第一章 緒論

一、研究背景

解碼基因為人類帶來了期望，因為可藉此了解遺傳及環境引起的人類進化、疾病和 DNA 內部的構造功能。解碼人類基因組計畫的目標即是決定全部的核酸序列，其計畫首度在 1985 年成形，在 1990 年美國國家衛生研究院才正式擬定人類基因組計畫 Human Genome Project (HGP)。迄今，在十年間，目前人類染色體 21 及 22 對已經被定序完成(Hattori, 2000; Dunham, 1999），隨著基因組定序已逐漸完成，愈來愈多的基因組序列會逐漸增加，開發從基因組序列中預測基因的工具也愈來愈受到重視，因為透過基因預測工具可幫助發現基因組序列中潛在基因的位置，以協助作為基因發現研究的第一步。

許多的基因組序列已被定序，研究出基因辨識的工具，可以幫助跨物種基因結構及功能的比較，將基因分類幫助了解未來疾病與基因功能相關，了解基因編碼區以外的基因組序列對基因功能是否有幫助的角色(Nowak, 1994)。定序的成功雖使我們能更詳盡地調查基因構造及功能，但是全基因組散彈法定序速度較慢，使得原本在 1990 年宣示所有人類基因組序列將在 2005 年前要完成定序的目標並不確定可以如期完成(Pennisi, 1998)，為加速如此龐大的基因鹼基序列計算量定序基因，需仰賴電腦科技運算的基因預測程式，因此許多程式陸續

地被開發。然而，基因預測工具相當多，目前這些程式仍無法精確的尋找基因，尚未能完全真正解決許多基因結構預測的問題。

本研究之基因預測的程式整合了電腦運算及統計技術，可在配合實驗室研究下幫助生物學家分析新定序的序列，迅速搜尋基因，並致力克服人類基因組複雜的結構，使得基因預測在實驗研究過程中化繁為簡，達成生物學家生物實驗研究先機，對於醫學研究影響深遠。

二、問題陳述

生物序列的特性與資料結構相當特殊，因為人體的遺傳及演化訊息，總共可以以四個字母來表達，所以研究中所使用的序列資料組成就這四種字母而已，其全部長度共約 30 多億的鹼基對。因為目前生物學界對基因組序列功能了解有限，所以雖然 DNA 資料庫中資訊(information)很多，但知識(knowledge)卻有待挖掘，因為研究 DNA 和蛋白質相關資料或論文的累積數目相當大，資料成長也隨著時間加倍，現有許多生物序列的資料庫所提供的資料量更是相當地龐大，由以真核生物中的人類基因組愈列最為複雜。真核生物基因複雜的特徵，可歸為以下兩點:

1. 真核生物基因結構較原核基因複雜

基因組由 DNA(去氧核糖核酸)本身兩條互補的鹼基鏈扭成雙螺旋

所構成，而每一條鹼基鏈又是由一個一個的核苷酸(nucleotide)接成的鏈型大分子，每一個核苷酸上附著一個鹼基，鹼基共有四種分別是:腺嘌呤(adenine)、鳥嘌呤(guanine)、胸腺密啶(thymine)、鳥密啶(cytosine)，通常一條鹼基鏈則以此四種鹼基以 A、T、G、C 四個英文字母表示其組成。人類基因組有 23 對染色體，基因組全長約 30 億對鹼基對，其中人類基因約有 30000 到 40000 個，如表 1-1。目前人類基因的平均長度 27Kb，編碼區序列平均 1,340bp，exon 的數量平均 8.8，中間 exon(internal exon)平均長度 145bp，intron 長度平均 3,365bp(Lander, 2001），依基因結構組成比例來看，基因組中有 30%是基因，每條基因有 5%是編碼序列。

表 1-1 生物基因體之定序與基因密度

各種生物基因體之定序與基因密度		
真核有機生物	基因組的大小與預測基因的數量	文獻
Saccharomyces Cerevisiae (Budding yeast)	12 Mb, 5570-5651 genes	(Goffeau, 1997)
Caenorhabditis elegans (worm)	97 Mb, ~19099 genes	(The C. elegans Sequencing Consortium., 1998)

Drosophila melanogaster (fruit fly)	137 Mb, 13601 genes	(Adams, 2000)
Arabidopsis thaliana (musturd weed)	125 Mb, 25498 genes	(The Arabidopsis Genome Initiative, 2000)
Homo Sapiens (human)	3200 Mb, 30000-40000 genes	(Lander, 2001)
Schizosaccharomyces pombe (fission yeast)	12.5 Mb, 4824-4940 genes	(Wood, 2002)

2. 真核生物基因組的特徵

研究基因預測可將下列基因組的特徵分成二類:

A. 訊號

尋找訊號可找尋基因組的功能區域，如起始及停止基因碼、剪接點、促進子、轉錄因子調控。

轉譯起始區(translation initiation sites)的 AUG 碼是大部分真核生物的轉譯起始區的起始基因碼(start codons)，約有 90-95%的脊椎動物 mRNA 起始都是此三核苷酸(Kozak,M)；另一種公認已知的轉譯終點區終止碼是 TAA、TAG、TGA；而剪接點中的捐獻區 AG/GT 亦

是預測基因位置的訊號,因為 AG 位在 exon 終止的雙核苷酸,而 GT 位在 intron 起始的雙核苷酸(Thanaraj, 2001)，目前也有相關程式合併使用 exon 及 intron 的統計特性去預測剪接點組成,計算剪接點的局部特徵，如 HSPL 程式使用線性鑑別法(Thanaraj, 2001)。

B.基因組結構內容及其組成

這類資訊以統計上非編碼區及編碼區的統計特徵為主,將編碼區域的特徵進行統計分析，為編碼區的測量(coding measures)或可能的編碼區(coding potentials)，許多編碼分析都是以觀察編碼區塊(codon)的使用情形為主，國外有研究是以 6 個鹼基為一組用發生頻率，，每次重複長度以大於 12 bp 為重複的判斷，分別統計分析內部可能的 501 種核酸重複序列，去測量編碼區塊和 intron 的特徵(Subramanian, 2003)。

另外亦有人用 CpG 島或 CG 核苷酸的含量去看,如圖 1-2 即可以以 GC 含量去判斷 exon 及 intron 的不同組成特徵(Lander, 2001)。

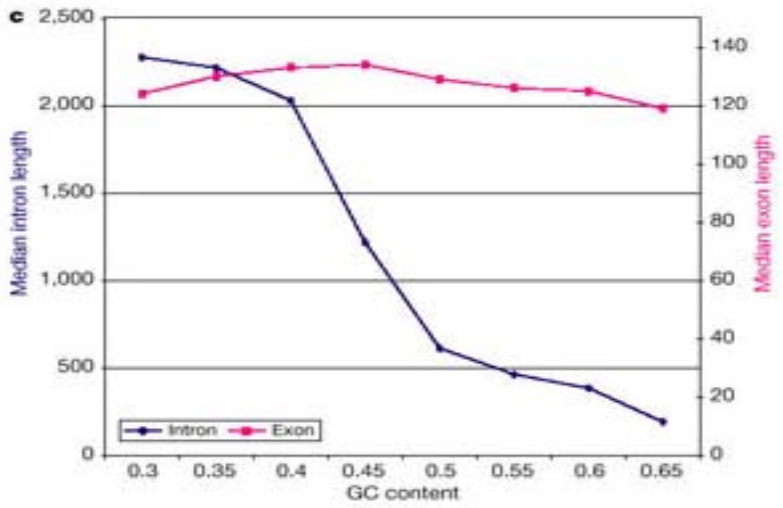


圖 1-1 exon 和 intron 有不一樣的 GC 比例

三、研究問題

原核生物和真核生物就有所不同，原核生物基因組較小的多，但有較高的基因密度，如 Haemophilus influenza 的基因組中有 85%的基因密度(Fraser, 1995)。原核生物沒有 intron，所以一條基因就是一條蛋白質、一條閱讀框架；真核生物和原核生物不同且較複雜的地方，在於必須要剪接基因(spliced gene)，切除 intron 的部分，過程即是在轉錄到 mRNA 上時，重新合成的 mRNA 會自己剪掉 intron 的鹼基序列，並且將連接 exon 的部分，這個過程稱 RNA 接合。因此由 DNA 轉錄成為原始的轉錄產物，即 mRNA 前體，再經過剪接以後，編碼蛋白質的外顯子部分即可連接成為一個連續的整體可譯框(open reading frame, ORF)，通過核孔進入胞質。如圖 1-1:

