生物資訊在結構基因體學的應用

Application of Bioinformatics in Structural Genomics

呂平江、劉益忠、賴思明
Ping-Chiang Lyu, Yi-Chung Liu, Si-Ming Lai

結構基因體研究的目的是要以實驗及理論計算的方法來決定蛋白質的三度空間結構。蛋白質的結構資訊不僅對功能的註解有用，對新藥的研發也很重要，另一個目標是希望提供每一個蛋白質家族至少有一個代表性的結構。如何在眾多的候選者選出目標來決定結構是非常重要的，因此利用生物資訊的方法來分析序列及結構的資料，以提供目標蛋白的選擇就成了結構基因體計畫的第一步。我們可以藉由比較已知的資料庫資料，預測及分析目標蛋白，來幫助結構學家選擇目標，並提供一個溝通及協調的平台。

The structural genomics project aims at determination of the 3D structure of all proteins experimentally and theoretically. Structural information of a protein is valuable in functional annotation and powerful in new drug-discovery. Another goal is to provide at least a representative structure for each protein family. Sequence and structure data will be combined using bioinformatics methods (a) to compare with the known data in the database, (b) to predict and analyze targets, (c) to help target selection, (d) and to provide a platform to communicate and coordinate.

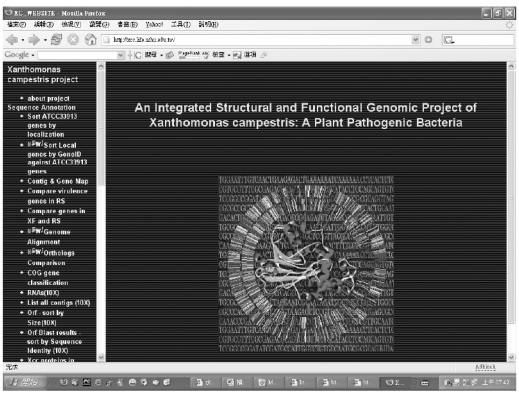
一、前言

從人類基因體計畫完成至今，也陸續完成了水稻、狗、大鼠及小鼠的基因體計畫，微生物方面則有將近三百個基因體被定序完成。面對這樣龐大的基因體序列資料，生命科學家接下來最重要的工作就是尋找這些基因的功能。方法之一就是透過與已知基因的比對和分析來對這些新定序的基因加以註解，但是還是有很多註解為所謂的「假設蛋白(hypothetical protein)」之基因無法經由比對和分析來了解其功能，因此世界各國紛紛成立結構基因體聯盟 (Structural Genomics Consortium)，希望藉由共同的努力能大量解出基因體計畫完成後可能基因之蛋白質的三維結構，再經由蛋白質結構的比對和分析能註解基因的功能，結構基因體計畫另外一個目的則是希望對每一個蛋白質家族都能提供一個代表性的結構，因為結構解析的速度遠比不上定序的速度。目前在基因資料庫的序列已超過百萬筆，但是在蛋白質資料庫的結構不超過三萬五千個，因此同一家族尚未解出結構的成員可以藉由此一代表性的

結構得到結構的資訊甚至模型。

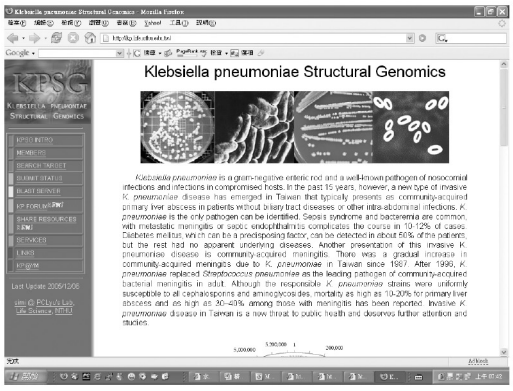
台灣目前共有四個結構基因體計畫進行中，這些計畫分別為：*Xanthomonas campestris* (XCC 計畫)、*Helicobacter pylori* 26695 (HP 計畫)、*Klebsiella pneumoniae* (KP 計畫) 及 *Stenotrophomonas maltophilia* (SM 計畫)，本實驗室有幸參與這四個計畫，並負責其中生物資訊的工作，我們為這四個計畫分別建立了網路分析管理平台 (見圖 1)，包括了電腦軟硬體的建置、分析工具之開發與目標蛋白的分析比對。希望透過全基因體的分析幫助結構生物團隊能快速的選擇目標蛋白進行實驗。本文係以目前在清華大學生物資訊中心建立的四個結構基因體計畫網站當作例子，來解說生物資訊在結構生物學上的應用。

