行政院國家科學委員會專題研究計畫　成果報告

複合式蛋白質序列分群演算法之研究

計畫類別：個別型計畫
計畫編號：NSC93-2218-E-002-149-
執行期間：93 年 10 月 01 日至 94 年 09 月 30 日
執行單位：國立臺灣大學生物產業機電工程學系暨研究所

計畫主持人：陳倩瑜

計畫參與人員：陳冠豪、劉育廷、鐘文欽、張天豪

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 94 年 12 月 25 日

行政院國家科學委員會補助專題研究計畫 ☑ 成 果 報 告 □期中進度報告

複合式蛋白質序列分群演算法之研究

計畫類別：☑ 個別型計畫　　□ 整合型計畫
計畫編號：NSC 93－2218－E－002－149
執行期間：　　93 年　10 月　1 日至　　94 年　9 月　30 日

計畫主持人：陳倩瑜
計畫參與人員：陳冠豪、劉育廷、鐘文欽、張天豪

成果報告類型(依經費核定清單規定繳交)：☑精簡報告　□完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
☑出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：國立臺灣大學生物產業機電工程學系暨研究所

中　華　民　國　94　年　12　月　25　日

計畫中文摘要（五百字以內）

本計畫將針對複合式蛋白質分群演算法進行研究，目的在於開發以統計模型為基礎的階層式分群演算法，使其產生之蛋白質階層在不同高度具有不同的特性。在本主持人最近的研究成果中，成功地將統計模型與階層式分群演算法結合，所開發之蛋白質分群演算法不僅比傳統階層式分群演算法享有較低的複雜度，同時利用統計模型簡化原有二元階層的節點數，提供生物學家更為精確與實用之分群資訊。

階層式蛋白質分群演算法目前仍有一些困難需要克服始能提高蛋白質階層之正確率。本計畫將延續本主持人之前之研究，首先針對傳統階層式分群演算法不容許同一蛋白質分屬階層之不同位置之特性加以改良，進一步則必須針對蛋白質家族的特性，設計適當的控制機制，使所得之蛋白質階層在不同高度，分別滿足不同大小之蛋白質家族的需求。

本計畫為一年期之計畫，前半年將集中在研究如何利用特定蛋白質與其他蛋白質之相似度分布曲線，辨識需要在階層底層被複製的蛋白質集合。後半年則研究如何結合已開發之統計檢測，在各個分群階段使用不同的分群準則，以達其研究目的。

計畫英文摘要 （五百字以內）

The objective of this project is to study the hybrid hierarchical protein sequence clustering algorithms. The project aims to provide biologists a protein hierarchy that matches different sizes of proteins in the different levels of the hierarchy. In the recent study, we have successfully employed the statistical models to improve the efficiency of the traditional hierarchical clustering algorithms for protein family analysis. The proposed statistical model based algorithm also provides users a summarized hierarchy that the size of which is much smaller than the original binary tree generated by the traditional hierarchical clustering algorithms.

There are still some challenges for protein sequence clustering. In this project, we will continue our recent study to design a hybrid hierarchical clustering algorithm based on statistical models. In order to satisfy the demand of protein family analysis, the first problem we need to tackle is some multi-function proteins should be placed at more than one position in the protein hierarchy. Next, different sizes of protein families possess different properties. Smaller families ask for the property of homogeneity, while the larger families need to utilize the property of transitivity in order to find remote homology. The hierarchical clustering algorithm should hybridize different criterions for controlling the formation of new clusters.

The duration of this project is one year. In the first half of the year, we plan to recognize the proteins that should be duplicated in the bottom level of the hierarchy by examining the distribution of the similarities between a particular protein and all of the other proteins. In the remaining half of the year, different controlling criterions are designed and used in the different stages of clustering process to generate the hierarchy that matches the protein families better.

報告內容

蛋白質序列分群需要階層式的分群演算法以建構蛋白質家族的相關性分析，但由於現今最為常用的單一連結階層式分群演算法(Single-linkage algorithm)容易受到雜訊的影響，而分群品質較好的平均連結分群演算法[21](Average-linkage algorithm)卻因其較高的時間與空間複雜度而面臨無法處理大量資料的窘境。因此在此研究中，我們利用先前開發的統計檢定模型，幫助階層式分群演算法判斷某些並不適合在演算法初期使用的關聯性，在演算法的第一階段只允許同質性蛋白質群(Homogeneous clusters)的形成，第二階段才積極利用蛋白質序列的同源遞移性(Transitivity property of homology)來尋找遠親同源關係，建構非同質性的蛋白質家族。此演算法的貢獻，不僅提供更為準確的蛋白質階層，更使得小規模的實驗室也能利用一般個人電腦，對十幾萬條蛋白質序列進行分群研究，建構分群階層。分群過程僅需數百萬位元組(MB)的記憶體使用量，並可在數小時內完成分群動作；此等運算規模在未來不僅可以加速更優良的分群演算法之開發，也奠定了本實驗室利用分群結果進行下一階段序列資料庫分析的基礎。

