國立交通大學

生物資訊及系統生物研究所

碩士論文

利用空間相關結構片段快速尋找蛋白質

結合片段與環境之研究

Space-Related Pharmamotifs for fast search

protein binding motifs and environments

研 究 生：張力仁

指導教授：楊進木　教授

中 華 民 國 九 十 九 年 七 月

利用空間相關結構片段快速尋找蛋白質結合片段與環境之研究

Space-Related Pharmamotifs for fast search protein binding motifs and environments

研 究 生：張力仁　　Student：Li-Zen Chang

指導教授：楊進木　　Advisor：Jinn-Moon Yang

國 立 交 通 大 學

生 物 資 訊 及 系 統 生 物 研 究 所

碩 士 論 文

A Thesis Submitted to Institute of Bioinformatics and Systems Biology

College of Biological Science and Technology

National Chiao Tung University in partial Fulfillment of the Requirements

for the Degree of Master in

Bioinformatics and Systems Biology

July 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年七月

利用空間相關結構片段快速尋找蛋白質結合片段與環境之研究

學生：張力仁　　指導教授：楊進木

國立交通大學 生物資訊及系統生物研究所

摘　　要

了解藥物或是化合物其潛在具有結合能力的蛋白質是相當重要地。如在早期藥物發展，即能夠偵測出可能造成的副作用而避免無謂的藥物開發成本與時間；同時，對於已知功能的藥物也能予以提供新的治療用途；並且，目前對於治療具有複雜機制的疾病如癌症及糖尿病, 同時針對疾病中多個標靶蛋白的藥物也是現今藥物發展的一個新策略。所以，尋找可能與化合物結合的潛在標靶蛋白質(potential target proteins)對於學術研究以及臨床藥物開發上都是一大重要議題。

一般相信擁有相似結合環境的蛋白質能與相同或相似結構的化合物產生交互作用。以往研究者常利用比對蛋白質序列或搜尋相似蛋白質結構來尋找潛在的蛋白質，然而過去研究顯示，能結合相同化合物的蛋白質在序列或整體結構上並非都有明顯的相似性及演化關係，而是只在結合環境(binding environment)上相似。因此，針對尋找擁有相似結合環境的蛋白質，我們提出空間相關結構片段(Space-Related Pharmamotif, SRP)的概念，藉由 SRP 來搜尋擁有相似結合環境的潛在蛋白質，進一步瞭解蛋白質－化合物交互作用與結合環境的關係。

SRP 是由一組鄰近蛋白質－小分子結合位(binding site)、且三度空間中不連續的蛋白質結構片段所組成，相較於以往比對完整序列或整體結構相似度來尋找潛在蛋白質的方法，我們藉由 SRP 來描述蛋白質與小分子的結合環境，並以 3D-BLAST 將三級結構片段轉換編碼成一級序列，對目前所有已知蛋白質結晶結構進行快速搜尋，尋找擁有類似 SRP 的蛋白質。進一步結合結構比對工具如 DALI，標定相似結合環境在潛在蛋白質中的空間位置。

我們蒐集 530 個蛋白質－藥物分子結晶結構，共 187 種美國食品藥品監督管理局(FDA)核准的藥物，以此建構個別 SRP 並對蛋白質結構資料庫(PDB)進行搜尋。對於搜尋相同蛋白質或共結晶相同化合物的結晶結構，SRP 的覆蓋率(recall)分別為 80%和 54%。針對搜尋整體或區域結構相似的蛋白質，SRP 準確率(precision)可以達到 82%，說明我們的方法在尋找相似結合環境上可以提供可靠的預測結果。同時在本研究中，我們以治療流感的藥物瑞樂沙(Zanamivir)為例，說明 SRP 對結合環境的敏感性以及可能如何應用於蛋白質分類問題上，另一實例以治療慢性骨髓細胞白血病藥物 Imatinib，說明 SRP 應用在「舊藥新用」議題上的可能性。

最後，我們建置一網站提供 530 個蛋白質－藥物結晶結構所建構的 SRP 資訊以及搜尋結果，用以觀察討論 187 種藥物可能的潛在結合蛋白質。本研究利用空間相關結構片段(SRP)來尋找相似結合環境的蛋白質，期望能對瞭解蛋白質－化合物交互作用與結合環境的關係、以及探討藥物於舊藥新用或副作用的研究有所幫助。

Space-Related Pharmamotifs for fast search protein binding motifs and environments

Student: Li-Zen Chang Advisor: Dr. Jinn-Moon Yang

Institute of Bioinformatics and System Biology
National Chiao Tung University

Abstract

It is important to understand the potential target proteins for a chemical compound. During the early drug discovery stage, for example, it could avoid the unnecessary developing cost and time by detecting the potential harmful side effects. On the other hand, it could provide the new usages for old drugs. Recently, multiple target drugs give a new paradigm for diseases with complex mechanism such as cancers and diabetes. Therefore, discovering potential target proteins of a given compound is a valuable issue in bioinformatics and drug development.

Previous studies indicate that similar compounds enable to bind the proteins with similar binding environment. Researchers usually search similar proteins by aligning the given protein sequence or global protein structure in sequence or structure databases. However, previous works show that in some cases proteins bound the same ligand may not have significant evolutionary relationship in both sequence and global structure but in their binding environments. In this study, we introduce a concept named Space-Related Pharmamotif (SRP) to discover the proteins with similar binding environment in protein databases.

SRP is composed of a set of spatially discontinuous peptide segments, which surround the ligand-binding site. Compared with the previous methods of finding proteins with similar sequence or global structure, SRP focuses on protein-ligand interacting environment. By transforming the 3D structure segments into 1D structural alphabet sequences through 3D-BLAST, we can search the potential target proteins with similar binding environment against Protein Data Bank (PDB) rapidly. Furthermore, we use the structure alignment tool, such as DALI, to precisely locate the possible binding environment in these target protein structures.

We collect 530 protein-drug co-crystallized complexes, in which contain 187 different FDA-approved drugs. We build SRPs and screen PDB for each protein-drug complex. For searching the proteins with the same UniProt accession number and the same ligand, the recall achieves 80% and 54%, respectively. Proteins classified into the same homologous superfamily of CATH can be predicted with a precision of 82%. Our results demonstrate that SRP provides a reliable performance in searching the potential target proteins with similar

binding environment. We give an example of Zanamivir to describe how SRP can identify slight structural difference of the binding environments between proteins. In another example, we preliminarily discuss the issue of “new use for old drugs” about Imatinib, which is a marking drug known to against disease chronic myelogenous leukemia and gastrointestinal stromal tumor.

Finally, we build a web server to represent the SRP information and the searching results from 530 protein-drug complexes for helping to identify the potential binding protein of 187 known drugs. In this study, we supply evidence to present that SRP is reliable for searching the potential target proteins with similar binding environment. In the future, we will develop SRP to be useful to understand protein-ligand interactions and helpful for drug design.

誌謝

在交大研習碩士學位的兩年中，對於曾經給予熱忱協助以及諄諄教誨的良師益友，力仁由衷地感激與感恩，少了任何人在這當中的扶持是沒有辦法順利完成這趟學習之旅。

力仁十分感謝指導教授楊進木老師，對於學生的驚頓您仍給予耐心的指導及提點、適時導正學生的研究方向，力仁始終感激於心；而您於研究上所抱持的熱忱與專注，以及追求真理的執著，讓力仁體會到做研究應該持有的態度，感謝老師樹立這般學者風範；對於人生的道路上，認真、負責、態度這些再基本不過的觀念，卻是學生在老師身上所體認到最重要的人生哲理，謝謝老師！

同時，感謝實驗室的學長姐和各位同學們，感謝在研究上同為一組的其樺、章維和一原學長們在研究上與力仁的討論交流與幫助，力仁受益良多。感謝俊辰、阿甫、PIKI、宇書及志達學長們在研究上的種種幫助和論文的修訂。感謝怡馨學姊，妳始終是帶給實驗室愉悅氣氛的康樂股長。峻宇、超哥和韋帆，感謝我們是同一屆，讓我有伴打電動跟打球！也感謝學弟妹們在實驗室中給予的協助，謝謝大家在研究上和生活上的關心與幫助，力仁銘記於心。

當然，感謝我最親愛的家庭，沒有你們在背後給予最大的支持與關愛，我是走不完的。謝謝親愛的老爸與老媽，能有你們一對這麼愛孩子的父母，是我最大的幸福！也謝謝我親愛的老弟，平時的噓寒問暖始終是讓我再努力向前的動力。

最後感謝盧錦隆老師與楊昀良老師能參與力仁的碩士學位口試，給予力仁指導與建議。很幸運能來到 BioXGEM 這個大家庭，在這度過兩年成長磨練的時間，力仁十分地感恩，謝謝！

總目錄

表目錄

圖目錄

壹、緒論

一、研究背景

蛋白質－化合物交互作用在學術研究以及藥物研發上一直都是十分重要的議題。過往研究指出，開發一個新的治療疾病藥物，需要五億乃至二十億美金的高額研發經費[1]。然而，近三分之一研發中的藥物因會與某些非目標蛋白質結合(off-target receptors)，造成無法接受、具嚴重傷害性的副作用而使得藥物無法上市[2]。對於可能與未知蛋白質發生交互作用而導致研發中藥物喪失開發價值，無論在經濟層面或是臨床藥物研究上都是一大損失。

