國防大學國防醫學院

(90)年度

出國短期進修回國報告書

執行單位：生物化學科

執行期限：中華民國 90 年 6 月 15 日至 90 年 12 月 19 日止

計畫名稱：國外短期進修

短期進修人：胡 光 宇　　副教授

附件二

行政院及所屬各機關出國報告提要

出國報告名稱：短期進修

頁數 15 含附件：□是■否

出國計劃主辦機關：國防大學國防醫學院

聯絡人：楊素足　　電話：87923100 轉 18111

出國人員姓名：胡光宇

服務機關：國防大學國防醫學院　　單位：生物化學科

職稱：副教授　　電話：87923164 轉 18805

出國類別：□1 考察■2 進修□3 研究□4 實習□5 其他

出國期間： 90 年 6 月 15 日至 90 年 12 月 19 日止　　出國地區：美國

報告日期：91 年 3 月 6 日

關鍵詞：種緣關係分析、磷酸轉移酵素系統。

內容摘要：(二百至三百字)

種緣關係分析是定義或找到新基因的有效方法，且藉由分群的結果常可有效的預測出基因的功能。然而由於基因體計劃的發展，基因庫資料增加速度非常驚人，如何有效率又正確的分析種緣關係是很重要的課題。故此次出國選擇至對生物資訊學很有經驗及成果的實驗室短期進修，並於半年時間內以磷酸轉移酵素系統（PTS）為模型成功的開發出 5 個不同用途的程式，改進基因庫資料搜尋、擷取及格式轉換的缺失，且成功的建立具整合性資料的種緣關係樹狀圖，並藉分析定義出 PTS 中 HPr kinase 基因的 Signature 序列及其可能的功能及調節機制。改進後的種緣關係分析方法將可提供未來分子生物研究方向之指引，並可縮短未來新基因發現所需的時間。

磷酸轉移酵素系統之種緣關係分析 報告書

壹、中文摘要

種緣關係分析是定義或找到新基因的有效方法，且藉由分群的結果常可有效的預測出基因的功能。近來由於基因體計劃的發展，基因庫資料增加速度非常驚人，如何有效率又正確的分析種緣關係是很重要的課題。然而該分析方法目前仍有許多需要改進的空間，主要的有三方面：一、基因庫資料的搜尋、擷取及格式轉換，耗時且費力；二、多序列排列時，程式無法自動依各序列相關蛋白質活性中心保留序列進行排列，而手動排列又費時費力；三、傳統的種緣關係樹狀圖，僅列出蛋白質及其來源生物名稱，故所提供資訊有限，常造成判讀上的困難。因此職此次出國短期進修即選擇對生物資訊學很有經驗及成果的實驗室，期望能藉由一些定義已相當清楚的系統著手，針對上述種緣關係分析之三方面缺失進行改進。在進修半年中，本人選擇磷酸轉移酵素系統(PTS)為模型，成功地開發出 5 個不同用途的程式，以加速進行基因庫資料的搜尋、擷取及格式轉換。同時，成功地以已建立好的蛋白質功能區域資料為指引，使程式自動將所有序列相關蛋白質依其活性中心保留序列排列在一起。並以自行設計的巨集，將蛋白質名稱、大小、功能區域、以及其來源生物名稱和分類等資料整合在種緣關係樹狀圖中。目前已利用改進的方法成功地分析 PTS 中 HPr kinases、EIs、及 HPrs 等的種緣關係，並定義出 PTS 中 HPr kinase 基因的 Signature 序列及其可能的功能及調節機制，未來將進一步分析 PTS 其它的蛋白質，並將持續改進種緣關係的分析方法。此一研究成果將提供未來分子生物研究方向之指引，並可縮短新基因發現所需要的時間。

目次

貳、 短期進修目的

近年來由於基因體計劃的發展，人類及許多其他生物的DNA序列已漸漸被定序出來，基因庫資料增加的速度非常驚人。然而序列的了解只是研究的開始，定義出序列中具意義的基因並了解基因的作用及互相影響的機轉，進而應用在基因治療等領域，才是最終的目的。種緣關係分析就是定義或找到新基因的有效方法，而藉由分群的結果常可有效的預測出基因的功能。

