藉辨認模體來預測水解酶蛋白質之功能

Predicting the function of hydrolase proteins based on the recognition of motifs

曾怜玉[1,2]　王冠棋[2]　蔡杰松[2]

[1]國立中興大學資訊網路與多媒體研究所　[2]國立中興大學資訊科學與工程學系

{lytseng, s9456040, phd9603}@cs.nchu.edu.tw

摘要

蛋白質在生物體中扮演重要的角色，蛋白質有多種功能例如：運輸、運動、酵素、免疫、支持、保護等功能，而自從人類基因體計畫的完成，蛋白質功能的發現成為現在重要的工作之一，而蛋白質功能的發現,最正確的方法是屬於使用生物實驗所得到的結果，但是面對龐大數量的蛋白質，使用生物實驗必須消耗大量的金錢與時間，所以配合電腦預測蛋白質功能也成為重要方法之一，這些方法基本上分成三類，第一種是以序列為基準，第二種是以結構資訊為基準，最後是混合以上兩者，現在有龐大的蛋白質序列資料，大約還有三分之一沒有發現功能,但是有結構資訊的蛋白質又過少,所以本論文重點放在序列為主的蛋白質功能預測，藉著辨認出蛋白質的模體，判斷出蛋白質的功能,對於水解酶蛋白質之功能的辨識正確率已經達到 99%,未來將進行對其他種類的蛋白質進行預測。

關鍵詞：蛋白質，模體，功能預測，水解酶

ABSTRACT

Protein plays an important role in organisms. Protein has many kinds of functions, for example: transportation, sports, enzymes, immunization, support, protection, etc. After the Human Genome Project had been completed, the discovery of protein function becomes an important work. The method that uses biological experiments can achieve the most accurate results in finding the function of a protein. But lots of money and time are needed to find the functions of so many proteins. Therefore, predicting the function of a protein by computer is an important research issue. The computer methods in predicting protein function can be divided into three classes: methods based on sequences, methods based on structures, and methods that combine the previous two. In this thesis, a method based only on sequences to predict the protein function was proposed. This method predicts functions by recognizing the motifs containied in the sequence. In order to test the proposed method, an experiment has been conducted to predict the seven subfunctions of hydrolase proteins. The accuracy achieved is above 99%.

Keywords ： Protein, motif, function prediction, hydrolase

一、前言

蛋白質在生物體中，扮演著重要的角色，在大部分的生命化學作用之中，蛋白質都扮演不可或缺的角色，例如：物質運輸、作為酵素、運動、支架、保護等等，許許多多在生物體中的反應都跟蛋白質有關，甚至蛋白質也已經研發成為藥物可以治療疾病，所以了解更多的蛋白質是目前重要的工作，但是蛋白質是一個複雜的巨大分子，而且種類很繁複，所以要徹底的了解蛋白質，不是簡單的工作。

在人類基因體計畫之下，人類染色體的所有序列都已經被解碼完成，但是在找到龐大的基因之後，真正的工作才正要開始，根據UniProtKB已經收集了 4,448,557 蛋白質序列(2007/7/12)，但是根據PDB也只收集了 44,018 個蛋白質立體結構(2007/7/12)，遠遠少於已找到的序列，而這些序列中有Gene Ontology (GO)功能註解的，在UniProtKB有 3,048,127(2007/6/8)，PDB[32] 也有GO註解但也不是全部，而真正的工作就是要完全解出蛋白質的功能，及應用在人體醫療疾病方面，而在這個序列大量被發現的時候，蛋白質功用預測一般就是以實驗的方法探知蛋白質功能，是最準確可以知道蛋白質功能的方法，但是其缺點就是需要消耗大量時間與金錢，除了實驗以外，使用電腦來預測其功能也是方法之一，雖然沒有實驗上的缺點，但是卻有一個致命的缺點，就是預測結果不盡理想，所以希望在預測蛋白質功能的準確率上能夠更進一步。

蛋白質的功能跟它的立體結構有很大的關係，所以許多在預測功能的研究上能夠讀取結構的資訊，作為預測功能的資訊，但是利用實驗解出來的結構資訊的數量卻遠少於現在龐大的序列資訊，而且若要使用結構作為預測的參考，必定需要了解蛋白質的結構資訊，所以現階段我們希望能夠提升以序列為基準來預測蛋白質功能的準確率。

二、文獻探討

在蛋白質功能預測有許多分法，但是有一個問題存在，就是對蛋白質功能有許多觀點，許多預測方法是否在同一個觀點上比較，而功能的層次上可從生化功能到生物程序，更會大到器官、生物體等[27]，所以蛋白質也可以分成不同的功能，對蛋白質的功能註解而比較有名的，有下列兩種資料庫 Enzyme Commission(EC)[27] 、 Gene Ontology(GO) [27] 。 Enzyme Commission(EC)[27] 是一個知名的酵素分類的資料庫，將酵素分成六大類，如：