過去大部分的蛋白質分群演算法仍採用序列比對[1, 18, 22]所得之分數作為分群過程的重要依據。也就是說，序列相似度高的兩個蛋白質，在演算法的執行過程中將會盡可能被放在同一個群集中，因為它們有非常大的可能性會有相同的三級結構或是功能性[7, 10, 16]。一般認為，蛋白質間的序列相似度在高數值的情形下的確可以代表蛋白質之間的高度相關性，但低序列相似度的兩個蛋白質卻並能不代表彼此完全不相關。這個在相似度分布上的模糊地帶被稱為 Twilight zone of homology [20]，Rost 在此篇論文中提到，當序列相似度超過 40%的時候，相似度的數值將可以明確地反應出這個兩個蛋白質的相似性(同源性)；但是，當序列相似度介於 20%~35%的時候，相似度的高低將無法直接作為相似度的依據。更有實驗數據指出，只有 15%序列相似度的兩個蛋白質也會產生相同三級結構，並具有相同的功能。由於同源(或同功能)的蛋白質有可能只有微弱的序列相似度，許多分群演算法必須透過同源的遞移特性(transitivity property of homology)來尋找具有遠端同源性的蛋白質群集。所謂的同源遞移性指的是對於相似度低的兩蛋白質 A、B，可以透過一個蛋白質 C，其與蛋白質 A、B 分別存在著足以代表同源關係的高相似度，來推論蛋白質 A 與 B 亦存在著同源關係[4]。

現今所使用的蛋白質序列分群演算法都盡量善用同源遞移性來尋找具有同源性的蛋白質群集。傳統的單一連結階層式分群演算法(Single-linkage algorithm) [11]即為一廣為使用的生物序列分群演算法[6, 9, 13, 15, 23]。此演算法不僅善用其鏈結效應尋找遠端同源之蛋白質，它所產生的階層式分群結果對蛋白質資料庫的各種分析亦提供很大的幫助。ClusTr 資料庫[15]即提供網頁服務讓使用者查詢他們使用單一連結階層式分群演算法對 Swiss-Prot 資料庫[3]中所有蛋白質分群的結果。另一類被廣為使用的序列分群演算法則以圖論(Graph theory)為基礎，此類演算法將一個蛋白質序列視為一個單一節點(vertex)，兩個蛋白質之間的相似關係則以邊(edge)、有向邊(directed edge)或是加以權重的邊(weighted edge)來描述。以圖論為基礎的分群演算法經過設計之後可以善加利用同源遞移性，因此這幾年在蛋白質序列分群的問題上被廣為研究[4, 8, 14, 17, 24]。

多數分群演算法在利用同源遞移性尋找遠端同源蛋白質的時候都會面臨到同源模糊地帶所造成的問題。以單一連結分群演算法為例，此演算法雖然利用鏈結效應可以尋找到遠端

同源性之蛋白質，但是也可能由於鏈結效應的不當利用，導致所產生的群集不能維持良好的明確性。鏈結效應之所以會造成問題，是由於分布於模糊地帶的相似度值有時候並不能代表真正的同源關係(true homologues)，分群演算法試圖利用同源遞移性尋找遠端同源蛋白質的同時，也可能將不同源的蛋白質囊括進來。此外，當蛋白質資料庫快速成長時，序列比對的樣本增加，相對地雜訊程度(noise level)也增加，隨機比對所得的相似度也可能比真實相關之蛋白質間的相似度還高[19]。

我們延續之前於簡化階層結構方法的研究，針對蛋白質序列特性提出利用統計模型進行階層結構簡化的方法[6]。此簡化階層結構方法可應用於傳統階層式分群演算法所產生的二元數狀結構，輸出較為精簡的分群階層；此方法不僅提供使用者更為明確的分群資訊，簡化後的群集資訊亦有利於遞增性分群演算法快速地為新進的蛋白質序列加以分群。我們之前利用這個簡化階層結構方法有效地將 Swiss-Prot 資料庫中所有蛋白質加以分群，並輸出有利於進行蛋白質家族分析的階層架構，而我們所提出的複合式分群演算法(此論文已被國際期刊 Pattern Recognition 接受[5])，即以此簡化階層中最重要的資訊–同質性同源蛋白質群集(homogeneous homology protein clusters) –來做為新的改善方法的基礎。

