出國報告（出國類別：進修）

赴美國短期進修心得報告書
（生物資訊學之新觀念與研究技術）

服務機關：國防醫學院生物化學科（所）
職稱：　　上校教師(副教授)
姓名：　　胡光宇
派赴國家：美國加州大學聖地牙哥分校
出國期間：94.6.28~94.12.29
報告日期：95.1.28

I. 摘要

此一爲期半年短期進修的主要目的是學習生物資訊學的新觀念與研究技術，以充分掌握大量的生物資訊及網路資源，並提升這方面的教學與研究品質。在國際知名的賽米頓教授實驗室中，本人選擇了在醫學上相當重要的運載子 OstA 爲模型，應用當今的各式資料庫、生物資訊分析工具、及本實驗室過去所發展的軟體工具來探討一些關於同源 OstA 基礎且重要的問題。由於 OstA 與一種和細菌內毒素生物活性有關的脂多醣體成分的運送有關，研究結果將有助於設計新的抗菌藥物。此外，此一研究建立一個生物資訊的分析模式，強化各種生物資訊技術的連結，去蕪存精，並增加新的概念，期能讓未來分析的速度可以趕上資訊累積的速度，使生物資訊的運用可以更有效，結果更正確。

目次 頁碼

II. 目的

在此為期半年的短期進修的主要目的在學習生物資訊學之新觀念與研究技術。生物資訊學為一利用電腦資訊科技來解決生物問題的科學。近年來，由於不同物種基因體解碼的陸續完成，提供了基因體大量的資訊，所產生的各式生物資料庫與網路資源，多不勝舉，非傳統分析方法所能充分掌握與運用。有感於生物資訊學之教學研究需要，其發展的日新月異，加上個人興趣，故前往在此研究領域享有盛名且有世界級先進超級電腦(http://www.top500.org/)的美國加州大學聖地牙哥分校進行短期進修，並選擇對生物資訊學很有經驗及成果的賽米頓 (Milton Saier) 教授實驗室，以研究合作的方式，學習國際一流研究單位的新觀念和方法，以提升在生物資訊學方面的教學及研究品質。

賽米頓教授是國際著名的生物運載子(Transporters)系統分析專家，因此，在進修半年中，本人選擇在醫學上相當重要的運載子 OstA 為模型，進行生物資訊學的學習與研究。蛋白質 OstA 為 Organic solvent tolerance protein 的簡稱，也有人稱其為 Imp (Increased membrane permeability)。先在大腸桿菌(*Escherichia coli*)的研究發現 OstA 在細菌細胞外膜的生合成中扮演重要的角色[1]。最近在腦膜炎雙球菌(*Neisseria meningitidis*)內的研究[2]，發現 OstA 是細菌的脂多醣體(Lipopolysaccharide；簡稱 LPS)運送到細胞外膜外面所必需，在缺乏 OstA 的突變株中，運送到細胞膜外的 LPS 劇烈減少。而 LPS 與致病菌致病的主要成分之一內毒素(Endotoxin)的生物活性有關聯，LPS 會誘發動物體內的各種發炎反應，產生廣泛的非特異性生理及病理學反應，如發燒、白血球數目改變、瀰漫性血管內凝血反應、腫瘤壞死、低血壓、休克和死亡[3]。而關於產生 OstA 蛋白質之基因研究中，前人的觀察認為同源 OstA (homolog)存在於大部分格蘭氏陰性菌(Gram-negative)中，但不存在格蘭氏陽性菌中[1]。因此，負責 LPS 運送到細胞外膜的 OstA 被認為是設計新的抗格蘭氏陰性菌藥物的良好目標物[2]。

