行政院國家科學委員會專題研究計畫　成果報告

蛋白質交互作用及 Pathway 預測系統之研究與建構(2/2)

計畫類別： 個別型計畫
計畫編號： NSC93-2113-M-009-010-
執行期間： 93 年 08 月 01 日至 94 年 07 月 31 日
執行單位： 國立交通大學生物科技學系(所)

計畫主持人： 楊進木

報告類型： 完整報告

處理方式： 本計畫可公開查詢

中　華　民　國 94 年 10 月 27 日

中文摘要

關鍵字: 蛋白質-蛋白質交互作用、交互相似性、基因表現相關性、DAPID、蛋白質結合位、GEMDOCK、能量函式

蛋白質-蛋白質交互作用在細胞與功能性蛋白質體的許多調控作用中擔任核心角色。蛋白質-蛋白質交互作用與其結合區的預測能對巨分子辨識與生化機制研究提供重要的線索。在本計畫中我們已發表二篇SCI期刊論文：1) **J.-M. Yang*** and T.-W. Shen, "A pharmacophore-based evolutionary approach for screening selective estrogen receptor modulators," *Proteins: Structure, Function, and Bioinformatics,* vol. 59, pp. 205-220, 2005. 2) **J.-M. Yang*** Y.-F. Chen, T.-W. Shen, B. S. Kristal, and D. F. Hsu, "Consensus scoring criteria for improving enrichment in virtual screening," *Journal of Chemical Information and Modeling,* vol. 45, pp. 1134-1146, 2005. 另二篇論文正準備投稿中。

本計劃的主要研究成果分成三部份，第一部份是我們根據「交互相似性（interologs）」提出一個新的觀念：「結構功能區塊交互相似性（3D-domain interologs）」以預測蛋白質交互作用，我們亦能夠預測交互作用蛋白質對的交互作用功能區塊（interacting domain）。在酵母菌中，我們所預測的蛋白質-蛋白質交互作用和DIP資料庫達到18.6%重疊。我們將這些預測結果結合DIP、PDB等大量資料，建構了「功能區塊註解的蛋白質交互作用資料庫（DAPID）」。DAPID提供使用者觀察蛋白質蛋白質交互作用的交互作用功能區塊、實際接觸的胺基酸對，以及其它分子層次的詳細資訊。目前DAPID共收錄1008對結構功能區塊交互作用及135535對經過結構功能區塊註解的蛋白質蛋白質交互作用。DAPID已上線運作，可於 http://gemdock.life.nctu.edu.tw/dapid取得。

第二部份，我們發展了一套預測蛋白質結合位（binding site）的工具，透過分析蛋白質表面及結合位上胺基酸和二級結構的出現機率，再加入疏水性的概念，並且使用GEMDOCK的核心演算法最佳化（optimize）原子型態參數（atomic parameters），以預測蛋白質的結合位。在包含104個蛋白質的訓練資料組中，平均預測成功率為65.38%，其中Enzyme-inhibitor成功率為79.54%，Antibody-antigen成功率60.52%，其他類型的蛋白質成功率為45.45%。我們的程式在預測上具備了一定的可靠性，而此研究成果將來也可用做篩選前述第一部份所預測的蛋白質配對，再提升第一部份的預測準確度。第三部份是我們使用GEMDOCK進行蛋白質-蛋白質嵌合之預測，我們引入以知識為基礎的運算方式，將二十種胺基酸的原子細分為167種，可涵蓋五種主要的蛋白質交互作用力（氫鍵、靜電力、凡得瓦力、疏水性作用力、雙硫鍵）。為了提昇準確度，因此我們並額外加入了物理能量近似值的計分方式。在這兩種計分程式的預測表現上，167種原子分類在前兩百名平均預測成功的次數分別達到150次（bound structure）及98.17次（unbound structure）。物理能量近似值的計分方式在前兩百名平均預測成功的次數則為181.29次。

未來我們會將第二部份的研究（結合位的預測）整合至GEMDOCK中，藉此提高預測能力並縮短程式執行時間。最後，我們將整合本計劃的研究成果，完成一套自動化的蛋白質-蛋白質交互作用預測系統以及資料庫，提供相關服務。

Keywords: protein-protein interactions、interologs、Correlation coefficient of the gene expression profiles、DAPID、protein binding site、GEMDOCK、Scoring function

Protein-protein interactions play a central role in numerous processes in a cell and are one of the main issues of functional proteomics. Prediction of protein-protein interactions and binding sites is crucial to provide insight into the nature of macromolecular recognition and biochemical mechanisms. In this project, we have published two SCI journal papers: 1) **J.-M. Yang*** and T.-W. Shen, "A pharmacophore-based evolutionary approach for screening selective estrogen receptor modulators," *Proteins: Structure, Function, and Bioinformatics*, vol. 59, pp. 205-220, 2005. 2) **J.-M. Yang*** Y.-F. Chen, T.-W. Shen, B. S. Kristal, and D. F. Hsu, "Consensus Scoring Criteria for Improving Enrichment in Virtual Screening," *Journal of Chemical Information and Modeling*, vol. 45, pp. 1134-1146, 2005. We have also have prepared two papers to submit.

We have achieved three main results. In the first part, we proposed a new concept of "3D-domain interologs", which is similar to "interologs", to predict protein-protein interactions from 3D protein complexes. The mean correlation coefficient of the gene expression profiles of our predicted interactions is significantly higher than that for random pairs in *S. cerevisiae*. In addition, we find several novel interactions which are consistent with the functions of the proteins. Then we construct a Domain-Annotated Protein Interactions Database (DAPID) including DIP database, PDB and our predicting protein-protein interactions. The DAPID data model allows users to visualize the 3D interacting domains, contact residues, and molecular details of any predicted protein-protein interactions. The DAPID currently holds 1008 3D-interacting domain pairs and 135535 predicted 3D-domain annotated protein-protein interactions. It is available at http://gemdock.life.nctu.edu.tw/dapid.

In the second part, we developed a tool for predicting protein binding site. In order to predict protein binding site, we analyzed the probability of residue and secondary structure on protein surface and binding site, and integrated the hydrophobic concept. Besides, we used evolutionary strategies from GEMDOCK to optimize the atomic parameters. The average successful rate of predictions in training set (104 proteins) arrived 65.38%. Among, enzyme-inhibitor was 79.54%, antibody-antigen was 60.52%, and others were 45.45%. We can improve the accuracy of first part by our tool. In the third part, we used GEMDOCK to predict protein-protein docking and to model the protein-protein interactions based on important binding factors. We also use empirical-base scoring function to advance the predicting accuracy. The average numbers of hits in top 200 from knowledge-based scoring function were 150 (bound structure) and 98.17 (unbound structure). The average numbers of hits in top 200 from empirical-based scoring function is 181.29 (bound structure).

In the future, we will integrate the research of the second part (binding site prediction) into GEMDOCK. Finally, we will construct a new protein-protein interaction system and database, and expend our tool as an automatically pipeline system to provide related services.

目錄

第一部份：功能區塊註解的蛋白質交互作用資料庫

前言

大多數的生物運作機制都包含蛋白質蛋白質交互作用。欲了解這些機制，必須要全面性的研究蛋白質交互作網路。目前有許多大型的蛋白質交互作用網路資料庫，舉其大者如 DIP[1]，BIND[2]，MIPS[3] and STRING[4]，儲存由不同實驗方法及電腦預測的蛋白質蛋白質交互作用。由於用實驗方法偵測蛋白質蛋白質交互作用十分耗時而且昂貴，因此科學家發展了許多預測蛋白質蛋白質交互作用的方法，如「基因表現側寫(gene expression profiles)」[5]，「演化樹側寫(phylogenetic profiles)」[6]，「蛋白質結構複合體(know structure protein complex)」[7-9]，「交互相似性 (interologs)」[10] 及「功能區塊對側寫(domain pair profile)」[11] 等，希望能節省實驗所需成本，並且修正大規模實驗方法所產生的誤差。這些方法所預測蛋白質交互作用主要分為「物理性交互作用 (physical interaction)」，蛋白質之間有直接接觸，以及「功能性交互作用 (functional interaction)」，兩個蛋白質執行相似生物功能。

儘管有很多預測蛋白質交互作用的策略，但只有一小部分的方法是利用已知蛋白質結構複合體或者結構交互作用功能區塊資料庫，如 3did[12] 和 ipfam[13]. 一般說來，已知 3 級結構的交互作用蛋白質對，能提供科學家在原子的層次上觀察蛋白質交互作用並了解此交互作用如何產生。如果一蛋白質對能夠找到其同源「已知結構蛋白質複合體 (known structure protein complex)」，那麼用「比較模擬(comparative modeling)」的方法可以為此蛋白質對建立一個交互作用模型，並預測原子層次上蛋白質的交互作用情形[7-9]。

在本計畫中我們利用一個新的觀念「結構功能區塊交互相似性 (3D-domain interologs)」預測蛋白質蛋白質交互作用。結構功能區塊交互相似性的核心構想為：**在一個已知三級的結構的蛋白質複合體中，若蛋白質 A 上功能區塊 a 會與蛋白質 B 上功能區塊 b 發生交互作用，則它們的同源蛋白質 *A'* 以及 *B'* 也很有可能利用功能區塊 a 及功能區塊 b 產生交互作用**。根據結構功能區塊交互相似性的想法，我們預測了許多物理性代蛋白質蛋白質交互作用，並且用原子層次上蛋白質交互作用情形反應出蛋白質交互作用的專一性。我們的方法經由三種方式驗證，包括 TP/FP ratio，enrichment 以及基因表現側寫(gene expression profile)。

我們將這些預測結果結合 DIP 的資料以及 PDB 的資料建構一個「功能區塊註解的蛋白質交互作用資料庫 Domain-Annotated Protein Interaction Database (DAPID)」。DAPID 是第一個利用 3D-interologs 來預測有功能區塊註解的蛋白質蛋白質交互作用並且提供結構功能區塊交互作用的資訊以及任何其它分子層次的詳細訊息。目前 DAPID 共收錄 1008 對結構功能區塊交互作用及 135535 對經過結構功能區塊註解的蛋白質蛋白質交互作用。其中 32913 對 (24.3%) 從 DIP 資料庫收錄，1111 對 (0.8%) 從 PDB 蛋白質複合體收集，101511 對 (74.9%) 從 3D-domain interologs 推測而來(表一)。DAPID 內包含八

種常見物種的蛋白質蛋白質交互作用資訊，包括 *Homo sapiens, Mus musculus, Rattus norvegicus, Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae, Helicobacter pylori* 以及 *Escherichia coli*。

表一. DAPID 資料庫中八個物種蛋白質蛋白質交互作用統計表

Species	Our method	DIP[a]	PDB[b]
Homo sapiens	53669	1227	437
Mus musculus	39689	240	297
Rattus norvegicus	4461	91	70
Drosophila melanogaster	857	11847	4
Caenorhabditis elegans	941	3835	1
Saccharomyces cerevisiae	1158	14779	219
Escherichia coli	603	760	82
Helicobacter pylori	133	134	1
Total	101511	32913	1111

[a] 從 DIP 資料庫收錄的蛋白質蛋白質交互作用。

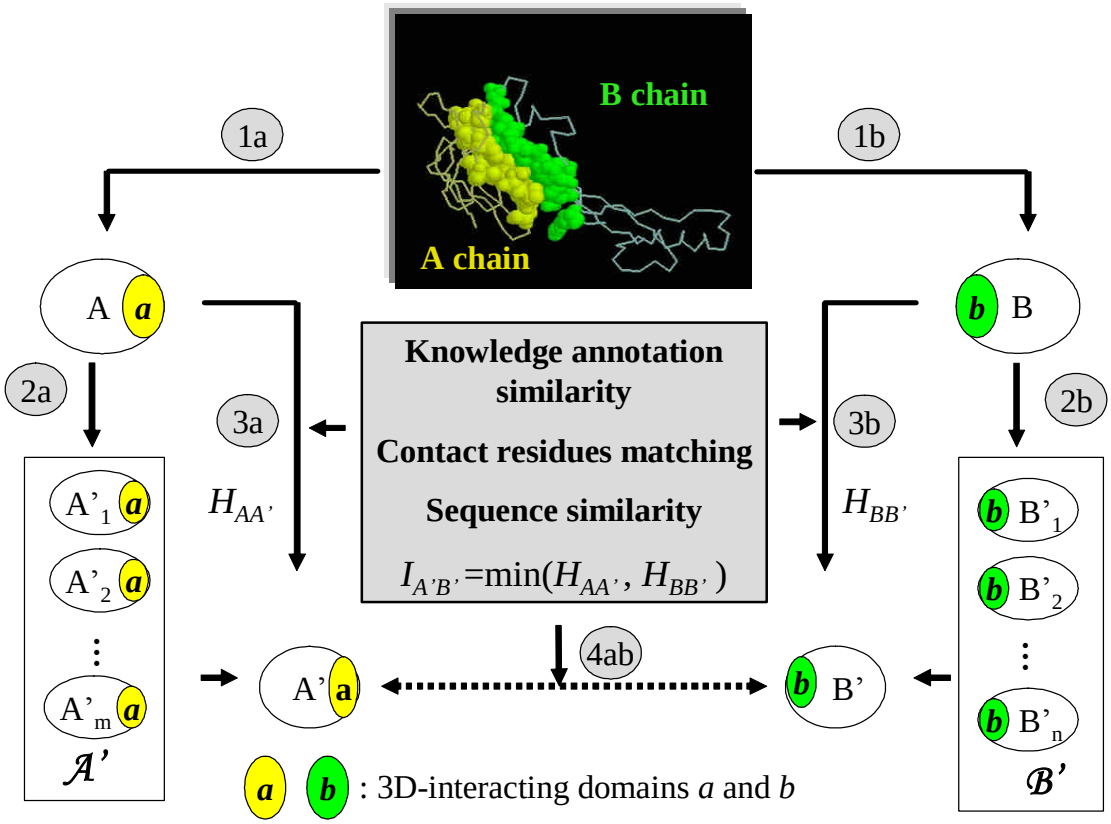
[b] 從 DIP 資料庫收錄的蛋白質蛋白質交互作用。

研究目的

在本計畫中我們提出一個新的觀念為「結構功能區塊交互相似性 (3D-domain interologs)」，利用高解析度的蛋白質複合體結晶結構，配合一個計算同源性分數的新計分函式來預測蛋白質蛋白質交互作用。另一方面我們整合預測的蛋白質蛋白質交互作用以及 DIP 和 PDB 資料庫，建構出「功能區塊註解的蛋白質交互作用資料庫」(Domain-Annotated Protein Interaction Database (DAPID))，它提供使用者觀察兩個互相作用蛋白質之間的結構功能區塊交互作用(3d-interacting domain)，以及任何其它分子層次的詳細訊息。