回顧過去幾年在微生物及真核生物種緣關係分析上，已累積了一些實驗結果和經驗，也深深地感受到在種緣關係分析(Phylogenetic analysis)上，其方法和可供使用之程式工具，目前尚有許多需要改進的空間，主要的有下列三方面：

一、 基因庫資料的搜尋、擷取、和資料格式的轉換，耗時且費力，無法快速產生可供分析之資料組。

二、 多序列排列(Multiple sequence alignment)時，程式無法自動依各序列相關蛋白質活性中心保留(Conserved)序列進行排列，而手動排列又非常耗時且費力，無法快速產生可供種緣關係分析之多序列排列。

三、 以傳統方法所建構之種緣關係樹狀圖(Phylogenetic tree)僅列出蛋白(或基因)及其來源生物名稱，所提供之資訊有限，常造成判讀上的困難。

如果能有所改進，提高效率及正確性，對台灣基因體計劃的推動，基因治療的發展均應有極大的助益。

由於種緣關係分析是一個整合的領域，必須有對資料庫有設計經驗的

電腦人才，也要有對生物系統十分了解的生化人才，並有實驗人才來作實驗驗證分析模型的正確性，能夠有這樣的條件的研究室是職出國短期進修選擇的目標。美國加州大學聖地牙哥分校的 Milton H. Saier, Jr 教授在磷酸轉移酵素系統(Phosphotransferase system;簡稱 PTS)及其它運載子(Transporters)方面，已有長期之研究經驗，是這一方面之權威。其研究室分為資訊研究及分子生物實驗研究兩部分，具備的不同領域的研究人才，在生物資訊學的領域佔有相當的地位。而其研究室發展出一運載子分類系統(Transporter commission;簡稱 TC)，將運載子依其功能及種緣關係分為 150 個 families (http://132.239.144.23 /tcdb/)，最近已被國際生化及分子生物聯盟(International Union of Biochemistry and Molecular Biology; IUBMB)採用，此運載子分類系統(TC)與酵素之分類系統(EC)分法類似，不同的是 TC 系統除了功能外，亦考慮到種緣關係及蛋白質序列等資料。由於 Saier 教授在種緣關係分析的經驗與成就，故職選擇進入其研究室短期進修，期望能已定義已相當清楚的 PTS 為模型系統，針對上述所提目前種緣關係分析方法的缺失做些改進，並進一步利用這些工具分析其它有趣的基因，進而提供未來分子生物學研究方向之指引，縮短新基因發現所需要的時間。

參、進修過程

如前所述，Dr. Saier 的研究室主要以研究運載子為主，運載子是自然界中最重要的蛋白質之一，其基因數約在佔所有已定序完成微生物及真核生物基因體的 10%。運載子之功能相當多，包括：(1)養分吸收；(2)終產物

輸出；(3)保護細胞免於有毒代謝物及藥物之攻擊；(4)細胞間化學及電子之傳遞；(5)細胞器間(Interorganellar)及細胞質與細胞器間之聯繫；(6)代謝及代謝流(Metabolic flux)之調節；(7)其它細胞基本生命現象之調節，如細胞分裂、DNA 複製、及巨分子合成等。而 Dr. Saier 將其研究室分為兩部份：一為 Dry Lab，以 database 管理、處理生物資訊分析為主；二為 Wet Lab，即以分子生物學的方法進行實驗，以獲取實驗數據來驗證一些預測，或補充資料庫之資料。

由於近幾年來基因體計劃的陸續完成，而 *E. coli* 及 *B. subtilis* 這兩種微生物中控制其 PTS 蛋白質表現之 Operons 也已被找到。雖然有關 PTS 蛋白質之種緣關係分析也於 1995 分析過，實仍有必要加入新的資料重新分析。所以本人即投入其 Dry Lab，針對前述種緣關係分析上的問題，以 PTS 為模型系統，進行改進，希望達成的目標為：