於此計畫中，我們所提的複合式分群演算法即利用單一連結分群演算法搭配兩個統計指標來進行初步的分群工作。偏差度及尖銳度[12]是統計學上常用的評量方法，偏差度是用來評量一群資料分布曲線的對稱程度，或是不對稱程度；尖銳度則是比較某一分布曲線是否比常態分布更為集中。對於群集 C，S_C 代表群集 C 中所有的的相似度集合，故$|S_C| = |C|\times(|C|-1)/2$。群集 C 的偏差度 $Skew(C)$及尖銳度 $Kurt(C)$的定義如下：

$$Skew(C) = \frac{|S_C|}{(|S_C|-1)\times(|S_C|-2)} \sum_{x_i \in S_C} (x_i - \bar{x})^3 / s^3$$

$$Kurt(C) = \left[\frac{|S_C|\times(|S_C|+1)}{(|S_C|-1)\times(|S_C|-2)\times(|S_C|-3)} \sum_{x_i \in S_C} (x_i - \bar{x})^4 / s^4 \right] - 3\frac{(|S_C|-1)^2}{(|S_C|-2)\times(|S_C|-3)},$$

其中 $\bar{x} = \frac{1}{|S_C|}\sum_{x_i \in S_C} x_i$ ，而 $s = \sqrt{\frac{1}{|S_C|-1}\sum_{x_i \in S_C}(x_i - \bar{x})^2}$ 。

我們利用這兩個統計指標來決定新群集的形成與否。同一家族內的蛋白質由於功能或特性相似，我們稱它們之間具有同質性。根據我們之前在論文[6]中的定義，我們定義一同質性群集(homogeneous cluster)為滿足下列條件的群集：

