行政院國家科學委員會專題研究計畫 成果報告

研發一套以與受體結合為基礎的快速虛擬篩選方法
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 98-2221-E-009-122-
執 行 期 間 ：98 年 08 月 01 日至 99 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：黃慧玲

計畫參與人員：碩士班研究生-兼任助理人員：林意哲
碩士班研究生-兼任助理人員：林玉祥
博士班研究生-兼任助理人員：洪瓊慧

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 99 年 10 月 31 日

研發一套以受體結合為基礎的快速虛擬篩選方法
Developing a fast docking-based virtual screening method

計畫編號： NSC 98-2221-E-009 -122 -
執行期限：98 年 08 月 01 日至 99 年 07 月 31 日
主持人：黃慧玲　　國立交通大學
hlhuang@mail.nctu.edu.tw

摘要

本計劃提出一套以受體結合為基礎的快速虛擬篩選方法來尋找資料庫中藥物設計所要的化學小分子。研究進行方式是改善先前發展的 SODOCK 蛋白質嵌合演算法為基礎，並以四個方向來進行研究，以達成目標: (1)提出一套改良 SODOCK 的方法，稱為 PSODOCK，使用直交實驗設計來對粒子群最佳化演算法 PSO 的初始化做高效能取樣，(2)與 Autodock 軟體工具做分子嵌合模擬實驗比較效能來提升 PSODOCK 效能，(3)探討蛋白質序列的物化特性來了解蛋白質鍵結機制，有助於發展蛋白質嵌合的快速演算法，(4) 研究以繪圖處理器(GPU)平行運算及 CUDA 的技術來幫助 PSODOCK 發展快速篩選的演算方法。本計畫研究進度順利，已達預期目標，並有相關期刊論文及研討會論文發表。

關鍵詞：蛋白質嵌合、資料探勘、粒子群最佳化、基因演算法、參數最佳化、蛋白質功能預測、物化特性、特徵選取、因素分析、機器學習

Abstract

This project proposes a fast docking-based virtual screening method to filter compounds for drug discovery process. This study bases on improving our developed docking method SODOCK and achieves the project goal by way of the following four aspects. 1) Propose an improved docking method, named PSODOCK, by using an orthogonal experimental design based initialization of particle swarm optimization which is the core algorithm of SODOCK. 2) Compare PSODOCK with existing software tool Autodock using large molecular docking applications to advance the performance of PSODOCK. 3) Investigate informative physicochemical and biochemical properties of protein sequences to understand the binding mechanism of proteins that is helpful in developing fast docking-based virtual screening method. 4) Study the parallel computation of Graphics Process Unit (GPU) and techniques of CUDA to enhance PSODOCK and the screening method. The goal of this project is achieved and part of the good achievements was published in several international conference and journal papers.

Keywords: Protein Docking, Data Mining, Particle Swarm Optimization, Genetic Algorithm, Parameter Optimization, Protein Function Prediction, Physicochemical Properties, Feature Selection, Factor Analysis, Machine Learning.

一. 前言

電腦進行虛擬篩選(virtual screening）在藥物設計過程中便扮演相當重要的功用。虛擬嵌合篩選（virtual docking screening）乃是針對受體的結合位置，對資料庫中大量分子做篩選以求發現與受體可以結合的分子。資料庫篩選最大的好處在於節省合成新先導化合物的時間，如何提高虛擬嵌合篩選的效能是本計畫的研究主題，研究進行方式是改善先前發展的SODOCK蛋白質嵌合演算法為基礎。

二. 研究目的

以受體結合為基礎的快速虛擬篩選方法來尋找資料庫中藥物設計所要的化學小分子，我們以提升蛋白質嵌合的效能為主，方向是改善嵌合的核心演算法SODOCK。另一方向是研究以繪圖處理器(GPU)平行運算及CUDA的技術來幫助改善的PSODOCK發展快速篩選的演算方法。

具體的研究目的以四個方向來進行研究,以達成目標: (1)提出一套改良SODOCK的方法,稱為PSODOCK,使用直交實驗設計來對粒子群最佳化演算法PSO的初始化做高效能取樣,(2)與Autodock軟體工具做分子嵌合模擬實驗比較效能來提升PSODOCK效能,(3)探討蛋白質序列的物化特性來了解蛋白質鍵結機制,有助於發展蛋白質嵌合的快速演算法,(4) 研究以繪圖處理器(GPU)平行運算及CUDA的技術來幫助PSODOCK發展快速篩選的演算方法。

三. 文獻探討

近年來基因演算法已經被應用蛋白質與化合物嵌合問題中,國外著名的方法包括了:GOLD [1]和AutoDock [2]等。在較早期的應用中,評估函式中的參數大多是由專家根據其經驗與知識來設計。而最近的研究之共通特點是以基因演算法取代傳統人工設計或數值方法,求得一組參數,使評估函式可以更準確的反應嵌合的狀態。GOLD使用傳統的基因演算法當作其搜尋的工具,並利用能量評估函式,結合了不同分子間的作用力,進行嵌合結果的評估。

DOCK[3]由 Kuntz 等人設計開發而來,已經非常廣泛地被使用在各種蛋白質與化合物的組合上。其做法為在蛋白質的活性位置範圍區域,自動產生配體分子的各種可能方位及構形,pocket 的形狀由許多圓球代表,圓球的中心代表化合物原子的可能位置,至少四個原子必須吻合各個圓球才算作有效的化合物。

AutoDock [2]有公開的原始碼,是很多研究虛擬篩選學者很喜歡的一套嵌合演算法與發展環境,目前由3.05版晉升為4.2版。AutoDock使用了基因演算法,加上一種優良的local search演算法來增加傳統基因演算法的效果,並利用能量評估函式,評估嵌合的結果,根據實驗出來的構型進行下一步的分類,選出得分較高的結果當作參考的依據,並且利用格子為基礎(grid-based)加速整體運算的速度。

FlexX [4]先利用幾何方法找到合適的化合物後,使用幾何評估法針對蛋白質與化合物間的構形進行評分,找到合適的構形後,再以能量評估法進行驗證,經過數次的交叉驗證後,求得一組較佳的構形。

DOVIS是一個做虛擬篩選的軟體工具,它是基於使用AutoDock4.0環境所發展的多CPU (256個CPU) Linux cluster,利用平行搜尋方式對大量的配體(ligand)做嵌合計算。DOVIS使用AutoDock4.0的蛋白質與配體的嵌合計算,並沒有使用自行發展的嵌合計算方法。為了提高虛擬篩選的效能,DOVIS在四個方面做改善(1)改善計算流程平行化效能、(2) 最小化在共同系統的檔案運算工作、(3) 與其他模組程式介面的溝通效率、(4)利用 third-party 軟體對配體與受體的化合物做分數的重新評分。