目前許多研究發現由於 Metallo β-lactamase(MBL) 基因的存在，格蘭氏陰性菌的多重抗藥性成爲一個愈來愈嚴重的問題[4-8]。Chitnis 等人於印度的醫院中所做研究發現[5]，由病人體中分離出來的 1533 個格蘭氏陰性菌菌株中，有 11%對除了 Meropenem 之外的抗生素均有抗藥性，而有 6%的菌株對所有抗生素均有抗藥性，顯示出其嚴重性。尤其目前抗格蘭氏陽性菌的新藥有很大進展，如 Linezolid 及 Daptomycin 等陸續推出，但抗格蘭氏陰性菌的新藥卻付之闕如[8]，尋找一個新的途徑，來突破目前格蘭氏陰性菌抗藥性的問題就變得十分重要。OstA 蛋白質作用機制的發現，無疑給了這個問題一線解決的曙光，故而對 OstA homologs 完整的研究更顯重要。然而過去對 OstA homologs 的生物資訊研究並不是很多，再加上基因體研究技術的進展，基因體資料累積迅速，目前已累積超過 165,000 種生物超過 100 gigabases 的資料，且已有三百多個基因體已完成定序[9]。利用新的基因體資料對 OstA homologs 做一個完整的分析實有其必要性。然而，面對此一龐大且快速增加的生物資料庫，如何汲取日新月異的生物資訊原理與技術，以利找到可信的研究方向，對當今生物科學家而言也是一項挑戰。

因此，在此一短期研究最主要的目的乃是要應用當今的各式生物資料庫、生物資訊分析工具、及我們過去幾年所發展的軟體工具來探討一些關於 OstA homologs 基礎且重要的問題。例如考慮到未來抗菌藥物的專一性，瞭解到底有哪些生物具有 OstA homologs?是很重要的。除了格蘭氏陰性菌外，真核生物(Eukaryote)和古生物(Archaea)有 OstA homologs 嗎? 是否所有的格蘭氏陰性菌都有 OstA homologs 呢? 哪些沒有？沒有的是否有什麼共同的特性？ 其次，是否所有的格蘭氏陽性菌都沒有 OstA homologs? 若不是，哪些有？其共同特色爲何？這些最基本的問題，利用龐大的基因體資料來分析將可得到較肯定而可信的答案。再者，利用生物資訊的技術，我們可以分析出在 OstA homologs 中有多少是異物種同源 (Ortholog)?有多少是同物種同源 (Paralog)? 然後以我們發展出的資訊充沛的種緣分析技術將這些 OstA homologs 蛋白質序列的種緣關係(Phylogenetic analysis)、功能區域(Domain)、蛋白質分子大小、鄰

近基因、操縱子(Operon)組成基因等資料完整的呈現。並藉這些基礎資料來回答一些更重要的問題，例如 OstA homologs 一級結構的 Signature 與 Concensus sequences 究竟爲何?以及可能的二級以上結構爲何？是 α-helix?還是 β-barrel? 其 transmembrane 區域在哪裡?等。最後希望能藉著解答以上問題來瞭解不同生物基因體 OstA homologs 的特色及關係。

總之，此一研究以 OstA homologs 爲研究標的，應用當今的各式生物資料庫生物資訊分析工具，以及我們過去幾年所發展的軟體工具，來進行深入的探討，期能對未來的 OstA 相關研究提供方向指引，如前述以 OstA 爲標的物設計新的抗格蘭氏陰性菌藥物的研究。並建立一個生物資訊分析的模式，讓未來分析的速度可以趕上資訊累積的速度，使生物資訊可以更有效的運用。

III. 過程

主要的研究課題、實行的方法、及其過程如下：

A. 具有 OstA homologs 生物及其分類?

先以 NCBI Psiblast (Position-specific iterated BLAST) [10,11]將包括序列相關的 OstA 蛋白質自 nr (non-redundant)蛋白質資料庫找出來。Psiblast 有別於 Blastp，能將 Blast 結果產生新的 Profile 來進行另一次的 Blast，可一直重複 Blast 至找不到新的爲止，最後將序列遙遠相關的也一併找出來。此一結果可以回答到底有哪些生物具有 OstA homologs? 除了格蘭氏陰性菌外，真核生物和古生物有嗎?等問題。由於，目前已累積超過 165,000 種生物超過 100 gigabases 的基因體資料，我們已開發了一個爪哇(Java)程式 IRT 來快速準確地進行資料擷取、格式轉換，以及自建的 VB(Visual basic)巨集指令在 Excel 進行資料的篩選和比對分析。