$$-\theta_s \le Skew(C) \le \theta_s \text{, and}$$

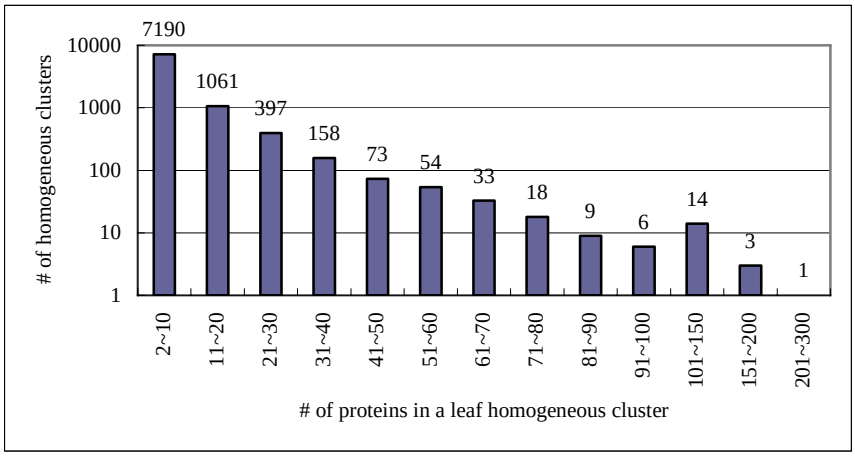
$$Kurt(C) \ge \theta_k,$$

其中，使用者可以自行決定 θ_s 和 θ_k 兩個參數的設定值。我們在過去的研究中有詳盡敘述這兩個參數的設定值對於分群結果整體效應的影響[6]。

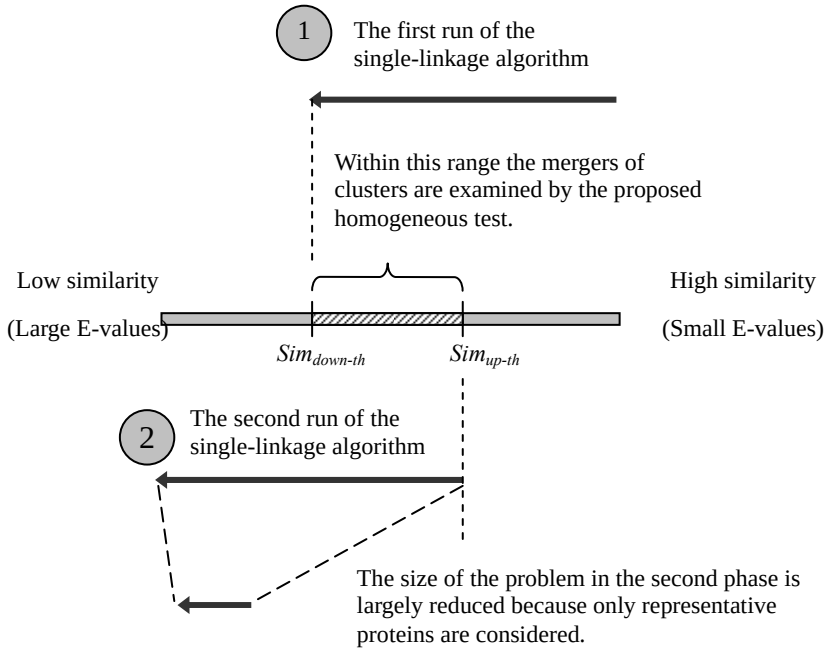
而當我們第一次執行單一連結分群演算法時，我們要求所有新形成的群集都要滿足同質性群集的條件，如此一來，當此一步驟結束的時候，我們將擁有大小不等的同質性群集。從論文[6]中的實驗數據得知，對於蛋白質家族而言，同質性群集的一致性相當高。如圖一所示，針對 Swiss-Prot 資料庫所有蛋白質(Release 41.0，2003 年，122564 個蛋白質)的分群結果，我們找到 9017 個同質性同源蛋白質群集，這些群集的大小不等，大者有兩百多個蛋

白質，不過大多數的同質性群集都集中在二十個蛋白質以下。經過與 InterPro 資料庫[2]的比對，這些蛋白質群集的同質性的確分常高，每一個群集所包含的蛋白質幾乎都具有相同的功能。

為了有效避免分群演算法受到雜訊的干擾，如圖二所示，我們搭配 2004 所提出的同質性統計檢測[6]，強迫不具同質性的蛋白質群集延後其產生的時間，如此一來，可以在分群過程的第一階段只產生同質性的同源蛋白質群集；如圖三所示，蛋白質階層的底層多數將為同質性同源蛋白質群集，這些同質性同源蛋白質群集未來可以作為序列探勘演算法的訓練資料，建構每一種功能的序列特徵。我們所提出的複合式分群演算法第二階段才積極利用蛋白質序列的同源遞移性(transitivity of homology)來尋找遠親同源關係，建構較為上層的蛋白質階層。在第二階段當中只要避免使用群集中特異的蛋白質序列作為代表序列，即可大幅度降低多重區域蛋白質或是多功能性蛋白質造成不良的影響。

