國立交通大學

生物資訊研究所

碩 士 論 文

以二階隱藏式馬可夫模型預測特定蛋白激酶磷酸化的位置

Predicting Protein Kinase-Specific Phosphorylation Sites Using Second-Order Hidden Markov Models

研 究 生：葉智國

指導教授：何信瑩　教授

中 華 民 國 九 十 六 年 七 月

以二階隱藏式馬可夫模型預測特定蛋白激酶磷酸化的位置
Predicting Protein Kinase-Specific Phosphorylation Sites Using Second-Order Hidden Markov Models

研 究 生：葉智國　　Student：Chih-Kuo Yeh

指導教授：何信瑩　　Advisor：Shinn-Ying Ho

國 立 交 通 大 學

生 物 資 訊 研 究 所

碩 士 論 文

A Thesis

Submitted to Institute of Bioinformatics

Department of Biological Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Bioinformatics

July 2007

Hsinchu, Taiwan, Republic of China

中華民國九十六年七月

以二階隱藏式馬可夫模型預測特定蛋白激酶磷酸化的位置

學生：葉智國　　指導教授：何信瑩

國立交通大學生物資訊研究所碩士班

摘要

蛋白質磷酸化是在蛋白質轉譯後修飾中很重要的機制,在調控基本的細胞進行過程像是新陳代謝、訊號傳遞、細胞分化和細胞膜穿透性等扮演重要角色。在過去，要標記已被磷酸化的蛋白質和被磷酸化的位置，即使透過如二維電泳分析和質譜儀分析等新技術的幫助，仍然要耗費大量的人力與資源。因此發展使用蛋白質序列資訊的電腦輔助預測軟體來預測磷酸化的位置與它們特定的激酶，可以提供一個關鍵的選擇步驟，用來減少實驗中候選者的數目。HMMer 是一個用來做蛋白激酶特定磷酸化位置預測的很好軟體，它使用一階隱藏式馬可夫模型(HMM-1)及 Plan7 的架構，在重要激酶 PKA、PKC、CDK 等現有資料集上，可分別達到 82%、74% 和 82% 辨識率。

本論文經由對 HMMer 的分析，希望提出一套以二階隱藏式馬可夫模型(HMM-2)為基底的改良演算法 iHMM (improved HMM)用來預測特定蛋白激酶磷酸化的位置，希望從序列中取得更豐富的前後文資訊來提升預測正確率。藉由搭配使用貝氏資訊準則(Bayesian Information Criterion)的模型參數選擇方法，使 HMM-2 在資料集不夠大時能盡量避免眾所周知的過度適化(over fitting)問題。本論文將 Phospho.ELM 資料庫與 Swiss-Prot 資料庫結合，將有蛋白激酶註解的資料依照蛋白激酶屬性 PKA、PKC、CDK 等分類建立十八個資料集，然後分別用 iHMM 建立預測模型，以 5-fold 交互驗證做 30 次獨立測試的結果，並經由與 HMMer 和傳統 HMM-1 的效能比較來評估 iHMM。實驗結果發現 iHMM 的辨識率與 HMMer 相比得到將近平均 4.3%的提升，而跟傳統 HMM-1 相比則得到將近平均 3.6%的提升。本論文並進一步探討比較 HMMer、傳統 HMM-1 和 iHMM 三種方法的特性與優缺點。

關鍵字：磷酸化、激酶、隱藏式馬可夫模型、貝氏資訊準則

Predicting Protein Kinase-Specific Phosphorylation Sites Using Second-Order Hidden Markov Models

Student: Chih-Kuo Yeh Advisor: Shinn-Ying Ho

Institute of Bioinformatics
National Chiao Tung University

ABSTRACT

Protein phosphorylation is an important mechanism of posttranslational modifications and it plays important roles in regulation of essential cellular processes such as metabolism, cell signaling, differentiation and membrane transportation. In the past, laboratory identification of phosphorylated proteins and phosphorylation sites is usually tedious and cumbersome. Recently, large-scale methods of two-dimensional gel analysis and mass spectrometry techniques were applied to efficiently detect phosphorylation sites. However, experimental identification of phosphorylation sites is still expensive. Therefore, computational prediction of phosphorylation sites with their specific kinases using protein's primary sequences can provide a crucial selection step to reduce the number of candidates. HMMer using first-order Hidden Markov Model (HMM-1) with the Plan7 architecture is a conventional tool for prediction of kinase-specific phosphorylation sites and the prediction accuracies of HMMer are 82%, 74% and 82% for existing data sets of the important kinases PKA, PKC and CDK, respectively.

From the analysis of HMMer, this thesis aims to propose an improved algorithm iHMM using the second-order HMM (HMM-2) with more context information of sequences to advance prediction accuracy. With the use of Bayesian Information Criterion on the selection of model parameters, iHMM tries to avoid the known over-fitting problem when the sizes of data sets are not large. This thesis established 18 data sets of annotated kinases family such as PKA, PKC, CDK, etc from the Phospho.ELM database and Swiss-Prot database. The performance of iHMM is compared with those of HMMer and the conventional HMM-1 using 5-fold cross validation for 30 independent runs. Simulation results reveal that iHMM can improve the average accuracies of HMMer and HMM-1 near to 4.3% and 3.6%, respectively. Furthermore, this thesis investigated the advantages and disadvantages of iHMM, compared with those of HMMer and HMM-1.

Keywords：phosphorylation, kinase, Hidden Markov Model, Bayesian Information Criterion

致謝

首先，我要特別感謝我的指導教授－何信瑩老師，讓我有機會進入這個生物資訊的領域，並且在研究陷入瓶頸時，不厭其煩很有耐心的指導我，在何老師身上學到了除了研究論文、學習到許多生物資訊相關技術以外，還有更重要的是做人處事的態度。

此篇論文要感謝義雄，冠維與俊維學長在論文寫作上幫我檢查並提出修正與改進的地方。還有特別感謝俊維，家達與凱迪在生物方面提供專業知識與寶貴意見。另外感謝孝邦與鍾錡讓我有豐富的研究所生活，同時也要感謝廖芹的祝福，使我可以順利畢業。感謝實驗室所有學長、學弟們，在每次實驗室開會時的批評與指教，以及精神上的鼓勵。這短短的兩年研究生活，不管在學習、運動、玩樂大家總是互相照應，是我這一生難忘的回憶。另外也要感謝黃憲達老師，以及李宗夷學長提供生物資料的協助，以讓我可以完成這次研究。

最後我要感謝我的父母的栽培與無時無刻的關心，你們是我最好的依靠，有你們的支持與鼓勵，讓我沒有後顧之憂，可以專心於學業上的研究！

目錄

圖目錄

表目錄

符號說明

Σ ：胺基酸字母的集合

$|\Sigma|$ ：胺基酸字母的總數

A ：狀態轉移機率矩陣

B ：狀態內部的符號觀測機率矩陣

π ：初始狀態機率向量

λ ：HMM 所有的參數的集合

a_{ij} ：First-order HMM 的狀態轉移機率

a_{ijk} ：Second-order HMM 的狀態轉移機率

$b_j(k)$ ：狀態 j 觀測到符號 k 的機率

S ：HMM 所有狀態的集合

s_i ：HMM 的第 i 個狀態

Q ：觀測的狀態向量

q_i ：時間 i 時的狀態

O^p ：positive 序列樣本集合

O^n ：negative 序列樣本集合

$O_i^{p(k)}$ ：positive 序列的第 k 條的第 i 個字母 (胺基酸)

$O_i^{n(k)}$ ：negative 序列的第 k 條的第 i 個字母 (胺基酸)

δ_1 ：門檻值取訓練資料正確率最高的那一點來當測試資料的門檻值

δ_2 ：門檻值取測試資料正確率最高的那一點來當測試資料的門檻值

第一章 緒論

1.1 研究動機

蛋白質磷酸化是在蛋白質轉移後修飾中很重要的機制，影響基本的細胞進程像是新陳代謝、訊號傳遞、細胞骨架重組、細胞運動、細胞凋亡、分化等和細胞膜穿透性等等。蛋白質常常被各式各樣的蛋白激酶(protein kinase) 磷酸化。在過去，要標記已被磷酸化的蛋白質和被磷酸化的位置，時常是費時費力的。最近在技術上有很大的近步像是二維電泳分析和質譜儀分析，然而，這些實驗依舊是很昂貴的。在這點上，發展僅使用蛋白質一級序列的資訊來預測磷酸化的位置與它們特定的激酶的電腦輔助預測工具，可以提供一個關鍵的第一步選擇，來減少實驗中候選的數目，並且可以取代領域專家用人工知識來篩選。

運用機器學習(machine learning)的技術來預測磷酸化位置已成為生物資訊中熱門的研究領域。而在這個問題中一些有名的預測系統，像是 Nikolaj Blom 等人發表的以類神經網路方法為主的預測系統 NetPhos [1]，以及最近 Rune Linding 等人發表的 NetworKIN 系統[2]，然而這些系統在 sensitivity 與 specificity 相差都很大，都是其中一個很高，另一個很低，而在眾多方法中以 2005 年黃博士等人發表的 KinasePhos 系統[3]，所使用的 HMMer[4] 在 sensitivity 和 specificity 均有不錯的表現，HMMer 是一個用來做蛋白激酶特定磷酸化位置的預測不錯的工具，但 HMMer 使用上有些限制跟缺點，第一，它有結構限制，第二，它在訓練前需要先將資料做過序列對齊，第三，它基本上無法訓練多類的資料，第四，它基本上是屬於一階隱藏式馬可夫模型(first-order Hidden Markov Model, HMM-1)，基於這些動機，本研究開發一個改良式的統計分析模型 iHMM (improved HMM)來改良上例缺點。

1.2 問題描述與研究方向

本論文題目主要以二階隱藏式馬可夫模型預測特定蛋白激酶磷酸化的位置，在這裡我們的將此預測問題的定義是輸入一級蛋白質(primary structure)序列片段，然後系統要能準確判斷出中間的位置是否會被特定的激酶磷酸化。

HMM 是一個富有彈性而且複雜的統計模型，HMM 發展自今以來已有相當多的變形，從傳統的 HMM-1 到目前生物上大家常在用的 profile HMM，到更強大的二階隱藏式馬可夫模型(second-order HMM, HMM-2)。本研究開發的 iHMM 主要是建構於 HMM-2 下的一個改良式 HMM。iHMM 在設計上使用一個簡單且快速的方法來幫助 Baum-Welch 演算法加速收斂到全域最佳解。以及運用貝氏資訊準則(Bayesian information criterion, BIC)的模型參數選擇方法，使 iHMM 在資料集不夠大時能盡量避免眾所周知的過度適化(over fitting)問題。本論文會探討傳統 HMM-1 跟 HMMer 和本篇論文主要使用的 iHMM 三種方法的特性與優缺點的比較，我們最後將用實驗結果來呈現不受結構限制的完全圖的 iHMM 要比 first-order 的 Profile HMMs 的效能要好，不過缺點就是在訓練階段時需要多一些計算時間。

圖 1.1 為我們的系統架構示意圖，本研究把 Phospho.ELM[5]資料庫與 Swiss-Prot 資料庫結合，將有蛋白激酶註解的資料依照蛋白激酶屬性 PKA、PKC、CDK 等分類十八個資料集，然後分別用 iHMM 建立預測模型。本研究將完成蛋白激酶預測系統，並且單純由一級序列來預測出特定蛋白激酶磷酸化的位置，當使用者輸入待測蛋白質序列

時，預測程式將預測出哪些位置將會產生磷酸化。以期望未來生物學家藉此系統將能深入了解蛋白質磷酸化過程之重要機制，亦或是以反應之重要特徵更深入研究。

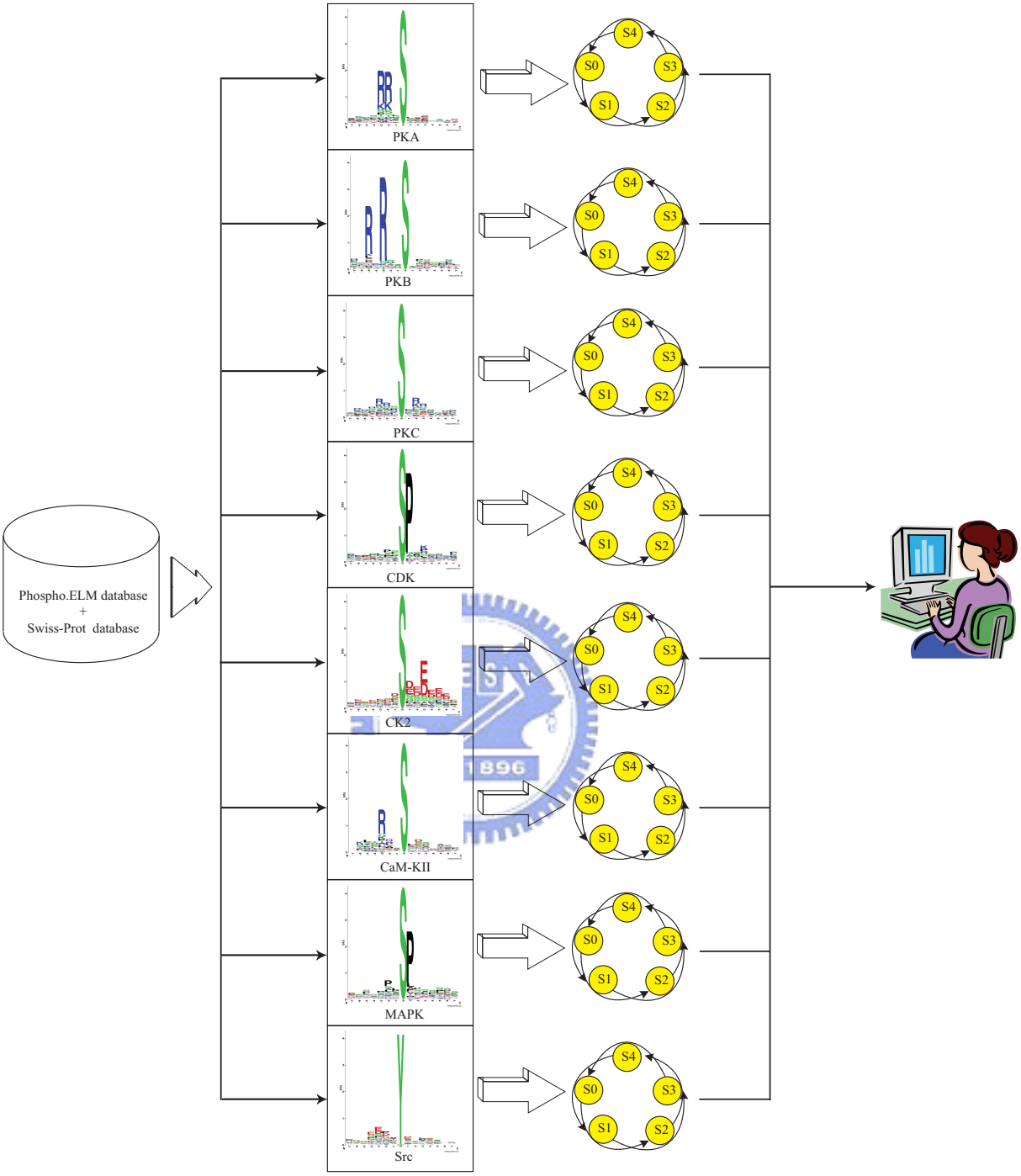

圖 1.1:系統架構示意圖

1.3 章節概要

本論文章節概要如下：

第二章介紹什麼是蛋白激酶，以及目前蛋白激酶主要有哪些分類，還有它們的活性，功能，結構做一個簡單的介紹。以及本研究所採用的激酶資料它的磷酸化的位置做一個簡單的整理。

第三章主要探討 HMM 的原理與方法。並分別介紹 HMM-1 和 Profile HMMs 和 HMM-2，以及 HMM 常使用的符號與解說，而在此章最後會講到 HMM-1 與 HMM-2 的訓練演算法。

第四章要介紹本論文所提出的 iHMM 方法，以及一般在使用 HMM 上可能會遭遇的困難與缺點，在這章節說明如何改良。以及 iHMM 如何將 HMM-2 與 BIC 結合，以及它的特性與優點。

第五章為實驗結果與討論，包括生物資料的擷取來源，以及如何處理還有如何實驗。最後我們將 HMM-1，HMMer，iHMM 三個方法做了正確率的分析與相關係數的比較。並且以 PKA 資料為例，以視覺化呈現的方式展示三者方法的結構。

第六章為歸納了本論文的結論，並提出針對磷酸化問題未來還可以繼續研究的方向與課題。

第二章 蛋白激酶基本介紹

2.1 蛋白激酶的分類

酵素上某些胺基酸的側鏈對於酵素反應的專一性及催化能力而言是很重要的，因為在酵素固有的構形中，會迫使這些旁鏈集中在一起，形成活性區 (active site)。活性區包含兩個重要的區域：受質結合區(binding domain)負責辨識並和受質結合，催化區(catalytic domain)則是在受質結合後負責催化反應。在某些酵素中，催化區就是受質結合區的一部分，而在其它酵素，這兩個區域的結構就如同它們功能一樣有很大的差異。蛋白激酶是激酶酵素，會在蛋白質的絲氨酸(serine, S)、蘇氨酸(threonine, T)或酪氨酸(tyrosine, Y)殘基由化學修正加入磷酸根（磷酸化），直接或是間接來調控蛋白質的功能。蛋白激酶作用絲氨酸(S)，蘇氨酸(T)和酪氨酸(Y)的磷酸化位置上的殘基具有保留性。它們有趣的研究是在分子演化的部份。在序列和結構的基本上，這些酵素形成關係密切的超家族(superfamily) 與組氨酸激酶和其它的磷酸轉移酵素截然不同 [6]。1988 年的時候 Steven K. Hanks 等人把真核細胞（eukaryotic cell）中的蛋白激酶分成五大類 AGC，CAMK，CMGC，PTK 和其它類[7, 8]。而在最近 G. Manning 等人的論文中[9]，則是把蛋白激酶依照催化域結構的差別、序列相似度，和生物上的功能分成七大類，並建立了蛋白激酶家族圖譜 ：

I. AGC 激酶家族：底下主要代表的是 PKA、PKG 和 PKC 三種激酶。AGC 激酶是比較寢向鹼性胺基酸的蛋白質，磷酸化主要做用於精氨酸(arginine, R)和賴氨酸(lysine, K)附近的絲氨酸(S)跟上蘇氨酸(T)上。
II. CAMK 激酶家族：底下包含了 CaMK、MARK 等許多激酶，CAMK 激酶是比較寢向鹼性胺基酸的蛋白質，在演化上與 AGC 激酶家族相近。
III. CMGC 激酶家族：底下主要代表的是 CDK、MAPK、GSK3 和 CLK 等依賴於細胞週期素的激酶，磷酸化主要做用於脯胺酸(proline, P)附近的絲氨酸(S)跟上蘇氨酸(T)上。
IV. TK 激酶家族：底下包含了 EGFR、INSR 和 SRC 等許多作用於酪氨酸(Y) 的激酶。
V. TKL 激酶家族：底下包含了 MLK、LISK、IRAK、Raf、RIPK 和 STRK 等激酶。
VI. CK1 激酶家族：底下包含了 CK1、TTBK 和 VRK 等許多激酶。
VII. STE 激酶家族：底下包含了 MAP2K、MAP3K、MAP 4K 等以 MAPK 組成級聯反應(cascade)的激酶。

2.1 蛋白激酶的功能

蛋白激酶作用在許多重要的信息傳遞途徑上，催化作用的緊密調控對於真核細胞的發展和維護至關重要，當它們活化時，不同蛋白激酶的催化區域採用顯著相似的結構，相較之下，未活化的蛋白激酶的結晶結在面對與特定調控區域或蛋白質的相互作用允許採用子然不同的構型蛋白激酶的區域暴露出可易見的可塑性 [6]。

重要的蛋白質磷酸化在真核細胞信號中反應出實際蛋白激酶的區域 2%全部在的真核基因中發現。特定絲氨酸(S)、蘇氨酸(T)或酪氨酸(Y)的磷酸化的空間與時間是控制細胞成長和發展的關鍵，激酶活化在錯誤的地方或是在錯誤的時間會有災害的後果，常常導致細胞轉型 (transformation) 或是癌症。其中以 TK 激酶家族與細胞信號通路有關，對基本細胞功能比如分裂、分化和抗凋亡信號進行調節。這些激酶雖可促進細胞分裂，但在非調控的啟動狀態下會導致各種腫瘤的形成。事實上已知的癌症基因中超過 70%就

是與酪氨酸激酶有關[6, 10]。

當幾百個真核細胞蛋白激酶，全部含有相同的催化支架(catalytic scaffold)，一些非常不一樣的調控機制已經演化成可以允許不同個體的家族輸入特殊訊號打開下游(downstream) 的功能 [6]。

蛋白激酶是分子開關可以採用至少兩個極端的構型，分別是啟動狀態與停止狀態。啟動狀態呈現最大活性，而停止狀態呈現最小活性。所有蛋白激酶的催化有同樣的反應，就是將 ATP 上的磷酸基轉移到絲氨酸(S)、蘇氨酸(T)或酪氨酸(Y)的殘基上的氫氧根群。這些活性他們全部只催化在結構非常相似的構型上。蛋白激酶在停止狀態時不需要受到必須達到活化狀態的化學限制，而且不同類別的蛋白激酶演化成採用不同的方式阻礙催化活性的構型來達成關閉狀態。

2.2 蛋白激酶的結構

為了要說明活性區如何和與特定受質結合並促進這個已結合的受質發生化學變化，我們來看環狀單磷酸腺苷依賴型蛋白質激酶(cyclic adenosine monophosphate dependent protein kinase)，現在通常被稱為 PKA，是真核細胞蛋白激酶領域中第一個用 X 射線結晶圖觀察到三級結構的例子，是第二訊息相關的酵素，主要做用在絲氨酸(S)，蘇氨酸(T)上，並且在細胞內處理(cellular processes)，包括轉錄(transcription)，新陳代謝及細胞的成長和凋亡扮演重要的角色。

圖 2.1: PKA 三級結構
(來源是Protein Data Bank http://www.rcsb.org 編號 2CPK)

圖 2.1 來源是Protein Data Bank (http://www.rcsb.org) 編號 2CPK，是由Knighton 等人結晶出來的 PKA 三級結構 [11]，此圖是由 PyMol 工具產生(http://www.pymol.org)。如圖 2.1，它的結構分成兩個小區域，小 N-端，或稱 N 葉(N lobe)，是一個五股的 β折板和一個明顯的 α 螺旋，稱為 helix αC。 C 葉(C lobe)是一

個大且顯著地螺旋狀。ATP (adenosine triphosphate)是包在兩葉深入的裂縫中，位置位於高保留的環 (loop) 連接者β1 和β2 兩股的下方。這個磷酸接合環，或 P 環，包含有保留的富含 glycine 基序(motif) (GxGxψG)這裡的ψ通常是酪氨酸(Y)或是苯丙氨酸(phenylalanine,F)。甘氨酸(glycine,G)殘基允許環靠 ATP 的磷酸非常接近且用骨架的交互作用協調他們，表留的芳香族的側鏈覆蓋磷酸轉移位置。 在缺乏 ATP的情況下甘氨酸(G)殘基使得 P 環變的非常易彎曲，事實上他促進小分子抑制物的接合。在環中，一些抑制物由保留的芳香族殘基的相互作用引起大結構的扭曲。

多肽型受質接合在核苷酸前端的延伸的構型，與 ATP 的γ-磷酸基接近。中心區域環是 “activation loop”，長通常約 20-30 個殘基，提供了一個平台給多肽型受質。 PKA 是最常見的蛋白激酶，當蛋白激酶活化時它的環是被磷酸化的。活化環的磷酸化使它穩定在開放和延伸的構型中允許受質接合。

PKA 的活化區是位在催化次單元上一個長達 240 殘基的「激酶核心 (kinase core)」，激酶核心在所有蛋白激酶都具有高度保留性，主要功能是和受質(ATP 及目標 peptide 序列)結合，並且將 ATP 上的磷酸基轉移到絲氨酸(S)、蘇氨酸(T)或酪氨酸(Y)的殘基上。激酶核心是由一個大的結構域及一個小結構域所組成，介於中間的是一個深的裂縫，而兩個結構域上都有構成活性區的殘基。

蛋白激酶通常保持停止狀態，而且催化區域的活化通常是包埋在多層的控制中，範圍從異位效應 (allosteric effectors) 的相互結合到亞細胞內位置轉移。當激酶在不活化與活化兩個狀態之間切換時，活化環有能力進行較大的構型改變。例如，胰島素受器激酶 (isuline receptor kinase, IRK) 是一個被三個 Y 殘基磷酸化與它的活化環活化。在沒有磷酸化狀態，活化環塌陷進入活化位置，凍結核苷酸和多肽型受質兩個接合。在磷酸化上面，活化環延著催化中心移動並且採用一個允許接合與催化的構型。在許多激酶，包括 IRK，在活化環的 N 端因為曲軸上的移動造成這些構型的改變，造成在 Asp-Phe-Gly 序列保留適當的方向有活化位置。

2.3 特定蛋白激酶磷酸化的位置

蛋白激酶是激酶酵素，所以他與其它酵素一樣在活性區俱有專一性辨識與結合基質的能力。我們將不同激酶特定磷酸化位置絲氨酸(S)、蘇氨酸(T)或酪氨酸(Y)的地方取出左右兩邊範圍各 7 個殘基，總長 15 建出序列圖案(sequence logos)，放在附錄中。這些序列圖案是由 WebLogo 工具 (http://weblogo.berkeley.edu [12])產生出來。序列圖案是一種基由 Shannon 訊息概念圖示化的技術，是用來呈現已經序列排列好常用的慣例，每一個字母的高度是與相對胺基酸出現頻率成比例，表示在那個位置的保留程度，如果某個位置全部都只出現那一個胺基酸，則 Shannon 訊息最高就是 $\log_2 20 = 4.32$ bits，相反的如果那個位置每個胺基酸出現的頻率都一樣，則就是 0 bits，這個觀念由 Schneider 在 1990 做了介紹[13]。我們將 13 個常見的激酶的位置保留性列在下表，紅色字體(S/T)表示在那個位置磷酸化，X 表示這個位置可以是任一個胺基酸。舉例來說我們可以觀查到 PKA 在位置 -2 和 -3 的地方是有很高頻率會出現精氨酸(arginine, R) 和賴氨酸(lysine, K)，而 PKB 在位置 -3 地方是有很高頻率會出現精氨酸(R)，在位置 -5 的地方有很高頻率會出現精氨酸(R) 和賴氨酸(K)，這些對於蛋白激酶特定磷酸化位置的預測是一個很重要的特徵。

表 2.1:蛋白激酶磷酸化的位置保留性

激酶	位置保留性
PKA	(R/K)-(R/K)-X-(S/T)
PKB	R-X-R-X-X-(S/T)
PKC	R-R-X-(S/T)-X-(R/K)
PKG	R-(R/K)-X-(S/T)
CaM-KII	R-X-X-(S/T)
CK1	S-X-X-(S/T)-X-E-S
CK2	(S/T)-(D/E)-(D/E)-(D/E)
CDK	(S/T)-P-X-(R/K)
MAPK	P-X-(S/T)-P
ATM	S-Q
EGFR	E-E-X-X-Y-V
INSR	D-Y-M-X-M
SRC	E-(E/D)-X-X-Y

第三章 方法

在眾多的預測模型中，像是決策樹(decision tree)、支持向量機器(support vector machine, SVM)以及類神經網路（neural network）等等，這些方法在解決不同類型的分類問題時都有各自的優缺點，然而我們這裡主要採用 HMM，主要是因為我們今天要預測的磷酸化問題與它周遭的胺基酸關係很密切，而 HMM 具備有良好處理時間序列的能力，把序列特徵中相關的胺基酸以時間來區分，找出某段時間內的可預期行為特徵，正好符合這些特性，因此我這裡主要探討以 HMM 為主的預測方法，分別是傳統的 HMM-1 和 Profile HMM 和 HMM-2。

3.1 First-order HMM

HMMs 是一個將有限狀態機與機率結合的一個統計模型，專門用來描述連續符號序列的條件概率。HMM 技術在語音辨識領域中已經有相當多年的歷史，是馬爾可夫模型的擴展。該模型由兩個隨機變量序列組成：一個是觀測不到的馬爾可夫鏈，另一個是可以觀測到的隨機序列。在數學上完整描述一個 HMM 是以起始狀態機率，狀態轉移機率，和狀態觀測機率三種機率矩陣所組成，分別以 π 、A 、B 表示，並且以 λ≡(A,B,π) 表是 HMM 所有的參數的集合[14]：

(1) 起始狀態機率

起始狀態機率是一個隱藏的隨機程序，它決定以某個狀態開始的機率，在數學上一般是 $\pi_i \equiv P(q_1 = S_i)$ 來表示，並且要遵循 $\sum_{i=1}^{N} \pi_i = 1$ 這個機率限制。

(2) 狀態轉移機率

狀態轉移機率是一個隱藏的隨機程序，它決定停留在原狀態或是遷移到下個狀態的可能性，狀態之間的連線表示它們的因果關係。在數學上一般是 $A \equiv [a_{ij}]_{NxN}$ 矩陣來表示，其中 $a_{ij} \equiv P(q_t = S_j \mid q_{t-1} = S_i)$，所有狀態都要遵循 $\sum_{j=1}^{N} a_{ij} = 1$ 這個機率限制，如：從某一狀態出發轉移到下個狀態的機率總和為 1。

(3) 狀態觀測機率

狀態觀測機率則描述在每個狀態觀察到某個現象的機率，在數學上一般是 $B \equiv \{b_j(k)\}$ 來表示，其中 $b_j(k) \equiv P(v_k at\ t \mid q_t = S_j)$，並且要遵循 $\sum_{k=1}^{N} b_j(k) = 1$ 這個機率限制。

3.2 Profile HMM

Profile HMMs 是一種「機率模型」的多重序列對齊技術(statistical models of multiple sequence alignments)，它捕捉在多重序列比對後每一行的位置特異性(position - specific)的資訊，包括有關保留性(conserve)，以及殘基的相似性。Profile HMMs 是由 Anders Krogh，David Haussler 等人在美國加州聖塔克魯斯大學首先引介給計算生物學 [15]，然而在生物學上 Krogh/Haussler 等人並不是第一個使用 HMM，從歷史上來看，HMM 的使用最早可以追朔到 1989 年，Churchill 為第一位將 HMM 用在異質基因型

(heterogeneous DNA) 序列的建模上[16]。生物學家一旦拿到未知功能的 DNA 或蛋白質序列時，最常做的事情就是從資料庫中進行序列比對，而 Krogh 的論文提出了一個震撼性的觀點，因為 HMM 技術是非常適合使用多重序列比對以廣泛性的 profile 方法進行資料庫搜尋取代單一序列搜尋。如圖 3.1，一般而言，Profile HMMs 有三個主要的狀態分別是 match (M)，insertion (I)和 deletion (D)，而機率的參數是從多重序列比對而來。HMMer 是其中有名的 profile hidden Markov model 一個工具程式[4]。

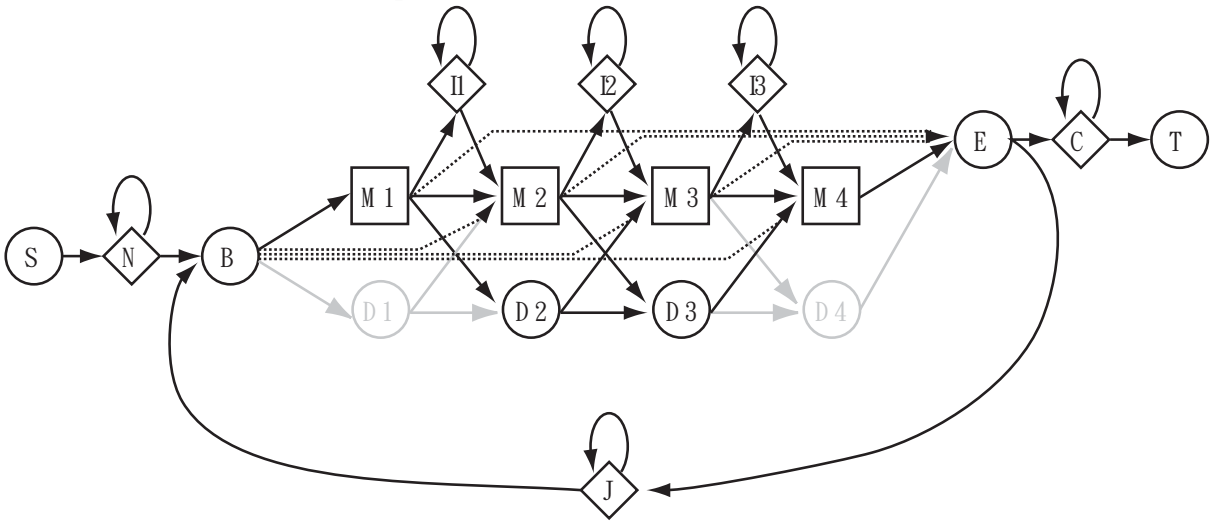

圖 3.1: Plan7 結構

HMM 是一個對於內文相關的序列統計分析相當有力的方法。Profile HMMs 參數的權重給定來自於多重序列比對，其演算法與 Plan7 架構的搭配可以說非常迅速而且乾淨俐落，然而 profile HMMs 由於參數的權重給定演算法和架構的限制有一些缺點，其中一個缺點是 HMMer 在訓練時無法直接處理長度不一的字串，簡單來說就是 HMMer 輸入訓練的資料必須是全部字串要都一樣長，若不一樣長則需要事先做多重序列比對，插入 gap 將所有字串對齊後才可以丟進去訓練，而在本論文提出的 iHMM 是不須要有這個限制，本論文提出的 HMM 訓練時可以接受長度不一的字串。另一個缺點是 Profile HMMs 是屬於 HMM-1 不能捕捉到微弱的相關訊號(high order dependency signals)。上圖 3.1 來說，從 M3 走到 M4 在 HMM-1 裡是已經把資訊壓縮掉，也就是不考慮前狀態，若我們把 M3→M4 的訊號放大來看，到底他之前的路徑 I2→M3→M4 還是 M2→M3→M4 還是 D2→M3→M4，然而在資料夠複雜的情況下，這些資訊的保留有時是很有用的。

3.3 Second-order HMM

HMM-2 是 HMM-1 的擴展，它與 HMM-1 的差別在於狀態轉移機率，在 HMM-1 中狀態轉移機率只有目前狀態與前一個狀態的關係。而在 HMM-2 中狀態轉移機率是目前狀態與前二個狀態的關係，它的狀態轉移機率定義是：$A \equiv \{a_{ijk}\}$，$a_{ijk} \equiv P(q_t = S_k \mid q_{t-1} = S_j, q_{t-2} = S_i)$ 此狀態所有 $1 \leqq i \leqq N$, $1 \leqq j \leqq N$ 轉移機率都要符合 $\sum_{k=1}^{N} a_{ijk} = 1$ 這個機率限制，這裡的 N 表示狀態數目，q_t 表示在時間 t 的狀態。如果以維度來看，HMM-1 的狀態轉移機率是一個二維的矩陣 $N \times N$，而 HMM-2 狀態轉移機率是一

個三維的矩陣 $N \times N \times N$ 。每一個 HMM-2 其實等價於二維狀態 SxS 空間展開的 HMM-1，但是它不像 HMM-1 會增加狀態數目，如果有足夠的 order ，訓練資料所沒有走過的路徑，在測試資料中也將不會出現，這可以有助於區別錯資料的能力。如下圖，在這裡我引用 1997 年 Jean-Francois Mari 等人在論文中 [17] 所舉的例子，圖 3.2 是 HMM-2 的結構，圖 3.3 是與圖 3.2 結構等價的 HMM-1，這兩個結構的符號觀測機率有其相同對印，也就是 S01 等於 S1，S12 與 S22 等於 S2，S23 與 S33 等於 S3，S34 等於 S4。待會我們會發現利用更高階的轉移機率來減少狀態數目，在訓練和預測時會有他的優勢。

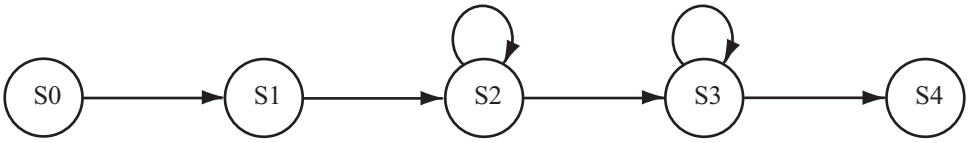

圖 3.2:原始 HMM-2

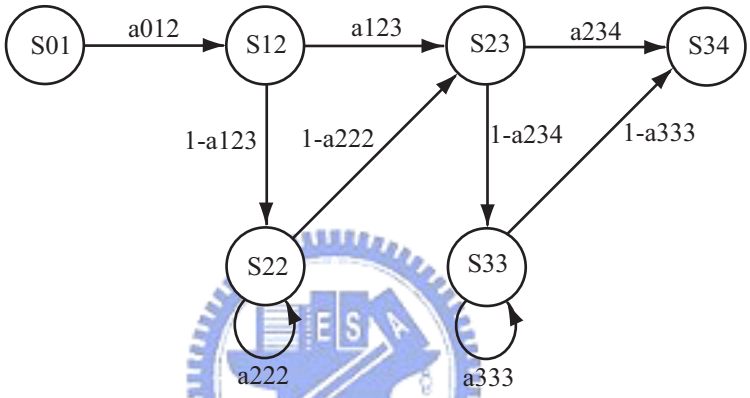

圖 3.3: HMM-1 等價模型

在語音辨認上已有越來越多的論文指出連續型的 HMM-2 的效能要比 HMM-1 好的多[18, 19]。生物資料這幾年正迅速的增加，higher-order 的統計模型也將越來越重要。當資料足夠多時 higher-order HMM 可以解決 HMM-1 所不能辨到的事情，因為 higher-order HMM 的轉移機率不只會參考前一個走過的狀態，而且還會考量過去之前二個以上所走過的狀態，higher-order HMM 可以比傳統 HMM-1 以更少的參數表達更複雜的模型，我們希望透過這種更複雜的前後文資訊(contextual information)幫助我們找出哪些蛋白激酶在哪些位置起作用。

3.4 分數計算

分數的計算是用來做為未知蛋白質序列判定接受或拒絕的標準，高於門檻值即接受，低於門檻值即拒絕。若 HMM 的參數 λ 全部都已知，給一條序列 $O=<O_1,O_2,\ldots,O_T>$ 與相對應一條所走的路徑 $Q=<q_1,q_2,\ldots,q_T>$，在 HMM-1 中我們可以很快的計算這條路徑下的機率：

$$P(Q \mid \lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T} \tag{1}$$

$$P(Q, O \mid \lambda) = \pi_{q_q} b_{q_1}(O_1) \prod_{t=2}^{T} a_{q_{t-1} q_t} b_{q_t}(O_t) \tag{2}$$

同樣的我們也可以很快的算出 HMM-2 一條路徑下的機率：

$$P(Q \mid \lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_1 q_2 q_3} \cdots a_{q_{T-2} q_{T-1} q_T} \tag{3}$$

$$P(Q, O \mid \lambda) = \pi_{q_q} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \prod_{t=3}^{T} a_{q_{t-2} q_{t-1} q_t} b_{q_t}(O_t) \tag{4}$$

然而我們要把一條序列所有可能走過的路徑的機率累加起來才是我們要的 $P(O \mid \lambda)$ 機率，此機率的計算是 HMM 中第一個基本問題，要計算這個機率我們可以由動態規劃(dynamic programming) 方法來快速求解，alpha 函數的計算待會在 3.5 節會有詳細的算法：

$$P(O \mid \lambda) = \sum_{Q} P(O \mid Q, \lambda) P(Q \mid \lambda) \tag{5}$$

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{6}$$

這裡我們在計算分數時與傳統的 Profile HMMs一樣 HMM-1 與iHMM均是採用log-odds的分數算法，log-odds 分數是用來評估這條序列相對於虛無模型 (null model)多少機率是由HMM產生。這個分數又稱做為log-likelihood ratio，這個公式描述在公式(7)。在此公式中 $P(O \mid \lambda_P)$ 是序列在positive HMM 中的機率。$P(O \mid Null)$ 是序列在自然界中的機率。在 HMMer 中，虛無模型是如同簡化一個狀態的 HMM，狀態裡的符號觀測機率是由已經統計好的自然界胺基酸出現的頻率。在我們的 iHMM 中所使用的自然機率，例在表 3.1，此自然機率表來自於 HMMer 原始碼。

$$score = \log \frac{P(O \mid \lambda)}{P(O \mid Null)} \tag{7}$$

表 3.1:自然界胺基酸出現的頻率

胺基酸代號	自然機率	胺基酸代號	自然機率	胺基酸代號	自然機率	胺基酸代號	自然機率
A	0.083534	G	0.070123	M	0.023504	S	0.070339
C	0.014532	H	0.022714	N	0.041622	T	0.055862
D	0.052473	I	0.0572	P	0.05025	V	0.065857
E	0.061821	K	0.052905	Q	0.040249	W	0.013155
F	0.039814	L	0.098249	R	0.055901	Y	0.029897

score 輸出是一個機率分數，表示 HMM 的機率相對於自然界中的機率兩者的距離，高的分數表示序列是傾向是會被磷酸化，反之底的分數表示序列是傾向是不會被磷酸化。如果一個新的序列進來得到的分數比給定的門檻值還高，則我們說這條序列是positive，反之這條序列是 negative 。

3.5 訓練演算法

我們令 Σ 是胺基酸字母的集合. |Σ| 是胺基酸字母的總數, 假設胺基酸的號碼是從 1 到 20，則標示的 |Σ| = 20. 我們令 positive 抽樣本 $O^P \equiv (O^{P(1)}, O^{P(2)}, O^{P(3)}, \cdots, O^{P(k)})$ 表示 k 條序列，並且假設每條序列都相互獨立並且具有相同分配(independent and identically distributed , i.i.d.)，其中 $O^{P(k)} \equiv < O_1^{P(k)}, O_2^{P(k)}, O_3^{P(k)}, \cdots, O_{Tk}^{P(k)} >$ 表示第 k 條由字母組成長度為 T_k 的觀查序列 (observation sequence)，其中 $O_{Tk}^{P(k)} \in \Sigma$。而在已知的 HMM 參數 λ≡(A,B,π)之下，得到的聯合機率密度函數，稱之為 likelihood

$$P(O \mid \lambda) = \prod_{i=1}^{N} P(O_i \mid \lambda) \tag{8}$$

最佳化機率模型的參數來精準描述所觀查到的序列是非常重要的。在這裡訓練的問題是要挑一組 HMM 參數 λ≡(A,B,π) 使得最大化抽樣資料的概似率(likelihood) ，最大概似率是求得一估計式使其抽樣資料的聯合機率密度函數的值為最大。正規的定義是

$$\lambda^* = \arg\max_{\lambda} \prod_{i=1}^{N} P(O^{P(i)} \mid \lambda) \tag{9}$$

其中期望值最大化演算法(expection maximization algorithm, EM algorithm)是一個廣泛用來解決最大概似率估計的(maximum likelihood estimation, MLE) 技術。它的廣義步驟如下：

步驟 1:隨機挑選一個初始參數 λ

步驟 2:E-step，求出輔助函數 (auxiliary function) $Q(\lambda, \overline{\lambda}) = \sum_{Q} P(Q \mid O, \lambda) \log\left[P(O, Q \mid \overline{\lambda})\right]$

步驟 3:M-step，根據訓練資料以步驟 2 的輔助函數計算 $\overline{\lambda}$ 期望值，$\overline{\lambda} = \max_{\overline{\lambda}} Q(\lambda, \lambda)$

步驟 4:以新的參數取代舊的參數 $\lambda = \overline{\lambda}$ ，然後重複步驟 2 跟 3，直到收斂為止

EM 演算法目前依然是解 MLE 最有效的方法，它保證每一次的疊代（iteration）後一定是 $P(O \mid \hat{\lambda}) \geq P(O \mid \lambda)$ 。在 HMM-1 裡，步驟 2 中輔助函數已被 Baum 和它的同僚所解出。它的詳細演算法在 Rabiner [14]已有很好的整理，在這裡我直接引用 Rabiner 論文裡所整理的 Baum-Welch 參數重估演算法，Baum-Welch 演算法主要是運用 forward probability (alpha 函數) 跟 backward probability (beta 函數) 來進行參數重估，forward probability 跟 backward probability 都可以用動態規劃(dynamic programming) 方法來求解，為了要方便計算我們另外還要 gamma 函數、跟 xi 函數。它們的完整公式如下

(1) alpha 函數

說明：alpha 函數，通稱為 forward probability，表示在模型 λ 底下，看到 $O_1O_2 \ldots O_t$ 並且於時間 t 時停在狀態 i 之下的機率。而我們剛才提到的 $P(O \mid \lambda)$ 的機率可以由此 forward 演算法求得，即 $P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$

定義：$\alpha_t(i) = P(O_1 O_2 \cdots O_t, q_t = s_i \mid \lambda)$

初始值：$\alpha_1(i) = \pi_i b_i(O_1)$,1≦i≦N

遞迴式：$\alpha_{t+1}(j)=\sum_{i=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})$

(2) beta 函數

說明：beta 函數，通稱為 backward probability，表示在模型 λ 底下，已知時間 t 時停在狀態 i 之下，看到 $O_{t+1}O_{t+2}\ldots O_T$ 的機率。

定義：$\beta_t(i)=P(O_{t+1}O_{t+2}\cdots O_T \mid q_t=s_i,\lambda)$

初始值：$\beta_T(i)=1,1\leqq i\leqq N$

遞迴式：$\beta_t(i)=\sum_{j=1}^{N}a_{ij}b_j(O_{t+1})\beta_{t+1}(j)$,t=T-1,T-2,…,1,$1\leqq i\leqq N$

(3) gamma 函數

說明：gamma 函數，通稱為 forward-backward probability，表示在模型 λ 底下，時間 t 時停在狀態 i 之下的機率。

定義：$\gamma_t(i)=P(q_t=S_i \mid O,\lambda)$

公式：$\gamma_t(i)=\dfrac{\alpha_t(i)\beta_t(i)}{P(O\mid\lambda)}$

(4) xi 函數

說明：表示在模型 λ 底下，在時間 t 由狀態 i 轉移到狀態 j 的機率。

定義：$\xi_t(i,j)=P(q_t=s_j,q_{t+1}=s_k \mid O,\lambda)$

公式：$\xi_t(i,j)=\dfrac{\alpha_t(i,j)a_{ij}b_k(O_{t+1})\beta_{t+1}(j,k)}{P(O\mid\lambda)}$

而由上面式子，我面又可以整理如下式

$\gamma_t(i)=\sum_{j=1}^{N}\xi_t(i,j)$

Baum-Welch 參數重估演算法，即藉由上例 alpha、beta、gamma 跟 xi 函數，進行以下計算，並且以疊代方式，不斷重估 HMM 的參數，直到收斂為止。

$$\overline{\pi_i}=\frac{\gamma_1(i)}{\sum_{i=1}^{N}\gamma_1(i)} \tag{10}$$

$$\overline{a_{ij}}=\frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \tag{11}$$

$$\overline{b_j(k)}=\frac{\sum_{\substack{t=1\\ s.t.O_t=v_k}}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)} \tag{12}$$

在 HMM-2 這裡我們使用推廣的 Baum-Welch 演算法，來做為我們訓練演算法[20]。為了要方便計算我們另外還要推廣的 gamma 函數、eta 函數跟 xi 函數。它們的完整公式如下

(1) 推廣的 alpha 函數

說明：alpha 函數，通稱為 forward probability，表示在模型 λ 底下，看到 $O_1O_2...O_t$ 並且於時間 t-1 跟時間 t 時分別停在狀態 j 跟狀態 k 之下的機率

定義：$\alpha_t(j,k) = P(O_1O_2\cdots O_t, q_{t-1} = s_j, q_t = s_k \mid \lambda)$

初始值：$\alpha_1(i,j) = \pi_i b_i(O_1) a_{ij} b_j(O_2), 1 \leqq i \leqq N$

遞迴式：$\alpha_t(j,k) = \sum_{i=1}^{N} \alpha_{t-1}(j,k) a_{ijk} b_k(O_{t+1})$

(2) 推廣的 beta 函數

說明：beta 函數，通稱為 backward probability，表示在模型 λ 底下，已知時間 t-1 和時間 t 分別停在狀態 i 跟狀態 j 之下，看到 $O_{t+1}O_{t+2}...O_T$ 的機率

定義：$\beta_t(i,j) = P(O_{t+1}O_{t+2}\cdots O_T \mid q_{t-1} = s_i, q_t = s_j, \lambda)$

初始值：$\beta_T(i,j) = 1$

遞迴式：$\beta_t(i,j) = \sum_{i=1}^{N} \beta_{t+1}(j,k) a_{ijk} b_k(O_{t+1})$

(3) 推廣的 eta 函數

定義：$\eta_t(i,j,k) = P(q_{t-1} = s_i, q_t = s_j, q_{t+1} = s_k \mid O, \lambda)$

公式：$\eta_t(i,j,k) = \dfrac{\alpha_t(i,j) a_{ijk} b_k(O_{t+1}) \beta_{t+1}(j,k)}{P(O \mid \lambda)}$

(4) 推廣的 xi 函數

說明：表示在模型 λ 底下，在時間 t 由狀態 i 轉移到狀態 j 的機率。

定義：$\xi_t(i,j) = P(q_t = s_j, q_{t+1} = s_k \mid O, \lambda)$

公式：$\xi_t(i,j) = \sum_{k=1}^{N} \eta_t(i,j,k)$

(5) 推廣的 gamma 函數

說明：gamma 函數，通稱為 forward-backward probability，表示在模型 λ 底下，時間 t 時停在狀態 i 之下的機率。

公式：$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$

推廣的 Baum-Welch 參數重估演算法，即藉由上例推廣的 alpha、beta、gamma 、xi 跟 eta 函數，進行以下計算，並且以疊代方式，不斷重估 HMM 的參數，直到收斂為止。

$$\overline{\pi_i} = \gamma_1(i) \tag{13}$$

$$\overline{a_{ijk}} = \frac{\sum_t \eta_t(i,j,k)}{\sum_{k,t} \eta_t(i,j,k)} \tag{14}$$

$$\overline{b_j(k)} = \frac{\sum_{\substack{t=1 \\ s.t. O_t = v_k}}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{15}$$