研究方法

(一) 方法概論：



1a Map binding area to Pfam domain *a* **1b** Map binding area to Pfam domain *b*
2a Find protein A'$_i$ that contains domain *a* **2b** Find protein B'$_j$ that contains domain *b*
3a Calculate homologous score between A and A' **3b** Calculate homologous score between B and B'
4ab Score and identify the protein-protein interaction A' and B' by joint scoring function $I_{A'B'}$

圖一. 結構功能區塊交互相似性 (3D-domain interologs)示意圖。

我們的方法利用「結構功能區塊交互相似性 (3D-domain interologs)」產生許多有功能區塊註解的蛋白質蛋白質交互作用(圖一)，其步驟如下：

I. 從已知結構蛋白質複合體(蛋白質 *A*,蛋白質 *B*)找出結構交互作用功能區塊(功能區塊 a, 功能區塊 b)以及實際接觸的胺基酸對。

II. 利用 Pfam[14] 資料庫定義的功能區塊區域，將結構交互作用功能區塊對應到 Pfam 功能區塊。

III. 從 Swiss-prot 資料庫 [15] 找出兩群分別含有功能區塊 a 的蛋白質家族 A 以及含有功能區塊 b 的蛋白質家族 B。

IV. 分別計算蛋白質家族 A 的成員 *A*'與模板蛋白質 A 以及蛋白質家族 B 的成員 *B*'與模板蛋白質 B 的"同源性分數(homologous score)。同源性分數有三個衡量標準:序列相似性，功能註解相似性和接觸胺基酸吻合度。

V. 計算蛋白質 *A*'與蛋白質 *B*'的連接分數(joint score)，如果連接分數大於某個門檻臨界值，我們就判斷 *A*'會與 *B*'產生交互作用。

(二)結構交互作用功能區塊 (3D-interacting domain)

為了建構結構交互作用功能區塊資料庫，我們從已知結構蛋白質複合體找出兩個 protein 之間的接觸胺基酸對。當蛋白質 A 上某個胺基酸的 C_β 原子和蛋白質 B 上另一個胺基酸的 C_β 原子距離小於 8Å 時(若胺基酸為 glycine 時，則用 C_α 代替)，這兩個胺基酸即為接觸鞍基酸對。接觸胺基酸對是結構交互作用功能的核心區域，若一個蛋白質有超過 5 個以上的接觸胺基酸，則我們認為這個交互作用的產生是合理的 [11]。我們搜尋整個 PDB 資料庫共找到 1649 對相異蛋白質複合體。為了定義結構功能區塊，我們利用 Pfam 資料庫所定義的功能區塊界線。我們將 Pfam domain 投影到 1649 對蛋白質複合體上得到 1008 對不同的結構交互作用功能區塊。

(三)預測可能的蛋白質蛋白質交互作用

我們利用「結構功能區塊交互相似性 (3D-domain interologs)」產生所有可能的蛋白質蛋白質交互作用(圖一)。前一個步驟已經找出每一對蛋白質複合體(蛋白質 A, 蛋白質 B)的結構交互作用功能區塊(3D-interacting domain)，然後我們從 Swiss-prot 資料庫中搜尋所有含有相同功能區塊的蛋白質為蛋白質家族 A 以及蛋白質家族 B。假設蛋白質家族 A 共有 m 個成員而蛋白質家族 B 共有 n 個成員，我們認為所有兩個家族的成員組合都有可能產生交互作用(即 m*n 種蛋白質交互作用組合)，因為它們都共同擁有結構交互作用功能區塊。如此一來，我們從 1008 對結構交互作用功能區塊中，預測出 24353269 對蛋白質蛋白質交互作用，其中有 845041 對蛋白質蛋白質交互作用的兩個蛋白質皆屬於同一個物種。

(四) 連接分數(joint score)

這個部分的目標是計算從"結構功能區塊交互相似性 (3D-domain interologs)"推測的蛋白質蛋白質交互作用的可靠度(reliability)。我們結合序列相似性，功能註解相似性和接觸胺基酸吻合度計算蛋白質家族成員 (*A'*或 *B'*)與模板蛋白質 (A 或 B) 之間同源性分數(homologous score)，進而設計了一套新的連接分數。兩個同源性分數之間較低者定義為 *A'*與 *B'*之間的連接分數($I_{A'B'}$)，如下所示：

$$I_{A'B'} = \min(H_{AA'}, H_{BB'}) \quad (1)$$

其中 $H_{AA'}$是蛋白質家族成員 *A'*與模板蛋白質 A 之間同源性分數；$H_{BB'}$是蛋白質家族成員 *B'*與模板蛋白質 B 之間同源性分數；模板蛋白質 A 與 B 是已知結構蛋白質複合體中的一個蛋白質，分別有交互作用功能區塊 a 與交互作用功能區塊 b。同源性分數 $H_{AA'}$定義如下：

$$H_{AA'} = 0.5(E_{AA'} + S_{AA'}) + K_{AA'} + D_{AA'} \quad (2)$$

其中 $K_{AA'}$是從 Swiss-prot 資料庫所定義的關鍵字(keyword)計算而得的分數；$E_{AA'}$ ($E_{AA'}$ = -log(e-value))與 $S_{AA'}$是用 Psi-blast 對蛋白質家族成員 *A'*與模板蛋白質 A 進行序列比對而得的 E-value 及序列相似度；$D_{AA'}$是根據蛋白質家族成員 *A'*與模板蛋白質 A 序列比對的結果，將接觸胺基酸的部分利用 BLOSUM62 substitution matrix 計算分數。三項分數皆先標準化至 0 到 1 之間以利正確合併。

為了計算功能註解相似性的分數($K_{AA'}$)，我們利用並且稍微修改在文件搜尋(document retrieval systems)中很常被運用的"*TF-IDF* scoring scheme"，其中 *TF* 是 一個特定蛋白質中某一個關鍵字出現的頻率，而 *IDF* 是一個特定關鍵字在全部 Swiss-prot 蛋白質資料庫出現的頻率的倒數 [16]。在 Swiss-prot 資料庫中，一個特定蛋白質的某一個關鍵字 i 的 TF_i 是 1 而 IDF_i 等於 $\log_2(N/n_i)$，其中 N 是 Swiss-prot 資料庫中全部蛋白質的數量(188477 in Release 47.5)。對一個蛋白質 A 來說，其中某一個關鍵字的 *TF-IDF* 加權比重(W_{Ai})被定義為 TF_i 乘上 IDF_i。給定一對蛋白質對 A 與 A'，它們的功能註解相似性分數($K_{AA'}$)計算方式為：

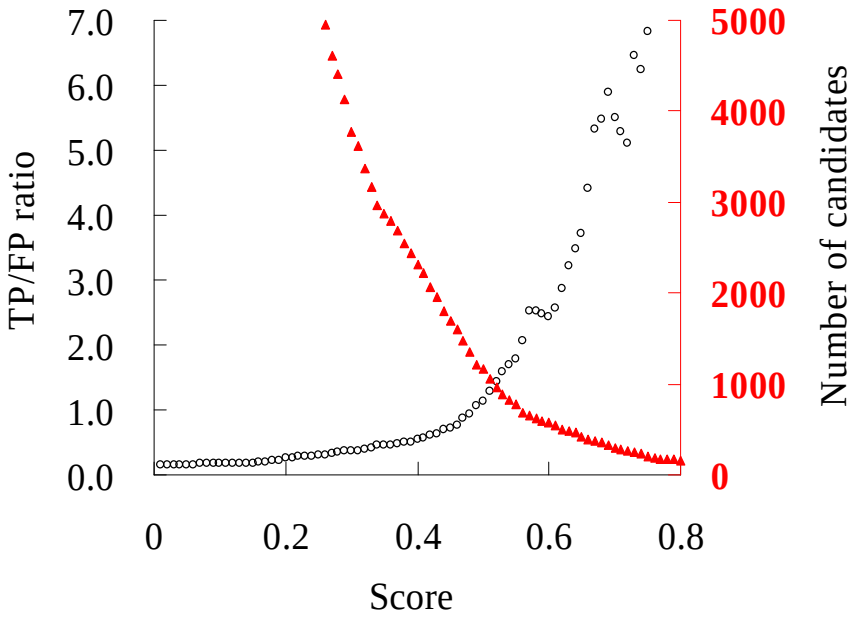
$$K_{AA'} = \sum_{i=1}^{M}(W_{Ai}W_{A'i}) \Big/ \sqrt{\left(\sum_{i=1}^{M} {W_{Ai}}^2\right) \times \left(\sum_{i=1}^{M} {W_{A'i}}^2\right)} \quad (3)$$

其中 *M* 功能是 Swiss-prot 資料庫中所有關鍵字的數目，W_{Ai} 與 $W_{A'i}$ 分別是關鍵字 i 對蛋白質 *A* 與 *A'*的加權比重。

結果與討論

(一) 評估蛋白質蛋白質交互作用預測準確性

我們用三種評估指標來驗證我們的方法的正確性，包括 TP/FP ratio、enrichment 及基因表現側寫。TP/FP ratio 定義為 A_h/F_h，其中 A_h 是我們預測的蛋白質對和已知會產生交互作用蛋白質對重複的數目；F_h 是我們預測的蛋白質對和已知不會產生交互作用蛋白質對重複的數目。Enrichment 定義為$(A_h/T_h)/(A/T)$，其中 T_h 是我們預測的蛋白質對數目；A 是已知會產生交互作用蛋白質對的數目，T 是全部蛋白質對的數目。

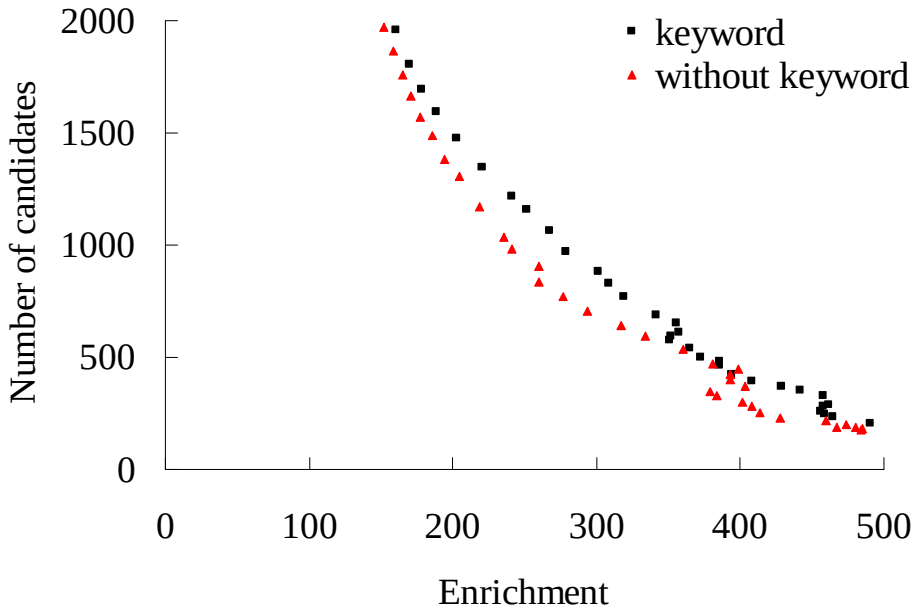


圖二. 連接分數的門檻臨界值與 TP/FP ratio、預測的交互作用蛋白質對數目的關係圖。連接分數的門檻臨界值設定越高的時候，TP/FP ratio 也會隨之上升。當連接分數的門檻臨界值設為 0.5 時，TP/FP ratio 等於 1.126 而預測的蛋白質蛋白質交互作用有 1158 對。

因為酵母菌(*S. cerevisiae*)是一個簡單而且被深入研究的動物模型，因此我們選擇在酵母菌系統中評估我們方法的正確性。科學家推測酵母菌細胞內約有 6000 種蛋白質，共可產生約 18000000 種蛋白質對。在 DIP 資料庫中共有 14779 對酵母菌的蛋白質蛋白質交互作用。另一方面，目前沒有實驗方法能夠驗證兩個蛋白質之間不會產生交互作用。Jansen *et al.* [5] 根據位於不同胞器的蛋白質不會產生交互作用的假設，預測 2599785 對不會發生交互作用的蛋白質對。