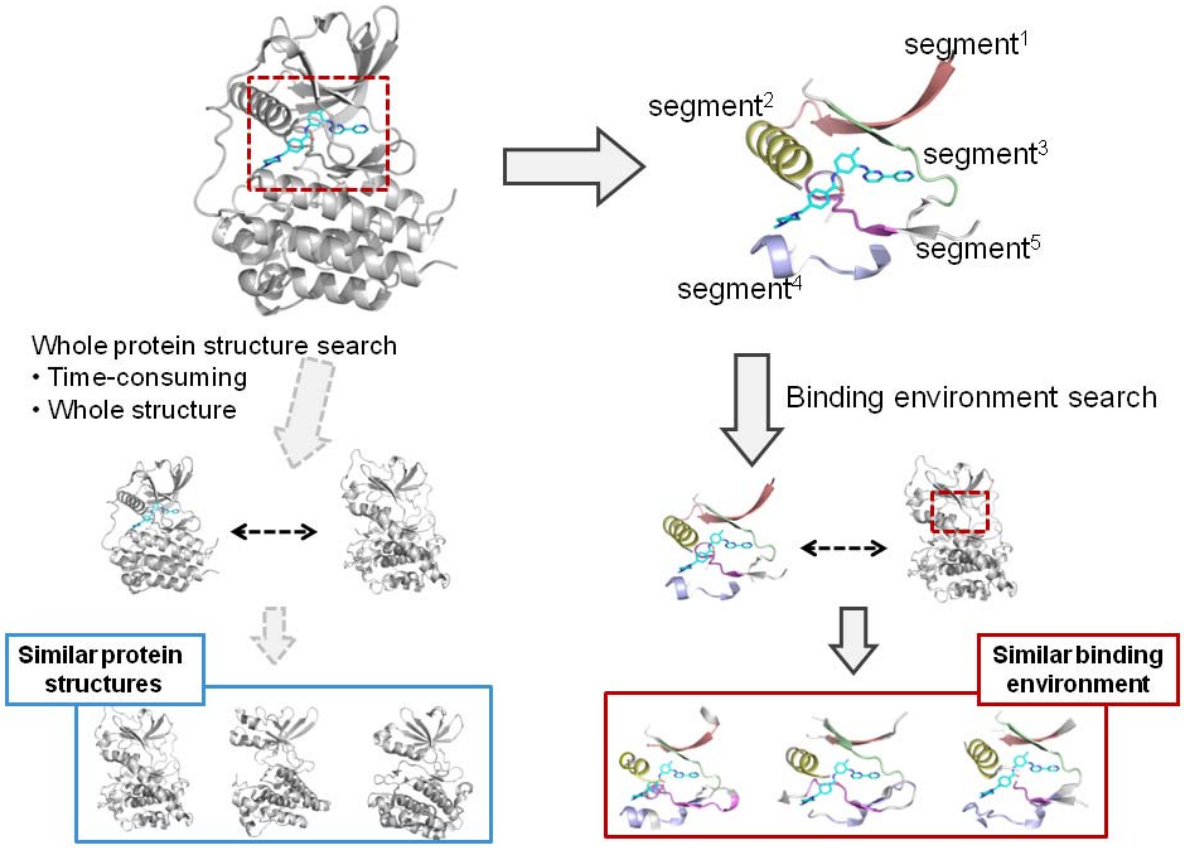
同時，以往在藥物開發時往往遵循「單一基因、單一藥物、單一疾病」的研發邏輯，著重於只針對一個標靶目標即可治療疾病的藥物理念。然而如此具有高度專一性的藥物可能容易因其標靶目標上一個胺基酸的改變以致降低藥物的結合能力而失去效用。相較於單一藥物抑制單一標靶目標，同時針對多個不同的標靶目標作用則不易失去效用，且研究指出具有多重藥理性質(polypharmacological)機制的藥物通常也能擁有著更好的治療疾病能力[3]。

因此，如何探索藥物其潛在可能結合的蛋白質？過去研究指出對於一個會與多個不同蛋白質結合的化合物，這群蛋白質可能在序列或是整體結構上並無明顯的演化關係[4]，因而不易藉由以往對蛋白質序列的比對，如蛋白質序列比對軟體 BLAST[5-6]，或是比對蛋白質整體結構，例如蛋白質結構比對軟體 CE[7]及 DALI[8]，來探索其潛在會與化合物結合的蛋白質。

面對這樣的問題，目前探討方向有二，其一是以化合物作為導向的方式[9]，假設存在一群化學結構相似的化合物，因為擁有相似的化學性質，故通常會與具相似結合環境的蛋白質產生交互作用[3, 10-12]。然而這種方法的限制在於，相似結構的化合物在某些情況下並不會與相似的結合環境產生交互作用；實際上，化合物的結構即使有些微改變都會對結合能力有極大的影響[3]。

再者是以蛋白質作為導向。然而過去研究表示蛋白質結構的演化相較於序列更具保留性[13-14]，故利用序列相似度僅能預測辨別出部分具有相似功能的蛋白質[15]。而根

據蛋白質整體結構相似度來預測功能則存在著即使蛋白質間整體結構相似，但在局部結構、具有催化功能的位置上有差異而產生不同功能但整體結構相似的蛋白質群，例如 TIM barrel 以及 Rossmann fold 等[13]。因此，針對無法從序列及整體結構辨認具有相似結合環境蛋白質的問題，本研究提出了空間相關結構片段(Space Related-Pharmamotif, SRP)，藉由描述蛋白質與化合物結合的環境來辨認其他擁有相似結合環境的蛋白質。我們蒐集 530 個蛋白質－藥物分子結晶結構，共計 187 種美國食品藥品監督管理局(FDA)核准的藥物，建構其個別的 SRP，並且在蛋白質結構資料庫(Protein Data Bank, PDB)中進行搜尋，找出具相似結合環境的潛在蛋白質；我們也已將這些資訊放上網路(http://e233.life.nctu.edu.tw/~neverfree/SRP/index.php)，製作易於使用的介面，為全球相關研究者提供服務。我們也以治療流感的藥物瑞樂沙(Zanamivir)做為實例，說明 SRP 對結合環境的敏感性如何反映在搜尋結果上，以及治療慢性骨髓細胞白血病的藥物 Imatinib，說明 SRP 應用在「舊藥新用」議題上的可能性。



圖一、傳統結構比對與 SRP 概念之比較。

二、研究動機

針對以上提出的議題，我們進行此研究。有別於傳統上以蛋白質序列或蛋白質整體結構來探討具有相似結合環境的蛋白質，本研究以蛋白質與化合物產生交互作用的結合環境為核心，建構數段空間上不連續的結構片段以描繪出蛋白質與化合物結合的環境，並且我們將這組結構片段結合實驗室先前的研究 3D-BLAST[16-17]，將三級結構片段轉換成帶有結構資訊的一級結構字元集序列，並在蛋白質結構資料庫中快速地搜尋得到可能擁有相似結合環境的蛋白質(圖一)。

同時我們建置一易於使用者瀏覽的網站介面，呈現 530 個蛋白質－藥物分子建構 SRP 的資訊以及對於蛋白質資料庫的搜尋結果，提供研究者研究探討對於已上市藥物我們如何能提供潛在具有結合藥物的可能標靶蛋白。

更進一步，未來我們將探討藉由如此觀點是否能對於過去以蛋白質整體結構或功能區塊(domain)之間的差異對蛋白質做分群的資料庫，如 SCOP[18]或 CATH[19]，能夠以結合環境的區別再做不同的、更細緻的分類。並且，本研究也期望這樣的概念對於藥物

發展能有所幫助，如舊藥新用的探討或是藥物副作用的了解。

三、論文總覽

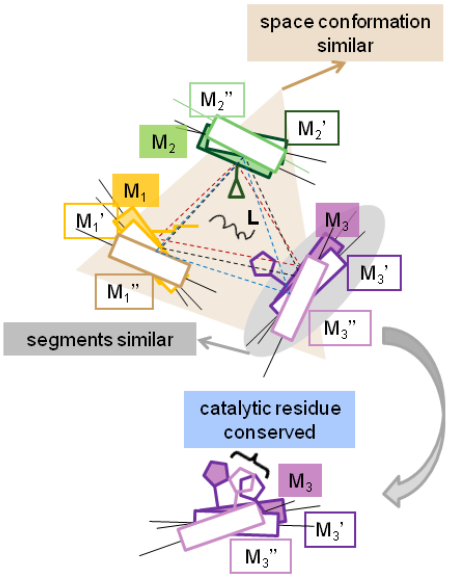
我們提出 Space-Related Pharmamotif 的概念並將其概要描述於第二章節中，主要表現 SRP 的核心概念以及如何使用 SRP 搜尋蛋白質結構資料庫，找尋潛在擁有相似結合環境的蛋白質。同時我們在第三章節中給予數個實例來初步說明 SRP 如何作用於蛋白質分類以及舊藥新用的研究。最後我們介紹對於 530 個蛋白質與已上市藥物的結晶結構所建構出的 SRP 資訊以及對於蛋白質結構資料庫的搜尋結果，我們建置一網站提供使用者觀察及研究已上市藥物潛在可能的標靶蛋白。

貳、研究材料與方法

一、SRP 概念

在本章節中我們呈現 SRP 的概念，對於一蛋白質-配體複合物(protein-ligand complex)建構出一組描述結合環境的結構片段，圖二說明建構 SRP 的核心概念。

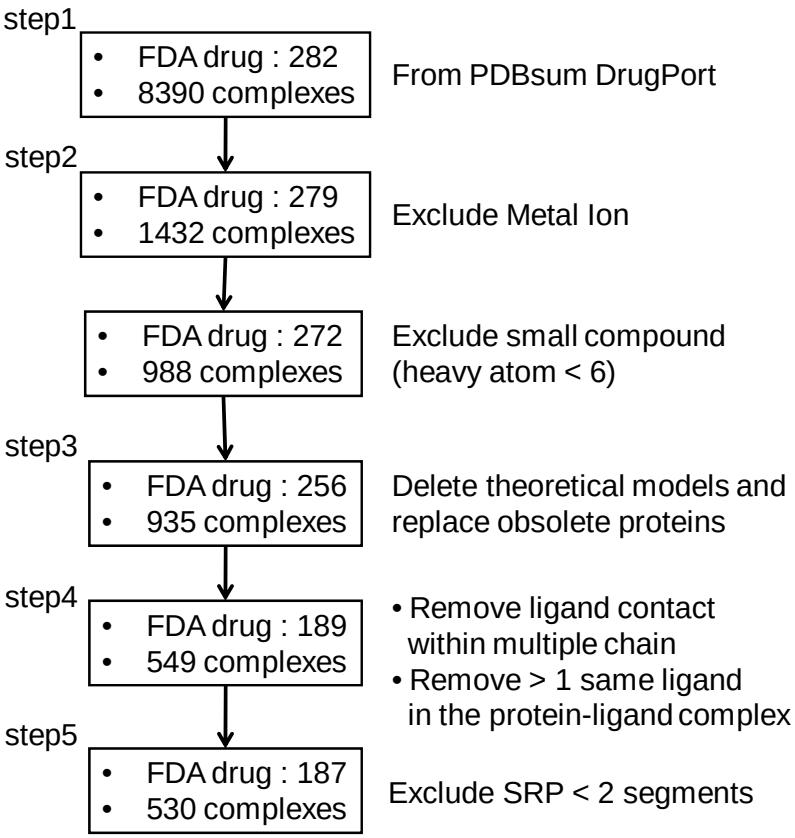
對於一個蛋白質(P)－配體(L)複合物，我們集中在其產生交互作用的結合位上，提出一組不同長度且空間上不連續的蛋白質結構片段來描述其結合環境。我們定義兩個蛋白質 P 以及 P'在滿足下列三項條件時，表示 P 與 P'可能擁有相似的蛋白質結合位環境，因此 P'可能與配體 L 結合：(1) 各蛋白質片段相似(例如 M_1 和 M_1'、M_1''相似、M_2 和 M_2'、M_2''相似)；(2) 空間上蛋白質片段分布相似(例如 M_1-M_2-M_3 和 M_1'-M_2'-M_3'相似)、(3) 擁有相似的催化胺基酸(catalytic residues)。(1)和(2)表示兩蛋白質在結合環境周圍擁有相似的結構摺疊以及空間以讓相同或相似結構的化合物進入，(3)則代表重要、用以催化化合物的胺基酸保留性，提供除了在環境上空間相似以外，同時注重具有直接產生交互作用的胺基酸，表示蛋白質－化合物交互作用的能力及可能性。



圖二、SRP 概念圖。M_1-M_2-M_3 片段組為對於一蛋白質－化合物(L)結晶結構所建構出的 SRP. 當 M_1-M_2-M_3 與另一蛋白質結構上 M_1'-M_2'-M_3'片段組在各片段相似、空間環境上分布相似以及在催化執行功能的胺基酸上具保留性，M_1'-M_2'-M_3'則有可能會與化合物 L 結合.