第四章 提出 iHMM 方法論

4.1 想法起源

HMMer 的演算法無論是在參數的權重給定演算法或是使用的 Plant7 結構，都具有其生物意義，然而我們在實驗中發現，HMMer 在小量資料時，常常會選到太長的結構，結果因為狀態太多，資料量太少，而導致過適化。用有限的資訊對未知現象作判斷與分析乃是統計學所面臨的必然問題。而 HMM-2 的特性正好可以利用更多的轉移機率來進行節省狀態的使用。我提出 iHMM 是將 HMM-2 加入二種機制，第一是並配合簡易解參數最佳化問題，第二是加上 BIC 的機制來防止訓練中過適化的問題。另外我們 iHMM 在設計上可以很方便進行第二類資料的訓練與分數計算。

本論文所提出的 iHMM 在訓練階段時不需要多重序列比對而且沒有結構限制，缺點是它需要花一些時間來訓練。然而，現在電腦速度這幾年越來越快，計算速度是越來越不重要的。我們提出的 iHMM 要比 profile HMMs 有更多的優點。其中一個優點是由完全圖可以捕捉到更多序列內文微弱的相關訊號。另一個優點是在少量的資料中，iHMM 可以比 profile HMMs 能更防止過適化的問題，基於這些原因，我們在測試資料上可以提升辨識率。

4.2 iHMM 之設計

Baum-Welch 演算法是一個眾所都知的用來解 HMM 問題的演算法，然而，在解空間中可能存在許多的區域最佳解。Baum-Welch 演算法不保證可以找得到全域最佳解。Baum-Welch 演算法對於一開始給定的初始點是很重要的，因為 Baum-Welch 演算法會與一開始參數的設定而掉入不同的區域最佳解

本論文提出一個族群搜尋式的 Baum-Welch 演算法。其主要的想法是使用多點式搜尋來取代單點搜尋。它將有更多跳脫區域最佳解的能力，這個演算法的虛擬程式碼如下：

Population-based Baum-Welch algorithm

Input: Training set, initial barrier ω, and population size k
Output: HMM model
Begin

$S=\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$

for i=1 to $|S|$
randomize $S(i)$ parameters
end i
while ($|S|>1$)
{
δ← 0
for i=1 to $|S|$
    do
    {
        $S(i)$ do E-step and M-step
        Δ ← (new likelihood of $S(i)$- old likelihood of $S(i)$)
    }
    while(Δ not converge to ω);
    δ←δ + Δ;
end i
ω←δ
sort S by likelihood
remove the last half λ_i of S, i= $|S|/2$ to $|S|$
}
return $S(1)$

End

這裡我們對演算法做一個描述，S 是一組 HMMs 候選解的集合，k 是一開始要灑的粒子數，在 7 到 9 行中，我們是用亂數隨機在解空間進行灑點，k 越大將越可以找到全域最佳解，但也相對要花越多時間搜尋，k 值是完全依據使用者自己選定。在 10 到 25 行 while loop 的迴圈每一次的重復，我每放入柵欄 ω 來同時控制所有點的行動。所有點的都由 Baum-Welch 演算法移動，並且當 likelihood 收斂到柵欄 ω 值則停下來。在演算法的第 16 行，E-step 和 M-step 是來自於 EM 演算法，它的補助函數 Q-step 已經由 Baum 和它的同僚所解出，詳細算法可以參考第三章 3.5 節的訓練演算法的部份，在演算法的第 22 行，我們每一代都會調控柵欄 ω 值，在演算法的第 24 行，為了加快搜尋速度，我們以 likelihood 做排名，然後只保留前半部目前較好的點，拿掉後面一半比較差的點，以減少計算時間，因為這些較差的點將來要再比最好的還要好的機率很小。然後一直重復直到只剩下只有一個點。換句話說，每一個點必須為了生存而競爭，這個如同達爾文的適者生存，不適者淘汰的觀念。這個演算法可以很方便改成平行處理，我們相信這個方法是很快的，可以幫助加速收斂到全域最佳解，而且很穩定。

在圖形識別(pattern recognition)課題上，過適化(over fitting)的現象一直是最令人棘手的問題，也是嚴重影響辨識率的主要原因之一。要挑選多少個狀態下去訓練是一個很難的問題，因為 HMM 會隨著狀態數目越多而會越有過適化現像，如下圖 4.1，我們在訓練時只能看的見訓練資料的正確率，就是圖 4.1 中實線的部份，虛線的變化是我們看不到的部份，然而我們必須要選擇正確的模型才可以最小化錯誤預測能力。

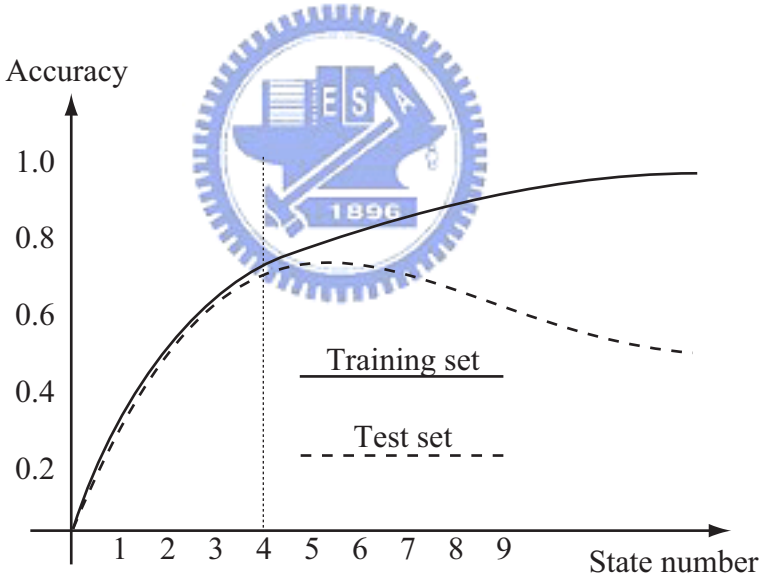

圖 4.1:狀態數目與辨識率之間的關係的示意圖

過適化的現象原因出在於當我們的訓練資料抽樣不平均，或是不夠足以代表母體樣本分佈，或著是資料量少於相對於模型的自由度，MLE的方法會讓我們很容易 fit 某一類的訓練資料，這就是為什麼說HMM通常會特別偏愛某些條序列。原因是雖然都是在同類 (positive) 的序列，但就是有些序列從HMM出來的機率特別的高，他們的分數會把孤立的序列分數拉下來，會導致系統不對稱，使得訓練 出來的 HMM profile 將來會特別偏愛某些條序列。有些論文指出用 $\max_{\lambda} \min_{i} P(O^i \mid \lambda)$ 的方法可能會比傳統

$\max_{\lambda} \prod_{i=1}^{N} P(O^i \mid \lambda)$ 的方法來的好，以用來避免過適化的問題[15, 21]。在 HMM 裡為了避免MLE的方法因為少量的訓練資料遭到過適化的問題。另一個方式則是採用最大事後機率(maximum a posteriori probability estimation, MAP)，取代MLE的方法，作為HMM 之評估準則，假設每個線性組合係數都是隨機且其事前機率為一個高斯分佈，我們可以推導出一組最大事後機率估測值。它的優點是可以避免因資訊的不足，而使得估測產生過大的誤差。根據貝氏定理，我們的事後機率分佈如公式 (16):

$$P(\lambda \mid O) = \frac{P(O \mid \lambda)P(\lambda)}{P(O)} \tag{16}$$

不幸的是，在公式 (16)這個 $P(\lambda)$ 事前機率很難求得。而在HMMer裡則使用混入 Dirichlet 事前機率[22]，並使用 Blocks9 表裡的芳香族、脂肪族，有極性，沒極性・・・等九個自然界胺基酸事前機率單元，這個混入Dirichlet 事前機率的效用如同模擬了很多與訓練資料可能的演化樣本，用來平滑化(smooth) HMM的參數，使得訓練出來的 HMM 不會遭到過適化。

廣義線性模型(generalized degrees of freedom)是傳統線性模型的延伸，使用 maximum likelihood 方法來估計參數，並假設反應變數的機率分佈屬於指數家族（exponential family），將廣義自由度的想法由常態分布推廣到更廣泛的指數族群分布，並研究廣義線性模型（generalized linear model）的變數選取問題。不同於傳統的狀態數目選取方式，我們不直接選取狀態數目，而是以廣義自由度方法選取具不同懲罰(penalty)參數的資訊準則（information criterion），然後再依照所選出的準則執行 HMM 狀態數目選取。我們用自由度的觀念修正變數選取過程對推論結果所產生的偏差。這個方法可隨資料背後的真實模型而自動調適，因而兼顧。

基本上有兩種公認標準的衡量方法叫做艾凱克資訊準則(Akaike information criterion, AIC) [23] 和貝氏資訊準則(Bayesian information criterion, BIC) [24]。而在磷酸化預測問題中，本論文首先提出離散型的 HMM-2 加上 BIC 模型選擇方法，來取代傳統 HMM-1 以解決生物資訊的問題。AIC與 BIC在公式上很相像，差別在於採用不同的懲罰方式，BIC在懲罰的地方則是比AIC多了一個訓練的資料個數的代價倍率，所以在這裡我們採用 BIC 來做為我們挑選 HMM 狀態數目的衡量準則。BIC 定義如式(17)：

$$BIC = -\log(L) + 0.5 \times K \times \log(N) \tag{17}$$

L 是統計模型的 likelihood 值
統計模型總共所使用的參數個數，或自由度
N 是所提供訓練的資料個數

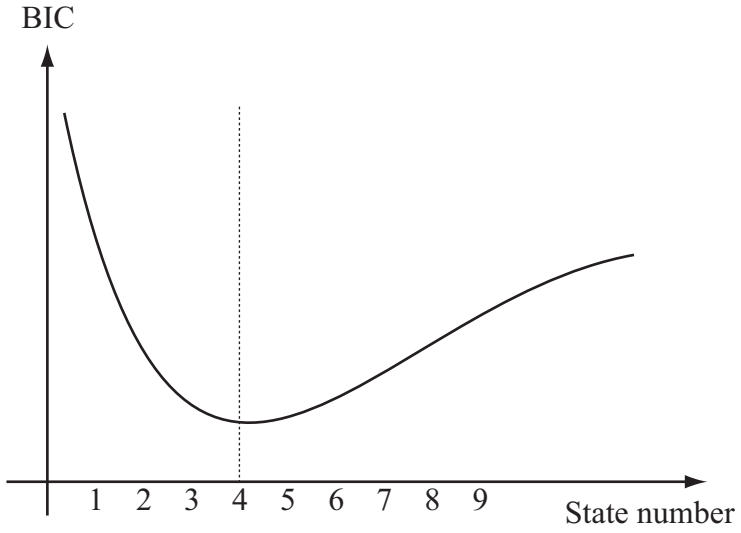

圖 4.2:狀態數目與 BIC 的示意圖

BIC 主要概念是運用少量參數，來建立一個統計模型，使它更具有強健性。如上圖4.2，我們的狀態數目選取，則是採用 BIC 最低點的位置。用 BIC 來選擇模型，除了用在 HMM-2 上，在許多文獻上 HMM-1 用 BIC 來選擇模型也有一也不錯的例子，像是要最佳化以微陣列基因表現的 HMM 模型[25]，以及最佳化預測蛋白質二級結構的 HMM 模型[26]等等。另外 1999 年 Brand 也發表了一篇論文[27]，他在文章提到說，將結構的一個參數趨近於零，可以減少模型一個自由度，並且減低模型的亂度(entropy)，模型的亂度越低，越有助於抵抗過適化的問題。同時參數使用的越少，將可以使得模型越可解讀。模型參數量的使用，其實跟信息熵有關，就好比將一份資料壓縮，我們希望在仍足以表達某程度資訊內容條件下，進行減少該模型的參數使用。

在語音辨識中，常會因為聲音受到背景雜音干擾而降低辨識率，於是人們常需要從不的錄音環境，如教室，車內或是室外中，將背景雜音與目標語者的聲音分離出來，所以要提高辨識率的一個方法就是另外再從背景聲音建一個模型。生物資料中也有相同的情況，positive 資料就好比是目標語者的聲音，而negative 資料就好比是背景雜音。理論上來說，隨著訓練樣本類別數的增加，辨識結果可以更加近似於實際結果。假設從positive 建立iHMM 統計模型以符號 λ_P來表示，而從negative 資料建立iHMM 統計模型，以符號 λ_N來表示，如果 negative 資料足夠亂，則 $P(O_i \mid \lambda_N)$ 就成為常數，我們就只要最大化 $\sum_i \log P(O_i \mid \lambda_P)$ 即可。相反的如果negative 資料不夠亂，則訓練negative的資料模型則會有用。兩個模型的分離成度可以用相對熵(relative entropy)來解釋，即有名的Kullback-Leibler (KL)量度，KL是用以量測兩個機率分佈的分離成度 (divergence)，它的定義在式(18)。

$$KL(\lambda_P, \lambda_N) = H(\lambda_P \,||\, \lambda_N) = \sum_i P(O_i \mid \lambda_P) \log \frac{P(O_i \mid \lambda_P)}{P(O_i \mid \lambda_N)} \tag{18}$$

然而KL並不是真正的距離公式，因為他在數學上不滿足交換率，如式 (19)，而在Matthew A. Siegler 的[28]論文中，則是把 KL 寫成對稱形式，如式(20)，即成為兩個模型的距離。我們就可以用 KL2 來衡量positive 模型是否與negative 模型是否有存在區別。

$$KL(\lambda_P \,||\, \lambda_N) \neq KL(\lambda_N \,||\, \lambda_P) \tag{19}$$

$$KL2(\lambda_P, \lambda_N) = KL(\lambda_P, \lambda_N) + KL(\lambda_N, \lambda_P) \tag{20}$$

第五章 實驗資料與結果討論

5.1 實驗資料

本實驗資料來原來自於 Phospho.ELM [5]資料庫第六版，這個資料庫是從 PhosphoBase 和 Swiss-Prot 資料庫裡依據現有的文獻做磷酸化注解，並且我們又從 2007 年最新的 Swiss-Prot 資料庫中收集了完整序列資料，值得注意的是在資料取得中 Swiss-Prot 有三種非實驗證據的狀態，分別是"potential", "probable"跟 "by similarity"。這些狀態有的是從相似序列比對出來的資料，尚未經過實驗驗證，我們為了確保資料的正確性，我們只將有標示實驗證據(evidence)的序列例為我的 positive 資料。我們將這兩種資料庫結合，去掉交集的部分，我們將 positive 和 negative 資料分別來自於資料庫有磷酸化注解，和沒有磷酸化注解。然而沒有磷酸化注解，不代表一定不會磷酸化，因為有可能是目前沒有合理和完整的證據或是缺乏實驗證據可以注解，所以目前不存在不會磷酸化地方資料庫。這點將有可能成為資料中的雜訊。磷酸化位置的生物化學的特性主要依靠附近鄰居胺基酸。在我的訓練資料是採用磷酸化位置絲氨酸(S)、蘇氨酸(T)或酪氨酸(Y)中間的地方取出左右兩邊範圍各 7 個殘基，總長度 15，來做為訓練資料。

5.2 實驗結果

我們已將詳細的實驗數據放在附錄中，並且每個蛋白激酶的資料均以 1:1 等比例的 positive 和 negative 做為計算，標準差是以括弧括號起來。我們以 30 次 5 等份交叉驗證（5-fold cross-validation, 5-CV）來比較 HMM-1、HMMer 和 iHMM 兩個演算法的效能。所謂 5-CV 是將原始資料均勻切成五等份，每次輪流用其中的四等分做為訓練資料，訓練完畢後，再以剩下的一第分做為測試資料，5 次做完之後，將 5 次測試資料的辨識率平均，及為此系統的辨識率。附錄一中的每條曲線均是 30 次 5-CV 平均後的結果（也就是 30x5 = 150 次實驗的平均），為了公平比較 HMM-1、HMMer 和 iHMM 的效能，全部都只訓練 "Positive" 這一類。我們例出兩種門檻值的結果，並以代號 δ_1 和 δ_2 來表示，δ_1 表示門檻值取訓練資料正確率最高的那一點來當測試資料的門檻值，δ_2 表示門檻值取測試資料正確率最高的那一點來當測試資料的門檻值。我將 δ_1 的門檻值下的正確率整理在表 5.1。所有數據都四捨五入到小數以下第四位。而在 PKG (S)，CK1 (S)跟 CK2 (S) 這三組資料由於 HMM-1 跟 iHMM 在 BIC 的選擇下都只用到一個狀態，由於在一個狀態中 first-order 跟 second-order 並無區別，所以這組資料的正確率和標準差是一樣的。

在表 5.1 中 HMMer 的平均正確率是 0.7439，HMM-1 的平均正確率是 0.7508，而我們提出的 HMM-1 的平均正確率是 0.7871。另外我們用標準的配對 t-test 來檢驗比較兩個演算法的 30 次 5-CV 正確率樣本的平均值之間是否存在顯著差異，若以 X_1 表示 A 演算法的正確率平均值，以 X_2 表是 B 演算法的正確率平均值，其虛無假設為 $H_0 : \overline{X}_1 = \overline{X}_2$，t-test 的機率與我們拒絕了該假設有關。統計學上 t-test 值小於 0.05 表示存兩組觀察樣本存在顯著差異，相反的則是不存在顯著差異。而我們可以看到 iHMM 在 PKA (S)，PKC (S)，CaM-KII (S)，CK1 (S)，CDK (S)，MAPK (S)，ATM (S)，PKA (T)，PKC (T)，CK2 (T)，CDK (T)，EGFR (Y)和 INSR (Y) 等資料相對於 HMMer 都有顯著改善。

表 5.1:效能比較表

殘基	激酶（資料筆數）	方法 正確率（標準差）			HMMer 與 HMM-1 的 T-test	HMMer 與 iHMM 的 T-test
		HMMer	HMM-1	iHMM		
S	PKA (308)	0.8250 (0.0089)	0.8319 (0.0091)	0.8444 (0.0063)	0.0034	6.92E-12
	PKB (81)	0.8579 (0.0169)	0.7942 (0.0202)	0.8561 (0.0160)	8.41E-14	0.6779
	PKC (304)	0.7419 (0.0121)	0.7448 (0.0081)	0.7624 (0.0080)	0.2890	9.59E-9
	PKG (30)	0.7072 (0.0373)	0.6983 (0.0320)	0.6983 (0.0320)	0.2571	0.2571
	CaM-KII (76)	0.6926 (0.0177)	0.6400 (0.0270)	0.7473 (0.0217)	1.5E-9	6.69E-12
	CK1 (49)	0.6074 (0.0301)	0.7487 (0.0234)	0.7487 (0.0234)	1.15E-18	1.15E-18
	CK2 (243)	0.8316 (0.0095)	0.8297 (0.0059)	0.8334 (0.0097)	0.3120	0.4622
	CDK (195)	0.8237 (0.0081)	0.8137 (0.0119)	0.8414 (0.0111)	0.0003	1.15E-7
	MAPK (207)	0.8024 (0.0122)	0.7887 (0.0145)	0.8228 (0.0090)	0.0002	3.65E-8
	ATM (77)	0.8628 (0.0280)	0.8257 (0.0155)	0.9234 (0.0223)	3.25E-7	9.43E-11
T	PKA (39)	0.6358 (0.0281)	0.7080 (0.0371)	0.8037 (0.0314)	5.71E-10	1.66E-21
	PKC (71)	0.6467 (0.0213)	0.6518 (0.0154)	0.6670 (0.0290)	0.2892	0.0010
	CK2 (42)	0.6510 (0.0211)	0.7904 (0.0234)	0.7904 (0.0234)	8.46E-23	8.46E-23
	CDK (113)	0.8493 (0.0196)	0.8127 (0.0114)	0.8810 (0.0147)	3.35E-11	7.55E-8
	MAPK (81)	0.8386 (0.0164)	0.8290 (0.0105)	0.8389 (0.0213)	0.0172	0.9424
Y	EGFR (46)	0.6223 (0.0247)	0.6549 (0.0178)	0.7059 (0.0446)	1.06E-5	4.76E-10
	INSR (58)	0.6730 (0.0195)	0.6780 (0.0223)	0.7017 (0.0286)	0.4213	2.57E-6
	SRC (143)	0.7219 (0.0161)	0.6741 (0.0135)	0.7013 (0.0180)	1.6E-18	4.07E-5
平均		0.7439	0.7508	0.7871		

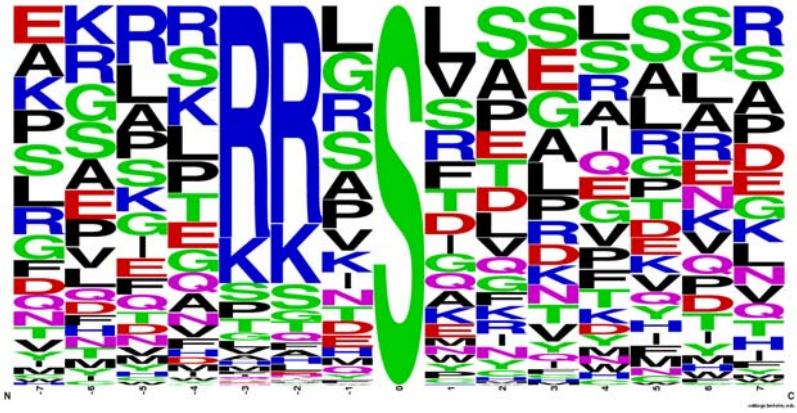

圖 5.1: PKA sequence logos

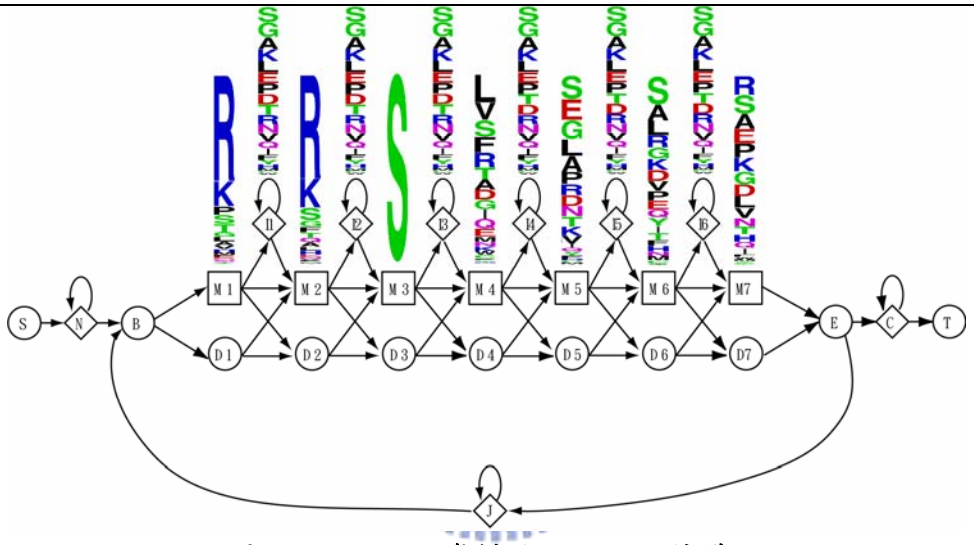

圖 5.2: PKA (S)資料的 HMMer 結構

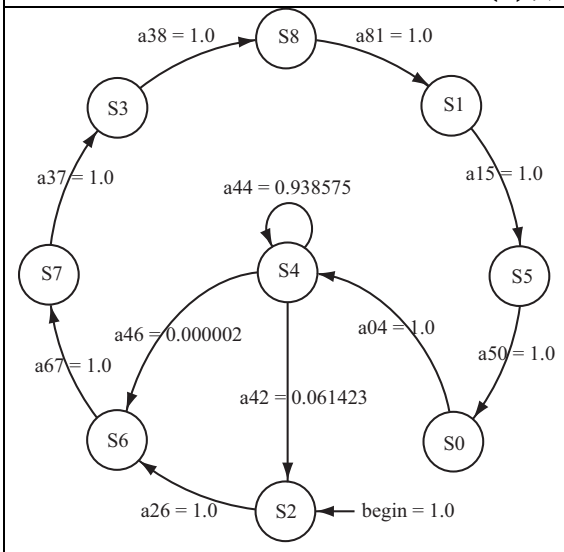

圖 5.3: PKA (S)資料的 HMM-1 結構

圖 5.4: PKA (S)資料的 HMM-1 結構的符號觀測機率圖

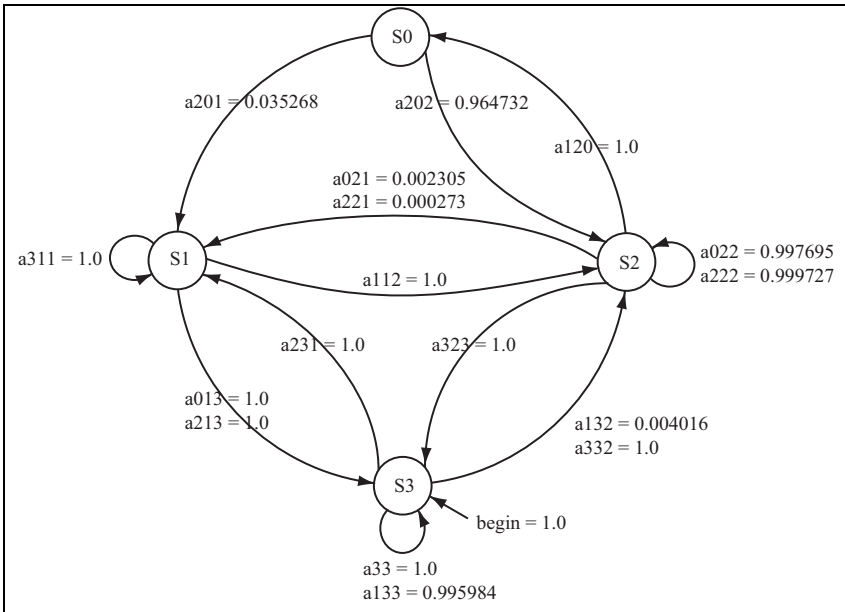

圖 5.5: PKA (S)資料的 HMM-2 結構

圖 5.6: PKA (S)資料的 HMM-2 結構的符號觀測機率圖

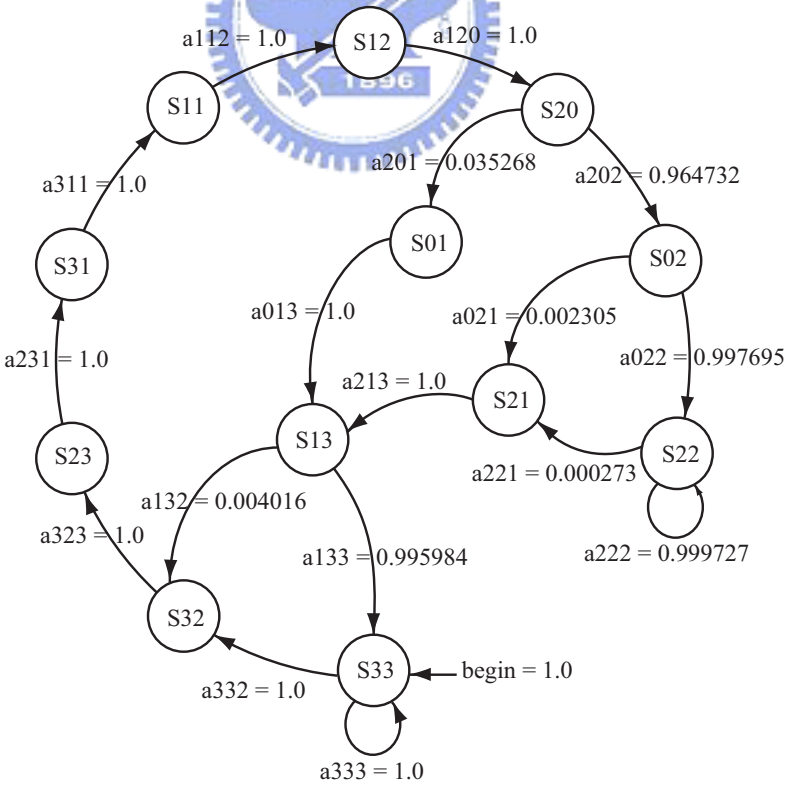

圖 5.7: HMM-1 等價展開圖

在這裡我們以 PKA 資料做為此章節的實驗結果的討論，我們將 HMMer 以實際資料訓練出來的結構圖放在圖 5.2，它的狀態內部的符號觀測機率 (observation probability) 已經在狀態的上面，另外也將 HMM-1 和 iHMM 的結構圖分別展現在圖 5.3 跟圖 5.5，而它們對應的狀態內部的符號觀測機率分別放在右邊的圖 5.4 跟圖 5.6，為了方便讀者閱讀，我已將符號觀測機率轉換成序列圖案(sequence logos)。圖 5.1 是 PKA 資料的序列圖案，左右兩邊各取 7 個殘基，中間絲氨酸(S)的地方歸零。圖 5.7 是圖 5.5 的 HMM-1 等價展開圖，其中狀態 S20 的符號觀測機率等於 S0；狀態 S01、S11、S21、S31 的符號觀測機率等於狀態 S1；狀態 S02、S12、S22、S32 的符號觀測機率等於 S2；狀態 S13、S23、S33 的符號觀測機率等於 S3，雖然兩個結構等價，但一般來說，以 HMM-1 沒有機會訓練出圖 5.7 的結構，因為很難找出一個好的模型選擇方式，如果以 BIC 來計算時，HMM-2 只用到 4 個狀態，其中狀態內部的符號觀測機率，S0 的符號觀測機率只用到一個參數 S=1.0 ，而其它三個狀態 state1,2,3 共用了 20*3 =60 個參數，HMM-2 的架構用了 18 個參數，總共用了 1+60+18 = 79 個參數，但如果 HMM-1 要訓練出圖 4.7 的結構時，則用到了 1 + 20*11 + 18=239 個參數，我們再回顧之前的 BIC = -log(L)+0.5*K*log(N) 公式，我們可以看到它的 K 暴增很多，它最多只會訓練到如圖 5.3 的結構。由此我們可以知道 HMM-1 在較多的資料時，要能選到適當的模型卻是非常困難的。我們由這個例子可以看到 HMM-2 配合 BIC 模型選擇上會比用 HMM-1 加上 BIC 更有力，在於 HMM-2 會比 HMM-1 用到更少的參數來建模型。另外我們也將 PKC (S)、CaM-KII (S)和 CDK (S) 訓練出來的 HMM-2 結構，分別放在圖 5.8，圖 5.10 和圖 5.12 做為例子，給讀者做為參考。

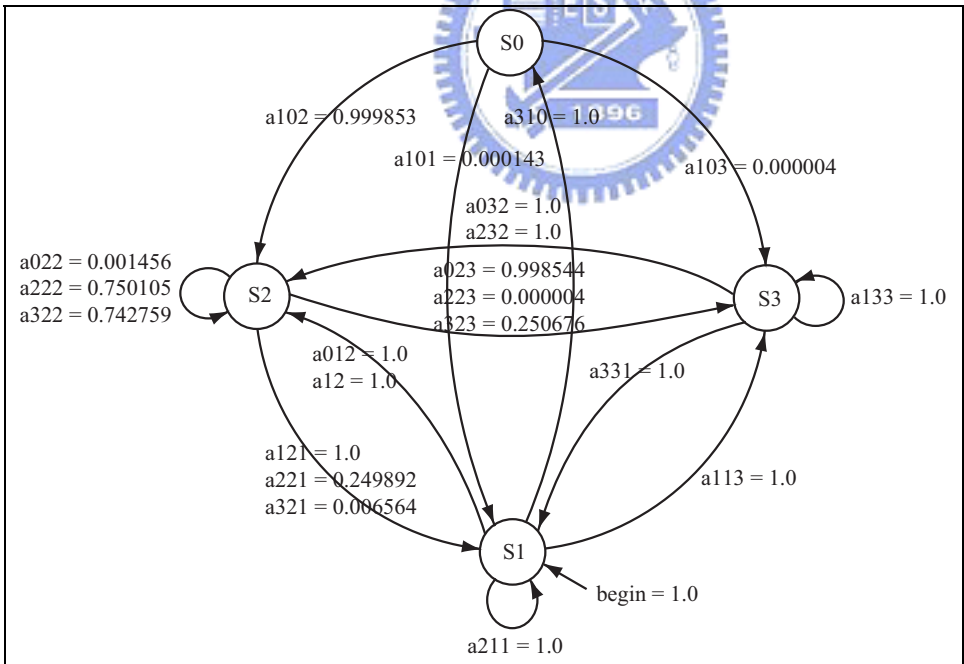

圖 5.8: PKC (S)資料的 HMM-2 結構

圖 5.9: PKC (S)資料的 HMM-2 結構的符號觀測機率圖

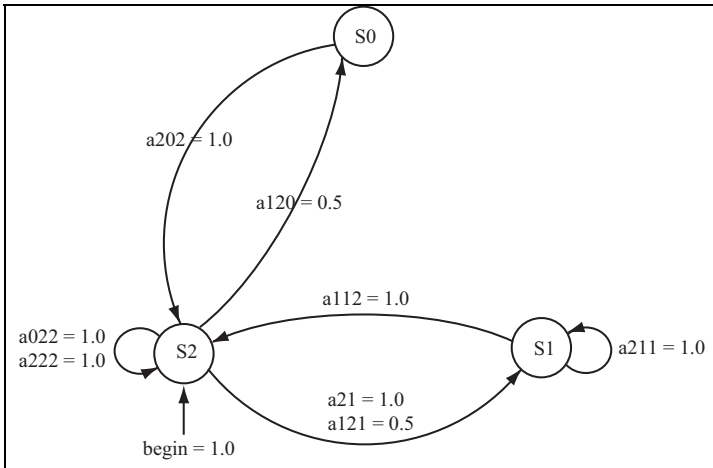

圖 5.10: CaM-KII (S)資料的 HMM-2 結構

圖 5.11: CaM-KII (S)資料的 HMM-2 結構的符號觀測機率圖

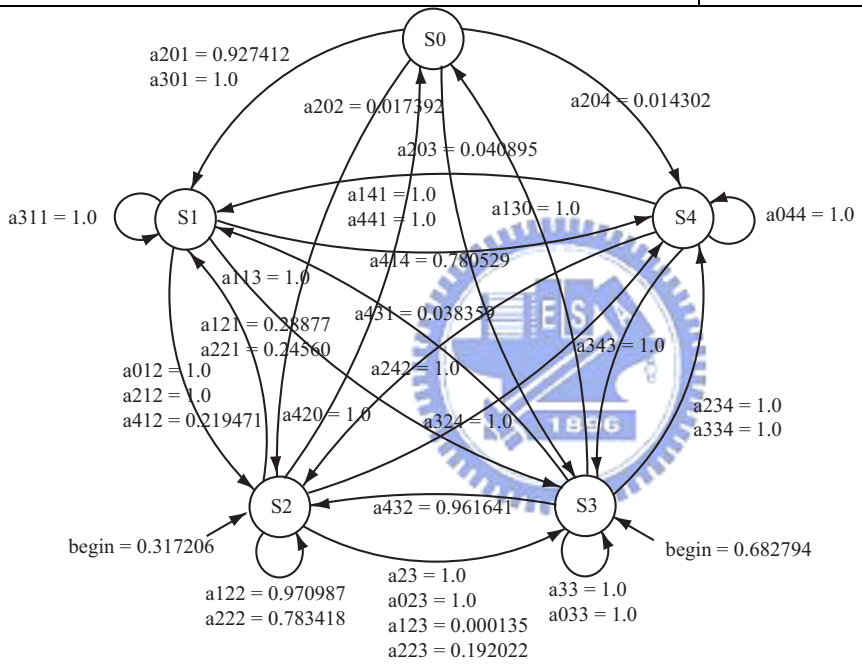

圖 5.12: CDK (S)資料的 HMM-2 結構

圖 5.13: CDK (S)資料的 HMM-2 結構的符號觀測機率圖

5.3 相關系數分析

在我們的研究中發現 ROC (Receiver Operating Characteristic Curve)面積越高不代表系統越可靠，一個系統的預測能力，因該要包含是否可以提供一個正確的門檻值，因為在少量訓練資料中，門檻值的決定就變的相當敏感，我們的目地不只是要在訓練資料得到一個高的正確率，還希望訓練資料正確率與測試資料的正確率之間的誤差最小，以減少估測上的錯誤及失真。因為在真實生活中，決策者常常需要從系統中就可以決定預測資料的情形。相關分析是利用來衡量兩個隨機變數之間「直線關係」的方向與強弱程度，因為對於研究變數所測量的尺度不同，需要搭配不同的相關係數(correlation coefficient)，才能進行正確的分析，而在本研究所採用的相關係數的公式是：

$$Correl(X,Y)=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} \tag{21}$$

由於其定義，相關係數的值恆介於-1 與+1 之間。+1 表示兩個變數為完全正向線性相關。-1 表示兩個變數為完全負向線性相關；若相關係數的值非常接近零，則表示兩個變數無線性關係。這裡的 X,Y 我們分別代入訓練資料隨著各各門檻值的正確率，和測試資料隨著各各門檻值的正確率，如圖 5.14，我們使用此函數計算兩個資料群的相關係數，以瞭解其性質間的相關性，若相關係數越高表示系統跑出來的數據越可信。我們將 18 個資料的相關係數整理在表 5.2，從此表中 HMMer 的平均相關係數是 0.8752 而 iHMM 的平均相關係數是 0.9869，可以看到本論文所提出的 iHMM 要比 HMMer 在訓練和測試資料更有顯著的正相關，此說明了 iHMM 所建立的模型要比 HMMer 穩定並具有良好的估計和預測能力。

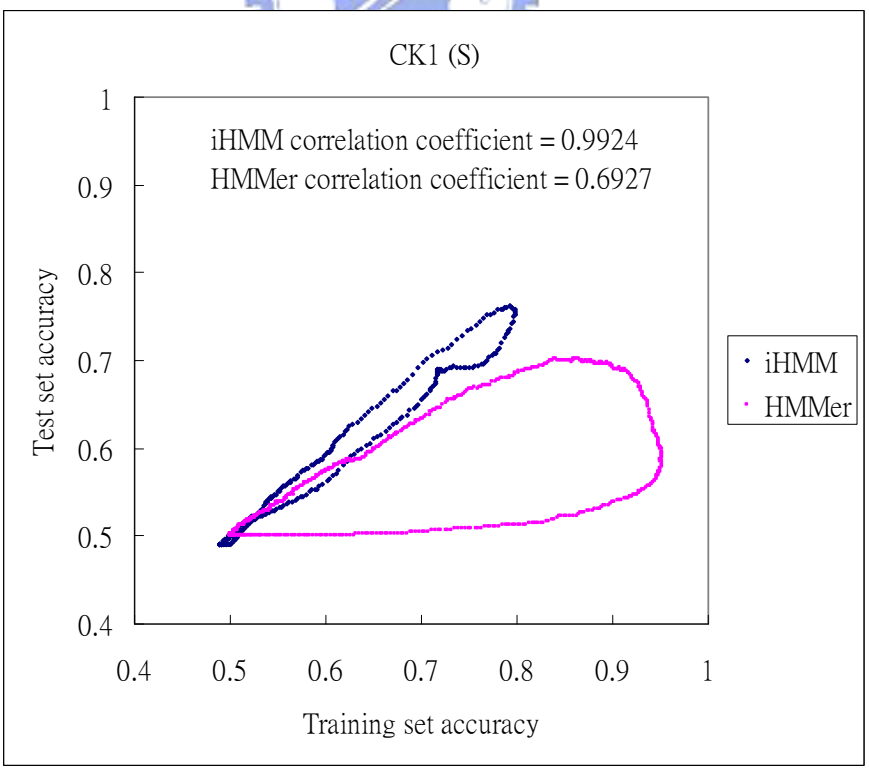

圖 5.14: CK1 (S)資料的相關係數分析圖

表 5.2:相關係數比較表

殘基	激酶 （資料筆數）	方法	
		HMMer	iHMM
S	PKA (308)	0.9980	0.9990
	PKB (81)	0.9280	0.9962
	PKC (304)	0.9965	0.9986
	PKG (30)	0.6653	0.9381
	CaM-KII (76)	0.9596	0.9908
	CK1 (49)	0.6927	0.9924
	CK2 (243)	0.9941	0.9989
	CDK (195)	0.9961	0.9970
	MAPK (207)	0.9924	0.9995
	ATM (77)	0.8662	0.9916
T	PKA (39)	0.6182	0.9972
	PKC (71)	0.9386	0.9582
	CK2 (42)	0.6613	0.9922
	CDK (113)	0.9474	0.9970
	MAPK (81)	0.8916	0.9970
Y	EGFR (46)	0.6568	0.9409
	INSR (58)	0.9676	0.9834
	SRC (143)	0.9824	0.9970
平均		0.8752	0.9869

5.4 訓練兩類的結果

在這之前，我們為了公平比較三者演算法都只訓練“Positive”這一類。而在這裡為了使正確率提高，我們訓練positive跟negative兩類的資料，不同於HMMer的log-odds算法，而是將之前的log-odds 分數式(7)進行修改如式(22)，是用來評估這條序列相對於 λ_N 多少機率是由 λ_P產生。

$$score = \log \frac{P(O \mid \lambda_P)}{P(O \mid \lambda_N)} \tag{22}$$

ROC 曲線是反映此敏感性和特異性連續變數的綜合指標，它的 X 軸與 Y 軸分別代表偽陽性與真陽性，它的曲線是將每個門檻值下得到的 sensitivity 與 1-specificity 所做的點狀圖，通常 ROC 曲線下的面積佔用來評估一個演算法對於資料的分類能力。我們將 iHMM 訓練兩類的結果的 ROC 圖放在圖 5.16 跟圖 5.17，而 negative 資料的 iHMM 結構於在圖 5.15，其正確率放在表 5.3。而從圖中我們可以看到，加上第二類資料的訓練，可以提供更精確的辨識效果。

表 5.3: PKA (S)訓練兩類資料的正確率表

PKA	Threshold	Sensitivity	Specificity	Precision	Accuracy
δ_1	-0.0483 (0.1389)	0.8713 (0.0170)	0.8454 (0.0167)	0.8516 (0.0128)	0.8584 (0.0095)
δ_2	-0.1617 (0.2100)	0.8934 (0.0196)	0.8580 (0.0208)	0.8659 (0.0157)	0.8757 (0.0068)

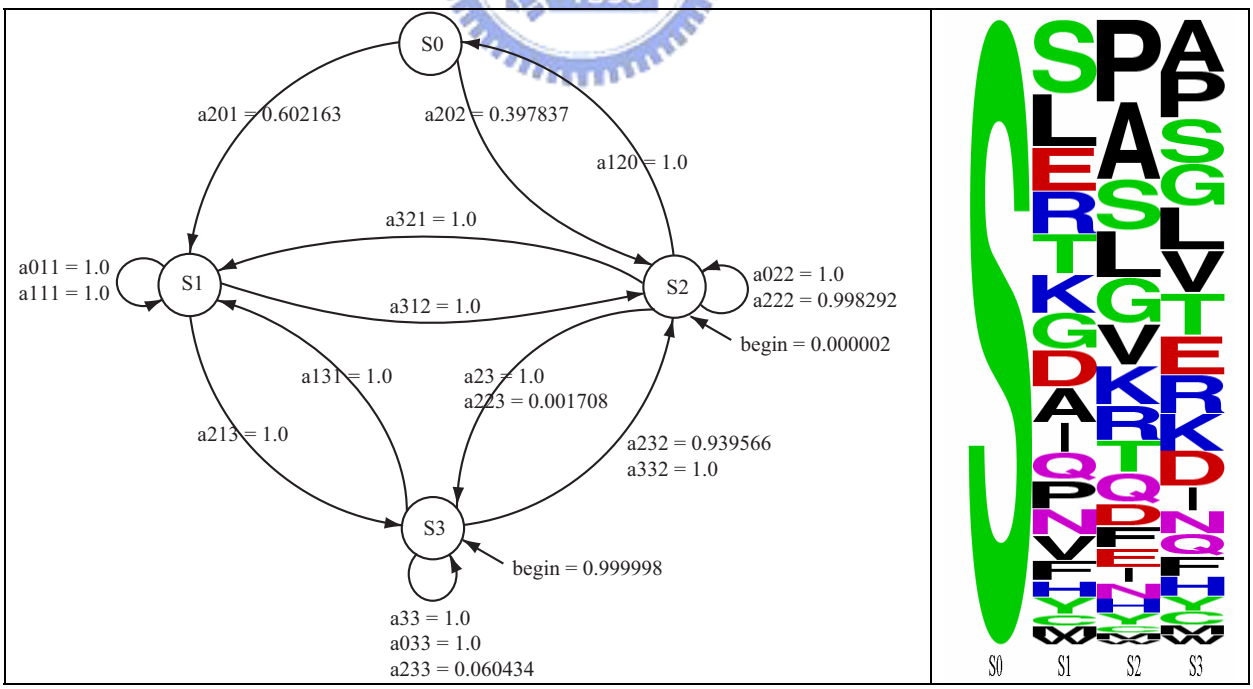

圖 5.15: PKA (S) negative 資料的 HMM-2 結構

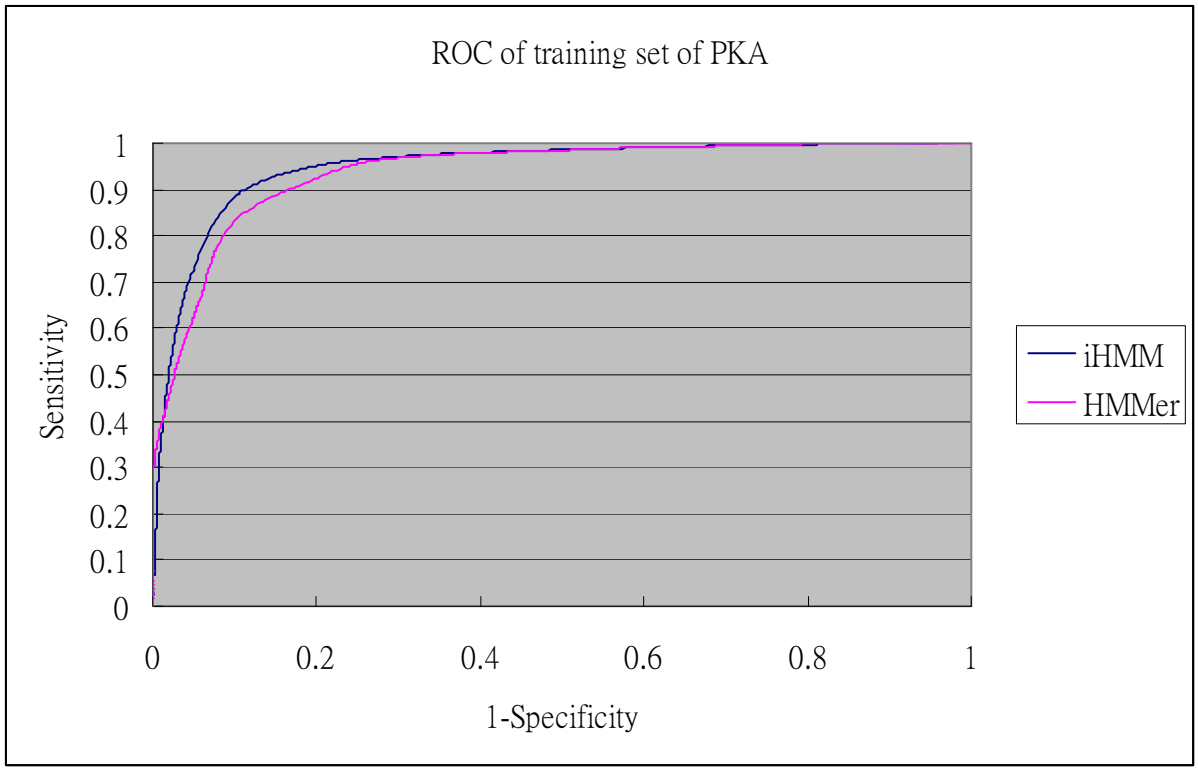

圖 5.16: PKA (S)兩類資料於訓練資料的 ROC 圖

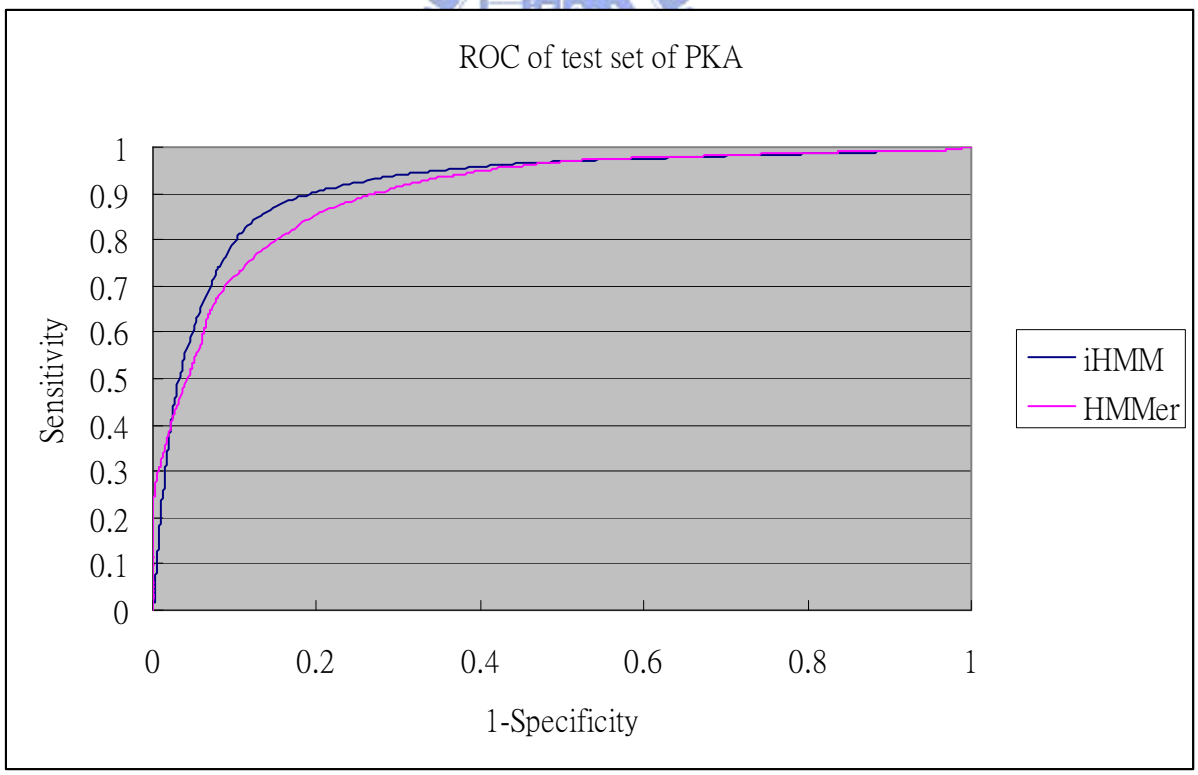

圖 5.17: PKA (S)兩類資料於測試資料的 ROC 圖

第六章 結論與展望

6.1 結論

實驗結果發現在同樣只訓練陽性(positive) 資料的公平比較之下，以 30 次的 5 等份交叉驗證（5-fold cross-validation, 5-CV）在測試資料上 iHMM 的辨識率與 HMMer 相比得到將近平均 4.3%的提升，而跟傳統 HMM-1 相比則得到將近平均 3.7%的提升。另外如果將 iHMM 同時訓練陽性(positive)跟陰性(negative)兩類的資料在 PKA 上與 HMMer 相比則得到將近平均 3.6%提升。依據本實驗將本研究的結論與貢獻整理如下：

1. HMM-1 加入 BIC 在測試資料上其最後平均辨識率會與 HMMer 差不多。
2. HMM-2 加入 BIC 在測試資料上最後平均辨識率要比 HMM-1 加入 BIC 更為有效。
3. 加入 negative 資料的訓練也可以提升辨識率。
4. 我們提出 iHMM 在訓練與測試資料比 HMMer 有更少的辨識率誤差。
5. 在少量資料時，因該要綜合考量訓練資料與測試資料之間的辨識率誤差，不應該只單存看 ROC 底下的面積。
6. 我們提出 iHMM 在測試資料上平均辨識率要比 HMM-1 和 HMMer 要好，不過美中不足的是我們的 iHMM 需要花比 HMMer 好幾倍的時間在訓練過程中。

結論，我們提出了一個新的演算法來預測預測蛋白激酶磷酸化的位置。在我們的實驗結果中可以看出一點小改進。iHMM 可以比 HMMer 找出比較正確的位置。雖然我們的演算法還有很多實驗要去做，但我們相信 HMM-2 搭配貝氏資訊準則(Bayesian information criterion, BIC) 並以多點搜尋式的 Baum-Welch 演算法在新的生物機率模型中將有重要影響，並且將是一個重要的工具，不只用來找磷酸化位置，還可以用來解決其它許多生物資訊相關問題。

「蛋白質磷酸化」近年來一直是生物科學中十分重要的環節之一，與生命維持的各項反應息息相關，一個細胞中有 30%~50%的蛋白質在任何特定時間都在進行磷酸化作用，然而對致癌激酶參與癌症形成及訊息傳遞的機轉不在本論文的研究範圍內。本論文研究主要目的是提出更好的預測方法，並將每個蛋白激酶建立電腦模型，以希望提供未來蛋白激酶磷酸化研究基本的電腦輔助工具，可以有助於未來新藥開發或臨床實驗設計研究，以減少因應新治療所需的高額成本等的課題。

6.2 未來展望

最近幾年來蛋白激酶磷酸化的問題已是疾病治療中最有潛力的新興研究領域[29]。像是酪氨酸激脢接受體 (receptor tyrosine kinase, RTK) 是細胞表面生長因子的一部份，它有者固有的配體控制的酪氨酸激脢 (tyrosine-kinase) 活性，同時在正常細胞內廣泛調控許多功能，並且是致癌基因的關鍵。在本研究中我們也預測了三個酪氨酸激酶 INSR，EGFR 跟 SRC。其中表皮生長因子受體(EGFR)的治療，是近年來出現確有療效的抗腫瘤治療。正常細胞的酪胺酸激酶活性一般會受到嚴密的調控，但是癌細胞往往具有很強的酪胺酸激酶活性，使得癌細胞內的酪胺酸激酶接受體過度表現，於是癌細胞不斷的進行分化、增殖、抗凋亡(anti-apoptosis)、血管新生以及轉移，因此酪胺酸激酶被認為和癌

細胞的增生有關。因此若能設法抑制酪胺酸激酶的活性，使酪胺酸激酶接受體不再過度表現，將有助於癌細胞的控制。在使用人化抗體和小分子藥物等干預療法中 RTKs 和生長因子這些配體已經變成合理的標靶，而以 RTK 為基礎的癌症治療，像是轉移性乳癌、胃腸道間質瘤和非小細胞肺癌等等，已經廣泛的應用在臨床上，並且藉此開起了基因治療研究發展的新動力[10]。另外週期素依賴性激脢 (cyclin-dependent kinases, CDK) 當作標靶已經成為[30]中風的診斷與治療策略，以及蛋白激酶抑制劑用來治療心臟衰竭症[31]，另外像是 p38 MAP 激酶也被認為是炎症性疾病治療的主要標靶[32]，這些是本研究未來要發展的方向。

參考文獻

[1] N. Blom, S. Gammeltoft, and S. Brunak, "Sequence and structure-based prediction of eukaryotic protein phosphorylation sites," *Journal of Molecular Biology,* vol. 294, pp. 1351-62, Dec 17 1999.