一、 結合當今各種不同特色的相關序列搜尋程式如 Position specific iterated blast (Psi-blast)及 Hidden markov models(HMM)等，來收集及整合各主要基因體資料庫，如 NCBI(http://www.ncbi.nlm.nih.gov/)、Pfam(http://www.sanger.ac.uk/Pfam/)、及 SwissProt (http://www.expasy.ch/sprot/sprot-top.html)等的資料，除去重複及片斷不全的序列資料，以產生適合進一步分析的資料組。並發展新的程式以加速上述資料的搜尋、擷取、及格式的轉換，以便進一步進行多序列排列(Multiple sequence alignment)的分析。

二、 並將蛋白質功能區域(Domains)的觀念及資訊導入多序列排列之程式中，使程式自動依序列相關蛋白質活性中心保留(Conserved)序列排列，以產生適合進行種緣關係樹狀圖(Phylogenetic tree)建構的序列組。

三、 在樹狀圖的分析中，除了傳統上所使用的蛋白質及其來源生物名稱外，嘗試加入蛋白質大小，其組成功能區域順序、以及來源生物的分類等資訊，使所產生的樹狀圖能提供更多可用的資訊，以利於樹狀圖種緣關係的判讀。並發展巨集或程式以加速此一過程的完成。

主要實行的方法及過程如下：

A. 資料擷取及格式轉換

美國 NCBI 基因體資料庫，至 2002 年 2 月止共已收集 800 個以上基因體的資料(包括已完成及未完成的基因體)，如下表所示：

生物	基因體或序列
Bacteria	296 genomes
Archaea	29 genomes
Eukaryotes	359 genomes
Viruses	672 genomes
Organelles	229 sequences

而且 NCBI 的資料量正以約每 18 個月就增加一倍速度成長。面對如此龐大資料，要將所有 PTS 相關基因資料搜尋擷取出來，以供進一步分析，是一相當耗時且費力的工作。雖然全自動基因資料擷取、分析、及註解(annotation)是生物資訊學研究的首務，但是目前其準確率仍不夠高，且所產生的錯誤註解，可因一再地被引用，而迅速地傳播開來。因此，在目前仍需不同領域的專家，利用半自動之分析工具來進行基因體資料的進一步分析及註解。故在此一研究中，嘗試設計新的程式，以加速基因體資料庫資料的擷取及格式轉換，以產生適合直接進一步分析的資料組。

B. 基因庫資料的搜尋

以有興趣的蛋白質序列，自資料庫中調出所有相關的序列，卻又不

調出非相關的序列，目前仍是一件不容易的工作。在此一研究中，同時採用 Psi-blast 及 HMM 等兩種目前最被廣泛使用的方法，自 NCBI 及 Pfam 資料庫有效地找出所有 PTS 相關蛋白質的資料。而能同時被兩種方法偵測到，且 e 值達一定標準的將優先放入進一步的種緣關係分析中。

C. 多序列排列

當所有與有興趣蛋白質序列相關的蛋白質找到後，接下來的另一挑戰就是如何準確地進行多序列排列(Multiple sequence alignment)。當有了正確的多序列排列，才能進一步獲得可信的種緣關係樹狀圖。而多序列排列的基本要求是能將各序列中蛋白質活性中心的保留(Conserved)序列排列在一起。為快速且正確地完成多序列排列，在此一研究中，將以利用 HMM 所建立的蛋白質功能區域(Domains)資料，導入多序列排列程式，如廣泛被使用的 Clustal X，作為進行多序列排列時的指引，使程式能自動將各蛋白質活性中心保留(Conserved)序列排在一起，產生適合建構種緣關係樹狀圖之多序列排列資料組。