結構基因體計畫首先要有全基因體的序列，如果是已經完成定序工作的基因體則可直接下載其全基因體序列進行基因分析及網站建置的工作，例如 HP 計畫及 SM 計畫就是如此。如果沒有全基因體序列的話則先要進行所謂的基因定序，即將每個染色體或質體上的核苷酸 (A、T、G、C) 順序排列出來，這個工作即稱為「定序」，例如 XCC 計畫及 KP 計畫就是由陽明大學蔡世峰教授所領導的定序團隊先進行基因體定序工作。由於目前的機器限制，不可能一次將所有的基因體從頭至尾一次定完，所以通常需要將整個基因組分段，予以個別定序後，再將重複定序的片段組合起來，接合成連續的序列。但因為許多片段序列的集合常常是原始序列長度的幾百倍甚至是幾千倍，所以此時便需要電腦來協助解讀這些序列的組合。不過這些序列中會有許多重複的片段，影響到電腦解析的效率，因此開發更強的演算法及分析軟體仍是一項非常重要的課題。目前較有名的軟體如：phrap 及 phred[1,2] 與

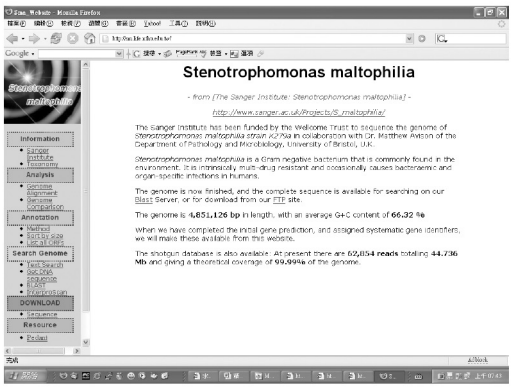


XCC 計畫

HP 計畫



KP 計畫



SM 計畫

圖 1. 四個進行中的基因體計畫網站。網址分別為: XCC - http://xcc.life.nthu.edu.tw/ ; HP - http://hp.life.nthu.edu.tw/ ; KP - http://kp.life.nthu.edu.tw/ ; SM - http://sm.life.nthu.edu.tw/。

consed[3] 等都是常用來做這些基因重組的軟體。

有了完整的序列資訊後，緊接的工作為找出真正的基因位置，並加上可能的功能註解，即我們所稱的基因預測和基因註解 (annotation)。在 XCC 計畫中陽明大學蔡世峰教授所領導的定序團隊完成了 10x 的定序工作，組成了 575 條 contigs 共 5,228,421 鹼基，針對這些序列作基因預測分析。基因的預測大都仰賴機器學習 (machine learning)，我們使用目前在微生物的註解上較有名的 Glimmer[5,6] 及 Orpheus[7] 兩個基因預測程式來做分析。另外 GeneMark[8-10] 程式也常被用來做基因預測，但因為已使用兩種預測程式交叉比對，所以未再使用第三種程式。此外在註解基因上，大都需要將未知基因與現有的基因資料庫做比對。較常使用的方式為：BLAST (basic local alignment search tool)[11]，利用小區域的字串比對尋找可能基因功能。另外有些資料庫也提供可能的結構區 (domain) 及模體 (motif) 比對，使註解的工作更準確。

而在有了基本的基因位置及序列後，在上千個基因中，如何找到適合的目標來進行研究的工作？由於目前用核磁共振 (nuclear magnetic resonance, NMR) 或 X 光繞射 (X-ray) 解結構的實驗方法有不一樣的考量，依照不同的方法，我們可以對目標的選擇加以篩選，這也是生物資訊可在此發揮其強大能力的地方，此步驟我們稱之為：目標選擇 (target selection)。下面我們簡單說明，從序列資訊如何藉由生物資訊軟體的分析，形成適合的篩選條件來找到適合研究的目標，以降低實驗的成本及花費的時間，這其中整合了公共資料庫、自建資料庫、網路分享軟體、自行研發程式及實驗團隊建議等不同區塊建置了結構基因體計畫分析管理網站 (圖 2)。