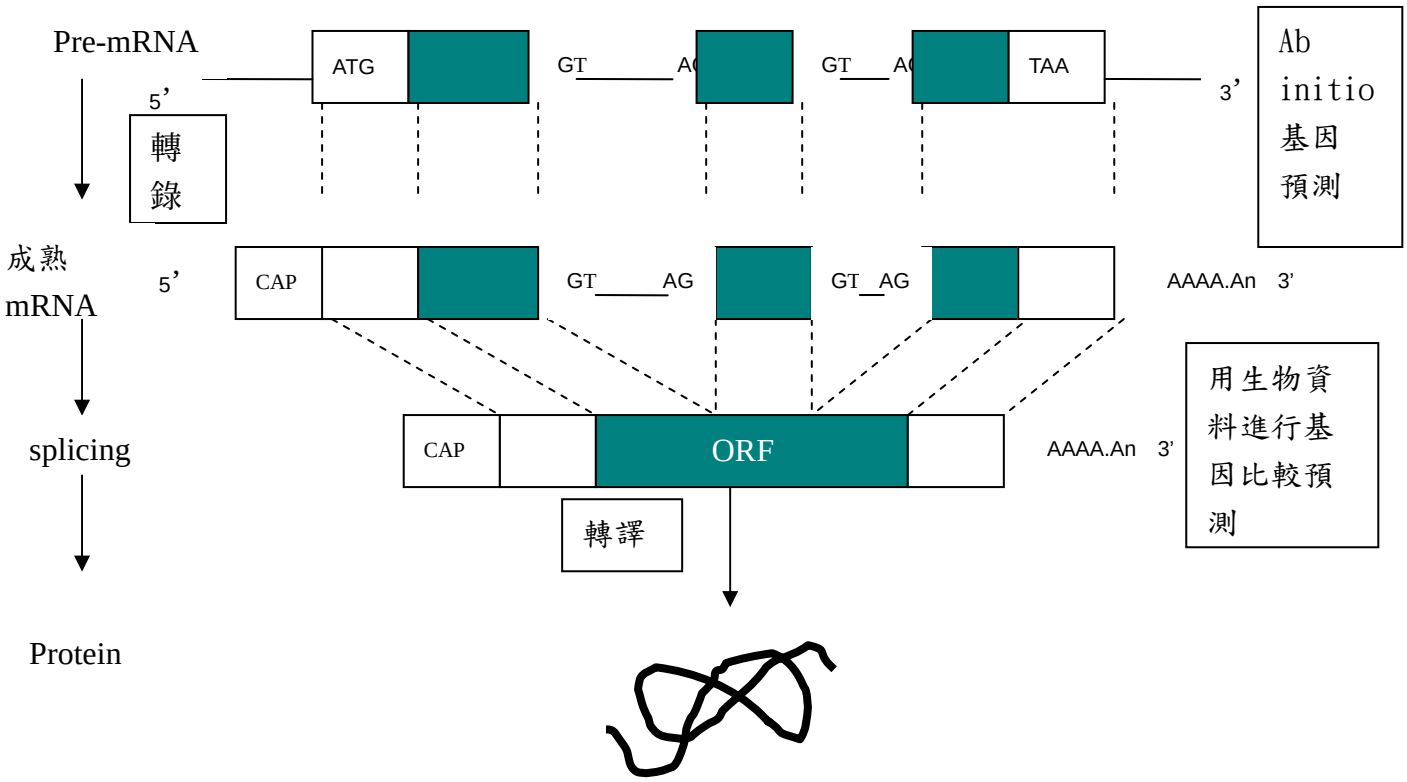


圖 1-2 人類基因表達的遺傳訊息

而原核生物不需進行切除 intron 進行 RNA 接合，所以在原核生物基因預測程式設計做法上只要搜尋開放閱讀框架就可以預測基因。相較之下真核生物編碼區密度較低，且每 kb 的基因數也較少，所以真核生物要進行基因預測較原核生物為複雜。當編碼區的長度縮短，exon 的預測變得更複雜，以程式設計而言，發現很多個短的 exon 比發現單一個長的 exon 要更為困難，這也是必須挑戰的問題。

四、研究目的

目前人類仍汲於探索基因醫學，我們的目標不僅僅在於知道何處有基因，還希望根據以往已知的基因鹼基序列來發現規則性，從規則與特徵推測基因的部分，希望能在尚未分析出特徵的基因組中，找尋潛在基因序列。因此發展一個友善的介面及有效率存取的工具，分析基因組資料，提供生物學家建立實驗假設，正是本論文的主要目的。

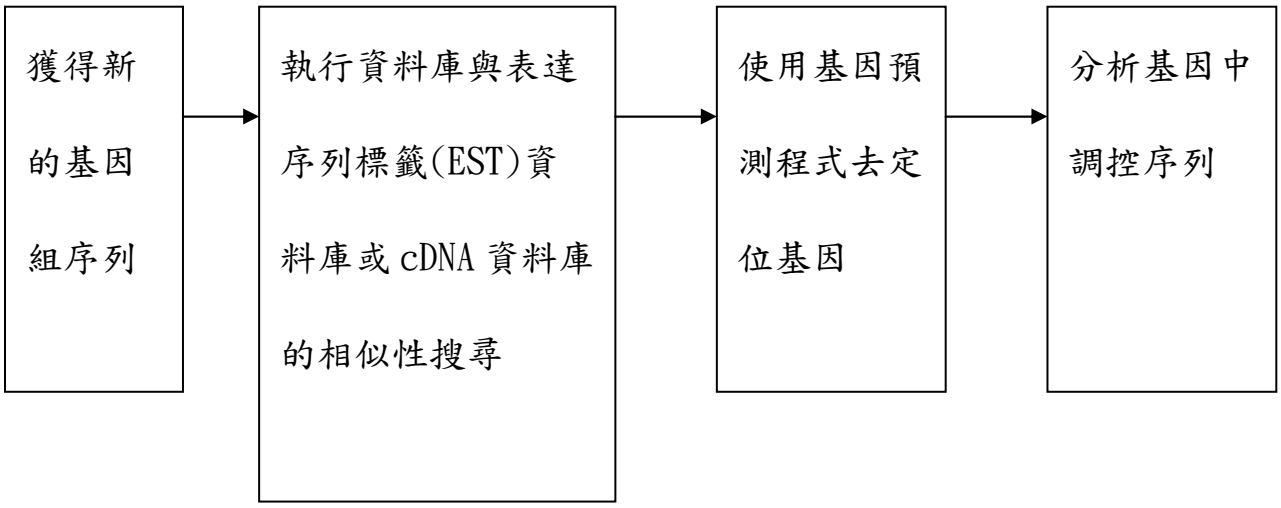
第二章 文獻探討

一、基因預測程式的探討:

1. 外部排比(Extrinsic):

亦稱同質法(Homology-based method; Similarity-based method)，直接使用經由實驗證實的註冊轉錄序列資料，和內部法不同的是，這類方法偵測先前已知的特徵，進行包括蛋白質、EST、cDNA等序列的相似性尋找，外部排比(Extrinsic)大多使用局部排比的工具去搜尋如 ESTs、mRNAs、cDNA 資料庫，再和基因組序列進行相似性排比(Mott, 1997; Florea, 1998;Bailey, 1998;Wheelan, 2001)。

外部排比法的限制是辨別不出資料庫中還未編碼成為蛋白質的基因，必須依賴先前所發表的生物序列資料才能產生預測，因此缺點是受限於先前存在的生物序列資料，完全仰賴資料庫中的序列，所以資料庫中的序列完全不能產生錯誤，如果序列相似度有限，尋找起來將會相當困難。在資料處理方面，則是計算時間長、儲存空間龐大。其資料處理流程如下:



2. 內部排比(intrinsic):

一般稱全始計算法(*Ab initio*)，不同於外部排比的地方在於預測基因時，並非直接使用實驗室所產生的表達序列，而是在進行大規模實驗室定序與生物實驗判定基因功能區域之前，事先就可以以統計為基礎計算序列的相似度，進行同質性基因區域(homogeneous)的區分，自動辨識出基因和蛋白質編碼的序列。

Ab initio 方法只使用序列中的資訊預測基因，直接以 exon、intron 和其他基因組序列的特徵屬性(computation properties)，整合編碼序列的統計資訊和訊號偵測，將所有的特徵進行統計計算，把序列想成一連串從 5 端到 3 端的線性特徵，去預測基因間際、編碼區及非編碼區，使用基因模式將基因特徵進行定位及計分， *ab initio* 方法預測結果偽陽率(false-positive)高(Hubbard, 2002)，但是 exon 預測敏感度佳，其方法有:隱馬可夫模式(HMM)、類神經網

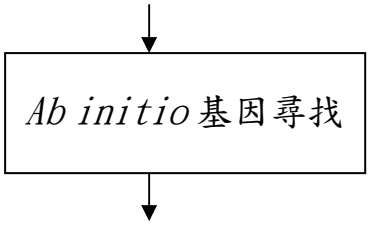
路、動態規劃演算法。目前許多基因預測的套裝軟體是使用隱馬可夫模式所設計的。

全始計算法(*Ab initio*)的操作原理:

Input:由四個鹼基{A,G,T,C}所組成的 DNA 序列。

Output:分別表示序列中每個核苷酸，哪些會經過編碼，哪些不會經過編碼。例如:

input: AAAGCATGACG TTACATGTCATC AG GA CTCCATACGTAA TGCCG



output: AAAGC *ATG ACG TTA CAT* GT CA TC AG *GA CTC CAT ACG TAA* TGCCG

而基於仰賴許多序列的訊號的 *ab initio* 法，目前必須要待其訊號正確性之證實，增加了此方法預測基因的不確定性。另外，它的問題是完全沒有生物方面的證據，在如此龐大的基因組序列中，不一定充分的就能代表預測結果，會有許多偽陽性的預測結果，容易高估。

A.隱馬可夫鏈模式

屬於全始計算法，目標在找出代表基因的一個統計模式，從基因組序列 5 端開始到 3 端結束,去找出分子序列的特徵架構、統計模型，以隱馬可夫鏈模式的訓練，用不同的狀態(state)計算來表示，如

intron、exon、intergenic 等區域，以序列或核苷酸的狀態再推估出基因組構造之全貌,和加權矩陣方式一樣可辨認出 gap,下圖是 HMM 中的隱藏狀態，其每個方框內分別是調控訊號的特徵、編碼起始區、intron 的接受區與捐獻區、轉譯停止區域，用這些內容來指出基因間際、exon 和 intron。這每一個方框的區域都有其核苷酸的頻率特徵，HMM 從許多已知基因的資料去獲得頻率訓練模式。

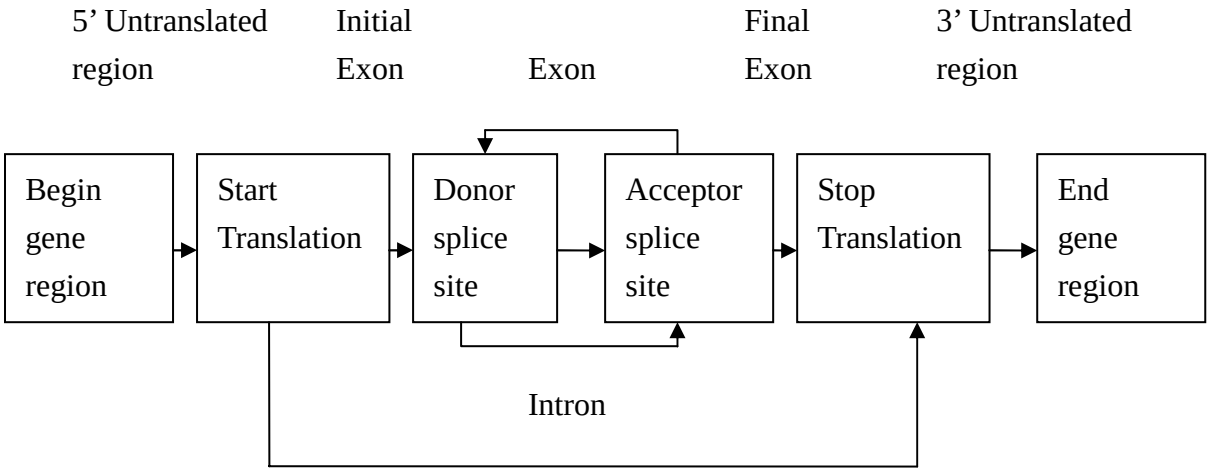


圖 2-1 HMM 中的隱藏狀態

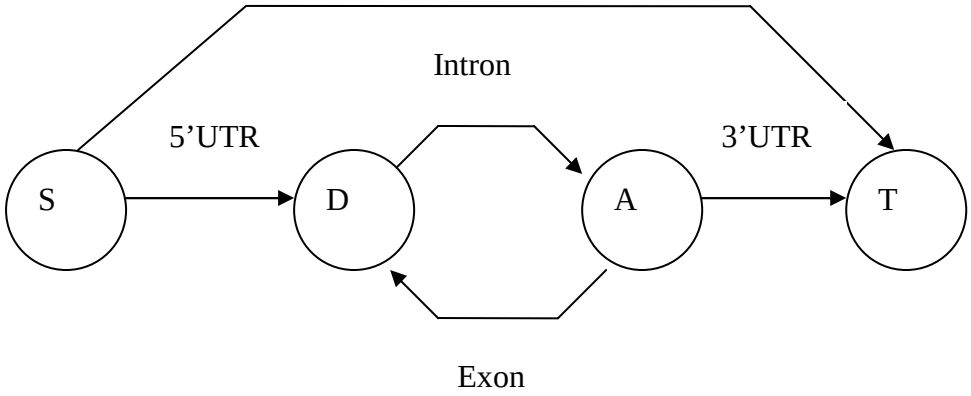


圖 2-2 簡單 HMM 模式

圖 2-2 是表示簡單的隱馬可夫鏈模式，S 和 T 表示起始碼及終止碼，D 和 A 是捐獻區和接受區，從 S 到 T 的單向弓形箭頭代表 HMM 所認定的單一個 exon 的基因，此模式可以不斷學習 DNA 序列中基因多種狀態轉換的機率模式。此法和其他演算法不同的是，不僅可找出全部基因，也允許找部分的基因，允許一股或兩股的序列中有多基因的發生，適用於脊椎動物、nematode、maize、arabidopsis、Drosophila，

