使用表面幾何特徵於蛋白質與配體結合位置之預測

Using Geometric Features to Predict Protein-ligand Binding Regions

陳明權、林煊庭、羅英倉、白敦文*
國立臺灣海洋大學資訊工程學系
*Email:twp@ntou.edu.tw

摘要–蛋白質和配體的結合是觸發與控制生物分子行為的關鍵，為了有效提昇生物學家在預測蛋白質與配體結合位置的研究分析，並延伸到藥物設計的應用開發，本研究使用蛋白質表面結構的幾何特徵包括立體角、凹洞深度與凹洞體積，提出一個可以快速且不失準確性地預測蛋白質和配體結合區的方法。由幾何立體角的概念，推導蛋白質表面結構的凹洞深度，透過掃描凹洞內的方向線，計算每個凹洞所對應的體積，再使用凹洞深度與體積兩個特徵組合進行結合區域的排名。預測系統的表現是使用 LigASite 所提供的測試資料集進行驗證，並與 SITEHOUND 與 MetaPocket2.0 系統進行比較，實驗結果顯示使用本系統的正確預測比例值達到 82.6%，明顯優於其他兩個知名的預測系統。除此之外，本系統的設計開發亦透過 CUDA 平行運算技術的引進來提昇系統整體計算效能，平均每個蛋白質的表面結構分析與結合位置的預測可以在 1 秒之內完成。

關鍵詞–蛋白質配體結合; 立體角; 凹洞深度；凹洞體積; 計算統一設備架構

ABSTRACT-**Protein-ligand interactions are key processes of triggering and controlling biological functions within cells. To improve the efficiencies of experimental approaches in drug discovery, a system employing geometrical features including solid angle, depth of cavity and volume of cavity is designed for protein-ligand binding region prediction. In this paper, we adopted the solid angle features to obtain the cavity depths and the scanned directions to acquire corresponding cavity volumes. Both depth and volume features were combined to rank predicted binding regions. To verify the system performance, a testing dataset including 388 structures from LigASite was applied and compared to two well-known prediction systems, SITEHOUND and MetaPocket2.0. The result has shown that the proposed system achieving an accuracy rate of 82.6%, and it outperforms these two existing systems. Additionally, the CUDA parallel computing architecture was designed in this prediction system to enhance the computational efficiencies. Each binding region prediction for a query protein structure only requires less than 1 second in general.**

Keyword: **protein-ligand binding; solid angle; depth of cavity; volume of cavity; CUDA**

一、簡介與背景

基於藥物設計開發的需要，蛋白質與配體間的互動結合是研究分子與分子間交互作用中最

重要的一種類型[15]。配體屬於小分子，它會尋求蛋白質巨分子的可結合區域進行結合，結合位置是蛋白質與配體形成具有其他功能化合物的區塊。預測這些可能結合的區域，對於以結構為主的藥物設計應用，可以提供生物學家進行更有效率及正確的生物實驗，對了解蛋白質與配體結合後所產生的生物功能也有很大的幫助。因此本論為希望能夠提供一個快速且不失準確性的技術來預測蛋白質與配體間的結合區位置。

成功預測蛋白質與配體間結合位置的重要因素可分為蛋白質的柔韌性(Protein flexibility)、配體的採樣(Ligand sampling)以及評分函數(Scoring function)等三大部分[10]。分別概述如下：

配體的結合通常是導致蛋白質構造產生改變的主因，改變範圍從小部分的側鏈結構到大區域的功能域移動都有可能。而大區域的移動以及蛋白質本身的柔韌性程度是預測結合位置時最大的難處。近年來對蛋白質柔韌性的研究方法大致上可分為四大類：軟性結合(soft-docking)、側鏈柔韌性程度(side-chain flexibility)、分子鬆弛度(molecular relaxation)以及蛋白質整體結合度(protein ensemble docking)。

配體的採樣是蛋白質與配體結合中最基礎的一環，給予一個蛋白質，執行配體的採樣演算法後就可以推定出結合的位置。配體的採樣也是蛋白質與配體間結合情形的研究成果中最為卓越的一塊。大致上將配體的採樣演算法分為三種：形狀比對(shape matching)、系統化搜尋(systematic search)及隨機決定演算法(stochastic algorithm)。

評分函數亦在蛋白質與配體結合演算法扮演著重要的角色，它直接代表該演算法的結果正確性。速度與準確性是評分函數最重要的兩項指標，理想的評分函數可以同時提昇計算效率及正確性。近年來，許多的評分函數已經陸續提出，大致上可依使用方法及起源分為三種：暴力解法、經驗法則及以知識為基礎的評分函數。基於上述方法的分析、研究及實驗證實後，目前大多的研究結果顯示配體幾乎都結合在蛋白質較深的凹洞中，故近年來許多基於蛋白質表面凹洞來預測蛋白質與配體間結合位置的方法陸續被發表，例如 fpocket[13]、LIGSITE[7]、LIGSITE^{CSC}[9]、SURFNET[11]、CAST[17]、PASS[2]與PocketPicker[16]等，這些系統都是使用純幾何性質進行分析，在不知道配體資訊的前提下進行預測。除了使用幾何形狀的互補特性外，也有使用表面結構的靜電化學特性來強化結合位置的預測分析[6]。