二、分析蛋白一級序列的資訊

由序列的一級結構，我們可以很容易地計算出序列的長度、分子量、胺基酸組成、Cystein 數目

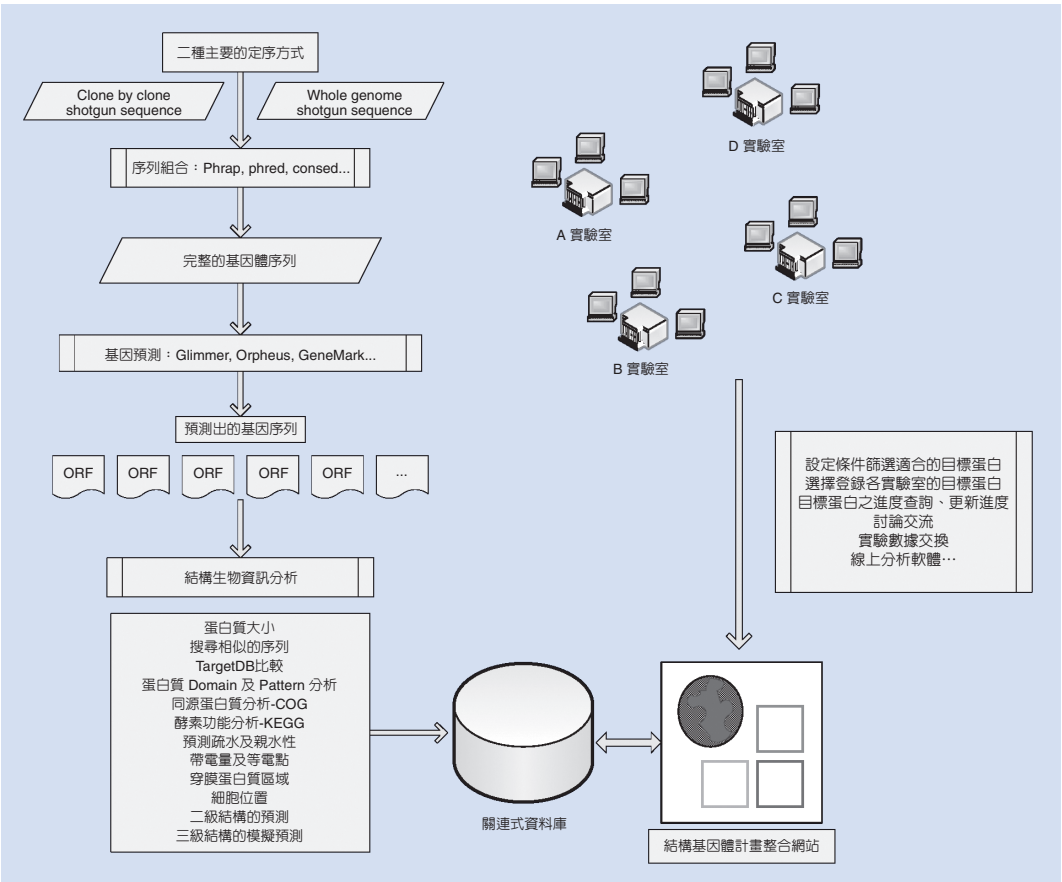


圖 2.
結構基因體計畫分析管理網站建置流程圖。

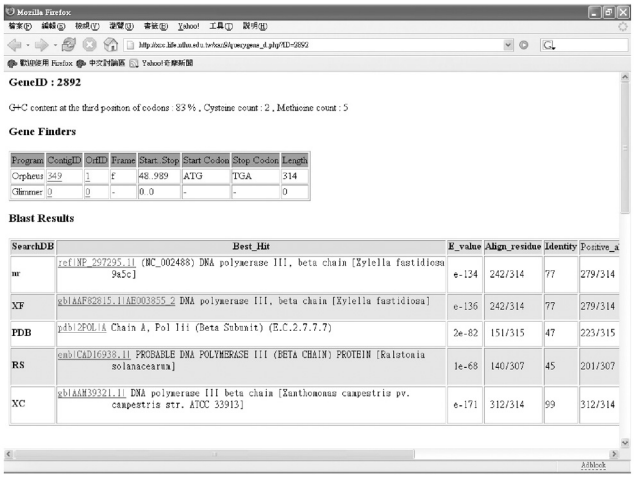


圖 3. 搜尋相似的序列 (XCC 計畫)。

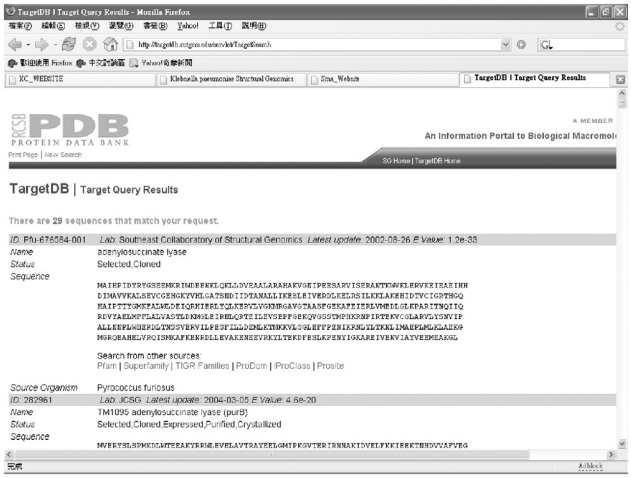


圖 4. TargetDB 搜尋結果 (KP 計畫)。

(考量雙硫鍵)、Methionine 數目 (考量解決 X 光繞射相角問題)、帶電量的組成等。常用的軟體如：商業的 GCG[12] 以及近來十分有名的開放軟體 EMBOSS[13] 套件，它們都包括了許多序列分析工具，可以有效地做大規模的計算分析。

1. 蛋白質大小 (長度及分子量）

在 NMR 的實驗上，蛋白質序列的大小 (size) 是實驗上的主要限制，通常我們可以針對已預測的蛋白依其分子量大小進行分類 (sorting)，將所有的 ORFs (open reading frame，開放讀碼框) 做排序後便於結構團隊挑選適合的分子進行實驗。

2. 搜尋相似的序列 (Similarity Search)

我們可以針對整個基因體中的所有序列，批次搜尋現有的序列資料庫，例如常見的 NR (non-redundant database) 資料庫，可以比對出相類似的基因出來。當相似度很高時可以作為功能的判斷；也可對於現今有的蛋白質結構資料庫 (PDB)[14]，看看是否已有相類似的蛋白質已經被解出結構，所有 ORF 比對的結果皆存在我們的資料庫，結構團隊不須再做比對即可直接點選參閱 (圖 3)。