B. Orthologs？Paralogs？功能區域?

上述 NCBI Psiblast 雖可將大部分序列遙遠相關的蛋白質找出來，但是對蛋白

質序列相似度低且分子小的，常會無法成功找到。此外，使用 Psiblast 或 Blastp 自 nr 資料庫搜尋時，找不到時，並無法確定該種生物沒有此一基因，因爲搜尋的資料庫是 nr 而非已完成基因體資料庫。類似地，若在一個生物找到許多個時，也無法斷定這些都是 Paralogs。因爲，nr 資料庫內可能是同種生物但是由不同實驗室定序並存入資料庫，且由於序列有些微不同，所以同時被保留在 nr 資料庫。爲了要有把握回答以下問題：是否所有的格蘭氏陰性菌都有 OstA homologs 呢? 哪些沒有？沒有的是否有什麼共同的特性？是否所有的格蘭氏陽性菌都沒有 OstA? 若不是，哪些有？其共同特色爲何？發現的 OstA homologs 中有多少是異物種同源 (Ortholog)?多少同物種同源 (Paralog)?等問題，我們使用了目前已完成的三百多個基因體資訊，同時配合使用不同的方法和資料庫來找出其 OstA homologs：

1. 以 NCBI 的 CDD (Conserved Domain Database) [12]所建立的 OstA/Imp 資料組 (COG1452)作爲起始資料組，由於該資料組是由 NCBI 電腦自動分析現有的基因體資料庫，所以，我們需要先進行資料整理並刪除重複或不正確的資料，然後此以資料組作爲 Query，於 NCBI genomic blast 一一搜尋三百多個已完成基因體的 OstA homologs (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi)。先使用 Genomic blastp 找出 OstA 蛋白質，對有疑問的基因體，則進一步以 Genomic blastn 檢視其核酸序列，是否有可能的 *ostA* 基因。利用此一方法，在初步研究中，我們發現格蘭氏陰氏菌 *Pseudomonas putida* 的已完成基因體中(GenBank accession number AE015451)，利用 Genomic blastp 找不到 OstA homologs，然而以 Genomic blastn 發現係因其基因體的 *ostA* 基因有一個經過確認的 Frameshift mutation 所致。相對於上述 Psiblast，此一方法，較能成功地將序列過度遙遠相關且分子較小的 OstA 蛋白質找出來。

2. 目前 NCBI genome blast 並不支援大量的批次(batch)分析，因此，下載了 NCBI blast (NCBI stand-alone version; ftp://ftp.ncbi.nih.gov/blast/)及已完成的三百多個基因體資料，一一進行批次 blast 分析，顯著縮短分析所需時間。同時，撰寫

Perl 的程式來進行分析結果的自動篩選及 Excel 下的 VB 程式進行資料整理分析比對。並且，與上述 Genomic blast 的方法配合使用，以有效降低搜尋失誤的機率。