本計畫主持人新近也參與發展出蛋白質與化合物嵌合系統SODOCK[5],這系統採用粒子群最佳化演算法的自動化,對於含有大量可旋轉鍵結的高度彈性化合物,會因為其極大的構型搜尋空間與參數間的強烈關聯性,使得彈性嵌合的最佳化問題變得極為困難。針對蛋白質與小分子嵌合問題,導入粒子群最佳化演算法成功克服了以往參數間強烈的交互作用的難題,使得預測精確度與計算效率皆優於現有之方法。此系統能夠快速決定分子結構的最佳構型,從而改善後續的蛋白質分析和預測結果之效能。SODOCK [5]聯合AutoDock的結合程式是第一套使用粒子群最佳化(PSO)的方法[6]。

四. 研究方法

4.1 改良SODOCK的嵌合方法

本計畫提出一套改良SODOCK的方法,稱為PSODOCK; 它是新增使用直交表設計來對PSO初始化的高效能取樣,是一套高度平行化的演算法。以PSO作為三維空間中配體構型的搜尋方法,就是考慮其位置 (tx, ty, tz)、旋轉角度 (rx, ry, rz)及各分支的旋轉角度(ai),來設粒子的向量,以 PSI-See remark 4為例,有 24個可旋轉分支,所以其參數數目為 30,算是一個高度彈性(highly flexible)的小分子[5]。如果能在初始化時,使用直交實驗,有系統的求出初步較佳解的粒子群,對正確解的搜尋有縮短時間的幫助[7]。直交表(orthogonal array) 是由 R. A. Fisher 最先提出的,直交所代表的意思是平衡,亦即

統計上的獨立 (statistically independence)，因此直交表中每一欄的各水準值 (level) 出現次數是相同的，使用直交表，事實上僅是進行部份因素實驗 (fractional-factorial experiment)，因此能較完全因素實驗 (full-factorial experiment) 節省大量執行的時間，且直交實驗具有系統推理的特性，因此只需進行部份因素實驗就可以求得最佳解的近似解 (near optimum) [8, 9]。

4.2 與Autodock做大分子嵌合比較

為了解在受體的蛋白質中可能接合的活化區之物化特性與分子鍵結機率，物化性質是一種容易被用於各種有效的預測方法，進而了解蛋白質的功能和特色。一般而言，領域知識(domain knowledge)來進行分析蛋白質是需要由生物學家來選擇有效的物化性質。

在這部份的探討先由可以鍵結的蛋白質的物化特性探討。Amino acid indices (AAindex) database中搜集了許多已經發表的氨基酸物化特性在最新的版本(AAindex 9.0版)[10]。

使用SODOCK軟體來分析大分子接合模擬預測分析。本實驗使用Beta-1, 4-Glycanase分子如圖一，將Beta-1, 4-Glycanase分子上的9個重要胺基酸活性位點分別突變如圖二所示，9個胺基酸分別是E127、E233、D235、H205、H80、K47、Q87、N44及W273，以直交實驗設計$L_{64}(8^9)$，大量找出最佳的胺基酸活性位點分析及預測大分子與小分子間接合，此方式可節省實驗上成本。

4.3 加入受體與分子結合的鍵結知識

為了改良Docking 的速度與精準度，擬加入受體與分子結合的鍵結知識(binding knowledge)，其中以物化特性(physicochemical properties)探討binding是最重要研究。我們使用智慧型基因演算法發展的特徵選取演算法來選取有觀蛋白質與DNA鍵結的有關物化特性，進一步了解較具有鍵結物化特性強度的binding sites，使得PSO可以因縮小可能的嵌合空間而加快速度。詳細方法已發表在下列期刊:
Hui-Ling Huang, S.-Y. Ho, I-C. Lin, Y.-F. Liou, C.-T. Tsai, K.-T. Hsu, W.-L. Huang and S.-J. Ho, "Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties." Accepted by BMC Bioinformatics, 2010.

4.4 使用GPU平行運算及CUDA技術

本計劃提出的虛擬篩選方加入平行計算可增快5~10倍速度。SIMD（Single Instruction, Multiple Data）一個指令可以同時操作不同數據，支持CUDA的GPU便可以看成是一個SIMD平行處理系統。CUDA的技術是將程序透過資料傳輸的方式交由繪圖處理器(GPU)執行，透過GPU上大量的算術邏輯單元(ALU)去做計算，雖然GPU對處理邏輯判斷的效能並不好，但如果是做單純的資料計算的話，它的效能是相當好的，而且價格便宜又不需要大量的電腦佔空間，也因此我們選擇了CUDA。

4.5 快速虛擬篩選方法

本計劃提出的快速虛擬篩選方法如下:

一、進行嵌合計算之前篩，先針對受體分析其可能配體的物化性質做為初步篩選的限制條件，例如分、子量、H-bound Acceptors、H-bound Donors、LogP，rotatable bound 等。

二、對符合物化性質條件的 ligand 做資料結構的前處理，符合 SODOCK 的規格(即 AutoDock 的規格)。例如對水分子、金屬離子、配體原子、有磁性的原子、電荷等的處理[5]。

三、設定 PSODOCK 與 AutoDock 的參數並準備受體與配體的資料檔。

四、執行 PSODOCK 對化合物的嵌合分數計算。評分函式採用 AutoDock 的函數。對嵌合計算結果輸出最佳嵌合分數的化合物構形(便於視覺化檢驗)、RMSD、cluster rank、嵌合分數。RMSD是嵌合的配體構形與參考配體結構的差值。

五、使用多個評分函式的一致評分法(consensus scoring)對候選的配體構形重新打分數。評分函式將採用有公開易於使用的常用函式。一致評分法在很多研究證明比單一評分函式有更好的效果[6]。