3. 跟世界上大型的結構基因體計畫目標做比較－TargetDB[15]

TargetDB (目標資料庫) 收集世界上各大結構基因體計畫的目標蛋白，並且標明各個目標蛋白質目

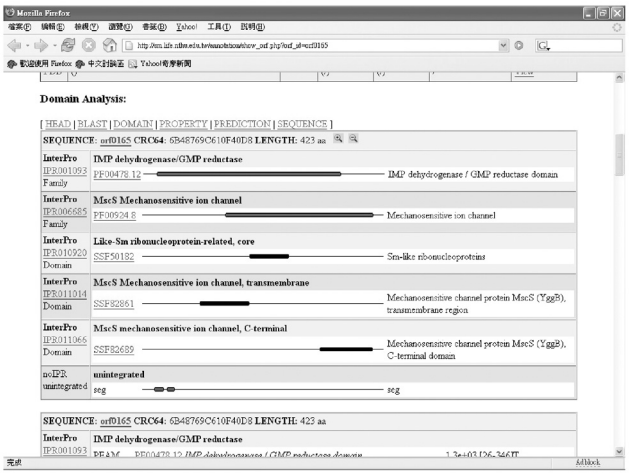


圖 5. Interproscan 分析結果 (KP 計畫)。

前的實驗進度。因此，我們在挑選蛋白質目標前，可以針對所有的蛋白質做全系列的搜尋，避免和其他國家的實驗室做相同的蛋白質，也可以知道是否選到了尚未有人投入研究的實驗目標，另外藉著參考相似序列的目標蛋白的實驗進度亦可推測實驗的難易度 (圖 4)。

4. 蛋白質 Domain 及 Pattern 分析

有些特殊功能或結構的蛋白質具有特別的序列特性，這時，也可據此同時大量分析我們感興趣的目標蛋白質，常用的結構區分析有 Interproscan[16]，這是一個整合型的分析套件，可同時針對多種特性作分析，另外，如 Scanprosite[17]，也可分析特殊的功能殘基序列，例如：磷酸化、糖化等 (圖 5)。

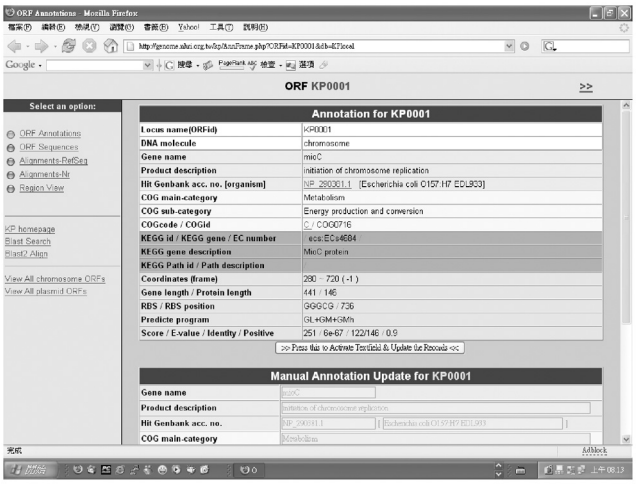


圖 6. COG 及 KEGG 分析結果 (KP 計畫)。

5. 同源蛋白質分析－COG (Cluster of Orthologous Groups of Proteins)[18-21]

此分析可依照不同基因在各個物種間的出現頻率加以分析，找出可能的功能分類。

6. 酵素功能分析－KEGG (Kyoto Encyclopedia of Genes and Genomes)[22-24]

許多有重要生物功能的蛋白質多為酵素，也是很好的目標蛋白。因此，在日本的一個組織建立了這個網站，將不同物種的酵素及生化途徑做分類，以便於比對出未知蛋白的酵素功能 (圖 6)。

三、以預測分析結果協助目標蛋白質的挑選

我們可以再進一步分析各種序列組成特性，依不同的演算法，預先了解此蛋白質是否容易表達、容易純化？是否會溶於水？是否為較難解結構的膜蛋白？會不會是分泌到細胞外的蛋白質或在細胞內表現？二級結構是 α-helix 為主？還是 β-strand 為主？這些問題，可以再當作挑選目標蛋白質的篩選條件，茲說明於下：

1. 預測疏水及親水性

一般而言，以核磁共振或 X 光繞射解蛋白質結構時，蛋白質的溶解度很重要，因此，以一級序列的組成作為分析依據，依照不同的殘基比重及視窗大小 (即 n 個殘基為一組) 給分，預測的結果可做為純化前的參考 (圖 7)。

2. 蛋白質帶電量及等電點 (pI) 分析

在純化蛋白質時，我們常用高效能液相層析 (high performance liquid chromatography, HPLC) 依蛋白的特性來純化。其中若是帶電的殘基較多或是等電點偏酸或鹼時，就可考慮用離子交換樹脂來進行純化工作；也曾有研究顯示，等電點值有助於判斷結晶時的 pH 值條件。

