一個以親緣樹差異分析為基礎的蛋白質同源檢索演算法

洪俊銘
崑山科技大學資訊管理系
助理教授
cmhung@mail.ksu.edu.tw

陳秉祥
崑山科技大學資訊管理所
研究生
bixictn@gmail.com

林運輝
崑山科技大學資訊管理所
研究生
hui731229@gmail.com

摘要

由於對蛋白質體學研究的需求，讓蛋白質序列親緣比對也成為蛋白質體學中熱門的研究之一。因為蛋白質的序列排列與其蛋白質結構存在相當程度的群聚關連，而具有群聚關連的序列親緣樹狀圖更是一般生醫學者判讀親緣結構的依據。因此本研究將提出一個方法，利用 Zhang 所提出的樹的編修距離演算法，將 ExPASy 蛋白質資料網站利用當中的 UniProt 資料庫所提供的蛋白質序列資料建立一個親緣樹，以演化的觀念配合樹的編修距離進行蛋白質序列比對。這個以親緣樹差異分析為基礎的蛋白質同源檢索演算法可應用在搜尋同一親緣或組合衍生的品種群聚相似性上，如早熟同型係發生率的判定。藉由此方法，使蛋白質序列比對更具有演化上的意義，來彌補單純利用序列相似度進行比對的不足。

關鍵詞：蛋白質序列、樹的編修距離、親緣樹。

Abstract

Phylogenetic analysis is one of popular topics in proteomics to the demand for study. Because of clustering correlation between protein sequence and structure, undoubtedly biomedical scientist discriminates genetic structure depending on the phylogenetic tree with the correlation between sequences. Therefore, this study designed a method to use the tree match algorithm with edit distance proposed by Zhang to compare phylogenetic trees. The phylogenetic trees built from the UniProt protein database in ExPASy were conducted the pairwise comparisons with evolutionary concept. The homology search algorithm based on discrimination of phylogenetic tree was applied to find the similar clusters of species of descended from a common ancestor or derivative combination, such as discrimination of homology of early mutants. Protein sequence alignment by this way exhibits more evolutionary meanings than traditional alignment by the other methods of similarity match.

Keywords: protein sequence、tree editing distance、Phylogenetic Tree

1. 前言

近年來，XML(eXtensible Markup Language，XML)[6]成為最熱門的網路應用之一，隨著網路的蓬勃發展數以萬計的網站如雨後春筍一般，而企業間需要廣泛交換跨平台之間的電子資料及 web service 的盛行，使得 XML 成為企業間電子資料交換的主流工具。再加上生物資訊研究的興起，綜合以上，無論是網頁、XML、生物資訊研究裡的 RNA 二級結構皆可視為有序的標籤含根樹。由於資料量的增加十分快速，使樹的編修距離的需求也日漸增多。

由於人類基因圖譜計劃的完成，生物學家對於功能基因體了解的必要性，造就了蛋白質體學研究的需求。1994 年澳洲生物學家 Wilkins 首次提出蛋白質體學(proteomics)[8]這個名詞，定義為“一種基因組所表達的全部蛋白質“[7]，使蛋白質體學研究開始受到廣泛的注意及重視。其中對於蛋白質序列比對也是蛋白質體學中熱門的研究之一，因為蛋白質的序列與其蛋白質結構存在相當程度的關連，想要了解其功能機制，蛋白質序列比對將是不可避免的步驟，一般常見的做法是使用序列比對軟體 BLAST 來進行蛋白質序列比對[1][4]，比對的分數配合得分矩陣(substitution matrix)，越高代表越相似，但是僅僅單純利用序列之間的相似度進行比對，對於生物學家了解蛋白質的結構和功能性可能會有一定程度的誤差。

因此，本研究將以分子生物學的角度提出一個方法，以演化的觀念配合樹的編修距離進行蛋白質序列比對，使用由瑞士生物資訊學研究所(Swiss Institute of Bioinformatics) 所建

立的蛋白質資料庫網站－ExPASy 並利用其中的 UniProt 資料庫所提供的蛋白質序列資料(ftp://ftp.uniprot.org/pub/databases/uniprot/knowledgebase/)做為本研究的資料集建立一個親緣樹(Phylogenetic Tree)，讓生物研究學者能藉由此方法，讓蛋白質序列比對具有演化的意義來彌補單純利用序列相似度進行比對的不足。