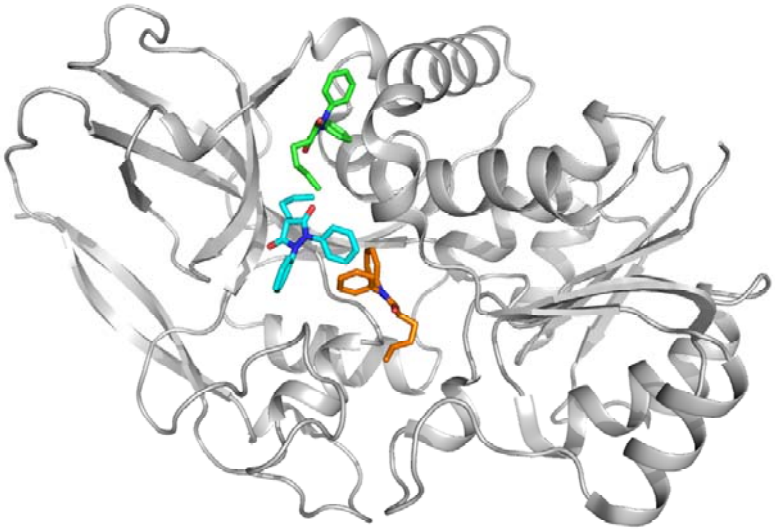
二、FDA 核准藥物資料集

為了解 SRP 對於蛋白質－配體交互作用的應用，我們根據 PDBsum 提供的 DrugPort 資料庫，總共蒐集了 282 個 FDA 核准藥物。對於 282 個化合物，我們希望先觀察傳統上認為可能是藥物的分子，故篩選掉符合下列幾項規則的分子(圖三)，(1) 金屬離子、(2) 重原子 (heavy atom)數目小於 6 的小分子，得到 272 個藥物分子。進一步我們蒐集已知與 272 個藥物分子共結晶的蛋白質結構，並且修正取代已被廢棄的蛋白質結構，共得到 935 個蛋白質結晶結構。由於目前我們只針對單一蛋白質鏈對單一化合物的交互作用環境做探討，因此我們對於配體與超過一個蛋白質鏈有接觸的蛋白質結晶結構予以篩除(圖四 A)。此外，對於一個蛋白質鏈結晶多個相同配體的結構，因為無法判別以何者作為交互作用的化合物已建構 SRP，故目前也先予以篩除(圖四 B)。並且我們認為結晶結構只能建構出一段蛋白質片段的 SRP 不合理而不列入討論的資料集內。經由此篩選流程後共計有 187 個 FDA 藥物分子符合條件(表六)，共 530 個蛋白質－配體結晶結構(簡稱 FDA-530)。

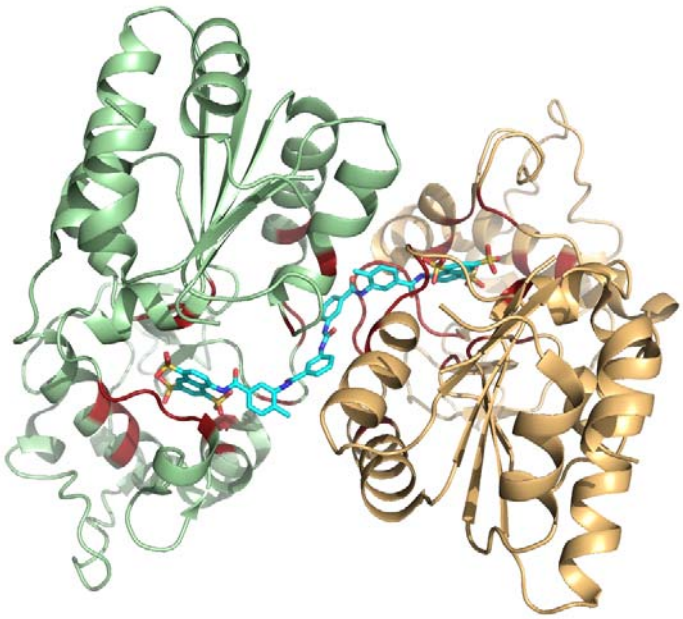


圖三、FDA-530 資料集準備流程。Step1：根據 PDBsum DrugPort 提供的藥物資料，蒐集 282 個 FDA 核准上市藥物；Step 2：過濾金屬離子及小分子；Step 3：刪除理論模型及修正已被廢棄的蛋白質結晶結構；Step 4：移除化合物與多個蛋白質鏈接觸以及超過一個相同化合物結晶於一蛋白質鏈的結晶結構；Step 5：排除建構 SRP 只有一段片段的結晶結構.

A



B



圖四、FDA drug 資料集篩選過濾實例。(A) 蛋白質結晶結構中包含超過一個相同化合物，此例中(PDB code 2w98)化合物為 P1Z；(B) 蛋白質結構中化合物與超過一個蛋白質鏈有接觸，此例中(PDB code 2nyr)化合物為 SVR，綠色結構部分為 A 鏈，褐色結構部分為 B 鏈，紅色區域結構為與化合物接觸的胺基酸。

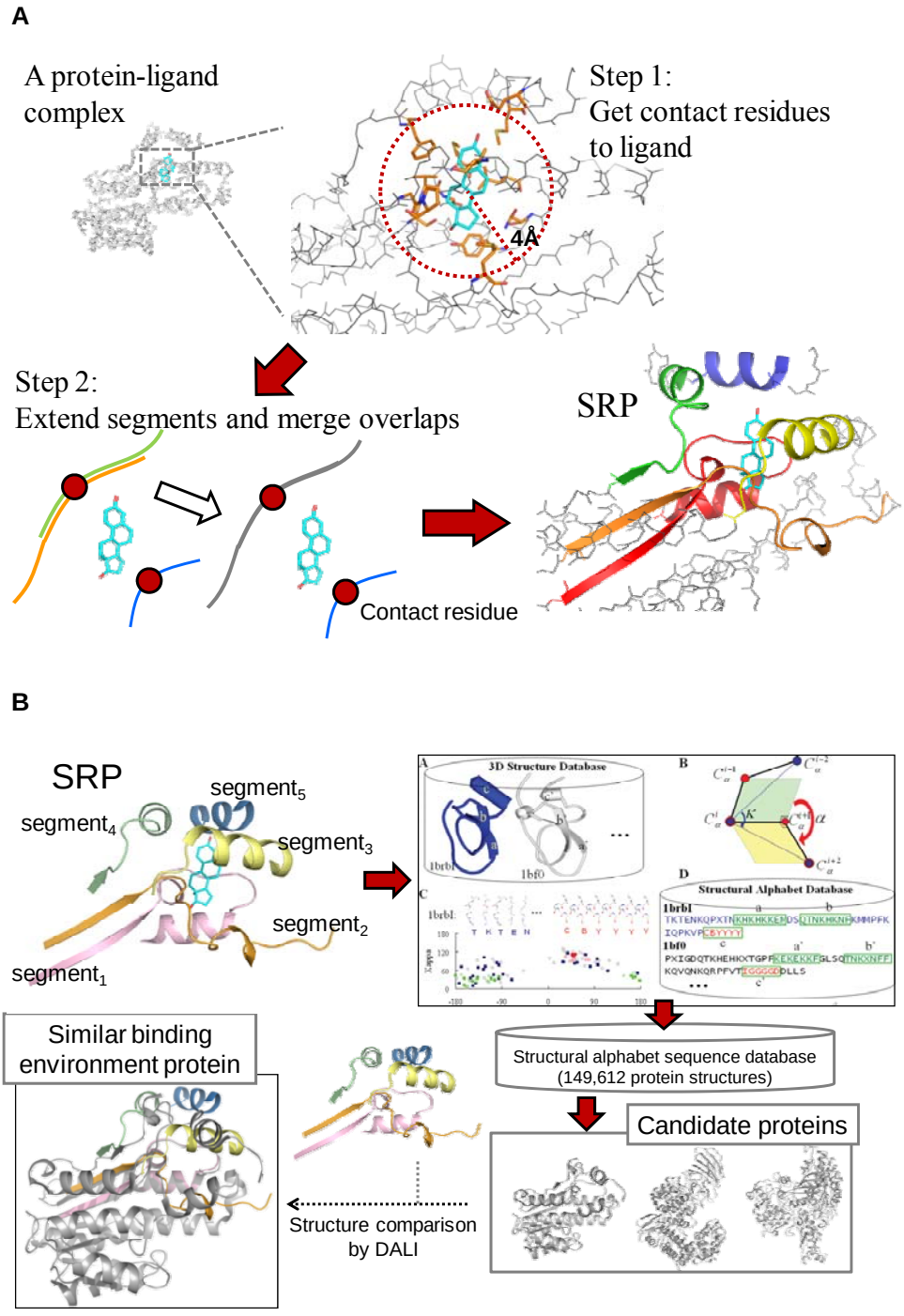
三、結構字元集資料庫

根據實驗室先前已發表的 3D-BLAST[16-17]概念，對於每一個蛋白質結構，我們能根據κ角度以及α角度將蛋白質三維立體結構編譯成一維的結構字元集序列(structural-alphabet sequence)，得到擁有結構資訊的一維結構字元集序列，而我們也將蛋白質結構資料庫(PDB)收錄的所有蛋白質結構編譯成序列並收集為字元集序列資料庫(structural alphabet sequence database, SADB)，本研究中使用的蛋白質結構資料庫收錄的蛋白質總共有 149,612 蛋白質鏈結構(更新至 2010/02/19)。並且，我們藉由實驗室於 3D-BLAST 中發展的結構字元集計分陣列[16-17]，我們能如 BLAST[6]以結構字元集序列對蛋白質結構資料庫做快速的搜尋。

四、建構 SRP 以及搜尋蛋白質資料庫

本章節描述我們如何建構 SRP 並以 SRP 對蛋白質資料庫進行搜尋。圖五 A 為 SRP 建構與搜尋的流程。對於建構出描述結合環境的片段組，我們以一蛋白質－配體結晶結構作為模板，分為兩個步驟：(1) 計算配體任一原子與蛋白質上各胺基酸任一原子間距離，小於 4 Å 即視該胺基酸為接觸胺基酸(contact residue)；(2) 以接觸胺基酸為出發點，向胺基酸前後各延伸 7 個胺基酸使其成為長度 15 個胺基酸的蛋白質結構片段；如果任兩個片段重疊則合併為一段，最後形成多段長度不同、空間上不連續、環繞在結合位周圍的蛋白質結構片段，即為該蛋白質－配體結晶結構的 SRP。

之後，我們根據實驗室以發表的 3D-BLAST[16]概念，將建構完成的 SRP 片段組各片段分別編碼成結構字元集序列(structural alphabet sequence)，並依序對同樣編碼成結構字元集序列的 PDB 結構資料庫做快速的序列搜尋比對。我們對各片段搜尋到的蛋白質鏈列表取其交集，得到可能擁有相似結合環境的蛋白質鏈名單。最後，我們再利用結構比對軟體 DALI 做更精確的結構比對，我們將不連續的 SRP 片段組對搜尋到的完整蛋白質鏈做結構比對，在此以結構均方根誤差(root-mean-square deviation, RMSD)小於 2 即認為兩結構擁有相似的結合環境，進而篩選出真正可能擁有相似結合環境的蛋白質鏈(圖五 B)。