第一類功能為 Oxidoreductases

第二類功能為 Transferases

第三類功能為 Hydrolases

第四類功能為 Lyases

第五類功能為 Isomerases

第六類功能為 Ligases

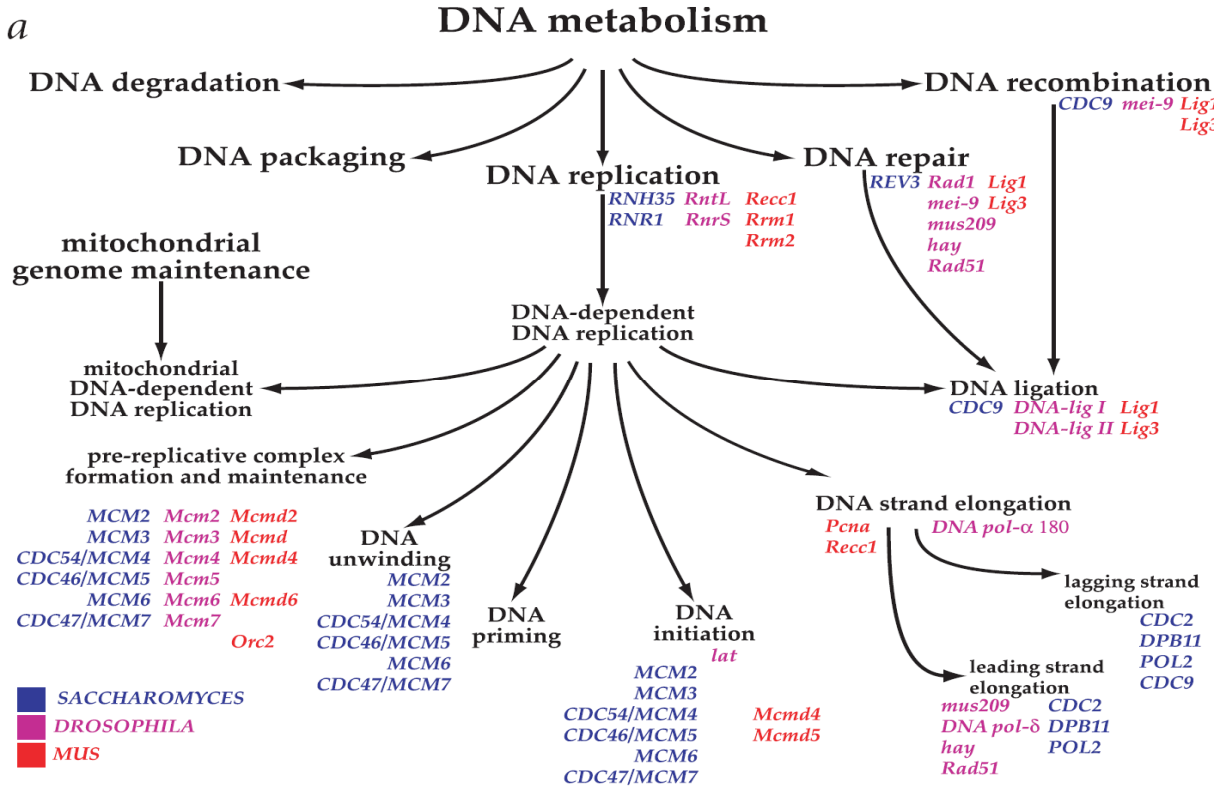
並給予一個專屬號碼來表示其功能，它有四個號碼，中間以”.”分隔，表示四種層次，例如 EC 1.1.1.1 是一種 alcohol dehydrogenase，執行這樣的反應：

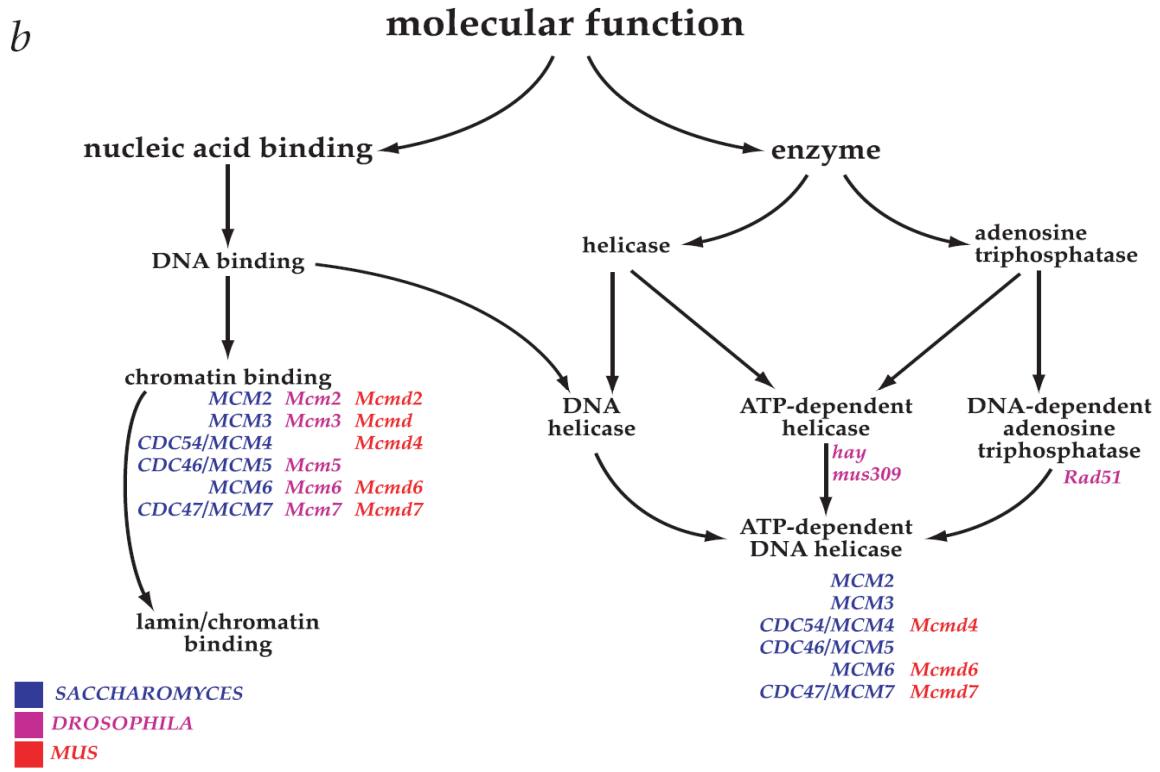
$$\text{an alcohol} + NAD^+ = \text{an aldehyde}$$

or

$$\text{ketone} + NADH + H^+$$

現在較常使用之蛋白質功能分類為Gene Ontology(GO) [27] ，GO收集許多不同資料庫的基因產物，然後將基因產物以biological processes、cellular components和molecular function註解，biological processes是收集一連串分子功能的集合，molecular function就是在分子層面討論個別基因產物的活性，cellular components記錄基因產物屬於在細胞中的那個部位。GO分類的架構如圖1，GO有兩個特點，它是公開的資源可供大家使用，而且它的資料都是機器可讀的，因為這兩個原因，使得GO成為在生物資訊界中重要的資料庫，而且在功能預測方面提供相當大的幫助。