現今在蛋白質結構數據庫(PDB)已經有超過75,000個蛋白質三維結構，該結構資料庫大部分是經由核磁共振、X光繞射或低溫電顯(cryo-EM)技術所解析。經由不同的實驗方式都可以提供蛋白質各個原子在空間中或表面結構的相對位置[1]。依照 PDB 資料庫所提供各個蛋白質內原子的相對位置座標，可將座標對應到數位化的三維網格空間，直接建構數位化的蛋白質結構物件，並透過蛋白質表面區域的特徵分析，判斷表面的網格點並進行後續各種計算及應用。許多預測蛋白質與配體結合位置的技術都是使用蛋白質表面的幾何特徵，例如 Connolly 是第一個使用立體角特徵來當作預測蛋白質表面結合的特性[4]，其中提到如果兩個點可以緊密結合的話，加總的立體角就是 4π ，如圖 1 所示。

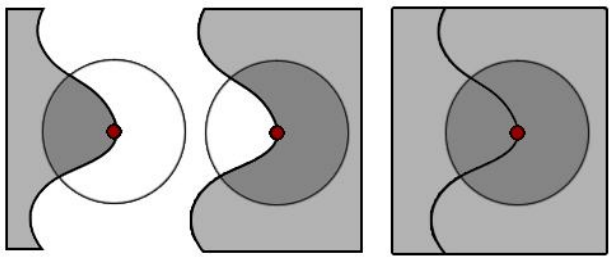

圖 1. 立體角結合示意圖。兩緊密結合的點，其立體角總和為 4π 。深灰色部分是代表蛋白質的互補結合區域。

本論文使用 Connolly 的立體角特徵為基礎進行凹洞深度的計算，並採用 MetaPocket2.0 中使用不同方向的掃描線進行凹洞體積的評估，使用凹洞深度與體積兩項數據來判斷是否為凹洞的依據，最後分別進行正規化後並依照權重分配對每個特徵點的特徵值加總，再依該數值預測可能的結合位置，詳細內容在下一章節介紹。

本系統亦開發 CUDA 平行運算的技術來加快立體角的計算。CUDA 是 NVIDIA 在 GPU 上提出的一個平行處理計算架構，全名是 **C**ompute **U**nified **D**evice **A**rchitecture。GPU 通常用來加速圖形的表現，它可以簡單且快速地產生許多執行緒，將大量重複計算進行平行化的過程，可以有效降低運算所需的時間。現今平行處理技術已經被廣泛的應用到各種不同的研究領域，如密碼學或生物資訊學等不同領域。

二、系統架構與測試資料集

本論文所提出的系統流程圖如圖 2 所示，首先讀取待預測蛋白質的結構檔名或檔案後，經過一系列的模組分析後，可以自動預測該蛋白質可能與配體的結合位置，各步驟的細節如下說明。

圖 2. 系統流程圖。

(一) 三維網格技術建構蛋白質立體結構

本模組的主要功能是快速模擬蛋白質的完整三維空間形狀及辨識蛋白質的表面區域，正確的表面區域定義才能提供後續有效計算表面結構的立體角、凹洞深度及體積等幾合特徵。依照蛋白質數據資料庫所提供的結構資訊檔案(.pdb)中，內容包含組成蛋白質結構各原子的空間相對座標，將其對應至三維座標的網格中(此空間網格的基本單位為 1Å^3)，並依照每個原子獨有的凡德瓦力作為該原子的延伸球半徑，再將每個原子分別對應到三維空間中，即可實際模擬整個蛋白質的構型，接著根據各原子在網格上分佈的位置，可明確定義蛋白質的表面區域，範例說明請見圖 3 與圖 4。

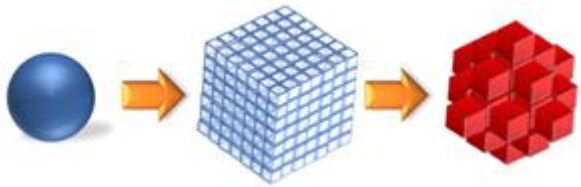

圖 3. 網格化流程圖。將單一原子離散化後對應到三維網格空間，並依每格所重疊的體積比例決定該網格是否屬於該原子所有。

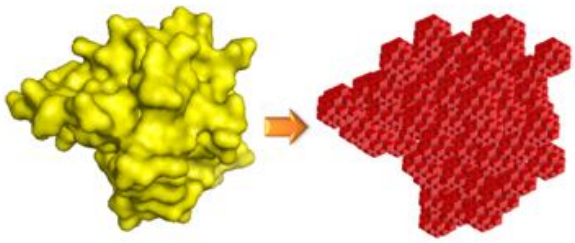

圖 4. 蛋白質經過網格建構後形成之圖形(PDB ID:1ACB)。

(二) 立體角計算

1986 年 Connolly 第一個提出蛋白質表面立體角特徵概念，立體角是從一特定點觀察或測量遠處物體的三維空間角度，定義某一物體投影在特定點的虛擬球上的面積，本論文使用的立體角