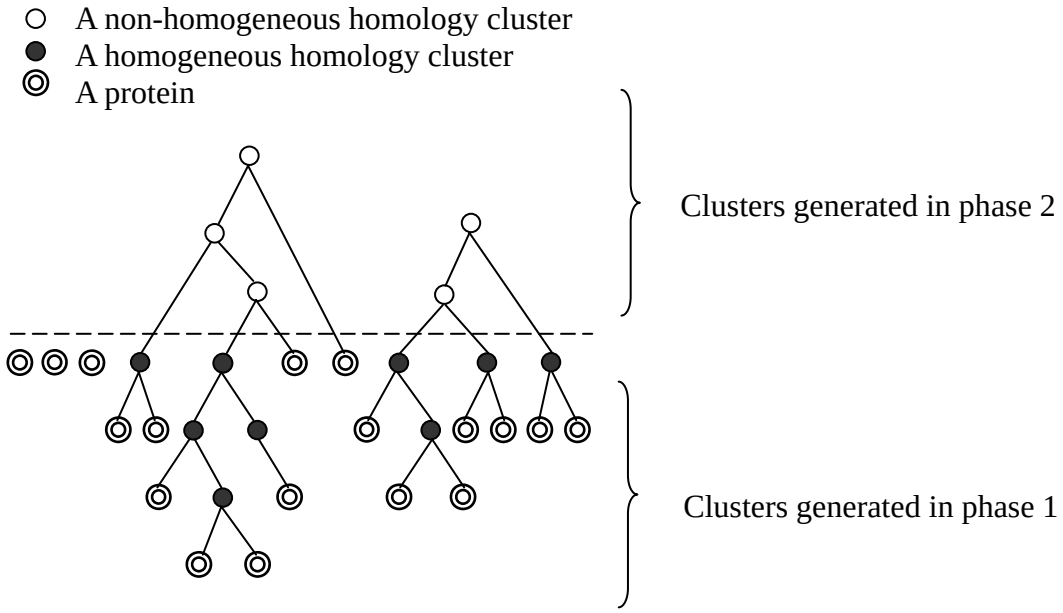


圖一. 於 Swiss-Prot 所有蛋白質的分群結果中，9017 個同質性同源蛋白質群集所包含的蛋白質個數統計分布圖。



圖二. 兩階段的複合式分群演算法

我們將我們所提的複合式分群演算法命名為 HomoClust。從表一及表二可以看出，於 Swiss-Prot 資料庫[3]中所有人類蛋白質的分群結果中，HomoClust 明顯地提昇了原有單一連結分群演算法的分群品質，其結果與平均連結分群演算法的分群品質差異不大，但從圖四中演算法效能之比較結果可以明顯看出，平均連結分群演算法並不適合使用於大量資料的分群問題上。接著我們比較 HomoClust 與單一連結分群演算法於 Siwss-Prot 整個資料庫的分群結果，從表三中可以觀察到，HomoClust 大福提昇了大型蛋白質家族的分群品質，此結果顯示我們所提的複合式分群演算法兼具同源遞移性與避免雜訊干擾的優勢。



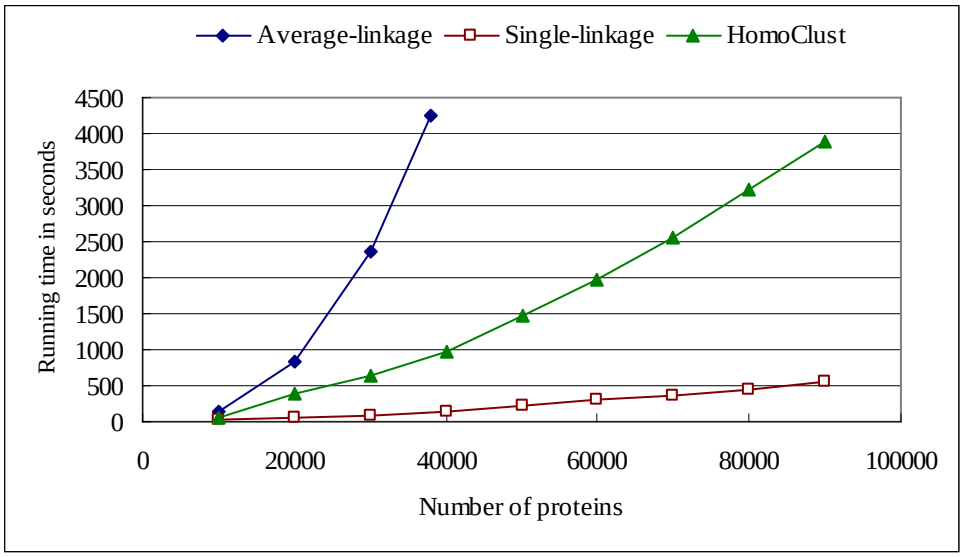
圖三. 蛋白質序列分群提供完整階層，下層為同質性同源蛋白質群，較上層的群集則為非同質性同源關係的蛋白質群

Family ID	Size of the family	Single-linkage	HomoClust	Average-linkage
IPR001314	75	75.79%	**82.95**%	82.76%
IPR005821	84	28.24%	57.78%	**72.16%**
IPR001254	92	84.00%	**91.40**%	91.30%
IPR001909	99	92.16%	91.26%	**95.19%**
IPR001806	116	83.62%	**98.28**%	86.36%
IPR001245	148	55.41%	61.59%	**62.75%**
IPR002290	212	72.73%	76.17%	**82.03%**
IPR000719	303	61.86%	**91.83**%	61.05%
IPR000276	307	67.43%	**99.03**%	97.08%
IPR003006	460	33.48%	**63.91**%	60.08%
Weighted average	-	59.64%	**80.59**%	75.28%

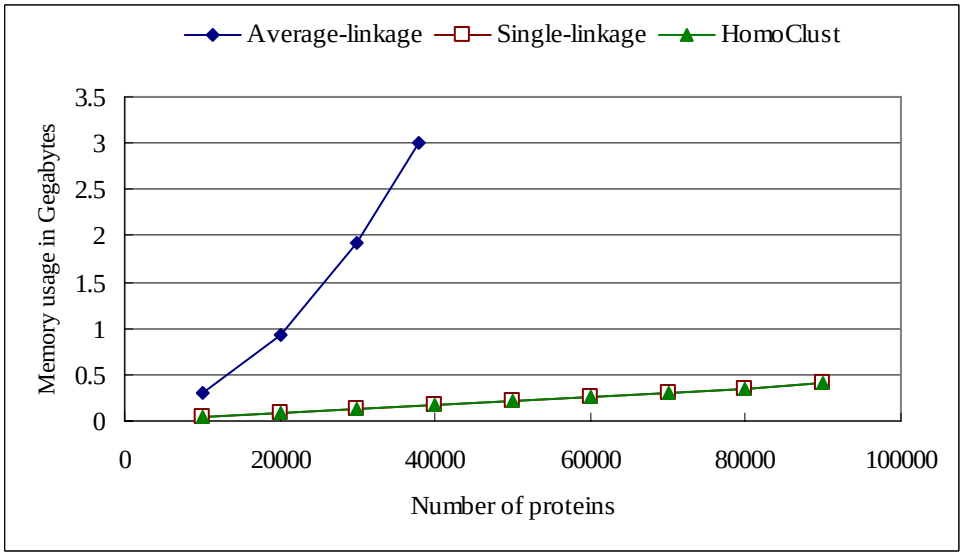
表一. 比較我們所提之複合式分群演算法(HomoClust)與單一連結和平均連結階層式分群演算法於十個最大蛋白質家族之分群品質，所使用的資料集為 Swiss-Prot 資料庫中所有的人類蛋白質