此法的缺點是必須仰賴信號及已知序列的正確性。著名的程式有 Genscan（Burge, 1997）、Genie（Reese, 2000）、HMMGene（Krogh, 1997）。使用此方式預測蟲(worm)的精準度高於人類，可正確預測單獨 90%蟲的 exons 而人類只有 70%，預測一基因的所有的編碼 exons 正確率在蟲方面達 40%而人類只有 20%(Reese, 2000）。

1. GENSCAN (Burge, 1997)

作者在設計此工具的目標是建立基因組序列的模型，去捕捉真核基因的功能單位組成特性: exon、intron、splice site、promoter，著名功能單位例如:許多任何真核促進子 TATA 盒子和 cap 位置，或是找轉錄調控因子的邊界，如 MyoD(Lassar, 1989)。GENSCAN 使用一個三週期第五順序的編碼區 Markov model，此模式計算人類基因組中不同的 C+G%組成區域(Bernardi, 1989; Duret, 1995)的基因密度和基因結構，作法類似於 Genie 程式(Reese, 2000)使用 Generalized Hidden Markov Model 方法，首先使用一個雙股的基因組序列模式，分析兩股中潛在的基因，第二，不同於許多基因尋找的程式，大部分程式都假設輸入的序列中確實存在著一條完整的基因，而 Genscan 可假設序列中含有部分基因、完整基因、多條完整的基因、多條部份的基因或根本序列中就沒有基因存在，並使用雙股模型和捕捉多元基因類型的基因數目，這些功能與特性都使 GENSCAN 更適用於人類基因組片段的分析。

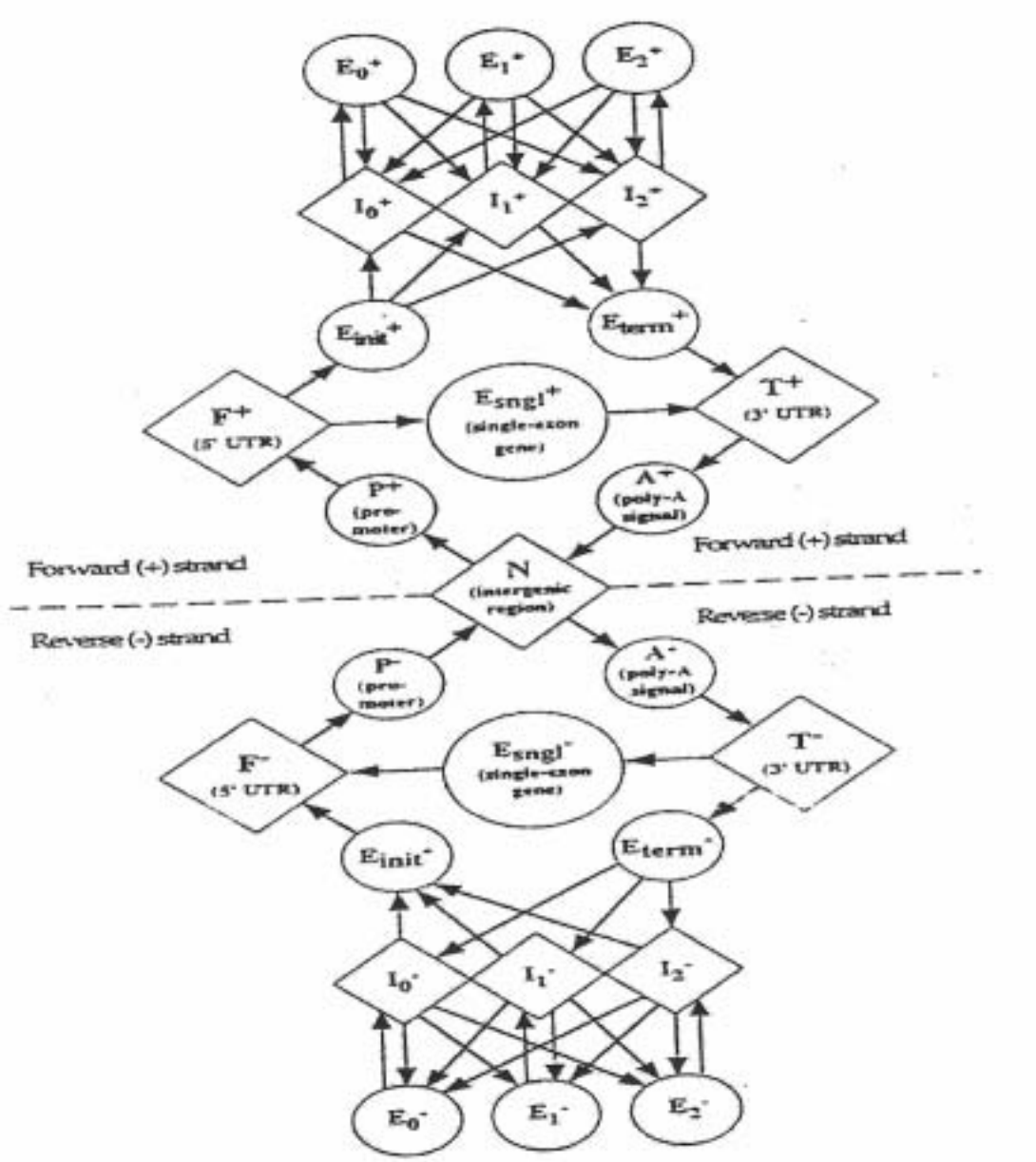


圖 2-3 GENSCAN 的一般基因組序列的結構模式

圖 2-3 是 GENSCAN 的一般基因組序列的結構模式，模式中的隱藏狀態是真核基因的功能單位，如 exon、intron、基因間際區，這些單位的發生有生物組成上的順序，introns 和 internal exons 在模式中，根據 phase 去區分最相關聯的 reading frame，因此，intron 在兩個鹼基之間表示為 phase 0，編碼子的第一個鹼基之後表示為

phase 1，編碼子的第二個鹼基之後表示為 phase 2，表示為I_0, I_1, I_2，internal exons 和 intron 根據 phase 去區分方法相似。反轉股(reverse strand)和向前股(forward strand)，以類似 GENMARK 程式(Borodovsky, 1993)的方式，其模式類似於 Genie，這裡模式則是更普遍的包含了(1)單個 exon 和多個 exon 的基因，(2)促進子、polyadenylation 訊號、基因間際序列，(3)DNA 兩股發生的基因，這個模式類似產生文法(parse)Φ，組成一系列的狀態，$\vec{q} = \{q_1, q_2, ..., q_n\}$，而和長度的一系列設定(持續的)，$\vec{d} = \{d_1, d_2, ..., d_n\}$，每個狀態的類型都使用機率模式，產生一段 DNA 序列 S 長度 $L = \sum_{i=1}^{n} d_i$，產生序列長度 L 的文法為下列:

(1) 根據一開始的狀態中的分佈$\vec{\pi}$選擇一個起始的狀態q_1，
$\pi_i = P\{q_1 = Q^{(i)}\}$，而$Q^{(i)} = (j = 1, ..., 27)$則是共有 27 種狀態索引

(2) 一個長度(維持的狀態)，d_i，和狀態q_1符合從長度分佈 $f_{Q^{(i)}}$ 所得的值$q_1 = Q^{(i)}$

(3) 產生一小段序列s_1的長度d_1，情境在d_1和狀態q_1的產生是根據一個適當的序列產生模式

(4) 子序列狀態q_2產生，情境是在q_1的值，從第一個順序的 Markov 狀態轉換矩陣T，$T_{i,j} = P\{q_{k+1} = Q^{(i)} | q_k = Q^{(i)}\}$

這些步驟一直持續，一直到加總$\sum_{i=1}^{n} d_i$出來狀態的持續期間(duration)

第一次超過長度L。這個模式有四個主要成分:起始機率的向量$\vec{\pi}$、狀態轉換機率的矩陣T、長度分佈f、產生序列的模式P，若固定序列的長度L，則空間是$\Omega = \Phi_L \times \mathrm{Y}_{L'}$，$\Phi_L$是所有可能文法的長度$L$，$\mathrm{Y}_{L'}$是DNA序列的長度$L$，模式$M$則是測量機率空間，因此部分機率$S \in Y_{L'}$，計算部分基因文法的情境機率為$\Phi_i \in \Phi_L$，使用Bayes Rule為:

$$P\{\Phi_i | S\} = \frac{P\{\Phi_i \cap S\}}{P\{S\}} = \frac{P\{\Phi_i \cap S\}}{\sum_{\Phi_i \in \Phi_L} P\{\Phi_i \cap S\}}$$

是預測外子和基因的程式，以GHMM模式為基礎，建構一個人類基因組的基因結構機率模式。此法中基因的機率模式分為幾種組成，預測完整的基因結構包括exon、intron、剪接點、促進區、和polyadenylation(PolyA)訊號，這些單位的發生，以此為搜尋基因模式的演算法。

預測序列的特徵計分是以log-odds用在測量部分(local)序列屬性特徵的品質，預測捐獻區域計分大於100分為「傑出的」,50-100分是「可接受的」,0-50分是「微弱的」,低於0分是「差的」,也就是分數愈少愈可能不是真的捐獻區)。

2.Genie程式(Reese, 2000):

亦使用GHMM統計模式，去看DNA序列之中規則。使用動態規劃演算法機器學習的方式,以最大機率模式判別序列所產生的路徑以及狀態,測試304個人類DNA基因,研究結果可辨識85%的蛋白質序列,

其準確率(specificity)80%。

B.類神經網路

使用類神經網路從5端到3端藉由訓練去確定基因的模式,所謂類神經網路的方法是給予訓練的資料,讓電腦找出訓練樣式,再讓電腦自己學習辨認樣式。

1.Grail II(Uberbacher, 1991)