虛擬球半徑為 10Å。如圖 5 所示，一物體 A 對於觀察點 O 的立體角，為該物體投影在單位球上的面積Ω，就電腦科學的計算方法，我們可以將立體角的計算使用下面的公式簡化：

$$SA = (inSP / nP) * 4\pi \qquad (1)$$

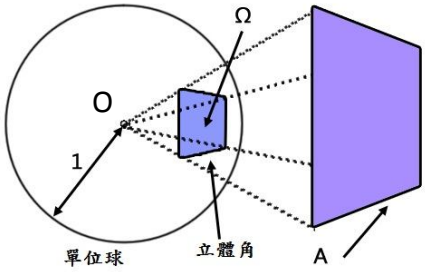

圖 5. 單位球體(球體半徑為 1)的立體角。

其中 *SA* 代表立體角，*inSP* 為虛擬球表面且在蛋白質內部的表面區域面積，用二維的示意圖即為圖 6 中的藍色實線，而 *nP* 則是整個虛擬球表面區域，即為圖 6 的黑色虛線部份。

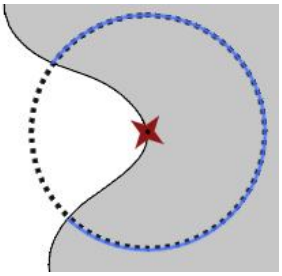

圖 6. 立體角計算的二維示意圖。黑色虛線為整個虛擬球的表面，藍色曲線為虛擬球座落在蛋白質內部的表面區域。

由於本步驟需要計算所有表面格點的立體角特徵值，且蛋白質的表面積隨著蛋白質的組成大小而有所改變，故當輸入的蛋白質結構較為龐大時，除了需要宣告大量的記憶體空間來生成整個蛋白質的模型外，還需要花費長久的計算時間處理表面特徵值的計算，有鑑於此，本研究加入CUDA平行處理技術，憑藉著圖形顯示晶片的快速執行緒產生、絕佳的數值處理能力及多核心平行運算等優點，大大縮短計算所等待的時間，在這裏，我們使用OpenGL來呈現立體角執行的結果，範例如圖7所示。

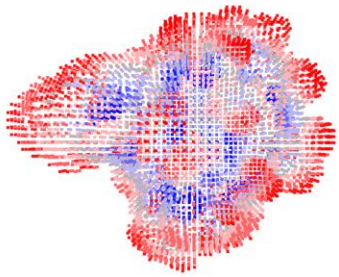

圖 7. 經由立體角運算後的所有表面點。紅點為立體角範圍在 0 至 1.5π 內，代表蛋白質表面較為凸出的區域，灰點為立體角範圍在 1.5π 至 2.5π 內，代表蛋白質表面屬於較平坦區域，藍點則是立體角範圍在 2.5π 至 4π 之間，代表較為凹陷的表面區域。

(三) 立體角分群

當所有表面點都已經算出各自的立體角後，接著將具有同一性質(凸、平、凹)的立體角且分布在限定距離內的表面點(本論文採用 10 Å)結合成為同一群，由於本研究只需辨識屬於凹洞的特徵點作為後續分析，於是僅取出凹洞部分的群組作為預測蛋白質與配體結合的目標，並且將該群中立體角最大的位置作為該分群的代表錨點(anchor)，範例如圖 8 所示。

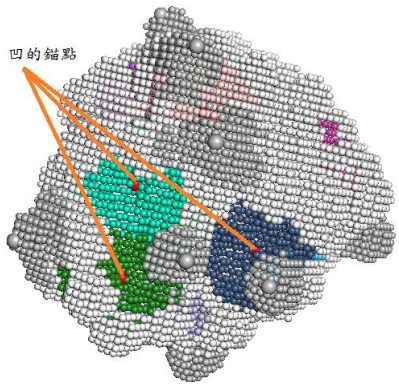

圖 8. 表面點分群示意圖。經分群後不同顏色代表不同群組，每一群組中立體角最大的表面點定義為該凹洞的錨點，如圖中凹洞群組標示為紅點的位置(較深灰色區域是代表凸出的表面群組及凸出群組的錨點位置)。

(四) 凹洞平均深度計算

根據觀察發現，有些錨點雖然具有極大的立體角，但其同一群中鄰近表面區域卻非如此凹陷，此種例外將造成後續程式在判斷時，對不可能與配體結合的區域，卻給予了高度的評分。為了避免此種例外的發生，本研究根據同一群表面點立體角的深淺凹凸特性，另外定義凹洞平均深度的特徵值，以下為特徵值的換算方式：

$$Depth = \begin{cases} 5 & if\ SA > 0.9*4\pi \\ 4 & if\ 0.8*4\pi < SA \le 0.9*4\pi \\ 3 & if\ 0.7*4\pi < SA \le 0.8*4\pi \\ 2 & if\ 0.6*4\pi < SA \le 0.7*4\pi \\ 1 & if\ 0.5*4\pi < SA \le 0.6*4\pi \\ 0 & else \end{cases} \quad (2)$$

SA 代表該點的立體角。依照特徵點分群處理後，再依據上式計算群內所有表面點的深度，最後再把該群所有表面點的對應深度加總並除以該群的總點數作為該表面凹洞的平均深度，範例如圖 9 所示。