以下各章節內容摘要分述如下。第二節對於蛋白質資料作相關的文獻探討。第三節對於親緣樹作相關的文獻探討。第四節對於樹的編修距離演算法作相關的文獻探討。第五節描述本研究的方法與進行步驟。第六節對本研究之實驗結果。第七節為本研究之結論。

2.蛋白質資料

ExPASy(http://www.expasy.org)是由瑞士生物資訊學研究所(Swiss Institute of Bioinformatics) 所建立的蛋白質資料庫網站，位於瑞士日內瓦大學，所強調的服務在於與蛋白質有關的資料，是蛋白質資料庫重要的資源網站。本研究的資料來源將利用 ExPASy 當中的 UniProt(Universal Protein Resource，UniProt)資料庫所提供的蛋白質序列資料，版本為 UniProt Knowledgebase Release 14.5，做為本研究的資料集。

UniProt(Universal Protein Resource，UniProt)是一個蛋白質資料庫，包含 SWISS-PROT 與 PIR(Protein Information Resource，PIR)兩個資料庫。現有的 UniProt 包括兩個部份：分別為 Swiss-Prot 跟 TrEMBL[3]，是現今蛋白質資料庫中最大且最完整的，其資料格式如圖 1。Swiss-Prot 是經過確認的資料庫，有完整的註解資料，並且除去重複資料。並且已跟其他 50 多個資料庫整合，例如核酸資料庫等。TrEMBL 則是電腦自動轉譯並且註解的蛋白質資料庫，這是為了快速整理目前發表十分快速的蛋白質序列資料。UniProt 是一個生物資訊研究相當重要的資料庫。[2]

```
AC   Q43495;
DT   15-JUL-1999, integrated into UniProtKB/Swiss-Prot.
DT   01-NOV-1996, sequence version 1.
DT   04-NOV-2008, entry version 44.
DE   RecName: Full=Protein 108;
DE   Flags: Precursor;
OS   Solanum lycopersicum (Tomato) (Lycopersicon esculentum).
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
OC   asterids; lamiids; Solanales; Solanaceae; Solanoideae; Solaneae;
OC   Solanum; Lycopersicon.
OX   NCBI_TaxID=4081;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA].
RC   STRAIN=cv. VF36; TISSUE=Anther;
RX   MEDLINE=94143497; PubMed=8310077; DOI=10.1104/pp.101.4.1413;
RA   Chen R., Smith A.G.;
RT   "Nucleotide sequence of a stamen- and tapetum-specific gene from
RT   Lycopersicon esculentum.";
RL   Plant Physiol. 101:1413-1413(1993).
CC   -!- SUBCELLULAR LOCATION: Secreted (Potential).
CC   -!- TISSUE SPECIFICITY: Stamen- and tapetum-specific.
CC   -!- SIMILARITY: Belongs to the A9/FIL1 family.
CC   ---------------------------------------------------------------------
CC   Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms
CC   Distributed under the Creative Commons Attribution-Noderivs License
CC   ---------------------------------------------------------------------
DR   EMBL; Z14059; CAA78466.1; -; mRNA.
DR   PIR; S26409; S26409.
DR   UniGene; Les.3896; -.
DR   GO; GO:0005576; C:extracellular region; IEA:UniProtKB-KW.
DR   InterPro; IPR003612; LTP/seed_store/tryp_amyl_inhib.
DR   Pfam; PF00234; Tryp_alpha_amyl; 1.
DR   SMART; SM00499; AAI; 1.
PE   2: Evidence at transcript level;
KW   Secreted; Signal.
FT   SIGNAL       1     30       Potential.
FT   CHAIN       31    102       Protein 108.
FT                               /FTId=PRO_0000000238.
FT   DISULFID     41     77       By similarity.
FT   DISULFID     51     66       By similarity.
FT   DISULFID     67     92       By similarity.
FT   DISULFID     79     99       By similarity.
SQ   SEQUENCE   102 AA;  10576 MW;  CFBAA1231C3A5E92 CRC64;
     RASVKSSSSS SSSSFISLLL LILLVFVLQS QVIECQPQQS CTASLTGLNV CAPFLVPGSP
     TASTECCNAV QSINHDCMCN TMRIAAQIPA QCNLPPLSCS AN
//
```