Grail II 分析蛋白質編碼區、poly(A)區域和促進區,以此建立基因模型預測解碼蛋白質序列以及資料庫搜尋能力。一開始找最符合exon 的候選 exon,列出所有可能的 exon,建立標準把不相似的候選序列排除掉,再使用類神經網路去評估最終最佳的候選 exon,然後再使用動態規劃演算法去定一最可能的 model。輸入 Grail II 的序列樣式有許多種不同型態的序列分析資料,如圖 2-4 包括用隱馬可夫模式去做基因判斷、用類神經網路評估兩種剪接位置以及用 6 個鹼基一組去計分候選區域,這 6 個一組的鹼基分數是隨著 GC 鹼基的組成增加而增加,Grail II 自動預測基因並從 GC 鹼基的組成密度去區分編碼區域,結果發現在基因組序列編碼區中,的確是可以成為這種方式可以成為 exon 的預測指標。

如圖 2-4 是說明其類神經網路有七個感應結點(sensor)的演算法,以每 99 個鹼基去看計算,輸出的結點提供給類神經網路去評估,

類神經網路的輸出整合 CRM 的輸出，反映出編碼的 exon 其位置的 likelihood 。

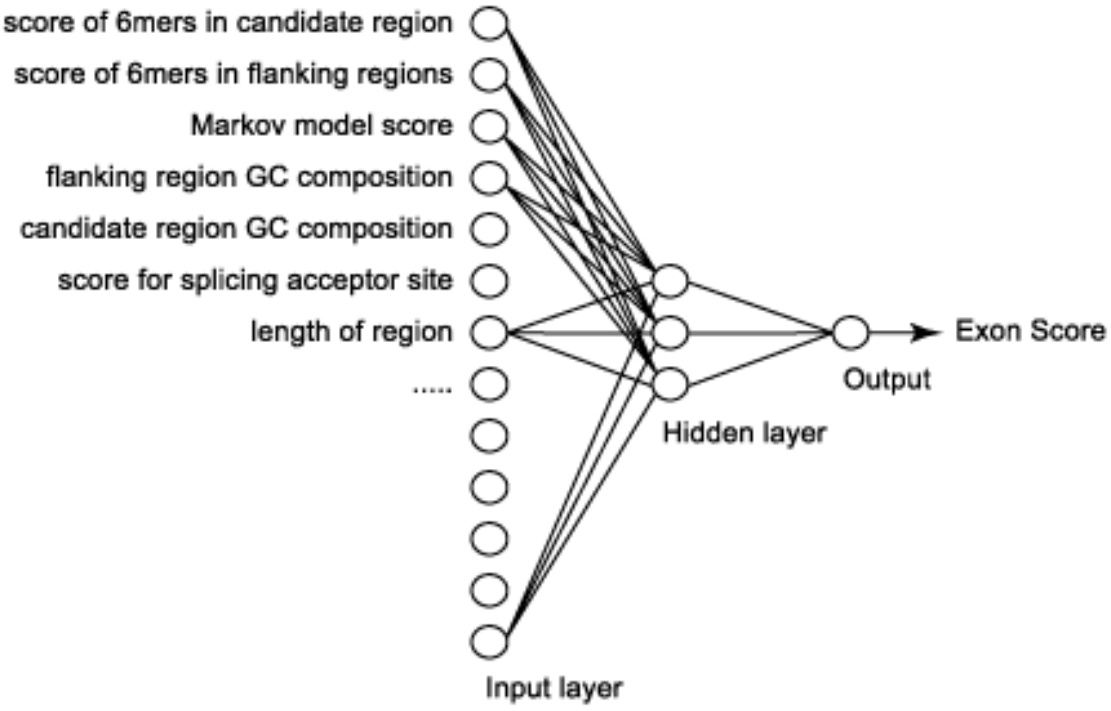


圖 2-4 Grail II 編碼區域的認可模式圖(Coding recognition module; CRM)

C.動態規劃演算法

動態規劃演算法是從有用的特徵去建立高的計分模式。

1.GeneParser(Snyder, 1995)

在基因結構內容統計指出 exon 或 intron 序列的所有小間隔後，此程式則依據此統計結果計分，去辨識 exon 及 intron 的邊界，這些資訊都是靠 exon 和 intron 的 log-likelihood 的近似程度以類神經網路給予權重，應用這些資料以動態規劃演算法去給予發現 exon 和 intron 組合發現最大的 likelihood 函數，exon 核苷酸預測的

correlation coefficient 為 0.89，而每一個 GC rich 基因的子集合的 correlation coefficient 為 0.94。

(1) 編碼區域統計特徵:把核苷酸看成是一連串六個鹼基為一組的重疊(overlapping)序列以六個鹼基為一組(in frame hexamers)產生編碼區頻率表格，統計 log-likelihood ratio，再以長度為 8 的寡核苷酸中鹼基發生次數的組成去看區域複雜度。

(2) 動態規劃演算法:計分是依據每個矩陣上述的 log-likelihood 分數,有四個矩陣(L-matrix)分別是 first exon, introns, internal exons, last exon，把這四個矩陣值加總，一直遞回的搜尋所有可能的最佳解。

(3) 找出最高分區段 type z 去看其位置 j，及最小的長度為何。

(4) 開始以類神經網路重複起始訓練。

(5) 每種型態的序列都會有一組初使的隨意權重，這些權重隨後會被調整直到訓練序列提供一個正確的基因結構，藉由類神經網路的方程式 $g(x)=\frac{1}{1+e^{-x}}-\frac{1}{2}$ ，觀察目標值到底是接近 0 或 1。

(6) 使用 G+C 內容去最佳化訓練資料的表現，因為 GC rich 有基因表達的含意(mouchiroud, 1991; bernardi, 1989) 。

(7) 當需要的精準度已達到，表示這方法已經可以用來使用基因組結構了。

D.使用規則預測基因

1.GeneID（Guigo, 1992,）

以人工智慧系統, 分析脊椎動物的 exon 和基因結構以階層式的規則為基因預測電腦系統之基礎(Hierarchical rule based approach)其策略是，給予一條 DNA 序列，辨識所有潛在的基因成分，如促進子區、轉譯起始碼及終止碼、剪接區域、poly(A)訊號，為上述的項目計算 likelihood，然後辨識和排序這些潛在的 exon 組合，理論上是優先評估預測的這些位置，然後過濾基因集結成分的評估，每一個集結的階段，都有切割值作為評估的門檻或作為過濾序列的最大敏感度，潛在基因的集結是階層式的，給予一段轉錄單位，使用絕對敏感的權重 profiles 去辨識起始碼、停止碼、捐獻區、接受區，從這些位置去獲得一組預測的 exon，然後計算這些預測基因在統計上是否顯著，exon 屬性是量化的，其每個測量的統計分佈可以作為真實 exon 的評估和使用距離函數去排序 exon(最佳距離或最頻繁的值)，門檻的值如真實 exon 有一個機率值在區間之外，從分佈中他們比很小的機率還小，這時就必須將之排除，在沒有顯著喪失敏感度的偽陽性數量減少也是成功的，目的是避免 exon 的組合暴量，藉由這樣過濾掉一組預測的 exon 而獲得預測基因的空間，如果 2 段 exon 是等值的，表示他們在同一個基因模型內，並在 2 段 exon 中的 intron

距離極小化，並定義每個基因模式都是一個等值 exon 分類(class)的線性排列，最後依函數值歸每個等值 exon 分類成分到每個潛在的基因當中，使用計分去評估基因模式中的空間排序，使用使用啟發的方式，以是否達到標準辨識候選的 exon，利用測量潛在的編碼區域去預測 exon，試著去預測脊椎動物基因組 DNA 和 exon，以一套規則將之集結為基因。

設計２４個變數去預測每個真實的 exon，第一個及最後一個 exon 分析之序列長度不得少於 100 和多於 20000 的核酸序列。為減少尋找潛在 exon 所花費的空間，使用類神經網路以刪減錯誤 exon 的方式節省花費的空間，並用以產生部分的鑑別函數，這些變數有：

變數 1 到變數 4 分別是 exon 值中的四個鹼基 A、T、C、G。

變數 5 到變數 8 是編碼位置的相關係數:第三個編碼子的位置和下一個編碼子的第一個位置的相關係數，和已被發現編碼區域中兩個連續編碼子中間位置的特徵(Smith, 1983)，在這兩者之中依雙核苷酸隨機分佈計算卡方檢定的變異數(deviation)，隨著 exon 長度校正卡方值。

變數 9 到變數 16 是在一開始的 exon 從變數 1 到變數 8 所產生出的變異數(deviation)，從變數 1 到變數 8 可獲得一開始 exon 的斜率(slope)計分分數。

變數 17-24 是最後的 exon 從變數 1 到 8 所生出的數字。GeneID 偵測真實 exon 的敏感度有 0.69，而預測編碼核甘酸的確是有進行編碼的部分是 0.89。

E.基因組結構內容統計特徵(Claverie, 1986)

以基因組序列結構統計特徵去預測編碼區域，早期是取代傳統把分子生物序列視為字串(string)或字元(character)，而把宏觀的分子序列視為一連串重疊的長度 k 個字一組(k-tuple)，k>1，如:DNA 序列 ATGCTCAGGATCAT…，而 2-tuples 有:AT, TG, GC, CT, TC, CA, AG, GG, GA, AT, TC, CA, AT,…，同樣的 3-tuples 有:ATG, TGC, GCT, CTC, TCA, CAG, AGG, GGA, GAT, ATC, TCA, CAT,…，其字母可為數目為有限的核苷酸或蛋白質，核苷酸的 k 為 1 到 9，蛋白質的 k 為 1 到 4，分割為 k-tuple 的數目 N(k)可表示為 N(k)=A^k，A 為 4 個核苷酸和 20 個胺基酸，例如:胺基酸 3-tuple 則有 20^3=8000 個不同的組合，而核苷酸 7-tuple 則是 4^7=16384 個不同的組合，每個序列的字元，其字母都有給予一個依序對照的 r，如:A=1, C=2, G=3, T 或 U=4，任何的 k-tuple 表示為一個向量，r = (r1, …,ri, …, rk)，這些向量在 1 到 A 的 k 次方之間轉為一個整數值 C(r)，

$C(r)=\sum_i (r_{i-1}) \bullet A^{(i-1)}+1$，以核苷酸為例，字母的大小就是以 2^2bit-wise 快速執行，用 k-tuple 的方式有下列優點:第一它提供關