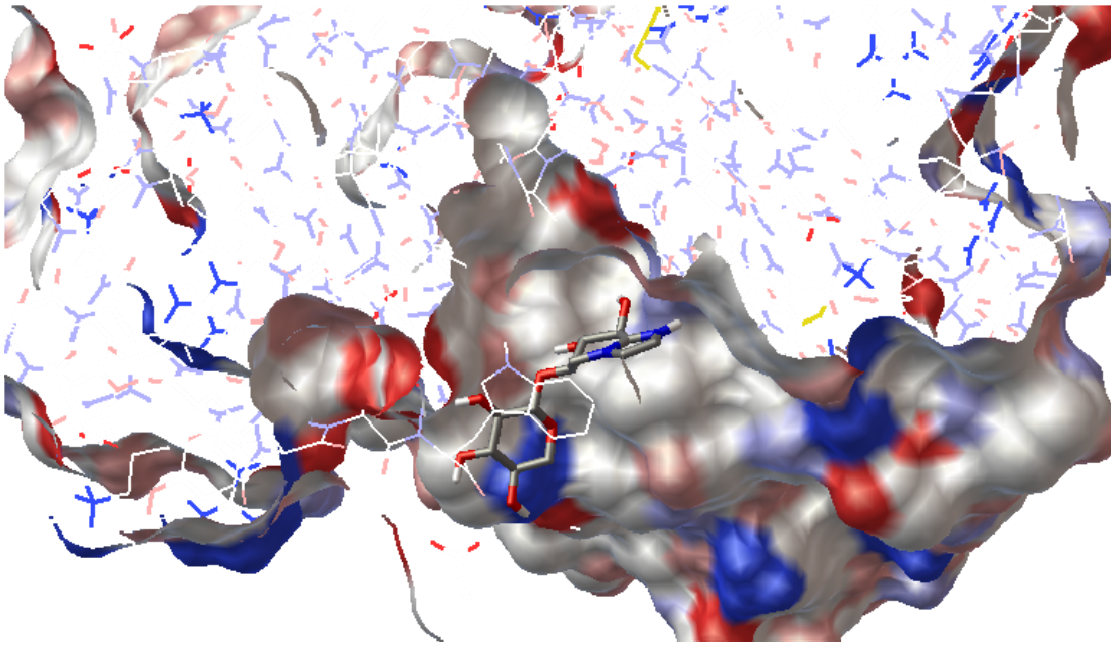
六、使用已知的生物領域知識對較佳排名的候選配體構形與授體(如 CDK2 蛋白質)的嵌合模式做進一步檢驗篩選。例如 CDK2 的 active sites 會提供很多不同的 inhibitor 嵌合模型,化合物在 ATP binding pocket 區域會與 Lys33、Leu83、Glu81、Ap86、Lys89……等行成氫鍵交互作用等。

五. 結果與討論

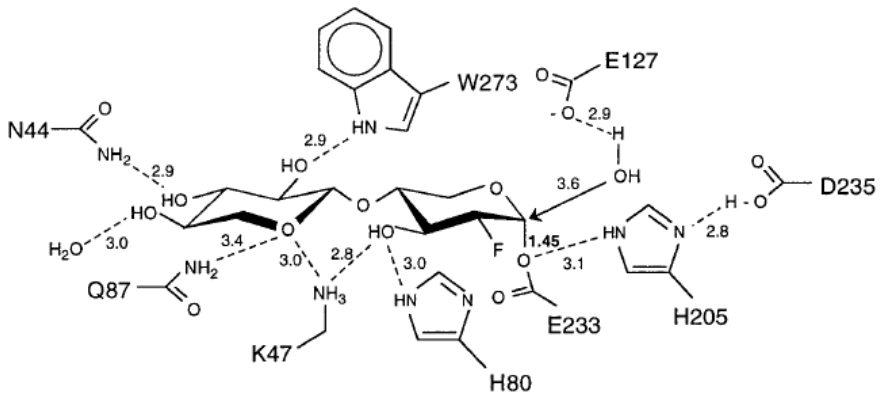
主要的計劃成果將由一些數據結果圖示與表格來呈現如下。.

5.1 大分子接合模擬預測分析

利用系統化推理出最佳突變置換氨基酸對換表。表一為 SODOCK、AUTODOCK 大分子接合模擬實驗結果,其中 Energy 部份以值越小越穩定,Histogram 值越大表示越穩定。由此表中我們可以看出使用 SODOCK 大分子接合模擬預測大部分預測結果都優於 AUTODOCK。



圖一. Beta-1, 4-Glycanase 分子



圖二. The Beta-1, 4-Glycanase, the nine activity sites (E127, E233, D235, H205, H80, K47, Q87, N44, W273) [11]。

表一、The Lowest Energy Result of SODOCK and AUTODOCK using $L_{64}(8^9)$.

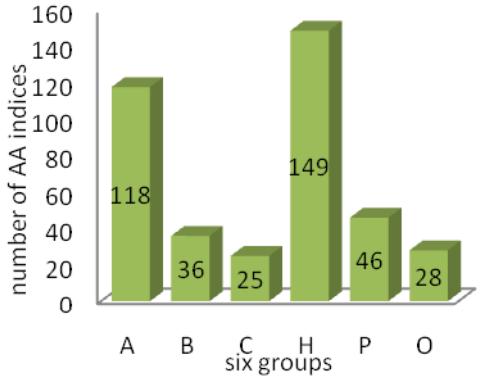
			SODOCK			AUTODOCK		
PDB	Ligand	Torsions	Energy	RMSD	Histogram	Energy	RMSD	Histogram
1FHD2-1	xyp_xim	7	-8.51	4.108	28	-7.21	4.043	11
1FHD2-2	xyp_xim	7	-8.75	1.180	44	-8.04	1.096	25