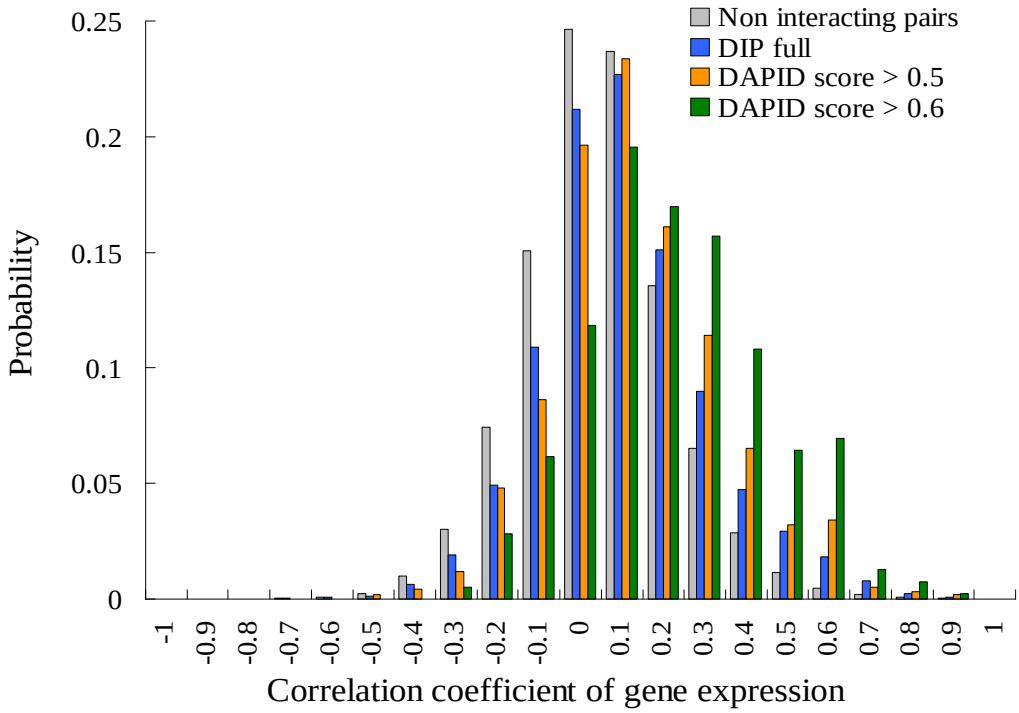
我們對連接分數設定門檻臨界值，若蛋白質對的連接分數大於臨界值，我們預測這兩個蛋白質會產生交互作用。當門檻臨界值設定為 0.5 時，則 A_h 等於 215，F_h 等於 191，T_h 等於 1158。在這個情形下，TP/FP ratio 等於 1.126 而 enrichment 等於 226。連接分數的門檻臨界值設定越高的時候，TP/FP ratio 也會隨之上升(圖二)，證明我們的連接分數

($I_{A'B'}$)對於描述蛋白質交互作用相當合適。若連接分數($I_{A'B'}$)門檻臨界值設定為 0.5，TP/FP ratio 等於 1.126，表示 *A'*與 *B'*產生交互作用的機率大於 50%，因此我們將門檻臨界值定為 0.5。在此情形下，我們從 845021 蛋白質對中，得到 101483 蛋白質蛋白質交互作用。當我們預測蛋白質蛋白質交互作用的數目為 1158 時，我們的準確率比隨機配對方法高出 226 倍(圖三)，另一方面圖三也顯示出功能註解相似性(公式 3)對於我們的方法的預測準確性，有很大的幫助。



圖三. Enrichment 與預測的交互作用蛋白質對數目的關係圖。比較計算同源性分數時，有考慮功能註解相似性與不考慮功能註解相似性的差異。

一般說來，兩個會產生交互作用的蛋白質，它們的基因表現側寫(gene expression profiles)也會非常相似[5]，我們根據此假說來驗證預測的蛋白質蛋白質交互作用是否合理(圖四)。我們利用 Hughes 等人[17]發表的酵母菌基因表現資料來計算兩個基因表現量的相關係數，測試四組蛋白質交互作用資料的基因表現相關係數分布情形，包括 DIP 資料庫中 14779 筆酵母菌的蛋白質蛋白質交互作用、2599785 筆不會有交互作用的蛋白質對，我們預測的酵母菌蛋白質交互作用其連接分數大於 0.5 (1158)以及連接分數大於 0.6 (575)。由圖四中顯示出我們所預測的蛋白質蛋白質交互作用其基因表現相關性明顯高於另外兩組蛋白質交互作用資料。T-test 的結果也顯示我們所預測的蛋白質對其平均基因表現相關性顯著高於不會有交互作用的蛋白質對。



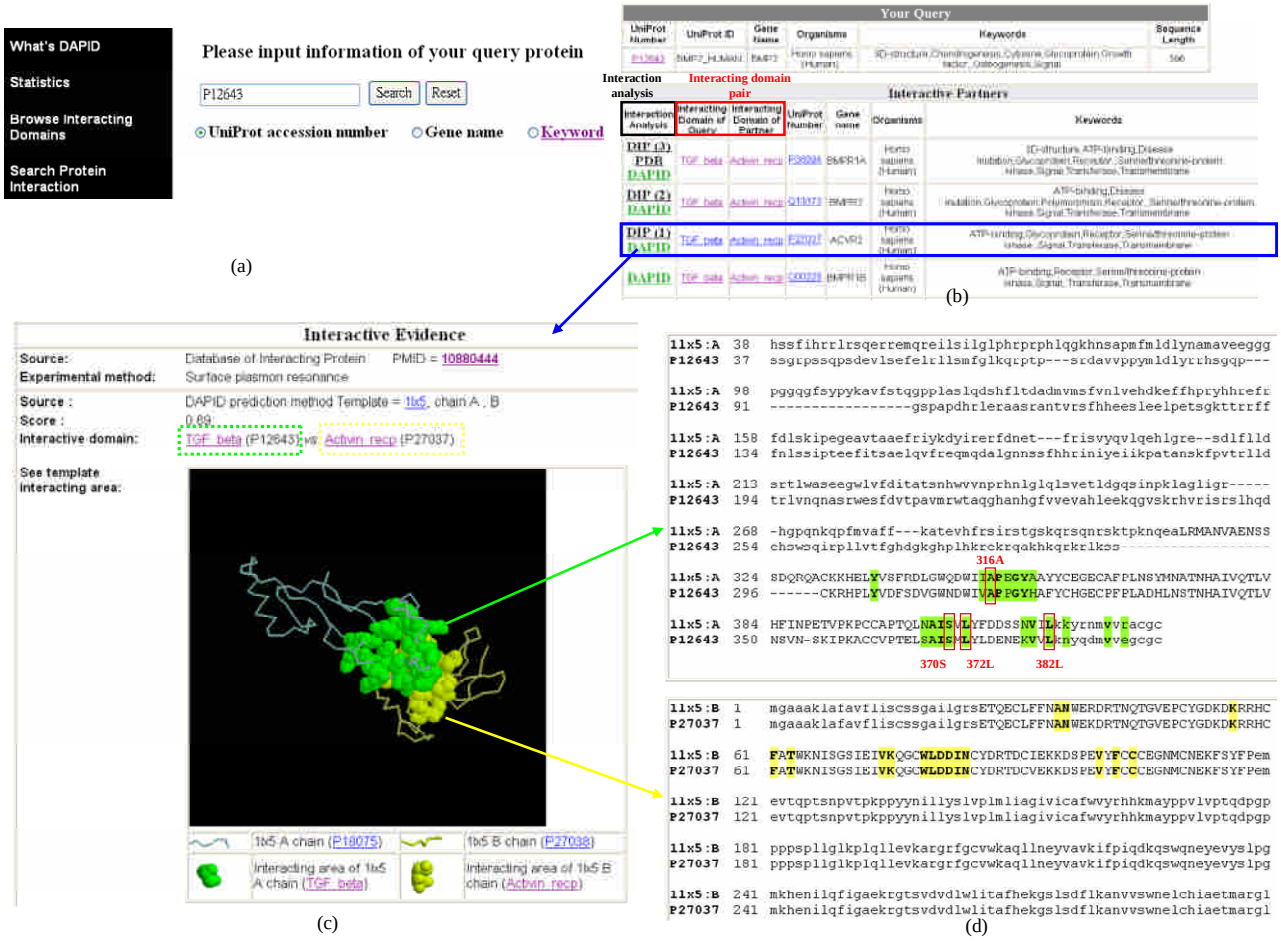
圖四. 四種蛋白質交互作用資料庫基因表現相關係數分布之比較，用我們的方法(DAPID)所預測的蛋白質蛋白質交互作用的基因表現相關係數明顯比另外兩組資料來的高。

(二) 功能區塊註解的蛋白質交互作用資料庫(DAPID)使用範例

使用者可以輸入 Swiss-prot 資料庫的序號(accession number)，基因名稱或者關鍵字來查詢 DAPID 資料庫。圖五(a)提供一個查詢 DAPID 資料庫的使用範例(查詢蛋白質的 Swiss-prot 資料庫的序號為 P12643，基因名稱為 bmp2)。我們的方法預測十個蛋白質會和人類的 bmp2 產生交互作用，其中三個蛋白質有紀錄於 DIP database 中，另外七個蛋白質之前並未被發現會和 bmp2 產生交互作用。搜尋之後的結果會出現 "交互作用分析(interaction analysis)"、 "結構交互作用功能區塊 (3D-interacting domains)"，"物種(Organisms)" … 等等蛋白質的基本資訊(圖五 b)。在交互作用分析欄位中，DAPID 顯示出這對蛋白質蛋白質交互作用是從 "結構功能區塊交互相似性"、DIP 資料庫或是 PDB 資料庫收集而得，而括號中的數字代表這對蛋白質蛋白質交互作用有幾對證據支持。如果這對蛋白質蛋白質交互作用是從結構功能區塊交互相似性(3d-domain interologs)推測而得，這個欄位會顯示成綠色(連接分數 > 0.6)或是橘色(0.6 > 連接分數 > 0.5)。

為了讓使用者觀察蛋白質之間分子層次上的交互作用情形，DAPID 提供結構交互作用功能區塊(圖五 c)以及接觸胺基酸(圖五 d)等資訊。例如從已知結構蛋白質複合體(PDB 序號 1lx5:A:B)預測人體內 BMP2 會與 ACVR2(Swiss-prot 資料庫的序號為 P27037)發生交互作用。DAPID 顯示許多資訊關於這對蛋白質蛋白質交互作用，如來源，文獻記載，連接分數以及結構交互作用功能區塊(TGF_beta 功能區塊和 Activin_recp 功能區塊)。已知結構蛋白質複合體的接觸胺基酸還會以圓球模型(SPACEFILL model)來呈現(圖五 c)。DAPID 利用 PSI-BLAST 分別比對 BMP2 與它的模板蛋白質(1lx5:A)以及 ACVR2 與它的模板蛋白質(1lx5:A)的序列。若接觸胺基酸的部分有比對成功會顯示綠色或黃色的標記，而且當兩的胺基酸相同時則會以大寫字體表示(圖五 d)。Kirsch 等人 [18] 利用實驗方法發現 BMP2 上有四個重要的胺基酸，316A, 370S, 372L 和 382L，對於和 ACVR2 產生交互作用相當重要，而我們也成功的預測出這些重要的胺基酸(圖五 d)。

COPII coat 在內質網(endoplasmic reticulum)運送物質的網絡中扮演重要的腳色，而它由 Sec23p/Sec24p, Sec13p/Sec31p 以及一個小的 GTPase 所組成 [19]。我們預測出 8 個蛋白質會和酵母菌的 Sec23p (Swiss-prot 序號為 P15303)發生交互作用，其中三個蛋白質(Sec24p, Sfb2p, and Sfb3p)於 DIP 資料庫中有紀錄，一個蛋白質(Sar1p)會和 Sec23p 形成蛋白質複合體，一個蛋白質(Sec23p)在 MIPS 資料庫 [3] 中有紀錄，而沒有實驗紀錄餘下三個蛋白質(Arf1p, Arf2p, and Arf3p)會和 Sec23p 產生交互作用。從分析蛋白質複合體(PDB 序號 1m2o:A:B)發現 Sec23_trunk 功能區塊會與 Arf 功能區塊產生交互作用。之前也有科學家研究指出 ARF 家族和 Sec23p 都和細胞內微泡運輸系統(vesicular transport pathway)[20] 有關。因此我們預測 Sec23p 會與 Arf1p, Arf2p 以及 Arf3p 產生交互作用是十分合理的。



圖五. DAPID 資料庫的使用範例。在此範例中我們使用 P12643(基因名稱為 bmp2)作為查詢蛋白質。(a) DAPID 資料庫可以用 Swiss-prot 資料庫的序號(accession number)，基因名稱或者關鍵字來查詢。(b)會與 BMP2 產生交互作用的蛋白質。(c) 從已知結構蛋白質複合體 (PDB 序號 1lx5:A:B) 推測 BMP2 會與 ACVR2 發生交互作用。BMP2 利用結構交互作用功能區塊 TFT_beta(綠色)與 ACVR2 的結構交互作用功能區塊 Activin_recp(黃色)發生交互作用。接觸胺基酸的部分用圓球模型來表現。(d) BMP2 與其模板蛋白質(1lx5:A)和 ACVR2(1lx5:B)與其模板蛋白質利用 PSI-BLAST 作序列比對的結果。位在結構交互作用功能區塊內的胺基酸已大寫表示，接觸胺基酸的位置用綠色(BMP2)與黃色(ACVR2)來表示，若兩個胺基酸相同的話，則將字體加粗。BMP2 上有四個重要胺基酸對於和 ACVR2 發生交互作用，分別是 316A, 370S, 372L 和 382L，在圖中以紅色方框標示。

(三)未來研究

已知結構蛋白質複合體(A 與 B)的同源蛋白質(*A'*與 *B'*)並不一定會發生交互作用。因此我們未來將結合蛋白質複合體的 empirical potential[7,8] 及其它種類型的功能註解(例如 GO 資料庫 [21])來加強我們的計分程式。另外，我們從已知結構蛋白質複合體中定義的結構交互功能區塊有時候和 Pfam 所定義的功能區塊有所差異，在下一步研究中，我們將會涵蓋其它資料庫的功能區塊定義，像是 SMART 和 ProDom。另外，我們也希望整合 DIP、BIND、STRING 等其它蛋白質資料庫到 DAPID 中，提供使用者更完整的蛋白質交互作用網路。