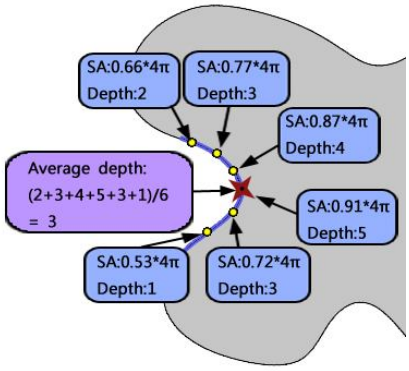

圖 9. 凹洞平均深度計算示意圖。紅點為該群的錨點，黃點為錨點鄰近的表面點，將各個黃點的立體角對應到的深度取平均值後作為錨點的平均深度特徵值。

(五) 凹洞體積計算

凹洞體積在預測配體結合位置時提供相當有效的鑑別力，本論文的計算方法是採用各個錨點當作球心，衍生成半徑為 10Å 的虛擬球，取出位於球內且為蛋白質外部的空間點，每個空間點各自延伸七個方向的座標軸(使用空間單位向量表示，分別為 $\langle 1, 0, 0\rangle$, $\langle 0, 1, 0\rangle$, $\langle 0, 0, 1\rangle$, $\langle \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\rangle$, $\langle -\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\rangle$, $\langle \frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\rangle$, 及 $\langle \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}\rangle$。若該方向座標軸延伸後的兩端皆有接觸到蛋白質本體，則屬於內部線，否則視為外部線。當設定的七條方向線中，若該點擁有超過四條以上的內部線，則該空間點即認定屬於凹洞內部的一部分，最後計算所有凹洞內部點的個數做為該特徵點所在的凹洞體積，範例的示意圖如圖 10 所示。

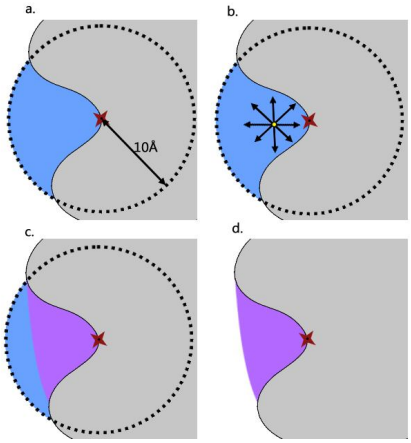

圖 10. 以二維概念說明凹洞體積的計算流程。a. 由特徵點(紅點)向外延展出半徑為 10Å 的虛擬球(虛線)，藍色區域為初始的外部點；b. 將藍色區域中的所有點延伸至七個方向的軸線並檢測交集點的數量；c. 紫色區域為所有通過篩選的點；d. 留下的紫色區域定義為該凹洞的體積。

(六) 預測配體位置

將各群的凹洞平均深度及凹洞體積分別正規化後，再將兩項數值以 8:2 的權重比例加總得到的值由大到小進行順序排名(權重比例是經過訓練資料集進行測試後建議的最佳比例值)，最後取前三名的群集作為本系統預測蛋白質與配體可能的結合位置。分數計算如以下公式所列：

$$RV(p) = \frac{CD(p)_{AVG}}{CD_{MAX}} * 0.8 + \frac{CV(p)_{AVG}}{CV_{MAX}} * 0.2 \quad (3)$$

其中 RV(p)(Rank Value)為單一錨點 p 最後排名的依據，$CD(p)_{AVG}$ 及 $CV(p)_{AVG}$ 為錨點 p 點的凹洞深度及凹洞體積，CD_{MAX} 及 CV_{MAX} 是指該蛋白質表面結構的所有錨點中最大的凹洞深度及凹洞體積值。

(七) 測試資料集

本研究實際測試的資料集是 2011 年 7 月由 LigASite[5] 釋出的第 9.5 版 (http://www.bigre.ulb.ac.be/Users/benoit/LigASite/index.php)資料集。總共包含 388 條不重覆且未結合的蛋白質結構(unbound proteins)。LigASite 不僅提供了蛋白質與配體結合前後的蛋白質 PDB 編號，該資料庫更精確的列出每一條蛋白質結合位置的胺基酸編號。

三、研究結果

(一) 評估效能說明

將 LigASite 提供的胺基酸編號分類為結合位置胺基酸以及非結合位置胺基酸，而預測的結果也分別表示某胺基酸座落在預測的結合位置及預測的非結合位置，由實際結合位置及預測結合位置的分佈情形可以合併形成四種可能的預測結果類型，並分別以 TP、FP、TN 及 FN 表示，定義範例如圖 11 所示。

由以上四種基本預測結果數據的組合，可以進行統計分析並判斷預測系統的表現。常用來評斷及驗證預測工具好壞的指標如下列五項，包括敏感性、特異性、準確性、PPV 值及馬修相關係數，這些項目也分別用來檢測本論文所提出的特徵好壞及系統表現。