D. 種緣關係樹狀圖的建構

如前面 C 所述完成多序列排列後，下一步是利用它建構出種緣關係樹狀圖(Phylogenetic tree)。傳統上，樹狀圖僅包含蛋白質及其來源生物名稱，因所提供資訊的有限，常使研究者在樹狀圖的判讀上捉襟見肘。在此一研究中，為使所建構的樹狀圖內容更豐富，更易於判讀，擬於所建構的樹狀圖中，除了蛋白質及其來源生物名稱外，將加入蛋白質的大小、蛋白質功能區域(Domains)順序，以及來源生物分類(Taxonomy)等

資料。並嘗試利用自行設計的巨集(Macros)以加速此一過程。

肆、進修心得

一、 初步研究結果

利用前述的 Psi-blast 及 HMM 等方法。在此半年短期進修中本人已成功地開發出 5 個不同用途的 Perl 程式，以加速使用 Psi-blast 方法自 NCBI 資料庫進行資料的搜尋、擷取、及格式轉換，篩選所得進行分析所需之不含片段及重複序列的 HPr kinases 及 EIs 資料。

進一步為產生適合用於建構種緣關係樹狀圖的多序列排列資料，本人經過不同的嘗試，目前已成功地以已建立好的蛋白質功能區域資料為指引，利用 Clustal X 程式，將所有序列相關蛋白質依其活性中心保留(Conserved)序列自動排列在一起。所得多序列排列，再以自行設計的巨集，將蛋白質名稱、大小、功能區域組成、以及其來源生物名稱和分類等資料整合在裡面，並進一步以此整合的資料，建構出資訊整合的種緣關係樹狀圖。

而以改進的種緣關係分析方法，得到了下列十分有用的結果：

A. HPr kinases

以新資料用前述新方法所建立之 HPr kinases 種緣關係樹狀圖。發現沒有一個生物有超過一個的 HPr kinase homologue，顯示這些 HPr kinases 都是 Orthologues。

特別引人注意的是一般 HPr kinases 全長大小為 304-320 amino acyl

residues(aas)，然而此新方法所建構的圖可清楚地發現來自 4 個不同 α-proteobacteria 菌株的 HPr kinase orthologues，其大小竟然都只有其它的一半，即 144-154 aas，且均無法以 Pfam 資料定義其功能區域(Domains)。再者，4 個之中僅 *C. crecentus* 在 GenBank 中被註解為 HPr kinase，其餘均無法定義出其可能之功能。而多序列排列分析顯示這 4 個半長的 HPr kinases 與全長的中央區域相似，即 139-143 至 300-309 的區域，4 個半長 HPr kinases 均少了全長的 N-端及 C-端區域。而由微生物 *L. casei* HPr kinase 的立體結構及功能分析，其結果顯示全長 HPr kinaes 中央區域為該酵素與 ATP 及 HPr 交互作用的重要區域，也正是這 4 個半長 HPr kinases 仍保留的區域，顯示這 4 個半長 HPr kinases 很可能仍具有功能。因此，我們有興趣進一步分析這 4 個不同 α-proteobacteria 所特有 HPr kinases 所在 Operons 的結構，以瞭解在其 Operons 中其它的基因為何？及其可能參與催化的生化反應為何？

這 4 個不同α-proteobacteria 所特有 HPr kinases 所在 Operons 的結構分析，結果顯示這 4 個微生物 HPr kinases 鄰近基因排列相當一致(圖二)。這四個 Kinases 分別與其它 5 個(甚至可能 6 個)的基因依序排列組成 Operons，這些 Operons 所包含的基因依序為：Transcriptional response regulator(RR)、Sensor kinase(SK)、半長的 HPr kinase(K)、負責 Mannose 運載之 PTS IIA 蛋白質(A)、及 HPr(H)，再下去較遠處是 3-adenosyl L-homocysteine hydrolase 的基因(SAHH)。最近已有證據顯示該 SAHH 可能作用於產生一種細胞外"Quorum sensing"訊息分子。而這些可能同屬一 Operon 的基因，其基因表現方向與鄰近的 PEP carboxykinase (PCK) 相左。綜合這些現象，我們推論屬同一 Operon 的 HPr kinase、HPr、及