[2] R. Linding, L. J. Jensen, G. J. Ostheimer, M. A. van Vugt, C. Jorgensen, I. M. Miron, F. Diella, K. Colwill, L. Taylor, K. Elder, P. Metalnikov, V. Nguyen, A. Pasculescu, J. Jin, J. G. Park, L. D. Samson, J. R. Woodgett, R. B. Russell, P. Bork, M. B. Yaffe, and T. Pawson, "Systematic discovery of in vivo phosphorylation networks," *Cell,* vol. 129, pp. 1415-26, Jun 29 2007.

[3] H. D. Huang, T. Y. Lee, S. W. Tzeng, L. C. Wu, J. T. Horng, A. P. Tsou, and K. T. Huang, "Incorporating hidden Markov models for identifying protein kinase-specific phosphorylation sites," *Journal of Computational Chemistry,* vol. 26, pp. 1032-41, Jul 30 2005.

[4] S. R. Eddy, "Profile hidden Markov models," *Bioinformatics,* vol. 14, pp. 755-763, 1998.

[5] F. Diella, S. Cameron, C. Gemund, R. Linding, A. Via, B. Kuster, T. Sicheritz-Ponten, N. Blom, and T. J. Gibson, "Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins," *BMC Bioinformatics,* vol. 5, p. 79, Jun 22 2004.

[6] M. Huse and J. Kuriyan, "The conformational plasticity of protein kinases," *Cell,* vol. 109, pp. 275-282, May 3 2002.

[7] S. K. Hanks, A. M. Quinn, and T. Hunter, "The protein kinase family: conserved features and deduced phylogeny of the catalytic domains," *Science,* vol. 241, pp. 42-52, Jul 1 1988.

[8] S. K. Hanks and T. Hunter, "Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification," *FASEBJ Journal,* vol. 9, pp. 576-96, May 1995.

[9] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam, "The protein kinase complement of the human genome," *Science,* vol. 298, pp. 1912-34, Dec 6 2002.

[10] A. Gschwind, O. M. Fischer, and A. Ullrich, "Timeline - The discovery of receptor tyrosine kinases: targets for cancer therapy," *Nature Reviews Cancer,* vol. 4, pp. 361-370, May 2004.

[11] D. R. Knighton, J. H. Zheng, L. F. Teneyck, V. A. Ashford, N. H. Xuong, S. S. Taylor, and J. M. Sowadski, "Crystal-Structure of the Catalytic Subunit of Cyclic Adenosine-Monophosphate Dependent Protein-Kinase," *Science,* vol. 253, pp. 407-414, Jul 26 1991.

[12] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner, "WebLogo: A sequence logo generator," *Genome Research,* vol. 14, pp. 1188-1190, Jun 2004.

[13] T. D. Schneider and R. M. Stephens, "Sequence Logos - a New Way to Display Consensus Sequences," *Nucleic Acids Research,* vol. 18, pp. 6097-6100, Oct 25 1990.

[14] L. R. Rabiner, "A Tutorial on Hidden Markov-Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE,* vol. 77, pp. 257-286, Feb 1989.

[15] A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler, "Hidden Markov-Models in Computational Biology - Applications to Protein Modeling,"

Journal of Molecular Biology, vol. 235, pp. 1501-1531, Feb 4 1994.

[16] G. A. Churchill, "Stochastic models for heterogeneous DNA sequences," *Bulletin of Mathematical Biology,* vol. 51, pp. 79-94, 1989.

[17] J. F. Mari, J. P. Haton, and A. Kriouile, "Automatic word recognition based on second-order hidden Markov models," *IEEE Transactions on Speech and Audio Processing,* vol. 5, pp. 22-25, Jan 1997.

[18] I. Shahin, "Improving speaker identification performance under the shouted talking condition using the second-order hidden Markov models," *Eurasip Journal on Applied Signal Processing,* vol. 2005, pp. 482-486, Mar 15 2005.

[19] I. Shahin, "Enhancing speaker identification performance under the shouted talking condition using second-order circular hidden Markov models," *Speech Communication,* vol. 48, pp. 1047-1055, Aug 2006.

[20] O. Aycard, J.-F. Mari, and R. Washington, "Learning to automatically detect features for mobile robots using second-order Hidden Markov Models," *International Journal Of Advanced Robotic Systems* vol. 1, pp. 231-244, Dec 2004.

[21] A. Krogh and G. Mitchison, "Maximum entropy weighting of aligned sequences of proteins or DNA," *Proceedings Third International Conference on Intelligent Systems for Molecular Biology* vol. 3, pp. 215-21, 1995.

[22] K. Sjolander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I. S. Mian, and D. Haussler, "Dirichlet mixtures: A method for improved detection of weak but significant protein sequence homology," *Computer Applications in the Biosciences,* vol. 12, pp. 327-345, Aug 1996.

[23] H. Akaike, "New Look at Statistical-Model Identification," *IEEE Transactions on Automatic Control,* vol. Ac19, pp. 716-723, 1974.

[24] G. Schwarz, "Estimating Dimension of a Model," *Annals of Statistics,* vol. 6, pp. 461-464, 1978.

[25] A. Schliep, I. G. Costa, C. Steinhoff, and A. Schonhuth, "Analyzing gene expression time-courses," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 2, pp. 179-93, Jul-Sep 2005.

[26] J. Martin, J. F. Gibrat, and F. Rodolphe, "Analysis of an optimal hidden Markov model for secondary structure prediction," *BMC Structural Biology,* vol. 6, p. 25, 2006.

[27] M. Brand, "Structure learning in conditional probability models via an entropic prior and parameter extinction," *Neural Computation,* vol. 11, pp. 1155-1182, Jul 1 1999.

[28] U. J. Matthew A. Siegler, Bhiksha Raj, Richard M. Stern, "Automatic Segmentation, Classification and Clustering of Broadcast News Audio," *DARPA Speech Recognition Workshop,* 1997.

[29] R. Sridhar, O. Hanson-Painton, and D. R. Cooper, "Protein kinases as therapeutic targets," *Pharmaceutical Research,* vol. 17, pp. 1345-1353, Nov 2000.

[30] H. Osuga, S. Osuga, F. H. Wang, R. Fetni, M. J. Hogan, R. S. Slack, A. M. Hakim, J. E. Ikeda, and D. S. Park, "Cyclin-dependent kinases as a therapeutic target for stroke," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 97, pp. 10254-10259, Aug 29 2000.

[31] C. J. Vlahos, S. A. McDowell, and A. Clerk, "Kinases as therapeutic targets for heart failure," *Nature Reviews Drug Discovery,* vol. 2, pp. 99-113, Feb 2003.

[32] S. Kumar, J. Boehm, and J. C. Lee, "p38 map kinases: Key signalling molecules as therapeutic targets for inflammatory diseases," *Nature Reviews Drug Discovery,* vol. 2, pp. 717-726, Sep 2003.

附錄: 詳細實驗數據

PKA (S)

全名：Protein kinase A

資料筆數：positive 跟 negative 各 308 筆資料

序列長度：15

磷酸化位置：中間的絲氨酸(S)

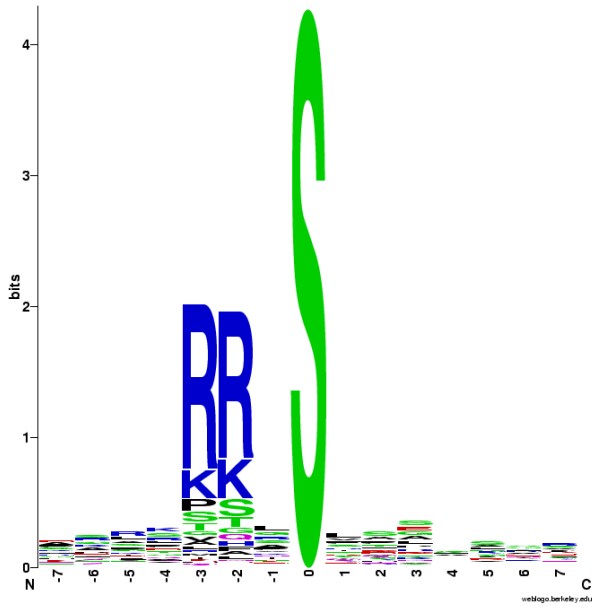

圖 A.1: PKA (S)資料的序列圖案

表 A.1: PKA (S)的 30 次 5-CV 於測試資料的效能比較表

PKA (S)		Threshold	Sensitivity	Specificity	Precision	Accuracy
δ_1	HMMer	-4.2667 (0.2835)	0.7823 (0.0212)	0.8676 (0.0212)	0.8591 (0.0183)	0.8250 (0.0089)
	HMM-1	3.3100 (0.1602)	0.7994 (0.0215)	0.8642 (0.0125)	0.8575 (0.0107)	0.8319 (0.0092)
	iHMM	3.1617 (0.1054)	0.8348 (0.0116)	0.8540 (0.0117)	0.8534 (0.0091)	0.8444 (0.0063)
δ_2	HMMer	-5.0157 (0.3400)	0.8534 (0.0218)	0.8469 (0.0273)	0.8538 (0.0212)	0.8503 (0.0073)
	HMM-1	3.2440 (0.3210)	0.8257 (0.0243)	0.8812 (0.0205)	0.8779 (0.0177)	0.8535 (0.0060)
	iHMM	3.3043 (0.2198)	0.8457 (0.0205)	0.8847 (0.0178)	0.8838 (0.0143)	0.8652 (0.0046)

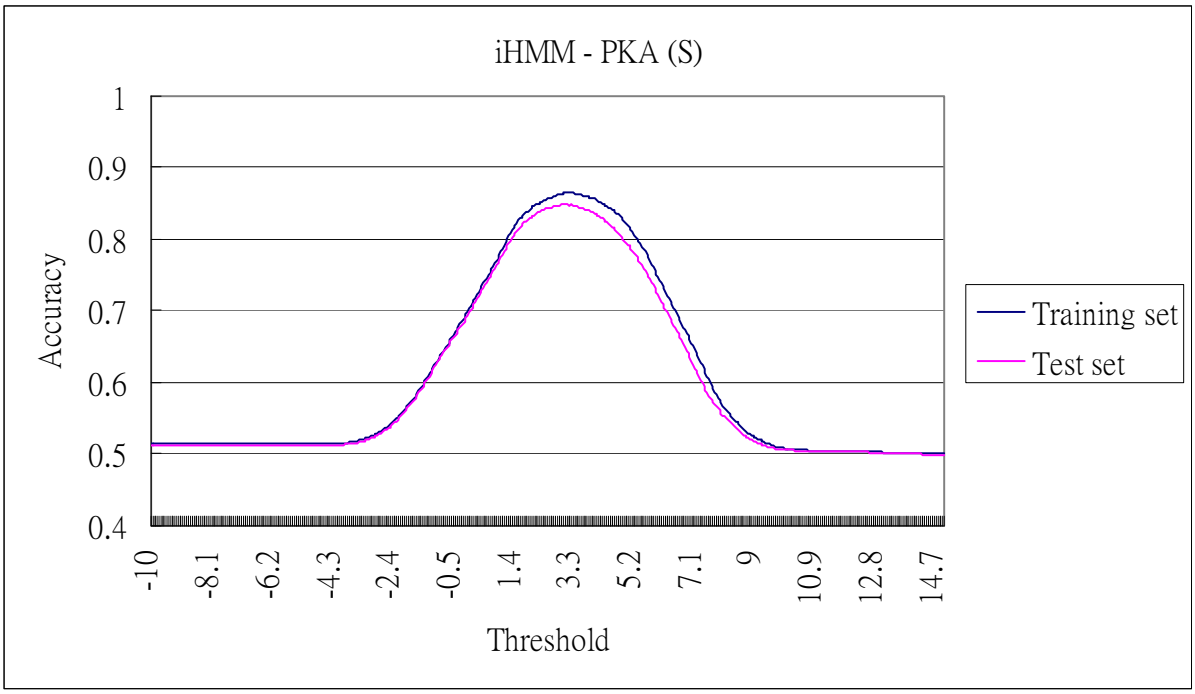

圖 A.2: PKA (S)資料於 iHMM 的門檻值與正確率對應圖

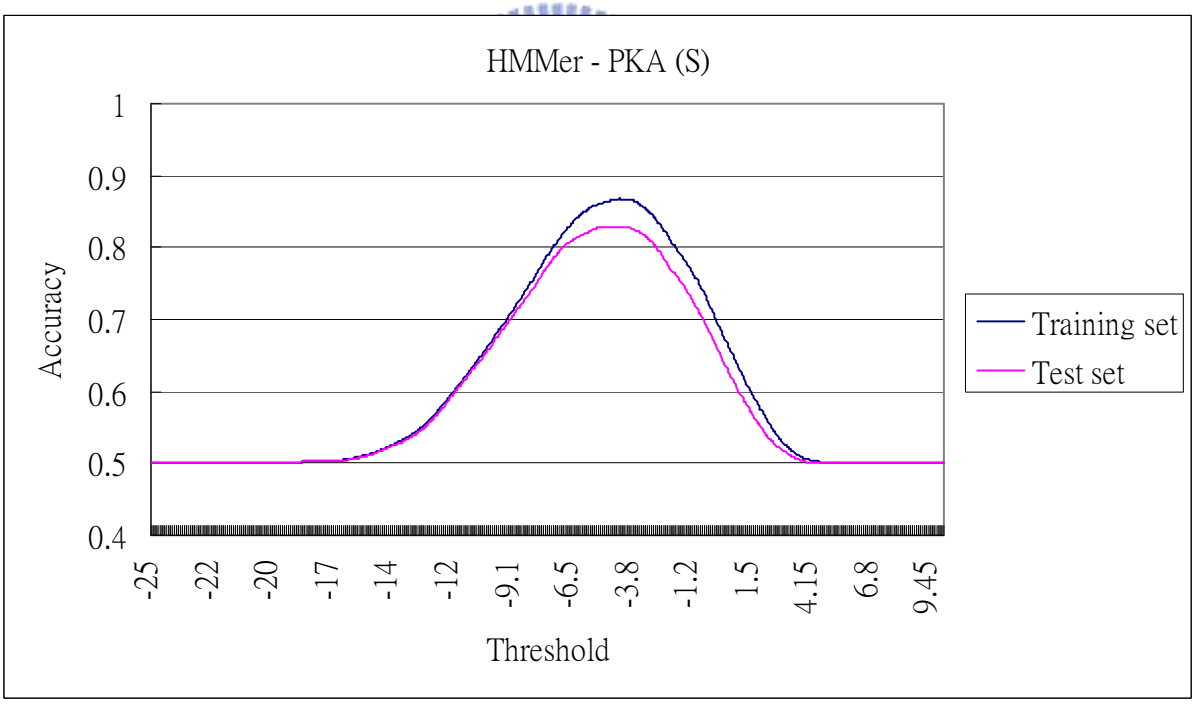

圖 A.3: PKA (S)資料於 HMMer 的門檻值與正確率對應圖

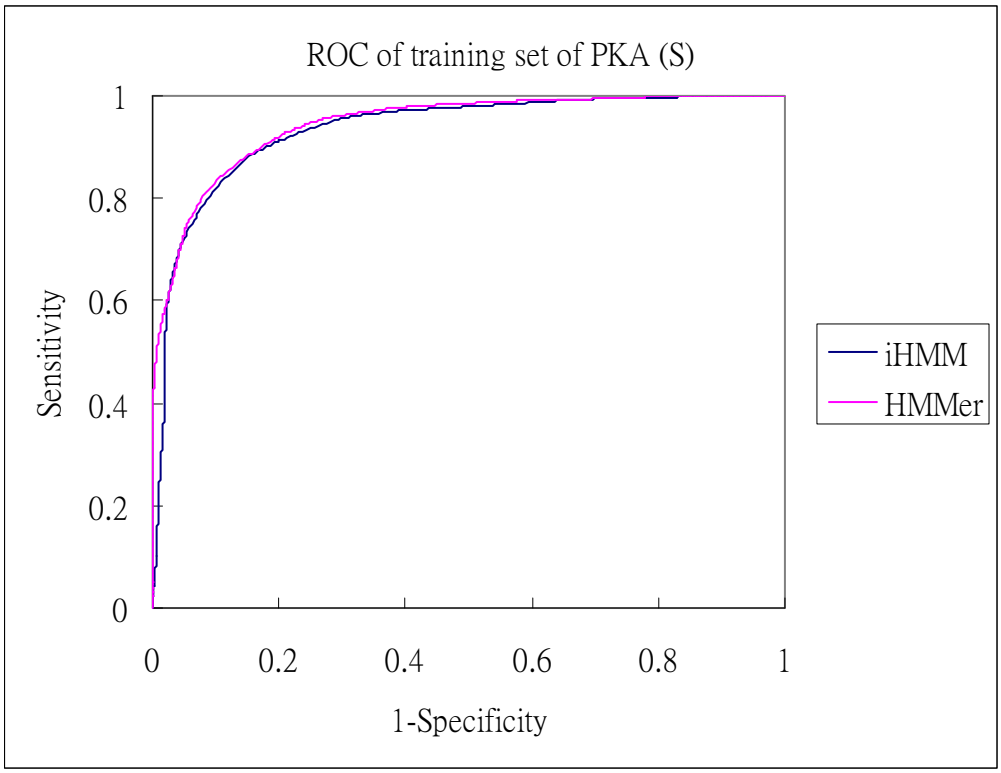

圖 A.4: PKA (S)於訓練資料的 ROC 圖

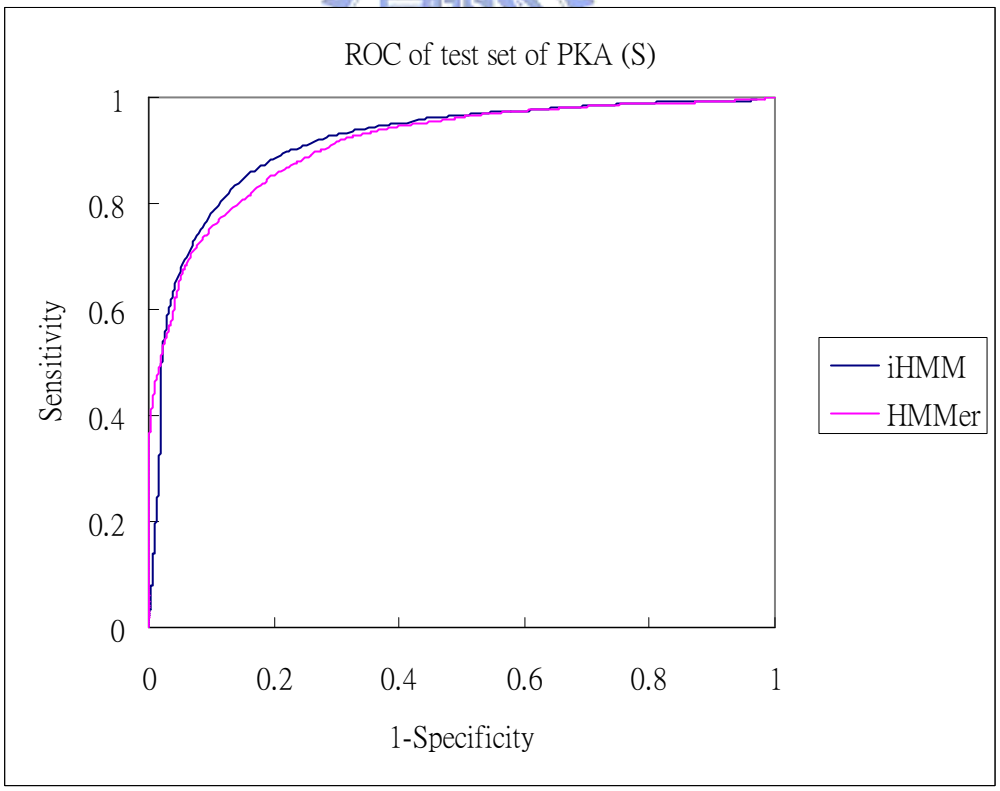

圖 A.5: PKA (S)於測試資料的 ROC 圖

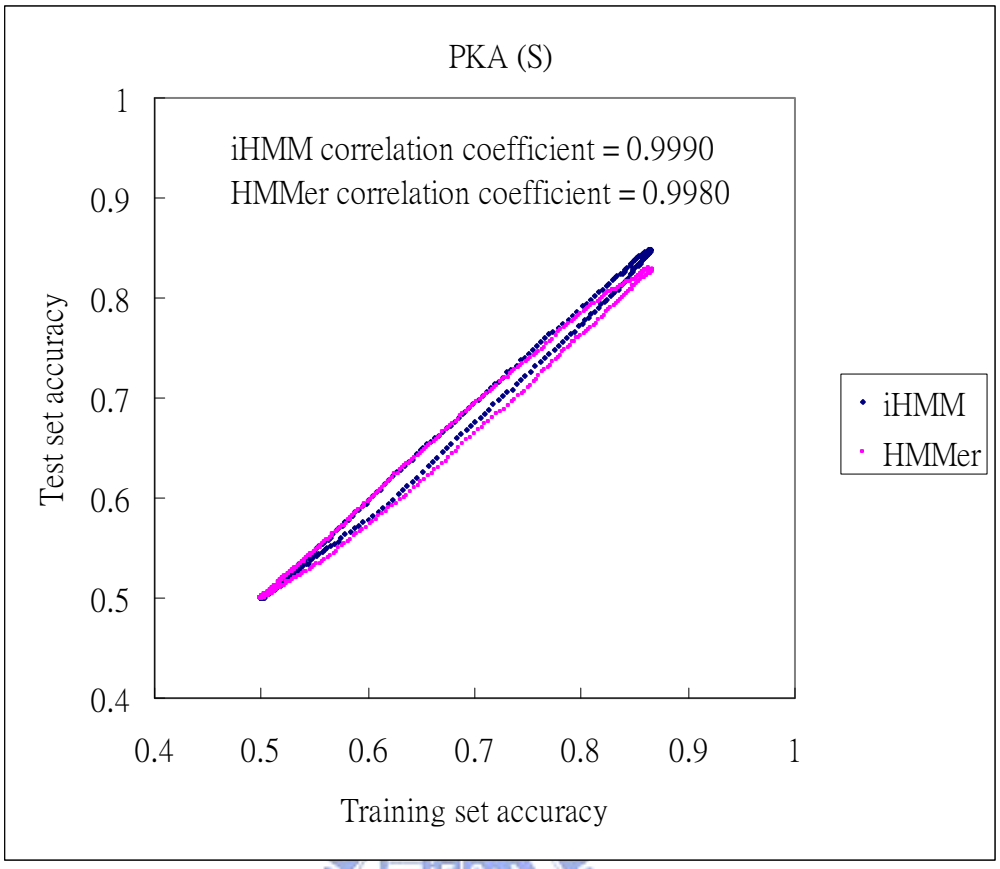

圖 A.6: PKA (S)資料的相關係數分析圖

PKA (T)

全名：Protein kinase A

資料筆數：positive 跟 negative 各 39 筆資料

序列長度：15

磷酸化位置：中間的蘇氨酸(T)

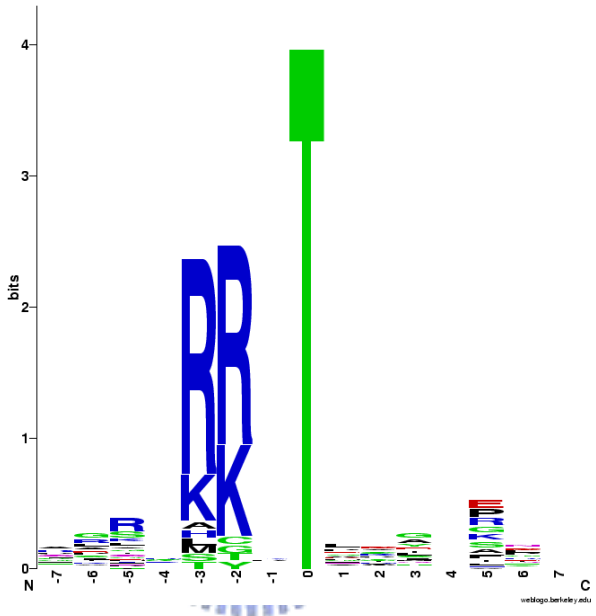

圖 A.7: PKA (T)資料的序列圖案

表 A.2: PKA (T)的 30 次 5-CV 於測試資料的效能比較表

PKA (T)		Threshold	Sensitivity	Specificity	Precision	Accuracy
δ_1	HMMer	-2.3400 (0.4343)	0.3027 (0.0520)	0.9698 (0.0201)	0.8536 (0.1144)	0.6358 (0.0281)
	HMM-1	0.4105 (0.2180)	0.6680 (0.0684)	0.7495 (0.0555)	0.7501 (0.0373)	0.7080 (0.0371)
	iHMM	3.2217 (0.3269)	0.7008 (().0594)	0.9074 (0.0272)	0.8941 (0.0367)	0.8037 (0.0314)
δ_2	HMMer	-12.0790 (1.0812)	0.8862 (0.0512)	0.8342 (0.0498)	0.8590 (0.0360)	0.8602 (0.0209)
	HMM-1	1.0458 (0.7838)	0.7211 (0.0786)	0.8381 (0.0549)	0.8485 (0.0437)	0.7793 (0.0271)
	iHMM	-7.1746 (28.0175)	0.7430 (0.0698)	0.9642 (.0.0187)	0.9595 (0.0187)	0.8538 (0.0315)

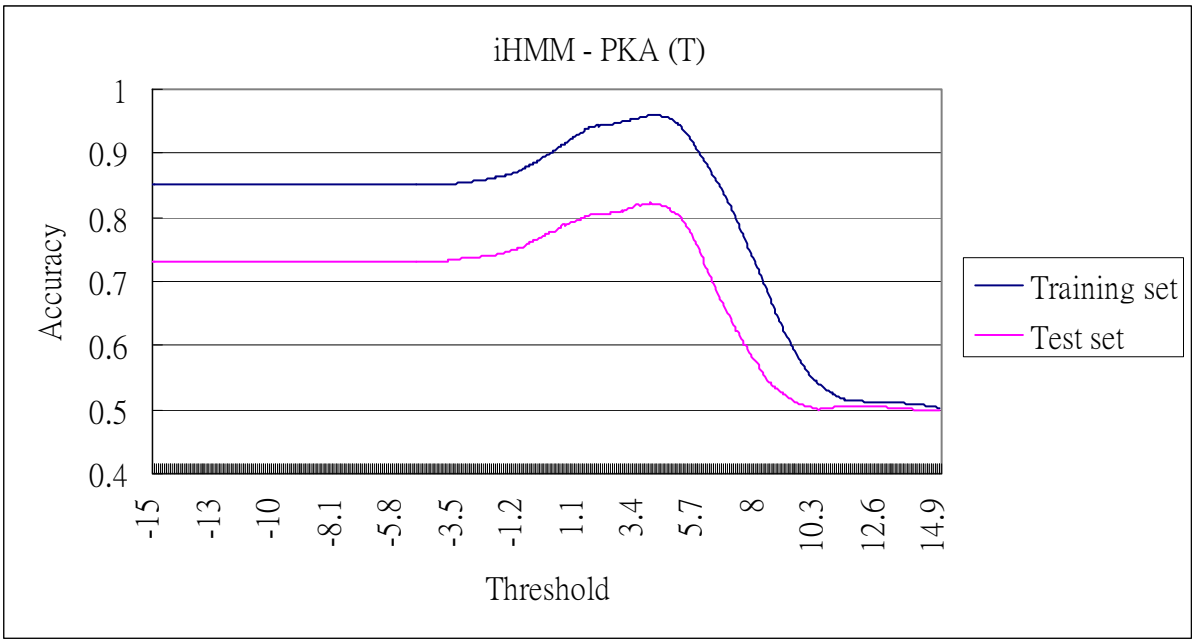

圖 A.8: PKA (T)資料於 iHMM 的門檻值與正確率對應圖

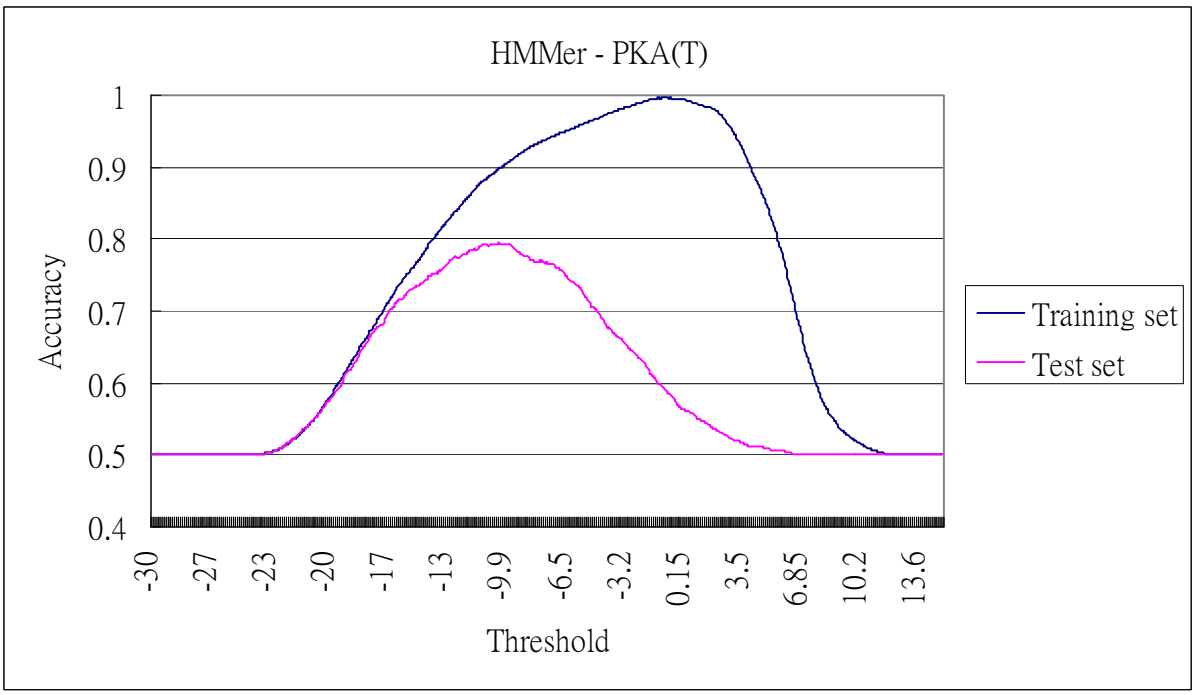

圖 A.9: PKA (T)資料於 HMMer 的門檻值與正確率對應圖

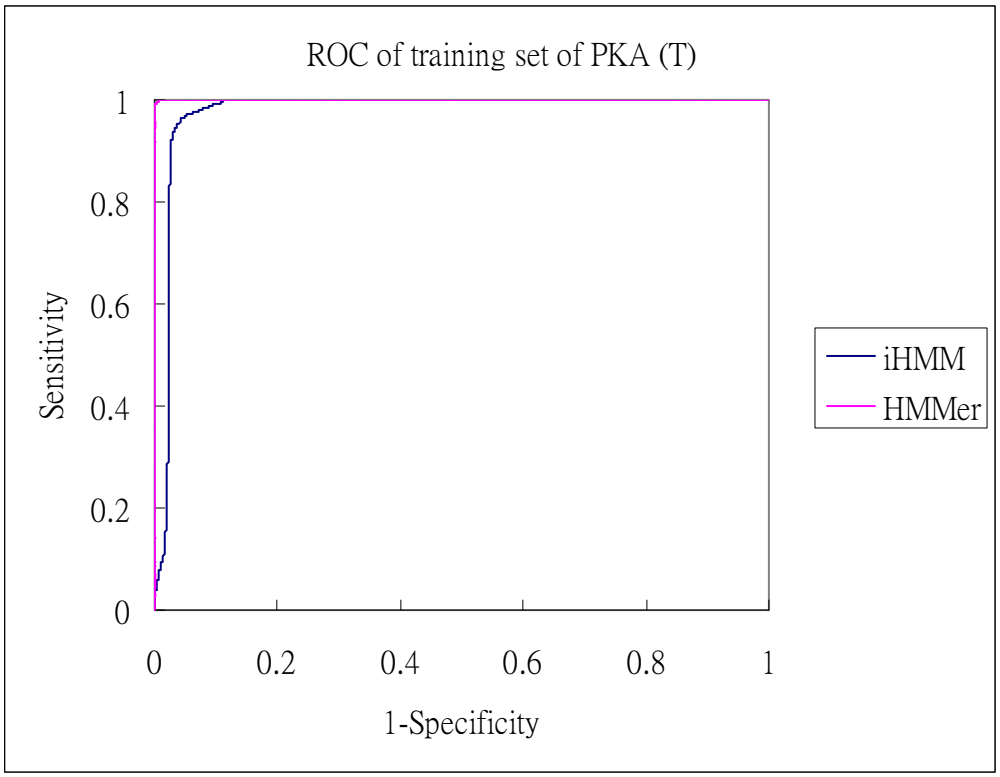

圖 A.10: PKA (T)於訓練資料的 ROC 圖

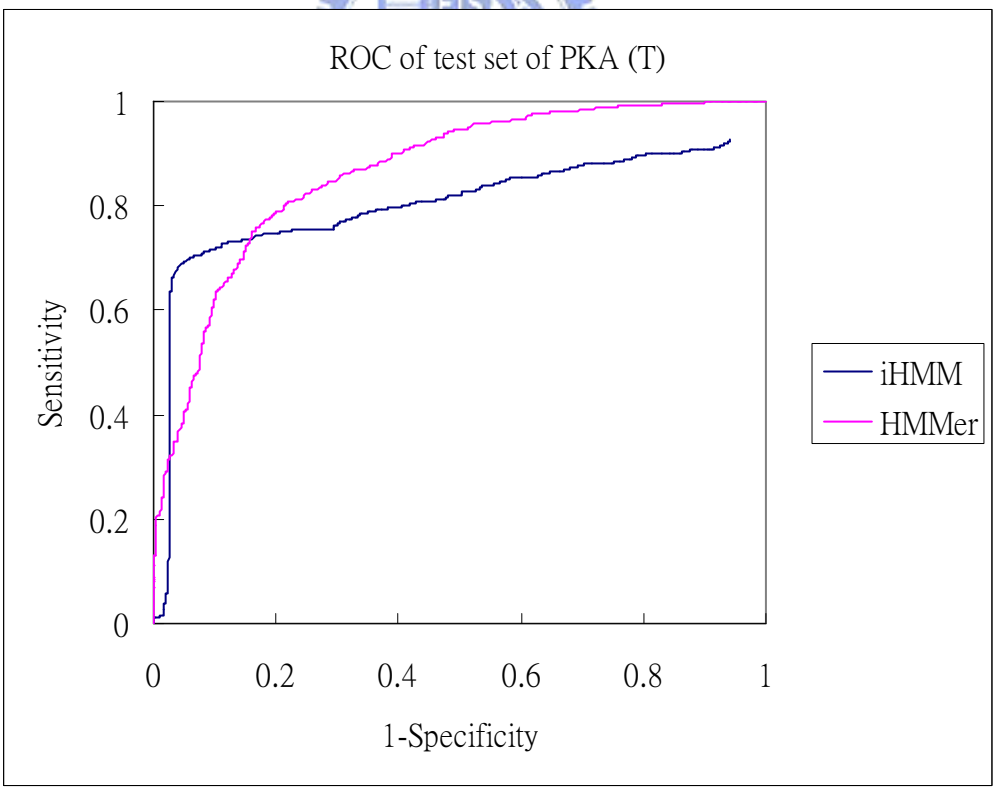

圖 A.11: PKA (T)於測試資料的 ROC 圖

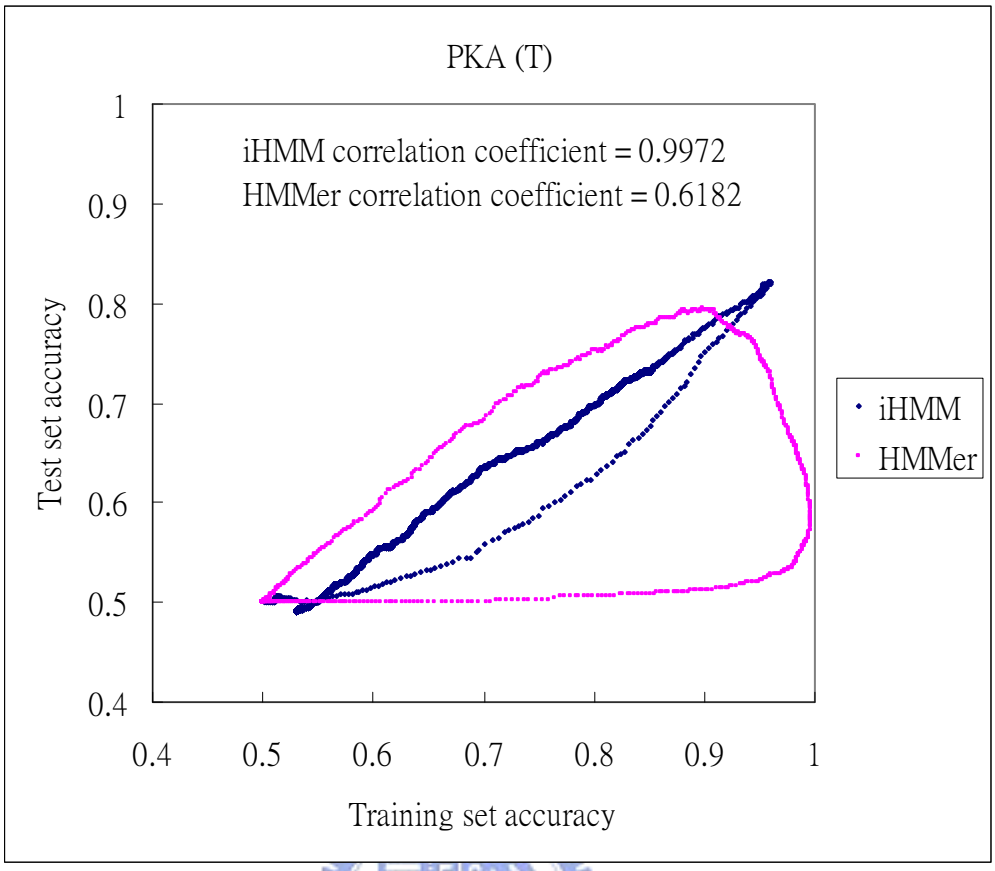

圖 A.12: PKA (T)資料的相關係數分析圖

PKB (S)

全名：Protein kinase B

資料筆數：positive 跟 negative 各 81 筆資料

序列長度：15

磷酸化位置：中間的絲氨酸(S)

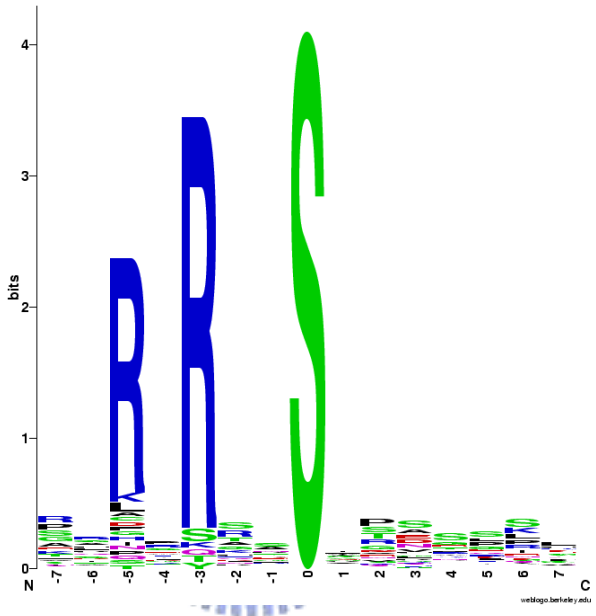

圖 A.13: PKB (S)資料的序列圖案

表 A.3: PKB (S)的 30 次 5-CV 於測試資料的效能比較表

PKB (S)		Threshold	Sensitivity	Specificity	Precision	Accuracy
δ_1	HMMer	-5.4843 (0.3306)	0.7516 (0.0338)	0.9639 (0.0159)	0.9568 (0.0175)	0.8579 (0.0169)
	HMM-1	2.2208 (0.3594)	0.6683 (0.0459)	0.9188 (0.0237)	0.9013 (0.0275)	0.7942 (0.0202)
	iHMM	1.8582 (0.1229)	0.8125 (0.0235)	0.8993 (0.0236)	0.8955 (0.0229)	0.8561 (0.0160)
δ_2	HMMer	-8.4843 (0.8688)	0.8882 (0.0249)	0.9261 (0.0288)	0.9310 (0.0248)	0.9078 (0.0104)
	HMM-1	-21.0125 (50.5841)	0.6953 (0.0507)	0.9479 (0.0297)	0.9431 (0.0262)	0.8230 (0.0181)
	iHMM	2.9525 (0.7226)	0.8350 (0.0245)	0.9474 (0.0198)	0.9481 (0.0167)	0.8918 (0.0103)