1FHD2-3	xyp_xim	7	-8.13	4.094	22	-9.25	4.115	16
1FHD2-4	xyp_xim	7	-8.40	0.942	50	-8.06	0.970	50
1FHD2-5	xyp_xim	7	-7.63	3.973	22	-7.57	3.085	20
1FHD2-6	xyp_xim	7	-8.62	0.669	50	-7.25	2.405	49
1FHD2-7	xyp_xim	7	-7.71	2.744	22	-8.14	2.757	21
1FHD2-8	xyp_xim	7	-7.02	2.943	29	-7.14	3.761	19
1FHD2-9	xyp_xim	7	-7.59	1.233	42	-8.76	4.000	41
1FHD2-10	xyp_xim	7	-7.78	4.211	12	-8.39	4.339	6
1FHD2-11	xyp_xim	7	-8.27	1.641	50	-8.40	1.523	40
1FHD2-12	xyp_xim	7	-7.36	3.072	45	-7.90	3.955	29
1FHD2-13	xyp_xim	7	-7.49	3.638	38	-7.14	4.559	16
1FHD2-14	xyp_xim	7	-7.66	3.779	43	-7.06	4.941	7
1FHD2-15	xyp_xim	7	-7.66	3.994	28	-7.02	4.023	24
1FHD2-16	xyp_xim	7	-8.94	4.171	49	-7.79	4.127	45
1FHD2-17	xyp_xim	7	-7.86	4.239	28	-7.15	3.150	7
1FHD2-18	xyp_xim	7	-8.42	1.460	42	-7.59	1.433	26
1FHD2-19	xyp_xim	7	-7.48	3.842	28	-7.11	3.810	14
1FHD2-20	xyp_xim	7	-7.83	4.172	18	-7.50	4.018	25
1FHD2-21	xyp_xim	7	-8.73	2.465	21	-8.72	2.485	17
1FHD2-22	xyp_xim	7	-8.12	0.876	49	-8.58	0.741	42
1FHD2-23	xyp_xim	7	-8.06	4.091	38	-7.87	1.420	20
1FHD2-24	xyp_xim	7	-8.56	2.780	18	-7.36	2.179	15
1FHD2-25	xyp_xim	7	-8.09	1.220	50	-8.62	1.167	43
1FHD2-26	xyp_xim	7	-8.00	4.063	30	-8.38	4.040	27
1FHD2-27	xyp_xim	7	-7.24	4.004	30	-7.68	4.005	16
1FHD2-28	xyp_xim	7	-6.46	4.109	43	-7.68	4.007	42
1FHD2-29	xyp_xim	7	-7.79	2.833	37	-7.76	1.172	30
1FHD2-30	xyp_xim	7	-7.52	1.033	50	-8.09	0.893	47
1FHD2-31	xyp_xim	7	-7.81	3.602	21	-8.33	3.606	7
1FHD2-32	xyp_xim	7	-6.49	1.160	50	-7.80	1.160	35
1FHD2-33	xyp_xim	7	-7.17	1.641	33	-8.20	1.669	41
1FHD2-34	xyp_xim	7	-8.11	2.191	48	-8.12	2.207	23
1FHD2-35	xyp_xim	7	-7.88	4.037	42	-7.87	3.984	44
1FHD2-36	xyp_xim	7	-7.62	2.273	42	-7.69	2.256	25
1FHD2-37	xyp_xim	7	-7.67	1.554	50	-8.00	1.479	44
1FHD2-38	xyp_xim	7	-7.83	4.065	45	-7.49	3.345	35
1FHD2-39	xyp_xim	7	-6.71	3.707	47	-8.02	1.429	47
1FHD2-40	xyp_xim	7	-6.97	4.132	40	-8.33	4.119	29
1FHD2-41	xyp_xim	7	-9.25	4.013	41	-7.53	4.062	38
1FHD2-42	xyp_xim	7	-6.30	1.275	50	-6.96	1.266	40
1FHD2-43	xyp_xim	7	-8.11	2.276	32	-7.87	2.194	41
1FHD2-44	xyp_xim	7	-7.26	2.796	45	-8.45	2.817	10
1FHD2-45	xyp_xim	7	-7.01	3.108	31	-7.26	3.856	23
1FHD2-46	xyp_xim	7	-8.21	3.941	17	-7.57	4.027	16
1FHD2-47	xyp_xim	7	-7.67	3.890	25	-6.76	3.842	15
1FHD2-48	xyp_xim	7	-7.50	3.375	14	-7.12	3.996	7
1FHD2-49	xyp_xim	7	-7.33	3.015	49	-7.71	0.886	48
1FHD2-50	xyp_xim	7	-7.22	4.155	9	-8.68	4.132	7
1FHD2-51	xyp_xim	7	-9.03	3.717	39	-7.82	3.353	21
1FHD2-52	xyp_xim	7	-8.33	2.766	24	-7.52	2.792	29
1FHD2-53	xyp_xim	7	-8.49	3.346	23	-6.68	4.535	8
1FHD2-54	xyp_xim	7	-8.03	4.067	36	-8.25	4.059	16
1FHD2-55	xyp_xim	7	-7.11	3.357	22	-7.74	2.796	16
1FHD2-56	xyp_xim	7	-7.55	2.305	16	-7.60	2.404	11
1FHD2-57	xyp_xim	7	-7.97	2.301	40	-7.55	1.617	30
1FHD2-58	xyp_xim	7	-6.89	3.066	45	-8.04	4.059	34
1FHD2-59	xyp_xim	7	-7.31	1.570	50	-8.24	1.700	40

1FHD2-60	xyp_xim	7	-8.03	4.562	10	-7.99	4.680	6
1FHD2-61	xyp_xim	7	-8.21	1.662	39	-8.20	1.646	35
1FHD2-62	xyp_xim	7	-9.59	2.895	31	-7.55	2.938	17
1FHD2-63	xyp_xim	7	-8.86	1.242	41	-8.10	1.294	41
1FHD2-64	xyp_xim	7	-8.00	2.227	28	-7.22	3.994	18

5.2 物化特性分析

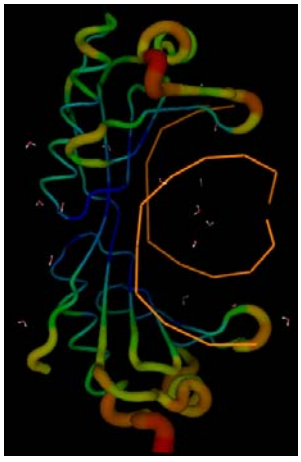
Tomii and Kanehisa[10]將 404 個物化特性分為 6 groups，圖三為 404 個物化特性統計表。本計劃用 fuzzy c-means(FCM)[13]將 531 個物化特性分成 20 群，表二為 531 個物化特性分的 20 群與分佈的 6 groups。



圖三.為 404 個物化特性統計表。

群別[14bmc] C_7、C_9、C_{10}、C_{16}、和 C_1 等 5 類別為本計劃提出最相關 binding 物化特性類別。表三中列舉一些典型 binding 相關的物化特性 hydrophobicity、secondary structure、charge、solvent accessibility、polarity、flexibility、normalized Van Der Waals volume、pK (pK-C, pK-N, pK-COOH 和 pK-a(RCOOH))等等。

Binding protein 如圖四，紅黃橙色為 binding 區 B-factor 值較大。



圖四. TATA-BOX BINDING PROTEIN (TBP)

表二、531 個物化特性分的 20 群與分佈的 6 groups。

Cluster	A	B	C	H	P	O	TOTAL
C_1					1	1	2
C_2					2		2
C_3				6			6
C_4					3		3
C_5			1	2	1		4
C_6	1			3	1	1	6
C_7	47	7	2	74	14	3	147
C_8					3		3
C_9	51	1	3	50	6	21	132
C_{10}	38	30	2	42	9	2	123
C_{11}					6		6
C_{12}				2			2
C_{13}			1				1
C_{14}				12	2	1	15
C_{15}				1			1
C_{16}	1		38	4			43
C_{17}				3			3
C_{18}	3			17	8		28
C_{19}				1		1	2
C_{20}					2		2
TOTAL	141	38	47	217	58	30	531

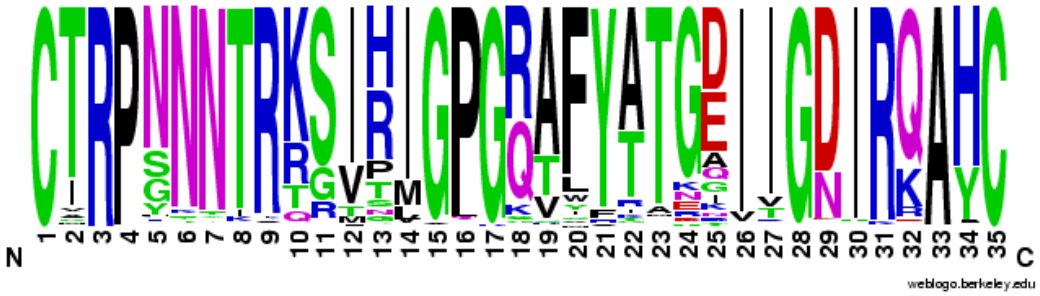
A: Alpha and turn propensities. B: Beta propensity. C: Composition. H: Hydrophobicity. P: Physicochemical properties. O: Other properties.