3. 以上述方法找到的可能 OstA homologs，其資料庫註解往往是僅是一個基因體計畫的編號，因此，需要進一步檢視其真僞，並瞭解其結構的異同，以電腦分析預測其功能區域。目前蛋白質功能區域(Domain)預測的網站有不少，各有其特色，包括：Pfam[13]、CDD[12]、SCOP [14]、SMART [15]、InterPro[16]、ProDom [17]、PRODOC[18]、PROSITE [19]、及 PRINTS [20]等。大量批次分析時，我們先使用有歷史且穩定發展的 Pfam[13]。一一比對時將優先使用發展穩定且較 Pfam 靈敏的 NCBI CDD [12]以及最近功能整合的 Hpred [21]。除了蛋白質功能區域預測外，有些如 NCBI 也提供其他許多的資訊蛋白質活性位置(Active site)、切割位置(Cleavage site)、基質結合位置(Substrate binding site)等資訊，並持續增加中，對蛋白質功能進一步分析，提供有用的參考。另外，當線上分析如 Pfam 批次分析服務過慢或暫停時，使用了我們在 Linux 電腦系統上建置的 Pfam 資料庫，當長度約 500-1500 個氨基酸的蛋白質 300-500 個時，其分析過夜即可完成，可以確保功能區域分析之順利進行。

C. 蛋白質二級及 3D 結構?

OstA 是 LPS 的運送子(Transporter)，許多運送子是膜(Membrane)蛋白質，因此，利用現有的資料庫及軟體來預測其蛋白質結構，可以進一步比較所找到的 OstA homologs 的結構特色及之間異同。目前已有不少蛋白質二級結構的預測網站服務工具，在此不一一列舉，使用比較後，我們優先使用有持續更新並支援批次分析的工具：(1) Helices 結構以 HMMTOP[22]、PHOBIUS [23,24]、TMHMM [25]及 SOUSI [26]； (2) Extended 或 Beta-strands 則以 TMB-Hunt[27,28]、BOMP [29]; (3) Random coils 則以 REPPER [30]。當需要綜合深入分析時，使用 HHpred [21]、

Alignment Viewer (http://protevo.eb.tuebingen.mpg.de/) 及 iMolTalk [31,32]，同時分析蛋白質功能區域及 3D 結構預測。此外，我們撰寫 Perl 的程式來進行不同網站服務分析結果的自動擷取及 Excel 下的 VB 程式進行資料整理分析歸納。

D. (資訊充沛)種緣關係?

前述 Blast 等方法，可以將序列相關蛋白質找出來，其 E-values 可以作爲 Query 與找到序列相似度的參考，至於是否是最接近的，則需要進行種緣分析(Phylogentic analysis)。因爲 Blast 所得結果中，分數最高(E-values 最低)的在種緣分析時，並不一定都是最接近的。

正確的種緣樹狀圖(Phylogenetic tree)來自正確的序列資料組。因此，我們資料的選擇上非常用心，使最後所得 OstA homologs 資料不要少放也不要多放。在此，我們選擇了資料完整正確的已完成基因體(Completed genomes)資料庫作爲最後基因序列的來源。但是爲了不要漏掉一筆資料，先以上述 A 的 Psiblastp 自 nr 資料庫調出所有的 OstA 序列（包括遙遠）相關的蛋白質找出，並與由 B1 所述的 Genomic blastp 的資料交差比對，同時，也以 B2 的 Batch blastp 方法進一步確認一基因體內所有的 OstA homologs 都有在。這些以 Blast 找到的序列，並不都是 OstA homologs，所以需要以 B3 的蛋白質功能區域分析，C 的蛋白質二級以上結構分析結果，來做確認和刪除的工作。如此所得的 OstA 資料組，才用來進行下面的種緣分析。而種緣分析的結果，也作爲進一步資料組確認與刪除的工作。這些複雜的分析比對整理的過程，同樣地使用已有或新建的 Perl 及 VB 程式來簡化，同時避免錯誤的產生。

種緣關係的分析，我們使用前面提到的，我們最近幾年持續發展的資訊充沛種緣分析(Information-rich phylogenetic analysis)，其所建構的樹狀圖內容更豐富，除了蛋白質及其來源生物名稱外，更包括了分類符號、分類組群、蛋白質功能區域順序、及蛋白質長度等。這種資訊整合的樹狀圖，較傳統樹狀圖易瞭解，可讓我