a
DNA metabolism
DNA degradation
DNA packaging
DNA replication
DNA repair
DNA recombination
CDC9 mei-9 Lig
Lig
RNH35 RntL Recc1
RNR1 RnrS Rrm1
Rrm2
REV3 Rad1 Lig1
mei-9 Lig3
mus209
hay
Rad51
mitochondrial genome maintenance
DNA-dependent DNA replication
mitochondrial DNA-dependent DNA replication
DNA ligation
CDC9 DNA-lig I Lig1
DNA-lig II Lig3
pre-replicative complex formation and maintenance
MCM2 Mcm2 Mcmd2
MCM3 Mcm3 Mcmd
CDC54/MCM4 Mcm4 Mcmd4
CDC46/MCM5 Mcm5
MCM6 Mcm6 Mcmd6
CDC47/MCM7 Mcm7
Orc2
DNA unwinding
MCM2
MCM3
CDC54/MCM4
CDC46/MCM5
MCM6
CDC47/MCM7
DNA priming
DNA initiation
lat
MCM2
MCM3
CDC54/MCM4
CDC46/MCM5
MCM6
CDC47/MCM7
Mcmd4
Mcmd5
DNA strand elongation
Pcna
Recc1
DNA pol-α 180
lagging strand elongation
CDC2
DPB11
POL2
CDC9
leading strand elongation
mus209 CDC2
DNA pol-δ DPB11
hay POL2
Rad51
SACCHAROMYCES
DROSOPHILA
MUS


b
molecular function
nucleic acid binding
enzyme
DNA binding
helicase
adenosine triphosphatase
chromatin binding
MCM2 Mcm2 Mcmd2
MCM3 Mcm3 Mcmd
CDC54/MCM4 Mcmd4
CDC46/MCM5 Mcm5
MCM6 Mcm6 Mcmd6
CDC47/MCM7 Mcm7 Mcmd7
lamin/chromatin binding
DNA helicase
ATP-dependent helicase
hay
mus309
DNA-dependent adenosine triphosphatase
Rad51
ATP-dependent DNA helicase
MCM2
MCM3
CDC54/MCM4 Mcmd4
CDC46/MCM5
MCM6 Mcmd6
CDC47/MCM7 Mcmd7
SACCHAROMYCES
DROSOPHILA
MUS

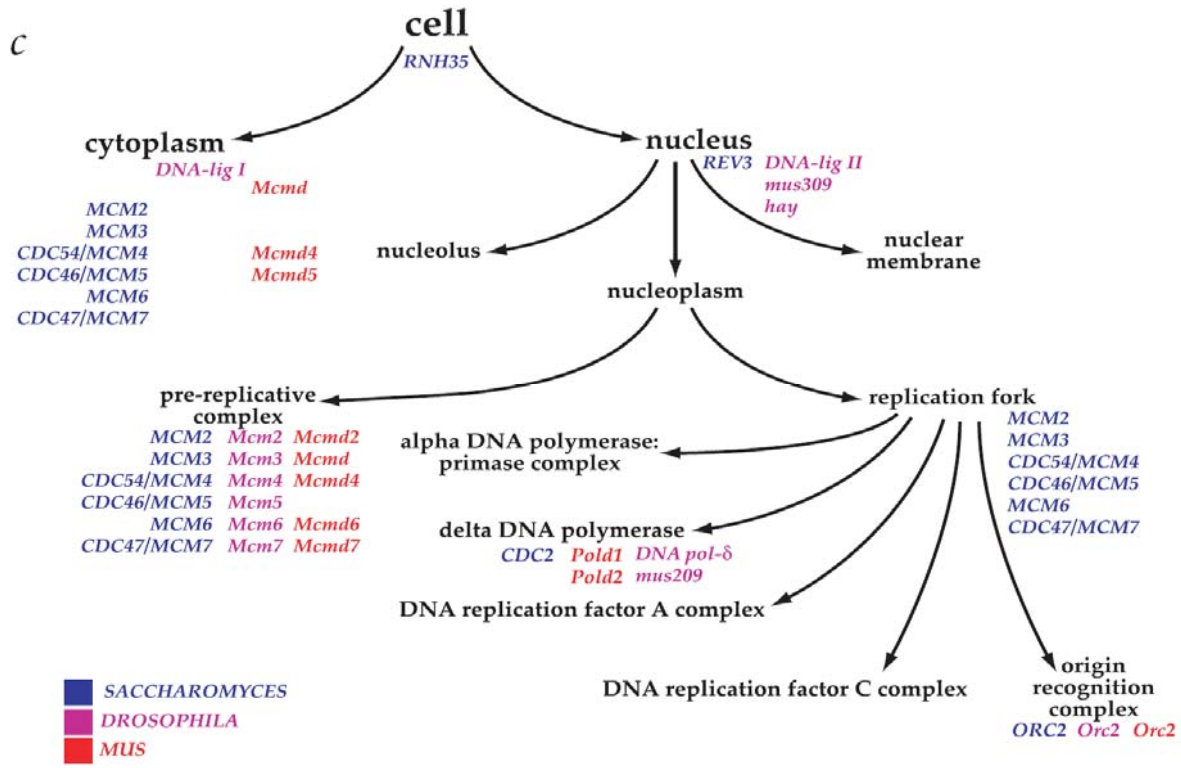


圖 1. GO 三種功能註解的例子(a)DNA(b)分子功能(c)細胞[27]

功能預測方法

以往利用蛋白質序列和結構預測蛋白質功能的研究，大致可以分成三類，分別為序列基準方法(sequence-based method)、結構基準方法(structure-based method)、混合式方法(combining method)，接下來詳細介紹這三類方法。

序列基準(sequence-based method)

序列基準方法(sequence-based method)在預測蛋白質的功能是基於序列的比對，最早的方法就是以序列與序列之間的比對，例如：BLAST或FASTA，接著出現以樣板為基礎（pattern-based）或輪廓(profile)比對的方法，最有名的就是隱藏馬可夫模型(Hidden Markov Model，HMM)[29]，而PSI-BLAST也是現在常用於預測的工具之一，這類的方法都屬於序列-樣板 (sequence-pattern)之間的比對，比序列-序列比對更具有敏感度(sensitivity)。

比序列-樣板比對更具有敏感度的方法就是profile-profile比對，有一種被稱為HHsearch[20] 的方法就是利用這種方式比較，比上述的幾種方法更快也更敏感。

結構基準(structure-based method)

在某種情形，以序列為基準的預測方法會失敗，有同源的序列卻執行不同的功能的蛋白質存在，而功能的產生通常跟它的3D結構有關，所以利用結構的資訊來預測蛋白質的功能，很直覺的被採用，不過利用結構來預測，有多一項限制就是必須知道待測的蛋白質結構為何，對於未知結構的蛋白質不是很有用。以下介紹幾種常見以結構為基準的預測方式。