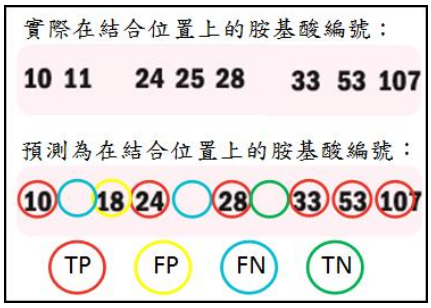

圖 11. 實際結合位置與預測結合位置之驗證方法說明。TP(true positive)指實際的結合位置同時也被正確預測為結合位置(如第 10, 24, 28, 33, 53, 107 號胺基酸)；TN(true negative) 指實際上不在結合位置同時也被正確預測為非結合區位置的胺基酸；FP(false positive) 指實際上不在結合位置卻被錯誤預測為結合區位置的胺基酸(第 18 號胺基酸)；FN(false negative) 指實際上是結合位置但未被正確預測為結合區位置的胺基酸(如第 11 及第 25 號的胺基酸)。

a) 敏感性(Sensitivity)

$$Sensitivity = \frac{TP}{TP + FN} \quad (4)$$

b) 特異性(Specificity)

$$Specificity = \frac{TN}{TN + FP} \quad (5)$$

c) 準確性(Accuracy)

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (6)$$

d) PPV(Positive predictive value)

$$PPV = \frac{TP}{TP + FP} \quad (7)$$

e) 馬修相關係數 (Matthew's correlation coefficient，MCC)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (8)$$

(二) 系統比較

本論文所提出的預測方法與兩個最近所發表的蛋白質配體預測系統進行比較。其中一個是2009 年 4 月由美國的 Mount Sinai School of Medicine 發表於 Nucleic Acids Research 的SITEHOUND 系統[8]，該系統將化學性質相近的視為同一群，再利用碳原子的探針球(Carbon probe)及磷酸鹽的探針球(Phosphate probe)來判定是否為結合位置，並整理出結合強度地圖(Affinity Maps)來表示蛋白質互動的情況，最後得到 TIE(Total Interaction Energy)值來進行排名，並決定哪一群為預測的結合位置 ，該系統提供有網頁版及單機版本。另一個系統是 2011 年 6 月由德國 Technical University of Dresden 發表於 Bioinformatics 的 MetaPocket2.0 系統[18]，結合了 LIGSITE^{CSC}, PASS, QsiteFinder[12], SURFNET, Fpocket , GHECOM, ConCavity[3]和 POCASA 共八種方法，LIGSITE^{CSC} 使用 Connolly 表面及保留的等級來判斷，PASS 用幾何的方式找出深埋的區域，並根據該區域的形狀、尺寸、延伸距離等來決定是否為結合位置，QSiteFinder 用反應的能量以及凡得瓦力的探針球來篩選出可能的結合位置，SURFNET 可以將蛋白質表面，凹洞等資訊進行視覺化的處理，Fpocket 使用 Liang and Edelsbrunner 提出的阿爾法球(alpha spheres)的概念[14]，GHECOM 根據數學型態學的理論(mathematical morphology)與不同大小的探針球來尋找凹洞，ConCavity 根據演化的序列保留性與三維結構來預測，POCASA 則是使用滾動的球體來偵測蛋白質上的口袋(pockets)與凹洞(cavities)。最後依據這些方法預測出的結果進行投票來決定最後預測的結合位置，該系統僅提供網頁版本，本論文將測試用的蛋白質資料集逐一上傳至該平台進行分析與比較。

本論文使用先前所敘述的資料集 LigASite，共有 388 條不重覆的蛋白質結構，分別就未結合(unbound)的蛋白質結構及對應之已結合(bound)結構為主要驗證資料，將欲驗證資料分別輸入到這兩個系統以及本系統分別進行預測並針對收結果進行比較。本論文之評估方式為定義單一蛋白質預測結果的敏感性必須大於 25%，才代表已經成功預測該蛋白質與配體結合的位置。

I. 未結合蛋白質資料集的測試

如表一所示，在 388 條未結合的蛋白質資料集中，本系統可成功預測出其中 312 條蛋白質可與配體結合的位置，成功率為 80.4%，而 SITEHOUND 僅有 224 條可以被成功預測出結合位置，成功率為 57.7%，而 MetaPocket2.0 的成功預測比例僅有 40.5%。本系統在可能結合位置的預測表現明顯優於其他系統。除此之外，在分析預測每筆蛋白質結構的處理時間上，本系統使用 CUDA 的平行計算架構設計，計算效率明顯的優於其他兩個系統。表二為本系統各別與其他系統預測表現的比較表格，由於在輸入 LigASite 資料集所提供的 388 個蛋白質，兩個預測系統皆有無法處理的蛋白質結構，為了公平的進行驗證比較，本系統僅挑選出兩系統皆能順利執行預測步驟的蛋白質進行驗證比較，在與 SITEHOUND 進行比較時，扣除 SITEHOUND 無法執行的 15 條蛋白質後，本系統的預測結果無論在敏感性、特異性、準確性、PPV 值及馬修相關係數的統計分析結果中，皆高於 SITEHOUND 系統，而與 MetaPocket2.0 進行比較時，扣除無法分析預測結果的 207 條蛋白質，僅剩下的 181 條個結構進行比較。在執行結合位置預測後進行統計各種實驗數據，我們的系統除了敏感性外，其他指標都優

於 MetaPocket2.0，其中原因是 MetaPocket2.0 系統偏向預測較多的結合位置胺基酸，故在敏感性指標可以大幅的提升，但是相對降低了其他的統計分析值表現。