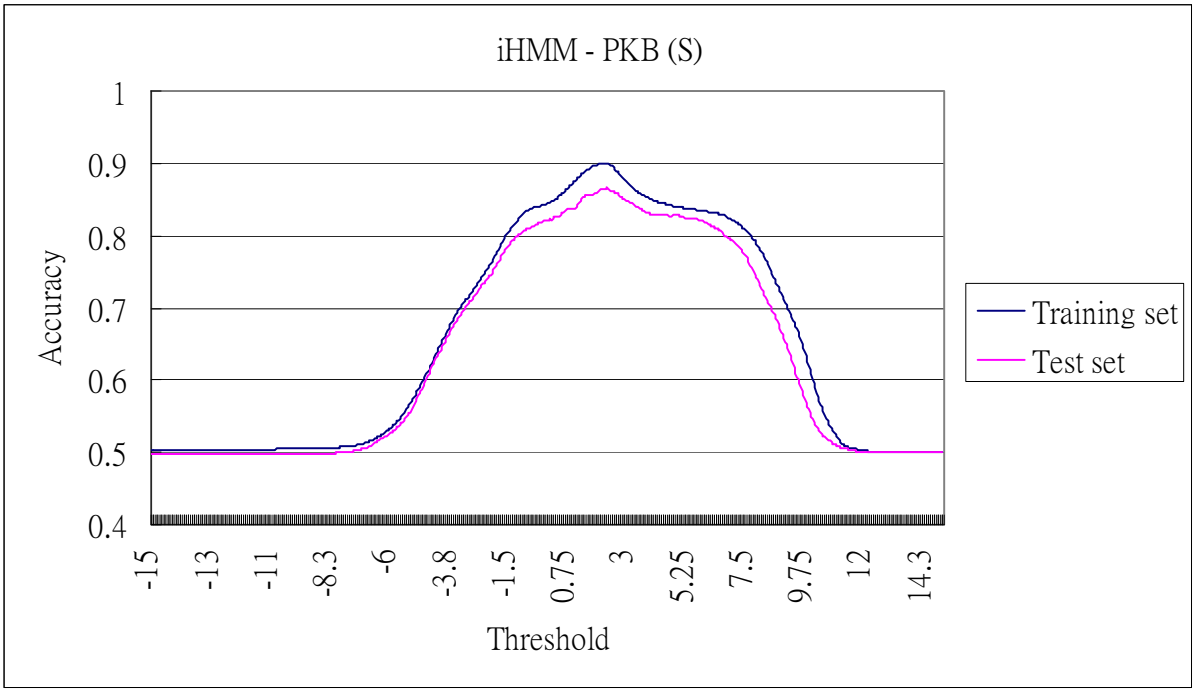

圖 A.14: PKB (S)資料於 iHMM 的門檻值與正確率對應圖

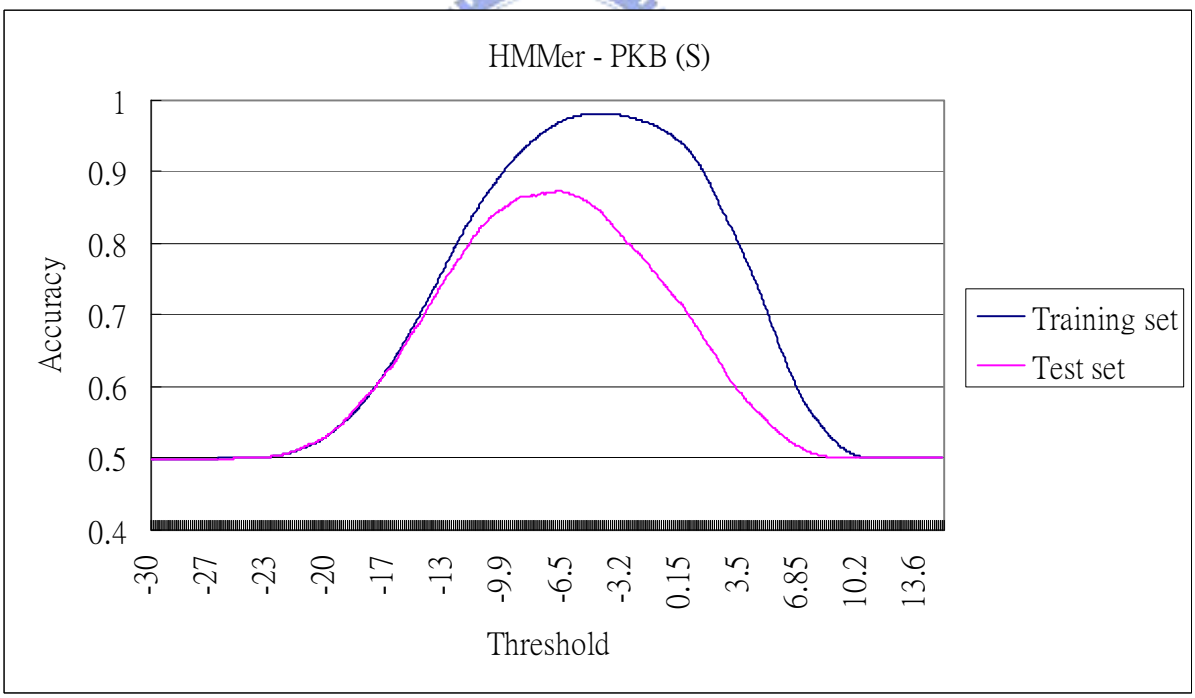

圖 A.15: PKB (S)資料於 HMMer 的門檻值與正確率對應圖

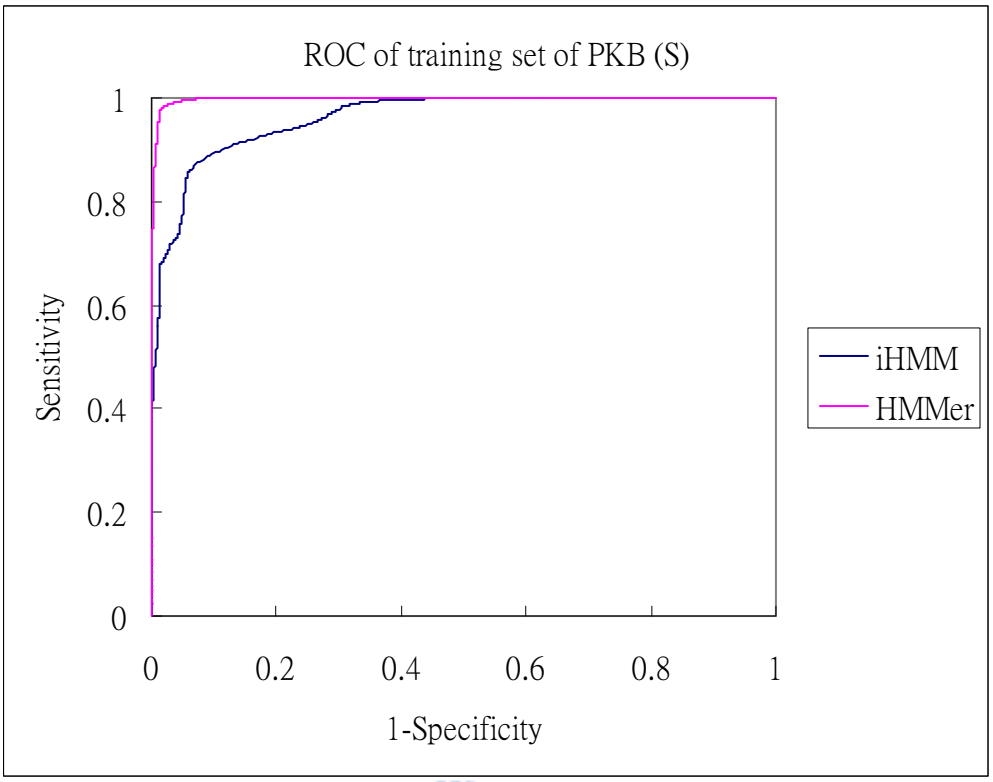

圖 A.16: PKB (S)於訓練資料的 ROC 圖

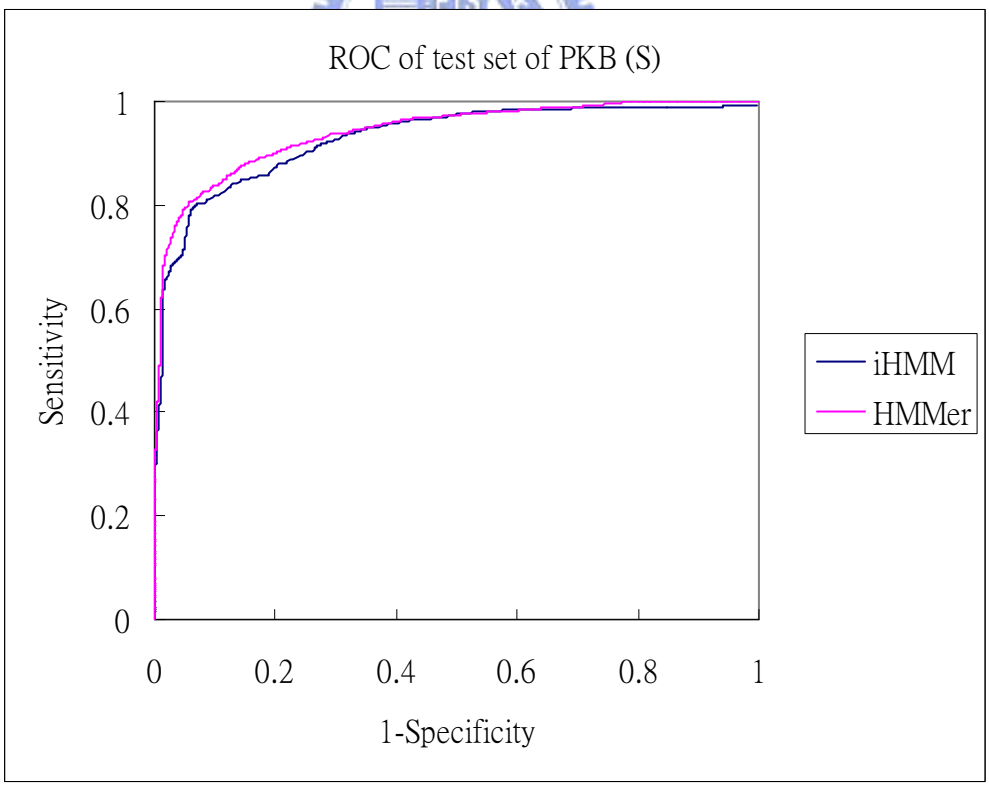

圖 A.17: PKB (S)於測試資料的 ROC 圖

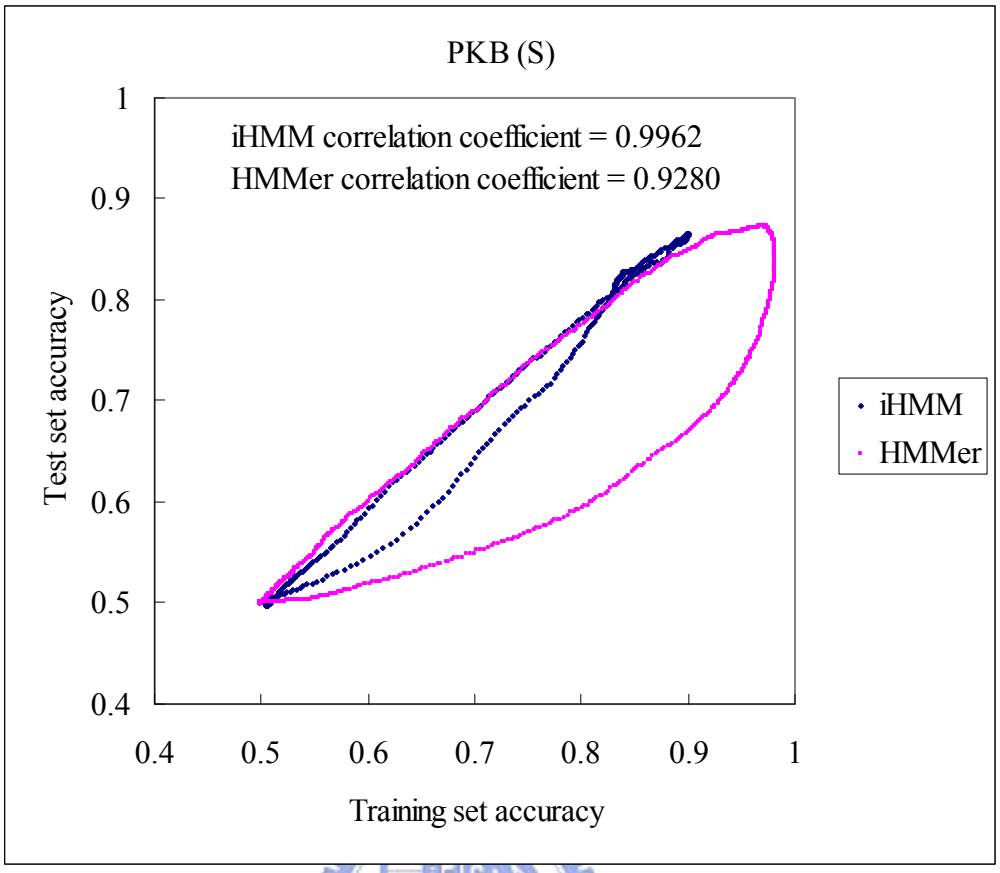

圖 A.18: PKB (S)資料的相關係數分析圖

PKC (S)

全名：Protein kinase C

資料筆數：positive 跟 negative 各 304 筆資料

序列長度：15

磷酸化位置：中間的絲氨酸(S)

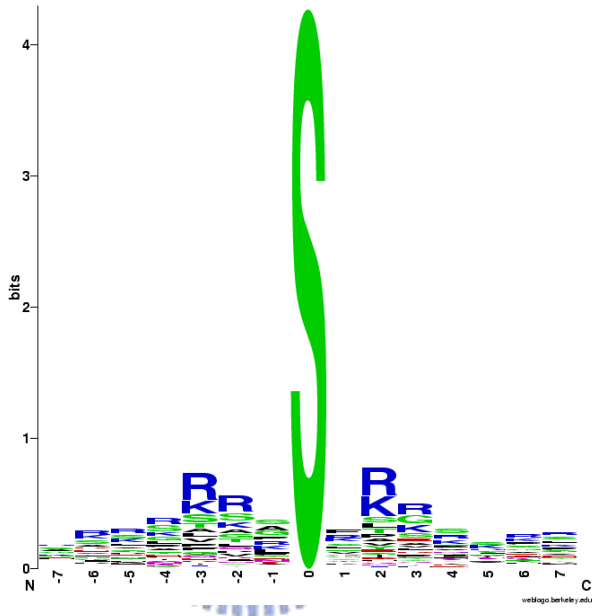

圖 A.19: PKC (S)資料的序列圖案

表 A.4: PKC (S)的 30 次 5-CV 於測試資料的效能比較表

PKC (S)		Threshold	Sensitivity	Specificity	Precision	Accuracy
δ_1	HMMer	-4.9333 (0.1246)	0.6999 (0.0228)	0.7840 (0.0210)	0.7689 (0.0164)	0.7419 (0.0121)
	HMM-1	2.5663 (0.1107)	0.7509 (0.0193)	0.7386 (0.0136)	0.7430 (0.0084)	0.7448 (0.0081)
	iHMM	2.6890 (0.0902)	0.7775 (0.0217)	0.7474 (0.0147)	0.7564 (0.0081)	0.7624 (0.0080)
δ_2	HMMer	-5.2620 (0.2602)	0.7554 (0.0276)	0.7788 (0.0291)	0.7787 (0.0190)	0.7671 (0.0080)
	HMM-1	2.5083 (0.2580)	0.7865 (0.0382)	0.7496 (0.0291)	0.7633 (0.0143)	0.7681 (0.0083)
	iHMM	2.7837 (0.1793)	0.7940 (0.0204)	0.7765 (0.0171)	0.7836 (0.0100)	0.7853 (0.0062)

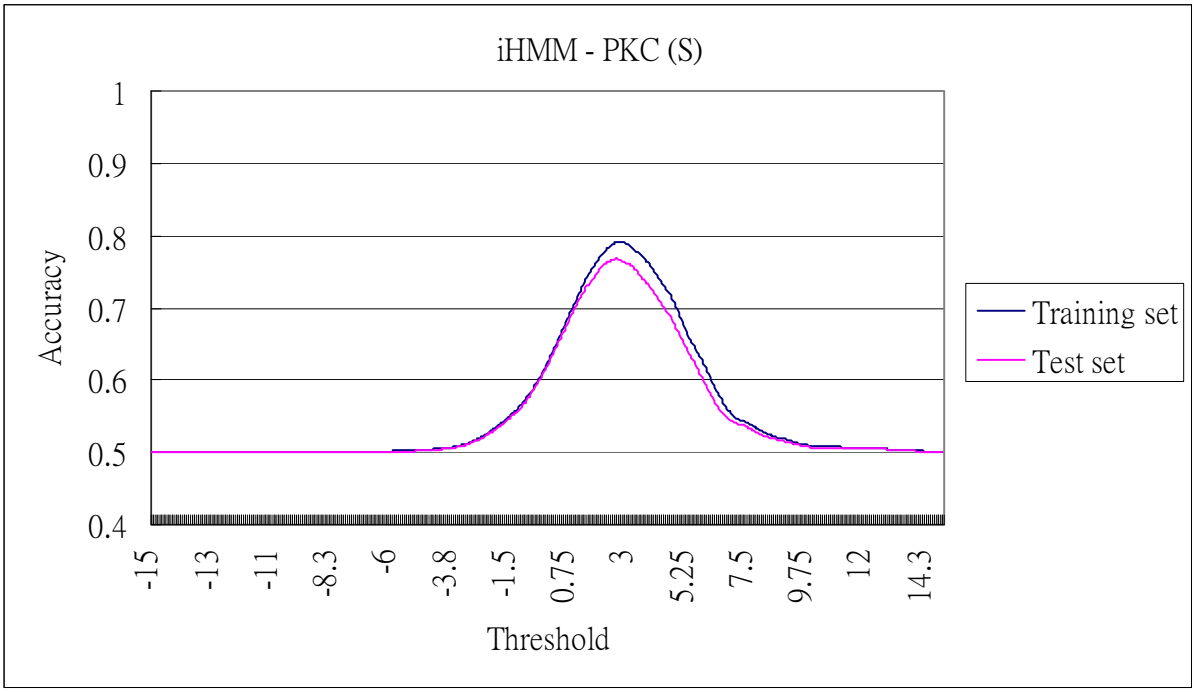

圖 A.20: PKC (S)資料於 iHMM 的門檻值與正確率對應圖

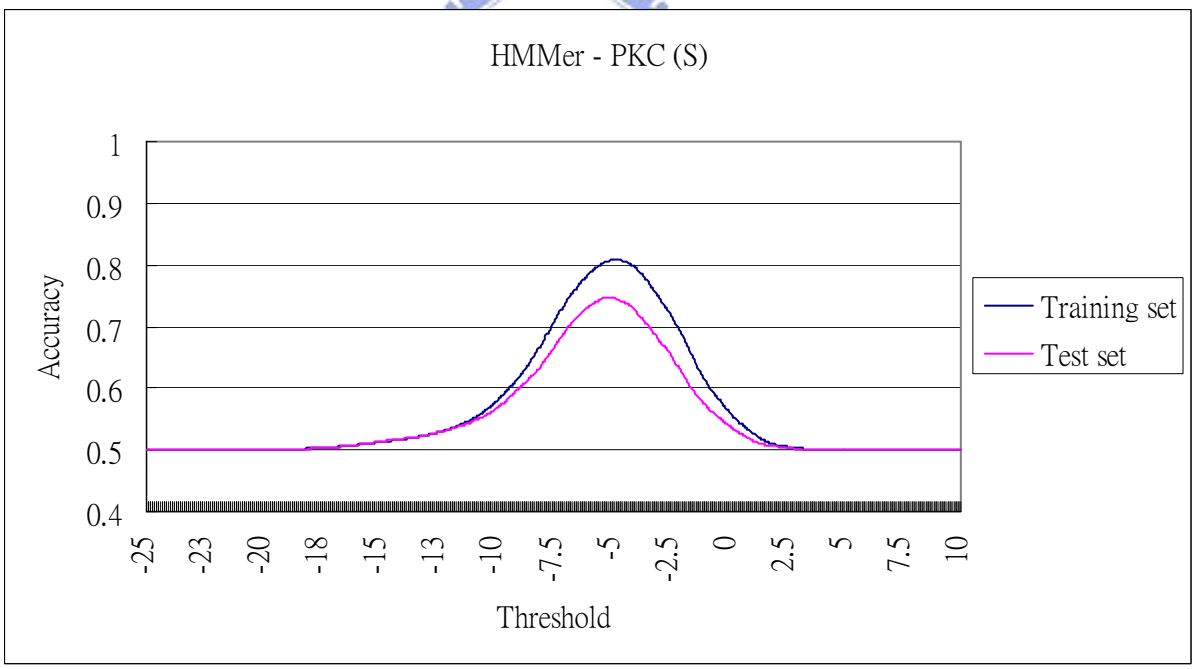

圖 A.21: PKC (S)資料於 HMMer 的門檻值與正確率對應圖

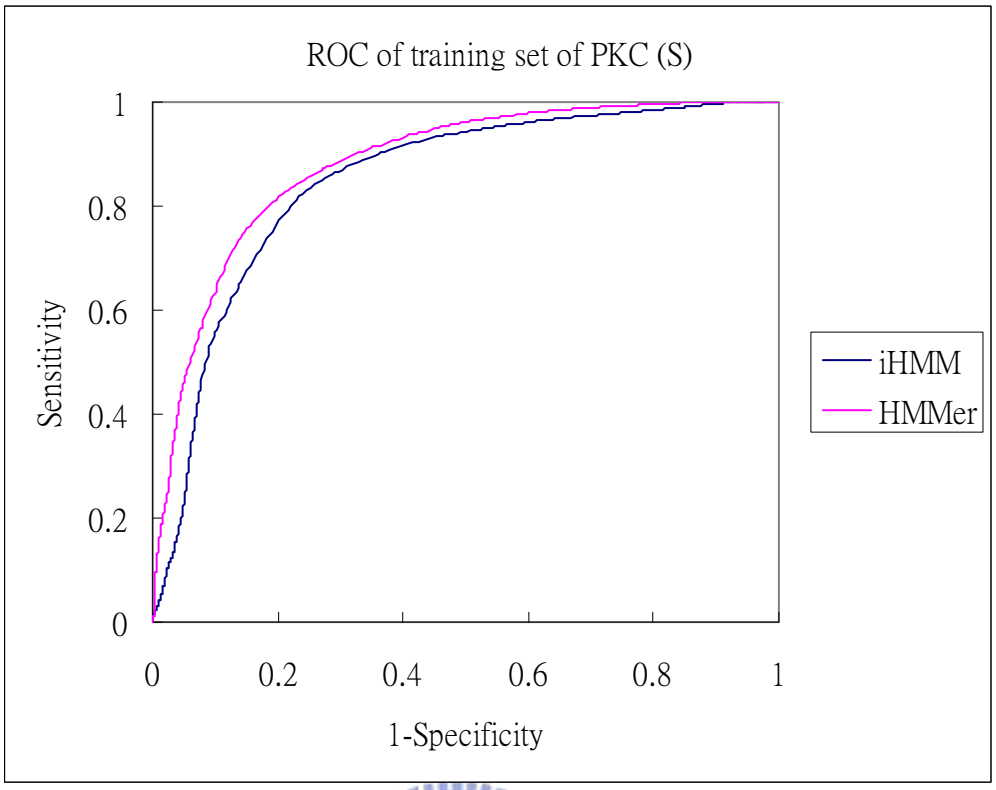

圖 A.22: PKC (S)於訓練資料的 ROC 圖

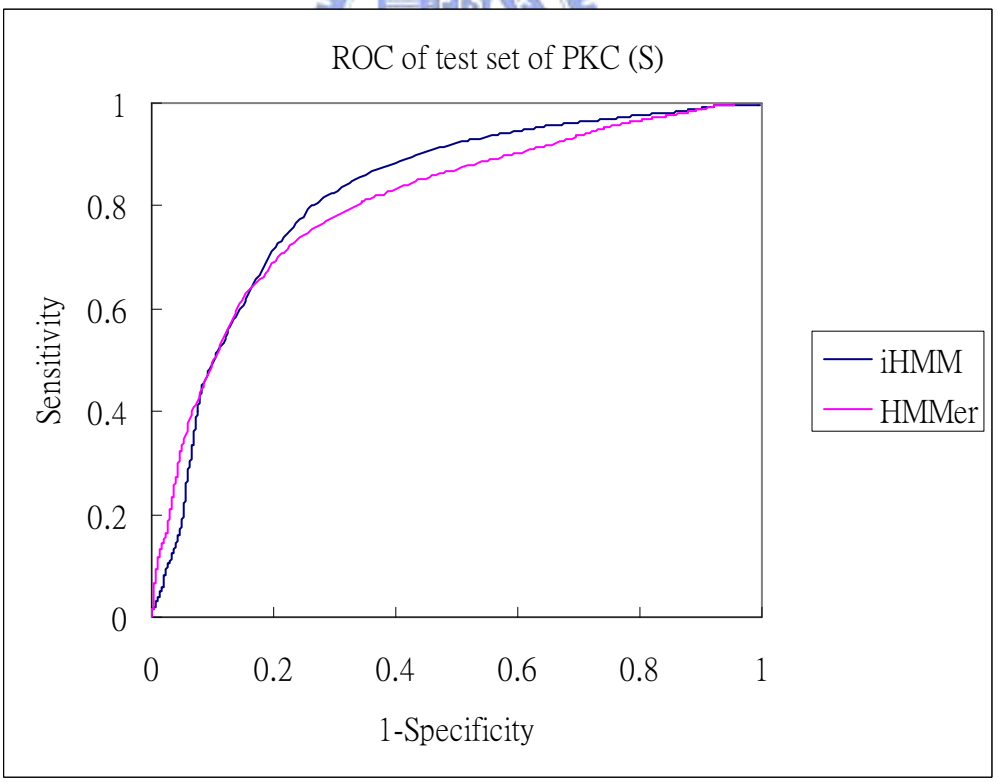

圖 A.23: PKC (S)於測試資料的 ROC 圖

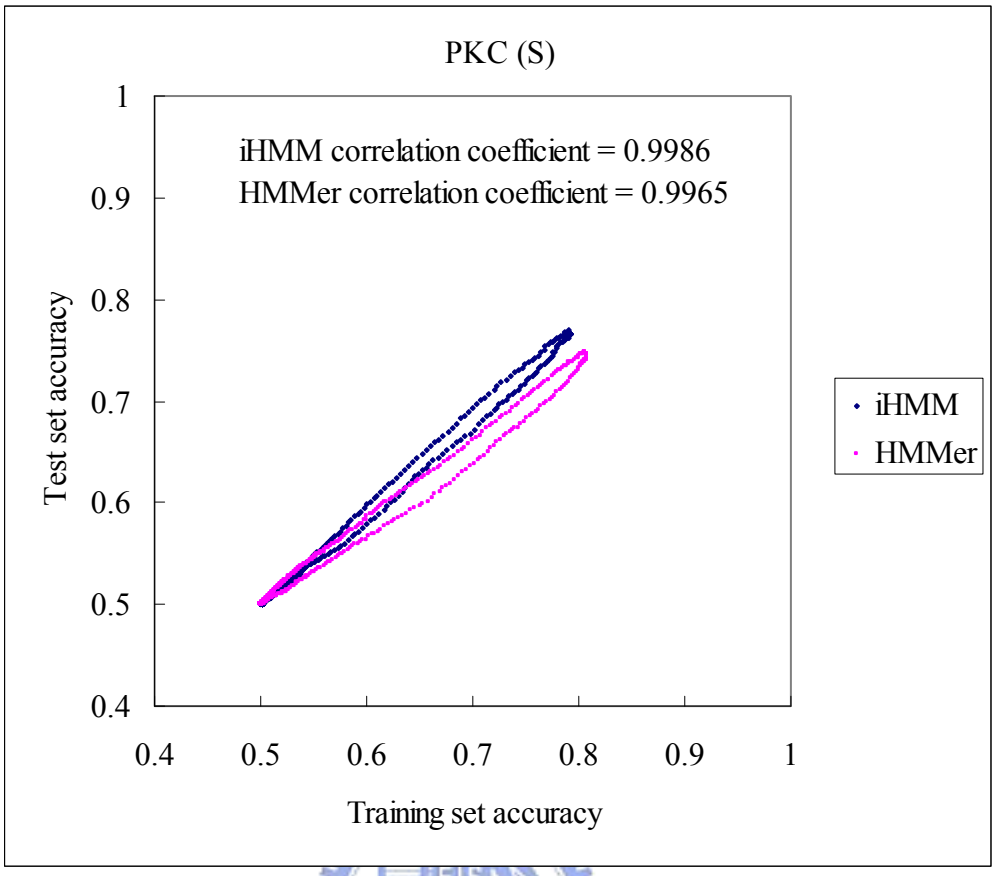

圖 A.24: PKC (S)資料的相關係數分析圖

PKC (T)

全名：Protein kinase C

資料筆數：positive 跟 negative 各 71 筆資料

序列長度：15

磷酸化位置：中間的蘇氨酸(T)

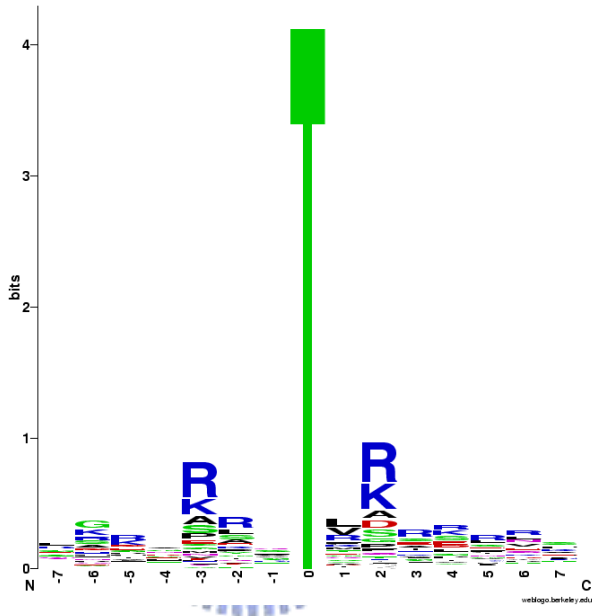

圖 A.25: PKC (T)資料的序列圖案

表 A.5: PKC (T)的 30 次 5-CV 於測試資料的效能比較表

PKC (T)		Threshold	Sensitivity	Specificity	Precision	Accuracy
δ_1	HMMer	-4.3507 (0.3379)	0.4290 (0.0364)	0.8646 (0.0289)	0.7737 (0.0454)	0.6467 (0.0213)
	HMM-1	-0.2551 (0.1877)	0.6948 (0.0482)	0.6128 (0.0402)	0.6471 (0.0178)	0.6518 (0.0154)
	iHMM	2.9929 (0.3658)	0.5505 (0.0542)	0.7845 (0.0549)	0.7351 (0.0463)	0.6670 (0.0290)
δ_2	HMMer	-8.3233 (1.2106)	0.7378 (0.0708)	0.7247 (0.0674)	0.7563 (0.0417)	0.7333 (0.0194)
	HMM-1	0.2503 (0.3221)	0.7106 (0.0628)	0.7561 (0.0537)	0.7715 (0.0342)	0.7365 (0.0153)
	iHMM	3.2749 (0.6441)	0.5968 (0.0784)	0.8580 (0.0536)	0.8414 (0.0552)	0.7292 (0.0250)

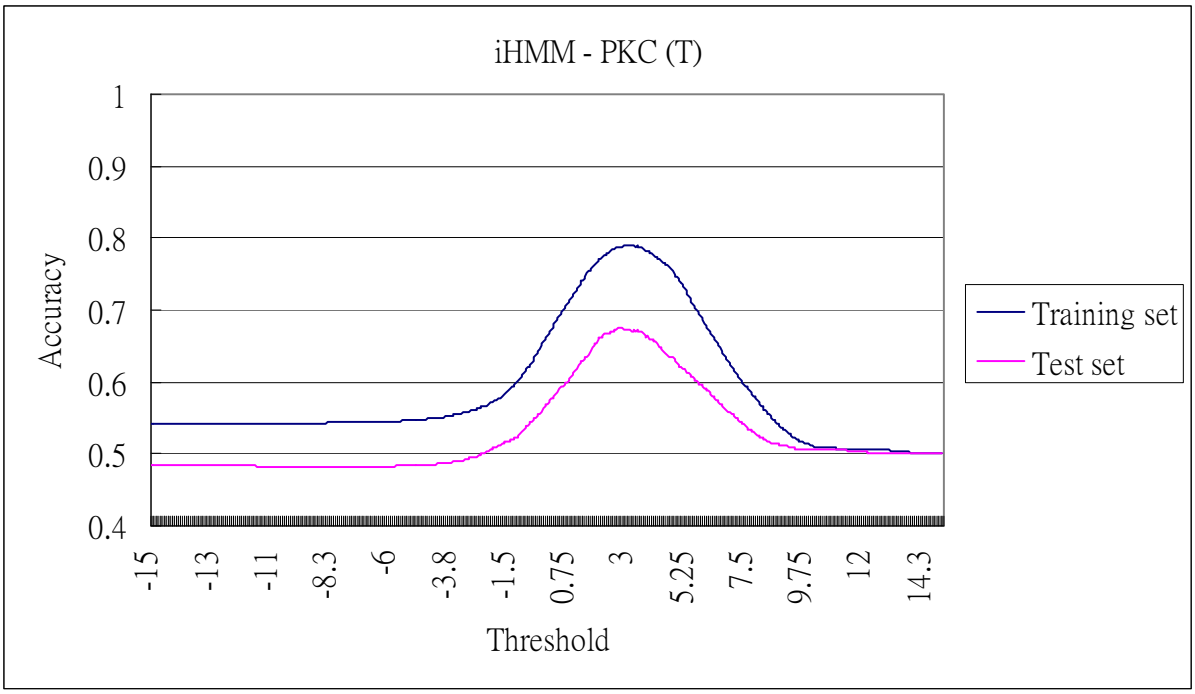

圖 A.26: PKC (T)資料於 iHMM 的門檻值與正確率對應圖

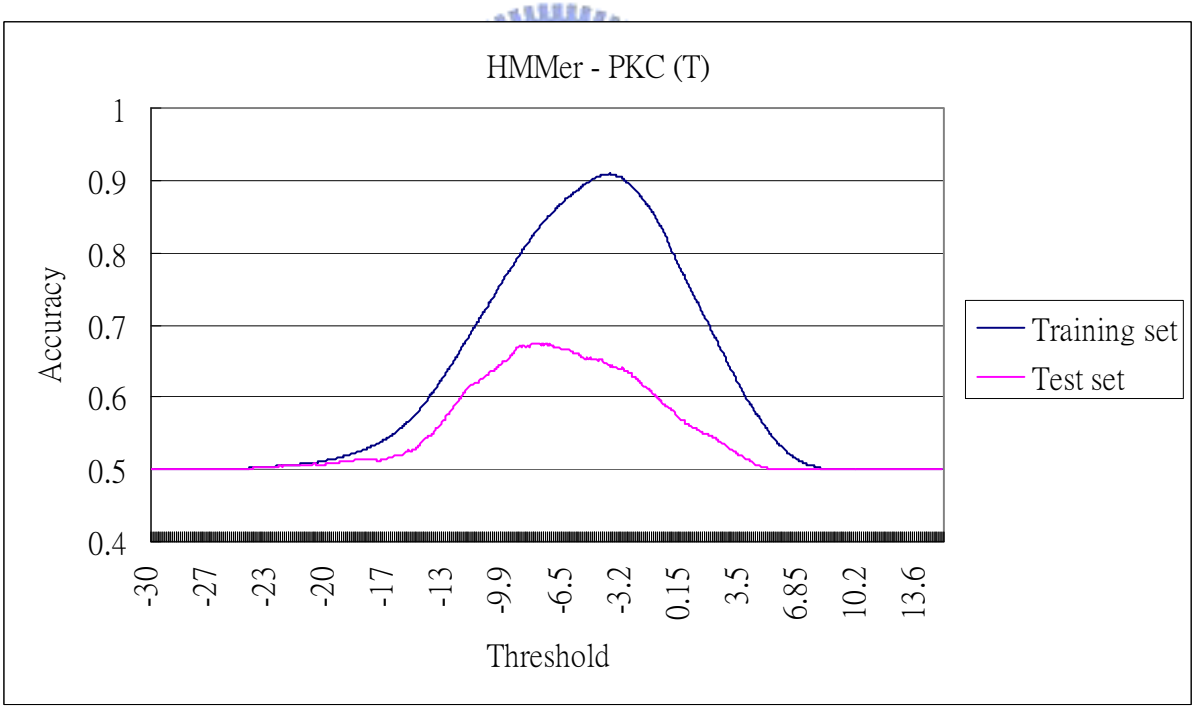

圖 A.27: PKC (T)資料於 HMMer 的門檻值與正確率對應圖

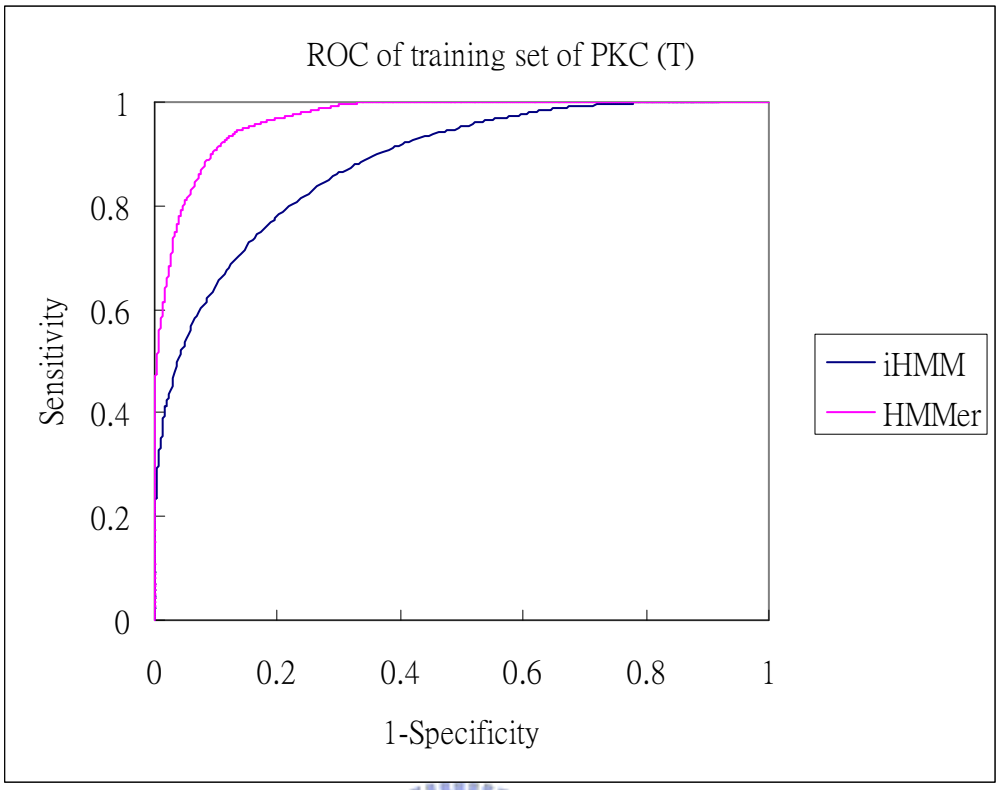

圖 A.28: PKC (T)於訓練資料的 ROC 圖

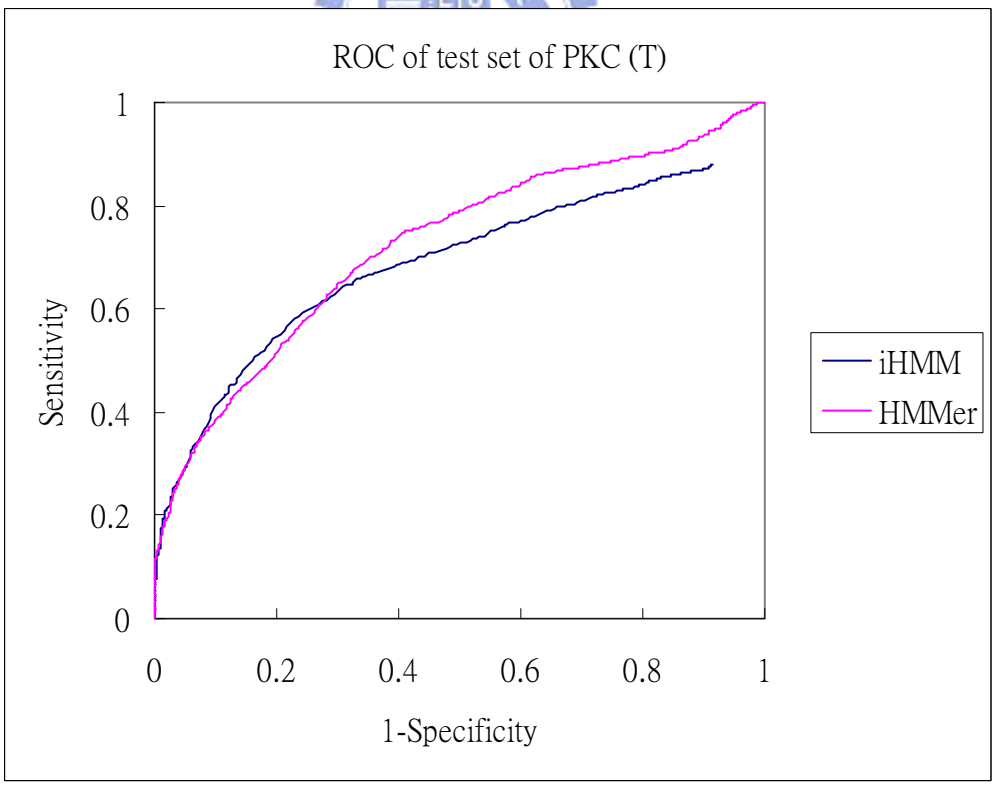

圖 A.29: PKC (T)於測試資料的 ROC 圖

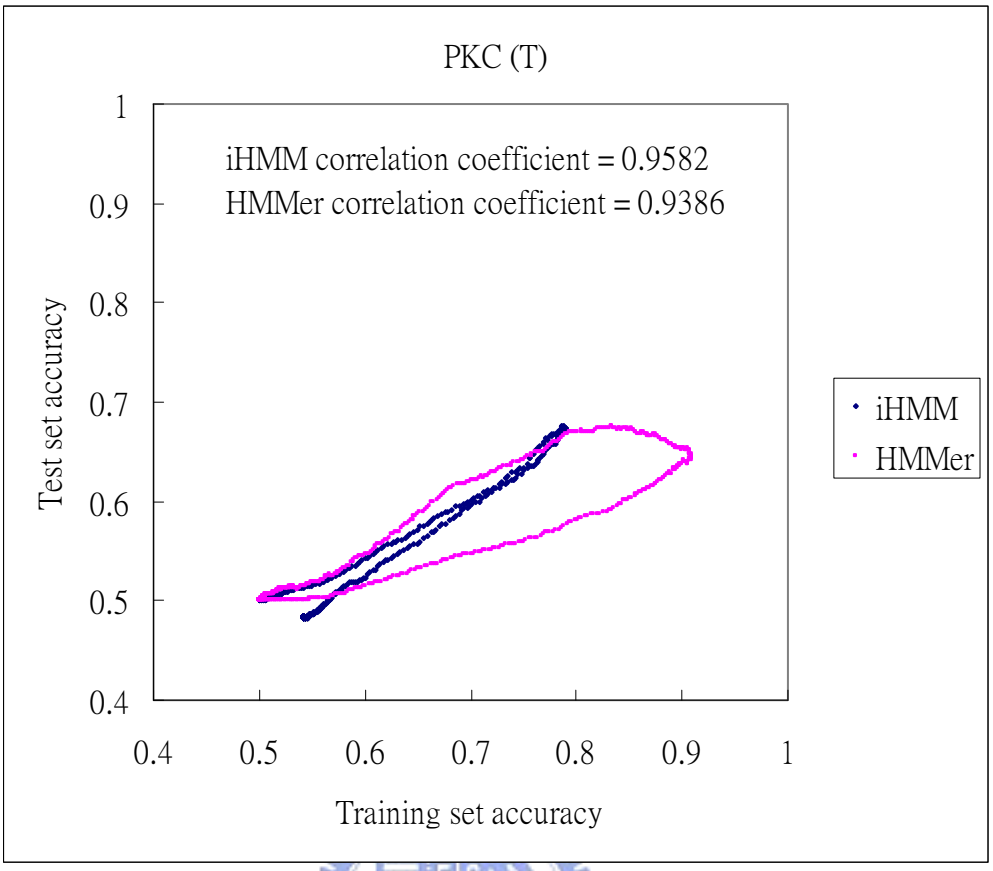

圖 A.30: PKC (T)資料的相關係數分析圖

PKG (S)

全名：Protein kinase G

資料筆數：positive 跟 negative 各 30 筆資料

序列長度：15

磷酸化位置：中間的絲氨酸(S)

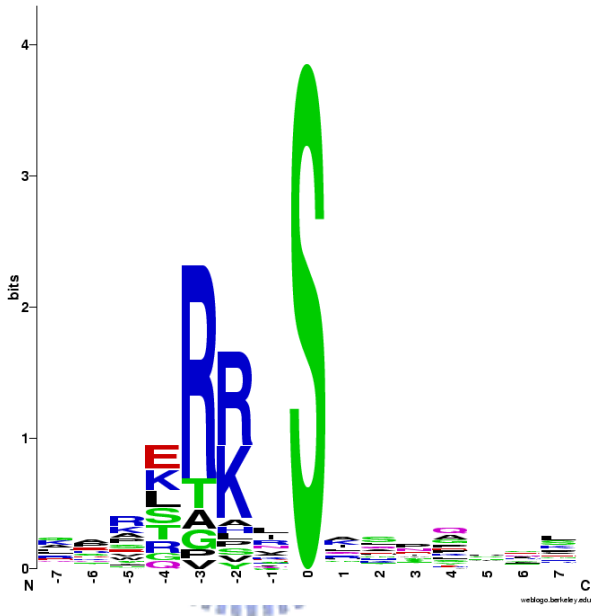

圖 A.31: PKG (S)資料的序列圖案

表 A.6: PKG (S)的 30 次 5-CV 於測試資料的效能比較表

PKG (S)		Threshold	Sensitivity	Specificity	Precision	Accuracy
δ_1	HMMer	-4.9190 (0.4089)	0.4931 (0.0675)	0.9176 (0.0350)	0.8607 (0.0822)	0.7072 (0.0373)
	HMM-1	-0.31667 (0.1448)	0.8620 (0.0645)	0.5312 (0.0458)	0.6544 (0.0296)	0.6983 (0.0320)
	iHMM	-0.31667 (0.1448)	0.8620 (0.0645)	0.5312 (0.0458)	0.6544 (0.0296)	0.6983 (0.0320)
δ_2	HMMer	-13.0930 (1.4398)	0.9030 (0.0542)	0.8055 (0.0752)	0.8551 (0.0489)	0.8658 (0.0243)
	HMM-1	1.4277 (1.3280)	0.7972 (0.1074)	0.7314 (0.0952)	0.7560 (0.0845)	0.7894 (0.0241)
	iHMM	1.4277 (1.3280)	0.7972 (0.1074)	0.7314 (0.0952)	0.7560 (0.0845)	0.7894 (0.0241)