圖 1 UniProt 資料格式

3.親緣樹

親緣樹(Phylogenetic Tree)是一種對各物種間演化關係進行分類的樹狀結構。其中包含節點（Nodes）與邊（Branches）兩大部份，節點代表序列，邊代表演化距離。親緣樹依表示方式可分為有樹根（Rooted）與無樹根（Unrooted）兩種。本研究所使用的是有樹根的親緣樹，以下只針對有樹根的建構方式做介紹。

有樹根的親緣樹是以階層式的架構來表示節點之間的演化關係，因此可看出節點的共同祖先。有根樹的建構方式為階層群聚法(hierarchical clustering)，先建立物種之間的距離矩陣，從距離矩陣中找出距離最低也最相似的物種合併為一個集合，每次合併二個最相似的物種，直到樹建立完成。距離的計算方式有以下三種方法：[4]

$$d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} |p - p'| \quad (1)$$

$$d_{max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} |p - p'| \quad (2)$$

$$d_{min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} |p - p'| \quad (3)$$

計算式(1)中，C_i 與 C_j 分別代表第 i 與 j 群，n_i 與 n_j 為第 i 與 j 群內資料的數量，P 與 P′為第 i 與 j 群內的代表點，它是計算兩群聚內所有資料點的距離平均數來表示兩群聚的距離。而在計算式(2)中，C_i 與 C_j 分別代表第 i 與 j 群，P 與 P′為第 i 與 j 群內的代表點，其評估方式是以兩群聚中最遠的兩資料點 P 與 P´ 的距離來代表群聚的距離。計算式(3)中，以 C_i 與 C_j 代表第 i 與 j 群，P 與 P′為第 i 與 j 群內的代表點，其計算方式是以兩群聚間最近的兩資料點 P 與 P′來代表群聚距離。

4.樹的編修距離演算法

4.1 樹狀結構之編修距離問題

此問題由學者 Tai 在 1979 年所提出，也稱作為 Tree-to-Tree Correction Problem[7]，但後面的介紹還是以 1989 年學者 Zhang 所提出的版本[11]為主。

4.1.1 基本操作

給定兩顆 S 與 T 的有序標籤含根樹和一個對 S、T 的操作集合，欲求在有限的操作之下，將 S 轉換成 T 的最小成本(即最短距離)。基本上，操作集合包含三個操作，交換兩節點(Change)，刪除一節點(Delete)、插入一節點(Inser

t)。以下這三個操作的意義將用圖來說明：

交換兩節點(Change)

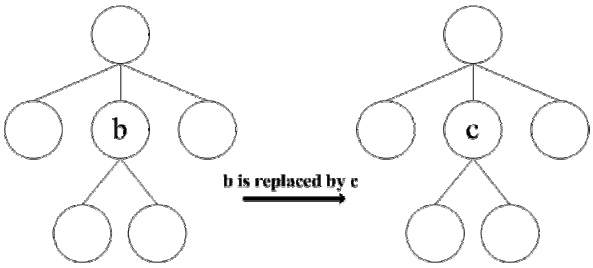

圖 2 交換兩節點

以 b→c，表示將節點 b 換成節點 c。

刪除一節點(Delete)

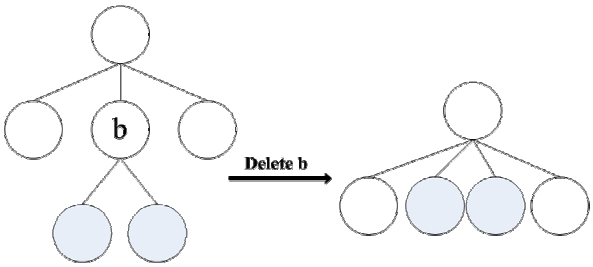

圖 3 刪除一節點

以 b→Λ，表示刪去 b 節點。Λ表示空節點(empty node)。

插入一節點(Insert)

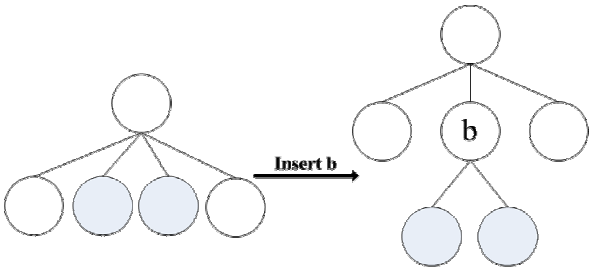

圖 4 插入一節點

以Λ→b，表示插入 b 節點。

以上三個操作中，每一個操作都有成本，分別以成本函式 Cost(b→Λ)、Cost(Λ→b)、Cost(b→c)，代表刪除 b 節點的成本、插入 b 節點的成本、交換 b 節點為 c 節點的成本。成本為一非負的實數。成本函式基本上為使用者定義。