圖五是利用全部序列作多重序列比對(multiple sequence alignment)的結果，並刪除含有空白胺基酸(gap)比例大於 80%的序列位置(site)，藉此得到只含有 35 個胺基酸長度的序列組[15]。使用 WebLogo 對序列資料作胺基酸組成分析圖，由圖可發現，經過序列比對以及刪除胺基酸較少出現的位置過後，序列資料的變異性還是非常的大。圖六表示將 Main dataset[14]中 binding 與 non-binding 的 sequences 分開統計，以 DNA-binding 而言，將其所有的胺基酸序列分成 20 個個別胺基酸統計 X 軸表示 20 個胺基酸的簡稱 Y 軸表示每個胺基酸數占全部的胺基酸數的百分比，故 non-binding 的以此類推做成的圖。

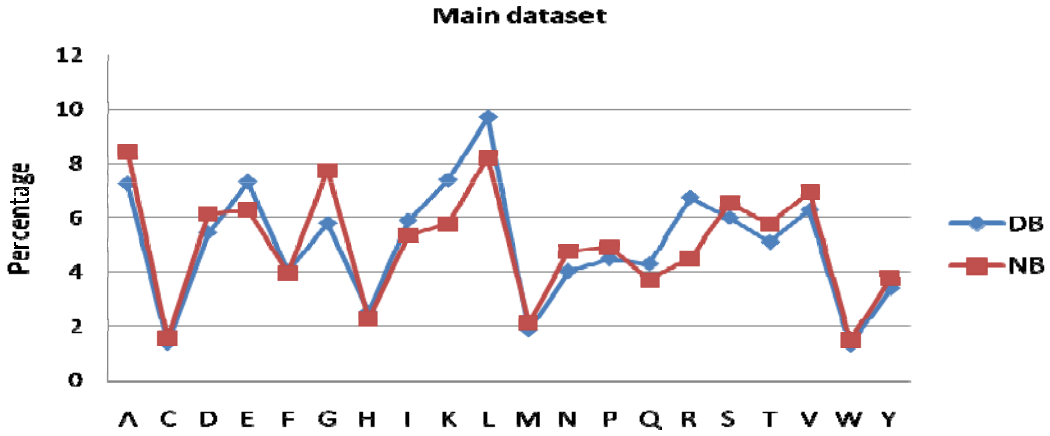
表三、由重要 5 類別中列舉一些典型 binding 相關的物化特性。

C_{id}	AAindex ID	PCP	C_{id}	AAindex ID	PCP
7	BHAR880101	Flexibility	10	FASG760105	pK-C
7	BURA740101	Secondary structure	10	JOND750102	pk- (-COOH)
7	CHOC760103	Solvent accessibility	10	RADA880108	Polarity
7	HOPT810101	Hydrophobicity	16	PRAM900101	Hydrophobicity
7	FAUJ880111	Charge	16	FUKS010104	Solvent accessibility
9	KARP850101	Flexibility	16	KUMS000103	Secondary structure
9	PALJ810115	Secondary structure	18	PONP800107	Solvent accessibility
9	ROSM880101	Hydrophobicity	18	GRAR740102	Polarity
9	KUHL950101	Solvent accessibility	18	FASG760104	pK-N
10	ZIMJ680101	Hydrophobicity	18	FAUJ880113	pK-a(RCOOH)
10	EISD860101	Solvent accessibility	18	FAUJ880103	Normalized van der Waals volume
10	GEIM800101	Secondary structure			

C_{id}: Cluster ID. PCP: physicochemical and biochemical property



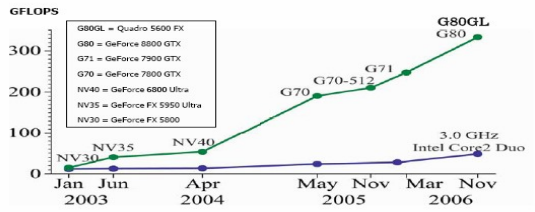
圖五. 利用全部序列作多重序列比對(multiple sequence alignment)的結果



圖六. Binding 與 non-binding 的 sequences 分開統計。

5.3 平行計算資源

圖七就大致說明了目前 GPU 和 CPU 的速度演進，很明顯的可以看出來， GPU 的成長速度已經算是 CPU 的數倍了，而這種利用 GPU 來完成圖形處理以外的計算的架構稱作 GPGPU(General-purpose computing on graphics processing unit，GPU 的通用計算)。CUDA(Compute Unified Device Architecture，統一計算設備架構) 即是 NVIDIA 的 GPGPU 模型。



圖七. 顯示晶片與中央處理器速度成長比較表[16]

「粒子群尋優化法」提出一個設計最佳平行運算負載平衡的方法，試將多處理器之平行計算工作負載效能最佳化。要使用對 GPU 來做計算來說，如果要發揮出它的

最大效益有幾個特點：

1. 加設執行緒，使同步進行的資料變多。
2. 平行化的方式，以資料可以不需等待上筆資料的循序處理為主。
3. 加設 shaerd memory，讓資料可以在 GPU 內部交換就好。
4. 多利用快取機制，可提高 Shared memory 的效益。

關於 AutoDock 軟體平行化執行在 NVIDIA GPU，可以到網站 http://autodock.scripps.edu/下載 CUDA 程式。關於 CUDA-PSO 可以到網站 http://www.ce.unipr.it/people/mussi/projects/CUDA-PSO-v1.0-html_documentation/main.html下載 CUDA 程式。圖八為提供程式名稱。

圖八. CUDA-PSO 程式名稱。

5.4 論文發表

這一年計劃皆有完成預定的目標。也如期投出期刊論文並且接受刊出論文。感謝國科會給予資源才能順利完成並且豐碩收穫。執行論文這一年期間，發表相關的論文如下列:

[1] K.-T. Hsu, **Hui-Ling Huang**, C.-W. Tung, Y.-H. Chen, and S.-Y. Ho, "Analysis of physicochemical properties on prediction of R5, X4 and R5X4 HIV-1 coreceptor usage," *International Journal of Biological and Life Sciences*, vol. 5, no. 1, pp. 208, Winter, 2009.

[2] C.-H. Hung, **Hui-Ling Huang**, K.-T. Hsu, S.-J. Ho and S.-Y. Ho, "Prediction of non-classical secreted proteins using informative physicochemical properties." *Interdiscip Sci Comput Life Sci* 2: 263-270, 2010.

