, **17**, S207-S214

9. Price, A., Ramabhadran, S. and Pevzner, P. A. (2003) Finding subtle motifs by branching from sample strings. Bioinformatics, 19, ii149-ii155

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

11. 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

12. Pevzner, P. A. and Sze, S. (2000) Combinatorial approaches to finding subtle signals in DNA sequences, *In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, 269-278

13. Griffiths et al. (1996) An Introduction to Genetic Analysis, 6th ed., WH Freeman and Co.

14. Agrawal R. and Srikant R. (1994) Fast algorithms for mining association rules, In Proceedings 199J International Conference VLDB, pp. 487-499