摺疊相似(Fold)：這種方式的基本假設是，執行相似功能的蛋白質，可能會有相似的摺疊方式，但是有時候在演化中蛋白質改變功能，但保持相同摺疊，所以有相同的摺疊方式的蛋白質會有不同功能，所以利用摺疊相似可能會造成錯誤，利用這種方法的有DALI[27] 、SSM[12] 、GEATH[7] 、VAST[27] 等。

表面凹陷(surface clefts)：蛋白質摺疊之後，在表面可能會形成凹陷，而這凹陷可能會與特定的物

質結合，這些特定的物質可能是輔因子(cofactors)，而藉由與表面凹陷結合，調節酵素的活性，所以分析這些表面凹陷，可以進一步了解蛋白質的功能，但是對於那些沒有輔因子相結合的蛋白質，可能就沒辦法分析表面凹陷，來預測其功能。常見利用這種方式預測蛋白質功能的有：pvSOAR[4] 、CASTp[27] 、SURFACE[6] 等。

殘基模板(residue template)：有些蛋白質的功能是因為較局部性的小片段殘基所形成，所以預測這些小片段殘基，也可預測蛋白質的功能，特別是在酵素或是DNA結合蛋白，這些種類的蛋白質的活性區域都是小片段的殘基，所以辨認出那些關鍵性的小片段殘基的構造，就可以辨認出蛋白質功能，這類的研究較多，以下是根據這樣的方式預測蛋白質功能的方法：Catalytic Site Atlas (CSA) [17] 、PDBSiteScan [9] 、DRESPAT[24] 、SuMo [12] 、ASSAM [21] 、RIGOR/SPASM [11] 、PINTS[27] 、PDNA-pred server[18] 、PreDS server[23] 。

混合式(combining method)

為了增加預測的準確率，現在都結合許多方法，而產生一個新方法，以下兩種方法都是使用這種策略：ProFunc[13] 、ProKnow[27] 。

模體(Motif)和蛋白質功能的關係

一般來說，超二級結構，又稱模體(motif)，是由數個二級結構所組成的一個小單位，被認為是蛋白質立體結構的功能組件，但在有些說法中模體(motif)是指在演化過程中被高度保留的序列區域，也是構成功能、結構的主要序列區域，所以預測蛋白質中存在的模體(motif)，可以作為預測蛋白質功能結構的方法[30] 。

序列模體(motif)非常適合用來預測某些蛋白質功能與序列分類，特別在一些酵素的活性區域，包含某些特定胺基酸序列的排列，甚至這些特定的胺基酸排列也可能出現在其他蛋白質，所以這種重複性很高，也具高保留性的序列模體(motif)，可以說是在預測生化功能上的重要利器[13]，本論文正是利用辨識模體來預測水解酶蛋白質的功能。

蛋白質資料庫

PDB

Protein data bank(PDB)[32] 是收集了大量的生物分子立體結構的資料庫，包括蛋白質和核苷酸，資料庫中包含很多物種，有人、鼠、酵母菌、植物等等。所收集的立體結構是利用X-ray、NMR、Electron Microscopy實驗解出來的，到 2007 年 7 月 12 日為止，PDB已經收集了 44,018 個蛋白質結構，現在持續在更新中。

UniProtKB/Swiss-Prot

UniProtKB/Swiss-Prot[28] 從 1986 年開始收集大量的蛋白質序列，它結合很多資料庫，包含結構功能的資料庫，並註解蛋白質序列，到 2007 年 7 月 12 日為止，UniProtKB/Swiss-Prot 收集了 270,778 序列，包含 99,412,397 個胺基酸，其中也包含許多物種(人、老鼠、酵母菌、植物、細菌、病毒等等)。

PRINTS

PRINTS[30] 是一個蛋白質模體(motif)的資料庫，裡面收集了 1,900 個fingerprints，其中包括 11,170 個模體(motif)(in year 2005)，fingerprint是一組模體(motif)，可用來預測有相似的模體(motif)的蛋白質之功能，所以可以利用fingerprint來預測蛋白質功能。

三、研究方法

方法流程

本研究方法有五大步驟，第一步是從PDB[32]選定GO ID16787 的蛋白質，第二步是找出對應的PRINTS的模體(motif)，第三步是利用clustalW[26]將每個模體(motif)分群，第四步是利用分群好模體(motif)的序列，建立記分矩陣，最後第五步使用交叉測試，預測出待測蛋白的功能，得到最後準確率，圖 2 是本實驗的流程圖，而在下面一節詳細介紹每一步驟。

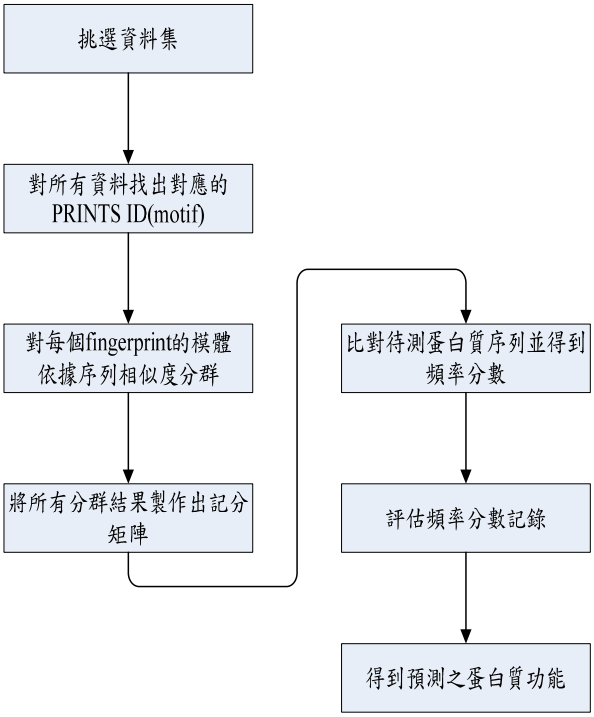


圖 2. 方法流程

資料庫的挑選

本實驗擷取 PDB 中，在 GO 的架構中屬於分子功能的水解酶(hydrolase GO ID16787)為註解的蛋白質，在 GO 的架構之下，把水解酶(hydrolase GO ID16787)分 9 種次功能分類，此九種次功能分類分別為 16788、16798、16801、16810、16817、16822、16824、19213、8233。以這九種功能分類為註解的 PDB 蛋白質，其數量有 6011 條，但經過挑選，可以找到有對應的 PRINTS ID 的 PDB 蛋白質，如附錄二所示，統計數量如表 1：總數量有 3519 條蛋白質。