3. 穿膜蛋白質區域 (Transmembrane Domain) 預測

一般而言，膜蛋白較不易表現和純化，因此在挑選目標蛋白質時，除了特別要針對膜蛋白的研究外，將膜蛋白剔除在目標蛋白外，有助於提高成功的機率，我們使用目前較準確的預測軟體 TMHMM (predicts trans-membrane helices based on hidden markov model)[25,26] 來預測所有 ORF 可能的穿膜蛋白質區域。

4. 細胞位置分析

此目的可以預測蛋白質的位置，在蛋白質表現及純化上具有參考價值，我們使用的工具為 PSORT (prediction of protein sorting signals and localization sites)[27]。

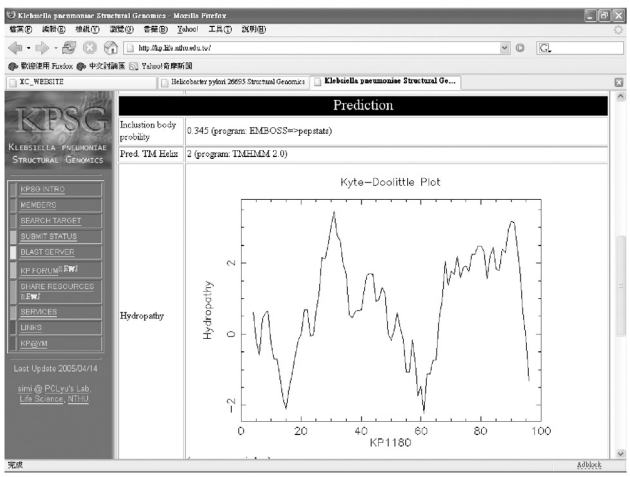


圖 7. 預測疏水及親水性結果 (KP 計畫)。

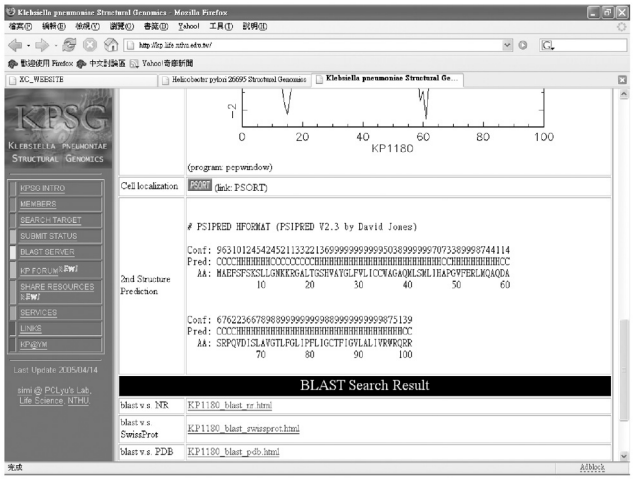


圖 8. 二級結構的預測 (KP 計畫)。

5. 二級結構的預測

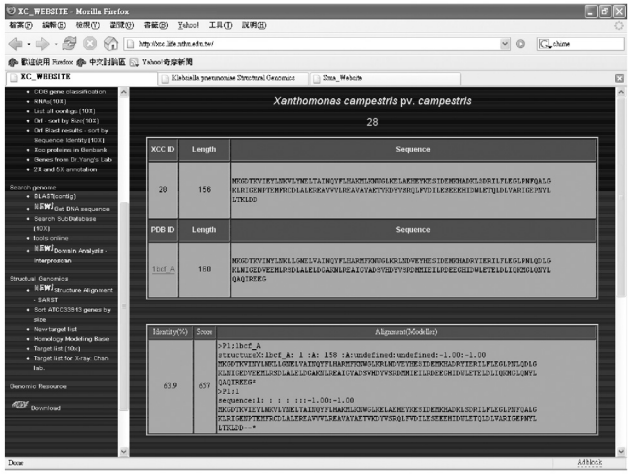
二級結構的預測對於尚未開始解結構的蛋白質先做預測，根據一級序列的組成，可預測 α-helix 及 β-strand 等結構，做為參考及目標選擇的考量。我們使用的預測程式為 PSIPRED (protein secondary structure prediction based on position-specific scoring matrices)[28] (圖 8)。

6. 三級結構的模擬預測

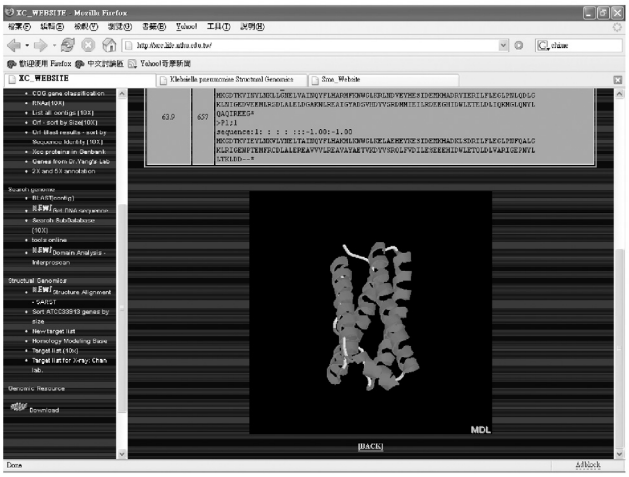
我們可以在結構基因體計畫中，大規模地利用現有已解出的結構為模板 (template)，模擬出想要解析的未知結構蛋白質。一般來說，當所要解析的蛋白質序列和模板間的相似性越高，所模擬出來的結構正確性、可信度也就越高。目前預測蛋白質結構的方法較常使用且準確度較高的為「同源模擬法 (homology modeling)」，例如：我們在 *Xanthomonas campestris* 計畫中將所有的蛋白質序列大規模地建立三級結構預測，選擇相似性高且成為目標蛋白機率低的蛋白質模擬出其三維結構 (圖 9)。