第二部份: 蛋白質與蛋白質結合位置之預測

前言

進入後基因體時代，對功能性基因體學的研究大量增加，與之相輔的蛋白質-蛋白質間交互作用與代謝路徑(metabolic pathway)等相關研究也愈形重要。蛋白質-蛋白質交互作用與蛋白質調控機制及其生物功能密切相關，並構成維持生命所必需的代謝路徑。在訊息傳遞路徑(signal transduction pathway)中需要靠蛋白質-蛋白質交互作用來傳遞訊息。而有些蛋白質必須要和其它的蛋白質形成複合體，才能執行正常的功能。蛋白質交互作用能直接影響生物體內的生化代謝反應，要了解生物體內的所有反應就要先了解每一個蛋白質-蛋白質之間的交互作用。利用傳統實驗yeast two-hybrid以及protein-complex purifications來研究蛋白質-蛋白質交互作用是相當普遍的作法，但是生化實驗耗時耗力。隨著已經證實有交互作用的蛋白質複合體(protein complex)快速增加，並且蛋白質複合體結晶結構越來越多，利用電腦結合資料庫將已知的資料分析，建構出可能的蛋白質-蛋白質交互作用配對，並利用預測蛋白質可能結合位置(potential binding sites)與計算蛋白質分子鉗合(protein docking)的運算模型，將有助於我們篩選出可靠的蛋白質-蛋白質交互作用配對，了解兩蛋白質間可能以何種方式進行結合。

現階段研究蛋白質交互作用的方法有很多，而分析蛋白質的結構資訊是其中一種了解交互作用的方法，在蛋白質結構上，並不是所有的蛋白質表面都有可能參與交互作用，只有一些特定的區域才會參與交互作用[22]，Neuvirth [23]等人透過分析這些特定的區域，得到以下的發現，從胺基酸的角度去分析，可以發現CYS、TYR、HIS、MET出現的機率最大，而THR、LYS、PRO、ALA、GLU是較不容易出現的，從原子的角度去分析，可以得到芳香環(aromatic ring)上的原子特別容易出現，從二級結構上去分析，可以發現loop以及beta strand出現機率較高，而alpha helix則不容易出現，在我們預測蛋白質可能結合位的方法論中，也參考了這些特性，發展了一套新的原子分類，應用在Fernandez-Recio [24] 等人的方法論上，在該論文中假設疏水性強的位置極有可能是蛋白質的結合位，並且利用計算蛋白質親疏水的特性來預測蛋白質的可能結合位。

在本計劃中，我們利用Fernandez-Recio計算蛋白質親疏水的方法，導入自己發展的原子分類以及利用GA進行最佳化的方式，發展了一套自動化的預測程式，為了要比較預測程式的可靠性，我們針對Neuvirth及Fernandez-Recio所使用的testing set做預測，在我們預測的結果上，56個蛋白質中的平均預測準確度為47.4%，雖然略低於Neuvirth的51.7%準確度，但是在預測結果的平均涵蓋率(涵蓋結合位的面積比)上，我們為15.7%，略優於Neuvirth的12.9%；而在Fernandez-Recio測試的50個蛋白質中，我們的平均預測準確度是40.5%，平均涵蓋率是34.9%，而Fernandez-Recio的平均預測準確度是37.8%，平均涵蓋率則未列出，詳細的比較結果將會列在結果與討論的內容中，在此不再贅述。

由上可知，我們所發展的預測程式在準確度上已經具備足夠的可信度，我們也將在

近期之內發表已有的成果，未來我們將利用這個預測程式，朝比較兩個蛋白質間的結合位所包含的結構資訊以及物化特性來判斷是否產生交互作用這部份做深入研究。

研究目的

從第一部份的研究結果中，我們可以得到數量龐大的蛋白質交互作用配對預測，接下來我們希望進一步的篩選出最有可能的交互作用配對，為了達到這個目的，我們透過統計胺基酸、二級結構在蛋白質表面以及在結合位(binding site)上的機率，並且加入疏水性的概念，來預測蛋白質可能的結合位，最終希望能從比較兩個蛋白質間的結合位所包含的結構資訊以及物化特性來判斷兩個蛋白質是否會產生交互作用。在本報告中，我們將會詳細說明預測蛋白質可能結合位的方法，而比較兩個蛋白質間的結合位所包含的結構資訊以及物化特性將是我們未來繼續努力的目標。

研究方法

在本計劃中，我們參考了 Juan Fernandez-Recio 計算親疏水性的方法，利用不同的原子型態(atom type)定義，將計算親疏水性的概念擴展成以知識為基礎(knowledge-based)的計算蛋白質結合位特定原子以及二級結構喜好出現的概念，我們定義的原子型態就可以包含物化特性以及結構資訊，在物化特性上定義了 14 種原子型態，在結構資訊上，加入了 4 種二級結構的資訊， 並且利用我們實驗室發展的 GEMDOCK[25] 這套軟體核心的演化式演算法(evolution strategies)作為最佳化演算法(optimal algorithm)，透過最佳化訓練資料(training data)的方式，合理的求出每一種原子型態所具備的參數值，並且應用在 56 個蛋白質的測試資料(testing data)上，以及與其他相關論文比較測試的結果，這些內容都將在下面一一說明。

Training Data Set

我們使用了 Protein-Protein Docking Benchmark [26] 中的 52 個已知的蛋白質複合體(protein complex)，包括了 Enzyme-Inhibitor (22 個)、Antibody-Antigen (19 個)以及 Others (11 個)，並且將其對應的 104 個蛋白質單體(protein monomer)拿來當做訓練資料，而這些單體都可藉由結構比對(structure alignment)回對應的蛋白質複合體上來定義出結合區的位置，104 個對應的蛋白質單體中，因為有 26 個蛋白質找不到對應，所以是以複合體的結構代表。

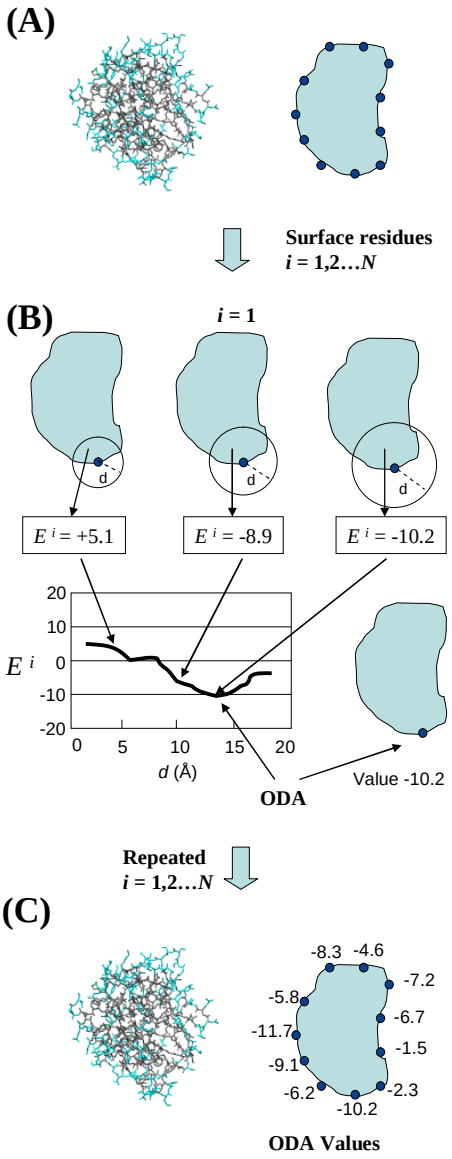
Testing Data Set

我們使用了 56 個蛋白質做預測，這是根據 Neuvirth 以及 Fernandez-Recio 使用的測試資料而來，其中 Neuvirth 使用了 56 個蛋白質， Fernandez-Recio 使用了 50 個蛋白質(Unbound database B)做預測，值得一提的是 Neuvirth 使用的蛋白質沒有包含任何的 Antibody-Antigen，這種資料跟我們使用的訓練資料組成有所不同，所以在測驗的結果

上也會受到影響。

Method

整個研究的流程圖(圖六)如下所示，總共由六個部份組成，**(A)**Surface and binding site residue、**(B)**Calculate ODA value and Z score、**(C)**Select Z score and smooth results、**(D)**Equations of ODA and Z score、**(E)**Optimal function、**(F)**Atomic parameters，接下來我們將針對每一個部份做詳細的說明。



(D)

$$E^i = \sum_i (\delta(Atom_j^i) \times ASA_j^i + Ss(Atom_j^i))$$

$$Z^i = (E^i - \overline{E}) / std(E)$$

(E)

$$Fit = \sum_k^N (1 - Rate_k + Success_k)$$

(F)

Atomic Chemical-Physical Parameters

σ	Atom Type
-41.732	C aliphatic backbone and C_β
-268.563	C aromatic
127.255	N uncharged
206.239	N charged in Lys^+
6.211	N1, N2 in Arg^+
2.662	O hydroxyl
-12.375	O carbonyl
102.601	O^- in Glu, Asp
-537.241	S in SH
-850.644	S in Met
33.616	C aliphatic side-chain
30.062	O double-bond in Asn and Gln
-254.692	N in His
70.029	C side-chain in Ala and Pro

Atomic Secondary Structure Parameters

SS	Atom Type
-7.738	Helix
-5.241	Strand
-9.485	Turn
-8.785	Loop

圖六. 預測蛋白質-蛋白質結合位置之流程示意圖，總共有六個步驟，**(A)**Surface and binding site residue、**(B)**Calculate ODA value and Z score、**(C)**Select Z score and smooth results、**(D)**Equations of ODA and Z score、**(E)**Optimal function、**(F)**Atomic parameters。

(A) Surface and binding site residue

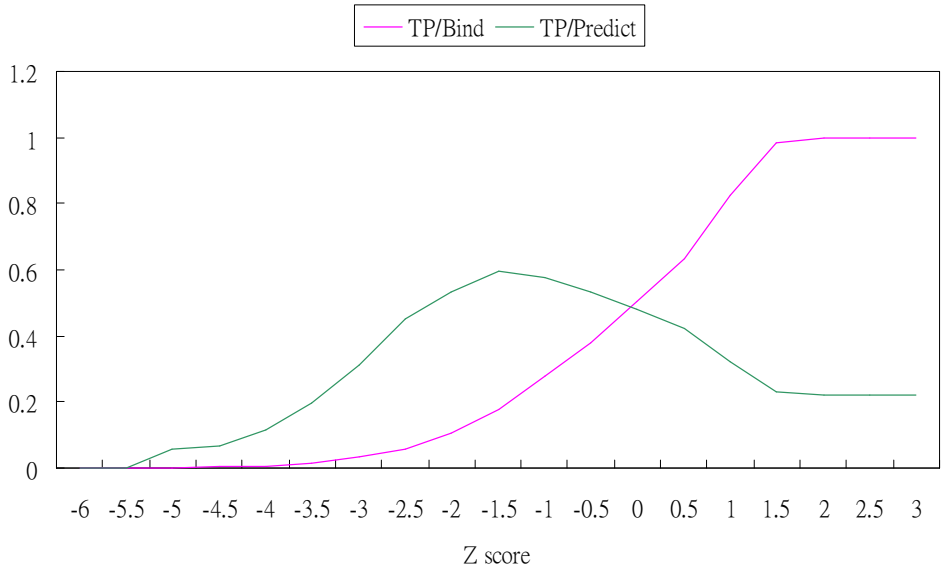
首先，我們必須要定義出蛋白質上哪些胺基酸是要拿來考慮的，因為只有蛋白質的表面會接觸到其他蛋白質，所以我們就以蛋白質表面的胺基酸當成要預測的目標，在此我們定義蛋白質上的胺基酸在 DSSP 的計算中曝露的面積超過 10%則為表面胺基酸。接著，我們要定義出結合位的胺基酸有哪些，在此我們是以距離為根據，當兩個蛋白質在交互作用時，彼此間的原子距離小於 6Å 的胺基酸，則定義為結合位的胺基酸。在訓練資料的 104 蛋白質中，表面的胺基酸平均為 128 個，結合位的胺基酸平均為 23 個。

(B) Calculate ODA value and Z score

我們在此步驟中要計算每一個表面胺基酸的 ODA value 並轉換成 Z score，最後依照 Z score 的大小來預測是否為結合位的胺基酸，詳細的公式將在(D)部份說明，在此只說明方法論的部份。首先我們以表面胺基酸的 C_B 原子為球心，畫一個半徑 20Å 的球，接著每隔 1Å 將其等分，最後得到半徑 1Å、2Å、3Å…20Å 的 20 個球，接下來我們計算每一顆球所包含到的每一顆原子，將原子的曝露面積乘上最佳化後的 atomic salvation parameter，並把它們加總起來，代表這一顆球的 ODA value，重複操作後得到 20 顆球的 ODA value，接著再取 ODA value 中的最小值作為該表面胺基酸所具有的 ODA value。依據這樣的方式將所有的表面胺基酸的 ODA value 計算出來，接下來再將這些 ODA value 轉換成 Z score，如此一來可以確保每一個表面胺基酸所具有的值不會差異過大，這在(C)中拿來預測哪些表面胺基酸是結合位時可以得到較佳的結果。

(C) Select Z score and smooth results

為了選擇合理的 Z score 大小來預測結合位的胺基酸，我們將自訓練資料組中求得的平均預測準確度(TP / Predict： 預測胺基酸是結合位且確實為結合位 / 所有預測是結合位的胺基酸總數)、平均涵蓋率(TP / Bind： 預測胺基酸是結合位且確實是結合位 / 結合位胺基酸的總數)與 Z score(-5.0 ~ +3.0)繪成下圖(圖七)，由圖可知，當 Z score 等於-1.5 的時候，平均預測準確度可以得到最高值 59.67%，而平均預測涵蓋率為 17.83%，在預測的成功率(預測準確度大於等於 50%的蛋白質稱為成功的預測)上則為 63.46% (66/104)，當 Z score 等於-1.0 的時候，平均預測準確度可以達到 57.80%，而平均預測涵蓋率為 27.76%，在預測的成功率上則為 65.38% (68/104)，因此我們就取 Z score 小於等於-1.0 的表面胺基酸當成預測為結合位的標準，在這個情況下我們可以得到最好的預測成功率以及不錯的平均預測準確度和平均涵蓋率。