表 1. 使用 GOID 註解之 PDB 蛋白質數量

GOID	16788	16798	16801	16810	16817
數量	685	1080	3	176	66
GOID	16822	16824	19213	8233	
數量	0	8	0	1501	

如附錄三所示，而統計的數量如表 2：總數量有 133 個 PRINTS 資料

表 2.相對應 GOID 註解之 PRINTS fingerprint 數量

GOID	16788	16798	16801	16810	16817
數量	37	27	1	6	12
GOID	16822	16824	19213	8233	
數量	0	1	0	49	

因為 GO ID16822 及 19213 中找不到有相對應的 PRINTS ID(fingerprint)所以接下來的動作就不執行。

分群與記分矩陣

在個別PRINTS的fingerprint包含許多組的模體 (motif)，而每一組的模體(motif)包含許多相關的序列，這些序列是從許多物種中萃取出來的，所以序列不太相似，因為要利用這些序列作出記分矩陣，為了使記分矩陣公平，不會因為數量不足而減少積分，所以先對每一組做分群，在對每一群作出記分矩陣[32] 。

本實驗在切分群時，是把 clustalW 所得到的演化樹，切到第一層，對應每個 GO ID 的分群個數如表 3 所示：其總數量為 2206。

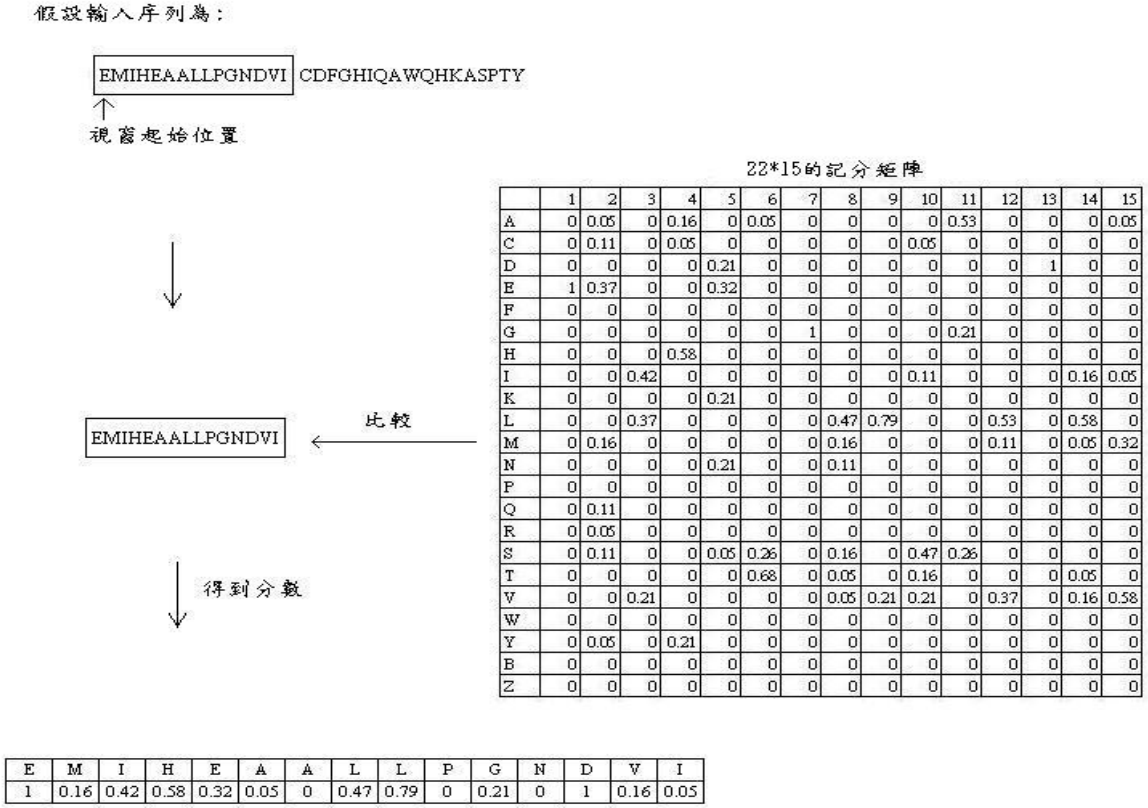
表 3. GOID 所含分群及記分矩陣數量

GOID	16788	16798	16801	16810
數量	597	535	15	87
GOID	16817	16824	8233	
數量	195	21	756	

製作記分矩陣的數量與分群相同。

預測方式

本實驗使用比對法的方式預測，假設想預測序列大小為 N 的蛋白質序列，先使用滑動視窗，從序列第一個開始，視窗的大小會根據記分矩陣的大小做改變，假設第一個記分矩陣為 22*15，則視窗大小為 15，接下來依據每個位置的胺基酸去檢查記分矩陣相對位置的胺基酸的分數，而得到這個位置的分數，圖 3 是一個記分矩陣為 22*15 的例子。



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	0	0.05	0	0.16	0	0.05	0	0	0	0	0.53	0	0	0	0.05
C	0	0.11	0	0.05	0	0	0	0	0	0.05	0	0	0	0	0
D	0	0	0	0	0.21	0	0	0	0	0	0	0	1	0	0
E	1	0.37	0	0	0.32	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	1	0	0	0	0.21	0	0	0	0
H	0	0	0	0.58	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0.42	0	0	0	0	0	0	0.11	0	0	0	0.16	0.05
K	0	0	0	0	0.21	0	0	0	0	0	0	0	0	0	0
L	0	0	0.37	0	0	0	0	0.47	0.79	0	0	0.53	0	0.58	0
M	0	0.16	0	0	0	0	0	0.16	0	0	0	0.11	0	0.05	0.32
N	0	0	0	0	0.21	0	0	0.11	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0.11	0	0	0	0	0	0	0	0	0	0	0	0	0
R	0	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0.11	0	0	0.05	0.26	0	0.16	0	0.47	0.26	0	0	0	0
T	0	0	0	0	0	0.68	0	0.05	0	0.16	0	0	0	0.05	0
V	0	0	0.21	0	0	0	0	0.05	0.21	0.21	0	0.37	0	0.16	0.58
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0.05	0	0.21	0	0	0	0	0	0	0	0	0	0	0
B	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