四、利用多種條件篩選出適合的目標蛋白

如圖 10 所示，在結構基因體計畫的網路管理平台上，可以整合不同條件篩選出適合的目標蛋白，包括分子量範圍、Cystein 數目、表現在包含



(a)



(b)

圖 9. 三級結構的模擬預測 (XCC 計畫)。

體 (inclusion body) 之可能性、transmembrane helices (穿膜螺旋結構) 的預測數目與已知結構之蛋白質相似度高低等條件，都可以藉由相關設定加以篩選。

五、結論

除上述之外，結構基因體計畫，通常需要多個實驗團隊來一同進行，為了避免選擇到重複的目標，一個整合的系統是必需的，利用網路四通八達的特性，可以達到資訊的即時更新與溝通，並可以隨時記錄各個團隊的實驗進度，加上線上的討論區及共享區，可以很快的做資源共享和意見交流 (圖 11)。當然，常用的分析工具也可以發展出網頁介

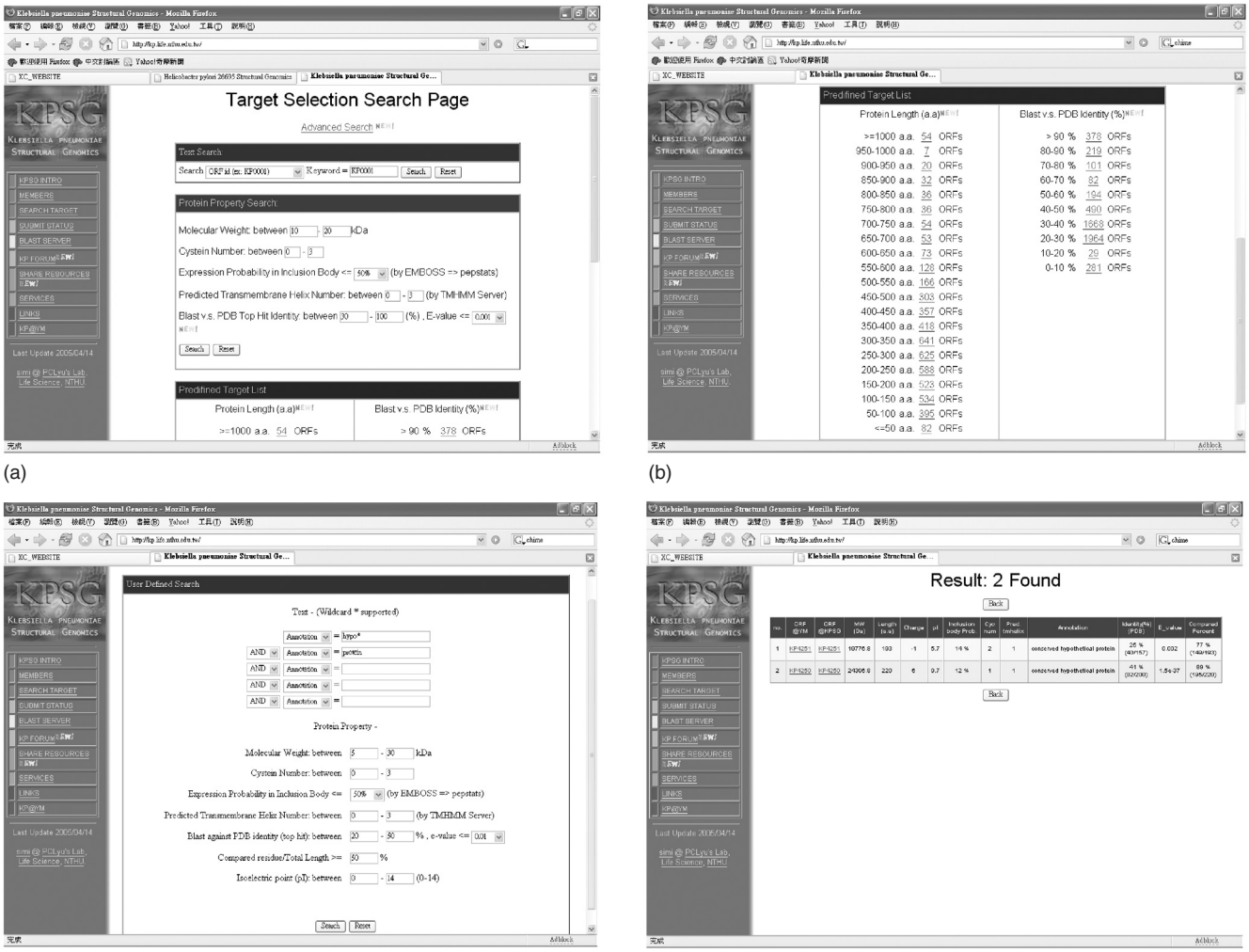
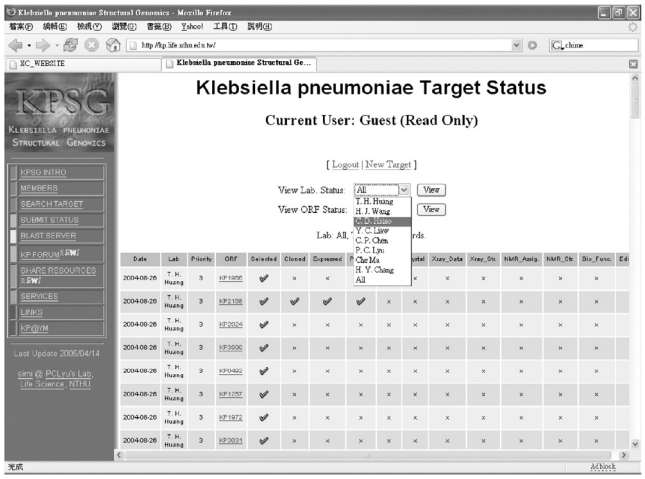