因此，我們要將樹 S 轉換成樹 T，需要經過 $e_1e_2e_3e_4\cdots e_n$ 個操作，每個 e_i (i=1,2,3,…)為一次操作。$E=e_1e_2e_3e_4\cdots e_n$ 為一連續的操作序列。所以，將 S 經由 E 轉換成 T 總共需要下述成本。以下公式便可求將 S 經由 E 轉換成 T 所需要的成本。

$$\sum \mathrm{Cost(E)} = \sum_{1}^{n} \mathrm{Cost}(e_i) = \mathrm{Cost}(e_1) + \mathrm{Cost}(e_2) + \ldots + \mathrm{Cost}(e_n)$$

4.2 樹的編修配對(Tree Editing Mapping)

給定兩顆有序標籤含根樹 S 與 T，為了求出 S 跟 T 之間如何轉換，S 經過了多少操作之後轉換成 T，我們在 S 跟 T 之間建立了一個配對的關係。利用下圖來說明：

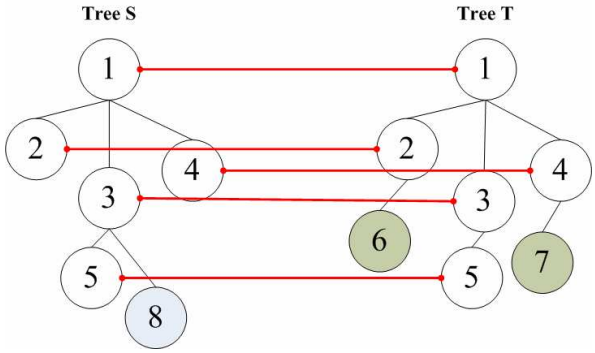

圖 5 Tree S 與 Tree T 之間的編修配對示意圖

在兩顆樹 S 跟 T 之間建立一個配對，我們稱之為 S 跟 T 的編修配對，以(M,S,T)表示。

在上述的編修配對中 Tree S 裡的節點 8 不在(M,S,T)中，表示被刪除的節點；Tree T 裡的節點 6 跟節點 7 不在(M,S,T)中，表示被插入的節點。因此根據所得的編修配對(M,S,T)來計算，將 S 轉換成 T 所需要的成本。令 I 為第一類的節點集合。J 為第三類的節點集合。第二類的點在(M,S,T)中。γ 表示節點操作的成本函式。

(M,S,T)的成本便可用下面公式來表示。

$$\mathrm{Cost((M,S,T))} = \sum_{i\in I} \gamma(s(i) \to \Lambda) + \sum_{s(i),t(j)\in(M,S,T)} \gamma(s(i) \to t(j)) + \sum_{j\in J} \gamma(\Lambda \to t(j))$$

5.研究方法與進行步驟

蛋白質序列比對是蛋白質體中熱門的研究之一，常見的做法都是利用序列之間的相似度來進行比對，但是對於想要了解其蛋白質結構與功能性的生物學家來說，可能會有某種程度的誤差。因此，本研究將嘗試以演化的特性配合樹的編修距離進行蛋白質序列比對，以期能彌補單純利用序列相似度進行比對的不足，下圖為本研究所提出的研究架構，將針對各部分進去說明。