於序列位置及相鄰位置的解碼資訊，第二它允許直接去注意表格(又稱為目錄)，去看每 k-tuple 的屬性，如出現頻率和位置，所有結果可被儲存，演算法以 k-tuple 為編碼主題可快速尋找群集序列的同質性(homology)(Wilbur, 1983; Karlin, 1983)或互補性(complementarity) (Dumas, 1982)，以及進行大型資料庫的快速搜尋(Lipman, 1985; Bishop，1984)並且快速產生目錄(catalog)及 k-tuple 在序列中的變化情形。測量基因和蛋白質中的頻率分佈鑑別真核生物基因的 intron-exon 區域和以特殊 pattern 找出蛋白質的位置和功能(claverie, 1986)，其鑑別指標(discriminate index) $d = P_{exon} \Big/ \left(P_{exon} + P_{\text{int}\,ron}\right)$，當所有的 globin gene d>0.5 時就是 exon，而 d<0.5 則是 intron，在這種方式可在尚未辨識的序列中，用啟發式方法(heuristic approach)幫忙找出新的結構和功能，有助於找出優先感興趣的分子基因組區域結構內容。

二、外部排比法及內部排比法的比較

許多方法普遍有下列相同的特點，包括目標都在於發現最佳切割區塊的方式，蒐集序列的統計特性；基於統計特性去發展計分函數，給予一段 DNA 區段(i, j)，計分這些區段 St(i, j)，t 是表示 exon 或 intron 或其他區段。但是綜觀這些方法仍有下列缺點要改進:

1. 外部排比法比較結論

雖然外部排比法累積先前的生物實驗為背景的強而有利的證明依據，因為序列明確清楚的呈現基因的表達，但是它只能辨識部分的 exon，僅侷限在基因結構中 mRNA 的局部資訊，必須顧忌人類的 EST 序列 40%有變位剪接(alternative splicing)偏誤問題 (Mironov, 1999)。因目前 EST 及 cDNA 的 library 還不完整，利用 EST 資料必須仰賴生物實驗序列的產生，相當緩慢，若是生物實驗序列資料不完整或不正確將影響到基因預測結果，所以必須仰賴 EST 序列的品質；除此之外，這種方法容易遺漏短的 exon(Mathé, 2002)。

2. 內部排比法比較結論:

至於內部排比法的全始計算法預測基因中運用機率 HMM 相當複雜，耗時且系統複雜，因為事實上基因並非連續的出現，基因不僅會被大塊的基因間際區(intergenic)區隔開，組成很少並且長度幅度較小的 exon，其組成約只佔基因組中的 5%，也被長度很長的 intron 所區隔，大於 50%的重複片段中亦有出現編碼區域(Lander, 2001)，如果重頭到尾進行序列結構的沙盤推演，增加了不少複雜與困難度。

基因組的特徵如前面所提及的分為兩種，一種是找尋訊號，一種是找尋基因內容的組成(Stormo, 1987)，基因結構內容統計應用於測量序列屬性是較長的序列時，幫助從其他序列的類型中區分

exon(Fickett , 1992)，而以訊號為基礎的尋找方式，則是用以搜尋短的共同保留序列成分。既然外部排比法容易遺漏短的 exon，因此，挑戰尋找 exon，傾向選擇內部排比法統計基因結構內容是較好的選擇。

內部排比法演算法設計內容多樣，有隱馬可夫模式(HMM)、類神經網路、動態規劃演算法等，其敏感度及精確度如下表:

表 2-1 內部排比法法各種基因預測程式精準度比較

		每個核苷酸的精準度				每條 exon 的精準度				
程式	序列	Sn	Sp	AC	CC	Sn	Sp	Avg	ME	WE
GENESCAN	570(8)	0.93	0.93	0.91	0.92	0.78	0.81	0.80	0.09	0.05
GeneID	570(2)	0.63	0.81	0.67	0.65	0.44	0.46	0.45	0.28	0.24
Genie	570(0)	0.76	0.77	0.72	n/a	0.55	0.48	0.51	0.17	0.33
Geneparser2	562(0)	0.66	0.79	0.67	0.65	0.35	0.40	0.37	0.34	0.17
Grail2	570(23)	0.72	0.87	0.75	0.76	0.36	0.43	0.40	0.25	0.11

資料來源:Burge, C., Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78-94.

表 2-1 中, Sn(Sensitivity)為敏感度、Sp(Specificity)為精確度、AC(Approximate correlation)、CC(Correlation coefficient) 、ME(Missing exons)、WE(Wrong exons)，每個核苷酸的精準度是和每條 exon 的精準度 GENESCAN 較高；敏感度方面 GENESCAN 最高，其次

是 Genie 和 Grail2；精確度方面 GENESCAN 較高，其次是 Grail2 和 GeneID；近似相關(Approximate correlation)以 GENESCAN、Grail2 和 Genie；每條 exon 的精準度敏感度方面 GENESCAN 最高，其次是 Genie 和 Grail2；精確度方面 GENESCAN 較高，其次是 GeneID 和 Genie；近似相關(Approximate correlation)以 GENESCAN、Grail2 和 Genie；Geneparser2 表現較好，ME(Missing exons)以 GENESCAN 和 GeneID 表現較好；WE(Wrong exons)以 Genie 和 GeneID 表現較好。

雖然內部排比法使用全始計算可觀得基因組結構全貌，以結果來看，隱馬可夫鏈具有較高的敏感度與精準度，但是較為耗費空間與時間，並且必須不斷重複地的進行資料的訓練，以訓練分數作為統計計分門檻去排除，難保不是今日偽陽率過高而造成高估的情形發生，而高估區段的內容為何，也需要了解與放大。為綜合這些缺點之後，考慮以每次位移的方式進行基因組序列比較的方式，透過基因組結構的比較去建立基因模式可以減少花費的時間，也可以簡化用 Ab initio 對每個基因結構訓練的過程，不僅明確找出 exon 區域相關鹼基組合出現規則的知識(knowledge)，更利用此知識去辨識基因組的序列，本方法不論 exon 區段計分結果如何，亦可詳細的調查與紀錄基因組序列中所有鹼基組合的發生情形，如果有高估的部分亦可放大詳細了解其結構變化。

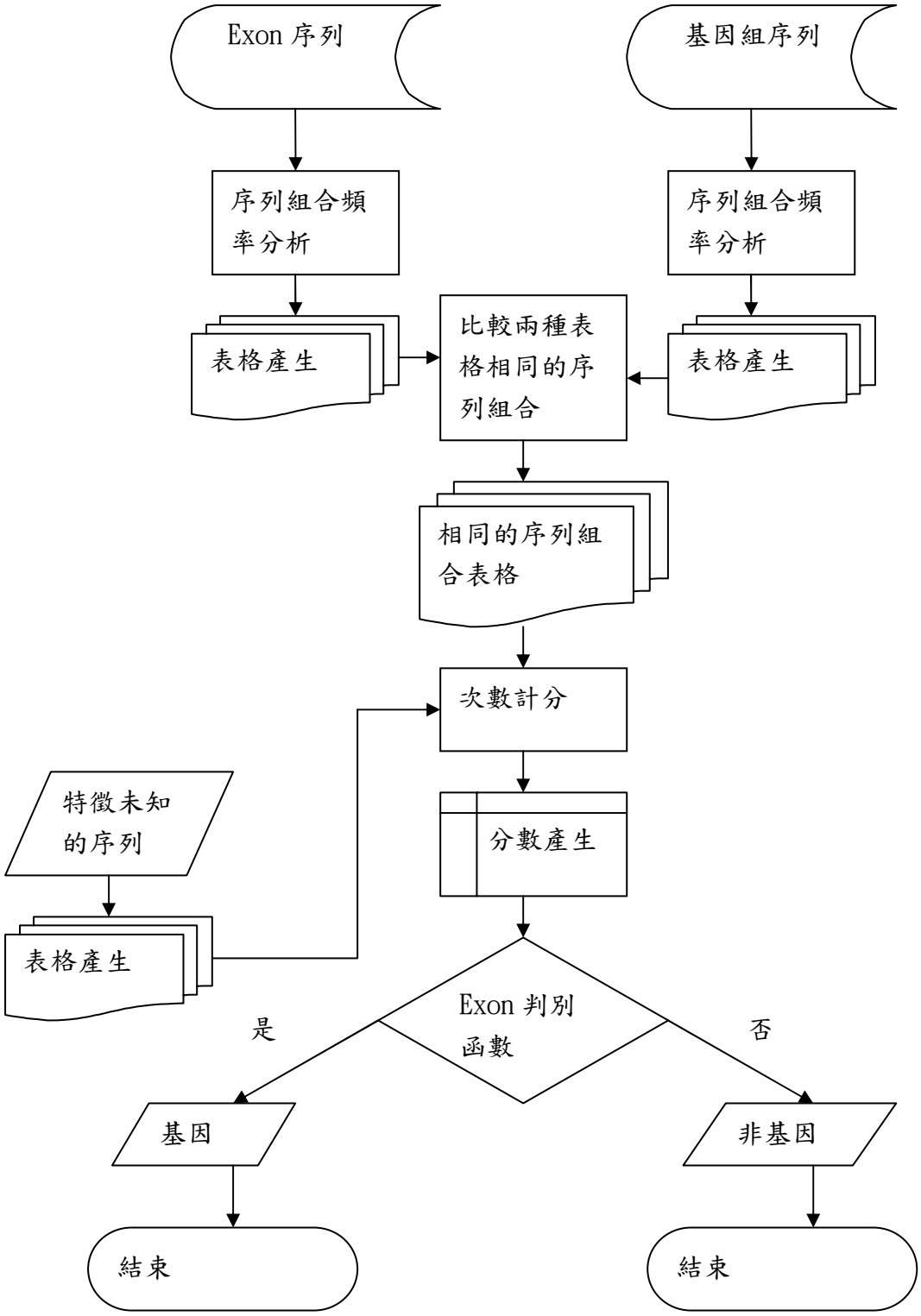
第三章 研究設計與方法

一、研究假設

本研究結合統計特徵的計算分析與候選 exon 區段計分演算兩種方式進行基因尋找。首先以每次平移一位鹼基統計基因組序列中的鹼基組合次數頻率分佈，接著再以每次平移三位鹼基以串聯(tandom)的方式，統計 exon 序列中的鹼基組合次數頻率分佈產生統計特徵表，以準備進行 exon 判別計分。

每段候選的 exon 會依 exon 序列與基因組序列共同發生的鹼基組合次數給予一個計分分數，最後加入觀察 exon 模糊區段的參數 Q 及 intron 模糊區段參數 J 作為 exon 邊界的判別函數。

二、研究模型


Exon 序列
基因組序列
序列組合頻率分析
序列組合頻率分析
表格產生
比較兩種表格相同的序列組合
表格產生
相同的序列組合表格
次數計分
特徵未知的序列
分數產生
表格產生
Exon 判別函數
是
否
基因
非基因
結束
結束