Z score	-5	-4.5	-4	-3.5	-3	-2.5	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2	2.5	3
TP/Bind	0	0	0.01	0.01	0.03	0.06	0.11	0.18	0.28	0.38	0.5	0.64	0.82	0.99	1	1	1
TP/Predict	0.06	0.07	0.12	0.2	0.31	0.45	0.53	0.6	0.58	0.53	0.48	0.42	0.32	0.23	0.22	0.22	0.22

圖七. 不同的 Z score 對於預測結果的影響，粉紅色是平均預測準確度(TP / Predict : 預測胺基酸是結合位且確實為結合位 / 所有預測是結合位的胺基酸總數)、綠色是平均涵蓋率(TP / Bind : 預測胺基酸是結合位且確實是結合位 / 結合位胺基酸的總數)與 Z score(-5.0 ~ +3.0)。

當我們利用 Z score 選出預測的表面胺基酸後，有可能會產生一些不合理的預測，例如預測的胺基酸對於構成結合位的數量不足(群組在一起的胺基酸個數小於 3)，或是預測的胺基酸太多，導致結合位大到超出平均值太多，因此在結合位的預測上，我們的程式會先將預測的胺基酸根據距離分為群組，詳細的步驟如下：首先我們先從 Z score 最小的胺基酸開始，判斷它與其他的預測為結合位的胺基酸之間的距離，只要胺基酸之間有任何一個原子距離小於 3 Å，我們就會將其歸為同一群，一直到找不到距離小於 3 Å 的胺基酸為止，接下來再從餘下的預測結果中挑選 Z score 最小的胺基酸，重覆之前的步驟，一直到所有的預測胺基酸都分完群，然後再去判斷每一群的胺基酸個數，如果小於 3 個則將這一群從預測的結果中去除，直到最後剩下來的群組在胺基酸的個數上都大於等於 3，以這些胺基酸當成預測的結果。用這樣子的方式可以去除掉孤立的預測點，增加預測的準確度。

(D) Equations of ODA and Z score

我們在計算每一個胺基酸所擁有的 ODA value 時，所使用的公式如下所示:

$$E^i = \min(\sum_d [\sum_j (\delta(Atom_j^i) \times ASA_j^i + Ss(Atom_j^i))] \qquad (4)$$

i 是要預測的胺基酸，*j* 是以第 *i* 個胺基酸的 C_B 為球心及以半徑為分 *d* (1, 2 …, 20Å) 的範圍之內所包含的原子，δ 是根據(F)所定義的原子物化特性參數，*ASA* 是原子的曝露面積，*Ss* 也是根據(in the following Section F)所定義的原子二級結構參數。當我們計算完半徑 1~20 Å 的 20 個 *E* 值之後，我們定義此群 *E* 值中最小者為此胺基酸的 ODA value。

當我們計算完所有的胺基酸的 ODA value 後，為了避免不同蛋白質間的 ODA value 差異過大，所以我們將這些 ODA value 轉換成 Z score，轉換的公式如下所示:

$$Z^i = (E^i - \overline{E}) / \sigma(E) \qquad (5)$$

i 是要預測的胺基酸，$\overline{E}$ 是所有 E 的平均，σ(.)是 *E* 標準差函式。

(E) Optimal function

我們是利用 GEMDOCK 的核心演化式演算法來調整 Atomic parameters 的數值大小，並且利用下面的公式來得到最佳化的結果。公式如下所示:

$$Fit = \sum_{k}^{N} (1 - Rate_k + Success_k) \qquad (6)$$

N 是訓練資料的蛋白質總數，$Rate_k$ 是指第 *k* 個蛋白質的預測準確度，當預測準確度超過 50%的時候，$Success_k$ 的值為-1，當預測準確度介於 0%到 50%時，$Success_k$ 的值為 0，當預測準確度為 0%，$Success_k$ 的值為 0.5，最後利用 GEMDOCK 的核心演化式演算法來調整 Atomic parameters 的數值大小（Fit 的值越小越好）。利用上面的公式可以使得 GEMDOCK 在調整參數時不止考慮到準確度的提升，也會考慮到蛋白質預測正確的數量，讓訓練資料的結果更好。

(F) Atomic parameters

我們根據相關的文獻以及知識，定義出了下列兩種原子型態，一種是原子的物化(Chemical-Physical)特性，另一種是二級結構(Secondary Structure)特性(表二)，這與我們參考的 Fernandez-Recio[24] 所使用的原子型態最大的不同在於我們定義的原子型態不僅僅是將親疏水性表現出來，也根據統計的資料將以知識為基礎的蛋白質結合位特定原子以及二級結構喜好出現的概念表現出來，這可以將原本只預測疏水性強的位置為蛋白質結合位的想法改進成為預測蛋白質結合位喜好出現的原子型態以及二級結構，這也可以避免某些蛋白質(ex. 7RSA:RNA binding protein)在功能上就是以親水性或是中性的位置作為結合位的例子。

表二. 原子型態參數，包含原子的物化(Chemical-physical)特性以及二級結構(Secondary Structure)特性

Atomic Chemical-Physical Parameters	
δ	**Atom Type**
-41.732	C aliphatic backbone and C_B
-268.563	C aromatic
127.255	N uncharged
206.239	N charged in Lys^+
6.211	N1, N2 in Arg^+
2.662	O hydroxyl
-12.375	O carbonyl
102.601	O^- in Glu, Asp
-537.241	S in SH
-850.644	S in Met
33.616	C aliphatic side-chain
30.062	O double-bond in Asn and Gln
-254.692	N in His
70.029	C side-chain in Ala and Pro

Atomic Secondary Structure Parameters	
SS	**Atom Type**
-7.738	Helix
-5.241	Strand
-9.485	Turn
-8.785	Loop

結果與討論

在結果與討論中，我們將先說明訓練資料的結果，並且比較我們與 Juan Fernandez-Recio 不同的原子型態定義對於結果的影響，接下來是利用我們的程式在測試資料上的預測結果，最後將針對預測結果不好的例子做說明。

Training Set Results:

我們根據前述的方法論，套用 Fernandez-Recio 所使用的 10 種原子型態以及我們所發展的 18 種原子型態，針對訓練資料的 104 個蛋白質去做最佳化，得到的結果如表三所示，在預測的成功率上，我們所定義的原子型態以及 Fernandez-Recio 所定義的原子型態在 Enzyme-inhibitor 的成功率都是最高的，而在 Antibody-antigen 上的表現比較差，在 Others 的表現是最不好的，這可以說明 Enzyme-inhibitor 的蛋白質在結合位的表現是比較一致的。而我們定義的原子型態在預測準確度的表現上也優於 Fernandez-Recio 所定義的原子型態，最主要的原因從圖八中可以得知， Fernandez-Recio 所定義的原子型態可能因太偏重疏水性的計算，導致預測的結果以芳香族居多，而忽略了其他的影響力，而我們定義的原子型態因為是從物化特性以及結構方面來看，則可以避免此類問題發生。

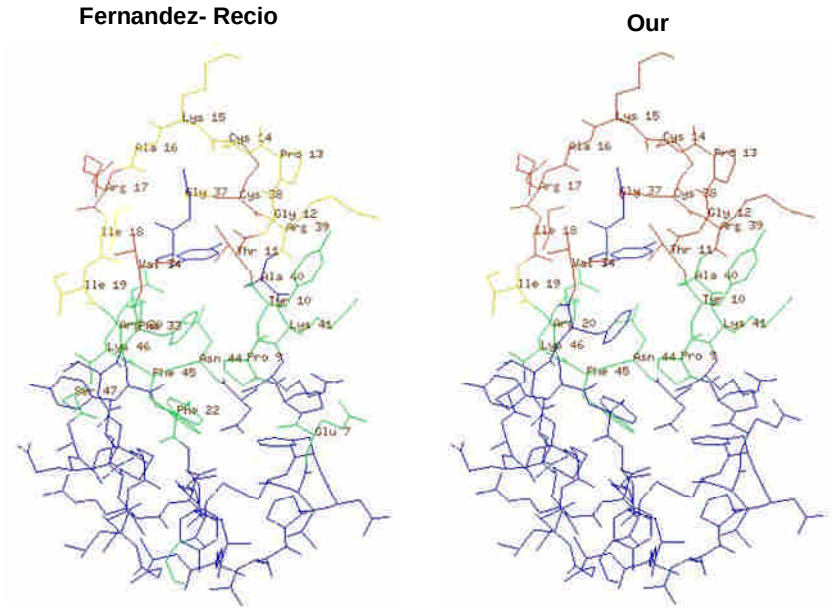
表三. 不同的原子型態在 104 個蛋白質中預測的結果

Protein type	Success [a]		TP/Prediction [b]		TP/Bind [c]	
	Our	Fernandez-Recio [24]	Our	Fernandez-Recio	Our	Fernandez-Recio
Enzyme-inhibitor	79.54% (35/44)	61.36% (27/44)	67.11%	54.2%	27.91%	39.15%
Antibody-antigen	60.52% (23/38)	42.11% (16/38)	54.34%	38.65%	33.16%	37.22%
Others	45.45% (10/22)	36.36% (8/22)	45.15%	32.81%	18.11%	25.19%

[a]:預測成功率，預測準確度大於等於 50%的蛋白質稱為成功的預測。

[b]:預測準確度，預測胺基酸是結合位且確實為結合位 / 所有預測是結合位的胺基酸總數。

[c]:預測涵蓋率，預測胺基酸是結合位且確實是結合位 / 結合位胺基酸的總數。



圖八. 蛋白質 2KAI，左邊是使用 Fernandez- Recio [24] 定義的原子型態所預測的結果，右邊是我們所預測的結果，紅色是預測正確的結合位，黃色是未預測到的結合位，綠色是預測錯誤的結合位，藍色是非結合位的部份。由圖可知，Fernandez- Recio 的結果幾乎都預測在芳香族的部份，最上方的 Lys15、Ala16、Cys14 等真正屬於結合位的部份則沒有預測到，而我們定義的原子型態則可以避免這樣的情況發生。

Testing Set Results:

我們針對了 56 個蛋白質做測試，並且比較了相關兩篇論文的結果，詳細的資料如表四所示，我們的平均預測準確度為 47.4%，比 Neuvirth [23] 的 51.7%稍微低了一點，但是比 Fernandez-Recio[24] 的 37.84%要好，而在平均涵蓋率上，我們是 15.7%，高於 Neuvirth 的 12.96%，而 Fernandez-Recio 則未列出涵蓋率，在預測成功率上，我們在 56 個蛋白質中成功的預測了 28 個，Neuvirth 在 56 個蛋白質中成功的預測了 35 個，Fernandez-Recio 在 50 個蛋白質中成功預測了 19 個，而且我們在預測失敗的蛋白質結果中，有 4 個蛋白質是超過 40%的準確度但還未到 50%，有 5 個蛋白質是超過 30%的準確度但還未超過 40%。由此可知，我們的預測程式在準確度上已經有不錯的水準，而在涵蓋率上則表現的更好。

表四. 56 個蛋白質預測結果與相關的兩篇論文(ProM [23] and ODA [24])之比較

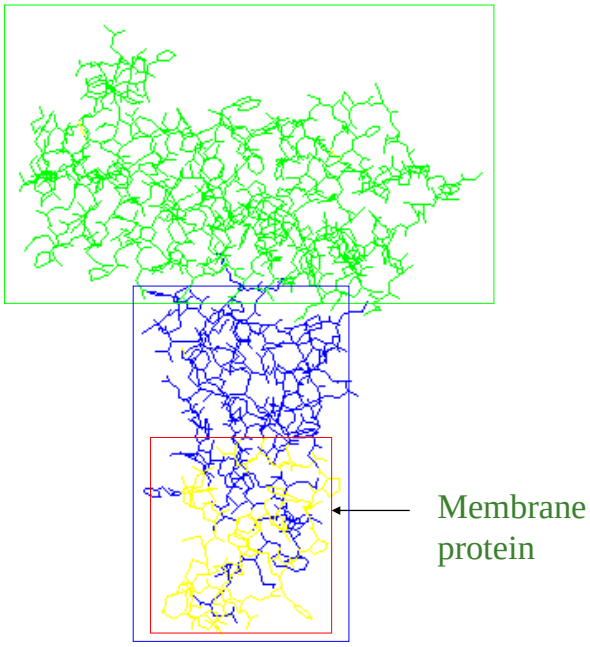
PDB	Coverage rate		Predicting rate			PDB	Coverage rate		Predicting rate			PDB	Coverage rate		Predicting rate		
	Our	ProM	Our	ProM	ODA		Our	ProM	Our	ProM	ODA		Our	ProM	Our	ProM	ODA
1a19A	0.263	0.29	1	1	0.89	1ez3A	0.167	0.06	0.5	1	---	1nos_	0	0	0	0	---
1a2pA	0.125	0.19	0.75	0.9	*	1eza_	0.031	0	0.067	0	*	1pco_	0	0.12	0	0.6	0.18
1a5e_	0.242	0.1	0.889	0.88	0.2	1eztA	0.136	0.13	0.333	0.54	0	1pne_	0.065	0	0.5	0	*
1acl_	0.455	0.14	0.625	0.24	0	1f00I	0	0	0	0	*	1poh_	0.211	0	1	0	0.8
1ag6_	0.214	0.16	1	0.7	*	1f5wA	0.154	0.06	0.667	1	1	1ppp_	0.083	0.3	0.333	0.91	0.3
1aje_	0.346	0.3	0.75	0.72	0.4	1fkl_	0.261	0.2	1	1	*	1qqrA	0.25	0.32	0.444	0.85	---
1ajw_	0.273	0.24	0.333	0.73	*	1flzA	0.167	0.19	0.556	0.52	0.9	1rgp_	0.042	0.05	0.083	0.5	1
1aueA	0.2	0.35	0.4	0.9	0.8	1fvhA	0.22	0	0.478	0	0.71	1selA	0.174	0.27	0.8	0.61	0.2
1avu_	0.368	0.29	0.7	1	0.7	1g4kA	0.147	0.21	1	0.78	1	1vin_	0.367	0	0.917	0	1
1aye_	0.158	0.24	0.6	0.54	---	1gc7A	0.09	0.06	0.429	0.78	*	1wer_	0.108	0	0.25	0	*
1b1eA	0.051	0.24	1	0.69	0.7	1gnc_	0	0.02	0	0.03	0.1	1xpb_	0.323	0	0.769	0	*
1bip_	0.047	0.27	0.667	1	1	1hh8A	0	0.02	0	0.5	*	2bnh_	0.325	0.04	0.65	1	*
1ctm_	0.207	0.12	0.316	1	0.63	1hplA	0.118	0.03	0.125	0.07	*	2cpl_	0.471	0.23	0.889	0.76	*
1cto_	0	0.29	0	0.36	---	1hu8A	0	0.02	0	0.05	0	2f3gA	0.261	0.12	1	1	*
1cye_	0	0	0	0	0.29	1iob_	0.04	0.03	0.25	0.31	*	2nef_	0.053	0.24	0.25	0.57	0.9
1d0nA	0.054	0.03	0.133	0.67	---	1j6zA	0.452	0	0.636	0	1	2rgf_	0.037	0.05	0.333	0.2	1
1d2bA	0.074	0.31	0.667	0.92	1	1jae_	0.483	0.13	0.737	0.5	1	3ssi_	0.125	0.24	0.667	1	*
1ekxA	0.04	0	0.167	0	0.22	1lba_	0.048	0.24	0.167	0.6	*	6ccp_	0	0	0	0	*
1ex3A	0.286	0.29	0.727	1	1	1nobA	0	0.03	0	0.07	*	Avg.	0.157	0.1296	0.474	0.517	0.3784