[3] **Hui-Ling Huang**, S.-Y. Ho, I-C. Lin, Y.-F. Liou, C.-T. Tsai, K.-T. Hsu, W.-L. Huang and S.-J. Ho, "*Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties*." Accepted by BMC Bioinformatics, 2010.

參考文獻

[1] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor, "Development and validation of a genetic algorithm for flexible docking," Journal of Molecular Biology, vol. 267, pp. 727-748, 1997.

[2]G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson, "Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function," *Journal of Computational Chemistry*, vol. 19, pp. 1639-1662, 1998.

[3]T. J. A. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz, "DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases," *Journal of Computer-Aided Molecular Design*, vol. 15, pp. 411-428, 2001.

[4]M. Rarey, B. Kramer, and T. Lengauer, "The particle concept: Placing discrete water molecules during protein-ligand docking predictions," *Proteins-Structure Function and Genetics*, vol. 34, pp. 17-28, 1999.

[5] H.-M. Chen, B.-F. Liu, H.-L. Huang, S.-F. Hwang and S.-Y. Ho, "SODOCK: Swarm Optimization for Highly Flexible Protein-Ligand Docking," *Journal of Computational Chemistry*, vol. 28, pp. 612-623, 2007.

[6] N. Moitessier et al., "Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go," British Journal of

Pharmacology, 153, s7-s26, 2008.
[7]Chen CP, “Designing an efficient general-purpose particle swarm optimization algorithm,” Thesis of Institute of Computer Science, National Chiao Tung University, 2007.
[8] M. S. Phadke, Quality Engineering Using Robust Design. Englewood Cliffs. NJ: Prentice-Hall, 1989.
[9] S. H. Park, Robust Design and Analysis for Quality Engineering. Chapman & Hall, 1996.
[10]S. Kawashima, *et al.*, "AAindex: amino acid index database, progress report 2008," *Nucleic Acids Res*, vol. 36, pp. D202-5, Jan 2008.
[11]Valerie Notenboomet al., Exploring the Cellulose/Xylan Specificity of the Beta-1,4-Glycanase Cex from Cellulomonasfimi through Crystallography and Mutation, 1998
[12]Tomii K, Kanehisa M: Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng 1996, 9(1):27-36.
[13]Bezdek JC: Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press 1981.
[14] Hui-Ling Huang, S.-Y. Ho, I-C. Lin, Y.-F. Liou, C.-T. Tsai, K.-T. Hsu, W.-L. Huang and S.-J. Ho, “Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties.” Accepted by BMC Bioinformatics, 2010.
[15]K.-T. Hsu, Hui-Ling Huang, C.-W. Tung, Y.-H. Chen, and S.-Y. Ho, "Analysis of physicochemical properties on prediction of R5, X4 and R5X4 HIV-1 coreceptor usage," International Journal of Biological and Life Sciences, vol. 5, no. 1, pp. 208, Winter, 2009.
[16] NVIDIA http://www.nvidia.com.tw/page/home.html

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■ 達成目標

□ 未達成目標（請說明，以 100 字為限）

□ 實驗失敗

□ 因故實驗中斷

□ 其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：■已發表 □未發表之文稿 □撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：（以 100 字為限）

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）（以 500 字為限）

協助開發更有效的藥物來提高生命品質，為近年生物資訊研究當紅的重要議題。許多新藥都與人體內的疾病相關蛋白質有關，但是蛋白質分子是由成千上萬個原子所組成，而合適的藥物小分子三維立體結構會有多種構形，通常需從各種已知的龐大化學資料庫中找尋，蛋白質和藥物小分子可能結合的部位也非常多，當藥物小分子與蛋白質分子有效結合，才能達到治療效果。由電腦進行虛擬篩選（virtual screening）在藥物設計過程中便扮演相當重要的功用。本計劃提出一套以受體結合為基礎的快速虛擬篩選方法來尋找資料庫中藥物設計所要的化學小分子。研究進行方式是改善先前發展的 SODOCK 蛋白質嵌合演算法為基礎。本計畫使用直交表設計來對 PSO 初始化的高效能取樣，是一套高度平行化的演算法。PSODOCK 只要改良參數設定便可以有效地設計使用有繪圖處理單元的平行化版本。

國科會補助專題研究計畫項下出席國際學術會議心得報告

日期：99年10月31日

計畫編號	NSC 98 － 2221 － E － 009 － 122 －		
計畫名稱	研發一套以受體結合為基礎的快速虛擬篩選方法		
出國人員姓名	黃慧玲	服務機構及職稱	交通大學生物科技學系 副教授級專案教學人員
會議時間	98年10月9日至 98年10月11日	會議地點	上海
會議名稱	(中文)2009 計算與系統生物學國際會議 (英文) The International Conference on Computational and Systems Biology 2009-ICCSB		
發表論文題目	(中文)利用物化特性預測非典型性外泌蛋白 (英文) Prediction of non-classical secreted proteins using informative physicochemical properties		

一、參加會議經過

本次會議2009-ICCSB為IEEE EMBS (Engineering in Medicine and Biology Society), IASIA(International Association of Scientists in the Interdisciplinary Areas),上海交通大學(Shanghai Jiaotong University)聯合舉辦。會議地點為Howard Johnson Hotel Songjiang Shanghai,日期2009/10/9至2009/10/11。共邀請50多個來自各國演講者進行演說,同時有5間會議室進行來自30多個國家的海報張貼與口頭報告發表學術論文。因中國大陸進年來積極擴展國際化交流,因此本會議自中國大陸各省大學學者與各國優秀學者到此聚會共同交流在生物工程與科技的發展,促進眾多學術和科學界的領導組織共同合作。本人因系所規定不得收研究生，因此加入生資所何副院長信瑩的研究團隊。本人參與發表一篇論文口頭報告，是在 10/10 日下午2 點的廬山廳，都是以該研究團隊的核心技術，即計算智慧所衍生的生物資訊論文。會議期間三天裡，一早由中正機場搭直航班機90分鐘後底達上海機場,再搭30分鐘接駁車到神旺酒店(台灣旺旺公司所有),飯店至會場，從入住的飯店搭乘新開通的9 號線捷運前往會場，9 號線捷運的終點站便是松江新城，再搭計程車前往會場,30分鐘車程,算相當方便。團隊報告當天早上主辦單位安排了四位演講者，後兩位講者的題目我們團對比較有興趣，分別是"Understanding disease-associated amino acid mutations with computational models"與"Physico–chemical principles governing biological processes"，尤其後者與我們所發表的論文題目密切相關，是關於生物作用的物化特性，而我們所發的論文便是利用物化特性來預測蛋白質是否屬外泌蛋白，若能了解蛋白質的位置，不但對於推測蛋白質的功能很有幫助，將來也可以