II. 已結合蛋白質資料集的測試

如表三所示，對於 388 條已結合蛋白質資料集的測試中，本系統仍然可以正確預測出大部分蛋白質(329 條)的結合位置，成功預測的機率為 84.7%；而 SITEHOUND 只有 292 條可以被成功預測出正確的結合位置，成功率為 75.3%；MetaPocket2.0 的預測結果只有 144 條可以成功被預測出結合位置，成功比例僅有 36.1%。由上述的預測結果，本系統無論蛋白質結構在未結合或已結合的情況下，皆有很好的辨識能力。

在表四中顯示與 SITEHOUND 的效能比較，扣除 SITEHOUND 無法執行的 14 條蛋白質後，本系統各項分析的指標皆優於 SITEHOUND，代表本論文所提出的系統對於已結合蛋白質的配體結合位置預測的準確率優於 SITEHOUND。相對於 MetaPocket2.0，該系統無法順利執行 240 個蛋白質結構的預測，數量遠多於可以順利執行預測的 144 個結構，若僅使用該 144 個結構進行統計分析，本系統在特異性、準確性、PPV 這三項的結果仍優於 MetaPocket2.0 的表現。

(三) CUDA 加速成果

圖 12、13 為單純使用 CPU 的計算與加入 GPU 加速計算後所使用的運算時間比較。測試的環境如下：CPU 為 Intel® Core™2 Duo Processor E8400、4GB DDR2 記憶體、顯示卡為 NVIDIA Geforce GTS450、作業系統為 Windows XP 32-bit。測試的蛋白質結構為前述已結合與未結合的兩組 388 個蛋白質結構。在 388 個已結合之蛋白質資料集中，所含胺基酸個數範圍從 58 到 4,521 個，原子個數範圍從 530 到 34,156 個，表面點個數範圍從 4,513 到 162,159 ，只使用 CPU 計算每個蛋白質的立體角，預測所需的平均時間為 7.03 秒，若加入 GPU 的平行計算架構，平均時間僅為 0.64 秒，平均加速的倍數為 10.98 倍，如圖 12 所示。

圖 13 是針對未結合的 388 個蛋白質結構進行系統效率分析。胺基酸個數範圍從 58 到 4,520 個，而蛋白質的原子個數範圍從 454 到 34,186 個，表面點個數範圍從 4,510 到 141,201。僅使用 CPU 計算每個蛋白質的立體角，預測所需的平均時間為 6.51 秒，若使用 GPU 計算架構強化運算過程，平均時間可降至 0.59 秒，平均加速的倍數可提昇 11.03 倍。

四、總結

立體角於 1986 年被正式提出後，生物學家藉由此幾何特性觀察蛋白質表面上的凹凸情況，本論文根據此方法偵測蛋白質表面上的凹陷部分，進一步運用附近表面原子的立體角，算出該特徵點所在凹洞的平均深度及凹洞體積作為主要與配體結合位置的預測依據。本篇論文的研究目的在於提供一個快速且不失準確性的方式來預測蛋白質與配體間的結合區域，在系統驗證所使用的資料集方面，我們採用 LigASite 提供的未結合與已結合蛋白質結構各 388 條不重複的結構作為測試資料集，其中本系統對未結合蛋白質結構資料集的正確預測率為 80.4%，已結合蛋白質結構資料集的正確預測率為 82.6%，兩者皆明顯超越兩個知名系統的預測率。除此之外，透過 CUDA 的 GPU 平行運算技術，平均每個蛋白質結構僅需 0.65 秒即可全自動完成該蛋白質與配體結合位置的分析與預測。

致謝：本論文感謝國科會(計畫編號:

NSC98-2221-E-019-031-MY2 及 NSC100-2321-B-019-004)及國立台灣海洋大學海洋生物科技及環境生態中心之計畫經費贊助。

五、參考文獻

[1] A. M. Bonvin, R. Boelens, and R. Kaptein, "NMR analysis of protein interactions," *Current opinion in chemical biology,* vol. 9, pp. 501-8, Oct 2005.

[2] G. P. Brady, Jr. and P. F. Stouten, "Fast prediction and visualization of protein binding pockets with PASS," *Journal of computer-aided molecular design,* vol. 14, pp. 383-401, May 2000.

[3] J. A. Capra, R. A. Laskowski, J. M. Thornton, M. Singh, and T. A. Funkhouser, "Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure," *PLoS computational biology,* vol. 5, p. e1000585, Dec 2009.

[4] M. L. Connolly, "Measurement of Protein Surface Shape by Solid Angles," *Journal of Molecular Graphics,* vol. 4, pp. 3-8, Mar 1986.

[5] B. H. Dessailly, M. F. Lensink, C. A. Orengo, and S. J. Wodak, "LigASite--a database of biologically relevant binding sites in proteins with known apo-structures," *Nucleic Acids Research,* vol. 36, pp. D667-73, Jan 2008.

[6] H. A. Gabb, R. M. Jackson, and M. J. Sternberg, "Modelling protein docking using shape complementarity, electrostatics and biochemical information," *Journal of molecular biology,* vol. 272, pp. 106-20, Sep 12 1997.