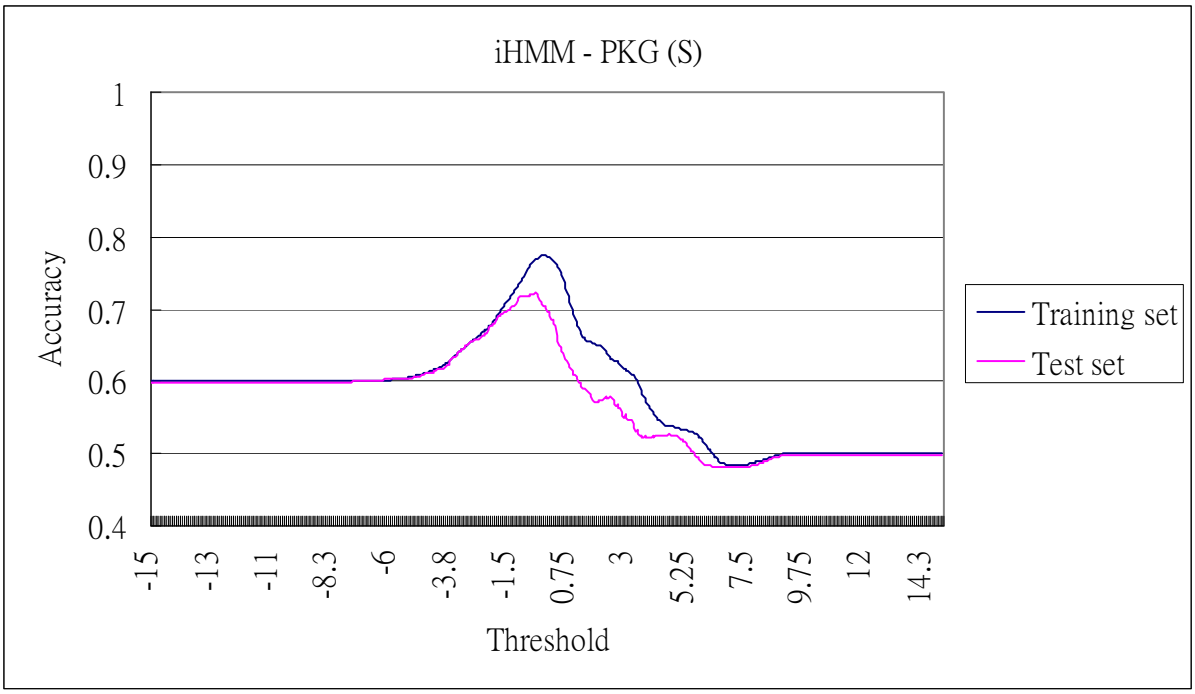

圖 A.32: PKG (S)資料於 iHMM 的門檻值與正確率對應圖

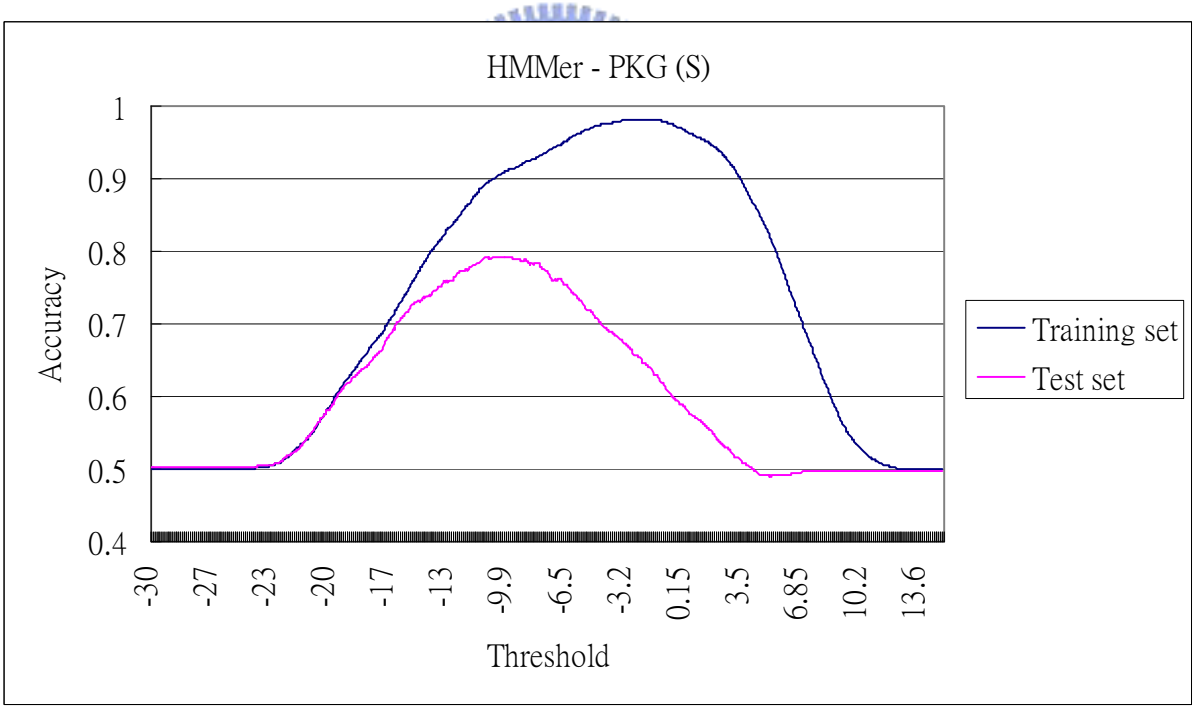

圖 A.33: PKG (S)資料於 HMMer 的門檻值與正確率對應圖

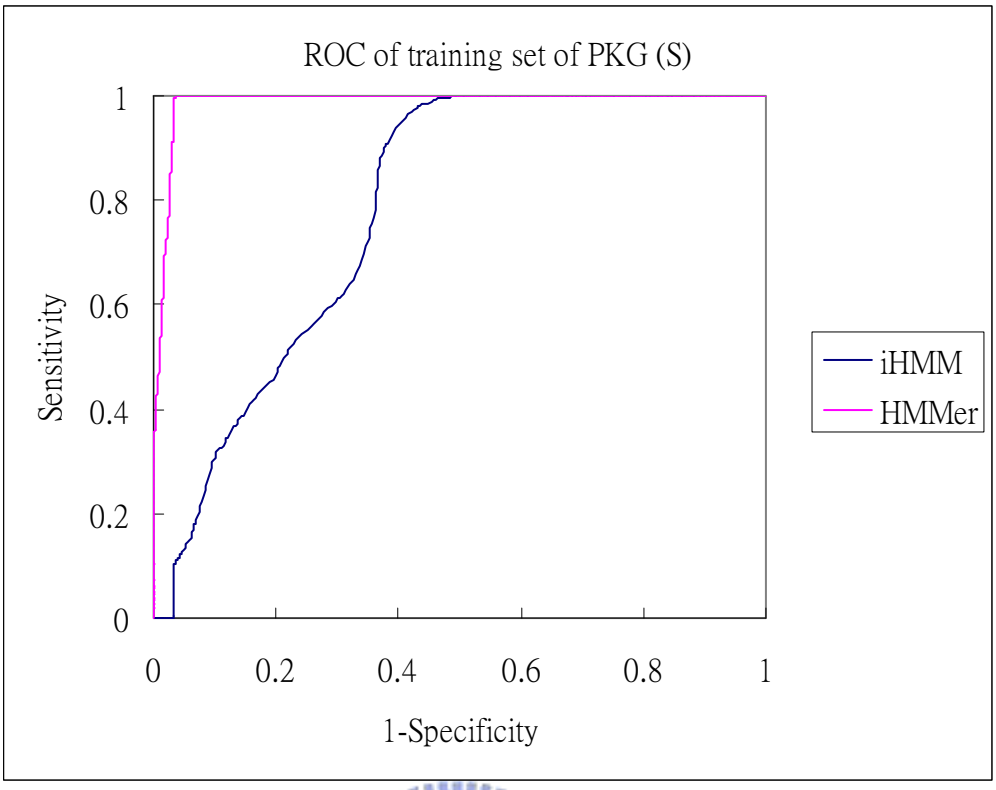

圖 A.34: PKG (S)於訓練資料的 ROC 圖

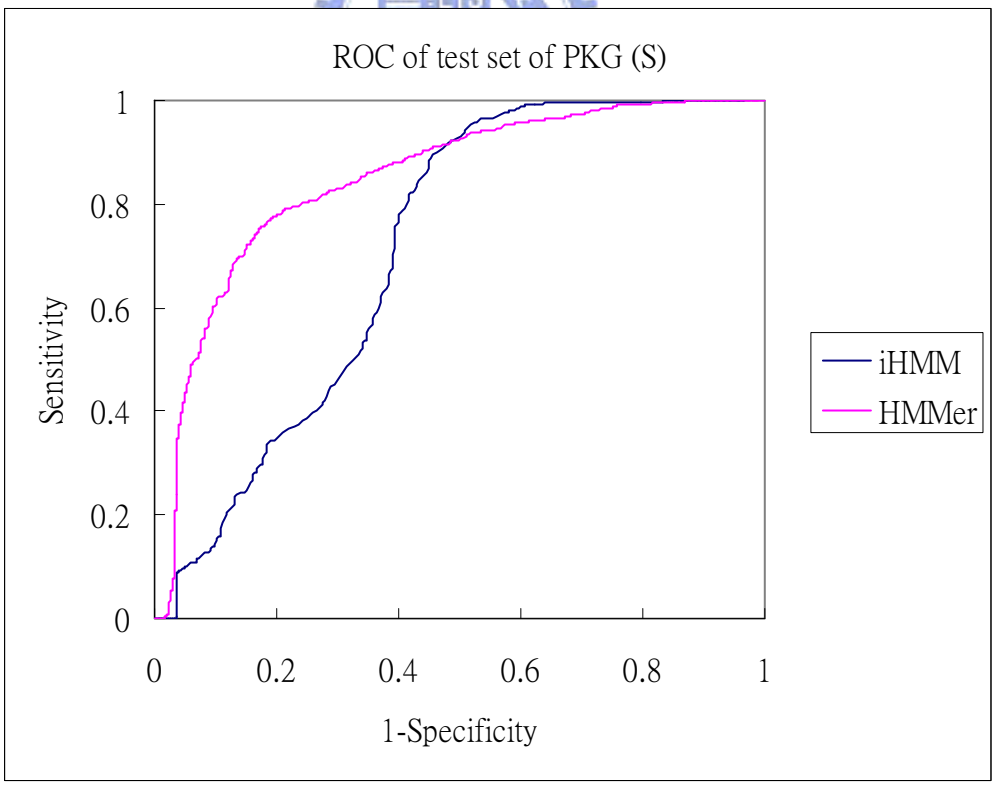

圖 A.35: PKG (S)於測試資料的 ROC 圖

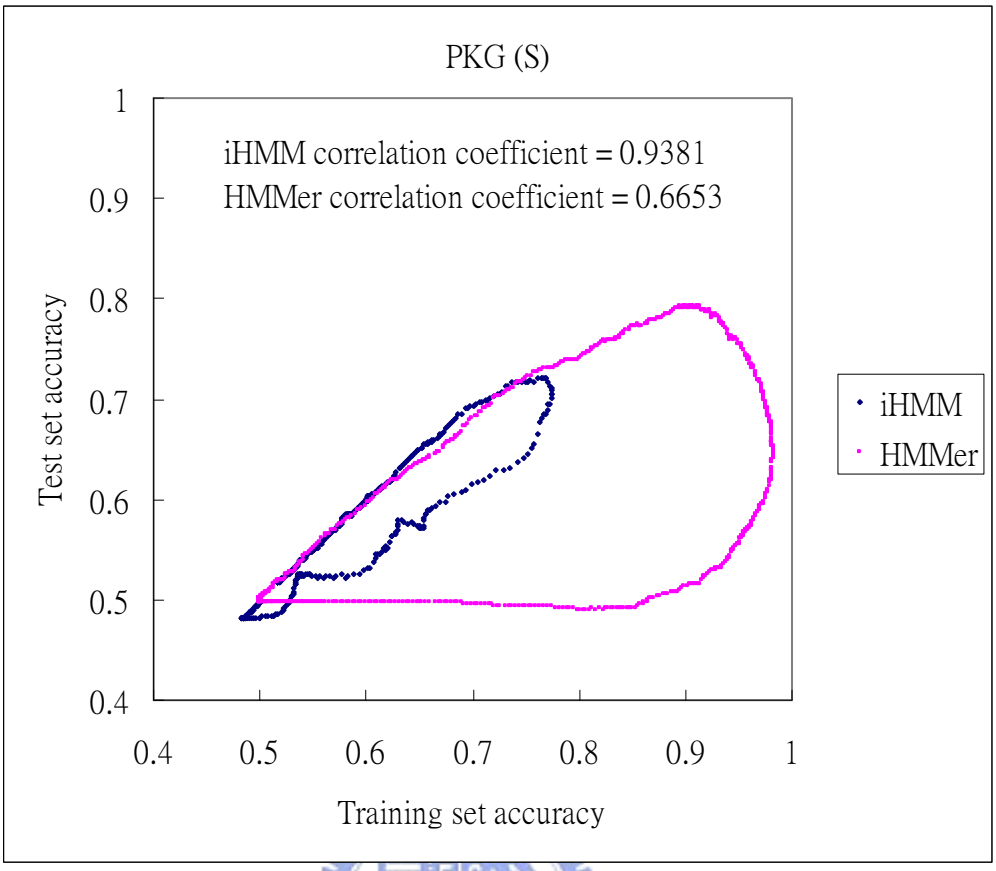

圖 A.36: PKG (S)資料的相關係數分析圖

CDK (S)

全名：Cyclin-dependent kinase

資料筆數：positive 跟 negative 各 195 筆資料

序列長度：15

磷酸化位置：中間的絲氨酸(S)

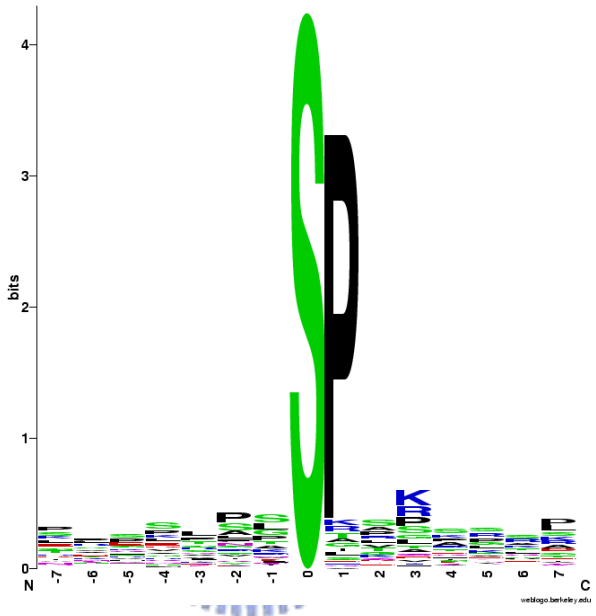

圖 A.37: CDK (S)資料的序列圖案

表 A.7: CDK (S)的 30 次 5-CV 於測試資料的效能比較表

CDK (S)		Threshold	Sensitivity	Specificity	Precision	Accuracy
δ_1	HMMer	-4.0093 (0.2652)	0.7731 (0.0220)	0.8739 (0.0135)	0.8634 (0.0119)	0.8237 (0.0081)
	HMM-1	1.8995 (0.2686)	0.8117 (0.0233)	0.8155 (0.0232)	0.8186 (0.0169)	0.8137 (0.0119)
	iHMM	3.3438 (0.2076)	0.8218 (0.0209)	0.8615 (0.0189)	0.8591 (0.0142)	0.8414 (0.0111)
δ_2	HMMer	-5.0437 (0.4899)	0.8456 (0.0232)	0.8558 (0.0214)	0.8588 (0.0152)	0.8507 (0.0073)
	HMM-1	1.9771 (0.3029)	0.8326 (0.0227)	0.8361 (0.0190)	0.8387 (0.0142)	0.8344 (0.0090)
	iHMM	3.6848 (0.2914)	0.8279 (0.0244)	0.8979 (0.0174)	0.8929 (0.0150)	0.8632 (0.0089)

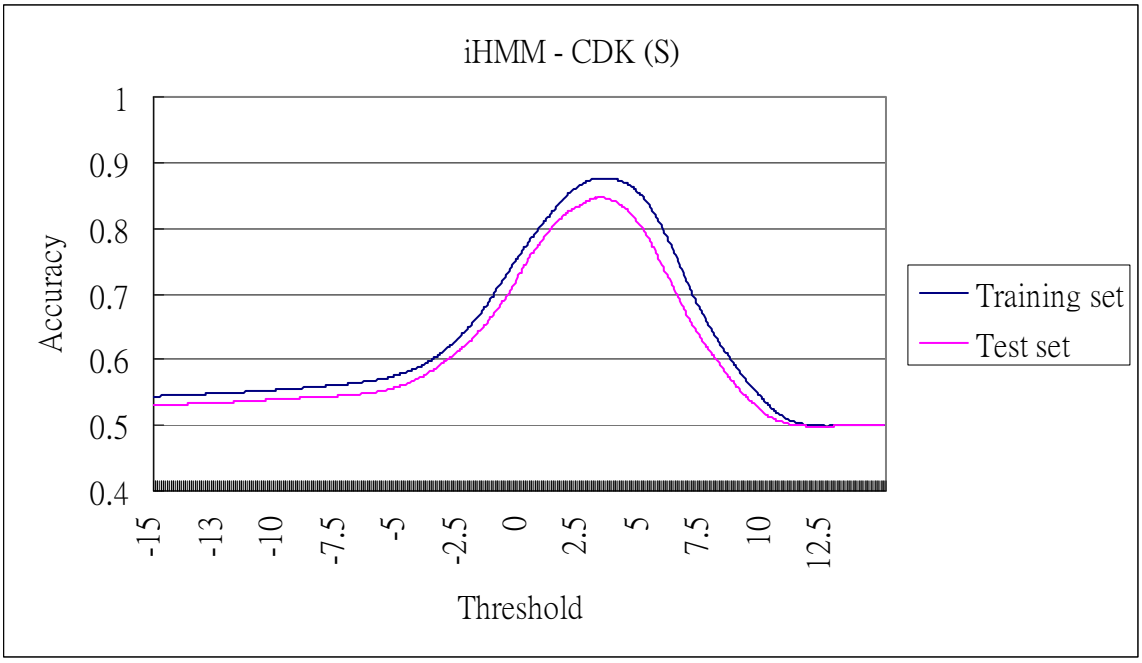

圖 A.38: CDK (S)資料於 iHMM 的門檻值與正確率對應圖

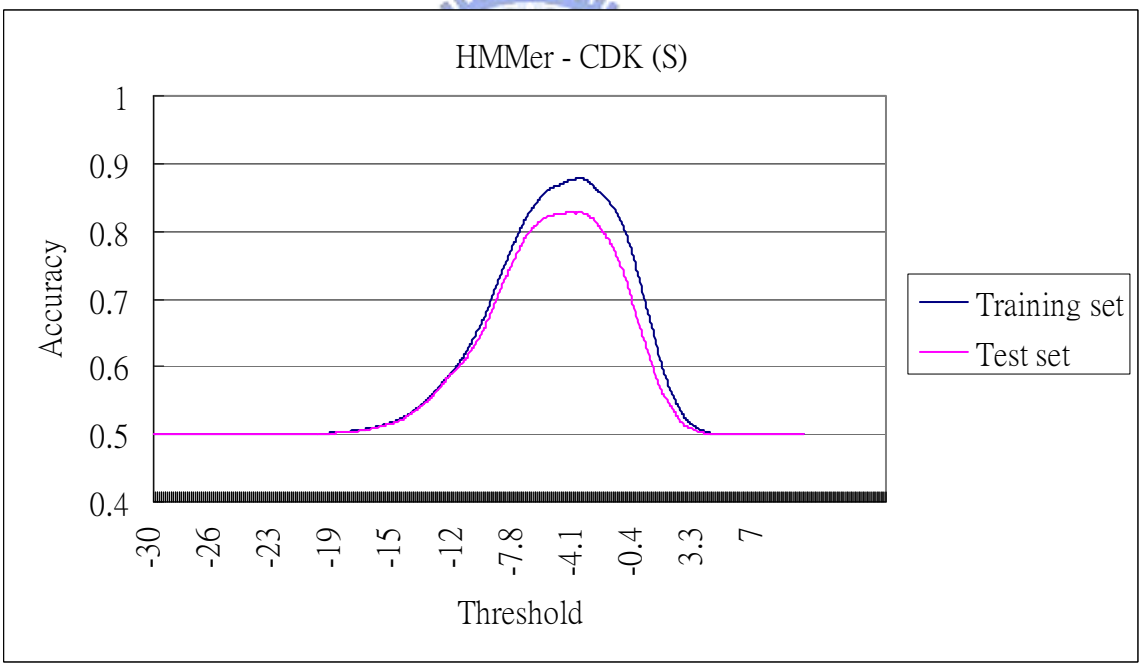

圖 A.39: CDK (S)資料於 HMMer 的門檻值與正確率對應圖

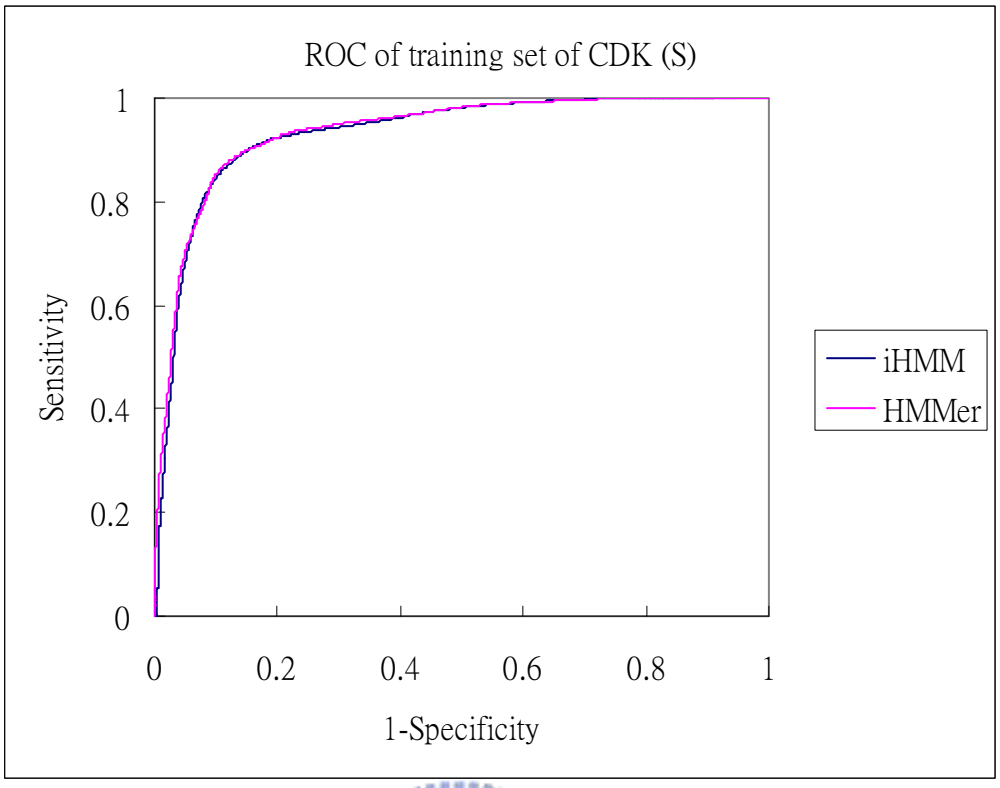

圖 A.40: CDK (S)於訓練資料的 ROC 圖

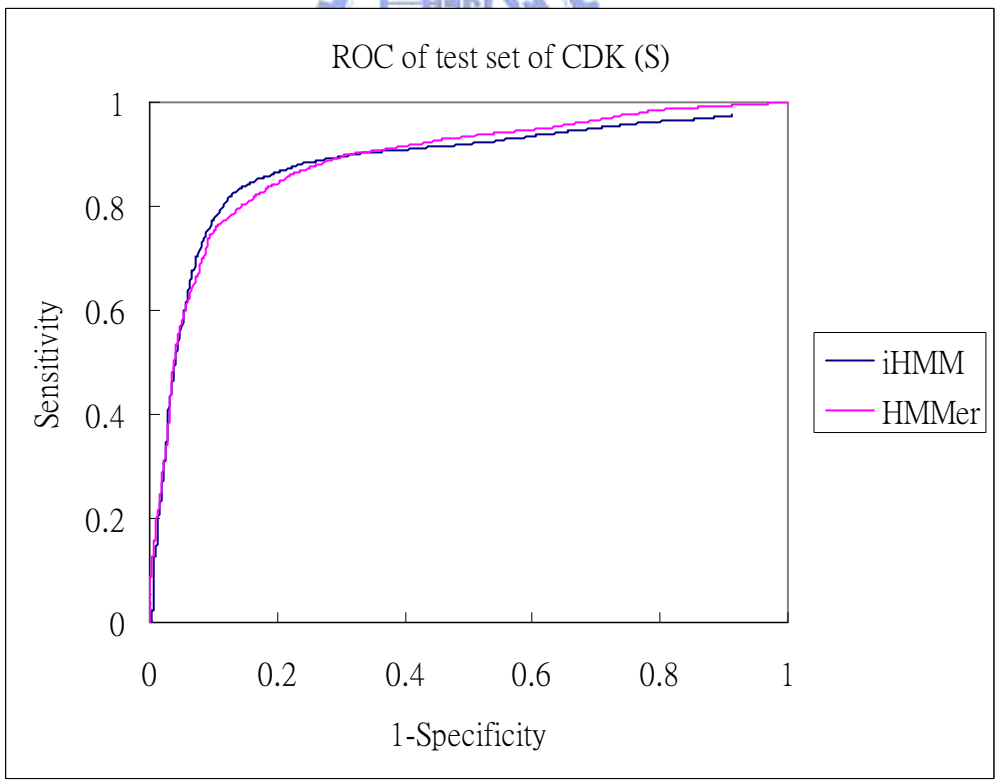

圖 A.41: CDK (S)於測試資料的 ROC 圖

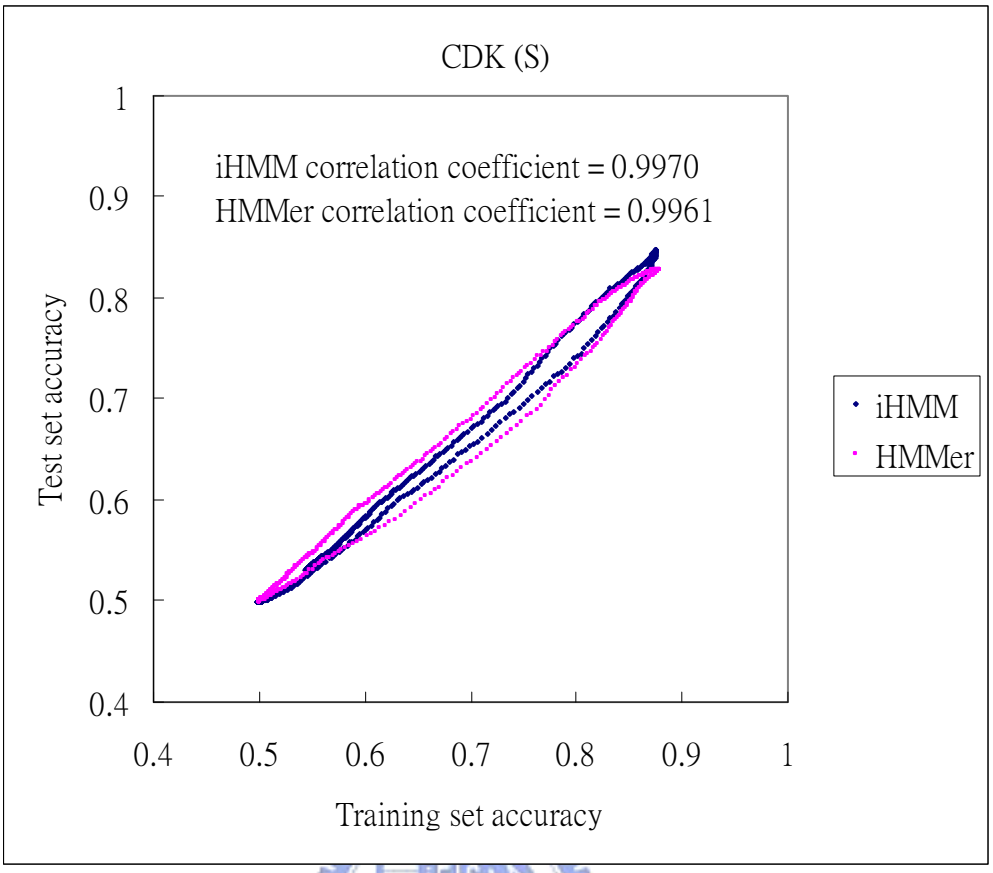

圖 A.42: CDK (S)資料的相關係數分析圖

CDK (T)

全名：Cyclin-dependent kinase

資料筆數：positive 跟 negative 各 113 筆資料

序列長度：15

磷酸化位置：中間的蘇氨酸(T)

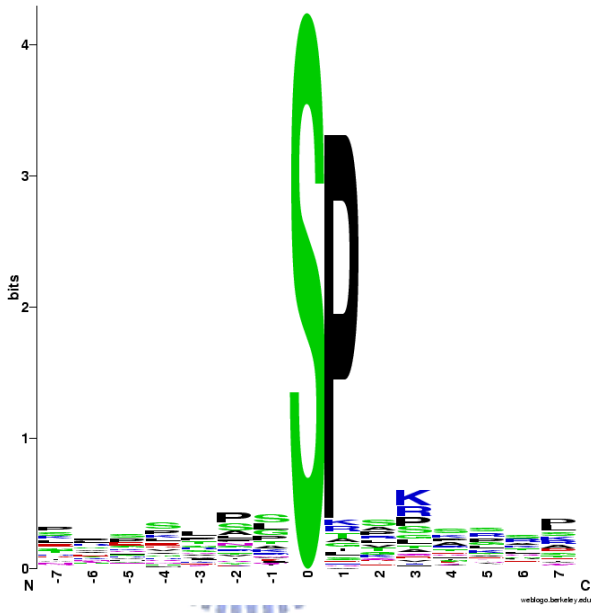

圖 A.43: CDK (T)資料的序列圖案

表 A.8: CDK (T)的 30 次 5-CV 於測試資料的效能比較表

CDK (T)		Threshold	Sensitivity	Specificity	Precision	Accuracy
δ_1	HMMer	-4.6577 (0.5060)	0.7846 (0.0400)	0.9136 (0.0161)	0.9068 (0.0159)	0.8493 (0.0196)
	HMM-1	1.0918 (0.1558)	0.8360 (0.0220)	0.7897 (0.0189)	0.8031 (0.0139)	0.8127 (0.0114)
	iHMM	2.5223 (0.3060)	0.8666 (0.0236)	0.8950 (0.0202)	0.8967 (0.0174)	0.8810 (0.0147)
δ_2	HMMer	-7.3247 (0.6255)	0.9194 (0.0251)	0.8813 (0.0252)	0.8916 (0.0202)	0.9007 (0.0128)
	HMM-1	1.6198 (0.2431)	0.8333 (0.0278)	0.8618 (0.0218)	0.8662 (0.0145)	0.8480 (0.0066)
	iHMM	2.6073 (2.1345)	0.8810 (0.0338)	0.9314 (0.0200)	0.9312 (0.0176)	0.9064 (0.0147)

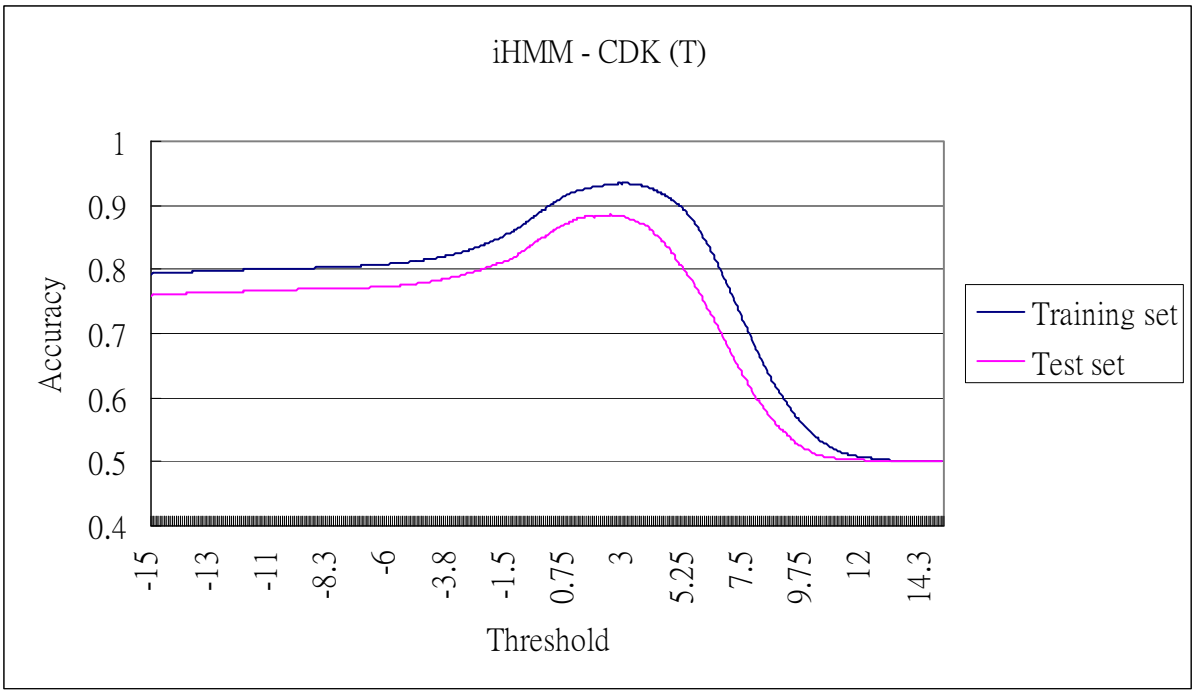

圖 A.44: CDK (T)資料於 iHMM 的門檻值與正確率對應圖

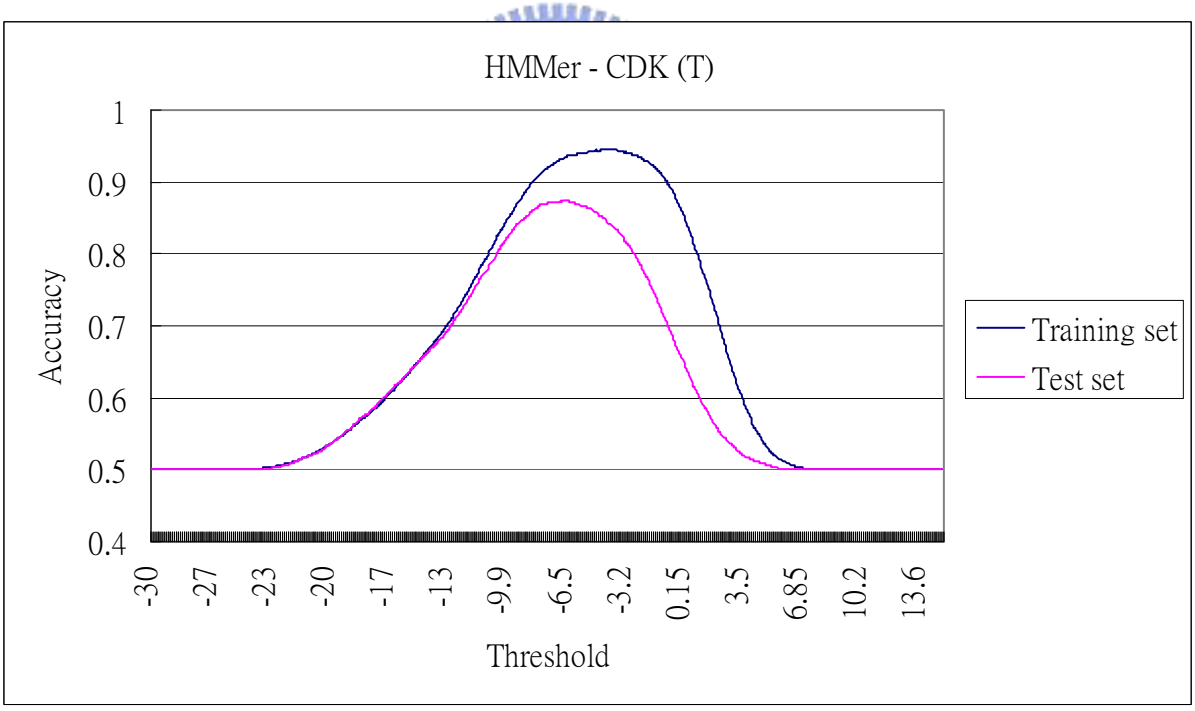

圖 A.45: CDK (T)資料於 HMMer 的門檻值與正確率對應圖

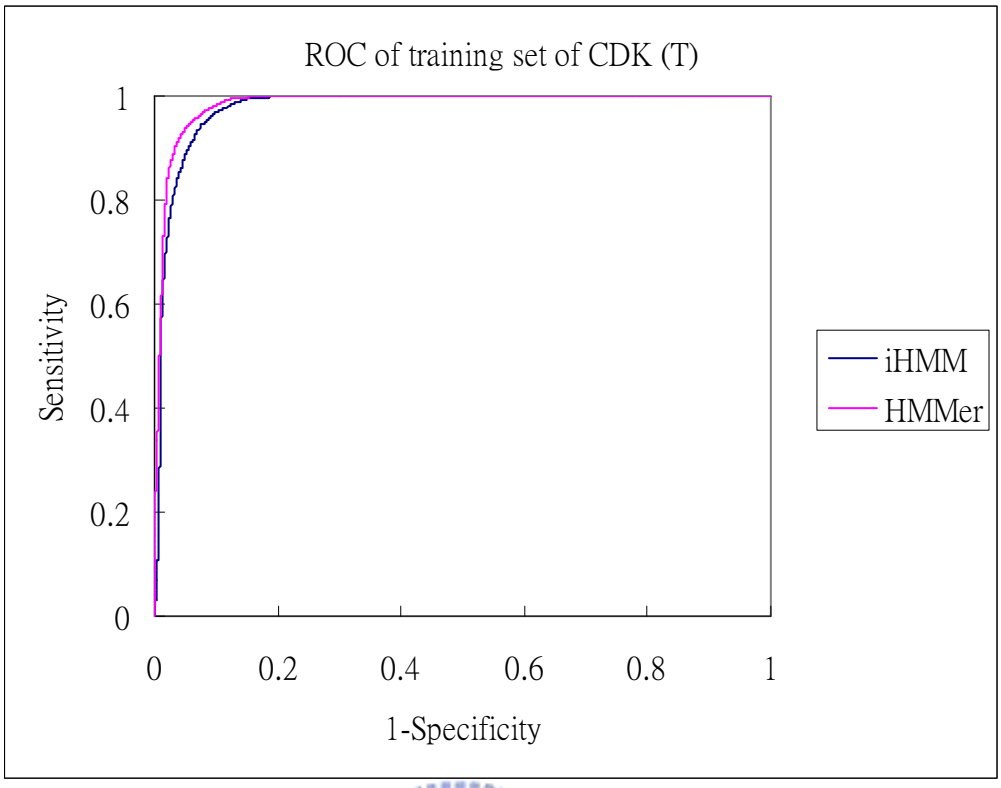

圖 A.46: CDK (T)於訓練資料的 ROC 圖

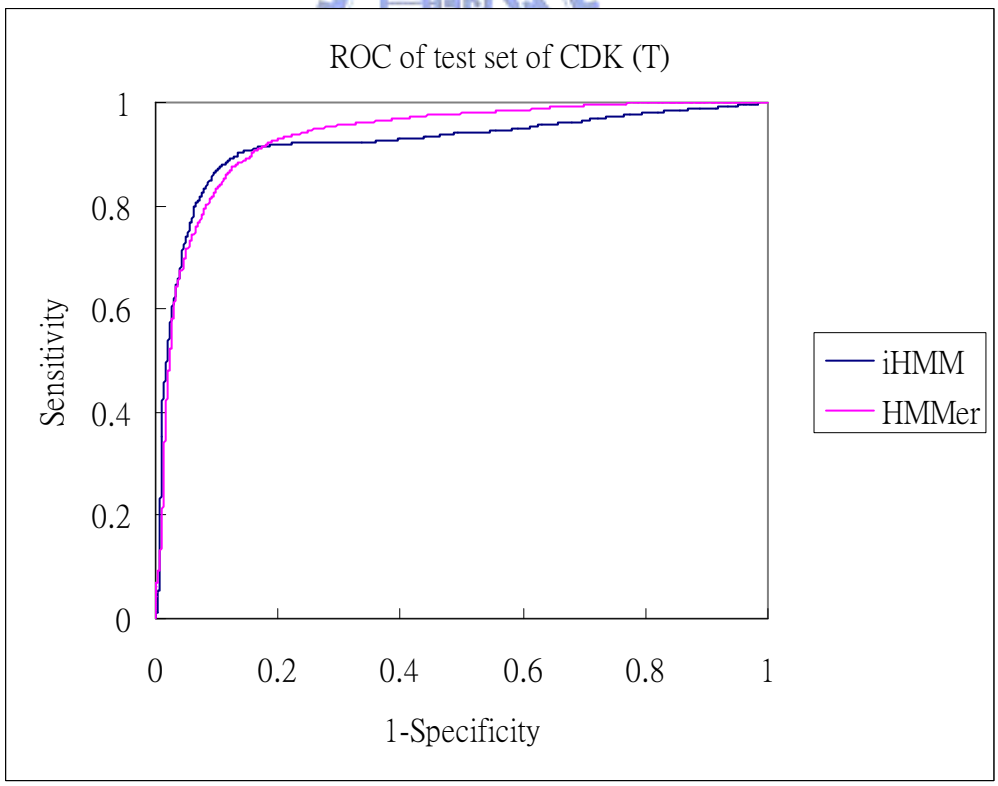

圖 A.47: CDK (T)於測試資料的 ROC 圖

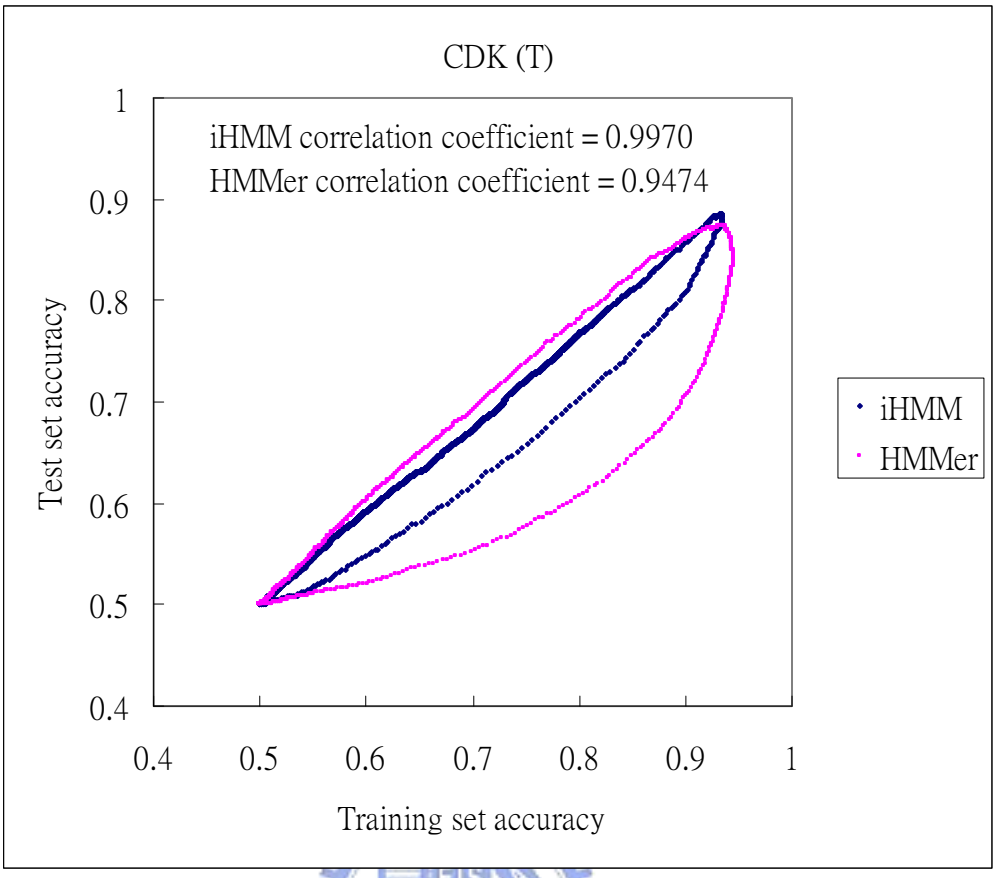

圖 A.48: CDK (T)資料的相關係數分析圖

CaM-KII (S)

全名：Calcium/calmodulin-dependent protein kinase II

資料筆數：positive 跟 negative 各 76 筆資料

序列長度：15

磷酸化位置：中間的絲氨酸(S)

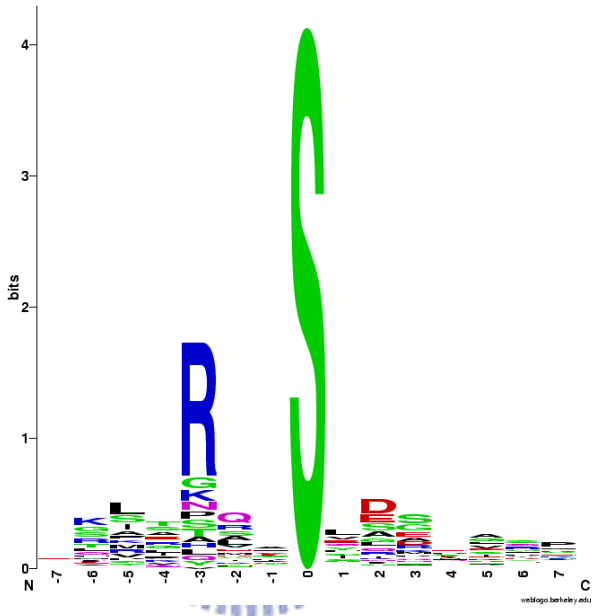

圖 A.49: CaM-KII (S)資料的序列圖案

表 A.9: CaM-KII (S)的 30 次 5-CV 於測試資料的效能比較表

CaM-KII (S)		Threshold	Sensitivity	Specificity	Precision	Accuracy
δ_1	HMMer	-4.4003 (0.3006)	0.4832 (0.0358)	0.9017 (0.0218)	0.8432 (0.0343)	0.6926 (0.0177)
	HMM-1	0.5795 (0.1389)	0.6551 (0.0466)	0.6270 (0.0460)	0.6497 (0.0286)	0.6400 (0.0270)
	iHMM	1.4928 (0.1104)	0.7088 (0.0329)	0.7869 (0.0353)	0.7775 (0.0299)	0.7473 (0.0217)
δ_2	HMMer	-8.0937 (1.0199)	0.7672 (0.0650)	0.7739 (0.0712)	0.8018 (0.0491)	0.7740 (0.0148)
	HMM-1	0.8168 (0.4232)	0.6911 (0.0613)	0.7166 (0.0633)	0.7244 (0.0475)	0.7078 (0.0187)
	iHMM	1.7641 (0.2312)	0.7269 (0.0432)	0.8659 (0.0340)	0.8572 (0.0294)	0.7971 (0.0137)

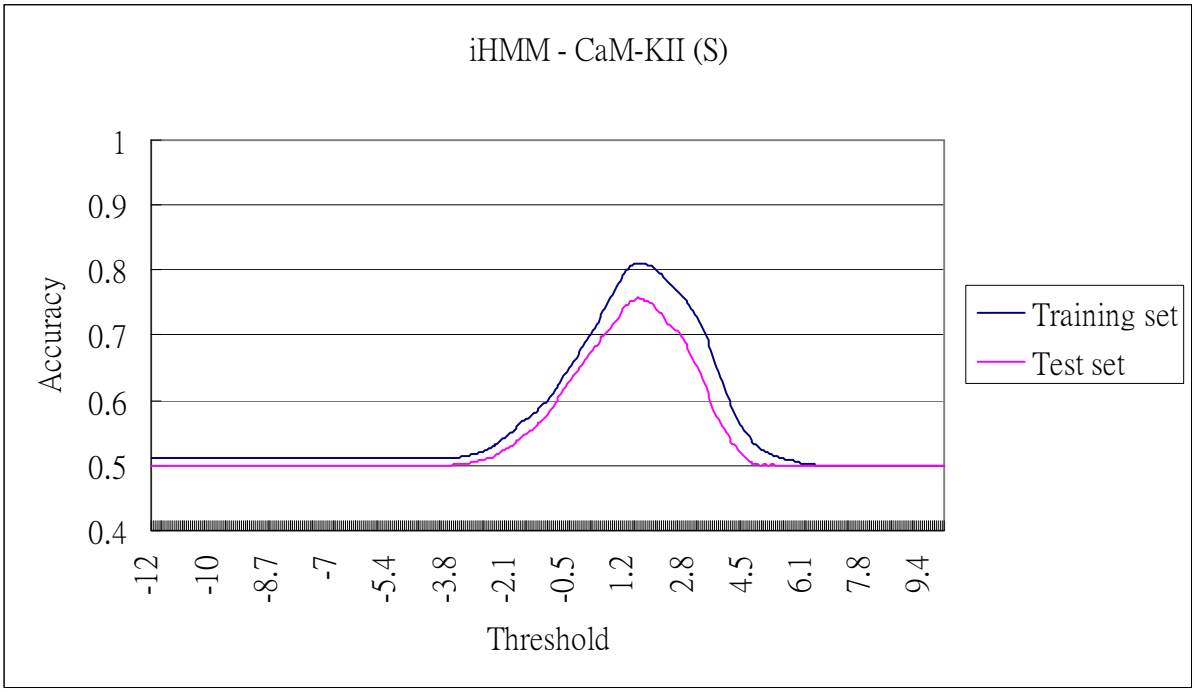

圖 A.50: CaM-KII (S)於 iHMM 的門檻值與正確率對應圖

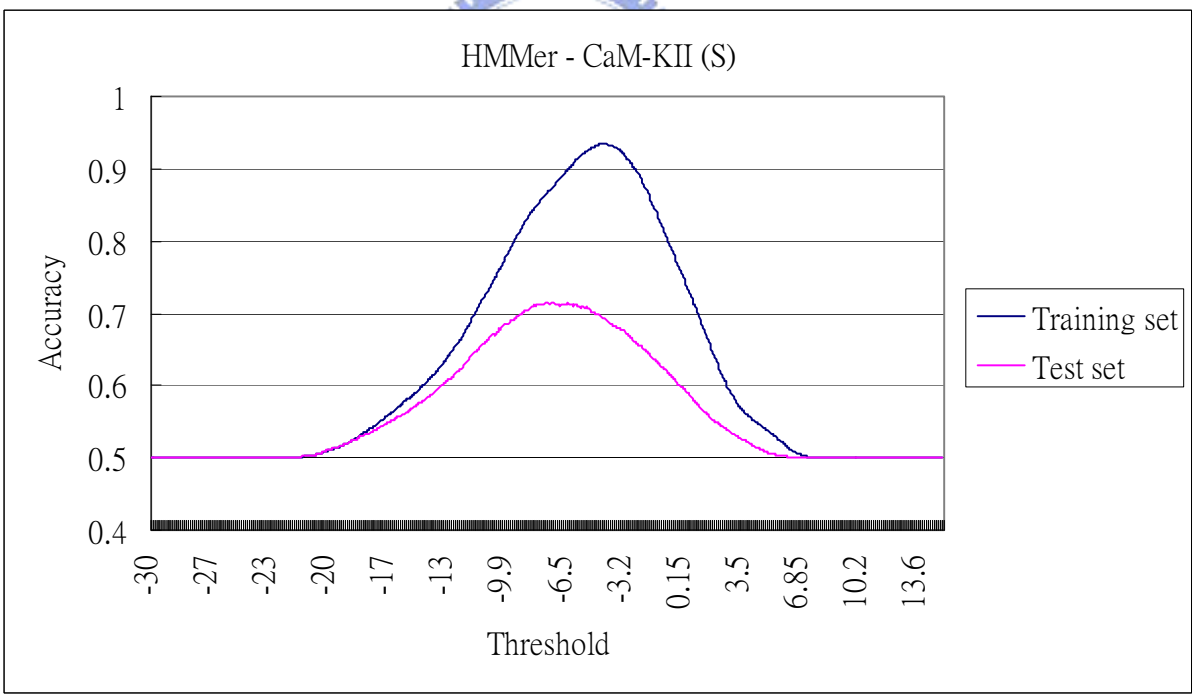

圖 A.51: CaM-KII (S)於 HMMer 的門檻值與正確率對應圖

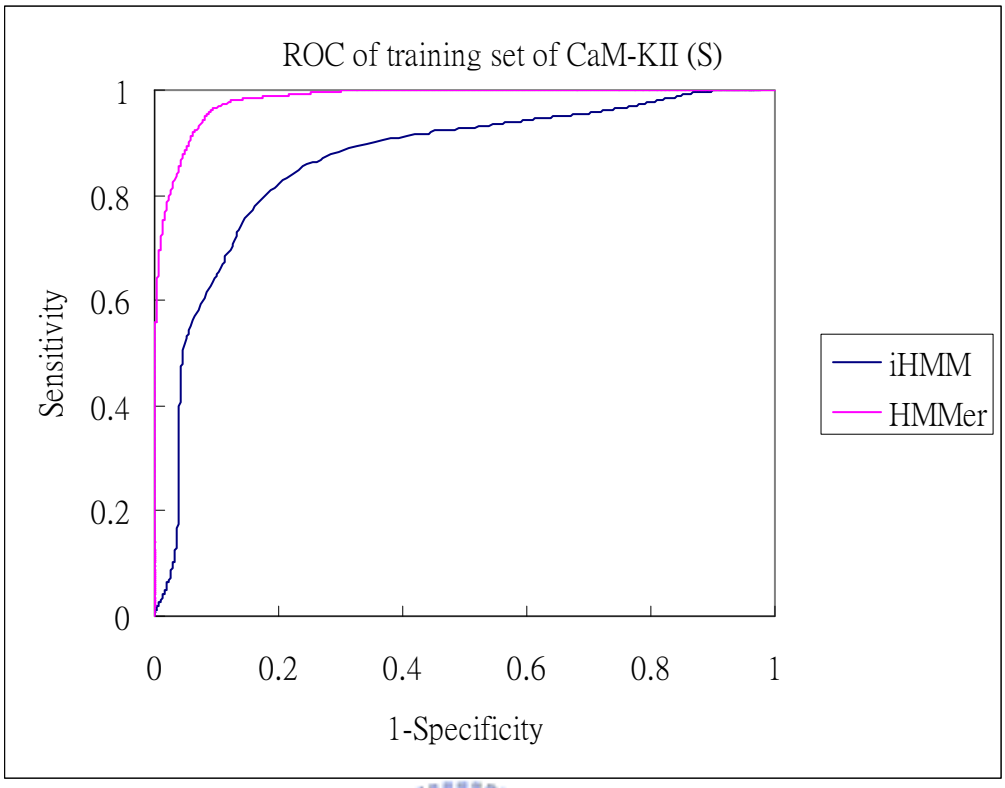

圖 A.52: CaM-KII (S)於訓練資料的 ROC 圖

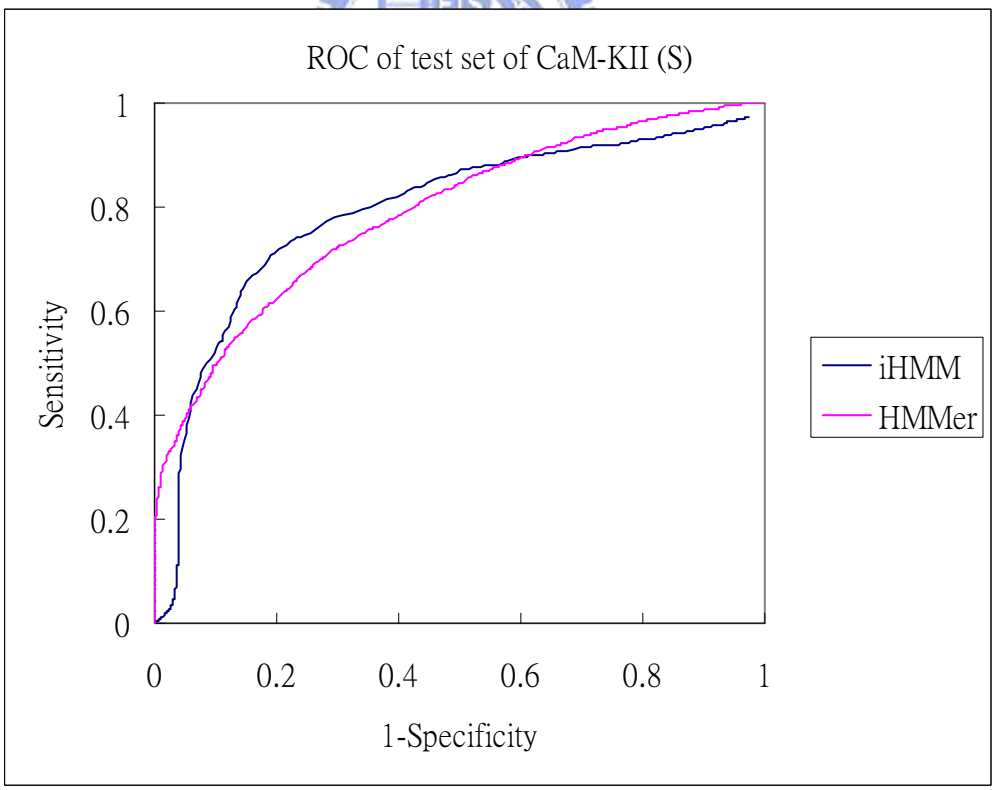

圖 A.53: CaM-KII (S)於測試資料的 ROC 圖

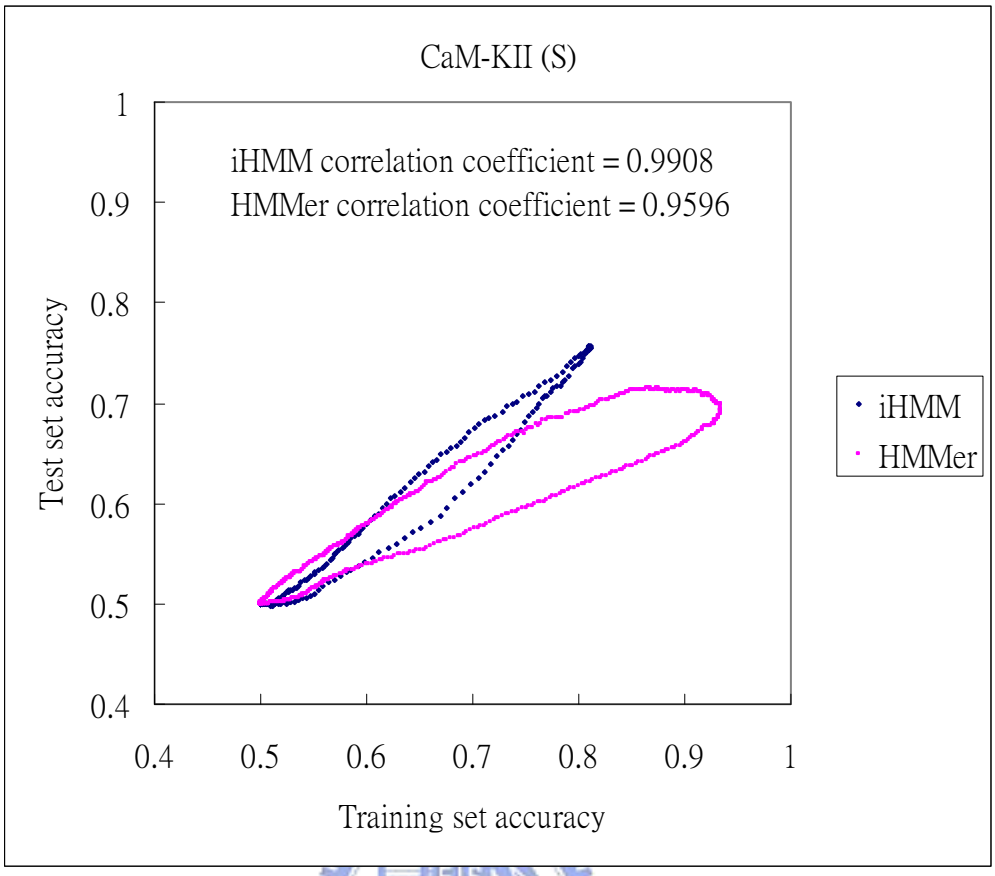

圖 A.54: CaM-KII (S)資料的相關係數分析圖

CK1 (S)