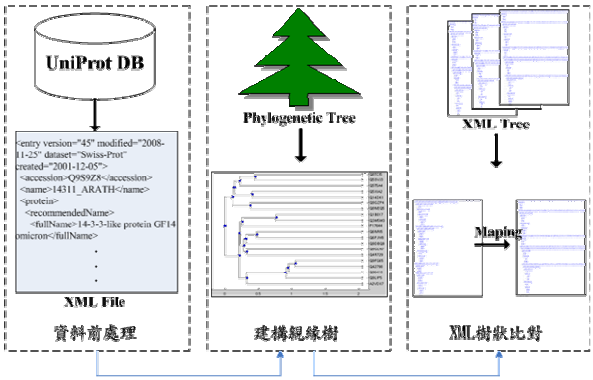

圖 6 研究架構

5.1 資料前處理

首先，將從 UniProt 資料庫裡取得的蛋白質資料總共 402,481 筆進行分類，本研究所使用的序列取自靈長類(Homo sapiens)蛋白質序列為共 25195 筆。分類完成後，接著進行資料前處理，並擷取研究所需的部份，如存取編號(Accession number)、蛋白質序列(Protein Sequence)，如圖 7，以便之後蛋白質序列親緣樹的建構。

```
60:Accession number:O57211 Sequnct:MPQQLSPINIETKKAISI
61:Accession number:B1IRD7 Sequnct:MDFNLNDEQELFVAG
62:Accession number:P59394 Sequnct:MDHLPMPKFGPLAGLR
63:Accession number:Q5PIL1 Sequnct:MSESLHLTRNGPILEITL
64:Accession number:A7ZVZ0 Sequnct:MKNEKRKTGIEPKVFI
65:Accession number:Q3ZBY3 Sequnct:MAGPQQQPPYLHLAE
```

圖 7 擷取後的存取編號與蛋白質序列

5.2 建構親緣樹

本研究將利用 MATLAB R2007b bioinformatics 工具箱，使用 UPGMA(Unweighted Pair-Group Method Using Arithmetic averages)群集分析來建構蛋白質序列親緣樹，但因總資料筆數過於龐大，使繪製親緣樹不宜，因此本研究將擷取部分序列為一親緣樹，並將其轉存成 XML 檔案格式，如圖 8，作為後面實驗測試比對所用。

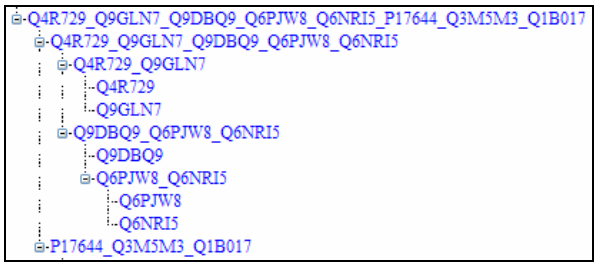

圖 8 親緣樹轉成 XML 檔案局部畫面

5.3 XML 樹狀比對

最後，本研究將利用 1989 年學者 Zhang 所提出的演算法進行實作，做為蛋白質序列親緣樹的 XML 樹狀比對，但是樹的編修比對常見的做法是應用在 XML 與 RNA 二級結構上，因此將蛋白質序列編碼成樹狀結構有其困難度，而親緣樹在各個節點上並沒有節點名稱，無法使用演算法進行樹狀比對，必須將其轉換成 XML 格式，因此我們將使用蛋白質序列的存取編號(Accession number)串接起來並以底線分隔開來，做為 XML 標籤名稱，以便進行蛋白質親緣樹的親緣結構比對，從所需的編修距離成本當中了解其結構的相似程度為何。下圖為二個蛋白質序列親緣樹的樹狀比對結果示意圖。

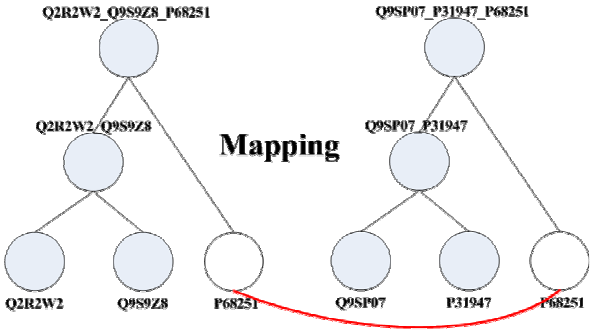

圖 9 樹狀比對結果示意圖

從上圖的樹狀比對結果來看，有紅線連接的節點為成功配對，其餘四個藍色節點無配對成功，必須進行編修操作，因為各節點結構階層相同，只需做交換(Change)的編修操作，因此這二棵樹的編修距離成本為 4。

6. 實驗結果

本研究的建置實驗環境為 Acer Veriton 7700GX 一台，其硬體規格為 CPU P4 650 3.4G、記憶體 DDR2 1G。軟體為 MATLAB R2007b、JDK 1.6。