粉紅色、灰色和藍色的部份是預測準確度未超過50%的蛋白質

＊: No ODA hot spots

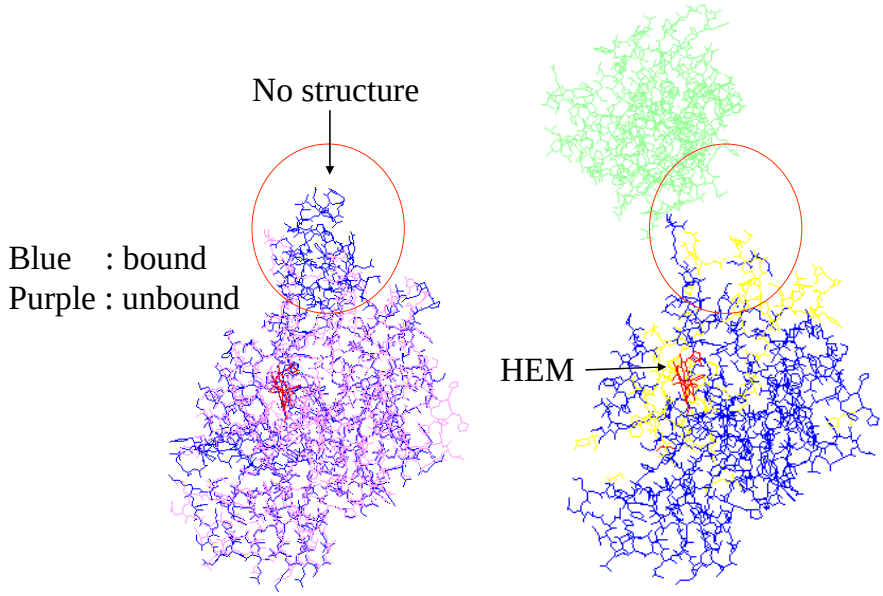
--- : No PDB prediction

在預測的結果中，有幾個蛋白質是完全沒有預測到結合位的，在此以 1CTO 以及 1NOS 做為例子說明，如圖九所示，1CTO 是屬於膜蛋白的一種，所以它是一端插在細胞膜上而另一端與其他蛋白質產生交互作用，在圖中黃色的部份是我們預測的結合位，雖然我們沒有預測到與綠色的蛋白質產生交互作用的結合位，但是我們預測到的是 1CTO 插在細胞膜上的部份，仍然屬於結合位的一種。



圖九. 蛋白質 1CTO，綠色是 1CD9:A，藍色是 1CTO，黃色是我們預測的結果，1CTO 是屬於膜蛋白的一種，所以黃色的部份是 1CTO 插在細胞膜上的一端，仍然屬於結合位的一種。

我們預測的蛋白質 1NOS 的結果如圖十所示，圖的左半部是以 1NOS 與其對應的複合體結晶 1NOC:A 做結構比對，從圖上可知，1NOS 本身在結合位的結晶就不完全，所以預測出來的結果自然不好，而 1NOC:A 中原本有一個 HEM，我們在預測 1NOS 時並未考慮 HEM，而預測結果如圖右半部所示，我們將 HEM 的位置預測成了結合位，所以雖然沒有辦法預測到真正與蛋白質交互作用的部份(因為結合位的結晶不完全)，但是我們仍舊可以預測到其他種類的結合位。



圖十. 蛋白質 1NOS，左半部的藍色是複合體 1NOC:A，紫色是 1NOS，右半部的藍色是 1NOS，綠色是 1NOC:B，黃色是我們預測結果，紅色是 HEM 的結構，由圖可知 1NOS 本身在結合位的結晶就不完全，所以預測出來的結果自然不好，但是我們將 HEM 的位置預測成了結合位，所以雖然沒有辦法預測到真正與蛋白質交互作用的部份(因為結合位的結晶不完全)，但是我們仍舊可以預測到其他種類的結合位。

未來研究

我們將持續發展改進預測程式的準確度，不只是從結構或是物化特性上來做預測，並希望能加入同源(homologous)的觀念來增加預測的準確度。另外我們也將整合預測程式與之前發展的蛋白質嵌合(docking)程式，希望透過預測的結合位實際做嵌合，由最後嵌合的結果是否合理，來判斷兩個蛋白質是否可能產生交互作用，最終的目的是幫助第一部份的研究，篩選出最有可能的蛋白質交互作用。

第三部份 :蛋白質與蛋白質嵌合之預測

前言

蛋白質-蛋白質之間的交互作用(protein-protein interaction)直接影響生物體內的生化代謝反應，是了解生物體內生化反應的鑰匙，但以傳統實驗方式研究蛋白質-蛋白質交互作用卻極耗時耗力。隨著已證實有交互作用的蛋白質複合體(protein complex)越來越多，且蛋白質複合體結晶結構快速增加，利用電腦分析已知資料，建構可用來預測蛋白質可能結合位(potential binding sites)與計算蛋白質分子鉗合(protein docking)的運算模型將有助於我們了解兩蛋白質間可能以何種方式進行結合[27-31]。以電腦模擬預測蛋白質可能結合位的方法主要包含兩大主軸：能將正確分子鉗合構形(docking conformation)與不正確分子鉗合構形區分開來的計分函式(scoring function)，以及快速搜尋構形解空間(conformational space)的演算法[32]。我們發展了一套以知識為基礎的計分函式(knowledge-based scoring function)，並配合估算蛋白質-蛋白質交互作用的演化式嵌合方法(evolutionary-based docking approaches)。我們將此模型應用在實際的蛋白質分子鉗合測試中，發現目前發展出的計分函式可區分正確與不正確的構形，這有利於我們進一步的研究。我們利用GEMDOCK[33-35]核心的演化式演算法(evolution strategies)作為搜尋演算法(search algorithm)，發展出一套新的預測蛋白質-蛋白質交互作用的程式。而為了加快程式的執行速度，我們亦加入兩個減少計算量的技巧，使得程式執行時間大幅降低為原本的十分之一。

研究目的

預測蛋白質交互作用乃是進入蛋白質體學時代的一項重要研究課題。生物體內的繁複生化運轉機制與代謝調控絕大多數由蛋白質交互作用網絡所控制與影響，預測蛋白質交互作用活性區將有利於藥物開發與預測潛在蛋白質交互作用。電腦模擬是研究蛋白質-蛋白質交互作用的有效方法。利用演算方法建構的模型(model)能幫助我們預測兩蛋白質間的互動關係，再加上分子複合體的工作模型(working module)，此二者對探討巨分子間識別以及生物化學機制都是非常重要的。欲準確預測蛋白質-蛋白質交互作用需要兩個要素，計分函式(scoring function)及快速有效率的構形解空間(conformational space)搜尋演算法。蛋白質產生交互作用的區域擁有複雜的物理化學性質，一一分析所有的性質將造成計算量激增，而分析蛋白質產生交互作用區域又可分為以原子特性為基礎與以胺基酸特性為基礎[36,37]兩類，利用原子特性來建構的計分函式較為精確，但是也相對需要更多計算時間。準確率與速度無法兼得，是以我們希望發展一種新的計分函式，能滿足快速與準確的需求。我們引入以知識為基礎的方式將二十種胺基酸的原子分類為167種，涵蓋五種主要的蛋白質交互作用力 (氫鍵、電荷、凡得瓦力、疏水性作用力、雙硫鍵)，可有效降低計算成本並兼顧準確度。我們希望將我們的記分函式搭配演化式嵌合

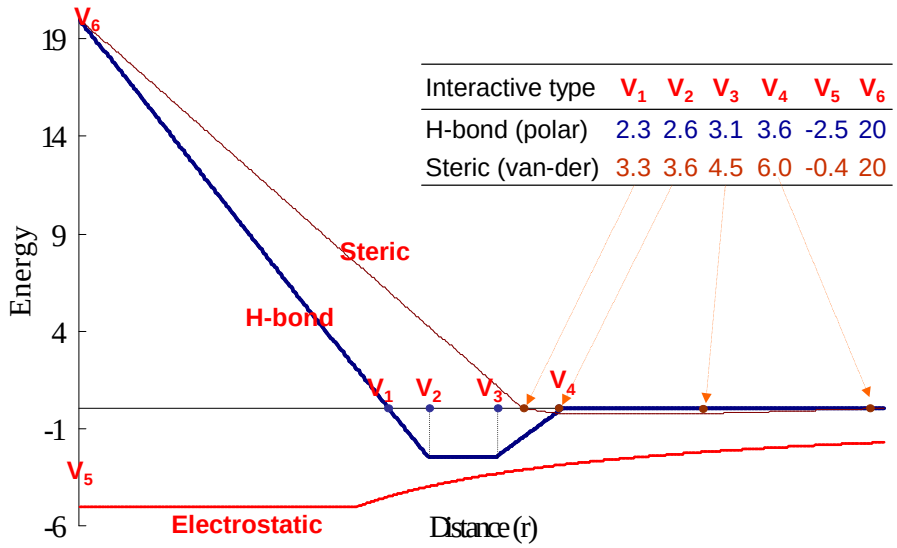
方法能提供一套自動化的蛋白質嵌合預測系統，可以用來預測蛋白質之間的交互作用，並在將來進一步研究生化反應調控。

研究方法

我們發展的 GEMDOCK[33-35] 已應用在超過 100 組的蛋白質-配體(protein-ligand)的分子識別與 estrogen receptor[38]、thymidine kinase[39] 的潛在藥物篩選。此計劃中本實驗室加強與修改 GEMDOCK 之功能，使其適合應用於蛋白質-蛋白質分子鉗合。我們以 GEMDOCK 這套軟體核心的演化式演算法(evolution strategies)作為搜尋演算法(search algorithm)，並配合先前發展的知識為基的計分函式(knowledge-based scoring function)以及新增的經驗為基的計分函式(empirical-based scoring function)，發展出一套新的預測蛋白質-蛋白質交互作用的軟體。為加快程式執行速度，我們也加入了兩個減少計算量的技巧，使得執行時間大幅縮短為原本的十分之一。以下將分別詳述新增的部份:

Scoring function

我們引入以知識為基礎的方式將二十種胺基酸的原子細分為 167 種，可涵蓋五種主要的蛋白質交互作用力 [40,41] (氫鍵、靜電力、凡得瓦力、疏水性作用力、雙硫鍵)。由於經驗為基的計分函式使用在程式中的結果仍有改進的空間，因此我們又加入了一個經驗為基的計分函式試圖增進準確度。因此法與先前發展的同樣皆以原子特性為基礎計算，故同時使用並不會增加太多時間。其參數設定與能量線性關係如下圖十一所示：



圖十一. 原子配對能量線性關係圖。電荷作用力(Electrostatic)、氫鍵(H-bond)、原子空間作用力(Steric)。

Surface search algorithm

為降低搜尋的解空間，我們藉由比較兩個蛋白質質心的位置，限定搜尋的範圍落在蛋白質表面的附近，如此可以大大縮短搜尋的時間，詳細步驟如下:

1. 計算兩個蛋白質的長寬高，取出三邊中最長的邊及最短的邊。
2. 將質心的距離限制在最長邊的和的一半以及最短邊的和的一半之間。
3. 若是質心距離大於最長邊的和的一半時，即表示這個解的蛋白質距離超過表面，程式即不再計算這個解的分數。
4. 若是質心距離小於最短邊的和的一半時，即表示這個解的蛋白質距離近到重疊在一起，由於不符合物理性質，所以直接給予極高的罰分。