	Single-linkage	HomoClust	Average-linkage
Number of false negatives	1238	842	**631**
Number of false positives	274	**204**	455
Total faults	1512	**1046**	1086

表二. 比較我們所提之複合式分群演算法(HomoClust)與單一連結和平均連結階層式分群演算法之分群品質，所使用的資料集為 Swiss-Prot 資料庫中所有的人類蛋白質



(a) 執行時間與資料集中蛋白質個數之關係



(b) 記憶體使用量與資料集中蛋白質個數之關係

圖四. 三種演算法計算效能之比較

Family ID	Size of families	Single-linkage (Matching rate)	HomoClust (Matching rate)
52 families	≥ 200	71.46%	**81.95**%
IPR007114	395	74.03%	**83.03**%
IPR005834	415	66.67%	**81.39**%
IPR000836	421	**78.86%**	77.99%
IPR000685	459	100.00%	100.00%
IPR001254	486	71.93%	**78.83**%
IPR001806	584	73.63%	**83.55**%
IPR001245	614	38.44%	**41.49**%
IPR001412	627	66.20%	**68.15**%
IPR000795	643	42.46%	**92.17**%
IPR001128	673	88.30%	**98.66**%
IPR000971	837	94.03%	**98.57**%
IPR002290	1069	42.77%	**72.70**%
IPR003006	1206	58.29%	**60.20**%
IPR000276	1321	82.74%	**95.22**%
IPR000719	1449	39.51%	**84.21**%

表三. 比較我們所提之複合式分群演算法(HomoClust)與單一連結階層式分群演算法於家族個數超過 200 的蛋白質家族之分群品質，所使用的資料集為 Swiss-Prot 資料庫中所有的蛋白質

計畫成果自評

本計畫順利於一年內完成，前半年將計劃重點放在如何辨識哪些蛋白質可能具有多功特性而需要被放置在蛋白質階層的不同位置，後半年則研究階層式分群演算法於不同階段適用之分群條件，以滿足蛋白質家族分析之需求。本計畫完成的成果如下:

1. 應用先前根據蛋白質家族特性所提出的統計模型於開發複合式階層分群演算法的適用性，此部份結果已彙整成期刊論文，投稿於國際期刊 Pattern Recognition，日前已被接受[5]；
2. 設計視覺化工具幫助蛋白質分群演算法之開發，其人性化的使用介面同時可提供生物學家分析分群結果之用。(http://mars.csie.ntu.edu.tw/~cychen/HC/)

計畫執行成果與預期相符。此演算法的貢獻，不僅提供更為準確的蛋白質階層，更使得小規模的實驗室也能利用一般個人電腦，對十幾萬條蛋白質序列進行分群研究，建構分群階層。分群過程僅需數百萬位元組(MB)的記憶體使用量，並可在數小時內完成分群動作；此等運算規模在未來不僅可以加速更優良的分群演算法之開發，也奠定了本實驗室利用分群結果進行下一階段序列資料庫分析的基礎。

參考文獻

[1] Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389-3402.

[2] Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D.R., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M., Servant, F., Sigrist, C.J.A., Zdobnov, E.M. (2000), InterPro - an inte-grated documentation resource for protein families, domains and functional sites, Bioinformatics, 16, 1145-1150.

[3] Bairoch, A. and Apweiler, R. (2000) The Swiss-Prot protein sequence database and its supplement TrEmBL in 2000, Nucleic Acids Res., 28, 45-48.

[4] Bolten, E., Schliep, A., Schneckener, S., Schomburg, D., and Schrader, R. (2001) Clusteirng protein se-quences-structure prediction by transitive homol-ogy, Bioinformatics, 17, 935-941.