PTS IIA 彼此作用來調節 Sensor kinase 的活性。而 Sensor kinase 進一步磷酸化 Response regulator，Response regulator 在接著去調節 PCK 基因的轉錄。已知 PCK 是 Glycolysis 和 Krebs cycle 中由 C_3 變 C_4 步驟的主要作用酵素。此外， HPr kinases 的多序列排列顯示只有一個區域是最被保留(Conserved)，此一區域在多序列排列的 149-185 位置。其 Consensus 序列為：

^{149}(L I V) H G (L I V)$_3$ D (L I V) (Y F) G (L I V) G (L I V)$_3$ (T Q) G X S G (L I V) G K S E T A L E L V K R G H R^{185}

而其 Signature 序列為：

^{160}(I V L G) (Y F G D L) (G E S T W) X G (L I V A) (L I V F M) (L I V F) X G X (S A) (G A) (L I V S A) (G E) K (S A T) (E N A D T) (T S C L A)^{172}

[X= 任何氨基酸; 括號內為該一位置可能之氨基酸]

此一區域具有與 ATP 結合功能的 Walker A motif。以此 Signature 序列搜尋 SwissProt 及 TREMBL 資料庫，只找出 HPr kinases，顯示此 Signature 序列為 HPr kinases 蛋白質所特有。

B. EI

以前述之新方法所建構之 EIs 種緣關係樹狀圖可得知 PTS 之 EIs 依其種類及來源生物種類分成不同的群(Clusters)。

首先，在 EI 依來源生物種類分成不同的群方面。所有的 Low G+C Gram-positive 細菌的 homologues 均群集在圖的左邊，而 Mycoplasma、γ-proteobacteria、α-proteobacteria、High G+C Gram-positive bacteria、及

Chlamydia 則各形成其它不同的群。值得注意的是 *Lactobacillus brevis* 之 EI 蛋白質，如同其 HPr kinase，不與其它的 Lactobacilli 群集在一起，此一現象與其它實驗室所得結果一致。

其次，在 EI 依其種類分成不同的群方面。由新方法所建構之種緣關係樹狀圖，可得到：不同種類的 EIs 有其特殊之功能區域(Domains)，如典型之 EI 為 P_PC；Fructose EI 為(HPr)P_Pc_A2 或 A2_HPr_P_Pc；Nitrogen EI 為 G_P_Pc；而 N-acetylglucosamine EI 則為 A1_HPr_P_Pc。值得一提的是，結果顯示沒有任何 Gram-positive 細菌或 Chlamydial 或 Sporochete 生物具有一個以上的 EI paralogues。然而在 Proteobacteria 卻發現多個 Paralogues，*E. coli* 有 5 個 EI paralogues，*Pseudomonas aeruginosa*、*Yersinia pestis*、*Mesorhizobium loti* 及 *Caulobacter crescentus* 各有 3 個，*Vibrio cholerae* 有 2 個，而 E. coli 的 5 個 EI paralogues，包括：1 個 EI、1 個 EI^{Dha}、2 個 EI^{Fru}(EI^{Frw} 及 EI^{Fry})、及 1 個 EI^{Ntr}。其中目前只有 EI^{Frw} 及 EI^{Fry} 之功能仍不清楚。

由 EI 的多序列排列，可定出其 Signature 序列：

^{1086}G X M (I V L) E (I F V T) P (A S) X(12) (S D K A) F (L I V F M) (S A) (L I V) G (T S)N D (L I V M) X (Q G) (Y F) X (L I V M F) (A G S) X (D S A) R^{1125}

以此序列搜尋 SwissProt 及 TREMBL 資料庫只會找到 PTS 的 EIs，顯示此 Signature 序列為 PTS 的 EIs 所特有。

HPr

我們也以上述方法分析 HPr 蛋白質的種緣關係，由於該蛋白質平均

只有 88 aas，相當短，故所能提供的資訊有限，。

綜合以上初步研究結果顯示，以定義已相當清楚的 PTS 為模型系統，我們已成功地改進種緣關係之分析方法及軟體工具，以基因體資料庫中之新資料建構出易於判讀之 HPr kinases、EIs 及 HPrs 等蛋白質之種緣關係樹狀圖。