[7] M. Hendlich, F. Rippmann, and G. Barnickel, "LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins," *J Mol Graph Model,* vol. 15, pp. 359-63, 389, Dec 1997.

[8] M. Hernandez, D. Ghersi, and R. Sanchez, "SITEHOUND-web: a server for ligand binding site identification in protein structures," *Nucleic Acids Research,* vol. 37, pp. W413-6, Jul 1 2009.

[9] B. Huang and M. Schroeder, "LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation," *Bmc Structural Biology,* vol. 6, p. 19, 2006.

[10] S. Y. Huang and X. Zou, "Advances and challenges in protein-ligand docking," *International journal of molecular sciences,* vol. 11, pp. 3016-34, 2010.

[11] R. A. Laskowski, "SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions," *Journal of Molecular Graphics,* vol. 13, pp. 323-30, 307-8, Oct 1995.

[12] A. T. Laurie and R. M. Jackson, "Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites," *Bioinformatics,* vol. 21, pp. 1908-16, May 1 2005.

[13] V. Le Guilloux, P. Schmidtke, and P. Tuffery, "Fpocket: an open source platform for ligand pocket detection," *BMC Bioinformatics,* vol. 10, p. 168, 2009.

[14] J. Liang, H. Edelsbrunner, and C. Woodward, "Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design," *Protein science : a publication of the Protein Society,* vol. 7, pp. 1884-97, Sep 1998.

[15] A. K. a. J. Ojanen, "Protein docking," Nov 2002.

[16] M. Weisel, E. Proschak, and G. Schneider, "PocketPicker: analysis of ligand binding-sites with shape descriptors," *Chem Cent J,* vol. 1, p. 7, 2007.

[17] T. Yamazaki, Y. Hamano, H. Tashiro, K. Itoh, H. Nakano, S. Miyatake, and T. Saito, "CAST, a novel CD3epsilon-binding protein transducing activation signal for interleukin-2 production in T cells," *J Biol Chem,* vol. 274, pp. 18173-80, Jun 25 1999.

[18] Z. Zhang, Y. Li, B. Lin, M. Schroeder, and B. Huang, "Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction," *Bioinformatics,* vol. 27, pp. 2083-8, Aug 1 2011.

表一. 388 條未結合蛋白質結構中成功預測出結合位置的個數統計表。Top1 表示預測為第一名的結果，Top2 及 Top3 以此類推。在預測結果為前三名的正確性以本論文所提出的系統表現最佳。

System	Top1	Top2	Top3	Total	Success Rates	Cost time(sec) (**單筆預測**)
本系統	**151**	**100**	**60**	**311**	**80.4%**	**0.2-5**
SITEHOUND	109	72	43	224	57.7%	60-600
MetaPocket2.0	113	31	13	157	40.5%	20-120

表二. 在 388 條未結合蛋白質結構中各項評比指標的比較表。本系統在整體比較的參數表現優於 SITEHOUND 及 MetaPocket2 系統

System	Sensitivity	Specificity	Accuracy	PPV	MCC
取其中 373 條預測結果(去除 SITEHOUND 沒有跑出結果的 15 條)					
本系統	**0.530**	**0.968**	**0.934**	**0.580**	**0.510**
SITEHOUND	0.379	0.955	0.912	0.399	0.332
取其中 181 條預測結果(去除 MetaPocket2 沒有跑出結果的 207 條)					
本系統	0.536	**0.967**	**0.935**	**0.566**	**0.504**
MetaPocket2	**0.710**	0.904	0.878	0.478	0.500

表三. 在 388 條已結合蛋白質結構中成功預測出結合位置的個數表。在預測結果為前三名的正確性以本論文所提出的系統表現最佳。

System	Top1	Top2	Top3	Total	Success Rates	Cost time(sec) (**單筆預測**)
本系統	145	**111**	**73**	**329**	**84.7%**	**0.2-5**
SITEHOUND	**159**	87	46	292	75.3%	60-600
MetaPocket2.0	92	27	21	140	36.1%	20-120

表四. 在 388 條已結合的蛋白質結構中各項評比參數比較表。

System	Sensitivity	Specificity	Accuracy	PPV	MCC
取其中 374 條預測結果(去除 SITEHOUND 沒有跑出結果的 14 條)					
本系統	**0.621**	**0.975**	**0.952**	**0.625**	**0.585**
SITEHOUND	0.538	0.970	0.943	0.541	0.496
取其中 148 條預測結果(去除 MetaPocket2 沒有跑出結果的 240 條)					
本系統	0.616	**0.973**	**0.948**	**0.627**	0.583
MetaPocket2	**0.861**	0.912	0.905	0.556	**0.634**