三、資料來源

本研究開發程式使用工具為 Visual Basic.NET，而人類基因組序列從 NCBI 的資料庫中得到，而人類 exon 序列則是從 the university of Queensland 序列資料庫中 Genbank 127 的檔案:human_exons.gb127.nr99.fna.gz 得到，三條研究範例的 exon 是:IDB400517_exon_8、IDB400517_exon_12、IDB400517_exon_2。

四、測量方法

1.理論步驟:

主要運用量化鹼基組合方式(莊振村, 民92)將序列資料做資料轉換，再對每一DNA序列分割成長度為*k*的Pattern，基因組序列每次分割平移1個鹼基，exon序列每次分割平移3個鹼基，於是每條DNA序列就形成由長度為*k*的Pattern所形成的集合，並計算這些小集合位於DNA序列中的位置與出現次數，再根據這些所獲得的資訊進行搜尋，簡單步驟如下。

步驟一: 首先將長度為 L_s 的基因組序列 S 和長度為 L_e 的 exon 序列 e 用函數轉換其資料型態，將 P_k 定義為長度 k 的 pattern，$P_k=\{B_1B_{2\cdots}B_k\}$，$B_i \in A,T,G,C$，$i=1\cdots K$，$K$ 可以根據使用者所需來自訂長度，基因組序列最多可能有(L_s-K+1)個組合，exon 序列最多可能有

(L_e-K+1)個組合，因配合胺基酸編碼故以 K=3 來切割。

步驟二:將基因組序列以每 k=3 個為一長度分割成許多 Pattern，每次分割平移 1 個鹼基，並設一個 Genome table 將基因組序列中所有可能構成的組合及出現次數和位置的資料紀錄到表格中，每次鹼基組合發生次數為 C，其發生位置為 P。

步驟三:將所有 exon 的序列以每 k=3 個為一長度分割成許多組合，每次分割平移 3 個鹼基，紀錄 exon 中所有可能構成的組合及出現次數，以便找出 exon 序列可能的規則，並設一個 exon table 將資料紀錄到表格中，每次鹼基組合發生次數為 N。

步驟四: 將 Genome table 和 exon table 兩者進行比對，找出兩者共同具有的組合，並將這些組合挑選出並紀錄到另一個表格 Genomexon table 中。

步驟五:再根據此表格所紀錄次數，來做為基因序列中的計分依據，並分別依照所紀錄的位置給予基因組序列一個得分值，當計算出來一段序列中所擁有的得分越高者表示越有成為 exon 區域的可能。

每段候選 exon 會依次數給予一個計分分數，其每個鹼基組合分數為 w_i，共有 i=1…n 個 exon 序列組合，算法為 $\sum_{i=1}^{n} w_i \Big/ n$ 。

步驟六:

每段 exon 會依次數給予一個計分分數，由於兩段 exon 之間，可

能因這模糊的區段出現的間隔而使得一條 exon 誤判為兩條 exon，因此，最後我們將所有 exon 具有分數的區段挑出使用表格整理，加入觀察 exon 模糊區段的容忍值參數 Q 於 exon 判別函數中作為判別。

步驟七:

以 GT/AG 訊號設計 intron 模糊區段的參數 J 加入 exon 的判別函數中，方便辨識 intron 範圍。

2.研究方法範例

假設有一條基因序列如:

tcgggtagga gagctccgat gccttcctct tggctgcaga cacagactgg aaggtaatca
agctctccgg cctaactagc actttacccg acacggtgct ctcccacctc ttcagccagg
tcatcatagg tggagtgcaa ctccaaactg gacccgacca ctacgacctt cctgaaggta at

，另外三條 exon 的基因序列分別是

>IDB400517_exon_2
gag agc tcc gat gcc ttc ctc ttg gct gca gac aca gac tgg aag
>IDB400517_exon_8
ctc tcc ggc cta act agc act tta ccc gac acg gtg ctc tcc cac ctc ttc agc cag

>IDB400517_exon_12
gtg gag tgc aac tcc aaa ctg gac ccg acc act acg acc ttc ctg aag

首先將基因序列以 k=3 分割為許多小集合，並這些組合的出現位置及發生次數紀錄於資料表中，其資料表名稱為 Genome table，資料型態如表 3-1，研究範例紀錄結果呈現於表 3-2:

表 3-1 Genome table 資料型態

欄位名稱	資料型態	摘要
Pattern	字串	以 K 個為一長度切割的 pattern
Count	數值	紀錄 pattern 在序列中的出現次數
Position	數值	紀錄 pattern 在序列中的出現位置

表 3-2 研究方法範例之 Genome table

pattern	count	position
tcg	1	1
cgg	3	2,68,94
ggg	1	3
ggt	6	4,53,95,119,129,177
gta	3	5,54,178
tag	3	6,77,127
agg	5	7,52,118,128,176
gga	4	8,49,132,150
gag	3	9,11,133
aga	3	10,38,44
gag	1	11
agc	4	12,61,78,114
gct	4	13,33,62,98
ctc	8	14,27,63,65,99,101,108,141
tcc	6	15,25,66,102,142,170
ccg	4	16,67,88,154
cga	4	17,89,155,164
gat	1	18
atg	1	19
tgc	4	20,35,97,136
gcc	3	21,70,115
cct	6	22,26,71,107,167,171
ctt	5	23,29,82,110,168
ttc	3	24,111,169
tct	4	28,64,100,109
ttg	1	30

tgg	4	31,48,131,149
ggc	2	32,69
ctg	4	34,47,148,172
gca	3	36,79,137
cag	4	37,43,113,117
gac	6	39,45,90,151,156,165
cac	5	41,80,92,105,159
aca	2	42,91
act	6	46,75,81,140,147,160
gaa	2	50,174
aag	3	51,60,175
taa	3	55,73,179
aat	2	56,180
atc	2	57,123
tca	4	58,112,121,124
caa	3	59,138,144
aag	2	60,175
cta	1	72,76,161
aac	1	74,139,146
ttt	1	83
tta	1	84
tac	2	85,162
acc	5	86,106,152,157,166
ccc	3	87,103,153
acg	2	93,163
gtg	3	96,130,135
cca	4	104,116,143,158
gtc	1	120
cat		122,125
ata	1	126
agt	1	134
aaa	1	145
tga	1	173

再將三條 exon 的序列資料也依照 k=3 進行分割，在 exon 的資料中只紀錄發生的次數紀錄於資料表中，其資料表名稱為 exon table，

資料型態如表 3-3，研究範例紀錄結果呈現於表 3-4:

表 3-3 exon table 資料型態

欄位名稱	資料型態	摘要
Pattern	字串	以 K 個為一長度切割的 pattern
occurrence	數值	紀錄 pattern 在 exon 序列中的出現次數

表 3-4 研究方法範例之 Exon table

pattern	occurrence	pattern	occurrence
gag	2	ccg	1
gct	2	cta	1
gat	1	ctc	4
gca	1	ctg	2
gcc	1	gac	4
aac	1	ggc	1
aag	2	gtg	2
aca	1	tcc	4
acc	2	tgc	1
acg	2	tgg	1
act	3	tta	1
agc	3	ttc	3
cac	1	ttg	1
cag	1	aaa	1
ccc	1		

再將所 exon 所得的資料與基因組所得的資料進行比對挑出兩者皆共同具有的組合，紀錄於資料表中，其資料表名稱為 Genomexon table，資料型態如表 3-5，研究範例紀錄結果呈現於表 3-6:

表 3-5 基因組與 exon 比較表，表格名稱 Genomexon table

欄位名稱	資料型態	摘要
pattern	字串	比對後符合條件的 pattern
occurrence	數值	Pattern 在 Exon table 中所紀錄的數值

Position	數值	Pattern 在基因序列中出現的位置

表 3-6 研究方法範例之 Genomexon table

pattern	occurrence	position	Exon occurrence
tcg	1	1	
cgg	3	2,68,94	
ggg	1	3	
ggt	6	4,53,95,119,129,177	
gta	3	5,54,178	
tag	3	6,77,127	
agg	5	7,52,118,128,176	
gga	4	8,49,132,150	
gag	3	9,11,133	2
aga	3	10,38,44	
gag	1	11	
agc	4	12,61,78,114	3
gct	4	13,33,62,98	2
ctc	8	14,27,63,65,99,101,108,141	4
tcc	6	15,25,66,102,142,170	4
ccg	4	16,67,88,154	1
cga	4	17,89,155,164	
gat	1	18	1
atg	1	19	
tgc	4	20,35,97,136	1
gcc	3	21,70,115	1
cct	6	22,26,71,107,167,171	
ctt	5	23,29,82,110,168	
ttc	3	24,111,169	3
tct	4	28,64,100,109	
ttg	1	30	1
tgg	4	31,48,131,149	1
ggc	2	32,69	1
ctg	4	34,47,148,172	2
gca	3	36,79,137	1
cag	4	37,43,113,117	1
gac	6	39,45,90,151,156,165	4

cac	5	41,80,92,105,159	1
aca	2	42,91	
act	6	46,75,81,140,147,160	
gaa	2	50,174	
aag	3	51,60,175	2
taa	3	55,73,179	
aat	2	56,180	
atc	2	57,123	
tca	4	58,112,121,124	
caa	3	59,138,144	1
aag	2	60,175	
cta	1	72,76,161	1
aac	1	74,139,146	1
ttt	1	83	3
tta	1	84	1
tac	2	85,162	
acc	5	86,106,152,157,166	2
ccc	3	87,103,153	1
acg	2	93,163	2
gtg	3	96,130,135	2
cca	4	104,116,143,158	
gtc	1	120	
cat		122,125	
ata	1	126	
agt	1	134	
aaa	1	145	1
tga	1	173	

再根據 exon 所得出來的次數作為計分的依據，首先先針對 exon 次數最高 pattern 進行計分，如 ctc 次數 4 次，在基因組序列上出現位置為 14、27、63、65、99、101、108、141，則在基因序列的 14、27、63、65、99、101、108、141 位置上紀錄 4 分，又如 gag 出現次數為 2 次出現位置為 9、11、133，則在 9、11、133 的位置上紀錄 2

分，其結果如圖 3-1。

Tcgggtagg^{2}a g^{2}a^{3}g^{2}c^{4}t^{4}c^{1}cg^{1}at^{1} g^{1}cct^{3}t^{4}cc^{4}tct^{1} t^{1}g^{1}g^{2}c^{2}t^{1}g^{1}c^{1}ag^{4}a
c^{1}ac^{1}ag^{4}ac^{2}t^{1}gg a^{2}aggtaatc^{1}a^{2} a^{3}g^{2}c^{4}tc^{4}t^{4}c^{1}cg^{1}g^{1} cc^{1}ta^{1}ac^{1}ta^{3}g^{1}c^{1}
act^{3}t^{1}ta^{2}c^{1}c^{1}cg^{4} ac^{1}a^{2}cgg^{2}t^{1}g^{2}c^{4}t c^{4}t^{4}c^{1}cc^{1}a^{2}cc^{4}tc t^{3}tc^{1}a^{3}g^{1}cc^{1}agg
tcatcatagg^{2} t^{1}gg^{2}ag^{2}t^{1}g^{1}c^{1}a^{1}a c^{4}t^{4}cc^{1}a^{1}a^{1}ac^{2}t^{1}g g^{4}a^{2}c^{1}c^{1}cg^{4}a^{2}cc^{1}a
c^{1}ta^{2}cg^{4}a^{2}cct^{3}t^{4} cc^{2}tga^{2}aggta at