運用於尋找潛在的生物標記，開發新的檢驗方法。報告者中有些也是工程背景來做生資,對我這工程背景者亦是一大鼓勵，吾有另一層的領域的體會與思考。

本次發表之國際會議論文，利用物化特性預測非典型性外泌蛋白，是利用先前研究特徵選取技術發現物化特性為特徵的預測系統。本論文主要有二項特色:（1）提出一套選取物化特性為特徵的演算法及一套包括 17 個物化特性的特徵集合，並且預測正確率高達 82%。(2)由選出的物化特性的特徵集合，進一步了解非典型性外泌蛋白之序列 N 端的是一重要特性。

由於生物資訊與生物工程是跨領域的研究，因此論文題目較分散，但也因此多廣泛學習，收穫頗豐。我們在會議結束後於 10 月 11 日午間同樣搭乘華航班機返台。

二、與會心得

感謝國科會補助參加國際會議之出國補助，使本人得以出席跨領域生物資訊國際會議，開拓眼界及促進國際觀。每次參加國計會議除了努力讓世界知道臺灣人在研究方面非常認真與相當有能力為心則。此次與生資所何副院長信瑩所帶領的研究團隊參加會議，擴大認識同一研究領域之學者，彼此討論如何改善論文方法之細節亦討論研究之發展方向，本人收穫甚多，相信團隊們也應有相同收穫。

此次同行之碩士研究生洪瓊慧，由口頭報告以及回答發問者之問題，這些種種經歷，本人相信對該生之國際交流能力與信心皆提升。也肯定所研究之預測非典型性外泌蛋白之研究方法是有趣且有意義之議題。

此行本人覺得如同將研究室移到國際，因為每次會後討論就如同在實驗室中方法步驟一一分析討論，進而熱切討論，激勵見解，有此境界當然要感謝生資所何副院長信瑩主導討論並提出精湛見解。

三、考察參觀活動(無)

四、建議

近年來國科會、教育部和學校積極鼓勵年輕研究人員，除鼓勵教師參與會議外，特別是博士班學生，參與大型國際會議，及早進入研究領域的核心，吸取國際研究經驗，以提高國人的研究水準。參加生物資訊國際會議對老師及學生是非常重要的，會議中不但可以得到相關研究的最新發展資訊，認識結交許多相關領域的學者，彼此交換研究心得，更可找到跨領域的學者國際合作，在跨領域的生物資訊研究更是重要。目前研究生已有多管道獲(部份)補助出席國際會議，建議繼續擴大進行。而與中國大陸密切的學術交流亦是趨勢，也能有所激勵國人學界能力與支援。

五、攜回資料名稱及內容

1. 期刊一本
2. 會議論文摘要集一本。

六、其他


International Conference on Computational
and Systems Biology 2009
Oct.9-11,2009
Shanghai Jiao Tong Univ
National Science Foudati
International Association o
Interdisciplinary Areas(IAS
Interdisciplinary Sciences
Chemical Physics and Biop
Acta Physico-Chimica Sinic
www.bioon.com
eeting.edu.cn/
上海交通大學
Shanghai Jiao Tong University
IASIA
10.10.2009

國科會補助專題研究計畫項下出席國際學術會議心得報告

日期：99 年 10 月 31 日

計畫編號	NSC 98 － 2221 － E － 009 － 122 －		
計畫名稱	研發一套以受體結合為基礎的快速虛擬篩選方法		
出國人員 姓名	黃慧玲	服務機構 及職稱	交通大學生物科技學系 副教授級專案教學人員
會議時間	99 年 2 月 26 日至 99 年 2 月 28 日	會議地點	新加坡 新達城會議廳 2F
會議名稱	(中文)2010 第二屆電腦與自動化工程國際會議 (英文) 2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE2010)		
發表論文 題目	(中文)設計一個預測器利用物化特性預測 DNA 結合蛋白 (英文) Designing predictors of DNA-binding proteins using an efficient physicochemical property mining method		

一、參加會議經過

本次會議2010-ICCAE為IEEE Computational Intelligence Society, IACSIT(International Association of Computer Science and Information Technology),電子科技大學(University of Electronic Science and techonlogy)聯合舉辦，此次會議為第二屆舉行，第一屆舉辦於泰國曼谷。本屆會議地點為1 Raffles Boulevard, Suntec City, Singapore 039593,日期2010/02/26至2010/02/28。邀請來自各國演講者進行演說，主要有7個演講者，同時有5間會議室進行來自各個國家的海報張貼與口頭報告發表學術論文。

在本會議中各國優秀學者到此聚會共同交流在自動化工程與電腦科技的發展,促進眾多學術和科學界的領導組織共同合作。本人因系所規定不得收研究生，因此加入生資所何副院長信瑩的研究團隊。本人參與發表一篇論文口頭報告，是在 02/27 下午6點的新達城會議廳3樓312室，都是以該研究團隊的核心技術，即計算智慧所衍生的生物資訊論文。會議期間三天裡，第一天早上由桃園國際機場搭長榮直航班機4小時後底達新加坡機場,再搭20分鐘接駁車到史丹福瑞士飯店(Swissôtel The Stamford),飯店至會場，從入住的飯店步行約10鐘即可到達會議現場,相當方便。團隊報告當天下午主辦單位安排了35位演講者，因為人數非常多，相對的聽很多演講者演講，同時也學到很多，而我們所發的論文是利用物化特性來預測DNA結合蛋白，若能了解蛋白質的特性與DNA結合蛋白的相關性，不但對於推測DNA結合蛋白的功能很有幫助，將來也可以運用於尋找潛在的生物反應訊息傳遞，開發新的預測訊息傳遞路徑方法。報告者中有些也是工程背景來做生資,對我這工程背景者亦是一大鼓勵，吾有另一層的領域的體會與思考。

本次發表之國際會議論文，利用物化特性預測 DNA 結合蛋白，是利用先前研究特徵選取技術發現物化特性為特徵的預測系統。本論文主要有二項特色: (1) 提出一套選取物化特性為特徵的演算法及一套包括 36 個物化特性的特徵集合，並且預測正確率高達 80%。(2)由選出的物化特性的特徵集合，進一步了解 DNA 與結合蛋白序列的功能是一重要特性。

由於生物資訊與生物工程是跨領域的研究，因此論文題目較分散，但也因此多廣泛學習，收穫頗豐。我們在會議結束後於 2 月 28 日午間同樣搭乘長榮班機返台。

二、與會心得

感謝國科會補助參加國際會議之出國補助，使本人得以出席跨領域生物資訊國際會議，開拓眼界及促進國際觀。每次參加國計會議除了努力讓世界知道臺灣人在研究方面非常認真與相當有能力為心則。此次與生資所何副院長信瑩所帶領的研究團隊參加會議，擴大認識同一研究領域之學者，彼此討論如何改善論文方法之細節亦討論研究之發展方向，本人收穫甚多，相信團隊們也應有相同收穫。