全名：Casein kinase I

資料筆數：positive 跟 negative 各 49 筆資料

序列長度：15

磷酸化位置：中間的絲氨酸(S)

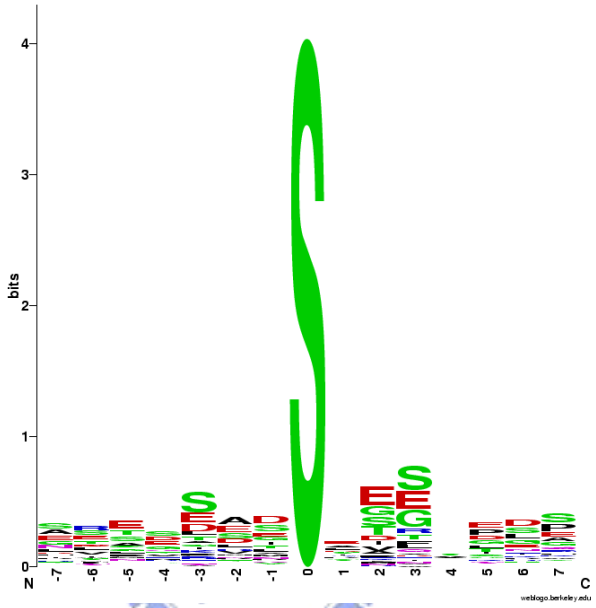

圖 A.55: CK1 (S)資料的序列圖案

表 A.10: CK1 (S)的 30 次 5-CV 於測試資料的效能比較表

CK1 (S)		Threshold	Sensitivity	Specificity	Precision	Accuracy
δ_1	HMMer	-4.6300 (0.4909)	0.2974 (0.0689)	0.9194 (0.0306)	0.8044 (0.0861)	0.6074 (0.0301)
	HMM-1	-0.3159 (0.1211)	0.8410 (0.0372)	0.6570 (0.0293)	0.7156 (0.0212)	0.7487 (0.0234)
	iHMM	-0.3159 (0.1211)	0.8410 (0.0372)	0.6570 (0.0293)	0.7156 (0.0212)	0.7487 (0.0234)
δ_2	HMMer	-10.7763 (1.1806)	0.8042 (0.0600)	0.7245 (0.0748)	0.7674 (0.0459)	0.7639 (0.0268)
	HMM-1	0.6243 (0.4793)	0.8221 (0.0632)	0.8019 (0.0558)	0.8304 (0.0377)	0.8131 (0.0151)
	iHMM	0.6243 (0.4793)	0.8221 (0.0632)	0.8019 (0.0558)	0.8304 (0.0377)	0.8131 (0.0151)

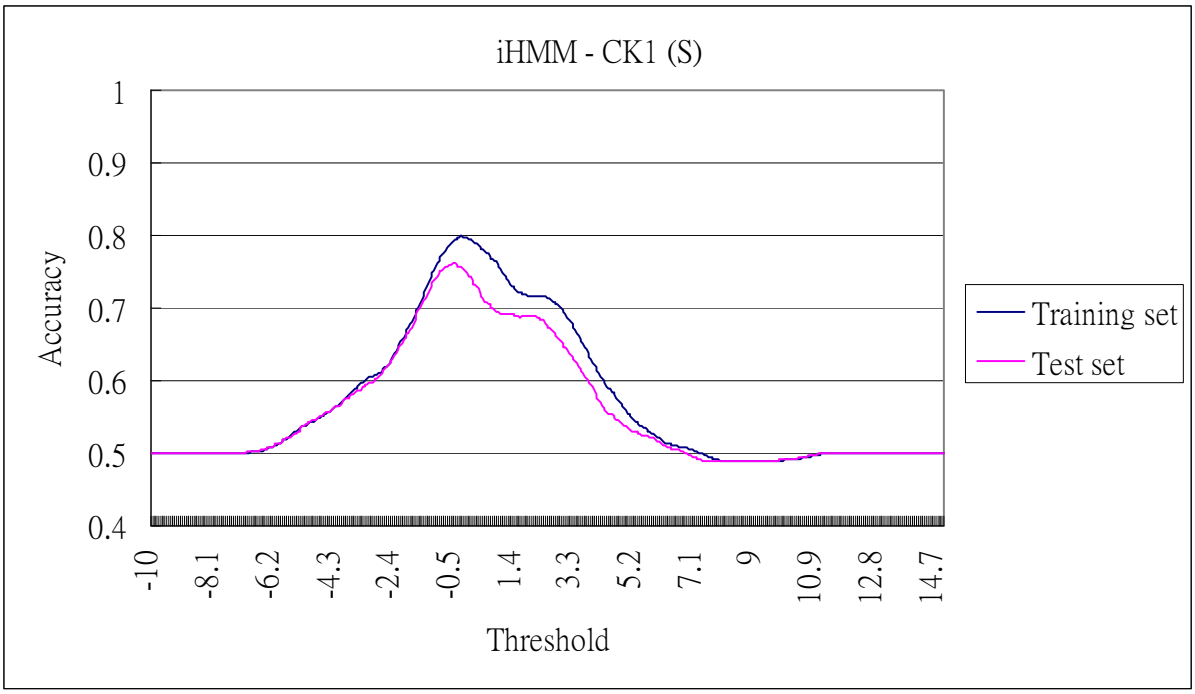

圖 A.56: CK1 (S)資料於 iHMM 的門檻值與正確率對應圖

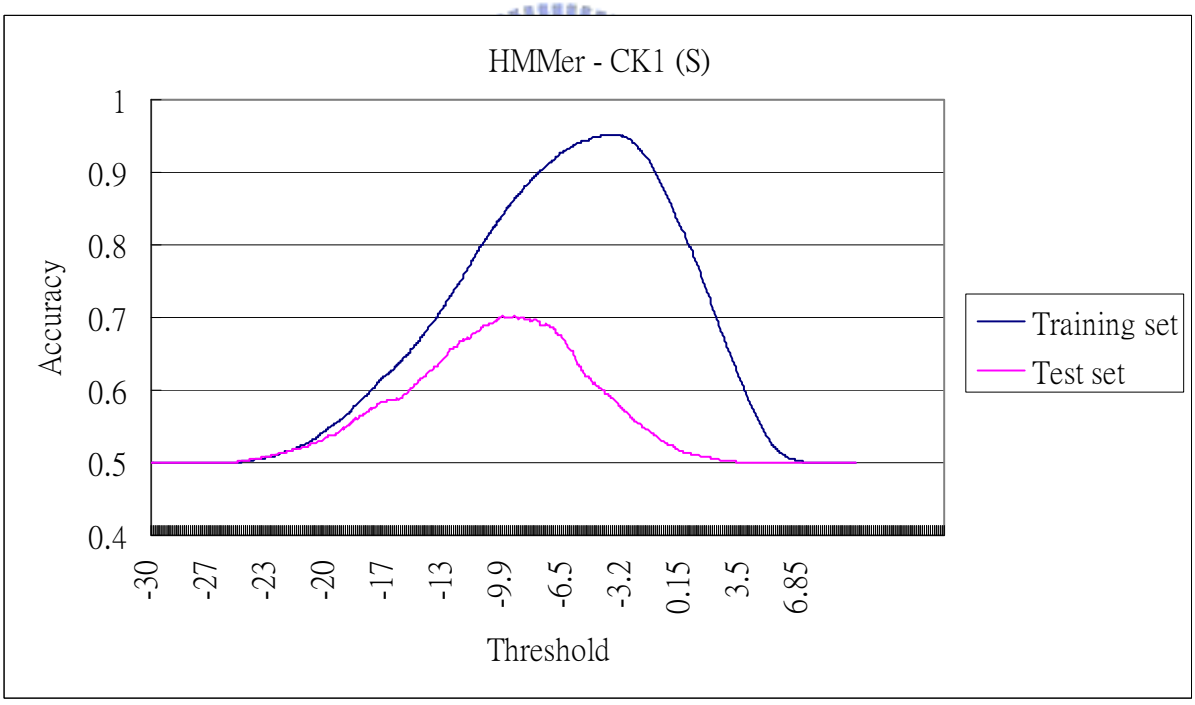

圖 A.57: CK1 (S)資料於 HMMer 的門檻值與正確率對應圖

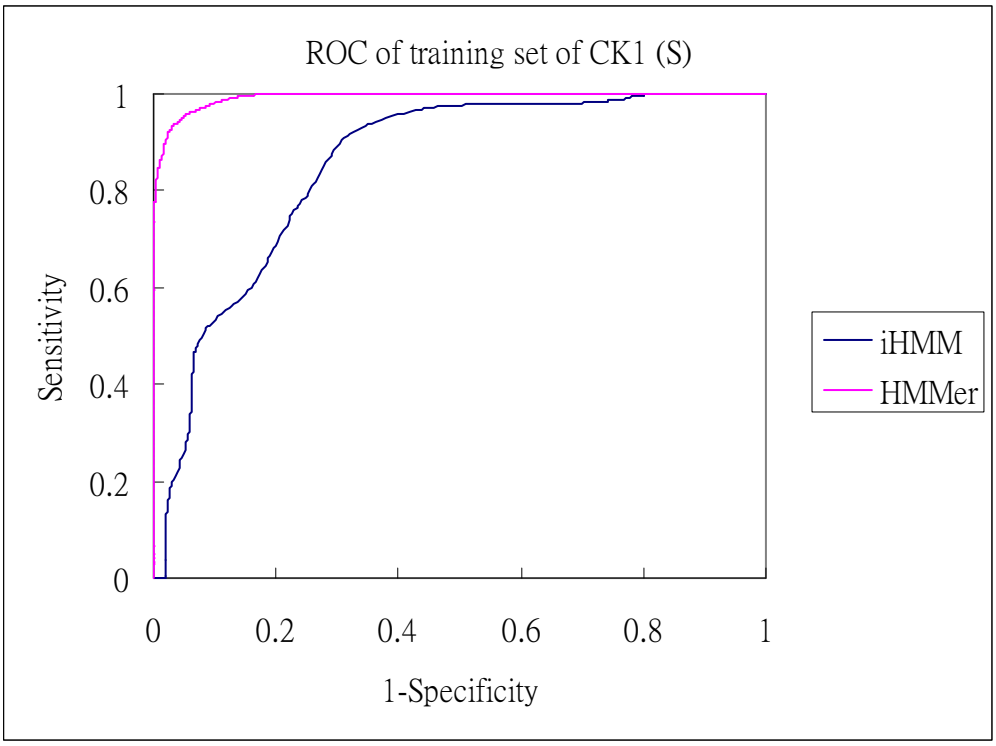

圖 A.58: CK1 (S)於訓練資料的 ROC 圖

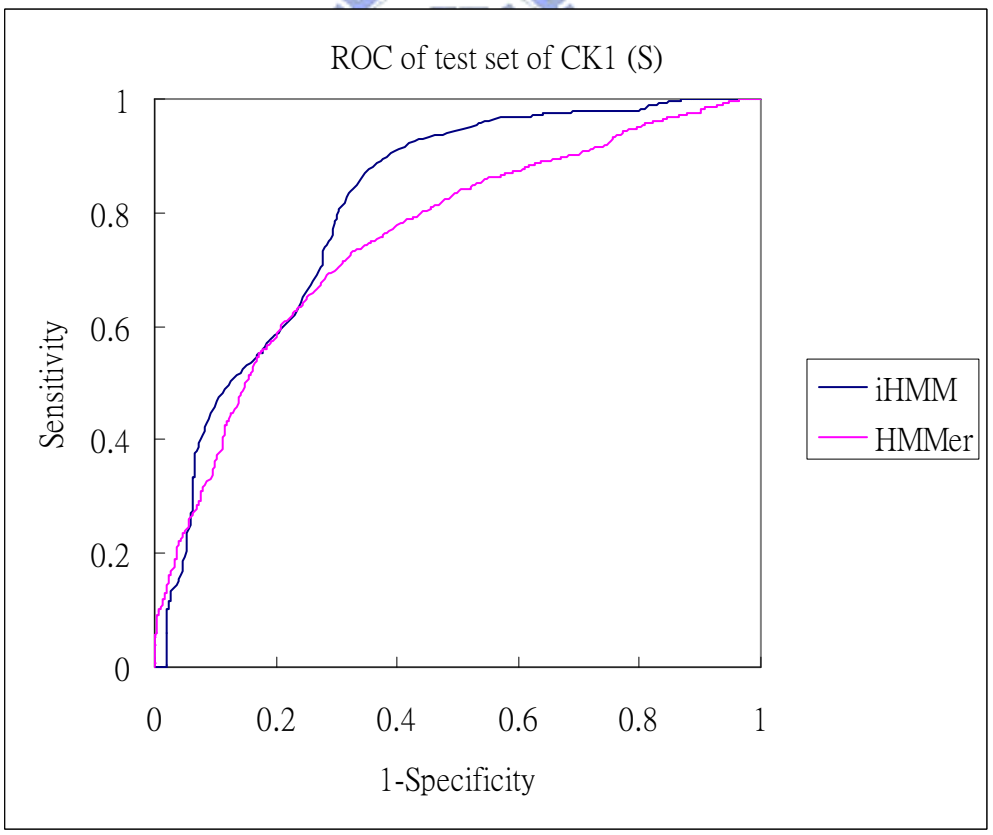

圖 A.59: CK1 (S)於測試資料的 ROC 圖

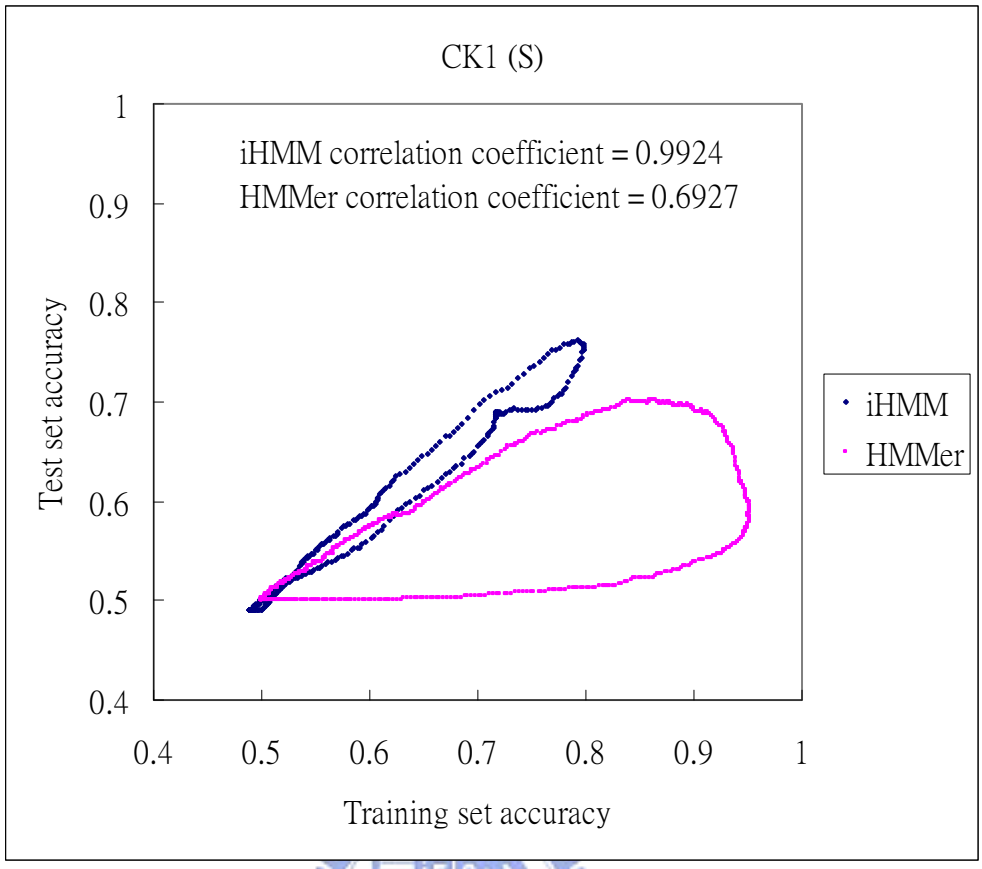

圖 A.60: CK1 (S)資料的相關係數分析圖

CK2 (S)

全名：Casein kinase II

資料筆數：positive 跟 negative 各 243 筆資料

序列長度：15

磷酸化位置：中間的絲氨酸(S)

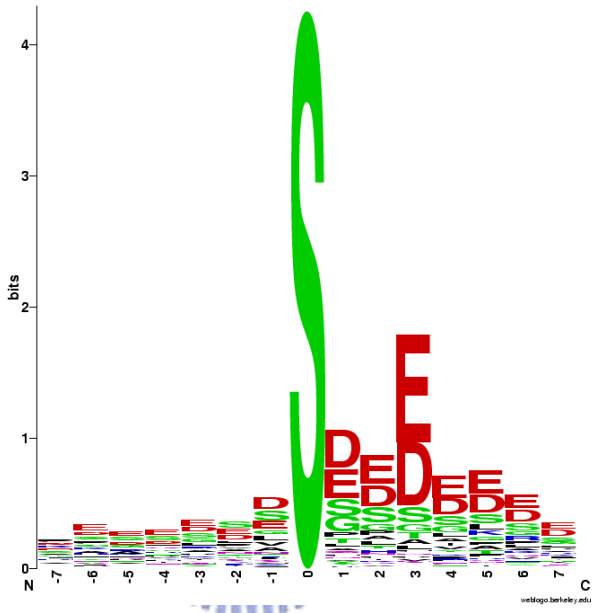

圖 A.61: CK2 (S)資料的序列圖案

表 A.11: CK2 (S)的 30 次 5-CV 於測試資料的效能比較表

CK2 (S)		Threshold	Sensitivity	Specificity	Precision	Accuracy
δ_1	HMMer	-5.3043 (0.2643)	0.7885 (0.0216)	0.8748 (0.0151)	0.8651 (0.0129)	0.8316 (0.0095)
	HMM-1	1.6534 (0.1154)	0.7926 (0.0121)	0.8668 (0.0128)	0.8591 (0.0107)	0.8297 (0.0059)
	iHMM	2.5544 (0.1951)	0.8135 (0.0230)	0.8535 (0.0152)	0.8494 (0.0116)	0.8334 (0.0097)
δ_2	HMMer	-6.2853 (0.5279)	0.8544 (0.0201)	0.8587 (0.0238)	0.8625 (0.0194)	0.8566 (0.0063)
	HMM-1	1.9828 (0.2709)	0.7998 (0.0214)	0.9056 (0.0176)	0.8995 (0.0144)	0.8527 (0.0051)
	iHMM	3.1994 (0.3847)	0.8043 (0.0237)	0.9091 (0.0178)	0.9045 (0.0157)	0.8568 (0.0059)

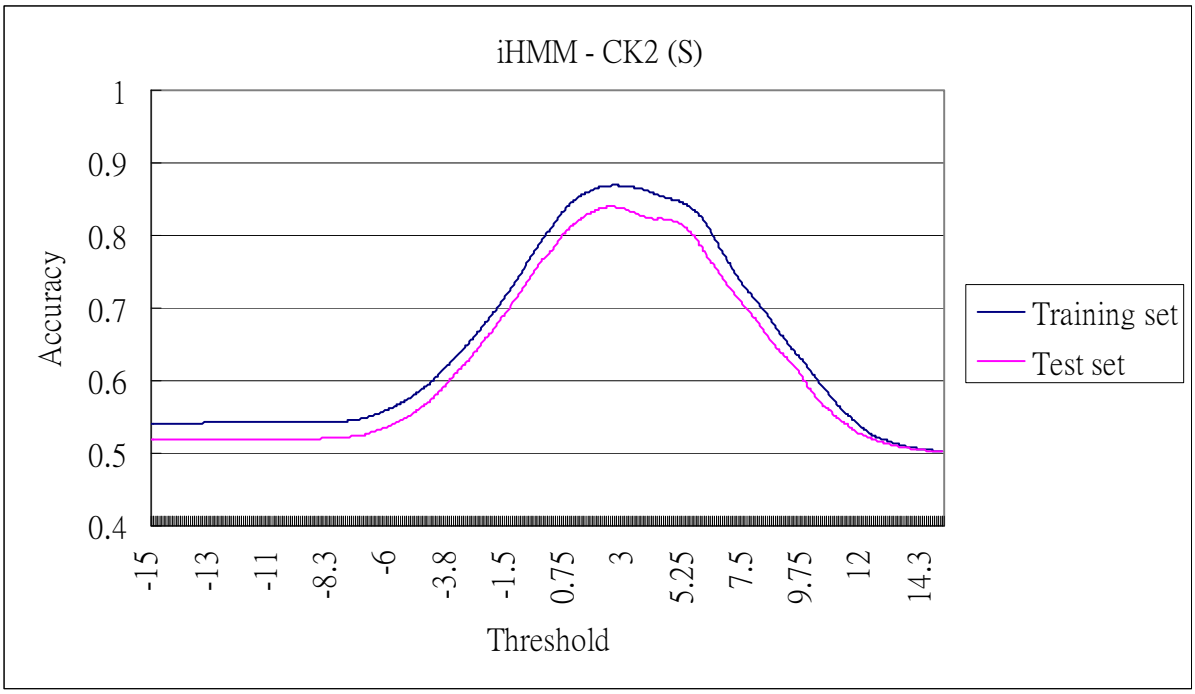

圖 A.62: CK2 (S)資料於 iHMM 的門檻值與正確率對應圖

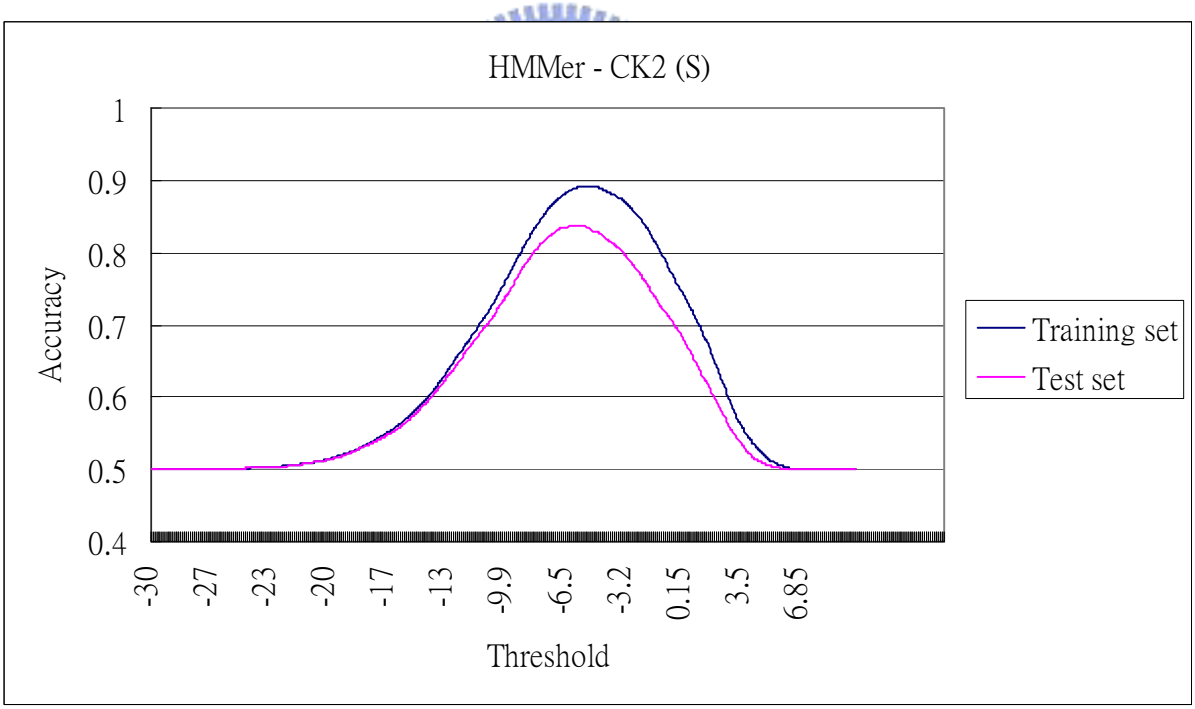

圖 A.63: CK2 (S)資料於 HMMer 的門檻值與正確率對應圖

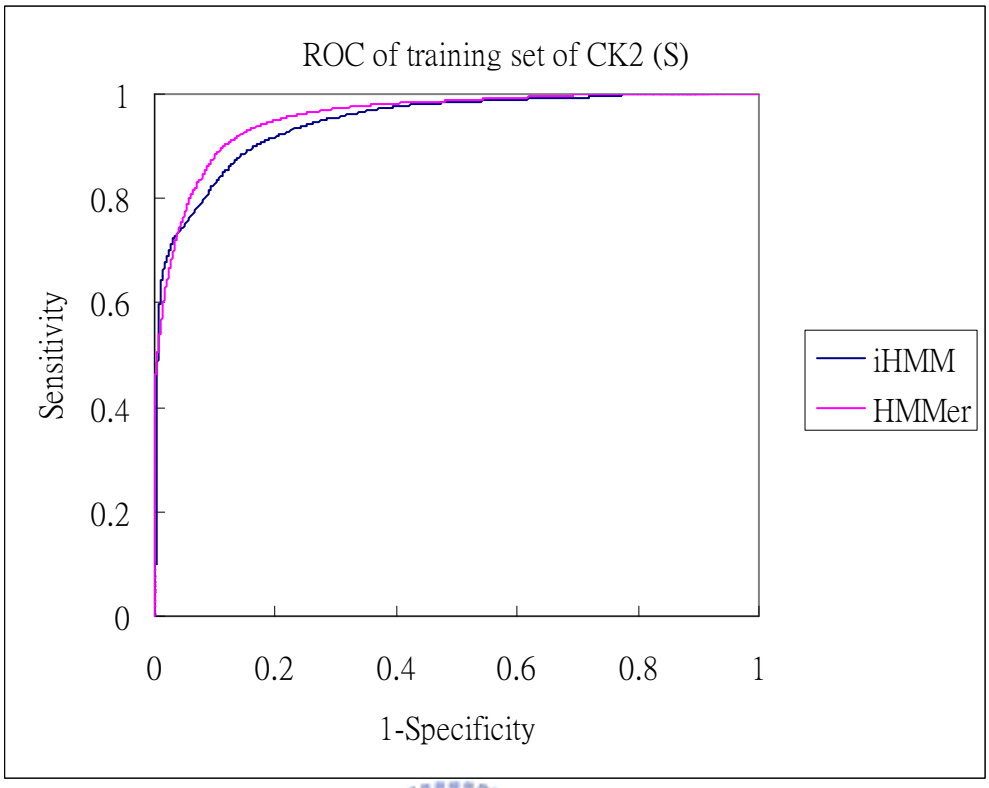

圖 A.64: CK2 (S)於訓練資料的 ROC 圖

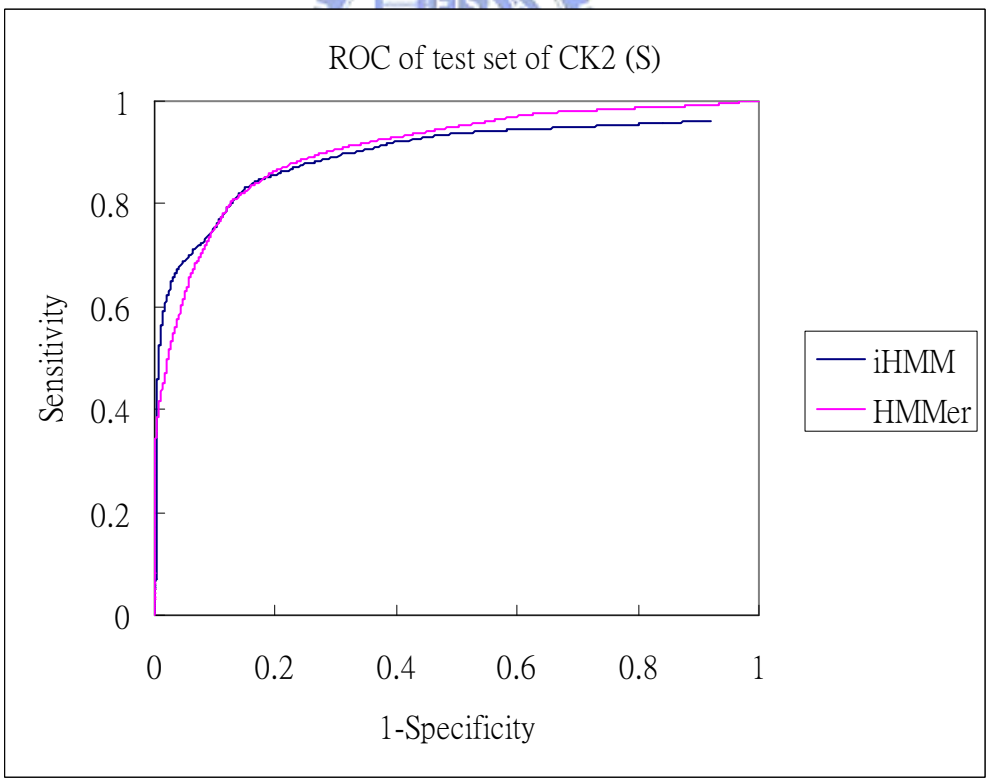

圖 A.65: CK2 (S)於測試資料的 ROC 圖

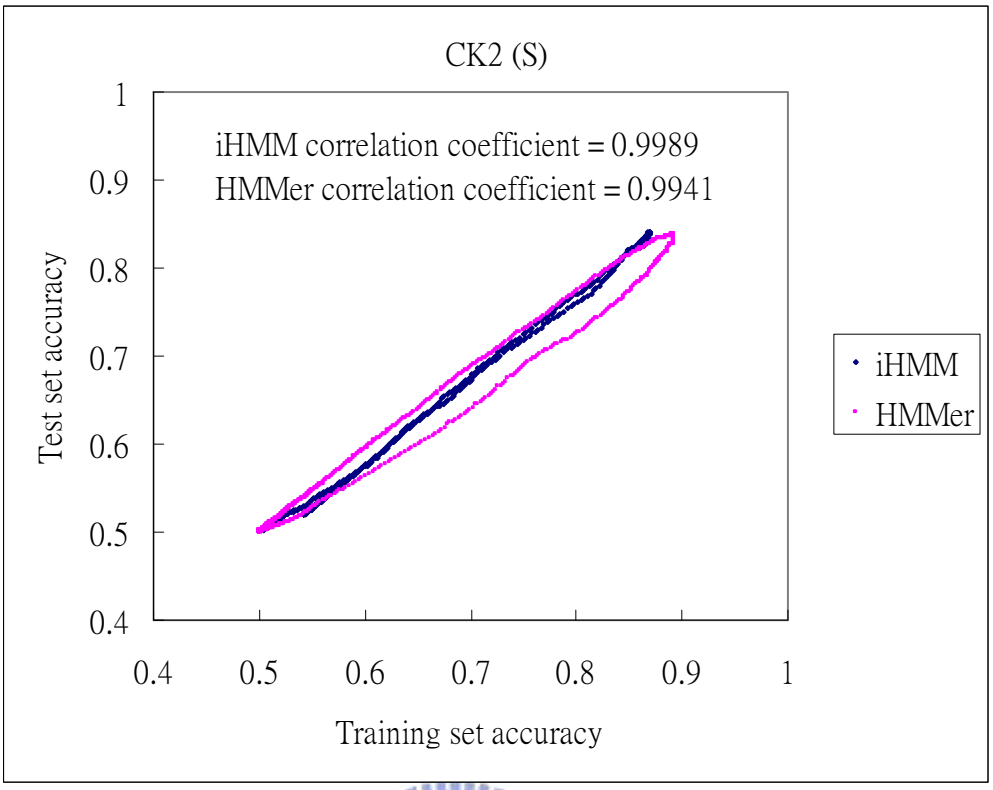

圖 A.66: CK2 (S)資料的相關係數分析圖

CK2 (T)：positive 跟 negative 各 42 筆資料

全名：Casein kinase II

資料筆數：positive 跟 negative 各 42 筆資料

序列長度：15

磷酸化位置：中間的蘇氨酸(T)

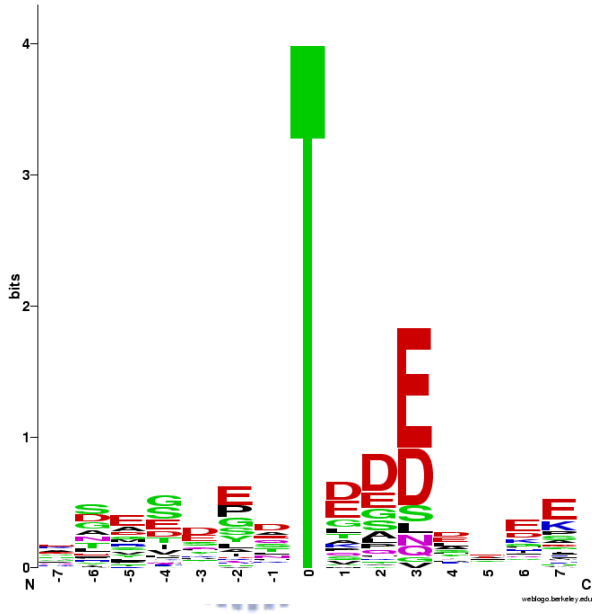

圖 A.67: CK2 (S)資料的序列圖案

表 A.12: CK2 (T) 的 30 次 5-CV 於測試資料的效能比較表

CK2 (T)		Threshold	Sensitivity	Specificity	Precision	Accuracy
δ_1	HMMer	-3.4897 (0.5563)	0.3424 (0.0432)	0.9610 (0.0273)	0.8577 (0.0868)	0.6510 (0.0211)
	HMM-1	0.3777 (0.2551)	0.7764 (0.0417)	0.8084 (0.0350)	0.8227 (0.0291)	0.7904 (0.0234)
	iHMM	0.3777 (0.2551)	0.7764 (0.0417)	0.8084 (0.0350)	0.8227 (0.0291)	0.7904 (0.0234)
δ_2	HMMer	-11.4433 (1.1646)	0.8753 (0.0530)	0.8072 (0.0653)	0.8418 (0.0479)	0.8434 (0.0191)
	HMM-1	0.9573 (0.4221)	0.8325 (0.0429)	0.9192 (0.0334)	0.9222 (0.0297)	0.8774 (0.0133)
	iHMM	0.9573 (0.4221)	0.8325 (0.0429)	0.9192 (0.0334)	0.9222 (0.0297)	0.8774 (0.0133)

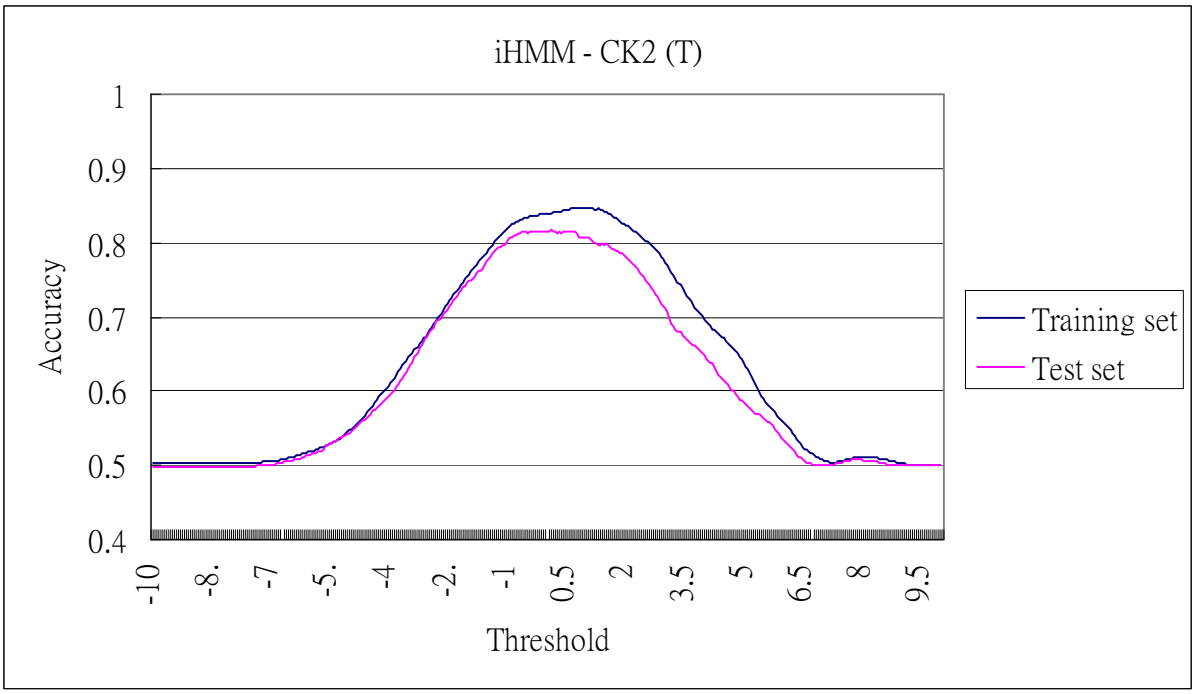

圖 A.68: CK2 (T)資料於 iHMM 的門檻值與正確率對應圖

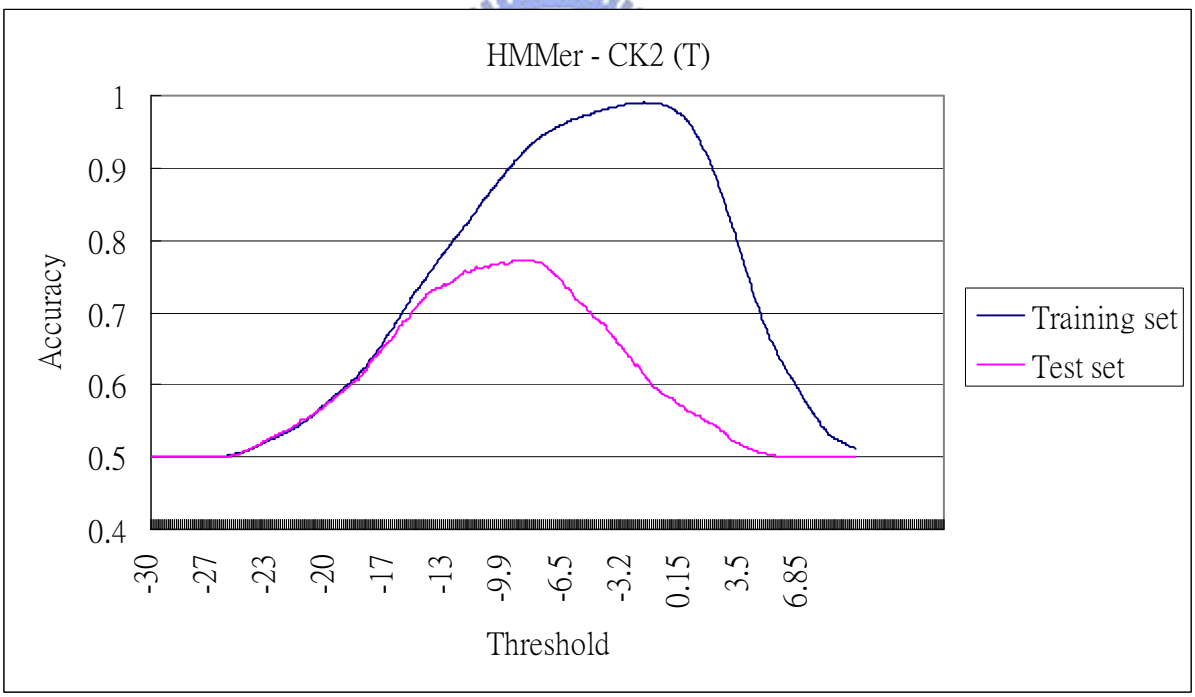

圖 A.69: CK2 (T)資料於 HMMer 的門檻值與正確率對應圖

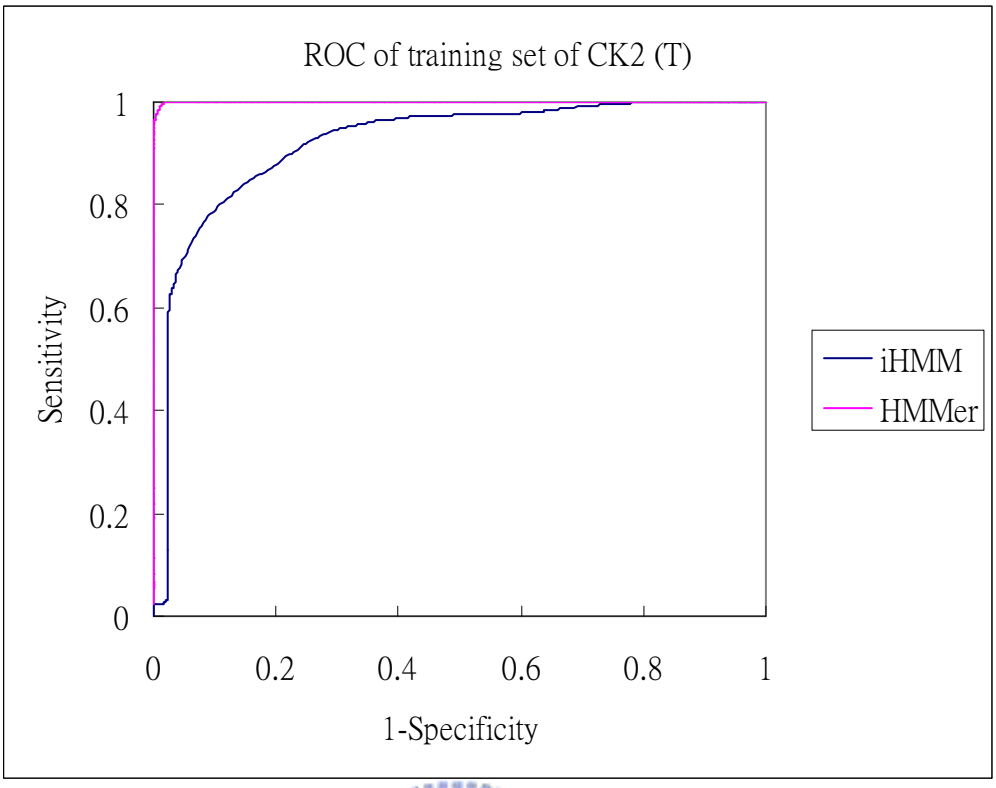

圖 A.70: CK2 (T)於訓練資料的 ROC 圖

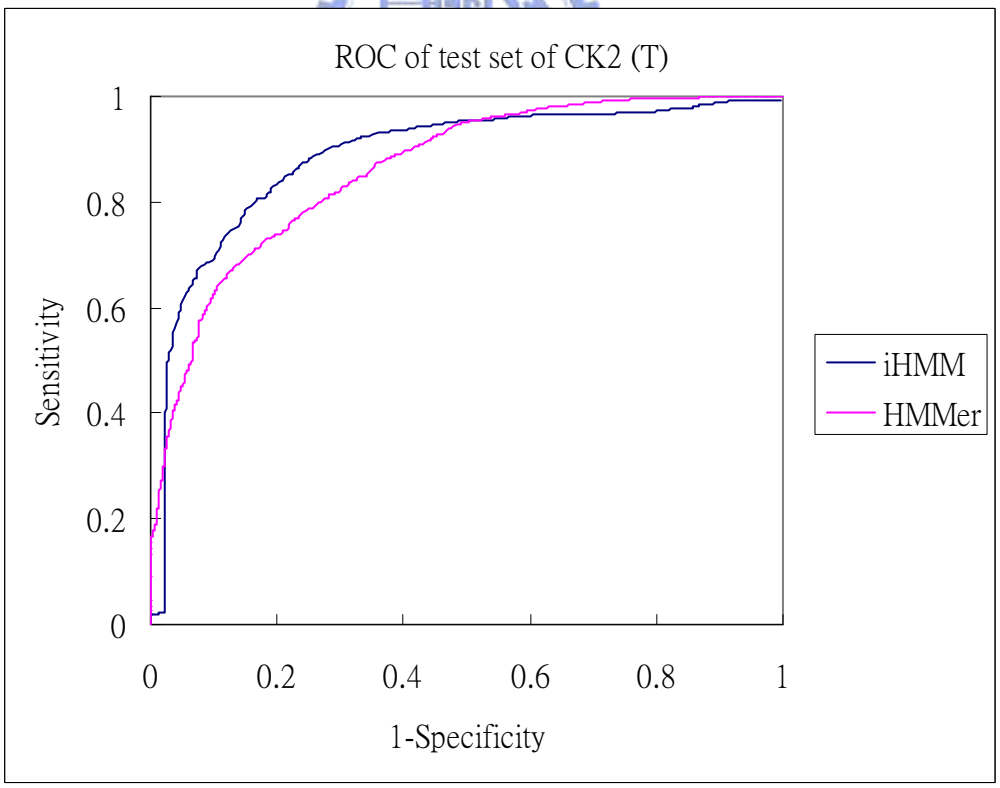

圖 A.71: CK2 (T)於測試資料的 ROC 圖

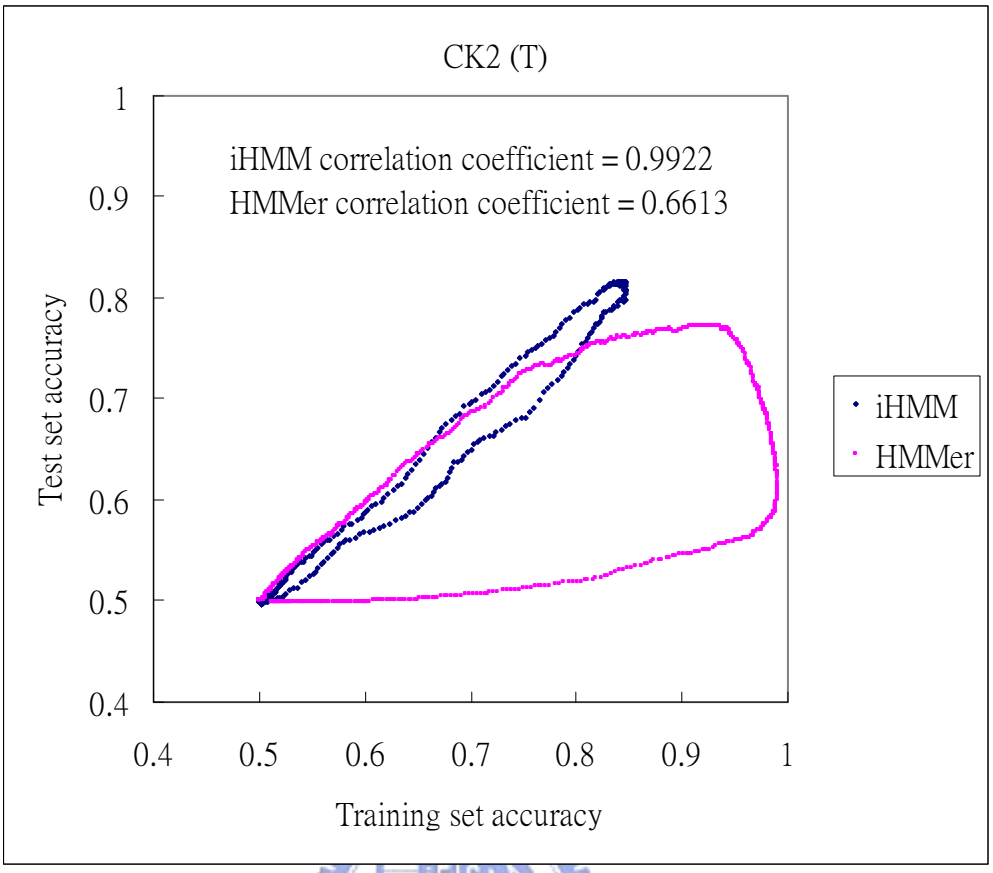

圖 A.72: CK2 (T)資料的相關係數分析圖

MAPK (S)