們輕鬆、快速、準確地分析序列相關蛋白質的種緣關係。我們亦持續發展並應用自行設計的巨集及統一的工具來加速此一過程。

多序列排比(Multiple sequence alignment)使用 Clustal X[33]或線上 Clustal W[34]，當序列相似度低結果不理想時，嘗試了較慢但較準確的 T-COFFEE[35]，或較新的 MUSCLE[36,37]。建構樹狀圖則使用 Mega3.1[38]或 PAUP(http://paup.csit.fsu.edu/)。所得的 OstA 基因樹(Gene tree)，與 16S rRNA 基因樹比較，由於後者接近所謂物種樹(Species tree)，前後兩者相似度可以作爲是否是 OstA orthologs 的判斷。多序列排比結果進一步以 Pratt 定出 OstA homologs 的 Signature sequences[39,40]，並以線上 ScanProsite 確認之[41]。

E. 其他

至於 OstA 鄰近的基因，或所在操縱子(Operon)的組成基因，則以最新發展出來的 IMG 資料庫來分析[42]。同時，我們持續回顧了最新的文獻及線上資訊，來了解最新的資料庫及程式，以更新並豐富我們的生物資訊分析。並繼續發展 Perl、VB、及 Java 等軟體工具，以期有效利用各種推陳出新的生物資料庫及工具程式。

IV. 心得

此次爲期半年的短期進修，其心得可以整理出下列幾點：

1、 將序列相關的 OstA homologs 以最新的基因體資料進行分類及比對分析，可確認 OstA homologs 在不同種類生物中的分佈情形，了解哪些菌種具有這種基因，並探究其共同的特性。且可知道有多少是 Orthologs? 哪些是 Paralogs?

2、 利用我們發展出的資訊充沛種緣分析，可清楚呈現 OstA homologs 蛋白質序列的種緣關係（Phylogenetic relationship）、功能區域（Domain）、蛋白質分子大小、鄰近基因、操縱子（Operon）組成基因等，有利於 OstA homologs 種緣關係的正確判讀。

3、 利用現有最新的生物資訊技術找出 OstA homologs 的 transmembrane 區域、其一級結構的 Signature 及 Concensus sequence，並預測其二級以上結構。

4、 以上工作項目的完成，可建立 OstA 的基因體相關之系統生物分析。在目前格蘭氏陰性菌抗藥性問題亟待解決之際，完整的瞭解不同分類菌種 OstA homologs 之間序列的異同等問題，對未來新抗菌藥物的發展及其專一性的評估有很大的幫助。

5、 此一短期進修研究中另一達成之目標，乃是就本人的興趣、專長和經驗，以 OstA 爲起點，應用當今的各式生物資料庫生物資訊分析工具，以及我們過去幾年所發展的軟體工具，建立一個生物資訊分析的模式，加速分析的過程，讓未來分析的速度可以趕上資訊累積的速度，使生物資訊的運用更有效，與不同研究室合作，有效地應用在不同基因的生物資訊分析上，對不同目標基因之功能、機制、來源及變異發展等提供完整訊息，並可縮短新基因發現所需的時間，對生物學、生物資訊學及流行病學的研究與教學均會有相當的助益。

6、 此次短修所費不皆，因聖地牙哥當地生活費極高，光是單人房每月房租就要 1300 美元，所幸，此次短修，習得不少新的生物資訊學將原理與方法，感到充實愉快，將有助於日後的生物資訊學的教學與研究。此外，職在此六個月期間的研究，已完成包括一篇第一作者的兩篇研究報告，目前已投稿至國際期刊。同時也很高興與美國賽米頓教授建立良好的研究關係，目前正持續進行我們共同有興趣的第三篇論文的研究。