圖五、SRP 建構與搜尋流程圖。(A)SRP 建置流程：步驟一，詢問一蛋白質－配體結晶結構；步驟二，計算配體與胺基酸原子的空間距離得到接觸胺基酸(contact residue)，從接觸胺基酸延伸長度得到蛋白質片段，經合併部分蛋白質片段後得到 SRP。(B)SRP 搜尋流程圖：我們將建構出的 SRP 片段經由 3D-BLAST 轉換成一維字元集序列並對蛋白質結構資料庫做搜尋，得到的蛋白質鏈名單進一步利用結構比對軟體 DALI 做更精確的結構比對，最後得到與詢問模板擁有相似結合環境的蛋白質。

五、評估 SRP 合理性

我們首先將 FDA-530 資料集中結構比對結果有誤之筆數去除，進而我們篩選同時具有 SCOP sccs id、CATH id 以及 UniProt AC 註解的資料，最後再將結晶相同藥物的蛋白質結構予以去除重複，留下 117 筆資料(簡稱 FDA-117, 表七)用以驗證評估 SRP 的合理性。我們從蛋白質結構層面觀察，使用 SCOP sccs id 與 CATH id 作為依據，驗證 SRP 能否辨認並搜尋到擁有相似整體或部分結構的蛋白質。在序列層面上，我們使用 UniProt 序列編號(accession number, AC)來衡量 SRP 對與擁有相同序列的蛋白質結構的辨別能力。同時，對於共結晶相同化合物的蛋白質結構，我們希望觀察 SRP 概念是否能搜尋到含有相同共結晶化合物的蛋白質結構。

六、評估方式

我們使用 precision 以及 recall 來測量 SRP 的搜尋方法。對於我們使用的蛋白質結構資料庫共 149,612 蛋白質鏈結構，以 SCOP sccs id 為例，TP(*True Positives*)為與詢問模板擁有相同 SCOP sccs id 的蛋白質且被 SRP 所辨認出來的個數，FP(*False Positives*)為與詢問模板擁有不同 SCOP sccs id 的蛋白質但被 SRP 辨認相同的個數，FN(*False Negatives*)為與詢問模板擁有相同 SCOP sccs id 的蛋白質但未被 SRP 所辨認出來的個數。由上述類推於 CATH id、UniProt AC 以及在辨認共結晶相同化合物的指標計算。

我們提供另一 Overall 的指標計算，其 TP 為與詢問模板擁有相同 SCOP sccs id、CATH id、UniProt AC 以及化合物任一項指標的蛋白質且被 SRP 所辨認出來的個數，FP 為與詢問模板擁有不同 SCOP sccs id、CATH id、UniProt AC 以及結合化合物指標的蛋白質但被 SRP 所辨認出來的個數，FN 為與詢問模板擁有相同 SCOP sccs id、CATH id、UniProt AC 以及化合物任一項指標的蛋白質但未被 SRP 所辨認出來的個數，以此衡量 SRP 在辨別相似結構、序列以及結合環境上的整體表現。

Precision 以及 recall 的公式如下：

$$precision = \frac{the\ number\ of\ True\ Positives}{the\ number\ of\ True\ Positives + the\ number\ of\ False\ Positives} \quad (1)$$

$$recall = \frac{the\ number\ of\ True\ Positives}{the\ number\ of\ True\ Positives + the\ number\ of\ False\ Negatives} \quad (2)$$

參、結果與討論

在本研究中，我們提出 SRP 以結合環境為導向的概念來探索已知藥物的潛在可能結合蛋白質，研究中使用了 530 個與 FDA 核准藥物共結晶的蛋白質結構來建構 SRP 並對蛋白質結構資料庫進行搜尋，進而了解 SRP 概念的運行以及對已知藥物尋找潛在結合蛋白的研究。對於說明 SRP 如何作用，首先我們給予一評估標準驗證 SRP 的合理性。其次，研究中提出實例說明將結構比對範圍縮小至只在結合環境這樣的概念與過往使用整體蛋白質結構的差異、SRP 片段數量以及片段的二級結構組成影響搜尋結果的問題。最後，對於 FDA-530 資料集的 SRP 建構與搜尋資料，我們建置網站收錄 FDA-530 的搜尋結果。

一、SRP 的合理性

我們以兩項常見描述蛋白質結構的資訊 SCOP sccs id、CATH id，以及描述序列關係上的 UniProt AC，從結構角度以及序列角度來觀察 SRP 在搜尋具有相同或相似蛋白質結構或序列時的表現。並且我們統計結晶相同化合物的蛋白質結晶結構數量，以此來評估 SRP 搜尋結合相同化合物結晶結構的表現。

從表一結果顯示對於描述蛋白質結構的兩項指標 SCOP sccs id、CATH id 在搜尋相同家族蛋白質結構，其覆蓋率(recall)的表現分別為 31%及 22%，在準確率(precision)的表現為 62%及 82%。結果表示對於被 SCOP 或 CATH 分成同一家族的蛋白質結構，以 SRP 的概念能部分找回相同家族的成員，而從準確率可以得知對於整體或部分區域相似的蛋白質，結果擁有一定程度可信的準確度。我們觀察到即使在 SCOP sccs id 或者 CATH id 將蛋白質鏈分在同一群家族的情形下，由於蛋白質鏈在結合位上仍有明顯差異，SRP

表一、SRP 評估指標表現

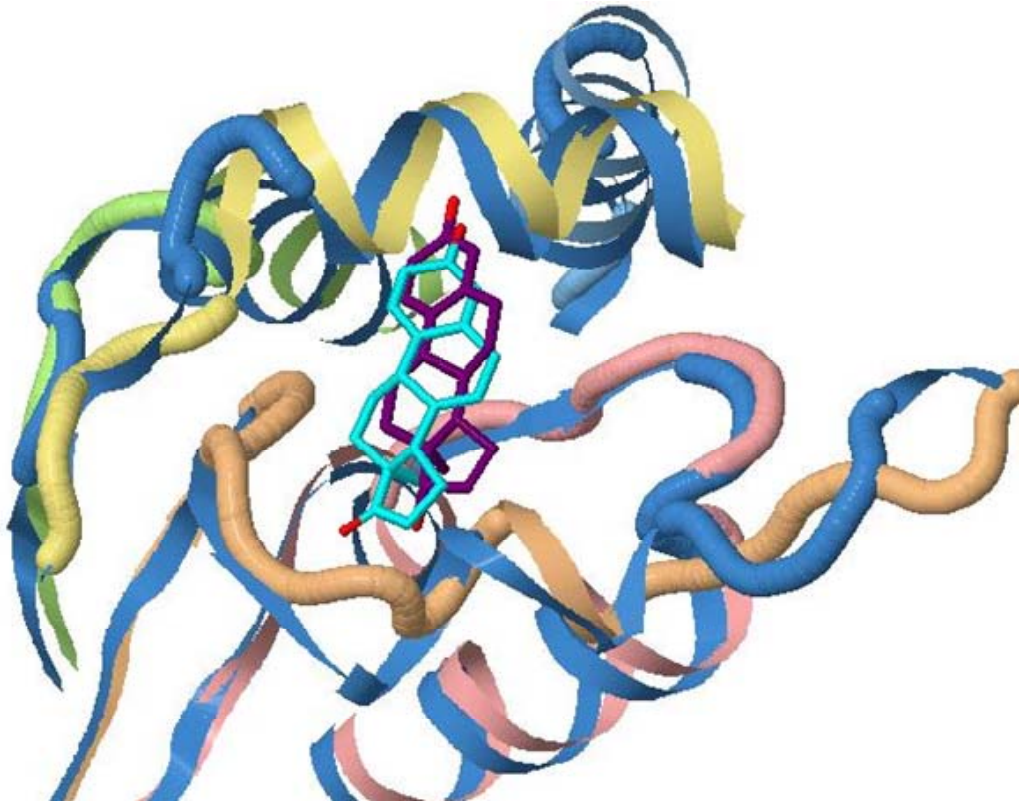
	Overall[a]	SCOP	CATH	UAC[b]	Ligand
Recall	0.22	0.31	0.22	0.80	0.54
Precision	0.88	0.62	0.82	0.36	0.03

[a] 綜合統計 SCOP, CATH, UAC, Ligand 註解
[b] 代表 UniProt AC

以結合環境的角度來對蛋白質做分類，能更敏銳地對整體結構或部分區域結構相似但結合環境不同的蛋白質再做區分？下一節會提供實例做進一步解釋。

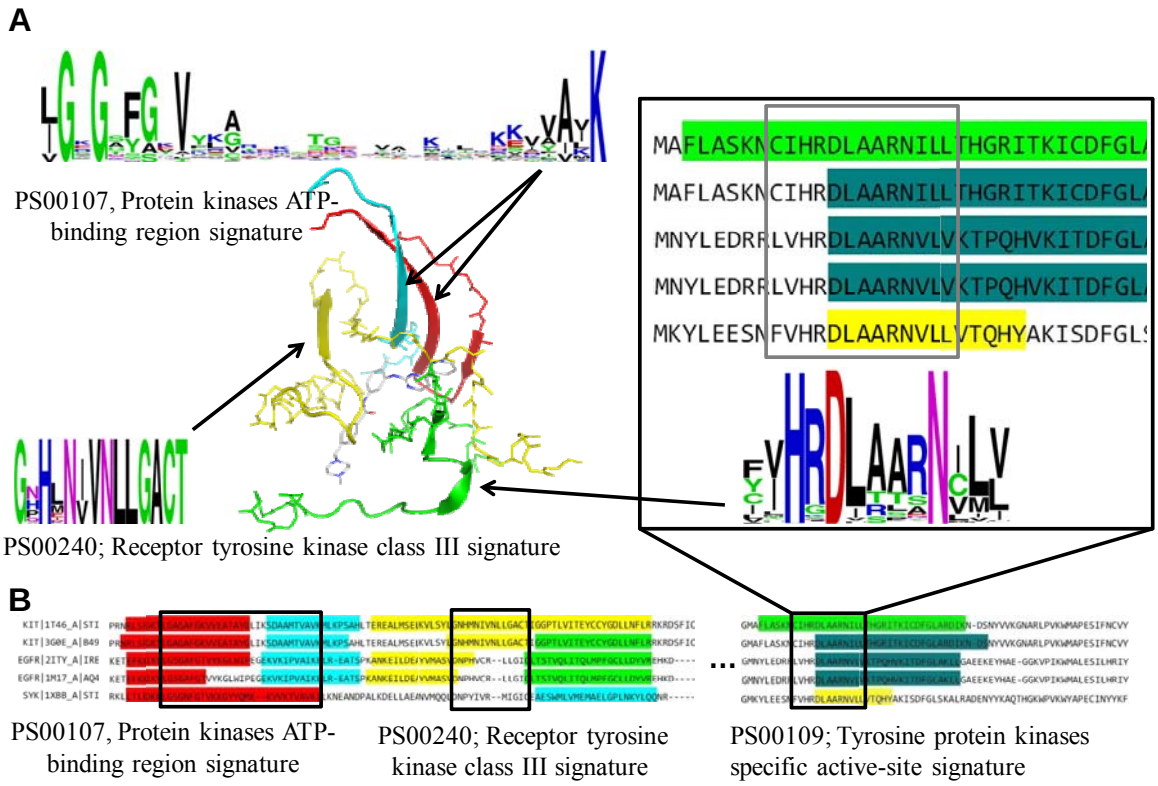
從 UniProt AC 的角度觀察，覆蓋率和準確率為 80%及 36%。結果顯示 SRP 以多段不連續的蛋白質片段組作為搜尋的概念，對於搜尋相同序列的蛋白質結構仍有很好的辨認性。然而對於 36%的準確率表現，觀察其原因為在計算準確率時我們只將與搜尋對象擁有完全相同UniProt AC的蛋白質鏈做計算，對於搜尋到在序列上仍高度相似但UniProt AC 的不同的蛋白質而不被列入，使得準確率下降。例如蛋白質 cCMP-specific 3',5'-cyclic phosphodiesterase(Gene name PDE5A)與 Viagra(HET group id VIA)的結晶結構(PDB code 1tbf, chain A)，其蛋白質 UniProt AC 是 O76074(*Human*)，以此結構做搜尋我們能夠找到蛋白質 cAMP and cAMP-inhibited cGMP 3',5'-cyclic phosphodiesterase 10A(Gene name PDE10A) 結晶結構(PDB code 2ouu, chain A)，以 DALI 進行結構比對 RMSD 0.7、Z-score 9.2，其 UniProt AC 是 Q9Y233(*Human*)。從 Gene name 以及結構比對結果均表示兩蛋白質來自相同祖先且結構高度相似，而因我們在計算時只針對完全相同的 UniProt AC，因此該蛋白質(PDB code 2ouu)不被我們列入計算。