全名：Mitogen-activated protein kinase

資料筆數：positive 跟 negative 各 207 筆資料

序列長度：15

磷酸化位置：中間的絲氨酸(S)

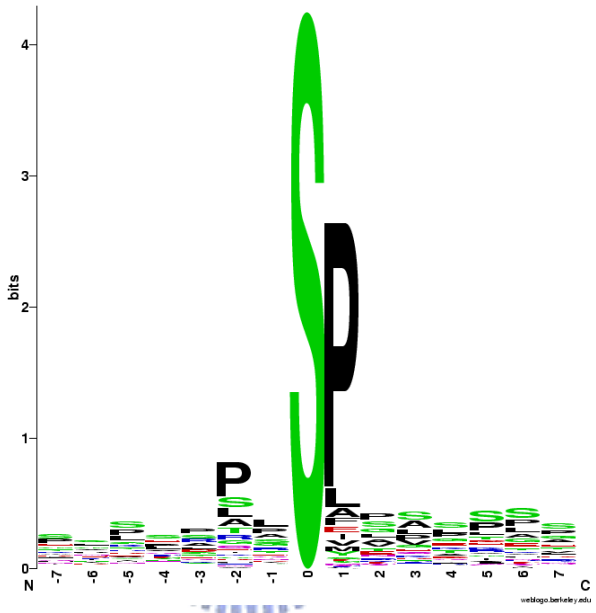

圖 A.73: MAPK (S)資料的序列圖案

表 A.13: MAPK (S)的 30 次 5-CV 於測試資料的效能比較表

MAPK (S)		Threshold	Sensitivity	Specificity	Precision	Accuracy
δ_1	HMMer	-4.8180 (0.1937)	0.7463 (0.0219)	0.8585 (0.0182)	0.8466 (0.0170)	0.8024 (0.0122)
	HMM-1	2.3696 (0.2825)	0.7562 (0.0246)	0.8210 (0.0258)	0.8150 (0.0203)	0.7887 (0.0145)
	iHMM	2.7627 (0.2064)	0.8121 (0.0189)	0.8335 (0.0217)	0.8338 (0.0164)	0.8228 (0.0090)
δ_2	HMMer	-5.7337 (0.3999)	0.8218 (0.0248)	0.8418 (0.0279)	0.8445 (0.0197)	0.8317 (0.0072)
	HMM-1	2.5986 (0.4247)	0.7660 (0.028)	0.8599 (0.0287)	0.8522 (0.0239)	0.8130 (0.0116)
	iHMM	3.2280 (0.3468)	0.8094 ().0228)	0.8871 (0.0212)	0.8831 (0.0164)	0.8483 (0.0077)

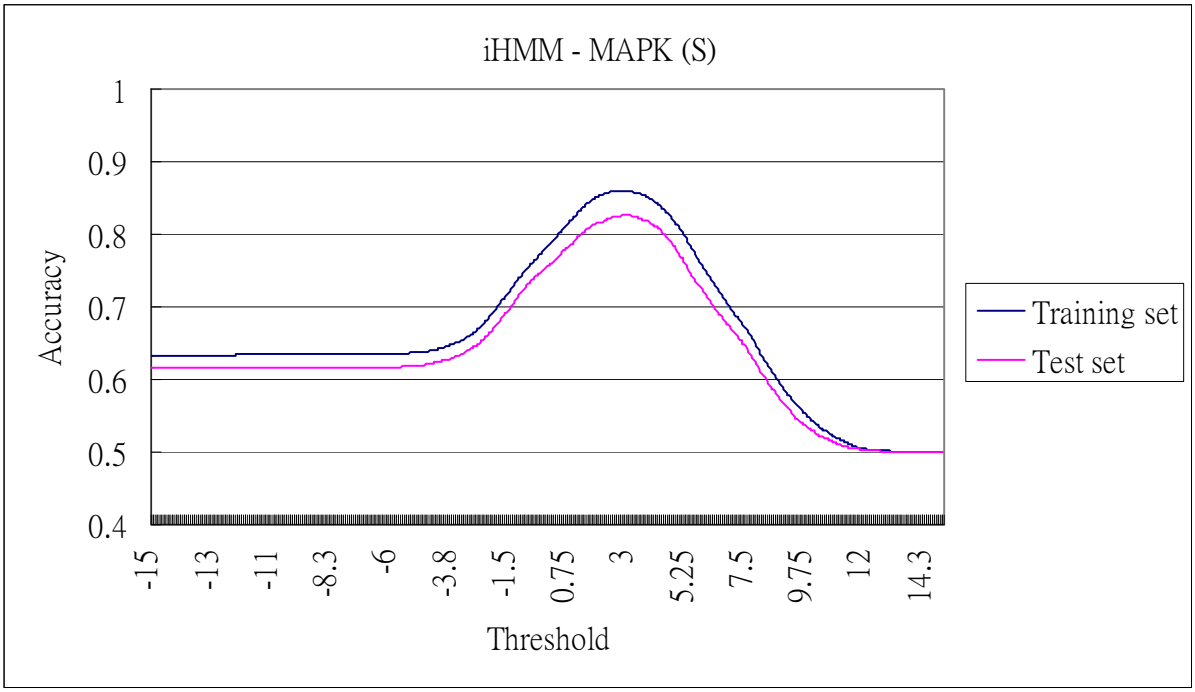

圖 A.74: MAPK (S)資料於 iHMM 的門檻值與正確率對應圖

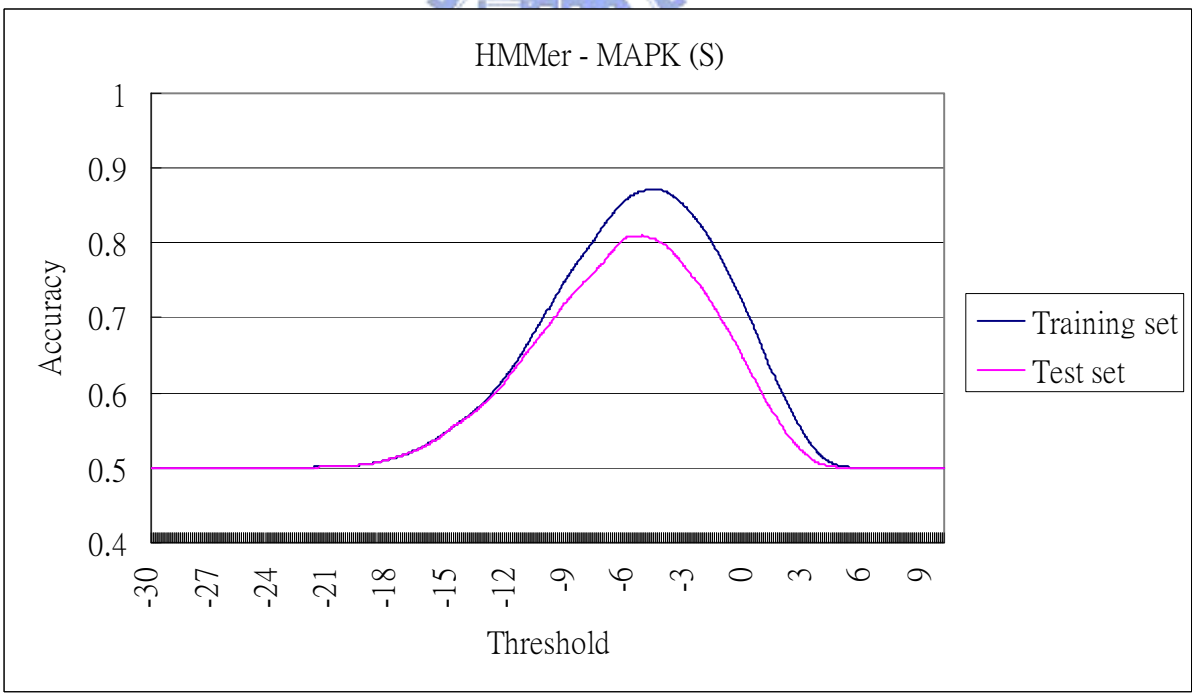

圖 A.75: MAPK (S)資料於 HMMer 的門檻值與正確率對應圖

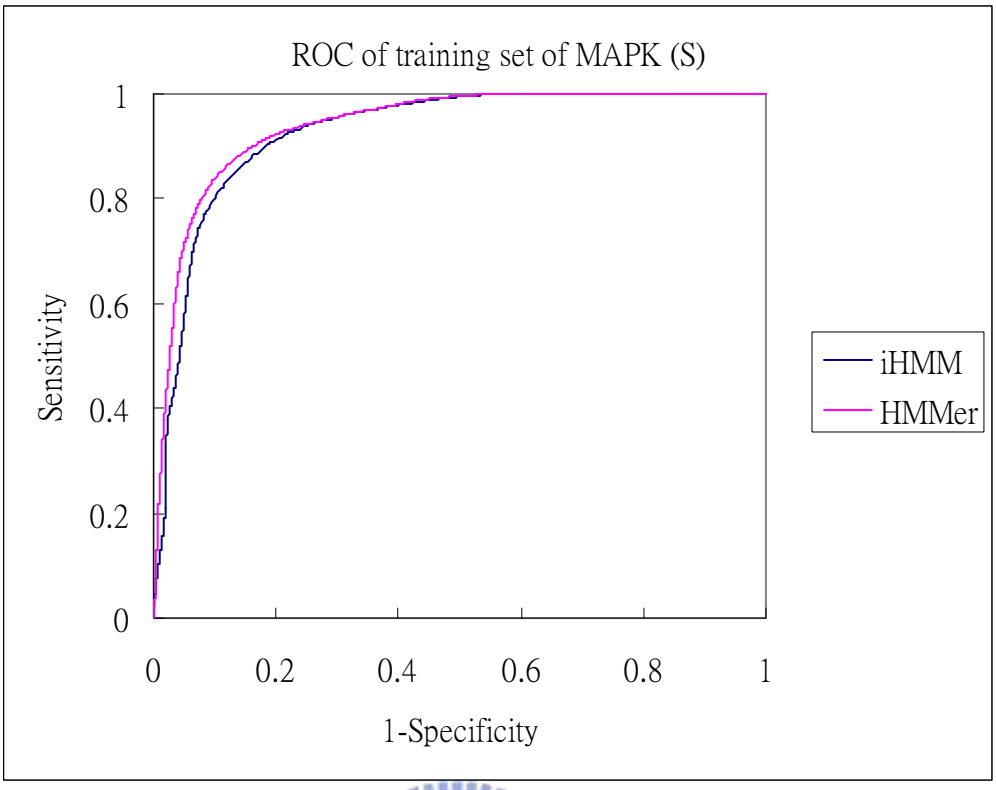

圖 A.76: MAPK (S)於訓練資料的 ROC 圖

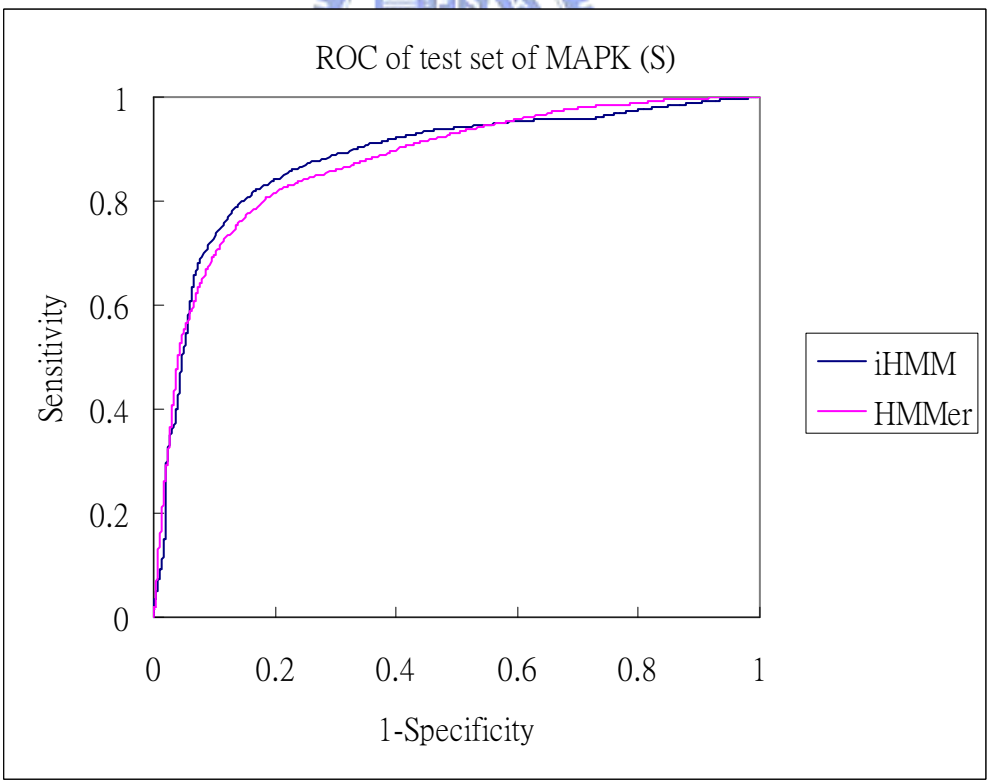

圖 A.77: MAPK (S)於測試資料的 ROC 圖

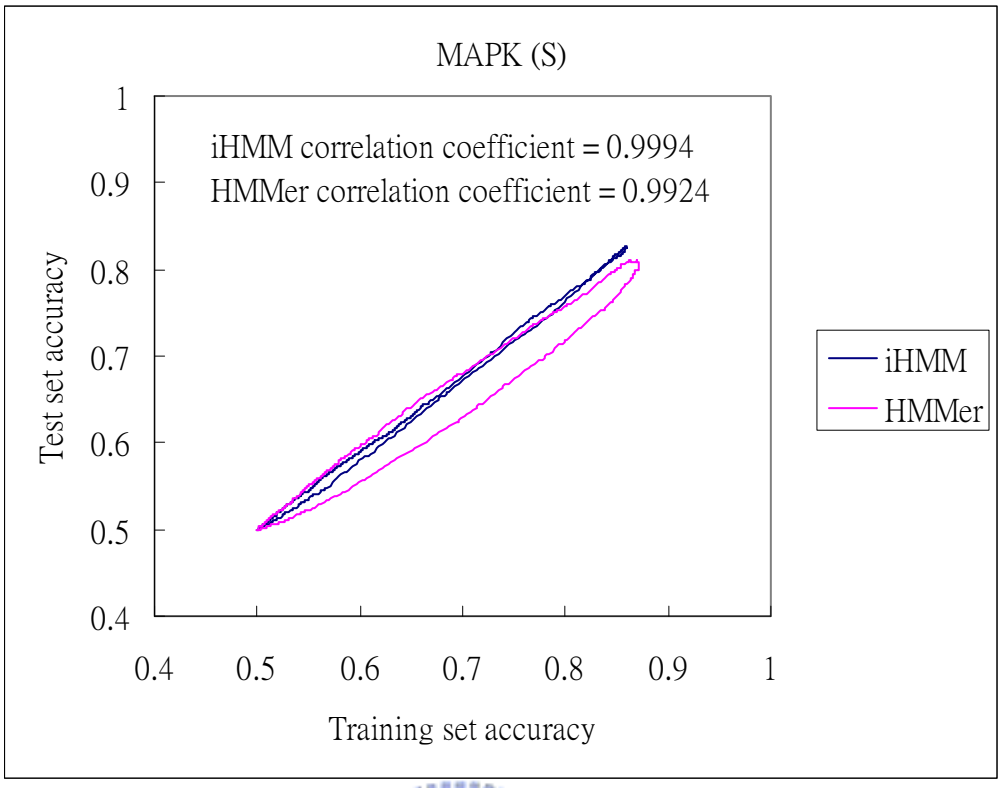

圖 A.78: MAPK (S)資料的相關係數分析圖

MAPK (T)：positive 跟 negative 各 81 筆資料

全名：Mitogen-activated protein kinase

資料筆數：positive 跟 negative 各 81 筆資料

序列長度：15

磷酸化位置：中間的蘇氨酸(T)

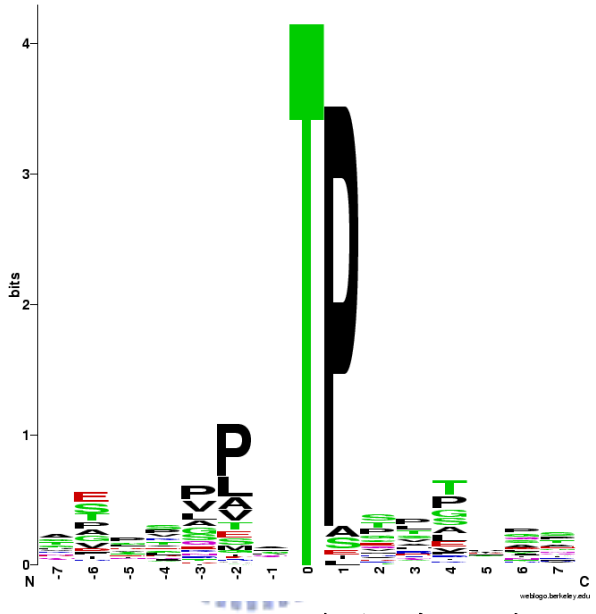

圖 A.79: MAPK (T)資料的序列圖案

表 A.14: MAPK (T)的 30 次 5-CV 於測試資料的效能比較表

MAPK(T)		Threshold	Sensitivity	Specificity	Precision	Accuracy
δ_1	HMMer	-4.0930 (0.3924)	0.7184 (0.0358)	0.9586 (0.0207)	0.9515 (0.0203)	0.8386 (0.0164)
	HMM-1	1.0038 (0.1177)	0.8644 (0.0283)	0.7957 (0.0241)	0.8142 (0.0135)	0.8290 (0.0105)
	iHMM	2.9692 (0.3611)	0.7886 (0.0361)	0.8895 (0.0300)	0.8850 (0.0277)	0.8389 (0.0213)
δ_2	HMMer	-7.9683 (0.7326)	0.9108 (0.0279)	0.8900 (0.0409)	0.9006 (0.0320)	0.9009 (0.0142)
	HMM-1	1.5768 (0.3129)	0.8689 (0.0296)	0.8721 (0.0261)	0.8804 (0.0178)	0.8718 (0.0087)
	iHMM	-2.4205 (9.3479)	0.8473 (0.0345)	0.9158 (0.0317)	0.9205 (0.0237)	0.8827 (0.0138)

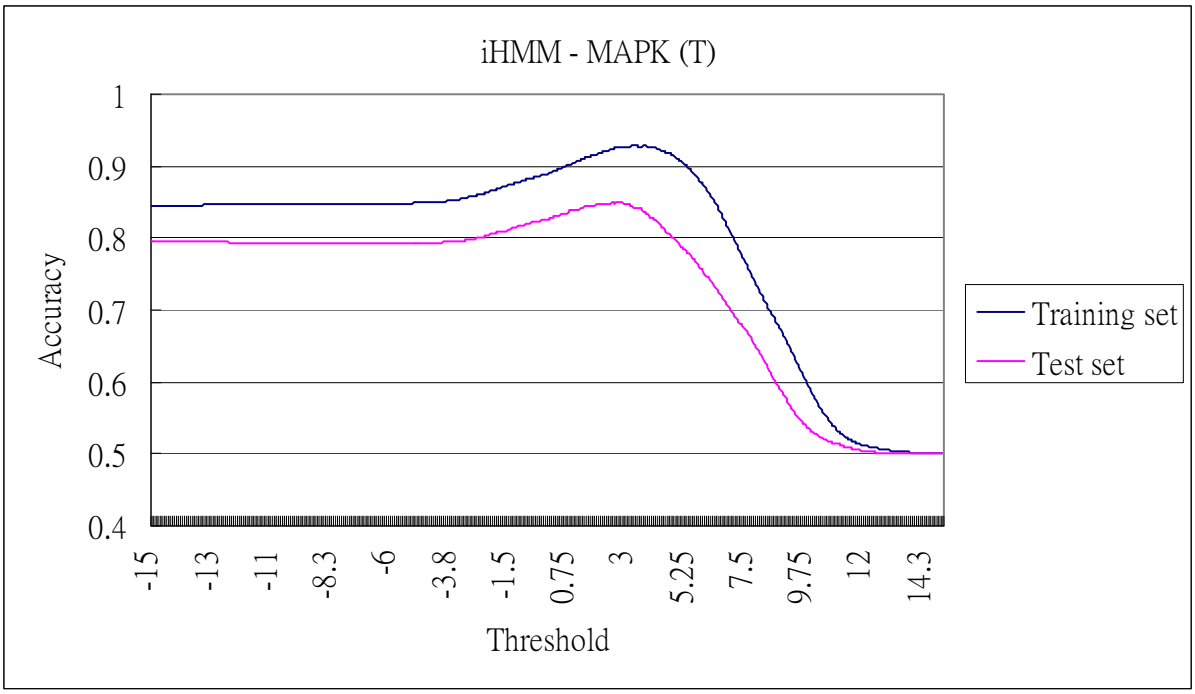

圖 A.80: MAPK (T)資料於 iHMM 的門檻值與正確率對應圖

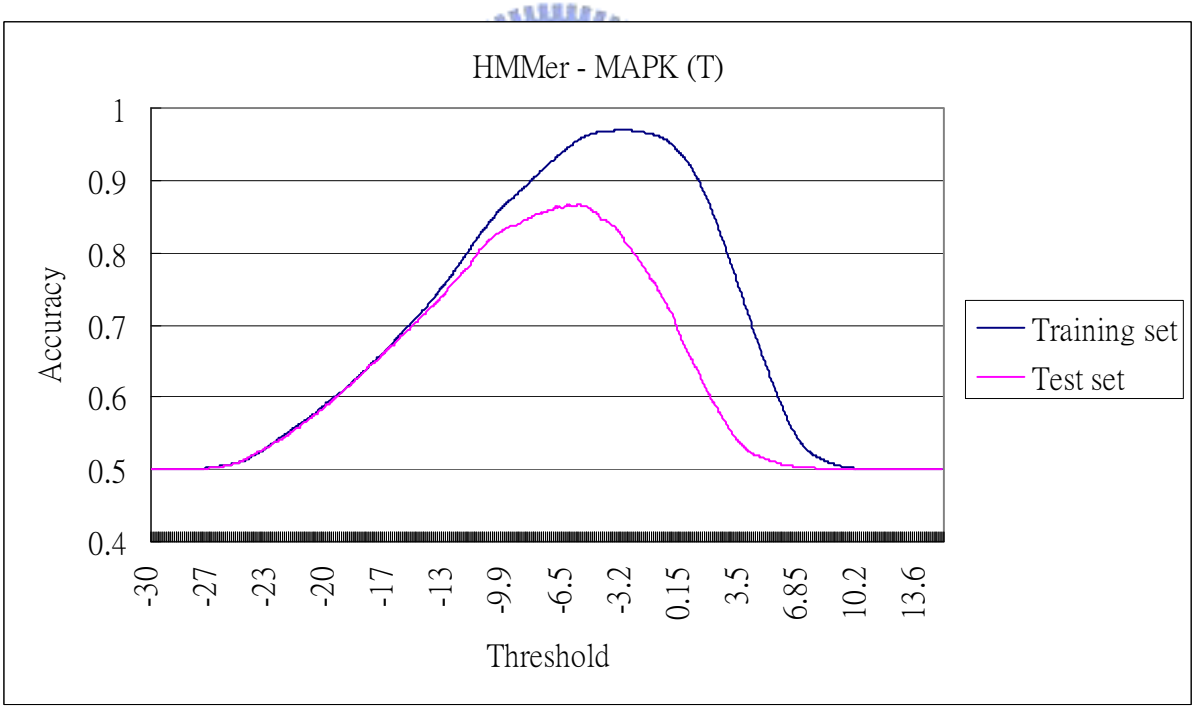

圖 A.81: MAPK (T)資料於 HMMer 的門檻值與正確率對應圖

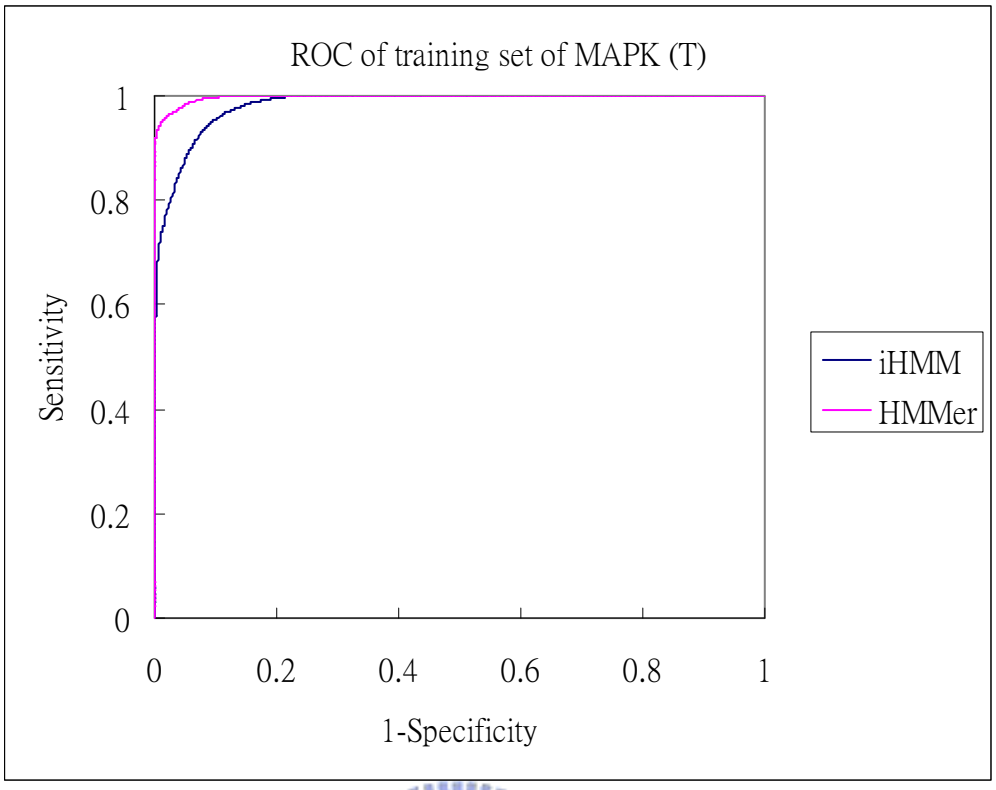

圖 A.82: MAPK (T)於訓練資料的 ROC 圖

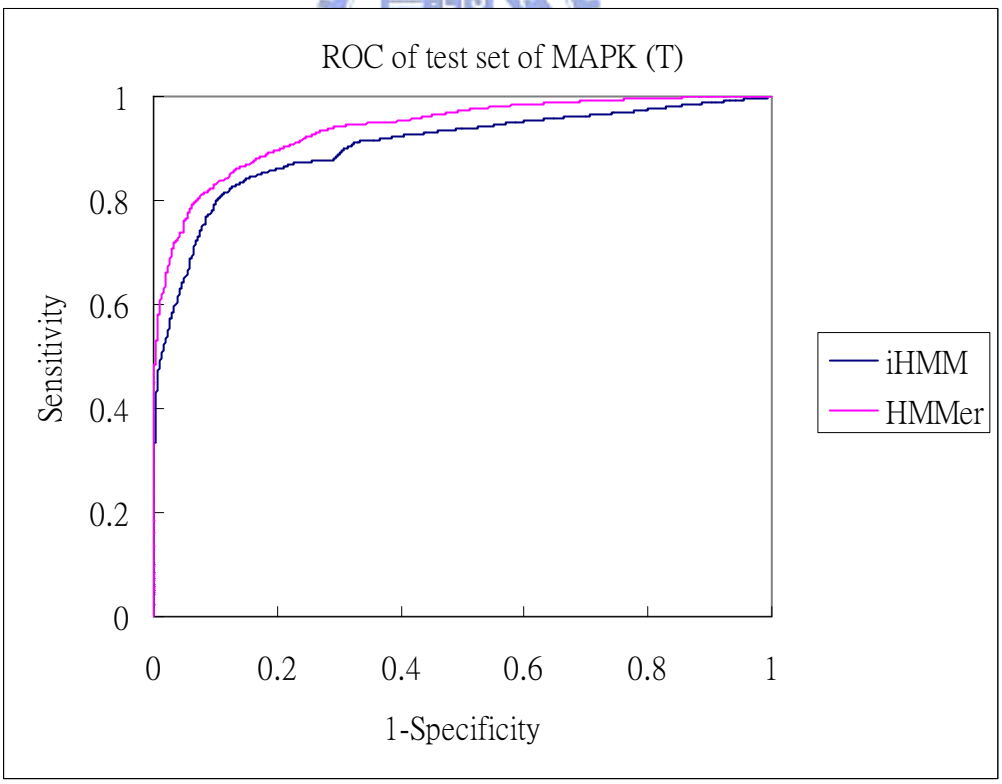

圖 A.83: MAPK (T)於測試資料的 ROC 圖

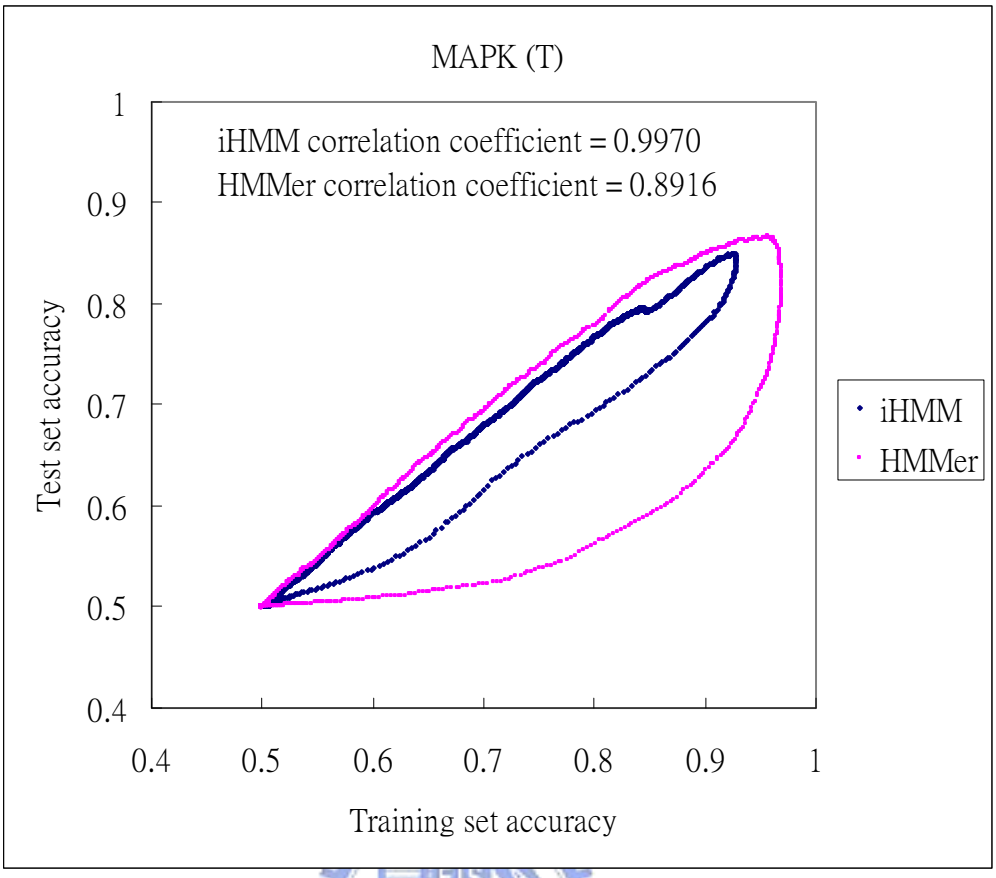

圖 A.84: MAPK (T)資料的相關係數分析圖

ATM (S)

全名：Ataxia telangiectasia mutated

資料筆數：positive 跟 negative 各 77 筆資料

序列長度：15

磷酸化位置：中間的絲氨酸(S)

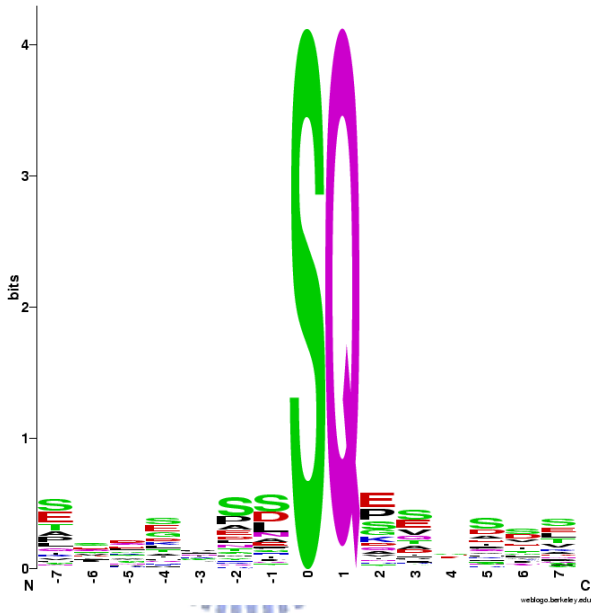

圖 A.85: ATM (S)資料的序列圖案

表 A.15: ATM (S)的 30 次 5-CV 於測試資料的效能比較表

ATM (S)		Threshold	Sensitivity	Specificity	Precision	Accuracy
δ_1	HMMer	-5.4557 (0.8528)	0.7495 (0.0567)	0.9772 (0.0131)	0.9766 (0.0111)	0.8628 (0.0280)
	HMM-1	1.1264 (0.1181)	0.8785 (0.0292)	0.7735 (0.0247)	0.8030 (0.0174)	0.8257 (0.0155)
	iHMM	-32.7786 (17.3377)	0.8934 (0.0411)	0.9534 (0.0178)	0.9546 (0.0158)	0.9234 (0.0223)
δ_2	HMMer	-11.2183 (0.6446)	0.9847 (0.0150)	0.9474 (0.0157)	0.9529 (0.0138)	0.9664 (0.0096)
	HMM-1	1.7204 (0.3527)	0.8812 (0.0419)	0.8651 (0.0341)	0.8770 (0.0238)	0.8733 (0.0095)
	iHMM	-16.8526 (15.5258)	0.9317 (0.0302)	0.9737 (0.0124)	0.9747 (0.0116)	0.9527 (0.0141)

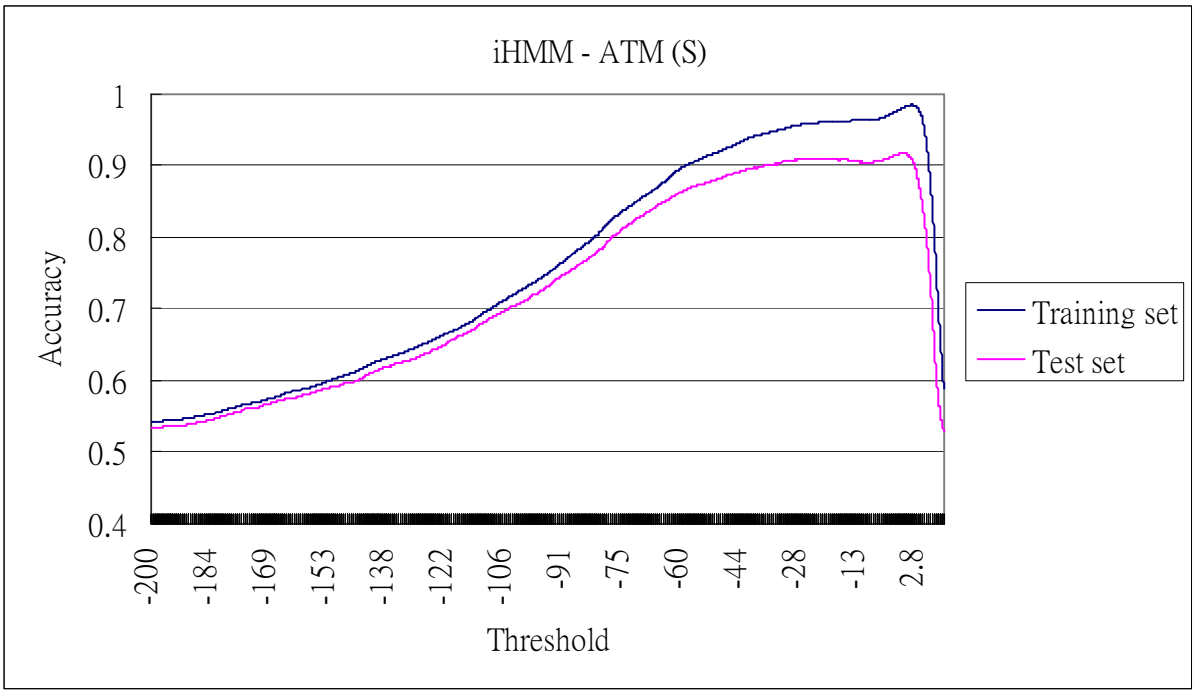

圖 A.86: ATM (S)資料於 iHMM 的門檻值與正確率對應圖

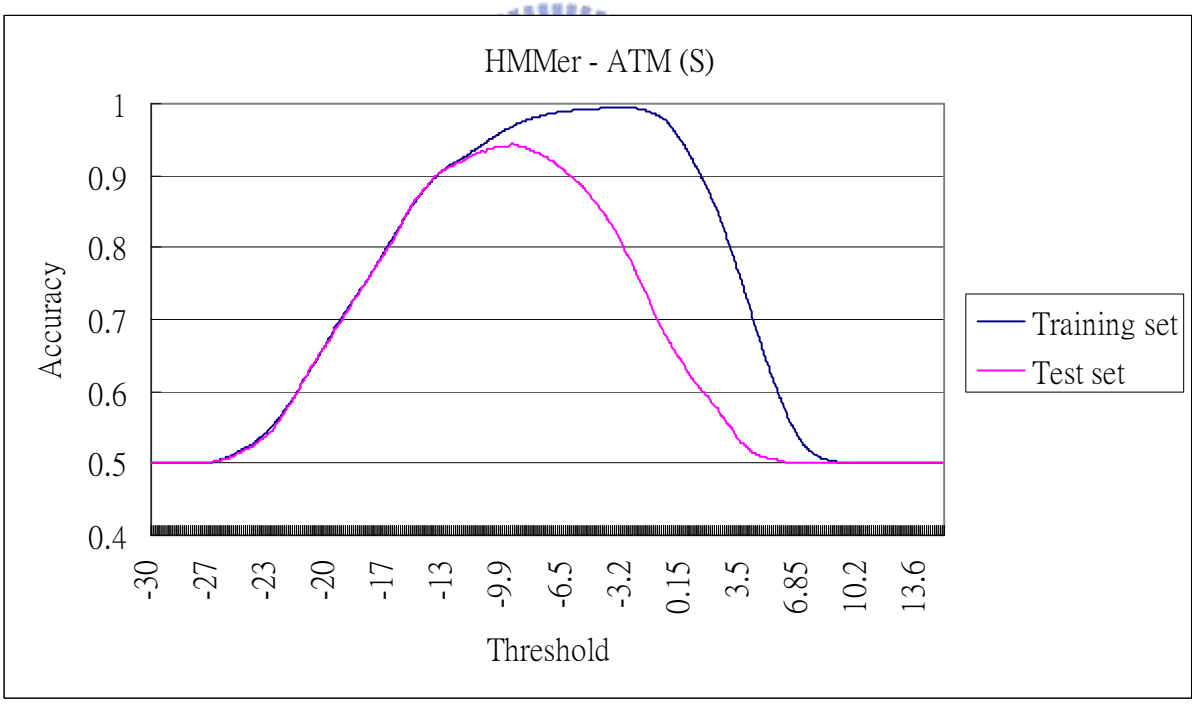

圖 A.87: ATM (S)資料於 HMMer 的門檻值與正確率對應圖

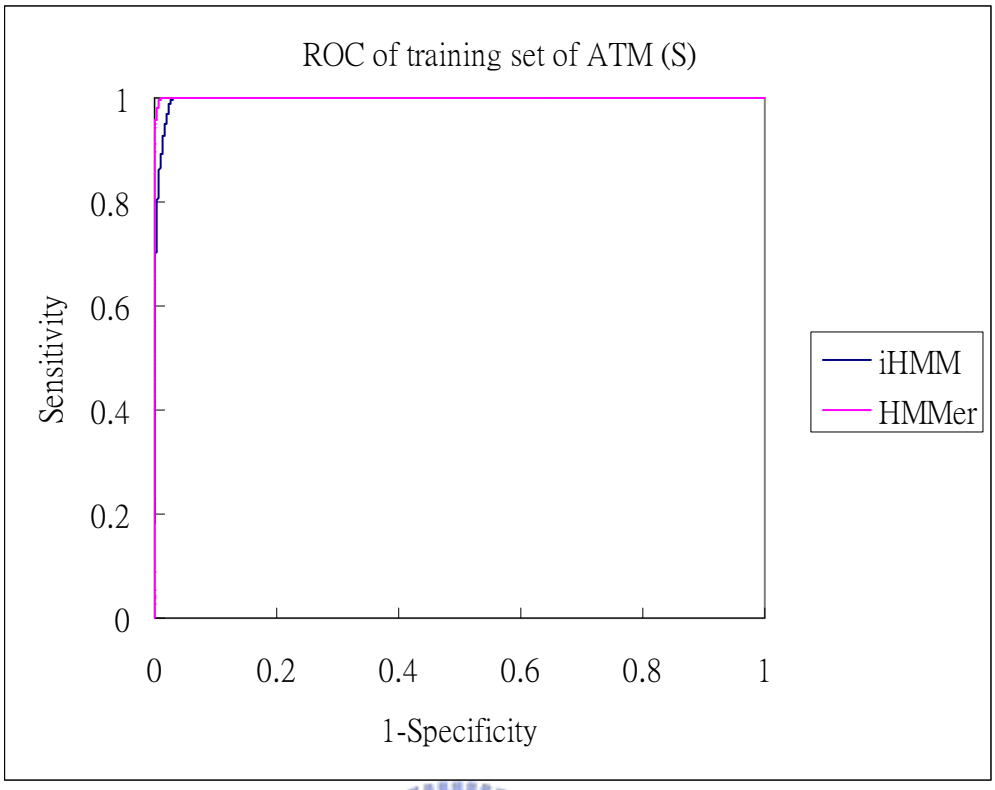

圖 A.88: ATM (S)於訓練資料的 ROC 圖

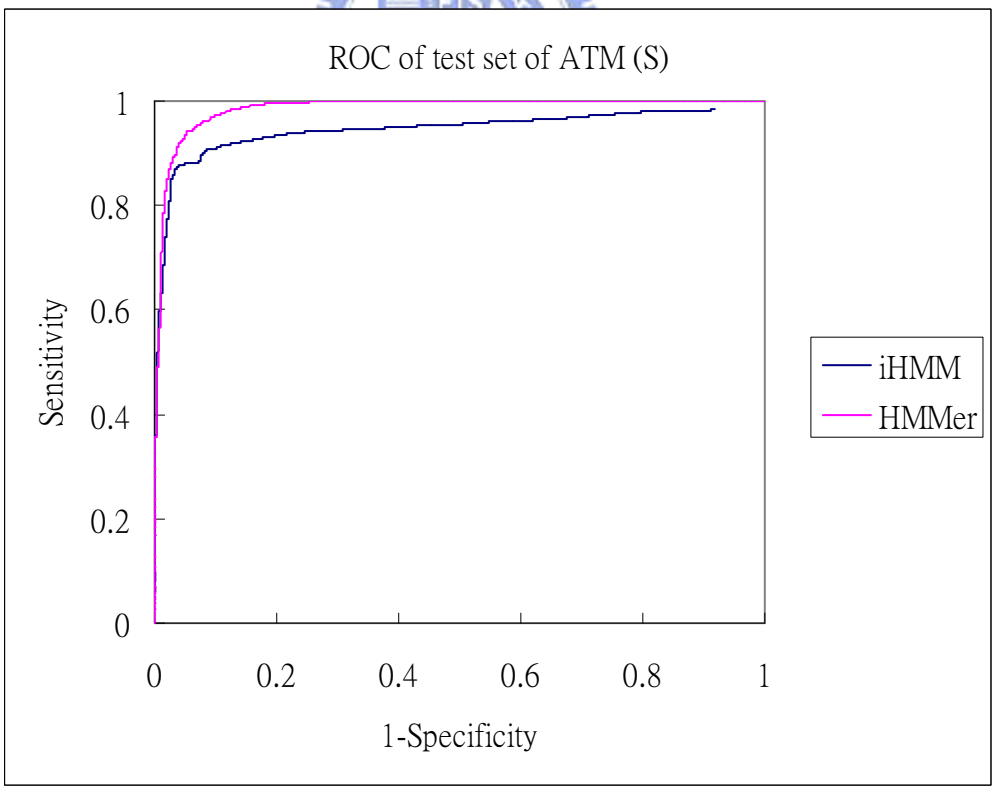

圖 A.89: ATM (S)於測試資料的 ROC 圖

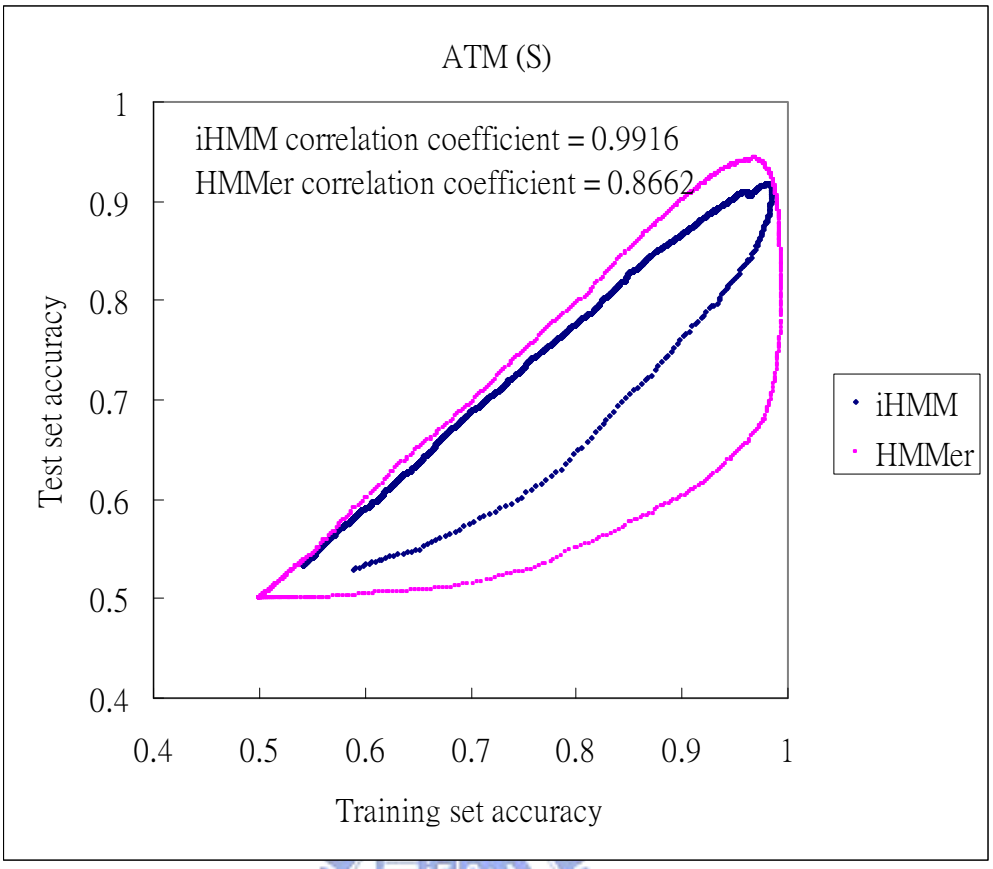

圖 A.90: ATM (S)資料的相關係數分析圖

EGFR (Y)