[5] Chen, C.-Y., Chung, W.-C. and Su, C.-T. Exploiting Homogeneity in Protein Sequence Clusters for Construction of Protein Family Hierarchies, accepted by Pattern Recognition, 2005.

[6] Chen, C.-Y., Juan, H.-F., and Oyang, Y.-J. (2004) Incremental Generation of Summarized Clustering Hierarchy for Protein Family Analysis, accepted by Bioinformaitcs and advance access published on May 6. 2004.

[7] Dayhoff, M.O. (1976) The origin and evolution of pro-tein superfamilies, Fed Proc, 35, 2132-2138.

[8] Enright, A.J., Dongen, S.V., and Ouzounis, C.A. (2002) An efficient algorithm for large-scale de-tection of protein families, Nucleic Acids Res., 30, 1575-1584.

[9] Enright,A.J. and Ouzounis,C.A. (2000) GeneRAGE: a robust algorithm for sequence clustering and domain detection. Bioinformatics, 16, 451-457.

[10] Hegyi, H. and Gerstein, M. (1999) The relationship between protein structure and function: a compre-hensive survey with application to the yeast ge-nome, J. Mol. Biol., 288, 147-164.

[11] Jain, A.K. and Dubes, R.C. (1988) Algorithms for clustering data, Prentice Hall.

[12] Jobson, J.D. (1991) Applied Multivariate Data Analy-sis, Springer-Verlag.

[13] Koonin, E.V., Tatusov, R.L. and Rudd, K.E. (1995) Sequence similarity analysis of Escherichia coli proteins - Functional and evolutionary implications, Proc. Natl Acad. Sci. USA, 92, 11921-11925.

[14] Kawaji, H., Yamaguchi, Y., Matsuda, H., and Hashi-moto, A. (2001) A graph-basd clustering method for a large set of sequences using a graph partition-ing algorithm, Genome Informatics, 12, 93-102.

[15] Kriventseva, E.V., Fleischmann, W., Zdobnov, E.M., Apweiler R. (2001b) CluSTr: a database of clusters of Swiss-Prot+TrEMBL proteins, Nucleic Acids Res., 29, 33-36.

[16] Lesk, A.M. (2002) *Introduction to bioinformatics*, New York : Oxford University Press.

[17] Matsuda, H., Ishihara, T., and Hashimoto, A. (1996) A clustering method for molecular sequences based on pairwise similarity, Genome Informatics, 7, 23-32.

[18] Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison, Proc. Natl Acad. Sci. USA, 85, 2444-2448.

[19] Pipenbacher,P., Schliep,A., Schneckener,S., Schonhuth,A., Schomburg,D. and Schrader,R. (2002) ProClust: improved clustering of protein sequences with an extended graph-based approach. Bioinformatics, 18, S182-S191.

[20] Rost B. (1999) Twilight zone of protein sequence alignments, Protein Eng. Vol. 12, No. 2, pp. 85-94.

[21] Sasson,O., Linial,N. and Linial,M. (2002) The metric space of proteins - comparative study of clustering algorithms. Bioinformatics, 18, S14-S21.

[22] Smith,T.F. and Waterman,M.S. (1981)Comparison of biosequences. Adv, Appl. Math., 2, 482-489.

[23] Watanabe, H. and Otsuka, J. (1995) A comprehensive representation of extensive similarity linkage be-tween large numbers of proteins, Comput. Applic. Biosci., 11, 159-166.

[24] Yona, G., Linial, N., Linial, M. (1999) ProtoMap: Automatic classification of protein sequences, a hi-erarchy of protein families, and local maps of the protein space. Proteins, 37, pp 360-378.

行政院國家科學委員會補助國內專家學者出席國際學術會議報告

94 年 9 月 10 日

附件

報告人姓名	陳倩瑜	服務機構及職稱	元智大學生物科技暨生物資訊研究所助理教授 自民國94年八月一日起轉任國立台灣大學生物產業機電工程學系助理教授
會議 (I) 時間 地點	94 年 7 月 31 日至 94 年 8 月 4 日 Canada, Montreal	本會核定補助文號	
會議名稱	(中文) 2005 國際類神經網路聯合會議 (英文) International Joint Conference on Neural Networks 2005		
會議 (II) 時間 地點	94 年 8 月 8 日至 93 年 8 月 11 日 美國 Stanford University	本會核定補助文號	
會議名稱	(中文) 2005 IEEE 計算系統生物資訊國際會議 (英文) 2005 IEEE Computational Systems Bioinformatics Conference		

報告內容：

一、參加會議經過

今年很高興可以參加國際類神經網路聯合會議，我們有兩篇論文被接受，其中一篇由我進行口頭報告(Oral representation)，另一篇則以海報(Poster)的方式報告。議程如下：