圖 10. 以不同條件篩選出適合的目標蛋白 (KP計畫)。(a) 目標蛋白搜尋介面；(b) 搜尋結果；(c) 目標蛋白進階搜尋介面；(d) 進階搜尋結果。

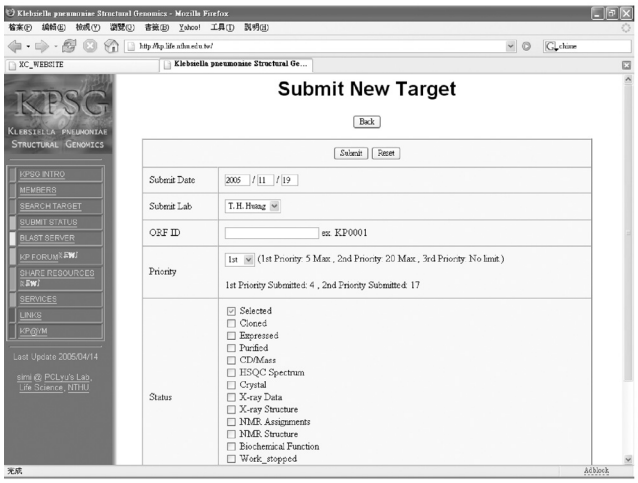
面，減低使用者安裝上的麻煩，因此建置交流及計畫控管的平台也是同樣重要的。為了實驗的保密性，網站必須加密，唯有在實驗團隊同意下才可釋出資料。生物資訊團隊在結構基因體計畫中雖扮演的是支援的角色，但是完善且功能強大的生物資訊分析管理平台支援，往往能使實驗團隊更方便及正確的做判斷，使計畫的進行更順利。

參考文獻

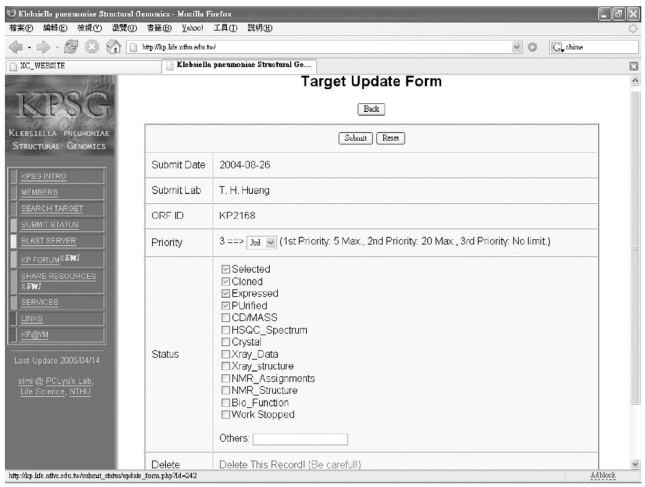
1. B. Ewing and P. *Green, Genome Res*., **8** (3), 186 (1998).
2. B. Ewing, L. Hillier, M. C. Wendl, and P. Green , *Genome Res*., **8** (3), 175 (1998).
3. D. Gordon, C. Abajian, and P. Green, *Genome Res*., **8** (3), 195 (1998).
4. E. D. Green, *Nat. Rev. Genet*., **2** (8), 573 (2001).
5. S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White, *Nucleic Acids Res*., **26** (2), 544 (1998).
6. S. L. Salzberg, M. Pertea, A. L. Delcher, M. J. Gardner, and H. Tettelin, *Genomics*, **59** (1), 24 (1999).
7. D. Frishman, A. Mironov, H. W. Mewes , and M. Gelfand, *Nucleic Acids Res*., **26** (12), 2941 (1998).
8. A. V. Lukashin and M. Borodovsky, *Nucleic Acids Res*., **26** (4), 1107 (1998).
9. J. D. McininchW. S. Hayes, and M. Borodovsky, *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **4**, 165 (1996).
10. Y. Almirantis and C. Nikolaou, *Comput. Biol. Med*., **35** (7), 627 (2005).