同樣的問題在搜尋結合相同化合物的表現上也觀察到，在搜尋結合相同化合物的覆蓋率及準確率為 54%和 3%。結果可見對結合相同化合物的結晶結構，透過 SRP 能找到一半以上的結構。然而，在準確率只有 3%的表現下，原因與我們以 UniProt AC 做觀察指標時觀察到的問題相同，我們只統計擁有與搜尋對象完全相同化合物名稱(HET group)的蛋白質鏈。如蛋白質 Estradiol 17-beta-dehydrogenase 1(Gene name HSD17B1)與 Estradiol(HET group id: EST)的結晶結構(PDB code 1a27,chain A)，我們對該結晶結構做搜尋後，搜尋結果中有一結構同為蛋白質 Estradiol 17-beta-dehydrogenase 1(Gene name HSD17B1)，其結晶的化合物為 Equilin(HET group id: EQI)的結晶結構(PDB code 1equ, chain A)，兩者擁有非常相似的化合物構型(圖六)，而由於我們只計算含完全相同化合物 HET group 的結構，因此該結構也沒有將之列入計算。同時除了結晶相似結構化合物的結構不被我們統計以外，沒有結晶化合物但擁有相似結合環境的結構也不被我們計算。最後，統整 SCOP sccs id、CATH id、UniProt AC 以及 Ligand 的綜合表現上，在準確率的部份因可互相彌補修正統計結果而有 88%的表現。在覆蓋率的部份雖然只有 28%，但也可觀察出在 UniProt AC 以及 Ligand 部分均在 50%以上，唯在結構表現 SCOP sccs id、CATH id 的部份為 31%和 26%表現較低。



圖六、SRP 結構比對結果。圖中為以蛋白質結構 PDB code 1a27(chain A)為模板，建構的 SRP 蛋白質片段組(粉紅、粉橘、黃、綠、粉藍)搜尋到 PDB code 1equ(chain A) 蛋白質鏈(深藍色蛋白質片段組)的結構比對結果。可見模板所結晶化合物 EST(紫)與搜尋結果所結晶化合物 EQI(青藍)結構相當相似。

同時，我們從 ExPASy Prosite[20]所提供已知的序列樣板與 SRP 做探討，來說明 SRP 建構出片段組的合理性。以在研究上較為所知的蛋白激酶做實例，藉由 FDA-530 中的酪氨酸蛋白質磷酸激酶 c-KIT 對藥物 Imatinib 建構 SRP 並比較對於該蛋白已知的序列樣板(圖七 A)，並且我們由 FDA-530 資料集中舉出由五個不同蛋白激酶-藥物的結晶結構，使用 CLUSTAL W 對其做多重序列排序並將所建構出的 SRP 資訊與 ExPASy Prosite 已知序列樣板做比較(圖 7B)。從不同蛋白激酶對不同藥物下建構的 SRP 及對已知序列樣板的比較，在蛋白激酶上的實例，SRP 建構出的片段組都能從蛋白質整體結構中擷取出包含 ExPASy Prosite 所定義的已知序列結構樣板，反映出 SRP 在建構片段組上的合理性。



圖七、SRP 與 ExPASy Prosite 片段序列之比較。(A)SRP 片段組[以 PDB code 1t46(chain A)為模板]，其中建構的 SRP 結構片段組包含了三項 ExPASy Prosite 定義區塊(cartoon form)，分別為 Protein kinases ATP-binding region signature(紅色以及天空藍)、 Receptor tyrosine kinase class III signature(黃色)、Tyrosine protein kinases specific active-site signature(綠色)。(B)以 FDA-530 資料集中五個蛋白激酶做多重序列排序，SRP 片段組(上色區塊)對 ExPASy Prosite 定義序列樣板之比較。

二、SRP 搜尋結果分析

本節我們以數個實例來說明 SRP 如何作用以及實作時所發現的問題。其一在某些實例上我們發現即使蛋白質鏈被分類在相同的 SCOP 家族底下，SRP 能因結合位的差異而區分出不同的"binding-site family"。同時我們初步討論 SRP 對於「舊藥新用」可行性。此外，我們列舉兩個實例探討在 SRP 建構出含有較多片段或者片段中含二級結構比例過高時，對於搜尋結果的影響。

（一）、蛋白質分類問題

病毒神經胺酸水解酶 (viral neuraminidase)與瑞樂沙(Zanamivir)

神經胺酸水解酶(neuraminidase)為 glycoside hydrolase enzyme (EC 3.2.1.18)，用以切開神經胺酸上的 glycosidic linkage 並此蛋白分佈於許多不同物種。一般最為認識的即為存在於流感病毒表面的病毒神經胺酸水解酶，促使病毒顆粒釋出宿主細胞。神經胺酸水解酶為一種已知治療流感病毒的目標蛋白質，當流感病毒在感染複製時，同樣存在於病毒表面上的流感血凝素(hemagglutinin)會辨認宿主細胞表面醣蛋白上的 sialic acid 而與宿主細胞結合，而為了讓病毒脫離宿主細胞，神經胺酸水解酶則會將宿主細胞表面醣蛋白上的 sialic acid group 予以切除，使病毒顆粒能離開宿主細胞繼續感染其他細胞。

瑞樂沙(Zanamivir, 上市名稱 Relenza)為近年被十九個國家核准用以治療及預防 A 型及 B 型流感病毒的口腔吸入式粉狀藥物，是第一個上市針對流感病毒在感染過程中必須的神經胺酸水解酶的競爭型抑制劑，其結構與神經胺酸水解酶的催化化合物 sialic acid 高度相似，作用機制為藉由與神經胺酸水解酶結合，抑制病毒脫離宿主細胞及感染其他細胞。

我們以流感病毒 A 型中的神經胺酸水解酶與藥物瑞樂沙(HET group id: ZMR)的結晶結構(PDB code 2htq,chain A)為模板建構 SRP，並對蛋白質結構資料庫做搜尋。表二中顯示 14 筆同樣與 ZMR 結合的蛋白質結晶結構，其中 11 筆為流感病毒的神經胺酸水解酶，CATH 分類為 2.120.10.10；2 筆為人類第三型副流感病毒蛋白，其蛋白質為同時擁有流感血凝素與神經胺酸水解酶功能的 Hemagglutinin-neuraminidase glycoprotein；1 筆為人類 sialidase-2(NEU2)，其結構與蛋白質功能與神經胺酸水解酶相似，用以催化 sialic acid 的水解。

14 個蛋白結構中我們搜尋到 11 筆，9 筆為流感病毒 A 型神經胺酸水解酶，2 筆為 B

型流感病毒神經胺酸水解酶。結果顯示對於結晶相同化合物 ZMR、擁有相同 CATH、相似物種的蛋白質鏈能被 SRP 找到，且結構比對軟體的結果也顯示都擁有相似結合位環境，且在序列保留性也相似(圖八)，說明我們的方法能辨認出結合環境相似的蛋白質結構；相對地，另外三筆蛋白為人類 Hemagglutinin-neuraminidase glycoprotein 以及 sialidase-2 沒有被 SRP 所辨認出來，而結構比對軟體結果也顯示這兩種蛋白在結合環境上與我們的模板病毒神經胺酸水解酶有較大的差別(RMSD 3.9 and 4.3)，且在序列保留性上也有較大的差異，且在病毒神經胺酸水解酶與瑞樂沙交互作用中可能產生氫鍵的胺基酸中僅有一個胺基酸 Arg^{118} 具有保留性(圖八)。

同時，由於近年來多起因使用克流感及瑞樂沙而導致神經精神系統失調和死亡的病例報導，使這些藥物被認為除了會與病毒神經胺酸水解酶結合以外，也可能會抑制參與調控 sialic acid 機制中的酵素，例如 sialidase、sialyltransferase 及 CMP-synthase。近期文獻針對此問題以克流感及瑞樂沙對目前已知存在於人類的四個 sialidase(NEU1－4)做活性測試，IC_{50} 實驗指出瑞樂沙需在 μM 等級的濃度才會對人類的 sialidase 有影響(NEU1: 2,713 μM, NEU2: 16.4 μM, NEU3: not determined, NEU4: 487 μM)，而對於病毒神經胺酸水解酶則只需要 nM 等級即可抑制(H1N1: 1.56 nM, H3N2: 2.66 nM, H5N3: 3.97 nM)[21]。

因此，從 SRP 搜尋的結果以及文獻的實驗結果可以發現，即使蛋白質因整體或部分區域結構相似而被已知的蛋白質結構分類方式如 SCOP 或 CATH 分成同一群，這些分類指標卻無法對於真正執行催化功能區域上的變異加以辨別區隔。某種程度說明著 SRP 對於結合環境有相當高的敏銳判別度，因此可能以「binding-site family」專注於蛋白質與化合物結合環境的角度對蛋白質做分類，如圖八中因流感病毒的神經胺酸水解酶與人類的 sialidase-2 在結合環境上的不同被分成兩個家族，從而提供更著重於交互作用的分類方法。