全名：Epidermal growth factor receptor

資料筆數：positive 跟 negative 各 46 筆資料

序列長度：15

磷酸化位置：中間的酪氨酸(Y)

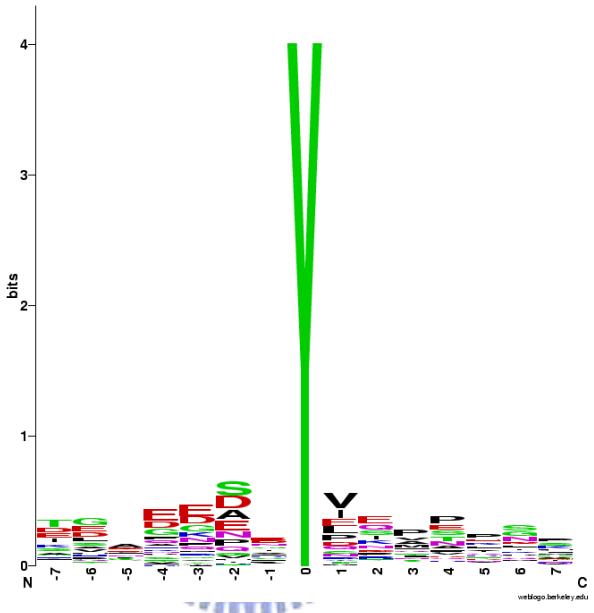

圖 A.91: EGFR (Y)資料的序列圖案

EGFR

表 A.16: EGFR (Y)的 30 次 5-CV 於測試資料的效能比較表

EGFR (Y)		Threshold	Sensitivity	Specificity	Precision	Accuracy
δ_1	HMMer	-3.5857 (0.5205)	0.3484 (0.0537)	0.8972 (0.0311)	0.7564 (0.0805)	0.6223 (0.0247)
	HMM-1	0.0004 (0.2334)	0.7294 (0.0515)	0.5845 (0.0492)	0.6437 (0.0233)	0.6549 (0.0178)
	iHMM	2.1990 (0.3385)	0.6399 (0.0609)	0.7726 (0.0574)	0.7536 (0.0622)	0.7059 (0.0446)
δ_2	HMMer	-12.1897 (1.3548)	0.8789 (0.0553)	0.6870 (0.0621)	0.7581 (0.0323)	0.7866 (0.0143)
	HMM-1	0.4404 (0.4183)	0.8617 (0.0659)	0.6555 (0.0730)	0.7353 (0.0496)	0.7615 (0.0194)
	iHMM	-10.0333 (34.4861)	0.7339 (0.0662)	0.8362 (0.0609)	0.8369 (0.0480)	0.7867 (0.0254)

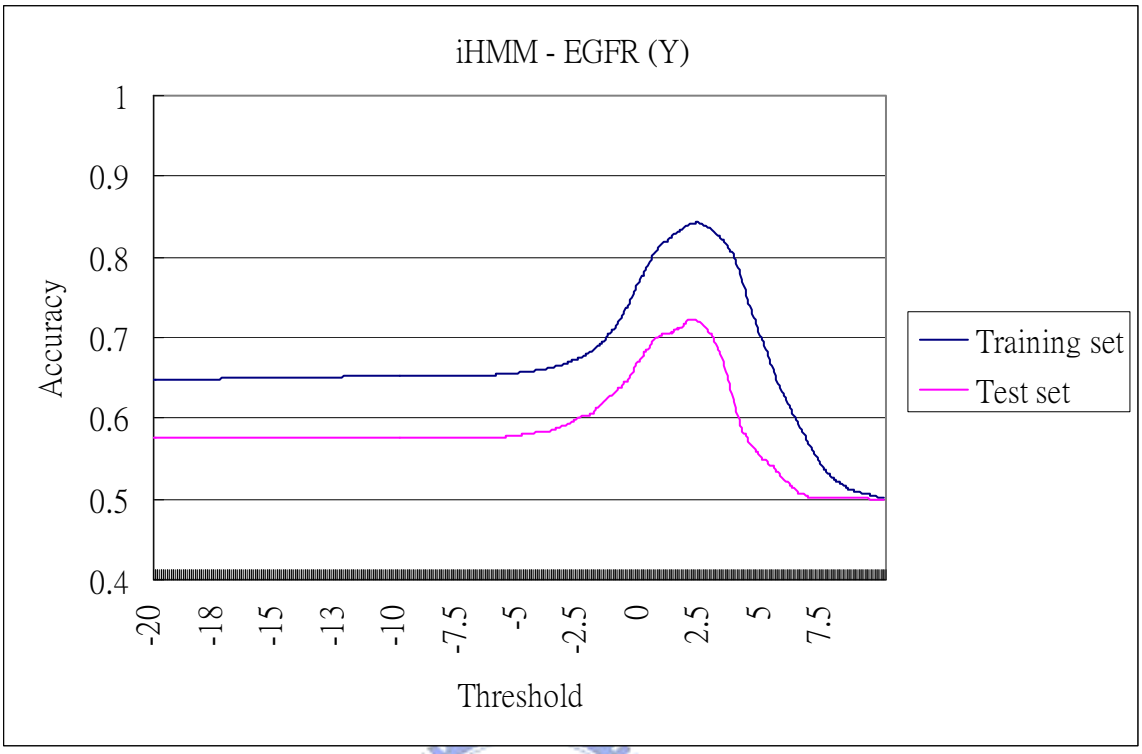

圖 A.92: EGFR (Y)資料於 iHMM 的門檻值與正確率對應圖

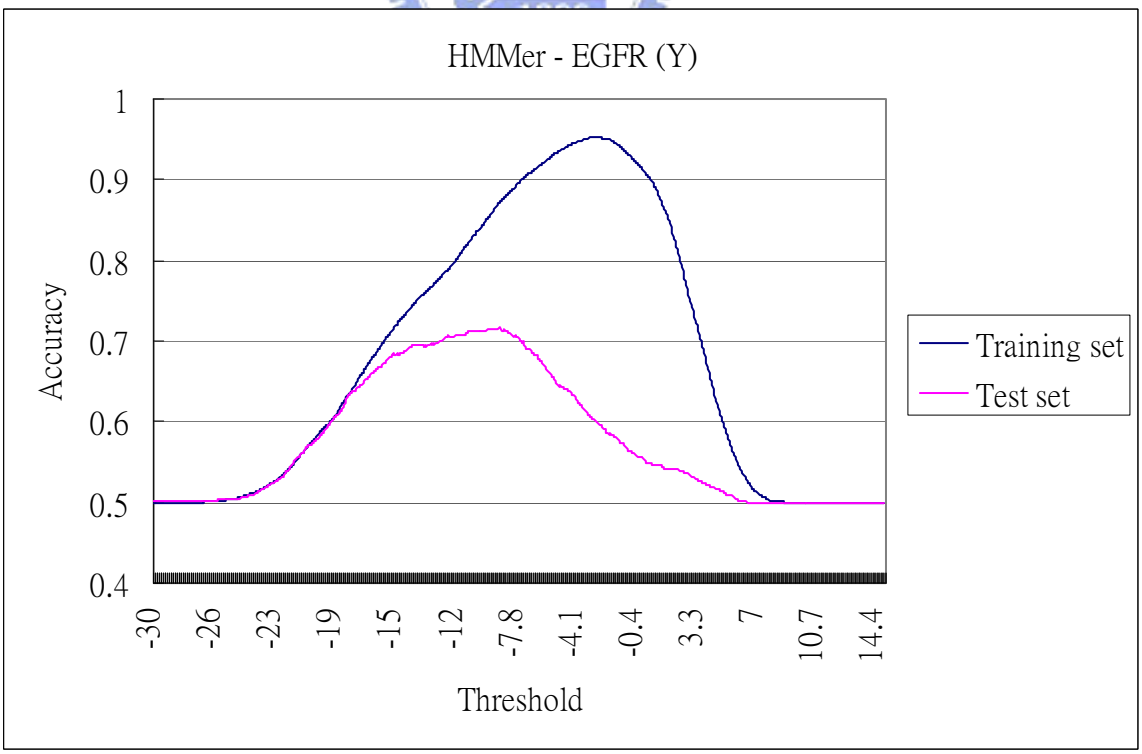

圖 A.93: EGFR (Y)資料於 HMMer 的門檻值與正確率對應圖

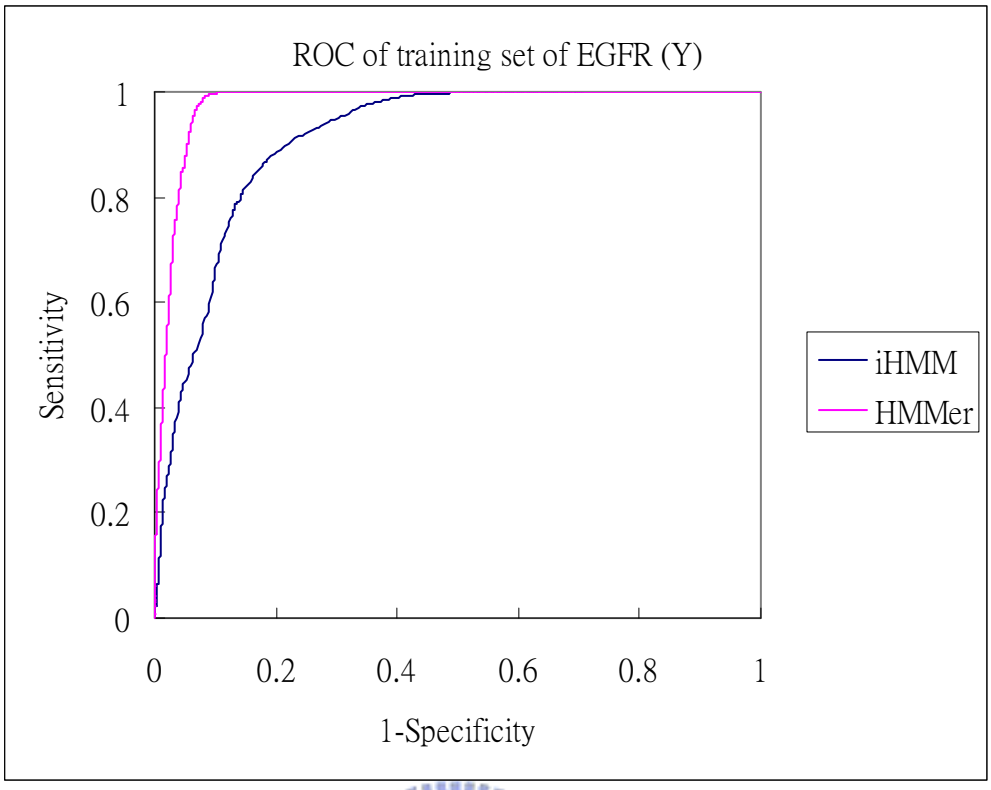

圖 A.94: EGFR (Y)於訓練資料的 ROC 圖

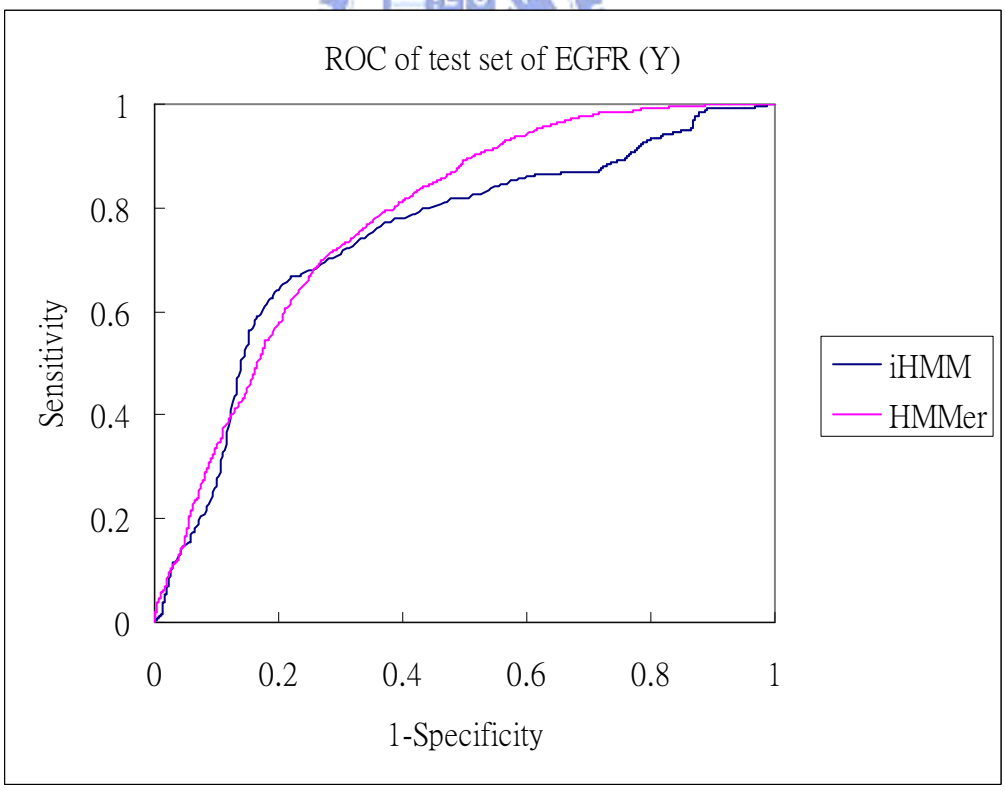

圖 A.95: EGFR (Y)於測試資料的 ROC 圖

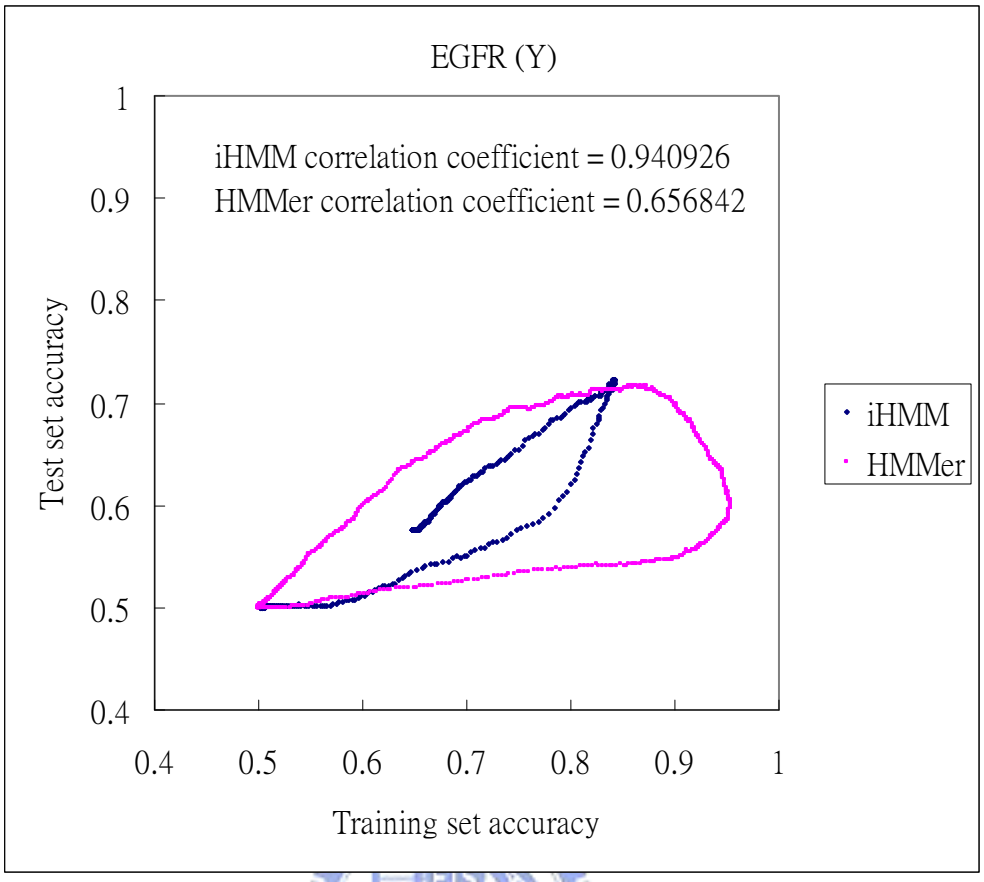

圖 A.96: EGFR (Y)資料的相關係數分析圖

INSR (Y)

全名：Insulin receptor

資料筆數：positive 跟 negative 各 58 筆資料

序列長度：15

磷酸化位置：中間的酪氨酸(Y)

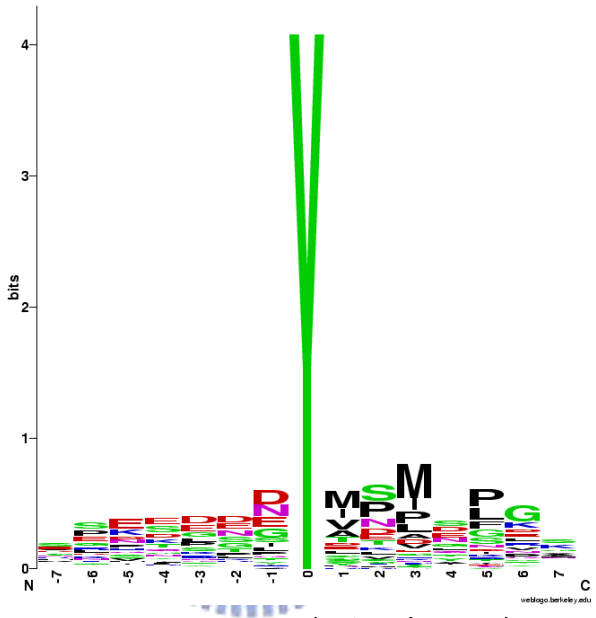

圖 A.97: INSR (Y)資料的序列圖案

表 A.17: INSR (Y)的 30 次 5-CV 於測試資料的效能比較表

INSR (Y)		Threshold	Sensitivity	Specificity	Precision	Accuracy
δ_1	HMMer	-3.6847 (0.3413)	0.4800 (0.0286)	0.8653 (0.0300)	0.7996 (0.0406)	0.6730 (0.0195)
	HMM-1	0.5323 (0.2534)	0.6727 (0.0428)	0.6842 (0.0631)	0.7030 (0.0367)	0.6780 (0.0223)
	iHMM	2.3958 (0.3739)	0.6864 ().0449)	0.7160 (0.0421)	0.7183 (0.0353)	0.7017 (0.0286)
δ_2	HMMer	-8.5620 (1.5519)	0.7484 ().0597)	0.7612 (0.0749)	0.7948 (0.0489)	0.7557 (0.0155)
	HMM-1	1.0020 (0.5906)	0.7073 (0.0714)	0.8176 (0.0708)	0.8213 (0.0537)	0.7639 (0.0156)
	iHMM	3.7868 (2.6263)	0.6601 (0.0663)	0.8865 (0.0505)	0.8864 (0.0406)	0.7748 (0.0192)

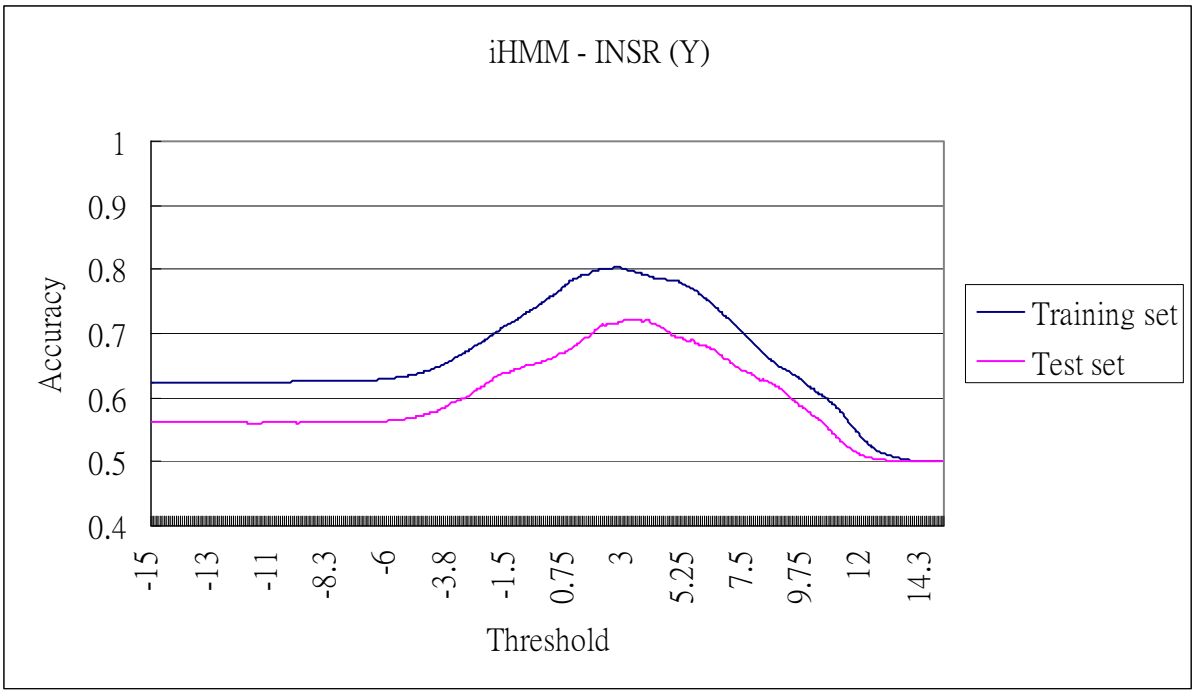

圖 A.98: INSR (Y)資料於 iHMM 的門檻值與正確率對應圖

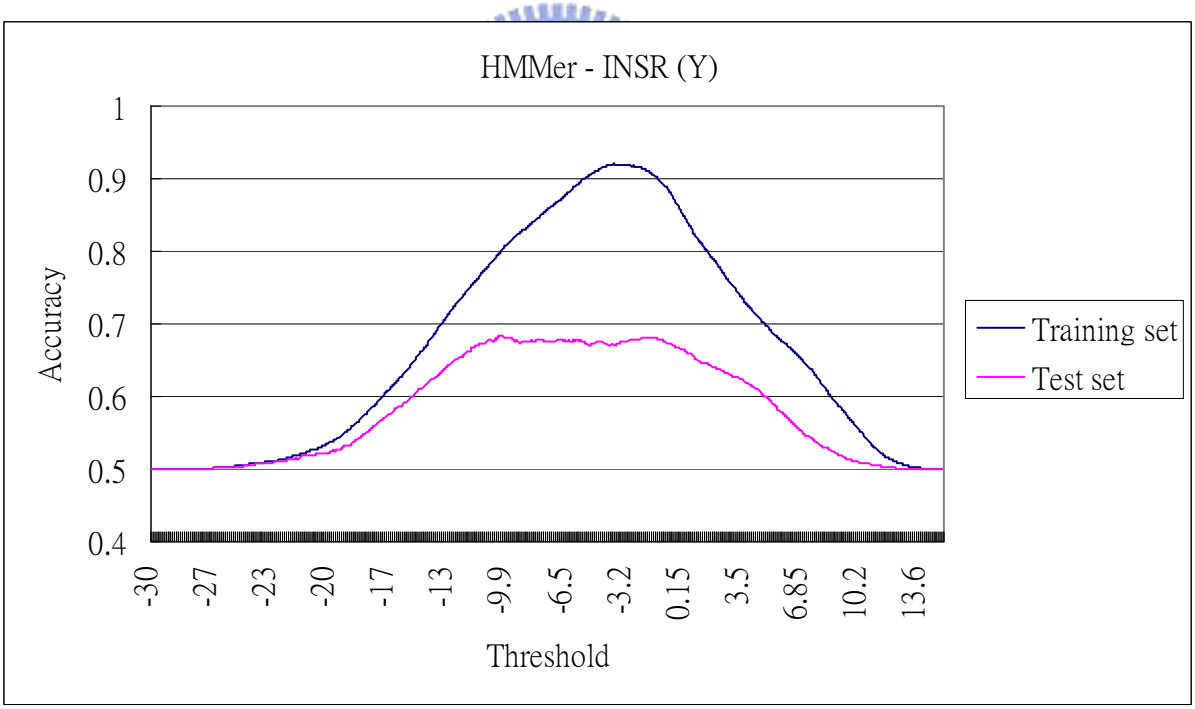

圖 A.99: INSR (Y)資料於 HMMer 的門檻值與正確率對應圖

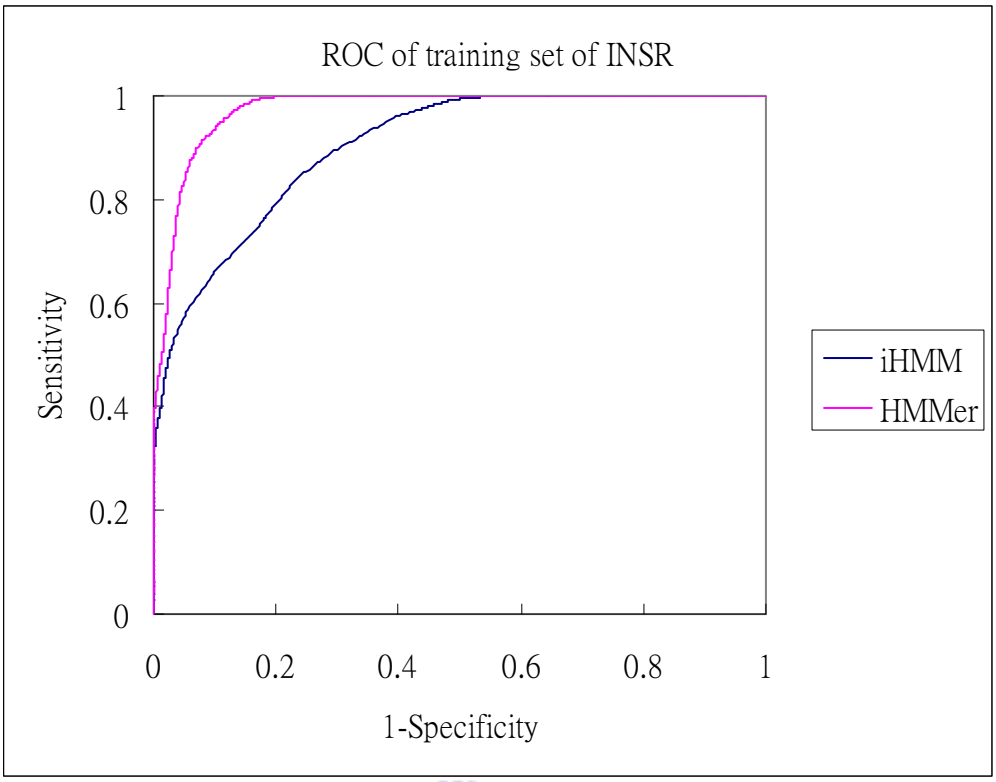

圖 A.100: INSR (Y)於訓練資料的 ROC 圖

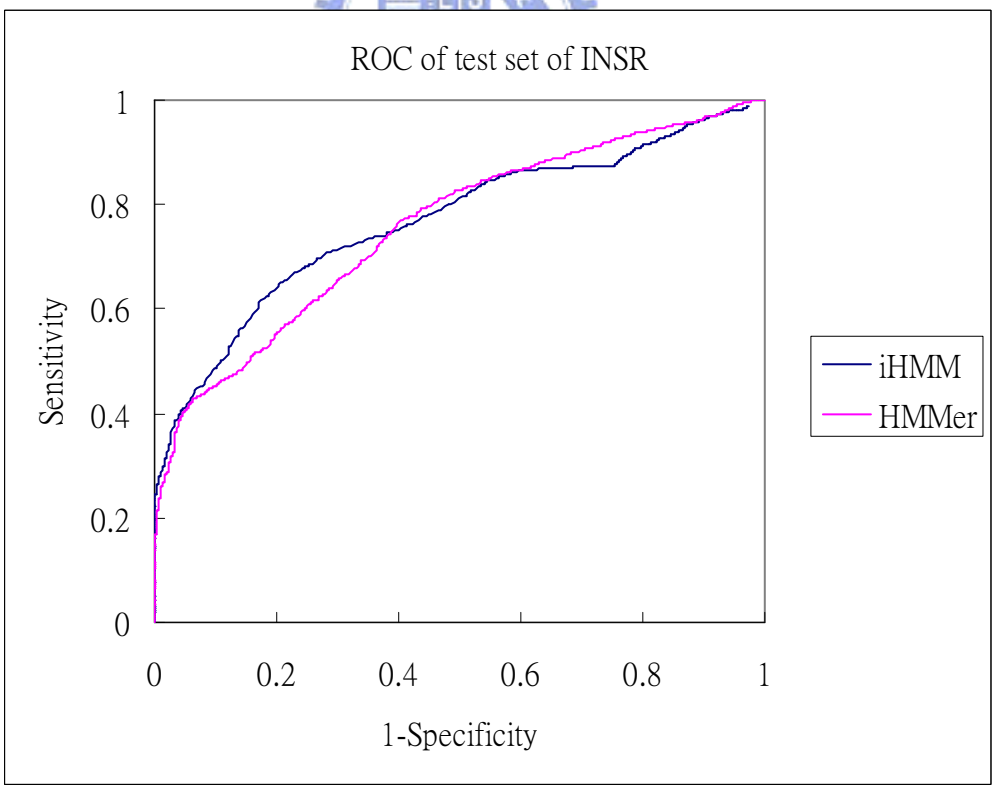

圖 A.101: INSR (Y)於測試資料的 ROC 圖

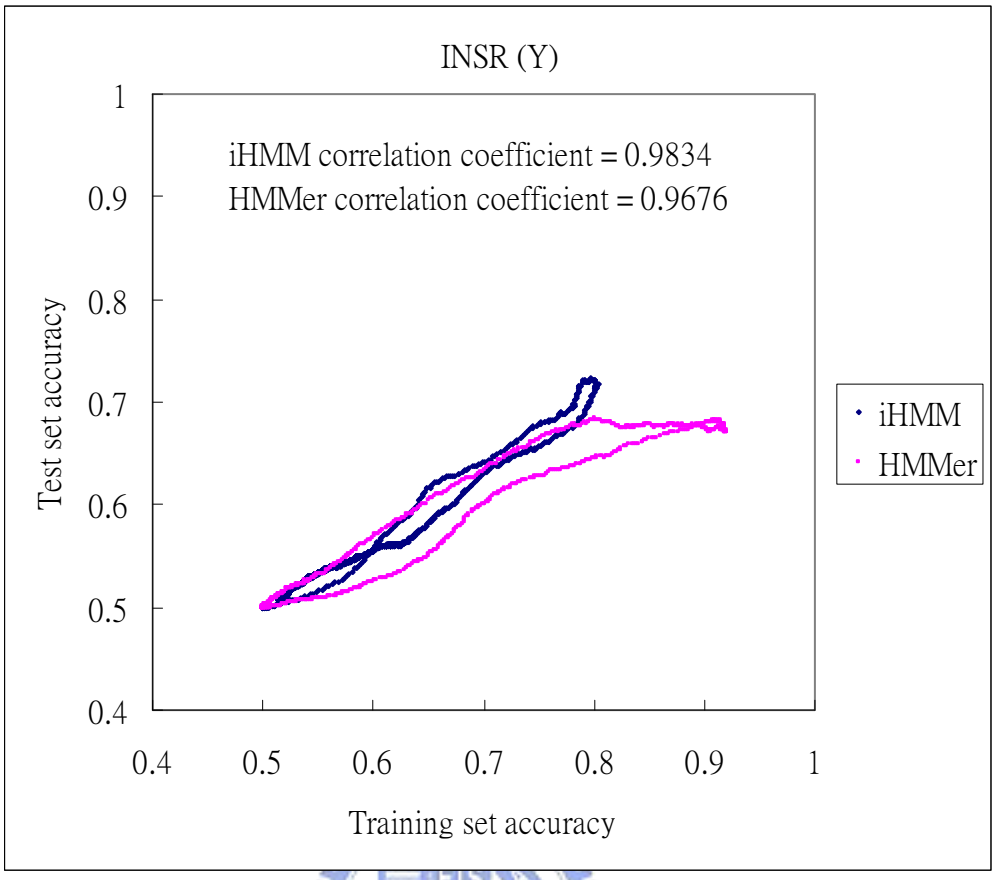

圖 A.102: INSR (Y)資料的相關係數分析圖

SRC (Y)

全名：Src kinase

資料筆數：positive 跟 negative 各 143 筆資料

序列長度：15

磷酸化位置：中間的酪氨酸(Y)

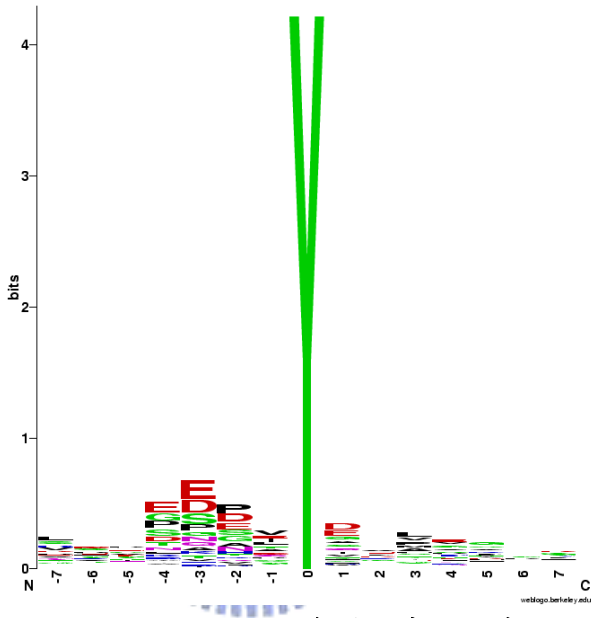

圖 A.103: SRC (Y)資料的序列圖案

表 A.18: SRC (Y)的 30 次 5-CV 於測試資料的效能比較表

SRC (Y)		Threshold	Sensitivity	Specificity	Precision	Accuracy
δ_1	HMMer	-4.4177 (0.1929)	0.6289 (0.0386)	0.8153 (0.0254)	0.7777 (0.0185)	0.7219 (0.0161)
	HMM-1	2.8829 (0.4552)	0.5883 (0.0459)	0.7599 (0.0399)	0.7196 (0.0220)	0.6741 (0.0135)
	iHMM	3.2385 (0.1364)	0.6497 (0.0366)	0.7528 (0.0285)	0.7290 (0.0207)	0.7013 (0.0180)
δ_2	HMMer	-5.5313 (0.6163)	0.7496 (0.0520)	0.7766 (0.0486)	0.7842 (0.0307)	0.7634 (0.0110)
	HMM-1	2.7656 (0.5988)	0.6563 (0.0621)	0.7862 (0.0571)	0.7803 (0.0396)	0.7216 (0.0120)
	iHMM	3.2253 (0.3075)	0.7012 (0.0346)	0.7856 (0.0347)	0.7819 (0.0266)	0.7438 (0.0140)

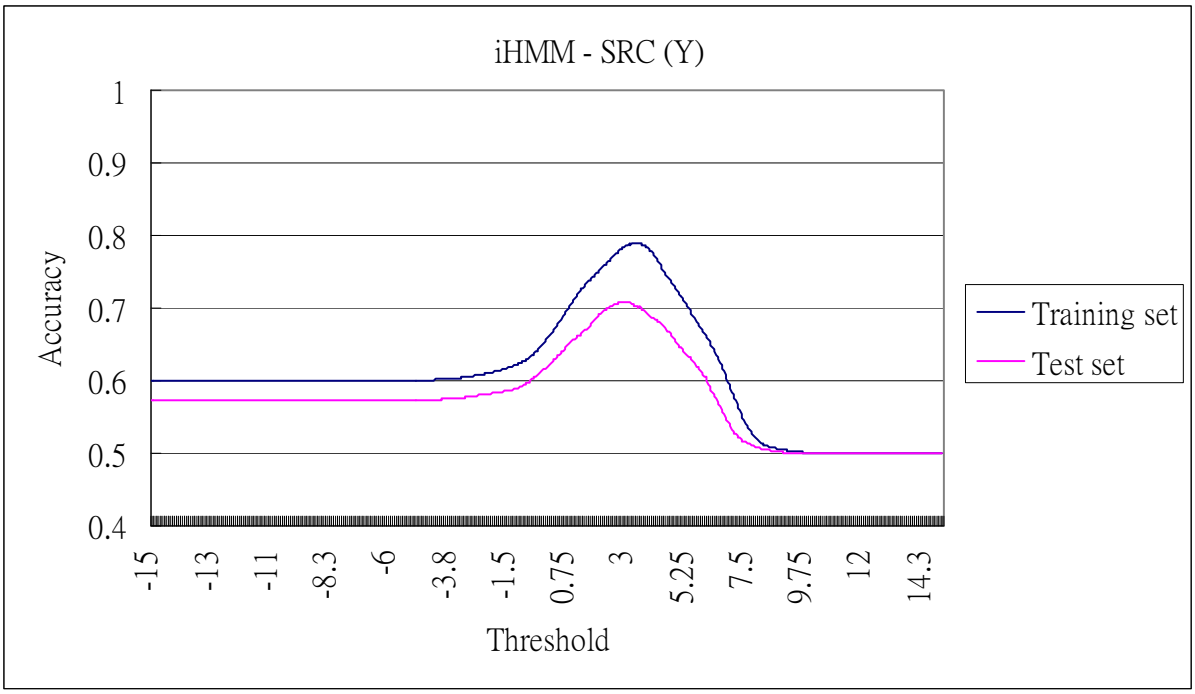

圖 A.104: SRC (Y)資料於 iHMM 的門檻值與正確率對應圖

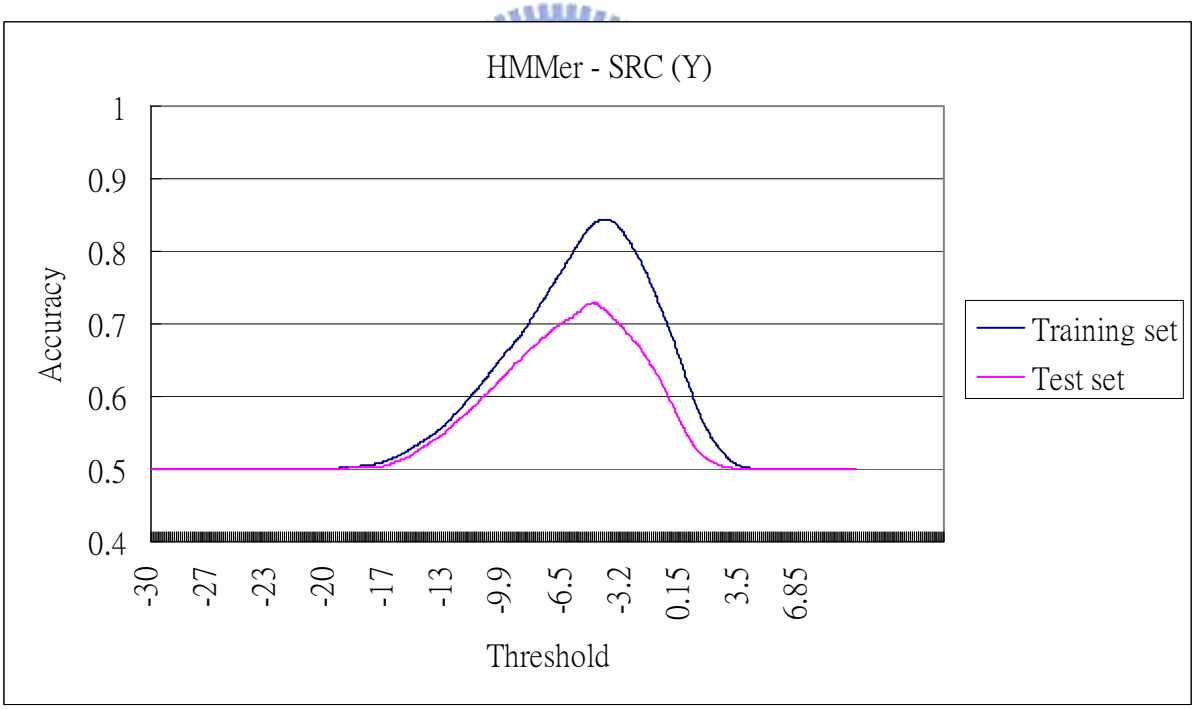

圖 A.105: SRC (Y)資料於 HMMer 的門檻值與正確率對應圖

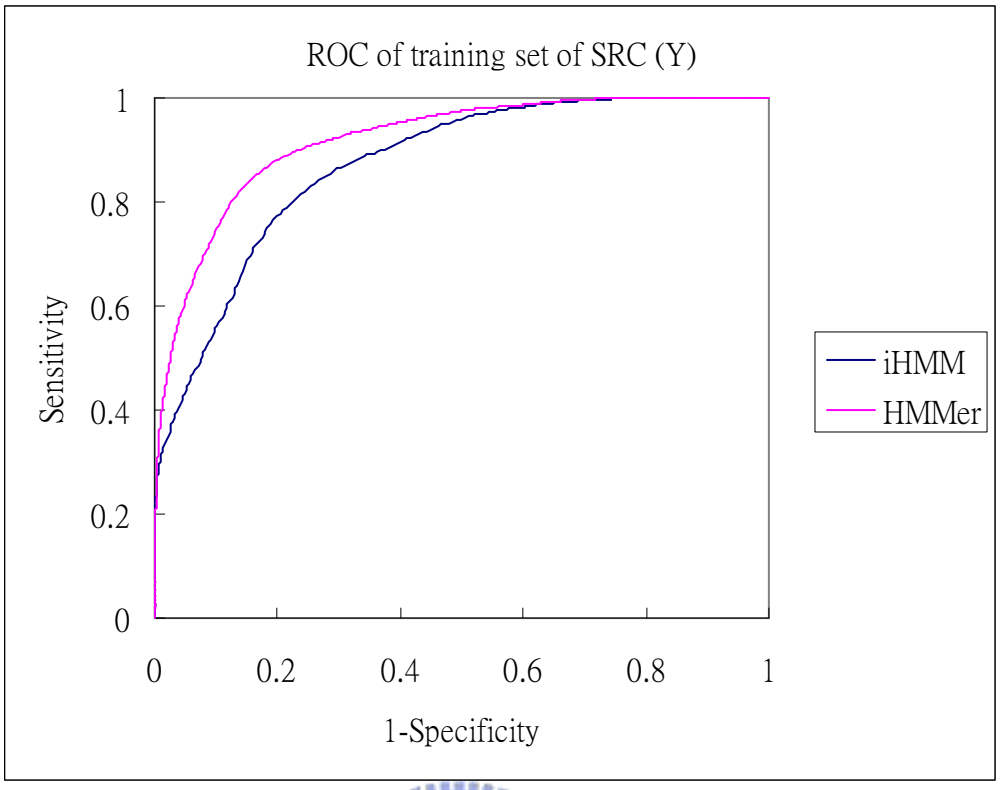

圖 A.106: SRC (Y)於訓練資料的 ROC 圖

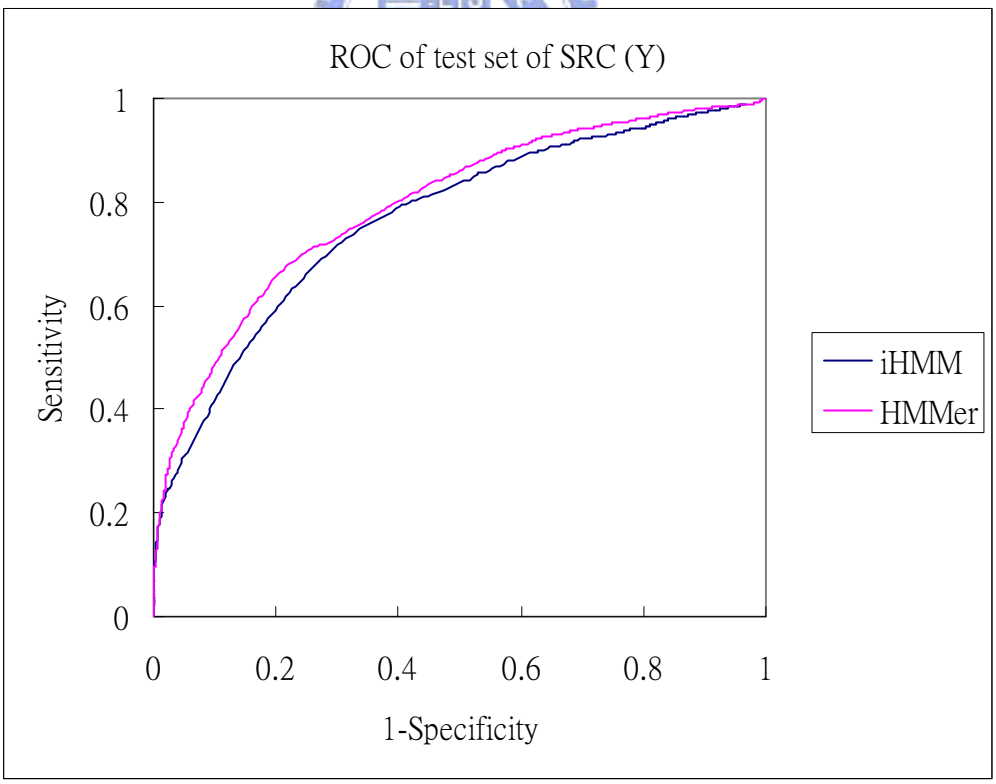

圖 A.107: SRC (Y)於測試資料的 ROC 圖

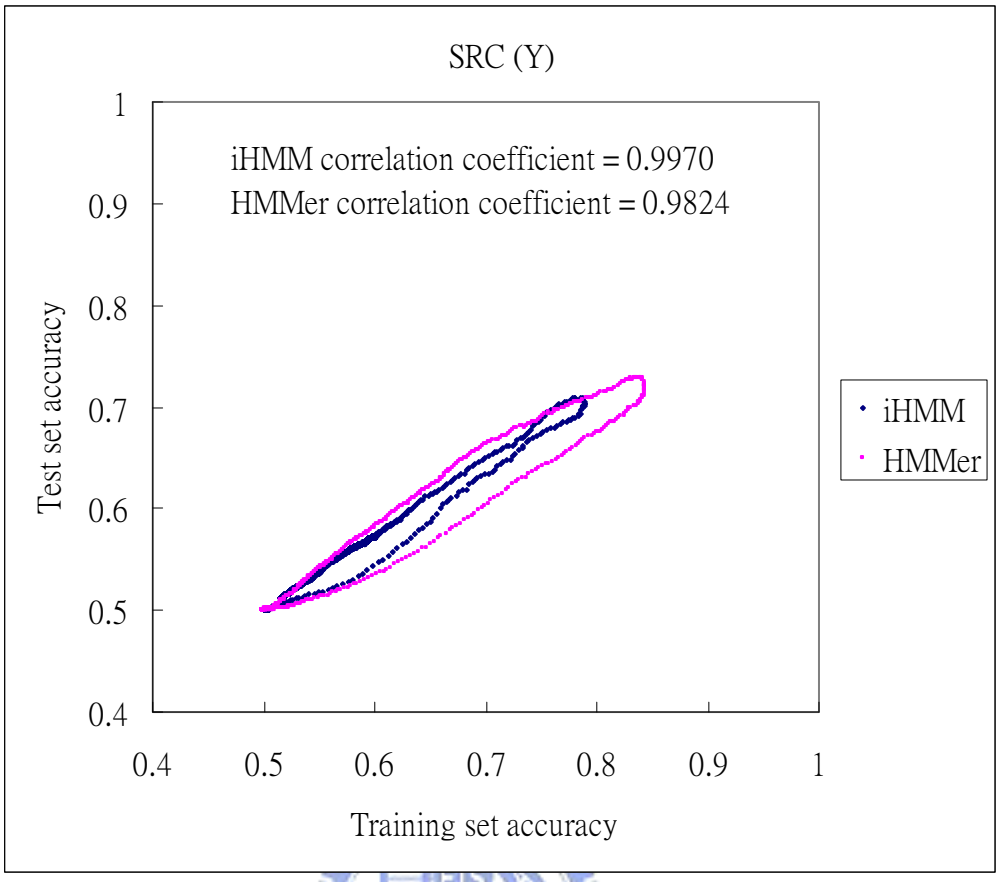

圖 A.108: SRC (Y)資料的相關係數分析圖