圖 3. 拿待測蛋白質去比對記分矩陣之流程

每個位置都比對完畢以後，再將所有位置的分數平均，成為這次比對的分數，接下來以序列第一個位置的視窗，會比較所有的記分矩陣，並將分數記錄下來，比較完以後，視窗會向下移動一位置，以序列第二個位置為首的視窗，開始做同樣的比對，並記錄下分數，依此類推，做完全部序列，我們只取分數前在 150 名的記錄。

這個記錄中的每一筆會包含 GO ID、PRINTS ID、這是第幾個模體(motif)、個別的 PRINTS 的 fingerprint 中包含幾組模體(motif)、分群結果的第幾群、在序列中的起始位置、在序列中的結束位置、分數，附錄一是一個例子，是 PDB ID 1ELL 比對的結果為例。

根據這個記錄決定出這次預測的結果，接下來看所有的預測出來的 PRINTS 的 fingerprints，計算出它們的分數，計算方法為，將在這紀錄中的屬於這個 fingerprint 的所有模體(motif)找出來，將所有模體(motif)中的最高分數做為這個模體(motif)代表，取所有模體代表的分數之平均值，最後再加上所有模體(motif)代表的分數最小的那一個，做為這個 PRINTS fingerprint 的分數，計算出全部的 PRINTS fingerprints 的分數之後，取最高值當作這次預測的結果。

根據附錄一的例子，找出所有 PRINTS fingerprints 的分數，在這裡列出前五名，如表 4：

表 4. PDB 1ELL 預測之 PRINTS 和 GO 前五名

GO ID	PRINTS ID	分數
GO-ID8233	PR00765	0.984867216
GO-ID16798	PR00911	0.251662887
GO-ID16788	PR00880	0.241269841
GO-ID16788	PR00377	0.220668221
GO-ID8233	PR00726	0.181502525

根據表 4 得到第一名為 GO ID8233 PR00756，所以預測結果為 GO ID8233。

實驗結果及分析

本實驗使用交叉測試，所在 PDB 所得到的以

GO 註解的水解酶(Hydrolase) 隨機挑選分成十份，每次拿一份作為測試集，剩下的九份則作為訓練集，這樣做十次每一份都會被當作是測試集一次，用訓練集進行實驗步驟，找出其對應的 PRINTS 的 fingerprints，對其模體(motif)進行分群，再將分群結果製作記分矩陣，對測試集則利用這些記分矩陣進行功能預測。這樣的步驟進行十次，將分別對所有功能計算敏感度(sensitivity)、識別度(specificity)、及正確率(accuracy)，及最後總結果，表 5 是這十次實驗結果：

表 5. 實驗結果

1

	TP	TN	FN	FP	sensitivity	specificity	accuracy
GO-ID16788	76	271	1	3	0.987012987	0.989051095	0.988603989
GO-ID16798	114	234	3	0	0.974358974	1	0.991452991
GO-ID16801	1	349	0	1	1	0.997142857	0.997150997
GO-ID16810	14	337	0	0	1	1	1
GO-ID16817	7	339	5	0	0.583333333	1	0.985754986
GO-ID16824	2	348	0	1	1	0.99713467	0.997150997
GO-ID8233	126	217	2	6	0.984375	0.97309417	0.977207977
平均					0.932725756	0.993774685	0.991045991

2

	TP	TN	FN	FP	sensitivity	specificity	accuracy
GO-ID16788	72	273	2	3	0.972972973	0.989130435	0.985714286
GO-ID16798	127	217	5	1	0.962121212	0.995412844	0.982857143
GO-ID16801	0	350	0	0	Na	1	1
GO-ID16810	12	338	0	0	1	1	1
GO-ID16817	1	348	1	0	0.5	1	0.997142857
GO-ID16824	1	349	0	0	1	1	1
GO-ID8233	126	214	3	7	0.976744186	0.968325792	0.971428571
平均					0.901973062	0.99326701	0.991020408

3

	TP	TN	FN	FP	sensitivity	specificity	accuracy
GO-ID16788	61	289	1	1	0.983870968	0.996551724	0.994318182
GO-ID16798	90	259	3	0	0.967741935	1	0.991477273
GO-ID16801	2	348	0	2	1	0.994285714	0.994318182
GO-ID16810	24	328	0	0	1	1	1
GO-ID16817	4	346	2	0	0.666666667	1	0.994318182
GO-ID16824	0	352	0	0	Na	1	1
GO-ID8233	162	183	2	5	0.987804878	0.973404255	0.980113636

平均					0.934347408	0.994891671	0.993506494
					4		
	TP	TN	FN	FP	sensitivity	specificity	accuracy
GO-ID16788	67	280	1	2	0.985294118	0.992907801	0.991428571
GO-ID16798	109	238	3	0	0.973214286	1	0.991428571
GO-ID16801	0	350	0	0	Na	1	1
GO-ID16810	15	334	1	0	0.9375	1	0.997142857
GO-ID16817	2	344	4	0	0.333333333	1	0.988571429
GO-ID16824	3	347	0	0	1	1	1
GO-ID8233	144	197	1	8	0.993103448	0.96097561	0.974285714
平均					0.870407531	0.993411916	0.991836735
					5		
	TP	TN	FN	FP	sensitivity	specificity	accuracy
GO-ID16788	60	287	0	4	1	0.986254296	0.988603989
GO-ID16798	103	246	2	0	0.980952381	1	0.994301994
GO-ID16801	0	351	0	0	Na	1	1
GO-ID16810	18	332	1	0	0.947368421	1	0.997150997
GO-ID16817	5	345	1	0	0.833333333	1	0.997150997
GO-ID16824	1	350	0	0	1	1	1
GO-ID8233	158	189	2	2	0.9875	0.989528796	0.988603989
平均					0.958192356	0.996540442	0.995115995
					6		
	TP	TN	FN	FP	sensitivity	specificity	accuracy
GO-ID16788	66	283	1	0	0.985074627	1	0.997142857
GO-ID16798	111	239	0	0	1	1	1
GO-ID16801	0	350	0	0	Na	1	1
GO-ID16810	16	334	0	0	1	1	1
GO-ID16817	9	341	0	0	1	1	1
GO-ID16824	0	350	0	0	Na	1	1
GO-ID8233	147	202	0	1	1	0.995073892	0.997142857
平均					0.997014925	0.99929627	0.999183673
					7		
	TP	TN	FN	FP	sensitivity	specificity	accuracy
GO-ID16788	67	279	0	4	1	0.985865724	0.988571429
GO-ID16798	106	240	4	0	0.963636364	1	0.988571429
GO-ID16801	0	350	0	0	Na	1	1
GO-ID16810	18	332	0	0	1	1	1