圖 3-1 根據 exon 觀察次數作為計分依據的範例

如此可以將基因序列中內含可能為 exon 的候選 exon 序列以次數出現為特徵標出，並加總每個連續出現組合的頻率，給予凡是出現 exon 組合特徵出現的區段一個分數做為判斷候選 exon 的依據，當分數越高代表此段的序列越有可能為 exon 區域，如下:

每段候選 exon 會依次數給予一個計分分數，其每個鹼基組合分數為 w_i，共有 i=1….n 個 exon 序列組合，算法為 $\sum_{i=1}^{n} w_i / n$ 。

由於兩段候選 exon 之間，可能因這模糊的區段出現的間隔而使得一條候選 exon 誤判為兩條 exon，因此，最後我們將所有候選 exon 具有分數的區段挑出使用表格整理，以 exon 判別函數，函數中加入觀察 exon 模糊區段的容忍值參數 Q 作為判別，如圖 3-2:

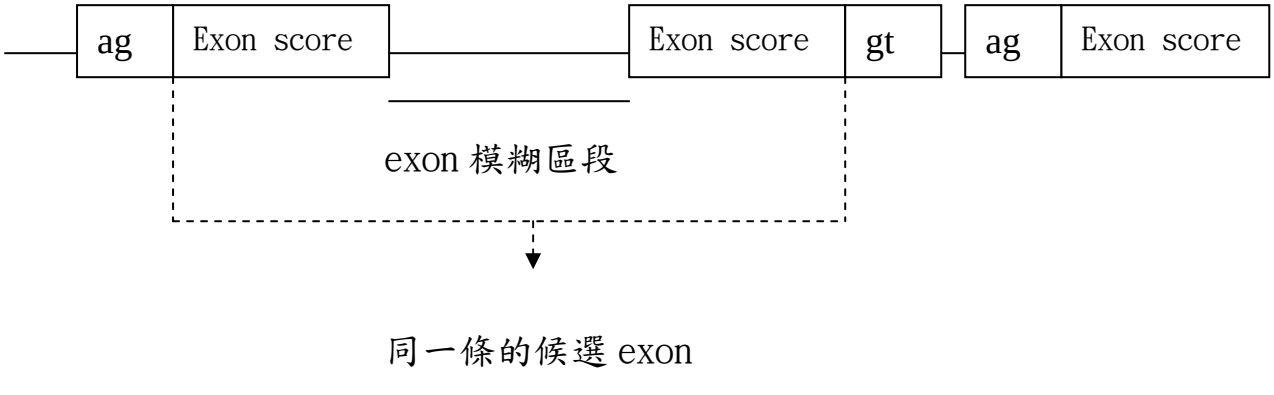


圖 3-2 函數中加入觀察 exon 模糊區段的容忍值參數 Q 的範例

假設一條基因組序列的計分結果是 $g^2t^1g^1c^1a^1attc^4t^4ccc^1a^1a^1ac$，出現兩條候選 exon，因此 Q 值以兩段候選 exon 中之間隙作為判別，整理 exon 特徵區段計分表的資料型態，表格名稱:exonscore table 如表 3-7，分別先後將找到的候選 exon 給予編號，紀錄每一段候選 exon 的起始組合和結束組合，起始位置和結束位置，每段候選 exon 以上述方法的計分分數，每一段候選 exon 和下一段候選 exon 的間隔為 Q 稱做 intron 可能長度，因基因組序列組合以每次位移一位，故 exon 序列組合在基因組序列之間連續出現時會發生重疊現象，因為序列組合以三個一組，若區段與下一段區段出現小於 3 個鹼基則計為同一段候選 exon，如 exon 序列計分結果 $g^2t^1g^1c^1a^1attc^4t^4ccc^1a^1a^1ac$ 中位置 10 的 tcc 和下一個特徵位置 13 的 caa 計為同一段候選 exon, 故每次 $Q \geq 1$，以才能計為下一段候選 exon，而候選 exon 區段的結束位置必須是此段候選 exon 最後一個序列組合的第三個鹼基，如 exon 序列計分結果中 $g^2t^1g^1c^1a^1attc^4t^4ccc^1a^1a^1ac$ 第 2 段候選 exon 區段結束的 AAC，候選 exon 區段結束位置為 17。

表 3-7 整理 exon 特徵區段計分表的資料型態，表格名稱:exonscore table

欄位名稱	資料型態	摘要
候選 exon	數值	依序將發現的 exon 編號

序號		
起始組合	字串	比對後符合條件的 pattern 的起始組合
結束組合	字串	比對後符合條件的 pattern 的結束組合
起始位置	數值	比對後符合條件的 pattern 的起始位置
結束位置	數值	比對後符合條件的 pattern 的結束位置
計分分數	數值	每條 Exon 中所計分的數值
intron 可能長度(Q)	數值	每一段 exon 和下一段 exon 的間隔

因此研究範例中，將挑出有 exon 序列計分分數結果的 $g^2t^1g^1c^1a^1attc^4t^4ccc^1a^1a^1ac$ 整理在表格 3-8，產生 2 段的候選 exon:

表 3-8 研究方法範例 $g^2t^1g^1c^1a^1attc^4t^4ccc^1a^1a^1ac$ 之 exonscore table

候選 exon 序號	起始組合	結束組合	起始位置	結束位置	計分分數	intron 可能長度
1	gtg	aat	1	7	6/5=1.2	1
2	ctc	aac	9	17	11/5=2.2	

以不同的容忍值 Q 作為例子對照:

例 1:如果 Q=1 時，那麼找出可能 exon 的區段將會有 2 段，如圖 3-3，其統計表格如同表 3-8:

$g^2t^1g^1c^1a^1at\ t\ c^4t^4ccc^1a^1a^1ac$

圖 3-3 容忍值 Q 為 1 時，會出現 2 段的候選 exon 的範例

例 2:如果 Q=5 時，那麼找出可能 exon 的區段將會有 1 段，如表

3-9:

表 3-9 容忍值為 5 時，表格合併結果

候選 exon 序號	起始組合	結束組合	起始位置	結束位置	計分分數	intron 可能長度
1	gtg	aac	1	17	3.4	0

在上個研究範例 $g^2t^1g^1c^1a^1attc^4t^4ccc^1a^1a^1ac$ 序列中，我們必須計算 intron 的可能長度，並將參數 Q 加入 exon 邊界的判別函數中，但這僅使用於同一條 exon 內部的候選 exon 區段邊界是否進行合併而判別。

在 exon 與下一段 exon 的邊界判別則是以 intron 的模糊區段進行辨識，如圖 3-4 位置 130 中 gtg 雖有訊號 gt 出現，但是以三個鹼基為一組的統計特徵可避免其誤判為 intron。不過在位置 54 的 intron 序列為 gtaat 應結束在位置 62 其 intron 應為 gtaatcaag 但卻結束在位置 58，因為統計序列組合時序列組合中有訊號成分的存在，所以會如圖 3-4 發生將 intron 判為 exon 的情形:

Tcgggtagg^{2}a g^{2}a^{3}g^{2}c^{4}t^{4}c^{1}cg^{1}at^1 g^{1}cct^{3}t^{4}cc^{4}tct^1 t^{1}g^{1}g^{2}c^{2}t^{1}g^{1}c^{1}ag^{4}a
c^{1}ac^{1}ag^{4}ac^{2}t^{1}gg a^{2}aggtaatc^{1}a^2 a^{3}g^{2}c^{4}tc^{4}t^{4}c^{1}cg^{1}g^1 cc^{1}ta^{1}ac^{1}ta^{3}g^{1}c^1
act^{3}t^{1}ta^{2}c^{1}c^{1}cg^4 ac^{1}a^{2}cgg^{2}t^{1}g^{2}c^{4}t c^{4}t^{4}c^{1}cc^{1}a^{2}cc^{4}tc t^{3}tc^{1}a^{3}g^{1}cc^{1}agg
tcatcatagg^2 t^{1}gg^{2}ag^{2}t^{1}g^{1}c^{1}a^{1}a c^{4}t^{4}cc^{1}a^{1}a^{1}ac^{2}t^{1}g g^{4}a^{2}c^{1}c^{1}cg^{4}a^{2}cc^{1}a
c^{1}ta^{2}cg^{4}a^{2}cct^{3}t^4 cc^{2}tga^{2}aggta at

圖 3-4 第 2 段 intron 為位置 54-58 的 gtaat

為解決此問題的產生，我們將設計 intron 模糊區段的參數 J 如圖 3-5 加入 exon 的判別函數中，方便辨識 intron 範圍:

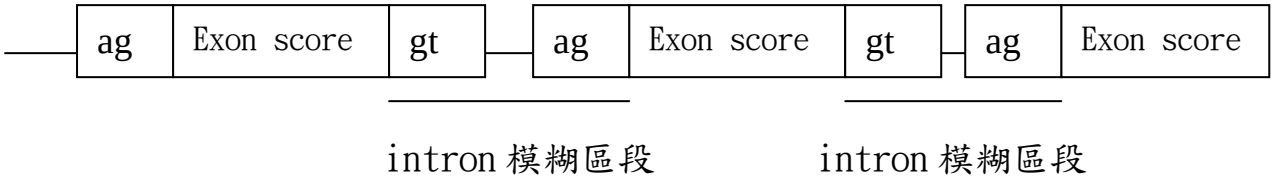


圖 3-5 函數中加入觀察 intron 模糊區段的範例

因此，以此方式從位置 58 往右邊平移尋找第一個訊號 ag 結束位置是 62，發現與候選 exon 區段重疊，因此如圖 3-6 中 intron 的區段將從位置 58 伸展到至位置 62，往左邊平移尋找 intron 的第一個開始訊號是 gt，因此發現符合。