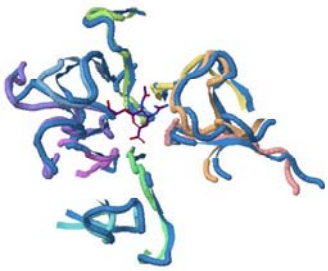
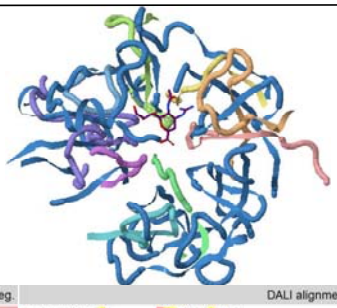
表二、SRP 搜尋結果 [以 PDB code 2htq(chain A)為模板]

PDBID[a]	#[b]	SCOP	CATH	UniProt AC	Gene name	Species	Protein description	DALI RMSD	Z score	SI(%)
2htq_A	3D	-	2.120.10.10	Q07599	NA	Influenza A virus (strain A/Duck/Ukraine/1/1963 H3N8)	Neuraminidase	0.0	19.0	100
3ckz_A	3D	-	2.120.10.10	Q6DPL2	NA	Influenza A virus (strain A/Viet Nam/1203/2004 H5N1)	Neuraminidase	0.5	16.7	67
3b7e_B	3D	-	2.120.10.10	Q9IGQ6	NA	Influenza A virus (strain A/Brevig Mission/1/1918 H1N1)	Neuraminidase	0.5	16.8	66
3b7e_A	3D	-	2.120.10.10	Q9IGQ6	NA	Influenza A virus (strain A/Brevig Mission/1/1918 H1N1)	Neuraminidase	0.5	16.9	66
2cml_A	3D	-	2.120.10.10	Q6XV27	NA	Influenza A virus (strain A/Duck/England/1/1956 H11N6)	Neuraminidase	1.1	15.7	62
2cml_C	3D	-	2.120.10.10	Q6XV27	NA	Influenza A virus (strain A/Duck/England/1/1956 H11N6)	Neuraminidase	1.1	15.7	62
2cml_D	3D	-	2.120.10.10	Q6XV27	NA	Influenza A virus (strain A/Duck/England/1/1956 H11N6)	Neuraminidase	1.1	15.7	62
2cml_B	3D	-	2.120.10.10	Q6XV27	NA	Influenza A virus (strain A/Duck/England/1/1956 H11N6)	Neuraminidase	1.1	15.6	62
1nnc_A	3D	b.68.1.1	2.120.10.10	P03472	NA	Influenza A virus (strain A/Tern/Australia/G70C/1975 H11N9)	Neuraminidase	1.2	15.6	56
1a4g_A	3D	b.68.1.1	2.120.10.10	P27907	NA	Influenza B virus (strain B/Beijing/1/1987)	Neuraminidase	1.4	14.5	42
1a4g_B	3D	b.68.1.1	2.120.10.10	P27907	NA	Influenza B virus (strain B/Beijing/1/1987)	Neuraminidase	1.4	14.5	42
1v3e_A	-	b.68.1.1	-	Q6WJ03	-	Human parainfluenza virus 3	Hemagglutinin-neuraminidase glycoprotein	3.9	3.1	14
1v3e_B	-	b.68.1.1	-	Q6WJ03	-	Human parainfluenza virus 3	Hemagglutinin-neuraminidase glycoprotein	3.9	3.1	14
2f0z_A	-	b.68.1.1	-	Q9Y3R4	NEU2	Homo sapiens	Sialidase-2	4.3	2.4	9

[a] 前四碼代表 PDB code, 最後一碼代表 chain

[b] 是否被 SRP 辨別為結合環境相似蛋白, 3D 為可能具相似結合環境的蛋白

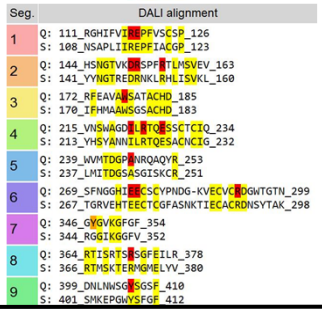
Query : 2htq (Influenza A virus)

	Family 1	Family 2
Protein Name	Neuraminidase	Sialidase-2
Species	Influenza B virus (strain B/Beijing/1987)	Homo sapiens
IC_{50}	1.56 – 3.97 nM	16400 nM
Structure alignment of DALI		

represent contact residue

represent identical residue

represent contact & identical residue



圖八、DALI 結構比對結果[以 PDB code 2htq(chain A)為模板]. 分別為模板對流感病毒 Neuraminidase (PDB code 1a4g) 結構比對以及對人類 Sialidase-2 (PDB code 2f0z) 結構比對結果。

（二）、舊藥新用應用探討

酪氨酸蛋白質磷酸激酶 c-Kit(Tyrosine-protein kinase c-Kit)與 Imatinib

c-Kit 是幹細胞因子(stem cell factor)受器，為酪氨酸蛋白質磷酸激酶的一種，負責觸發細胞生長與增生的訊息傳遞。過去研究了解對於 c-Kit 的突變會導致某些酪氨酸蛋白質磷酸激酶活化而引起胃癌(gastrointestinal stromal tumor) [22]。

Imatinib(上市名稱 Gleevec 或 Glivec)是一個針對酪氨酸蛋白質磷酸激酶家族的藥物，該家族成員包括 KIT、ABL、ABL-2、PDGF-R(血小板衍生生長因子接受體)。Imatinib 會與這些酪氨酸蛋白質磷酸激酶結合，抑制酪氨酸蛋白質磷酸激酶將 ATP 的磷酸鹽轉移到下游，進而抑制訊息傳遞。藉由這樣的方式，Imatinib 得以阻斷與突變型和野生型 KIT 蛋白有關的激酶活性。2001 年美國食品及藥物管理局核准使用 Imatinib 以治療慢性骨髓性白血病(Chronic myelogenous leukemia)，一年後又被核准用於治療胃腸道基質腫瘤(Gastrointestinal stromal tumor)。由於 Imanitib 可以對抗 BCR-ABL、c-Kit、PDGF-R 等標靶蛋白，因此它是一種多重標靶藥物(multi-target drug)。

在此，我們以人類細胞中的c-Kit與藥物Imatinib結晶結構(PDB code 1t46,chain A)為模板建構SRP並搜尋蛋白質結構資料庫。表三為前二十名具相似環境的蛋白質鏈，除模板的KIT蛋白質，SRP辨別到FLT1、CSF1R、KDR以及FGFR1等其他屬於酪氨酸蛋白質磷酸激酶家族的蛋白質，將其做結構疊合可看見在結合環境上皆高度相似(圖九A) ，而從參與催化反應的胺基酸保留性比較，在Glu^{640}、Thr^{670}、Cys^{673}以及Asp^{810}可能產生氫鍵的位置上在CSF1R均有保留(圖九B)。其中也有實驗證實[23] Imatinib對CSF1R的抑制能力為19 nM。因此，我們可以推測Imatinib可能與CSF1R產生交互作用，進而可能用以治療與CSF1R相關的疾病如急性骨髓性白血病(Acute myeloid leukemia)。

表三、SRP 搜尋結果[以 PDB code 1t46(chain A)為模板,Top20]

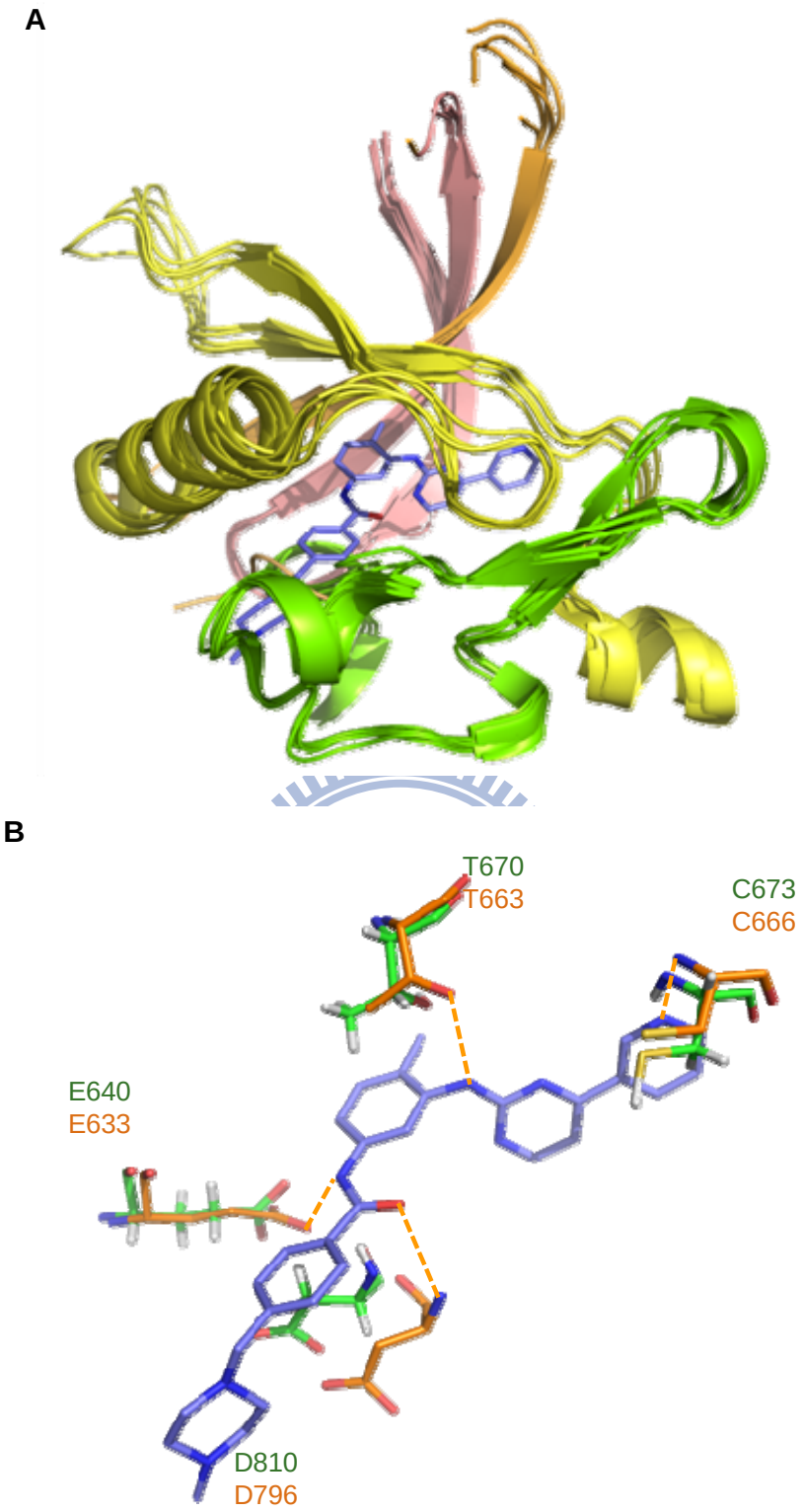
PDBID[a]	SCOP	CATH	UniProt AC	Ligand	Gene Name	Species	Description	RMSD	Z	SI(%)
1t46_A	d.144.1.7	1.10.510.10	P10721	PO4, STI	KIT	Homo sapiens	Mast/stem cell growth factor receptor	0	19	100
3g0f_A	-	1.10.510.10	P10721	B49, SO4	KIT	Homo sapiens	Mast/stem cell growth factor receptor	0.8	16	99
1t45_A	d.144.1.7	1.10.510.10	P10721	-	KIT	Homo sapiens	Mast/stem cell growth factor receptor	1	14.7	100
3hng_A	-	-	P17948	8ST, CL	FLT1	Homo sapiens	Vascular endothelial growth factor receptor 1	1	16	66
3dpk_A	-	1.10.510.10	P07333	8C5, SO4	CSF1R	Homo sapiens	Macrophage colony-stimulating factor 1 receptor	1	15.6	81
1agw_B	d.144.1.7	1.10.510.10	P11362	SU2	FGFR1	Homo sapiens	Basic fibroblast growth factor receptor 1	1	13.9	54
3b8q_B	-	1.10.510.10	P35968	900	KDR	Homo sapiens	Vascular endothelial growth factor receptor 2	1	14.5	64
2qu6_B	-	1.10.510.10	P35968	857	KDR	Homo sapiens	Vascular endothelial growth factor receptor 2	1	13.8	63
3b8r_A	-	1.10.510.10	P35968	887, EDO	KDR	Homo sapiens	Vascular endothelial growth factor receptor 2	1	12.7	63
3g0f_B	-	1.10.510.10	P10721	B49	KIT	Homo sapiens	Mast/stem cell growth factor receptor	1.1	15.7	99
3g0e_A	-	1.10.510.10	P10721	B49	KIT	Homo sapiens	Mast/stem cell growth factor receptor	1.1	15.1	100
2ogv_A	-	1.10.510.10	P07333	-	CSF1R	Homo sapiens	Macrophage colony-stimulating factor 1 receptor	1.1	17.7	80
1fgk_B	d.144.1.7	1.10.510.10	P11362	-	FGFR1	Homo sapiens	Basic fibroblast growth factor receptor 1	1.1	13.8	54