Faster rotation algorithm

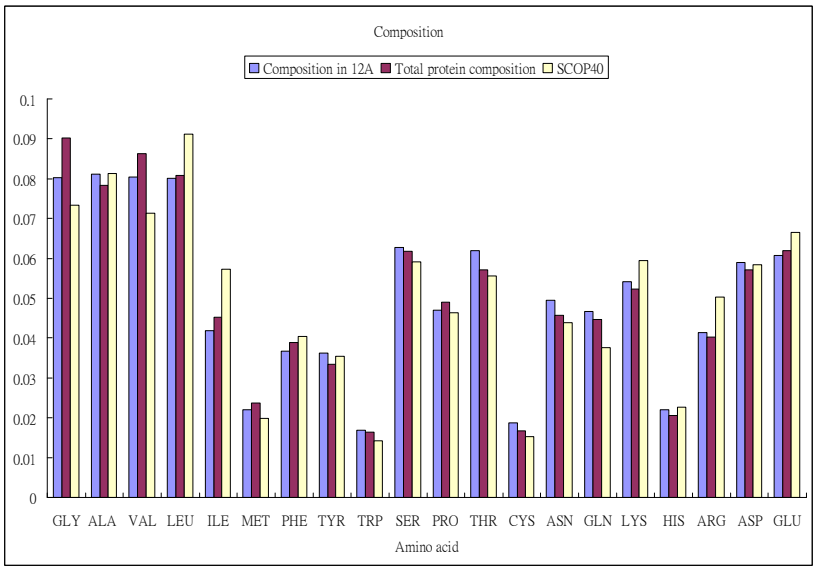
為了減少計算能量時做原子轉置(rotation + shift)的次數，我們將不必要的原子(距離太遠的原子)留在原地不動，只對轉置過後會跟結合區域產生作用的原子進行轉置，如此亦可大幅降低計算的時間。詳細步驟如下:

1. 將兩個蛋白質每邊均分為四等份，如此可將蛋白質切割為 64 塊立方格，接著計算每一格的中心，算法為每格內所含的胺基酸 C_β 的質量中心(Gly 為 C_α)。
2. 根據 Evolution strategies 所產生的轉置矩陣，轉置兩目標蛋白質中較小者上的 64 個質心，並計算轉置完畢後與較大的蛋白質上的 64 個質心間兩兩的距離，將小於 18 Å 的格子紀錄下來。
3. 轉置剛才紀錄的較小蛋白質每一格內的胺基酸 C_β 原子座標，並計算與較大蛋白質有紀錄的格子內胺基酸 C_β 原子間的距離，將距離小於 12 Å 的胺基酸紀錄下來。
4. 將有紀錄的較小蛋白質上的胺基酸中每個原子都經過轉置矩陣進行轉置，並與另一側較大蛋白質上有紀錄的胺基酸原子計算能量。

結果與討論

Data Set: 641 Protein-Protein Complexes

我們希望建構的預測模型(model)能用於預測所有類型的蛋白質-蛋白質交互作用，所以我們用來建構預測模型的資料集合(data set)是否具備足夠的代表性可代表自然界的蛋白質，將影響我們建構的模型是否有所偏差(bias)。我們挑選出的資料集合總共包含 641 對蛋白質複合體(protein complex)，經由這些 PDB[42] 檔案中的注解，我們可初步了解此資料集合中包含 G-proteins、antibody-antigen、DNA/RNA binding proteins、electron transport proteins、enzyme complex、viral proteins、transcription/translaction factors 等，涵括了自然界中大多數類型的蛋白質複合體。藉由統計 641 對蛋白質複合體資料集合的胺基酸組成(圖十二)，我們得知胺基酸的組成在蛋白質的產生交互作用的區域與整個蛋白質上並無顯著差異。這意味著，無法單純的只利用胺基酸組成差異來預測蛋白質的交互作用的活性區。



圖十二. 胺基酸組成分布。由左至右分別是 641 個蛋白質複合體 data set 的交互作用區(interface in 12 Å)、641 個蛋白質複合體 data set 的所有胺基酸組成、SCOP40 的所有胺基酸組成。

Performance of New Knowledge-based and Empirical-based Scoring Functions

從表五的結果中，我們可以看出使用 167 種原子分類的計分方式，在預測的準確度上優於 18 種原子分類跟 20 種胺基酸分類的計分方式，從結果我們可以推論 167 種原子分類的方式，不論是用在 bound 或是 unbond 的結構上，可較合理地反映出原子與原子間的交互作用以及重要的作用力(靜電力、凡得瓦力、氫鍵等)。而 18 種原子分類由於分

類的方式，所以無法反映出胺基酸配對時產生的能量關係，至於 20 種胺基酸分類則無法如 167 種原子分類那樣準確的反映出原子間物理的特性，是以我們可以說 167 種原子分類的方式，不但可以反映出原子間的物理特性，亦可以反映出胺基酸間的物理特性。

表五. 測試六種分類方式結果中前兩百名平均預測成功的次數，167 種原子分類、18 種原子分類以及 20 種胺基酸分類，加上 2 種距離分類方式(單一距離或是 10 種距離的分類)

	Average numbers of hits in top 200					
	167 atom type		**18 atom type**		**20 residue type**	
	method1[a]	method2[b]	method1	method2	method1	method2
Bound structure	117.82	**150**	0	143.41	103.58	7.76
Unbound structure	88.47	**98.17**	4.88	86.23	89.64	0

Hits are defined as docked structures with all main chain atoms RMSD ≦ 2.0 Å from the crystal complex.

[a] method1 : a contact distance cutoff R_c=6.0 Å

[b] method2 : a contact distance cutoff R_c=12.0 Å with 1.0 Å intervals

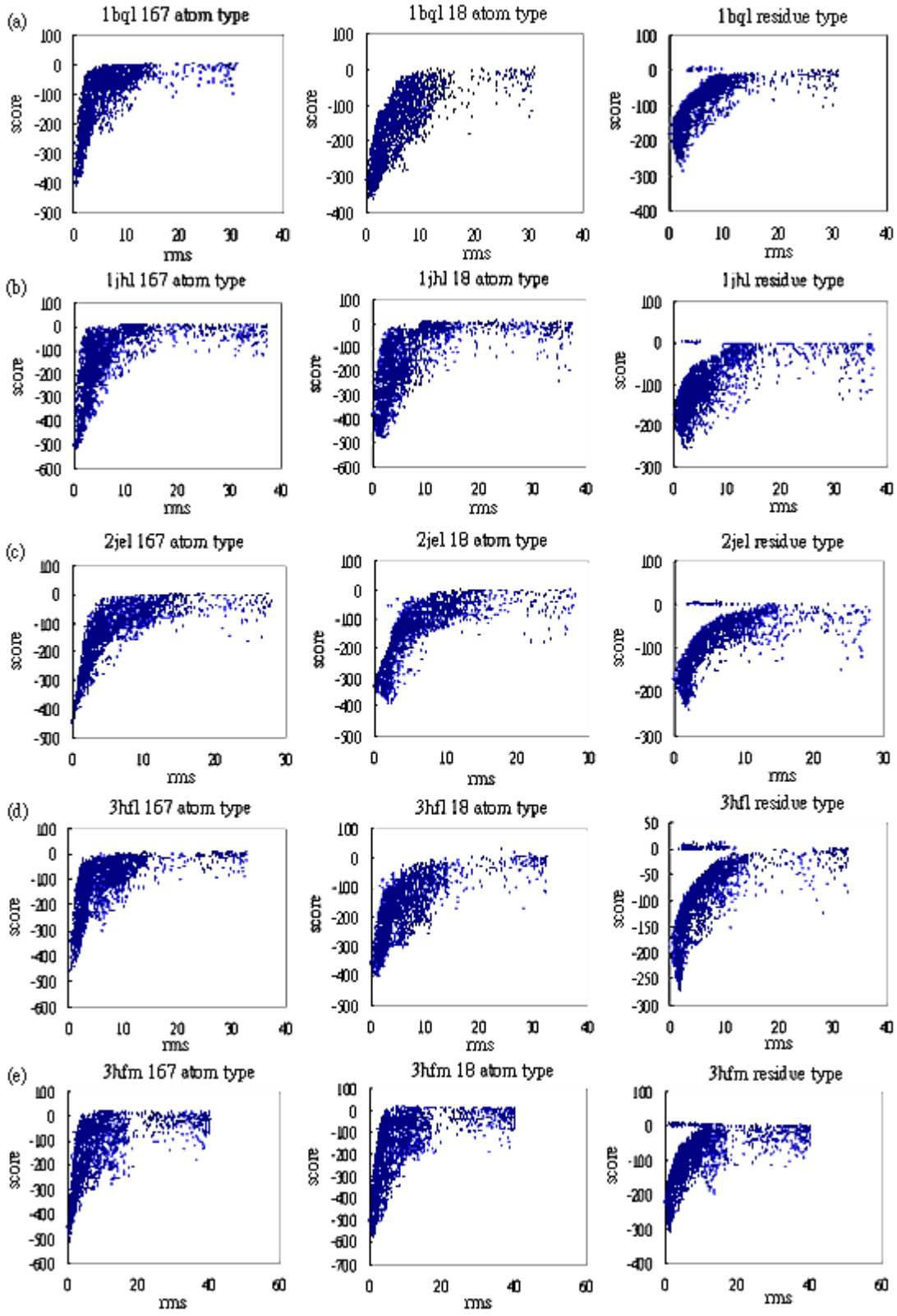
在表六中，我們加入了物理能量(empirical-base scoring function)近似值的計分方式(simpleS)的測試，我們可以從結果看出，在 bound structure 中，新加入的 simpleS 在預測的準確度上更優於之前的三種分類方式，也就是說 simpleS 雖然使用簡單的物理能量近似值，卻可以合理且準確的反映出原子間的交互作用力，因此我們可以推論原子間的物理特性對於蛋白質交互作用具有很大的影響力，而利用統計的方式做出的以知識為基礎的計分方式中，只要可以合理的反映出原子間的物理特性，在預測的準確度上也就會有一定的水準。另外我們也可以從結果觀察出，在以知識為基礎的計分方式中，距離的分類對於準確度有很大的影響，主要的原因是距離的分類隱含了表現作用力的影響，從原子的角度來看，距離 1 Å 相對於 0.5Å，更可以合理的表現出原子間的作用力，而距離 6Å 對原子來講就太大了，反而無法充分表現出原子的特性，可是對於胺基酸而言，0.5 Å 反而是太小了，6 Å 才適合表現它的特性，這也可以解釋表六中 20 種胺基酸分類在使用 0.5Å 時預測時為何結果較差。

表六. 加入物理能量近似值的計分方式測試結果，並比較不同的原子分類及不同的距離分類的優劣

Complex name (Bound structure)	Number of hits in top best 200 docked conformations								
	167 atom type			18 atom type			20 residue type		simpleS
	method1[a]	method2[b]	method3[c]	method1[a]	method2[b]	method3[c]	method1[a]	method2[b]	method4[d]
A. enzyme-inhibitor complexes									
1fss	119	107	109	0	136	123	97	0	**191**
1mah	112	89	86	0	119	114	0	0	**184**
1sbn	105	161	162	0	133	127	109	0	**184**
1udi	118	154	156	0	144	144	119	0	**182**
1ugh	131	191	185	0	159	156	131	0	**198**
2kai	119	133	158	0	137	132	117	0	**183**
2ptc	122	99	150	0	151	133	126	0	**184**
3sic	140	124	147	0	169	166	135	0	**181**
B. antibody-antigen complexes									
1bql	97	163	155	0	136	139	97	0	**191**
1jhl	82	**154**	149	0	116	112	83	0	151
2jel	125	156	158	0	128	126	125	0	**181**
3hfl	117	158	160	0	141	141	118	0	**174**
3hfm	131	**172**	164	0	149	148	0	132	155
C. other complexes									
1atn	110	159	155	0	136	137	130	0	**191**
1gla	72	168	168	0	138	134	79	0	**178**
2mip	144	174	190	0	168	169	140	0	**192**
3hhr	159	**188**	185	0	178	178	155	0	182

Hits are defined as docked structures with all main chain atoms RMSD ≦ 2.0 Å from the crystal complex.
[a] method1 : a contact distance cutoff R_c =6.0 Å (without distance class)
[b] method2 : a contact distance cutoff R_c =12.0 Å with 0.5 Å intervals (with distance class)
[c] method3 : a contact distance cutoff R_c =12.0 Å with 1.0 Å intervals
[d] method4 : simple empirical model system

根據文獻，抗原-抗體蛋白質的結合預測是目前 Protein-Protein 軟體的執行結果較不好，圖十三顯示測試了 5 種抗原-抗體蛋白質的 RMSD 與使用三種計分程式(167 種原子分類、20 種原子分類及 20 種胺基酸分類)的前 2500 名間的關係圖。從結果可知，167 種原子分類在計算蛋白質自然狀態下結晶的分數時，與其他不正確結構的分數可清楚區隔開來，且分數亦較其他兩種分類的方式要低得多。由此可知 167 種原子分類的計分方式，表現的確較好。

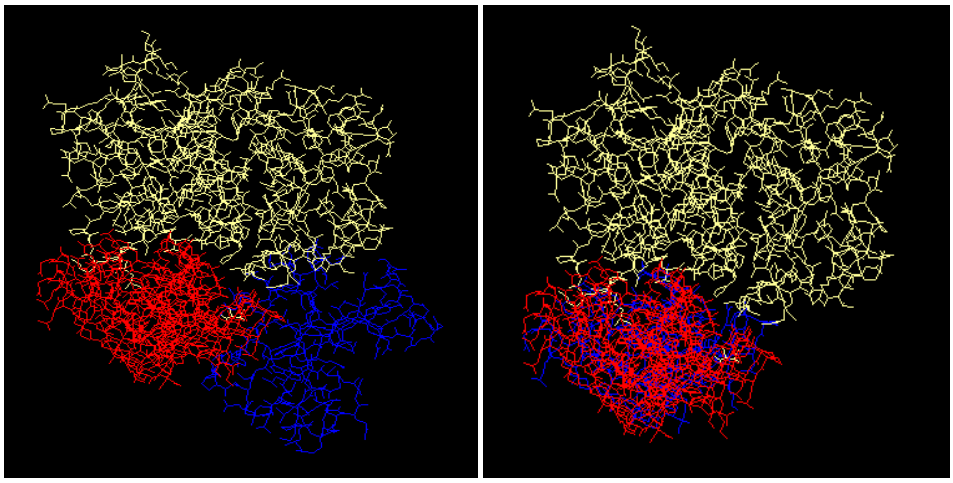


圖十三.抗原-抗體蛋白質的RMSD與使用三種計分程式(167種原子分類、20種原子分類及20種胺基酸分類)的前2500名間的關係圖(a) 1bql. (b) 1jhl. (c) 2jel. (d) 3hfl. (e) 3hfm。