GO-ID16817	6	343	1	0	0.857142857	1	0.997142857
GO-ID16824	0	350	0	0	Na	1	1
GO-ID8233	145	200	2	3	0.986394558	0.985221675	0.985714286
平均					0.961434756	0.995869628	0.994285714

8

	TP	TN	FN	FP	sensitivity	specificity	accuracy
GO-ID16788	61	288	0	1	1	0.996539792	0.997142857
GO-ID16798	102	245	3	0	0.971428571	1	0.991428571
GO-ID16801	0	350	0	0	Na	1	1
GO-ID16810	15	335	0	0	1	1	1
GO-ID16817	3	347	0	0	1	1	1
GO-ID16824	0	349	0	1	Na	0.997142857	0.997142857
GO-ID8233	165	182	1	2	0.993975904	0.989130435	0.991428571
平均					0.993080895	0.997544726	0.996734694

9

	TP	TN	FN	FP	sensitivity	specificity	accuracy
GO-ID16788	62	285	3	0	0.953846154	1	0.991428571
GO-ID16798	87	261	2	0	0.97752809	1	0.994285714
GO-ID16801	0	350	0	0	Na	1	1
GO-ID16810	19	331	0	0	1	1	1
GO-ID16817	7	340	3	0	0.7	1	0.991428571
GO-ID16824	0	350	0	0	Na	1	1
GO-ID8233	166	176	0	8	1	0.956521739	0.977142857
平均					0.926274849	0.99378882	0.993469388

10

	TP	TN	FN	FP	sensitivity	specificity	accuracy
GO-ID16788	69	278	0	3	1	0.989323843	0.991428571
GO-ID16798	100	246	4	0	0.961538462	1	0.988571429
GO-ID16801	0	350	0	0	Na	1	1
GO-ID16810	22	327	1	0	0.956521739	1	0.997142857
GO-ID16817	2	347	1	0	0.666666667	1	0.997142857
GO-ID16824	1	349	0	0	1	1	1
GO-ID8233	148	195	2	5	0.986666667	0.975	0.98
平均					0.928565589	0.994903406	0.993469388

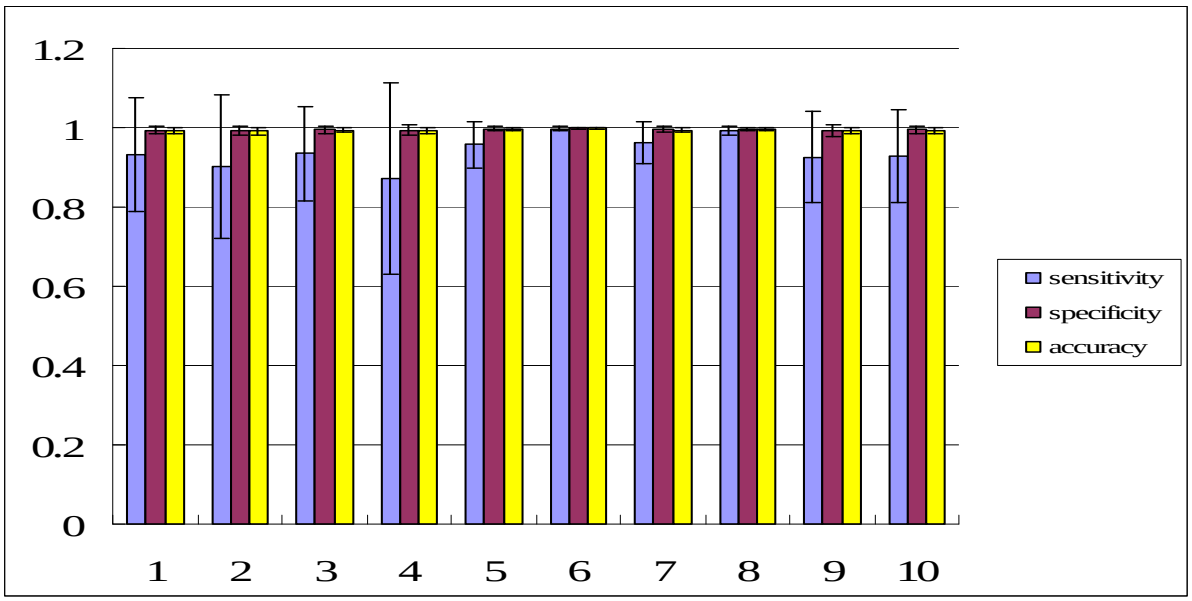


圖 4. 平均敏感度、識別度和正確率

圖 4 是整理每次實驗結果敏感度、識別度和正確率的平均及標準差，作本實驗最後總結果。而 10-fold 的總平均為：sensitivity：94.040%；specificity：99.533%；accuracy：99.397%。

討論

根據實驗結果利用記分矩陣辨認出模體(motif)判斷蛋白質功能，其準確率都在 99%以上，顯示本論文所提的方法是一可行而有效的蛋白質功能預測方法。實驗數據顯示出敏感度較差，其標準差也大，原因在於 GO ID16801、GO ID16817 與 GO ID16824 的實驗資料太少，因此用來訓練與測試的資料量都不夠，易造成較大的偏差。

四、結論及未來研究方向

結論

模體(motif)在蛋白質中重複且相似的出現，本研究利用這樣的特性建立出記分矩陣，並且利用這些記分矩陣預測蛋白質中有那些模體(motif)，並藉由模體(motif)與蛋白質功能的相關性，預測蛋白質功能，其準確率也相當高，達到預期的目標，表示利用模體(motif)預測蛋白質功能確實可行。本研究也將許多模體(motif)序列資訊轉換成記分矩陣，提供另一種方式表示蛋白質中模體(motif)，而不是傳統的序列，本研究所使用的模體(motif)是針對水解酶所作的預測，對於蛋白質而言，樣本有些不足，如果加入其他功能分類，則對於蛋白質功能分類預測，應更具代表性。另外本研究是 PRINTS 資料庫在 2005 年所建立的，至今已有許多新的蛋白質被發現，所以可能有新的模體(motif)並未收入在資料庫中，若加入這些新的模體(motif)，對於本研究的正確率應該會提升。