綜合上述的研究方法與進行步驟，我們將對由 3、5、10、20、50、100、200 條所繪製的蛋白質序列親緣樹進行兩兩樹狀比對，各集合共做 30 次的樹狀比對。從實驗中推估親緣樹的結構相似度為多少。如果兩個群集的蛋白質序列集合完全相同，形成完全相同的親緣樹結構，所以任何的編修成本皆為 0，並稱為「完全重疊取樣」。我們實驗設計是以蛋白質序列取樣完全不重疊做為測試「互斥取樣」編修距離成本的依據，而以一半重疊的取樣做為測試「半重疊取樣」編修距離成本的依據。

以下為本實驗的結果圖表，將針對其實驗結果的建構親緣樹的執行時間、樹狀比對執行時間和編修距離成本做詳細說明。

表 1 建構親緣樹平均時間

序列組成(條)	建構親緣樹平均時間(秒)
3	0.2221
5	0.7192
10	2.3494
20	11.4140
50	91.3920
100	353.1769
200	1401.1776

根據表 1 的實驗結果來看，本研究是利用 MATLAB® R2007b bioinformatics 工具箱中所提供的方法來建構親緣樹，發現其建構親緣樹的平均執行時間會成倍數成長，下圖為建構親緣樹平均執行時間的走向趨勢圖，x 軸為序列組合，y 軸為平均執行時間(秒)。

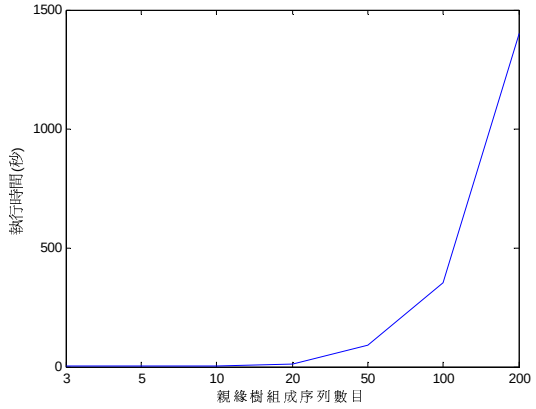

圖 10 建構親緣樹平均執行時間趨勢圖

表 2 比對親緣樹平均執行時間

組成序列(條)	互斥取樣編修時間(秒)	半重疊取樣編修時間(秒)	完全重疊取樣編修時間(秒)
3	0.0153	0.0166	0.0140
5	0.0361	0.0393	0.0396
10	0.1595	0.1740	0.1695
20	1.0427	0.9803	0.9568
50	10.7014	10.5447	10.7202
100	91.7522	94.4859	91.1506
200	1019.2070	1005.4810	1018.4068

根據表 2 的實驗結果來看，發現親緣樹比對的平均執行時間，無論是「互斥取樣」、「半重疊取樣」與「完全重疊取樣」上，其所耗費的時間皆沒有太大的差異，但是比對的執行時間都會隨著序列組合的增加而呈現倍數的成長。

表 3 編修距離成本實驗結果

組成序列(條)	互斥取樣距離成本	半重疊取樣距離成本
3	5	5
5	10	9
10	23	21
20	47	46
50	122	115
100	244	237
200	488	481