二、未來可繼續研究的項目:

在此一研究中，未來將努力的目標有下列三點：

A. EI-PPS-PPDK 種緣關係之分析

以 EI 蛋白質序列自基因體資料庫搜尋相關蛋白質序列時，除 EI 外，亦搜尋到 PEP synthases (PPSs)及 Pyruvate:phosphate dikinases (PPDKs)。這三種酵素的種緣關係，與先前所得之結果一致，這三種酵素依其功能不同於樹狀圖中形成了 3 個明顯不同的群，而與先前不同的是，以新資料利用新方法所建構的種緣關係樹狀圖，清楚顯示出這三種酵素其功能區域(Domains)及來源生物均有明顯的區隔：EIs 僅存在 Bacteria 中，其功能區域為 P_Pc 或 G_P_PC；PPSs 存在於 Bacteria 及 Archaea 中，其功能區域為 PDK_P_Pc，其中 PDK 係指 PEP/pyruvate binding domain；而 PPDKs 存在於 Bacteria 及 Eukaryotes 中，其功能區域為 PDK_P_PC。三種酵素的共同點是均含有功能區域 P_PC，以此區域所建構之種緣關係樹狀圖，或可提供瞭解 EI 如何演化來的重要線索。

B. Pfam 資料庫資料的搜尋、擷取及格式轉換

於初步研究中，本人所寫出 5 個不同用途的 Perl 程式，以加速使用

Psi-blast 方法自 NCBI 資料庫進行資料的搜尋、擷取及格式轉換。然而，目前以 HMM 方法自 Pfam 資料庫進行資料的搜尋、擷取及格式轉換仍相當費力耗時。因此，接下來的另一挑戰就是將嘗試改進此一問題。然而，由於目前網上之 Pfam 資料庫(http://www.sanger.ac.uk/Pfam/; http://pfam.wustl.edu/; http://www.cgb.ki.se/Pfam/)均沒有提供 Batch 服務，故可設置一台可長時間開機之伺服器在免費的 Linux 系統上，下載安裝免費的 Pfam 資料庫及軟件套件，並設計 Batch 程式來自動進行其資料的搜尋、擷取及格式轉換。

C. 其它 PTS 蛋白質之種緣關係分析

目前已完成 PTS 之 HPr kinases、EIs 及 HPrs 之種緣關係分析，將繼續進行其它的 PTS 蛋白質分析，包括：IIA、IIB、IIC 及 IID。

伍、 建議

這次出國短期進修所見，以國內外的研究環境而言，國內實驗室設施，絕對不比國外差，但研究成果卻仍遙落其後，可見問題並不出在設備差。我認為資源經費沒有有效率的整合、跨領域的合作不夠是主因。有些實驗室可能拿到許多不同的經費來做同一件事，經費遠超過其所需，有些實驗室可能有一些值得研究的方向，卻應經費的排擠效用，得不到支援，有些不同領域如電腦、物理、生物等可以各以專精的部分來合力解決一些研究上的問題，在國內卻很少看到這樣的整合。這些關於人力、經費的問題若能有合理的分配及整合，相信以國內人才的技術水準，要趕上歐美國家應是指日可待。

由於大部分研究做得較出色的地區，生活費都不低，以這次本人前往進修之以生物科技著稱的聖地牙哥為例，單人出租公寓的平均租金為美金一千元左右，每月美金九百元的生活費連房租都不夠付，遑論水、電、電話、食物、交通等必要開支。出國進修的補助，應是要使進修人員能在經濟壓力不會太大的情況下，安心努力的學一些有用的知識技能，回國貢獻，故建議檢討十多年來均未調整的生活補助費，以符合現實的需求，落實原本進修補助的美意。