[a] 前四碼代表 PDB code, 最後一碼代表 chain

表三、SRP 搜尋結果 [以 PDB code 1t46(chain A)為模板, Top20] (Continued)

PDBID[a]	SCOP	CATH	UniProt AC	Ligand	Gene Name	Species	Description	RMSD	Z	SI(%)
3efl_A	-	-	P35968	706	KDR	Homo sapiens	Vascular endothelial growth factor receptor 2	1.1	13.3	63
2fgi_B	d.144.1.7	1.10.510.10	P11362	PD1	FGFR1	Homo sapiens	Basic fibroblast growth factor receptor 1	1.1	13.6	54
3dtw_A	-	1.10.510.10	P35968	A96	KDR	Homo sapiens	Vascular endothelial growth factor receptor 2	1.1	13.4	64
2fgi_A	d.144.1.7	1.10.510.10	P11362	PD1	FGFR1	Homo sapiens	Basic fibroblast growth factor receptor 1	1.1	12.5	53
3cp9_A	-	1.10.510.10	P35968	C19	KDR	Homo sapiens	Vascular endothelial growth factor receptor 2	1.1	12.9	64
3cpb_A	-	1.10.510.10	P35968	C92	KDR	Homo sapiens	Vascular endothelial growth factor receptor 2	1.1	12.9	63
1vr2_A	d.144.1.7	1.10.510.10	P35968	-	KDR	Homo sapiens	Vascular endothelial growth factor receptor 2	1.2	13.5	63

[a] 前四碼代表 PDB code, 最後一碼代表 chain



圖九、結構比對結果[以 PDB code 1t46(chain A)為模板]。(A) 模板對蛋白質 FLT1、CSF1R、FGFR1 及 KDR 之結構比對 (B) 模板對 Imatinib 產生氫鍵的胺基酸(綠色)以及 CSF1R (PDB code 3dpk) 相對應胺基酸(橘色)。

（三）、SRP 搜尋問題

SRP 片段數量對於搜尋結果的影響

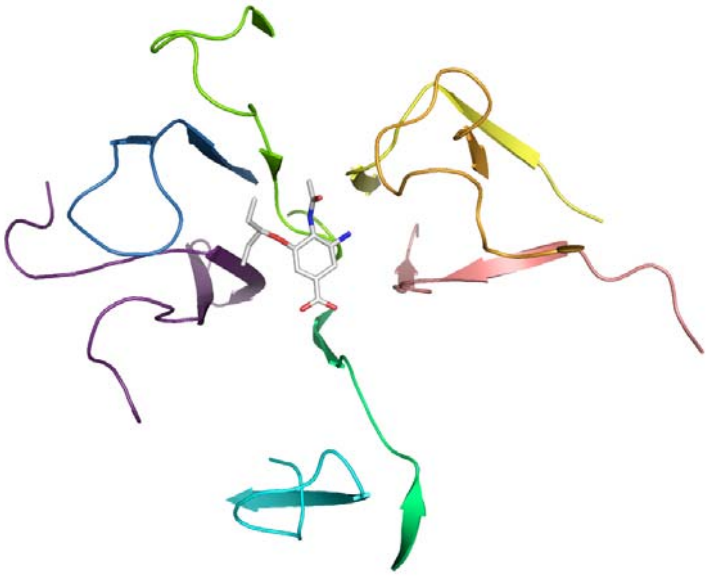
我們觀察到在某些實例中，建構出的 SRP 擁有相當多的片段，在 FDA-530 中 7 段以上的實例有 32 筆，平均為 4 段。同時我們也觀察到對於潛在蛋白的搜尋，我們的規則是每個片段各自搜尋到的蛋白質鏈清單取其交集得到可能的潛在蛋白名單。這樣的條件在擁有多個片段的實例中對搜尋結果造成影響。

我們以病毒神經胺酸水解酶(viral neuraminidase)與克流感(oseltamivir or Tamiflu)的結晶結構(PDB code 2ht7, chain A)作為模板說明，建立出來的片段共九段(圖十)。相同結有化合物 oseltamivir 的結晶結構有 22 筆，均為病毒神經胺酸水解酶且 CATH(2.120.10.10)以及 EC number (3.2.1.18)均相同，且結構比對顯示 22 筆的結合環境也都相似，但是 SRP 卻只辨認出 6 筆而已，有 16 筆無法被 SRP 所找回(表四)。進一步觀察我們發現，對於 16 筆沒有找回的資料，原因為某些片段在結構上的變動使得在搜尋時無法搜尋到 16 筆的蛋白質鏈，造成搜尋結果交集後被認為沒有搜尋到該 16 筆。對於這樣的問題，我們認為容忍結構上的些微變化，對結果只做部分交集可能是合理的；目前暫定的解決方式是對於 SRP 小於 4 段的結構要求全部交集，對於 SRP 擁有 4 段以上的結構，有 75%以上交集的蛋白質鏈即視為潛在可能的結合蛋白。經過這樣的規則後，22 筆資料都得以被 SRP 所辨認到。對於 FDA-117 中 SRP 擁有 4 個片段以上的搜尋結果，其覆蓋率從 13%升至 19%，準確率則從 94%變為 91%。

SRP 片段二級結構組成比例對於搜尋結果的影響

在觀察結果時我們也發現，部分建構出的 SRP 片段擁有高度二級結構的組成(α-helix 及 β-strand)，而由於 3D-BLAST 在將對結構片段編譯成字元集序列時只使用結構資訊，並無引用序列的資訊，故當 SRP 片段擁有高比例的單一二級結構時，相較於存在大量二級結構的蛋白質結構資料庫，這樣的 SRP 片段在進行資料庫搜尋時因無鑑別度而對搜尋結果有影響。以酵母菌蛋白質 Geranylgeranyl pyrophosphate synthase 與化合物 Zoledronic acid (HET group id: ZOL)結晶結構(PDB code 2e91, chain A)為例說明，其建構出的 SRP 片段如圖十一 A，可看出片段二(橘色)及片段四(綠色)均含大量的 α-helix。結晶相同化合物的蛋白質鏈共有 13 筆，從結構比對可發現 13 筆的結構十分相似，SRP 找

到其中 3 筆。觀察後發現部分沒被辨識出的原因在片段二及片段四並未辨認出這些蛋白質鏈。對於這樣的問題，我們推測當片段中擁有高比例的二級結構組成時，在搜尋蛋白質資料庫時不具鑒別度而無法提供有效的搜尋結果。因此，我們延長這種二級結構組成比例高的片段使得片段更具有特徵性，規則是對於經轉換後字元集序列，我們藉由辨認轉換後的序列字元來判斷是否具有二級結構，序列字元代表 α-helix 的為 A、Y、B、C 及 D，代表 β-strand 的為 E、F 及 H，如單一的二級結構組成超過片段序列的 80%則延長片段至組成低於 80%或片段長度達 30 即停止延長。藉由此規則，我們將模板重新建構 SRP 片段如圖十一 B，可以看見片段二以及片段四均延長 12 個胺基酸長度(圖十一 C 及 D)，搜尋結果能夠多找回其他相似結構蛋白質(表五)。



圖十、SRP 建構片段組[以 PDB code 2ht7(chain A)為模板]。由病毒神經胺酸水解酶與克流感的結晶結構為模板建構共九段蛋白質片段。

表四、SRP 搜尋結果 [以 PDB code 2ht7(chain A)為模板]

PDBID[a]	#[b]	SCOP	CATH	UniProt AC	EC number	Gene Name	Species	Description	RMSD	Z	SI(%)
2ht7_A	3D	-	2.120.10.10	Q07599	3.2.1.18	NA	Influenza A virus (strain A/Duck/Ukraine /1/1963H3N8)	Neuraminidase	0	17.7	100
2hu0_B	3D*	-	2.120.10.10	Q6DPL2	3.2.1.18	NA	Influenza A virus (A/VietNam/1203/2004(H5N1))	Neuraminidase	0.5	15.9	68
2ht8_A	3D	-	2.120.10.10	Q07599	3.2.1.18	NA	Influenza A virus (strain A/Duck/Ukraine /1/1963H3N8)	Neuraminidase	0.8	15.9	100
2hu4_E	3D	-	2.120.10.10	Q6DPL2	3.2.1.18	NA	Influenza A virus (A/VietNam/1203/2004(H5N1))	Neuraminidase	0.9	15.4	68
2hu4_G	3D	-	2.120.10.10	Q6DPL2	3.2.1.18	NA	Influenza A virus (A/VietNam/1203/2004(H5N1))	Neuraminidase	0.9	15.3	68
3cl2_B	3D	-	2.120.10.10	Q6DPL2	3.2.1.18	NA	Influenza A virus (A/VietNam/1203/2004(H5N1))	Neuraminidase	0.9	15.5	68
3cl2_G	3D	-	2.120.10.10	Q6DPL2	3.2.1.18	NA	Influenza A virus (A/VietNam/1203/2004(H5N1))	Neuraminidase	0.9	15.5	68
2hu4_A	3D*	-	2.120.10.10	Q6DPL2	3.2.1.18	NA	Influenza A virus (A/VietNam/1203/2004(H5N1))	Neuraminidase	0.9	15.3	68
2hu4_B	3D*	-	2.120.10.10	Q6DPL2	3.2.1.18	NA	Influenza A virus (A/VietNam/1203/2004(H5N1))	Neuraminidase	0.9	15.3	68
2hu4_C	3D*	-	2.120.10.10	Q6DPL2	3.2.1.18	NA	Influenza A virus (A/VietNam/1203/2004(H5N1))	Neuraminidase	0.9	15.3	68
2hu4_D	3D*	-	2.120.10.10	Q6DPL2	3.2.1.18	NA	Influenza A virus (A/VietNam/1203/2004(H5N1))	Neuraminidase	0.9	15.4	68
2hu4_F	3D*	-	2.120.10.10	Q6DPL2	3.2.1.18	NA	Influenza A virus (A/VietNam/1203/2004(H5N1))	Neuraminidase	0.9	15.3	68
2hu4_H	3D*	-	2.120.10.10	Q6DPL2	3.2.1.18	NA	Influenza A virus (A/VietNam/1203/2004(H5N1))	Neuraminidase	0.9	15.4	68
3cl0_A	3D*	-	2.120.10.10	Q6DPL2	3.2.1.18	NA	Influenza A virus (A/VietNam/1203/2004(H5N1))	Neuraminidase	0.9	15.3	68
3cl2_A	3D*	-	2.120.10.10	Q6DPL2	3.2.1.18	NA	Influenza A virus (A/VietNam/1203/2004(H5N1))	Neuraminidase	0.9	15.5	68