經過實際的編修距離比對實驗，比較「互斥取樣」與「半重疊取樣」的編修距離成本，根據表 3 的實驗結果發現，「互斥取樣」和「半重疊取樣」的編修距離成本相差不遠。

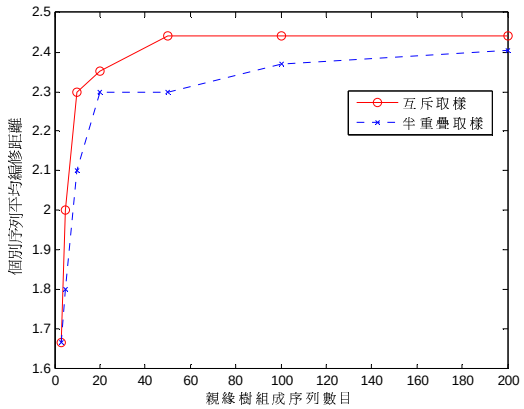

圖 11 親緣樹比對平均邊界距離比值圖

但是如圖 11 所示，比較親緣樹內個別序列平均編修距離發現，小於 100 條序列以內的親緣樹比較能區分出「互斥取樣」與「半重疊取樣」兩種不同組合在結構上的差異。

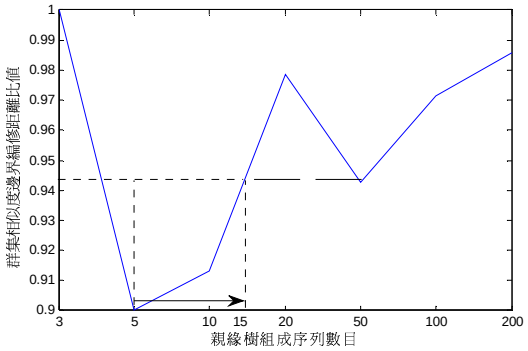

圖 12 群集相似度邊界編修距離比值圖

另外，如圖 12 所示，計算平均邊界距離比值(=50%不同序列距離/100%不同序列距離)之後的結果顯示，可推估一個親緣樹所組成的序列數目以 5 至 15 條最能表現此演算法區分蛋白質序列親緣樹的演化結構的差異。包含 5 條序列以下的親緣樹沒有區分結構差異的意義，而 15 條以上相似度的差異比較逐漸縮小(比值 0.95~1 趨近於 1)。也就是說，親緣樹包含 15 條以上「互斥取樣」與「半重疊取樣」的結構比較，沒有統計上的差異。以上的實驗結果可做為生物學者在蛋白質序列比對上的輔助參考資訊。藉由此方法，讓蛋白質序列在比對上具有演化的意義。

7. 結論

近年來，由於生物學家對於功能基因體了解的必要性，造就了蛋白質體學研究的需求，也因為蛋白質序列與其蛋白質結構存在相當程度的關連。因此，本研究嘗試以演化的特性配合樹的編修距離進行蛋白質序列比對，從實驗結果發現推估一個親緣樹所組成的序列數目以 5 至 15 條最能表現此演算法區分蛋白質序列親緣樹的演化結構的差異。希望能藉由本研究的方法，讓蛋白質序列在比對上具有演化的意義，可做為生物學者在蛋白質序列比對上的輔助參考資訊。在未來，由於建製和比對親緣樹需要耗費相當的執行時間，因此在往後的研究我們會將重點放在分散式處理的解決方案。

參考文獻

[1] Altschul, S.F., Gish, W., Miller. et al.. “Basic local alignment search tool,” ***J. Mol. Biol***, 215,403-410, 1990.

[2] Apweiler,R., Bairoch,A., Wu,C.H. et al. “UniProt: the Universal Protein knowledgebase,” ***Nucleic Acids Res,*** 32,D115–D119, 2004.

[3] Boeckmann B., Bairoch,A., Apweiler,R. et al. “The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003,” ***Nucleic Acids Res,*** 31,365–370, 2003.

[4] Han J. and Kamber M. “Data Mining: Concepts and Techniques,” ***Morgan Kaufmann***, 2000.

[5] Karlin, Samuel and Stephen F. Altschul. “Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes,” ***Proc. Natl. Acad. Sci,*** USA 87:2264-68, 1990.

[6] O’Farrell, P. H. “High resolution two-dimensional electrophoresis of proteins,” ***J. Biol. Chem,*** 250,4007-4021, 1975.

[7] Kuo-Chung Tai. “The tree-to-tree correction problem,” ***Journal of the Association for Computing Machinery (JACM),*** 26:422-433, 1979.

[8] Tim Bray, Jean Paoli, C. M. Sperberg-McQueen. et al. “Extensible Markup Language (XML) 1.0 (Fourth Edition)-Origin and Goals,” ***World Wide Web Consortium,*** 2006.

[9] Wasinger V C, Cordwell S J, Cerpa-Poljak C. et al. “Progress with geneproduct mapping of the Mollicutes:Mycoplasma genitalium,” ***Electrophoresis,*** 16,1090-1094, 1995.

[10] Wilkins M.R., Pasquali, C., Appel, R.D. et al.”From Proteins to Proteomes: Large scale protein identification by two-dimensional electrophoresis and amino acid analysis,” ***Bio/Technology,*** 14,61-65, 1996.

[11] K. Zhang and D. Shasha. “Simple fast algorithms for the editing distance between trees and related problems,” ***SIAM Journal of Computing,*** 18(6):1245-1262, 1989.