未來研究方向

本研究是採取 PDB 資料庫的蛋白質，而 PDB 資料中是收集蛋白質的結構相關的資料庫，收集 PDB 的蛋白質對於以後需要結構相關的資訊有很大的幫助，雖然本研究沒有用到結構相關的資訊，但是未來研究會加入結構相關的資訊，所以本研究採用 PDB 中的蛋白質，以便未來進一步的研究。

結構的資訊，在本研究中並沒有使用，未來面對蛋白質多種的功能，應該找出那一類的蛋白質功能會適合用哪一類型的預測方式，因為蛋白質的種類太多了，想要用一種方法預測出全部的蛋白質功能可能是沒有辦法的，所以在未來應該針對某類的蛋白質找出適合的結構特徵、序列特徵，來應用

於功能的預測。

本研究所對應的資料庫為PRINTS，這個資料庫是利用fingerprint的方法得到蛋白質的motif序列，在未來希望可以加入其他的資料庫，以增加功能預測的準確度，例如可以加入以模型為紀錄的資料庫PROSITE[31]，或者更進一步研發出更精準找蛋白質功能區域的方法，這樣可以更進一步提升蛋白質功能預測的準確率。

致謝

本論文之研究經費由國科會在計畫編號NSC96-2628-E-005-074-MY3 下部分補助，作者在此特別致上感謝。

五、參考文獻

[1] 蔡杰松, “利用蛋白質兩面角辨識α螺旋結構及β摺板結構”, 中興大學資訊科學系, 2006

[2] Garrett R.H. and Grisham C. M., *Biochemistry*, 2nd Ed. New York, 1999.

[3] A. Stark, S. Sunyaev and R.B. Russell, (2003) A model for statistical significance of local similarities in structure., *Journal of Molecular Biology*, **326**, 1307-1316.

[4] Binkowski TA, Freeman P and Liang J, (2004) pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins, *Nucleic Acids Research*, **32**, W555-W558.

[5] Binkowski TA, Naghibzadeh S and Liang J, (2003) CASTp: Computed Atlas of Surface Topography of proteins, *Nucleic Acids Research*, **31**, 3352-3355.

[6] Ferre F, Ausiello G, Zanzoni A and Helmer-Citterich M, (2004) SURFACE: a database of protein surface regions for functional annotation, *Nucleic Acids Research*, **32**, D240-D244.

[7] Harrison A, Pearl F, Sillitoe I, Slidel T, Mott R, Thornton JM and Orengo C, (2003) Recognising the fold of a protein structure, *Bioinformatics*, **19**, 1748-1759.

[8] Holm L and Sander C, (1993) Protein structure comparison by alignment of distance matrices, *Journal of Molecular Biology*, **233**, 123-138.

[9] Ivanisenko VA, Pintus SS, Grigorovich DA and Kolchanov NA, (2004) PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins, *Nucleic Acids Research*, **32**, W549-W554.

[10] Jambon M, Imberty A, Deleage G and Geourjon C, (2003) A new bioinformatic approach to detect common 3D sites in protein structures, *Proteins: Structure, Function, and Bioinformatics*, **52**, 137-145.

[11] Kleywegt GJ, (1999) Recognition of spatial motifs in protein structures, *Journal of Molecular Biology*, **285**, 1887-1897.

[12] Krissinel E and Henrick K, (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions, *Acta Crystallographica Section D Biological Crystallography*, **D60**, 2256-2268.

[13] Laskowski RA, Watson JD and Thornton JM, (2003) From protein structure to biochemical function? , *Journal of Structural and Functional Genomics*, **4**, 167-177.

[14] Madej T, Gibrat JF, Bryant SH, (1995) Threading a database of protein cores, *Proteins: Structure, Function, and Bioinformatics*, **23**, 356-369.

[15] McGinnis S, Madden TL, (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools, *Nucleic Acids Research*, **32**, W20–W25.

[16] Pal D and Eisenberg D, (2005) Inference of protein function from protein structure, *Structure*, **13**,1-10.

[17] Porter CT, Bartlett GJ and Thornton JM, (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data, *Nucleic Acids Research*, **32**, D129-D133.

[18] Shanahan HP, Garcia MA, Jones S and Thornton JM, (2004) Identifying DNA-binding proteins using structural motifs and the electrostatic potential, *Nucleic Acids Research*, **32**, 4732-4741.

[19] Shrager J., (2003) The fiction of function, *Bioinformatics*, **19**, 1934－1936.

[20] Soding J., (2004) Protein homology detection by HMM-HMM comparison, *Bioinformatics*, **21**,951-960.

[21] Spriggs RV, Artymiuk PJ and Willet P, (2003) Searching for patterns of amino acids in 3D protein structures, *Journal of Chemical Information and Computer Sciences*, **43**, 412-421.

[22] The Gene Ontology Consortium, (2000) Gene Ontology: tool for the unification of biology, *nature genetics*, **25**,25-29.

[23] Tsuchiya, Y., Kinoshita, K. and Nakamura, H., (2005) PreDs: a server for predicting dsDNA-binding site on protein molecular surfaces, *Bioinformatics*, **21**, 1721-1723.

[24] Wangikar PP, Tendulkar AV, Ramya S, Mali DN and Sarawagi S, (2003) Functional sites in protein families uncovered via an objective and automated graph theoretic approach, *Journal of Molecular Biology*, **326**, 955-978.

[25] W R Pearson and D J Lipman, (1988) Improved tools for biological sequence comparison, *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 2444–2448.

[26] clustalW

http://www.ebi.ac.uk/clustalw/

[27] Enzyme Commission

http://www.chem.qmul.ac.uk/iubmb/enzyme/

[28] ExPASy-UniProt Knowledgebase

http://www.expasy.org/sprot/

[29] Hidden Markov Model

http://www.soe.ucsc.edu/research/compbio/html_format_papers/hughkrogh96/node4.html

[30] PRINTS

http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/

[31] PROSITE

http://www.expasy.org/prosite/

[32] RCSB Protein Data Bank

http://www.rcsb.org/pdb/home/home.do

附錄

一、 http://oblab.cs.nchu.edu.tw/appendix_1.doc

二、 http://oblab.cs.nchu.edu.tw/appendix_2.doc

三、 http://oblab.cs.nchu.edu.tw/appendix_3.doc