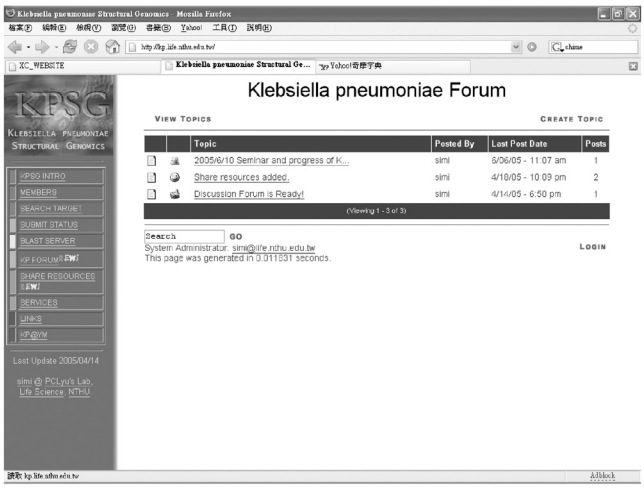
(a)



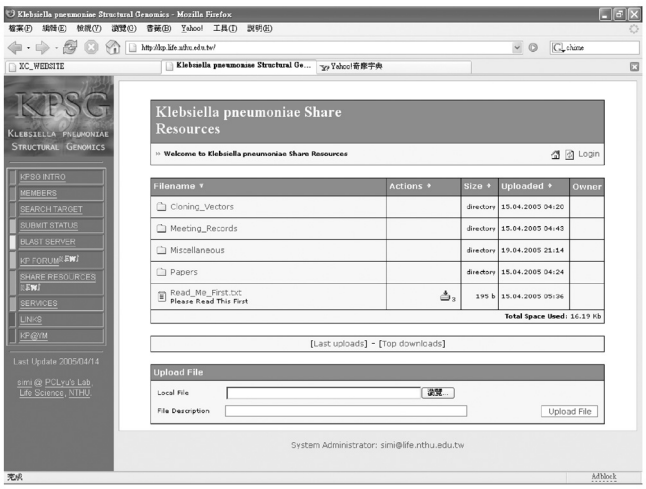
(b)



(c)



(d)



(e)

圖11. 結構基因體計畫管理交流平台 (KP計畫)。(a) 各團隊的目標蛋白選取更新及實驗進度記錄平台；(b) 傳送新目標蛋白介面；(c)實驗進度更新介面；(d) 線上討論區；(e) 資料共享區。

11. S. F. Altschul, W. Gish, W. Miller, E.W. Myers, and D. J. Lipman, *J. Mol. Biol.*, **215** (3), 403 (1990).
12. http://www.accelrys.com/products/gcg/.
13. http://emboss.sourceforge.net/.
14. http://www.rcsb.org/pdb/.
15. http://targetdb.pdb.org/.
16. http://www.ebi.ac.uk/InterProScan/.
17. http://www.expasy.org/tools/scanprosite/.
18. http://www.ncbi.nlm.nih.gov/COG/.
19. R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, D. A. Natale, *BMC Bioinformatics*, **4**, 41 (2003).
20. R. L. Tatusov, D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, E. V. Koonin, *Nucleic Acids Res.*, **29** (1), 22 (2001).
21. R. L. Tatusov, M. Y. Galperin, D. A. Natale, E. V. Koonin, *Nucleic Acids Res.*, **28** (1), 33 (2000).
22. H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, M. Kanehisa, *Nucleic Acids Res.*, **27** (1), 29 (1999).
23. M. Kanehisa and S. Goto, *Nucleic Acids Res.*, **28** (1), 27 (2000).
24. http://www.genome.ad.jp/kegg/.
25. E. L. Sonnhammer, G. von Heijne, and A. Krogh, *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 175 (1998).
26. http://www.cbs.dtu.dk/services/TMHMM/.
27. http://psort.nibb.ac.jp/.
28. http://bioinf.cs.ucl.ac.uk/psipred/

・呂平江先生為美國紐約大學生物物理博士，現任國立清華大學生命科學系教授兼系主任。
・劉益忠先生為國立台灣大學農學博士，現任國立清華大學生命科學系博士後研究員。
・賴思明先生為國立清華大學生命科學碩士，現為國立清華大學生命科學博士候選人。
・Ping-Chiang Lyu received his Ph.D. in biophysics from New York University, USA. He is currently a professor and chairman in the Department of Life Science at National Tsing Hua University.
・Yi-Chung Liu received his Ph.D. in agriculture from National Taiwan University. He is currently a postdoctoral research fellow in the Department of Life Science at National Tsing Hua University.
・Si-Ming Lai received his M.S. in life science from National Tsing Hua University. He is currently a Ph.D. candidate in the Department of Life Science at National Tsing Hua University.