Docking Results of GEMDOCK for Protein-Protein Docking

圖十四為利用 GEMDOCK 進行蛋白質-蛋白質分子鉗合測試的情況，程式在執行過程中，可以不斷的輸出演化過程中蛋白質相互的關係。以下為程式在執行時的例子，我們以 PDB[42] 編號 1udi 的 E chain (黃色)跟 1udi 的 I chain(紅色)進行分子鉗合試驗，在第 10 代及 100 代時輸出結果,則可以看出演化過程中的情況,黃色為 E chain 自然狀態時的位置，紅色為 I chain 自然狀態時的位置，藍色為程式演算過程中預測 I chain 的位置。在這個例子中,GEMDOCK 可以預測正確的分子鉗合構形,我們有系統地以 GEMDOCK 進行測試，我們已測試超過 20 個蛋白質複合體，其中也包含我們用來建構計分函式的蛋白質複合體。由於初步測試的結果有好有壞，因此接下來的工作是自測試的結果中找尋成功及失敗的原因，並且加入新的生物知識到計分程式上，以增加程式的準確度及穩定性。

執行時間上，若以只輸出最後的結果來計算，處理兩個平均含三百個胺基酸的蛋白質，population size 設為 1000，經過一百代演化的處理時間約為八到十分鐘，若未使用方法論中加速的技巧，則處理相同的情況所花的時間，約為 90 分鐘。整體來說，加入了前述的技巧，可以將執行時間減少至十分之一。就執行效率而言，我們新發展的方法較現有的方法更可節省時間 [43,44]。



10th generation　　100th generation

1udi:E chain and I chain

圖十四. GEMDOCK測試的結果，測試的資料為1udi的E chain(黃色)對上1udi的I chain(紅色)，預測的結構為藍色的部份。

計劃成果自評

I. 論文發表

In this project, we have published two SCI journal papers:

1. **J.-M. Yang*** and T.-W. Shen, "A pharmacophore-based evolutionary approach for screening selective estrogen receptor modulators," *Proteins: Structure, Function, and Bioinformatics,* vol. 59, pp. 205-220, 2005. (SCI, impact factor: 4.313) [NSC-92-2113-M-009-024 and NSC-93-2113- M-009-010]
2. **J.-M. Yang*** Y.-F. Chen, T.-W. Shen, B. S. Kristal, and D. F. Hsu, "Consensus scoring criteria for improving enrichment in virtual screening," *Journal of Chemical Information and Modeling,* vol. 45, pp. 1134-1146, 2005. (SCI, impact factor: 3.078). [NSC-93-2113- M-009-010]

We have also prepared two papers to submit: 1) DAPID: A 3D-Domain Annotated Protein-Protein Interaction Database, and (2) A fast and all-atom residue-specific energy model for protein-protein interaction prediction.

The title and acknowledge of the first paper are shown as follows:

A Pharmacophore-Based Evolutionary Approach for Screening Selective Estrogen Receptor Modulators

Jinn-Moon Yang* **and Tsai-Wei Shen**
Department of Biological Science and Technology, and Institute of Bioinformatics, National Chiao Tung University, Hsinchu, Taiwan

Grant sponsor: National Science Council of Taiwan; Grant numbers: NSC-92-2113-M-009-024 and NSC-93-2113-M-009-010. Grant sponsor: Veterans General Hospitals, University System of Taiwan; Grant number: VGHUST93-G5-05-3.

*Correspondence to: Jinn-Moon Yang, Department of Biological Science and Technology, and Institute of Bioinformatics, National Chiao Tung University, Hsinchu, 30050, Taiwan. E-mail: moon@cc.nctu.edu.tw

The title and acknowledge of the second paper are shown as follows:

Consensus Scoring Criteria for Improving Enrichment in Virtual Screening

Jinn-Moon Yang,[*,†,‡] Yen-Fu Chen,[†,‡] Tsai-Wei Shen,[†,‡] Bruce S. Kristal,[§,∥] and D. Frank Hsu[*,⊥,#]

Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, 30050, Taiwan, Institute of Bioinformatics, National Chiao Tung University, Hsinchu, 30050, Taiwan,

ACKNOWLEDGMENT

J.-M.Y. was supported, in part, by Grant NSC-93-2113-M-009-010 from the National Science Council of Taiwan and by Grant VGHUST93-G5-05-3 from Veterans General Hospitals University System of Taiwan. B.S.K. is supported

II. 主要成果

在本計畫中我們完成了三項研究蛋白質蛋白質交互作用的成果，第一部份是利用一個新的觀念「結構功能區塊交互相似性 (3D-domain interologs)」來預測蛋白質蛋白質交互作用。除此之外，我們還能夠預測交互作用蛋白質對的交互作用功能區塊。在酵母菌中，我們所預測的蛋白質蛋白質交互作用和 DIP 資料庫有 18.6% 的部份重疊。此外，我們預測出許多新的蛋白質蛋白質交互作用有相似的功能註解。我們將這些預測結果結合 DIP 的資料以及 PDB 的資料建構一個"功能區塊註解的蛋白質交互作用資料庫(DAPID)"，並且放上網路供人使用(http://gemdock.life.nctu.edu.tw/dapid)。

第二部份是發展了一套預測蛋白質結合位的工具，我們透過分析蛋白質表面及結合位上胺基酸和二級結構的出現機率，再加入疏水性的概念，並且使用 GEMDOCK 的核心演算法最佳化原子型態參數，來預測蛋白質可能的結合位。在 104 個蛋白質的訓練資料中，平均的預測成功率是 65.38%，在 56 個蛋白質的測試資料(testing set)中，平均預測成功率是 50%。我們的預測程式已具備了一定的可靠性，而這部份的研究成果將來也可用做篩選第一部份預測出來的蛋白質配對，讓第一部份的預測準確度再往上提升。

第三部份是我們使用 GEMDOCK 做蛋白質-蛋白質嵌合之預測，我們引入以知識為基礎的方式將二十種胺基酸的原子細分為 167 種，可涵蓋五種主要的蛋白質交互作用力(氫鍵、靜電力、凡得瓦力、疏水性作用力、雙硫鍵)。由於經驗為基的計分函式使用在程式中的結果仍有改進的空間，因此我們又加入了物理能量(empirical-base scoring function)近似值的計分方式增進準確度。在這兩種計分程式的預測表現上，167 種原子分類在前兩百名平均預測成功的次數是 150 次(bound structure)和 98.17 次(unbound structure)，物理能量近似值的計分方式在前兩百名平均預測成功的次數是 181.29 次(bound structure)。將來我們會把第二部份的研究(結合位的預測)整合到 GEMDOCK 中，藉此縮短程式執行時間與提高預測能力，最後，我們將整合本計劃三大部份的研究成果，完成一套自動化的蛋白質-蛋白質交互作用預測系統以及資料庫，為相關研究人員提供網路服務。

參考文獻

1. Salwinski, L. et al. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research* **32**, D449-D451 (2004).
2. Alfarano, C. et al. The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Research* **33**, D418-D424 (2005).
3. Mewes, H. W. et al. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Research* **32**, D41-D44 (2004).
4. von Mering, C. et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research* **33**, D433-D437 (2005).
5. Jansen, R. et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449-453 (2003).
6. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Science of the USA* **96**, 4285-4288 (1999).
7. Aloy, P. & Russell, R. B. Interrogating protein interaction networks through structural biology. *Proceedings of the National Academy of Science of the USA* **99**, 5896-5901 (2002).
8. Lu, L., Arakaki, A. K., Lu, H. & Skolnick, J. Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the Saccharomyces cerevisiae proteome. *Genome Research* **13**, 1146-1154 (2003).
9. Aloy, P. et al. Structure-based assembly of protein complexes in yeast. *Science* **303**, 2026-2029 (2004).
10. Yu, H. et al. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Research* **14**, 1107-1118 (2004).
11. Park, J., Lappe, M. & Teichmann, S. A. Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *Journal of Molecular Biology* **307**, 929-938 (2001).
12. Stein, A., Russell, R. B. & Aloy, P. 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Research* **33**, D413-D417 (2005).
13. Finn, R. D., Marshall, M. & Bateman, A. iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* **21**, 410-412 (2005).
14. Bateman, A. et al. The Pfam protein families database. *Nucleic Acids Research* **32**, 138-141 (2004).
15. Bairoch, A. et al. The universal protein resource (UniProt). *Nucleic Acids Research* **33**,

D154-D159 (2005).
16. Aung, Z. & Tan, K. L. Rapid 3D protein structure database searching using information retrieval techniques. *Bioinformatics* **20**, 1045-1052 (2004).
17. Hughes, T. R. et al. Functional discovery via a compendium of expression profiles. *Cell* **102**, 109-126 (2000).
18. Kirsch, T., Nickel, J. & Sebald, W. BMP-2 antagonists emerge from alterations in the low-affinity binding epitope for receptor BMPR-II. *The EMBO Journal* **19**, 3314-3324 (2000).
19. Schekman, R. & Orci, L. Coat proteins and vesicle budding. *Science* **271**, 1526-1533 (1996).
20. Barlowe, C. et al. COPII: a membrane coat formed by Sec proteins that drive vesicle budding from the endoplasmic reticulum. *Cell* **77**, 895-907 (1994).
21. Harris, M. A. et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* **32**, D258-D261 (2004).
22. Lim et al. Crystal structure and kinetic analysis of betalactamase inhibitor protein-II in complex with TEM-1 beta-lactamase. *Nature Structure Biology* **8**, 848-852 (2001).
23. Neuvirth, H., Raz, R. & Schreiber, G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *Journal of Molecular Biology* **338**, 181-199 (2004).
24. Fernandez-Recio, J., Totrov, M., Skorodumov, C. & Abagyan, R. Optimal docking area: a new method for predicting protein-protein interaction sites. *Proteins: Structure, Function, and Bioinformatics* **58**, 134-43 (2005).
25. Yang JM & CC., C. GEMDOCK: a generic evolutionary method for molecular docking. *PROTEINS: Structure, Function, and Bioinformatics* **55**, 288-304 (2004).
26. Chen, R., Mintseris, J., Janin, J. & Weng, Z. A protein-protein docking benchmark. *Proteins: Structure, Function and Genetics* **52**, 88-91 (2003).
27. Betts, M. J. & Sternberg, M. J. E. An analysis of conformational changes on protein-protein docking: implications for predictive docking. *Protein Engineering* **12**, 271-283 (1999).
28. Jackson, R. M. Comparison of protein-protein interactions in serine protease-inhibitor and antibody-antigen complexes: implications for the protein docking problem. *Protein Science* **8**, 603-613 (1999).
29. Jones, S. & Thornton, J. M. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 13-20 (1996).
30. Lo, C. L., Chothia, C. & Janin, J. The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology* **285**, 2177-2198 (1998).
31. Norel, R., Petrey, D., Wolfson, H. J. & Nussinov, R. Examination of shape complementarity in docking of unbound proteins. *Proteins: Structure, Function, and*

Bioinformatics **36**, 307-317 (1999).
32. Smith, G. R. & Sternberg, M. J. Prediction of protein-protein interactions by docking methods. *Curr Opin Struct Biol* **12**, 28-35 (2002).
33. Yang, J.-M. Development and evaluation of a generic evolutionary method for protein-ligand docking. *Journal of Computational Chemistry* **25**, 843-857 (2004).
34. Yang, J. M. & Chen, C. C. GEMDOCK: a generic evolutionary method for molecular docking. *Proteins: Structure, Function, and Bioinformatics* **55**, 288-304 (2004).
35. Yang, J. M. An evolutionary approach for molecular docking. *Lecture Notes in Computer Science* **2724**, 2372-2383 (2003).
36. Moont, G., Gabb, H. A. & Sternberg, M. J. E. Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins: Structure, Function, and Bioinformatics* **35**, 364-373 (1999).
37. Glaser, F., Steinberg, D. M., Vakser, I. A. & Ben-Tal, N. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins: Structure, Function, and Bioinformatics* **43**, 89-102 (2001).
38. Yang, J.-M. & Shen, T.-W. A pharmacophore-based evolutionary approach for screening estrogen receptor antagonists. *Congress of Evolutionary Computation (CEC 2004)*, 1028-1035 (2004).
39. Yang, J.-M., Shen, T.-W., Chen, Y.-F. & Chiu, Y.-Y. An evolutionary approach with pharmacophore-based scoring functions for virtual database screening. *Lecture Notes in Computer Science* **3102**, 481-492 (2004).
40. Chang, Y. S. in *Department of Biological Science and Technology* (National Chiao Tung University, Hsinchu, 2003).
41. Yang, J. M. & Chang, Y. S. A fast and all-atom residue-specific energy model for protein-protein interaction prediction. *Perpared to submit* (2004).
42. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242 (2000).
43. Juan, F.-r., Maxim, T. & Ruben, A. Soft protein-protein docking in internal coordinates. *Protein Science* **11**, 280-291 (2002).
44. Graham, R. S. & Michael, J. E. Prediction of protein-protein interactions by docking methods. *Current Opinion in Structural Biology* **12**, 28-35 (2002).