此次同行之碩士研究生林意哲，由口頭報告以及回答發問者之問題，這些種種經歷，本人相信對該生之國際交流能力與信心皆提升。也肯定所研究之預測 DNA 結合蛋白之研究方法是有趣且有意義之議題。

此行本人覺得如同將研究室移到國際，因為每次會後討論就如同在實驗室中方法步驟一一分析討論，進而熱切討論，激勵見解，有此境界當然要感謝生資所何副院長信瑩主導討論並提出精湛見解。

三、考察參觀活動(無)

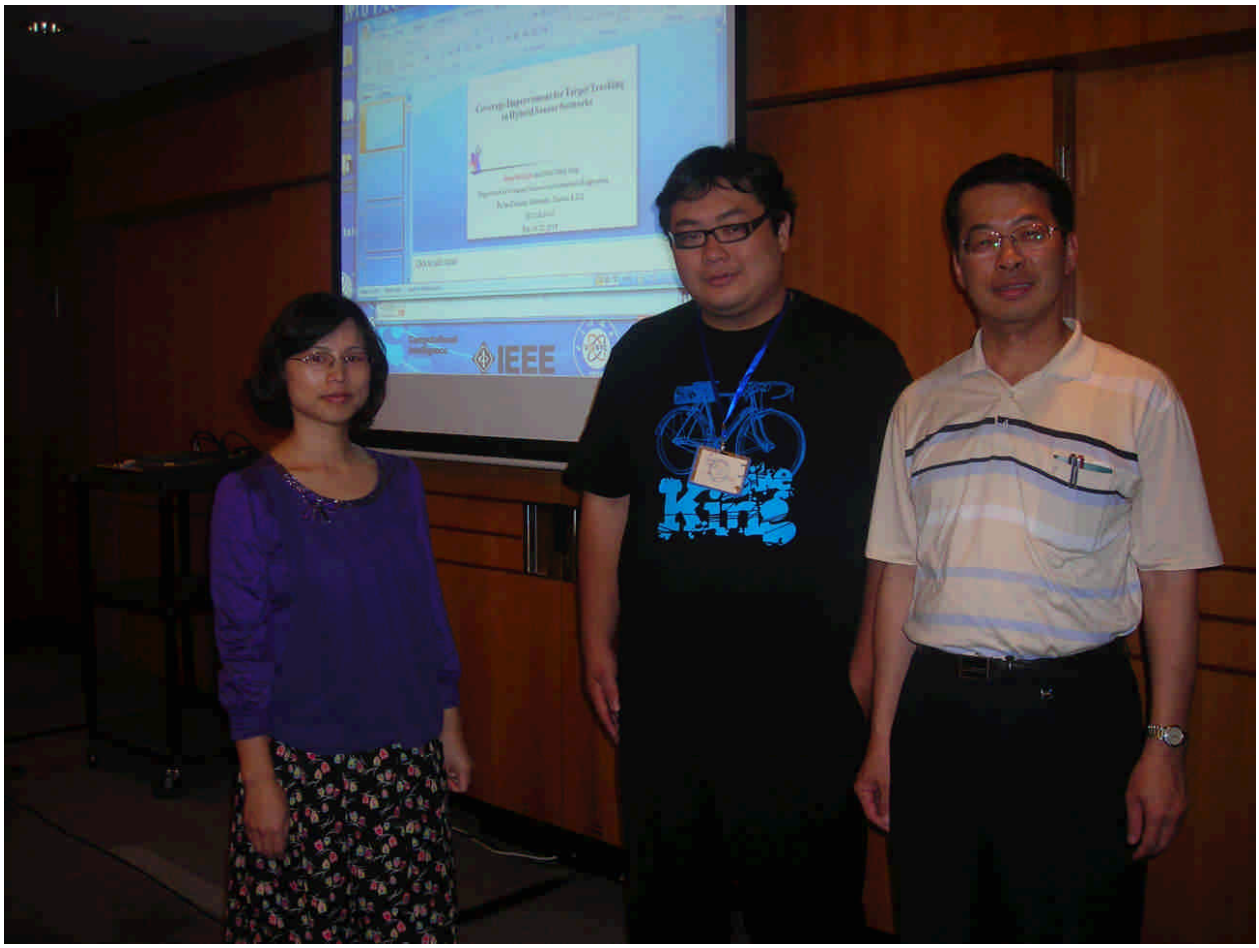
四、建議

近年來國科會、教育部和學校積極鼓勵年輕研究人員，除鼓勵教師參與會議外，特別是博士班學生，參與大型國際會議，及早進入研究領域的核心，吸取國際研究經驗，以提高國人的研究水準。參加生物資訊國際會議對老師及學生是非常重要的，會議中不但可以得到相關研究的最新發展資訊，認識結交許多相關領域的學者，彼此交換研究心得，更可找到跨領域的學者國際合作，在跨領域的生物資訊研究更是重要。目前研究生已有多管道獲(部份)補助出席國際會議，建議繼續擴大進行。而國際化的學術交流是往後的趨勢，也能有所激勵國人學界能力與國際觀。

五、攜回資料名稱及內容

1. 期刊一本
2. 手提公事包一個。

六、其他



無衍生研發成果推廣資料

98 年度專題研究計畫研究成果彙整表

計畫主持人：黃慧玲	計畫編號：98-2221-E-009-122-
計畫名稱：研發一套以與受體結合為基礎的快速虛擬篩選方法	

成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	2	2	100%	人次	
		博士生	1	1	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		
國外	論文著作	期刊論文	3	5	50%	篇	與其他跨領域生物資訊學者學術合作，為3個計畫共同成果,另有研究成果投稿中.
		研究報告/技術報告	0	0	100%		
		研討會論文	3	3	50%		與其他跨領域生物資訊學者學術合作，為3個計畫共同成果
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		

其他成果 (無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)	無

	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：■已發表 □未發表之文稿 □撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：(以 100 字為限)

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）(以 500 字為限)

協助開發更有效的藥物來提高生命品質，為近年生物資訊研究當紅的重要議題。許多新藥都與人體內的疾病相關蛋白質有關，但是蛋白質分子是由成千上萬個原子所組成，而合適的藥物小分子三維立體結構會有多種構形，通常需從各種已知的龐大化學資料庫中找尋，蛋白質和藥物小分子可能結合的部位也非常多，當藥物小分子與蛋白質分子有效結合，才能達到治療效果。由電腦進行虛擬篩選（virtual screening）在藥物設計過程中便扮演相當重要的功用。本計劃提出一套以受體結合為基礎的快速虛擬篩選方法來尋找資料庫中藥物設計所要的化學小分子。研究進行方式是改善先前發展的 SODOCK 蛋白質嵌合演算法為基礎。本計畫使用直交表設計來對 PSO 初始化的高效能取樣，是一套高度平行化的演算法。PSODOCK 只要改良參數設定便可以有效地設計使用有繪圖處理單元的平行化版本。