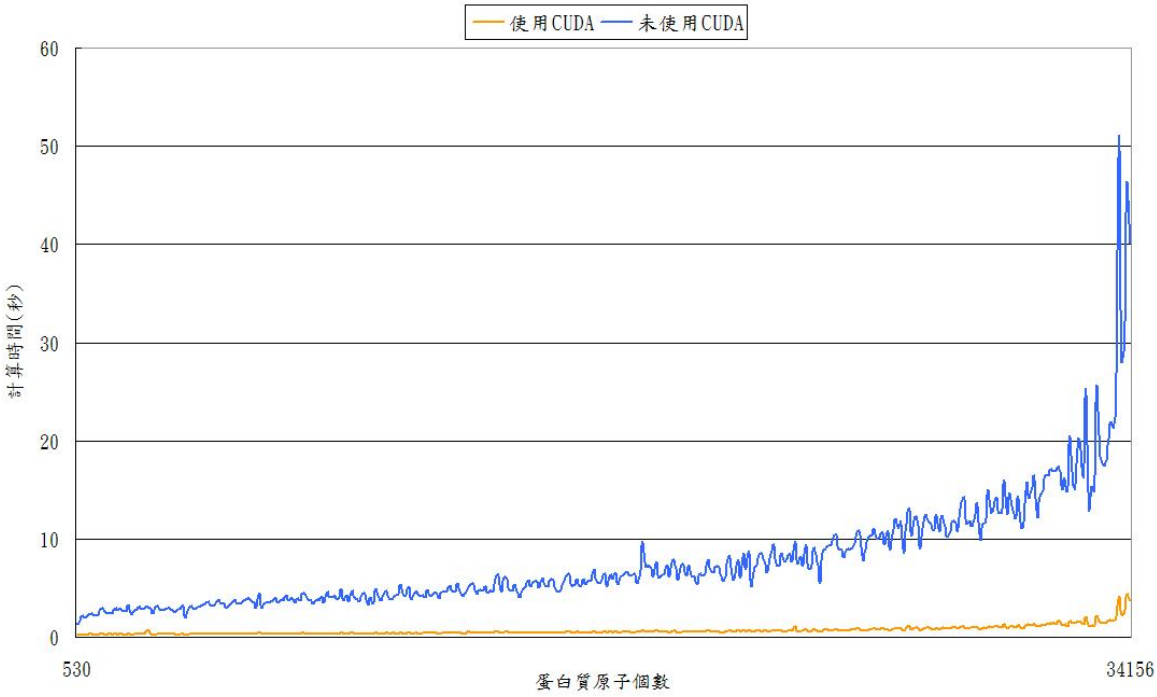

圖 12. 執行時間比較圖(已結合蛋白質)。對 388 個已結合蛋白質進行預測，使用(橘色)與不使用(藍色)CUDA 技術計算架構下，所需計算時間的比較折線圖。

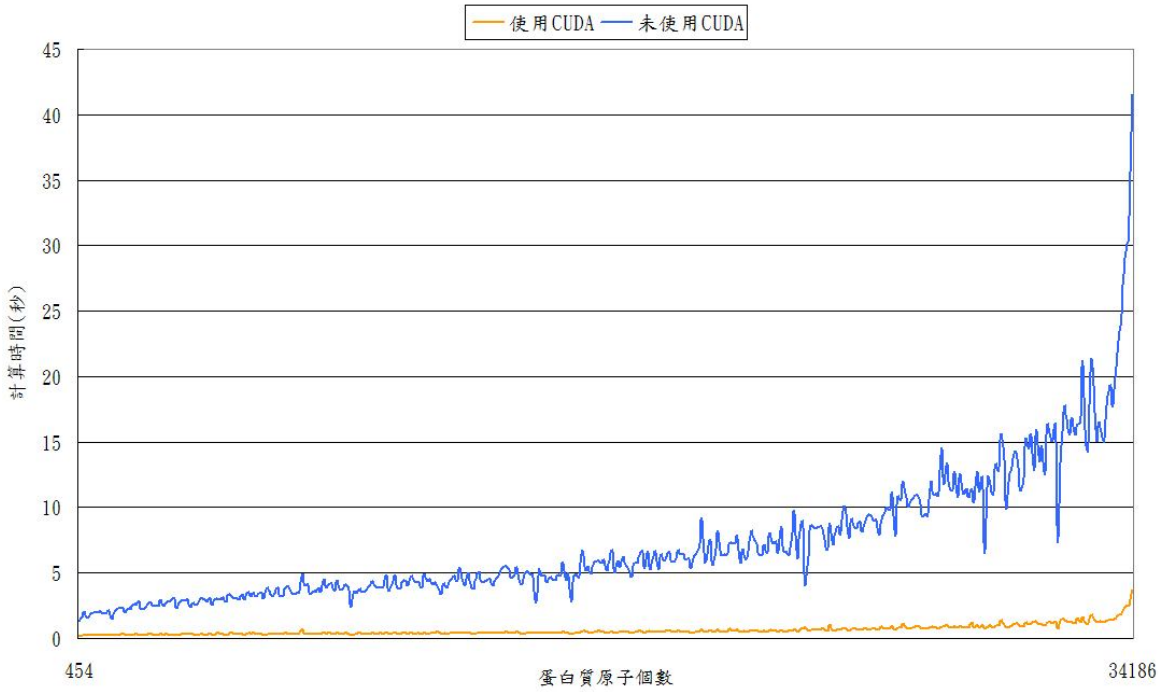

圖 13. 執行時間比較圖(未結合蛋白質)。對 388 個未結合蛋白質進行預測，使用(橘色) 與不使用(藍色)CUDA 技術計算架構下，所需計算時間比較折線圖。