7、 同時，出國六個月期間，在美國定時利用網路電話及電子郵件，與在台灣的研究生、台大合作研究室及科辦公室進行討論和聯繫，完成的教學研究相關事項包括：

(1) 指導研究生並完成執行中的法務部研究計畫，並於年 94 年 7 月 17 日提

出期中報告，12 月 23 日完成期末報告。

(2) 進行與台大生命科學系宋教授研究室合作的兩項研究，其中一項的成果於 12 月 19 日完成校稿，並投稿至國際期刊。

(3) 完成法務部法醫研究所 95 年度專題研究計畫申請，於 12 月 12 日提出書面申請。

(4) 進行奇美專案研究計畫申請，於 10 月 25 日送出申請書面資料，於 12 月 25 日電子郵件送出電子檔及個人 CV。

(5) 完成國科會 95 年專題研究計畫申請，於返國前 12 月 26 日完成線上申請繳交工作。

V. 建議

此次短期進修後，有下列兩點粗淺的建議：

1、 生物資訊學爲現今生命科學發展最快的領域之一，有需要定期申請員額送訓，以提升生物醫學服務、教學及研究之水準。

2、 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爲例，單人出租公寓的平均租金爲美金一千元三百元，每月美金一千元的生活費連房租都不夠付，遑論水、電、電話、寬頻上網、食物、交通、及昂貴的健康保險等必要開支。因此，對有心申請出過短修的，雖然在政府財政日益困難的情形下，無法增加生活費補助，在出國進修規定方面建議多以鼓勵方向發展，除了提升個人學識經歷外，也有利於服務單位及國家未來的持續發展。

VI. 文獻

[1] Braun M, Silhavy TJ. Imp/OstA is required for cell envelope biogenesis in Escherichia coli. Mol Microbiol 2002;45 (5):1289-302.

[2] Bos MP, Tefsen B, Geurtsen J, Tommassen J. Identification of an outer membrane protein required for the transport of lipopolysaccharide to the bacterial cell surface. Proc Natl Acad Sci U S A 2004;101 (25):9417-22.

[3] Raetz CR, Whitfield C. Lipopolysaccharide endotoxins. Annu Rev Biochem 2002;71:635-700.

[4] Andes D, Craig WA. Treatment of infections with ESBL-producing organisms: pharmacokinetic and pharmacodynamic considerations. Clin Microbiol Infect 2005;11 Suppl 6:10-7.

[5] Chitnis S, Chitnis V, Hemvani N, Chitnis DS. In vitro Susceptibility to Meropenem and Other Antimicrobial Agents among Gram-Negative Bacilli Isolated from Hospitalized Patients in Central India. Chemotherapy 2005;52 (1):43-5.

[6] Geng SN, Rui YY, Wang Q, Mou CH, Zhou XH, Zhang J. [Analysis of drug resistance spectrum and its mechanism in 1017 clinical bacterial isolates.]. Di Yi Jun Yi Da Xue Xue Bao 2005;25 (12):1529-32.

[7] Nishino K, Latifi T, Groisman EA. Virulence and drug resistance roles of multidrug efflux systems of Salmonella enterica serovar Typhimurium. Mol Microbiol 2006;59 (1):126-41.

[8] Walsh TR. The emergence and implications of metallo-beta-lactamases in Gram-negative bacteria. Clin Microbiol Infect 2005;11 Suppl 6:2-9.

[9] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2005;33 (Database issue):D34-8.

[10] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25 (17):3389-402.

[11] Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001;29 (14):2994-3005.

[12] Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S, Helmberg W, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pontius JU, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L, Yaschenko E. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2005;33 (Database issue):D39-45.

[13] Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. Nucleic Acids Res 2004;32 (Database issue):D138-41.

[14] Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 2004;32 (Database issue):D226-9.

[15] Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P. SMART 4.0: towards genomic data integration. Nucleic Acids Res 2004;32 (Database issue):D142-4.