Plenary Poster Session P1-Gf: *Neural network architectures and structures*
Monday, August 1, 7:00PM-11:00PM, Room: Fountaine, Chair: Program Chairs
…
A Novel Radial Basis Function Network Classifier with Centers Set by Hierarchical Clustering [1460] Yu-Yen Ou, **Chien-Yu Chen** and Yen-Jen Oyang
…

Session O30: *SVM II*
Wednesday, August 3, 9:30AM-11:30AM, Room: Outremount, Chair: Theodore Trafalis
…
10:30AM Data Classification with a Relaxed Model of Variable Kernel Density Estimation [1451] Yen-Jen Oyang, Yu-Yen Ou, Shien-Ching Hwang, **Chien-Yu Chen** and Darby Tien-Hau Chang
…

此次類神經網路聯合會議於加拿大蒙特樓舉辦，會議主題包含 PERCEPTUAL AND MOTOR FUNCTION，COGNITIVE FUNCTION，COMPUTATIONAL

NEUROSCIENCE，INFORMATICS，HARDWARE，NEURODYNAMICS，ADAPTATION AND DECISION MAKING，APPLICATIONS。其中，除了類神經網路的理論性探討吸引我的注意之外，我也特別重視這些理論於生物資料分析的應用上。同時，這次我除了參與類神經網路聯合會議之外，我特地於回程的時候順路前往 Stanford 大學參加 2005 年的計算系統生物資訊國際會議。

計算系統生物資訊國際會議每年均於美國 Stanford 大學舉辦，自 2005 年 8 月 8 日起至 11 日止，為期四天，該會議由 IEEE Computer Society 主辦，每年舉辦一次，今年是第四次舉辦。由於意識到計算生物(Computational Biology)及生物資訊(Bioinformatics)的重要性，這四年的會議都受到全世界各學術與研究單位高度的注視，也因次會議的水準相當高。

此次會議所涵蓋的層面非常廣，舉凡 whole genome analysis、gene expression analysis、protein motif analysis、pattern discovery、sequence search and alignment、protein family classification、protein structure and function prediction、molecular evolution and phylogeny、functional genomics 及 molecular biology databases and data mining 等主題都是本次會議之重點。這些議程的內容或從實務面、或從理論架構，都與「計算生物及生物資訊」的主題相關。

這次參與 CSB 雖然沒有發表論文，但也因此有更多的時間參與 Psoter 的討論，與去年不同的事，今年在會場碰到了許多從台灣前往參加的研究學者，也顯示出國內對於生物資訊的研究越來越重視，有所多資深教授也都跨足這個領域，將他們所熟悉的資訊理論一一應用於新的生物問題上，讓人倍感興奮。

二、與會心得

從我在台灣大學資訊工程學研究所唸博士班開始，即積極與歐陽彥正老師從事類神經網路於建構資料分類器上之研究，近幾年來已經發表數篇會議論文與期刊論文。今年更選擇這個會議發表我們近期的研究成果，藉這個機會和與會專家進一步討論我們的研究方法與理念，兩篇論文結果都吸引不少專家學者參予討論，對於我們後續的研究有極大的幫助。

這次同時參加系統生物資訊研討會，能與許多這個領域的專家有所接觸，感到非常開心。這次大會同樣涵蓋了下列主題，對於如何分析及處理分子層次之生物資訊，及以計算機、數學及統計模式分析分子生物現象，有相當多的主旨演講及論文發表。同時對於技術層面之技巧亦有著墨，例如資料結構之設計、機器學習、演化計算、模糊邏輯、類神經網路、訊息學及圖形識別。所以此次會議提對資訊工程與生物科學共同合作研究開發提供了一個良好之基石及遠景。

資訊處理 、數學模式、人工智慧、圖形識別、系統分析等都將成為未來生命科學研究之主流，缺乏這方面之認知，我們就無法培養出能因應未來這種研究趨勢的科學家。台灣的科學家應掌握「國際合作」及「跨領域合作」之趨勢，摒棄門戶之見，大家共同合作，並主動參與相關國際性事務，將有有助於我國生物資訊科技產業之精進。

三、建議

各國已投資相當多的經費於生物資訊的研發，我國則尚在起步階段，相關產業界在此領域的投入相當有限，如何促進產業的投入並加強學術合作或產學合作，如何整合國內各研究於生物資訊之研究都是值得進一步討論的課題。本人此行參與兩個盛會，深覺此行收穫良多，對於國科會贊助此行經費，深感謝意。

四、攜回資料

1. 研討會論文集；大會議程手冊