Tcgggtagg^{2}a g^{2}a^{3}g^{2}c^{4}t^{4}c^{1}cg^{1}at^{1} g^{1}cct^{3}t^{4}cc^{4}tct^{1} t^{1}g^{1}g^{2}c^{2}t^{1}g^{1}c^{1}ag^{4}a
c^{1}ac^{1}ag^{4}ac^{2}t^{1}gg a^{2}aggtaatc^{1}a^{2} a^{3}g^{2}c^{4}tc^{4}t^{4}c^{1}cg^{1}g^{1} cc^{1}ta^{1}ac^{1}ta^{3}g^{1}c^{1}
act^{3}t^{1}ta^{2}c^{1}c^{1}cg^{4} ac^{1}a^{2}cgg^{2}t^{1}g^{2}c^{4}t c^{4}t^{4}c^{1}cc^{1}a^{2}cc^{4}tc t^{3}tc^{1}a^{3}g^{1}cc^{1}agg
tcatcatagg^{2} t^{1}gg^{2}ag^{2}t^{1}g^{1}c^{1}a^{1}a c^{4}t^{4}cc^{1}a^{1}a^{1}ac^{2}t^{1}g g^{4}a^{2}c^{1}c^{1}cg^{4}a^{2}cc^{1}a
c^{1}ta^{2}cg^{4}a^{2}cct^{3}t^{4} cc^{2}tga^{2}aggta at

圖 3-6 調整後第 2 段 intron 位置應為 54-62 的 gtaatcaag

第四章 討論

本演算法使用統計表格直接並精準尋找序列全面的特徵，以序列組合頻率發生作為特徵計分，不需複雜的參數與模型推估基因的可能區段，僅在邊界的定義上以函數中加入模糊區段變數，使用序列全面特徵與規則進行區段分析與預測，除可挑出最感興趣的部分進行調查外，在統計表格中的發生位置紀錄更可方便調查鄰近位置的鹼基組合變化情形。

在本研究中是以三個鹼基為一組進行分析，但是鹼基組合變數可自由定義，將目前已知的 exon 區段進行特徵分析，紀錄 exon 鹼基組合次數與位置，嘗試找出 exon 序列可能的規則特徵，這樣的方式相較於運用類神經網路與馬可夫鏈要來的簡單，另外一般的演算方式利用 GT 和 AG 來作為尋找 exon 的訊號這樣的方式容易造成誤判，而以我們的方式可以避免這樣的訊號造成誤判，也能夠準確的去預測出 exon 區段所在。

基因序列雖仍然不斷仍有生物實驗進行定序之中，但是，儘管如此，等待定序的進度完成的時日仍是遠遠落後；而已定序的序列，也需要快速有效率，又相當準確的序列分析工具，基因搜尋的演算法希望能藉此幫助生物學家，並提供一種具有創見的序列分析工具方法。以下提出本研究在未來發展的兩種應用：

1.利用於物種中的特徵比較:

此方法的主要目的是定義 exon 序列及基因序列邊界，而其邊界是藉由比較基因序列區段的鹼基組成統計特徵，以及邊界的函數進行劃分。

未來，在基因搜尋上可帶來便利性外，可延伸此方法的應用特性，辨識同源基因的保留基因序列，找出物種中的重複保留基因片段，藉此了解世代生物的基因進化。

2. 調查 exon 中鹼基出現頻率規則:

表格中的 exon 序列鹼基分佈情形與組成，可進一步利用找尋 exon 序列中是否存在統計規則，如鹼基組合的發生是否具有週期性，或在 exon 中是否有哪些鹼基組合特別出現頻繁，而這些規則是否和 intron 有相異之處，亦是本演算法可應用的範圍之內。

以 exon 的統計特徵及邊界設定的函數，作為進行基因的搜尋原則，在這部分必須仰賴目前已找到的 exon 序列才能進行分析。

參考文獻

1. Adams M. D., et al. “The Genome Sequence of Drosophila melanogaster” Science, vol. 287, pp. 2185-2195, 2000.
2. Bishop M., Thompson E. ”Fast computer search for similar DNA sequences” Nucleic Acids Research, vol. 12 no.13 pp. 5471-5474, 1984.
3. Borodovsky, M. andMcIninch J. “GeneMark: Parallel Gene Recognition for both DNA Strands” Computers & Chemistry, vol. 17, pp. 123-133, 1993.
4. Burge, C., Karlin, S. “Prediction of complete gene structures in human genomic DNA” J. Mol. Biol., vol. 268, pp. 78-94, 1997.
5. Bailey, L. C. Jr, Sears, D. B. and Overton, G. C. “Analysis of EST-driven gene annotation in human genomic sequences” Genome Res., vol. 8, pp. 362-376, 1998.
6. Bernardi G. “The icochore organization of the human genome” Annual Review of Genetics, vol. 23, pp. 637-661, 1989.
7. Claverie J. M., Bougueleret L. “Heuristic information analysis of sequences” Nucleic Acids Research, vol.14, pp. 179-196, 1986.
8. Dumas J. P. and Ninio J. “Efficient algorithms for folding and computing nucleic acid sequences” Nucleic Acids Research, vol. 10 no. 1 pp. 197-206, 1982.
9. Duret, L., Mouchiroud, D., et al. “Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochors” J. Mol. Evol., vol. 40 , No. 3, pp. 308-317, 1995.
10. Dunham, I., et al. “The DNA sequence of human chromosome 22” Nature, vol. 402, pp. 489-495, 1999.
11. Florea L., Hartzell G., Zhang Z., Rubin G. M., and Miller W. A. “Computer Program for Aligning a cDNA Sequence with a Genomic DNA Sequence” Genome Res., vol. 8, No. 9, pp. 967 – 974, 1998.
12. Fickett J. W., Tung C. S. “Assessment of protein coding measures” Nucleic Acids Research, vol. 20, pp. 6441-6450, 1992.
13. Fraser C. M., Gocayne J. D., White O., Adams M. D., Clayton R. A., Fleischmann R. D., Bult C. J., Kerlavage A. R., Sutton G., Kelley J. M., et al.” The minimal gene complement of Mycoplasma genitalium.Science.” 20; vol. 270(5235), pp. 397-403, Oct 1995.
14. Goffeau, A., Aert, R., Agostini-Carbone, M. L., Ahmed, A., Aigle, M., Alberghina, L., Albermann, K., Albers, M., Aldea, M., Alexandraki, D., et al. “The Yeast Genome Directory” Nature, vol. 387, pp. 1–105, 1997.
15. Guigo R., Knudsen S., Drake N., Smith T. “Prediction of gene structure.” J. Mol. Biol., vol. 226, pp. 141-57, 1992.

16. Hattori, M., Fujiyama, A., Taylor, T. D., et al. “The DNA sequence of human chromosome 21” Nature, vol. 405, pp. 311-319, 2000.
17. Hubbard, T. et al. “The Ensembl genome database project” Nucleic Acids Research, vol. 30, pp. 38-41, 2002.
18. Kozak, M.”An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs” Nucleic Acids Research, vol. 15, pp. 8125-8148, 1987.
19. Krogh, A. “Two methods for improving performance of an HMM and their application for gene-finding” In Gaasterland, T., Karp, P., Karplus, K., Ouzounis, C., Sander, C. and Valencia, A. (eds), Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology, 1997.
20. Karlin S., Ghandour G., Ost F., Tavare S., and Korn LJ. “New approaches for computer analysis of nucleic acid sequences” Proceedings of the National Academy of Sciences USA, vol. 80, pp. 566O-5664, September 1983.
21. Lassar A. B., Buskin J. N., Lockshon D., Davis R. L., Apone S., Hauschka S. D., Weintraub H. “MyoD is a sequence specific DNA binding protein requiring a region of myc homology to bind to the muscle creatine kinase enhancer” Cell, vol. 58, pp. 823-831, 1989.
22. Lander E. S., et al.”Initial sequencing and analysis of the human genome” Nature, 15; vol. 409(6822), pp. 860-921, Feb 2001.
23. Lipman, D. J., Pearson, W. R. “Rapid and sensitive protein similarity searches” Science 227:1435--1441, 1985.
24. Mathe C., Sagot M. F., Schiex T., Rouze P. “Current methods of gene prediction, their strengths and weaknesses” Nucleic Acids Research, vol. 30, No. 19, pp. 4103-4117, 2002.
25. Mironov A. A., Fickett J. W., and Gelfand M. S.” Frequent Alternative Splicing of Human Genes” Geneome Res.,Vol. 9, Issue 12, pp. 1288-1293, 1999.
26. Mott, R. “EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA” Comput. Appl. Biosci., vol. 13, pp. 477-478, 1997.
27. Mouchiroud, D., D’Onofrio, G., Aissani, B., Macaya, G., Gautier, C., Bernardi, G. “The distribution of genes in the human genome” Gene, vol. 100, pp. 181-187, 1991.
28. Nowak, R. “Mining treasures from ‘junk DNA’” Science, vol. 263, pp. 608–610, 1994.
29. Pennisi E. “ HUMAN GENOME PROJECT: Funders Reassure Genome Sequencers” Science, vol. 280, pp.1185, 1998.
30. Reese, M. G., Kulp, D., Tammana, H. & Haussler, D. “Genie-gene finding Drosophila melanogaster” Genome Res., vol. 10, pp. 529-538, 2000.
31. Subramanian, S. Mishra, R. K. And Singh L. “Genome-wide analysis of

microsatellite repeats in humans: their abundance and density in specific genomic regions” Genome Biol., vol. 4, No. 2, R13, February 2003.
32. Stormo, G. D. “Identifying coding sequences. In Nucleic Acid and Protein Sequence Analysis” A Practical Approach, pp. 231-258, 1987.
33. Snyder, E. E., Stormo, G. D. “Identification of protein coding regions in Genomic DNA” J. Mol. Biol., vol. 248, pp. 1-18, 1995.
34. Smith, T. F., Waterman, M. S. and Sadler, J. R. “Statistical characterization of nucleic acid sequence functional domains” Nucleic Acids Research, vol. 11, pp. 2205-2220, 1983.
35. Thanaraj T. A. and Clark F. “Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions” Nucleic Acids Research, vol. 29, No. 12, pp. 2581-2593 , June 2001.
36. The C. elegans Sequencing Consortium. “Genome sequence of the nematode C. elegans: a platform for investigating biology” Science, vol. 282, pp. 2012-8, 1998.
37. The Arabidopsis Genome Initiative. “Analysis of the genome sequence of the flowering plant Arabidopsis thaliana” Nature, vol. 408, pp. 796–815, 2000.
38. Uberbacher E. C. and Mural R. J. “Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach” Proceedings of the National Academy of Sciences USA, vol. 88 , No. 24, pp. 11261–11265, 1991.
39. Wheelan, S. J., Church, D. M., and Ostell, J. M. “Spidey: a tool for mRNA-to-genomic alignments” Genome Res., vol. 11, pp. 1952–1957, 2001.
40. Wilbur, W. J. and Lipman, D. J. “Rapid Similarity Searches of Nucleic Acid and Protein Data Banks” Proceedings of the National Academy of Sciences USA, vol. 80, pp. 726-730, 1983.
41. Wood, V., et al. “The genome sequence of Schizosaccharomyces Pombe” Nature, vol. 415, pp. 871-880, 2002.
42. 莊振村，楊鎮嘉，黃梅芬，”人類第 22 號染色體各種序列組合頻率之統計分析”，慈濟醫學，第 16 卷，第三期，第 151-158 頁，民 93 年。