[16] Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowall J, Mitchell A, Nikolskaya AN, Orchard S, Pagni M, Ponting CP, Quevillon E, Selengut J, Sigrist CJ, Silventoinen V, Studholme DJ, Vaughan R, Wu CH. InterPro, progress and status in 2005. Nucleic Acids Res 2005;33 (Database issue):D201-5.

[17] Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S, Kahn D. The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res 2005;33 (Database issue):D212-5.

[18] Krishnadev O, Rekha N, Pandit SB, Abhiman S, Mohanty S, Swapna LS, Gore S, Srinivasan N. PRODOC: a resource for the comparison of tethered protein domain architectures with in-built information on remotely related domain families. Nucleic Acids Res 2005;33 (Web Server issue):W126-9.

[19] Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A. The PROSITE database, its status in 2002. Nucleic Acids Res 2002;30 (1):235-8.

[20] Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, Uddin A, Zygouri C. PRINTS and its

automatic supplement, prePRINTS. Nucleic Acids Res 2003;31 (1):400-2.

[21] Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 2005;33 (Web Server issue):W244-8.

[22] Tusnady GE, Simon I. The HMMTOP transmembrane topology prediction server. Bioinformatics 2001;17 (9):849-50.

[23] Kall L, Krogh A, Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. J Mol Biol 2004;338 (5):1027-36.

[24] Kall L, Krogh A, Sonnhammer EL. An HMM posterior decoder for sequence feature prediction that includes homology information. Bioinformatics 2005;21 Suppl 1:i251-i7.

[25] Moller S, Croning MD, Apweiler R. Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics 2001;17 (7):646-53.

[26] Hirokawa T, Boon-Chieng S, Mitaku S. SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics 1998;14 (4):378-9.

[27] Garrow AG, Agnew A, Westhead DR. TMB-Hunt: a web server to screen sequence sets for transmembrane beta-barrel proteins. Nucleic Acids Res 2005;33 (Web Server issue):W188-92.

[28] Garrow AG, Agnew A, Westhead DR. TMB-Hunt: an amino acid composition based method to screen proteomes for beta-barrel transmembrane proteins. BMC Bioinformatics 2005;6 (1):56.

[29] Berven FS, Flikka K, Jensen HB, Eidhammer I. BOMP: a program to predict integral beta-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. Nucleic Acids Res 2004;32 (Web Server issue):W394-9.

[30] Gruber M, Soding J, Lupas AN. REPPER--repeats and their periodicities in fibrous proteins. Nucleic Acids Res 2005;33 (Web Server issue):W239-43.

[31] Diemand AV, Scheib H. iMolTalk: an interactive, internet-based protein structure analysis server. Nucleic Acids Res 2004;32 (Web Server issue):W512-6.

[32] Diemand AV, Scheib H. MolTalk--a programming library for protein structures and structure analysis. BMC Bioinformatics 2004;5:39.

[33] Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 1997;25 (24):4876-82.

[34] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22 (22):4673-80.

[35] Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000;302 (1):205-17.

[36] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32 (5):1792-7.

[37] Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 2004;5:113.

[38] Kumar S, Tamura K, Nei M. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 2004;5 (2):150-63.

[39] Jonassen I. Efficient discovery of conserved patterns using a pattern graph. Comput Appl Biosci 1997;13 (5):509-22.

[40] Jonassen I, Collins JF, Higgins DG. Finding flexible patterns in unaligned protein sequences. Protein Sci 1995;4 (8):1587-95.

[41] Gattiker A, Gasteiger E, Bairoch A. ScanProsite: a reference implementation of a PROSITE scanning tool. Appl Bioinformatics 2002;1 (2):107-8.

[42] Markowitz VM, Korzeniewski F, Palaniappan K, Szeto E, Werner G, Padki A, Zhao X, Dubchak I, Hugenholtz P, Anderson I, Lykidis A, Mavrommatis K, Ivanova N, Kyrpides NC. The Integrated Microbial Genomes (IMG) System. Nucleic Acids Res 2006;34 (Database issue):D1-D5.