[a] 前四碼代表 PDB code, 最後一碼代表 chain

[b] 是否被 SRP 搜尋為結合環境相似蛋白, 3D 為加入規則前被辨別為具相似結合環境的蛋白, 3D* 為加入規則後被辨別為具相似結合環境的蛋白

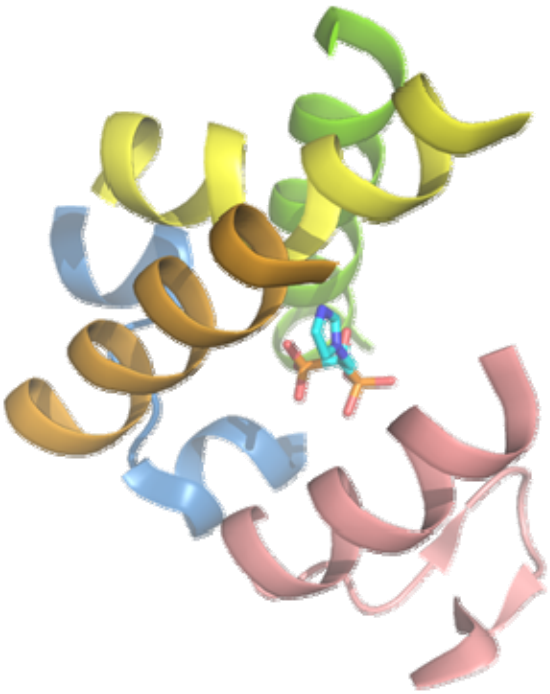
表四、SRP 搜尋結果 [以 PDB code 2ht7(chain A)為模板](Continued)

PDBID[a]	#[b]	SCOP	CATH	UniProt AC	EC number	Gene Name	Species	Description	RMSD	Z	SI(%)
3cl2_C	3D*	-	2.120.10.10	Q6DPL2	3.2.1.18	NA	Influenza A virus (A/VietNam/1203/2004(H5N1))	Neuraminidase	0.9	15.5	68
3cl2_D	3D*	-	2.120.10.10	Q6DPL2	3.2.1.18	NA	Influenza A virus (A/VietNam/1203/2004(H5N1))	Neuraminidase	0.9	15.5	68
3cl2_E	3D*	-	2.120.10.10	Q6DPL2	3.2.1.18	NA	Influenza A virus (A/VietNam/1203/2004(H5N1))	Neuraminidase	0.9	15.5	68
3cl2_F	3D*	-	2.120.10.10	Q6DPL2	3.2.1.18	NA	Influenza A virus (A/VietNam/1203/2004(H5N1))	Neuraminidase	0.9	15.5	68
3cl2_H	3D*	-	2.120.10.10	Q6DPL2	3.2.1.18	NA	Influenza A virus (A/VietNam/1203/2004(H5N1))	Neuraminidase	0.9	15.5	68
2qwh_A	3D*	b.68.1.1	2.120.10.10	P03472	3.2.1.18	NA	Influenza A virus (strain A/Tern/Australia /G70C/1975 H11N9)	Neuraminidase	1.4	14.5	56
2qwk_A	3D*	b.68.1.1	2.120.10.10	P03472	3.2.1.18	NA	Influenza A virus (strain A/Tern/Australia /G70C/1975 H11N9)	Neuraminidase	1.4	14.4	57

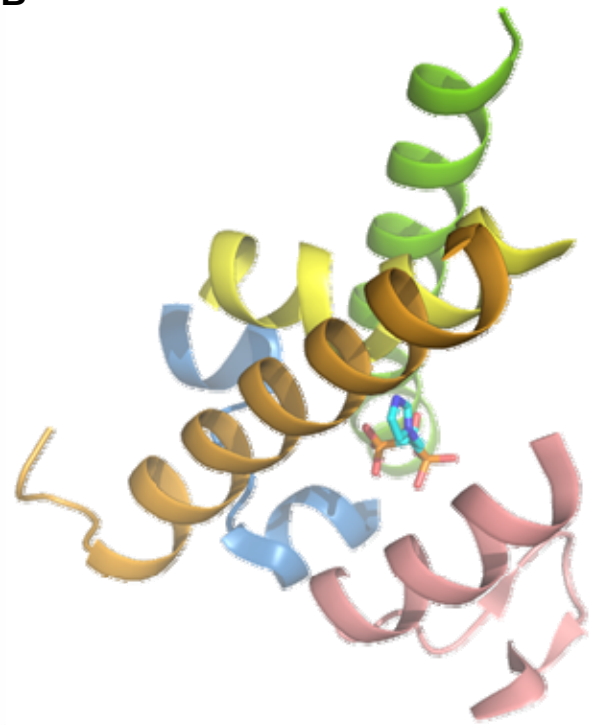
[a] 前四碼代表 PDB code, 最後一碼代表 chain

[b] 是否被 SRP 搜尋為結合環境相似蛋白, 3D 為加入規則前被辨別為具相似結合環境的蛋白, 3D* 為加入規則後被辨別為具相似結合環境的蛋白

A



B



C

Seg#	Length	Start	End	Contact#	Sequence
1	24	73	96	3	HNSSLLIDDIEDNAPLRRGQTTSH
2	15	140	154	1	LINLHRGQGLDIYWR
3	16	167	182	2	YLNMVMNKTGGLFRLT
4	18	204	221	2	NLLGIIYQIRDDYLNLKD
5	15	231	245	1	AEDITEGKLSFPIVH

D

Seg#	Length	Start	End	Contact#	Sequence
1	24	73	96	3	HNSSLLIDDIEDNAPLRRGQTTSH
2	27	134	160	1	TIFNEELINLHRGQGLDIYWRDFLI
3	16	167	182	2	YLNMVMNKTGGLFRLT
4	30	198	227	2	LVPFINLLGIIYQIRDDYLNLKDF(
5	15	231	245	1	AEDITEGKLSFPIVH

圖十一、SRP 建構片段組[以 PDB code 2e91(chain A)為模板]。(A)未考慮建構片段二級結構組成比例 SRP 結構圖 (B)考慮建構片段二級結構組成比例 SRP 結構圖 (C) 未考慮建構片段二級結構組成比例 SRP 序列資訊 (D) 考慮建構片段二級結構組成比例 SRP 序列資訊。

表五、SRP 搜尋結果 [以 PDB code 2e91(chain A)為模板]

PDBID[a]	#[b]	CATH	UniProt AC	Gene name	Species	Description	RMSD	Z	SI(%)
2e91_A	3D	-	Q12051	BTS1	Saccharomyces cerevisiae	Geranylgeranyl pyrophosphate synthase	0	11.6	100
2e91_B	3D	-	Q12051	BTS1	Saccharomyces cerevisiae	Geranylgeranyl pyrophosphate synthase	0.2	11.2	100
2f8z_F	3D*	1.10.600.10	P14324	FDPS	Homo sapiens	Farnesyl pyrophosphate synthase	1.2	7.8	36
2f9k_F	3D*	1.10.600.10	P14324	FDPS	Homo sapiens	Farnesyl pyrophosphate synthase	1.3	7.9	36
1zw5_A	3D*	1.10.600.10	P14324	FDPS	Homo sapiens	Farnesyl pyrophosphate synthase	1.3	8	36
3ez3_A	3D*	-	A5K4U6	-	Plasmodium vivax	-	1.3	7.6	35
3iba_A	3D*	-	Q8WS25	-	Trypanosoma cruzi	Farnesyl diphosphate synthase	1.3	7.6	30
2f8c_F	-	1.10.600.10	P14324	FDPS	Homo sapiens	Farnesyl pyrophosphate synthase	1.3	7.9	36
3ez3_C	-	-	A5K4U6	-	Plasmodium vivax	-	1.3	7.6	35
3ez3_D	3D	-	A5K4U6	-	Plasmodium vivax	-	1.4	7.6	35
3ez3_B	-	-	A5K4U6	-	Plasmodium vivax	-	1.4	7.6	35
2q58_A	-	-	Q5CR09	-	Cryptosporidium parvum Iowa II	-	1.5	7.5	32
2q58_B	-	-	Q5CR09	-	Cryptosporidium parvum Iowa II	-	1.5	7.6	32

[a] 前四碼代表 PDB code, 最後一碼代表 chain

[b] 是否被 SRP 搜尋為結合環境相似蛋白, 3D 為加入規則前被辨別為具相似結合環境的蛋白, 3D* 為加入規則後被辨別為具相似結合環境的蛋白

三、SRP 網站

對於 FDA-530 資料集建構 SRP 的資訊以及對蛋白質結構資料庫搜尋的資訊，我們提供網站呈現搜尋分析結果。圖十二 A 顯示 FDA-530 各蛋白質鏈的蛋白質資訊摘要以及 SRP 搜尋資訊總表，圖十二 B 為 SRP 資訊以及搜尋結果主頁面，分成上半部 SRP 資訊區塊以及下半部搜尋結果區塊，主要呈現我們如何以一蛋白質－配體結晶結構來描述出 SRP 與配體的相對關係，同時提供多個常用以描述蛋白質資訊的註解如 SCOP sccs id、CATH id、UniProt AC、EC number、Gene name、共結晶化合物以及蛋白質物種等資訊，用以協助瞭解該蛋白質是否有可能為潛在能結合配體的蛋白質，並且我們給予結構軟體 DALI 比對結構所提供 RMSD、Z-score 以及 Sequence Identity，來衡量搜尋到的蛋白質鏈是否真正與我們建構的 SRP 的模板擁有相似的結合環境。圖十二 C 是對於搜尋結果蛋白質的細節資訊，提供 DALI 對於 SRP 與搜尋結果的細部結構比對資訊，包含蛋白質疊合的結構資訊以及序列資訊。

網站超連結為： http://e233.life.nctu.edu.tw/~neverfree/SRP/FDA.php

四、方法限制

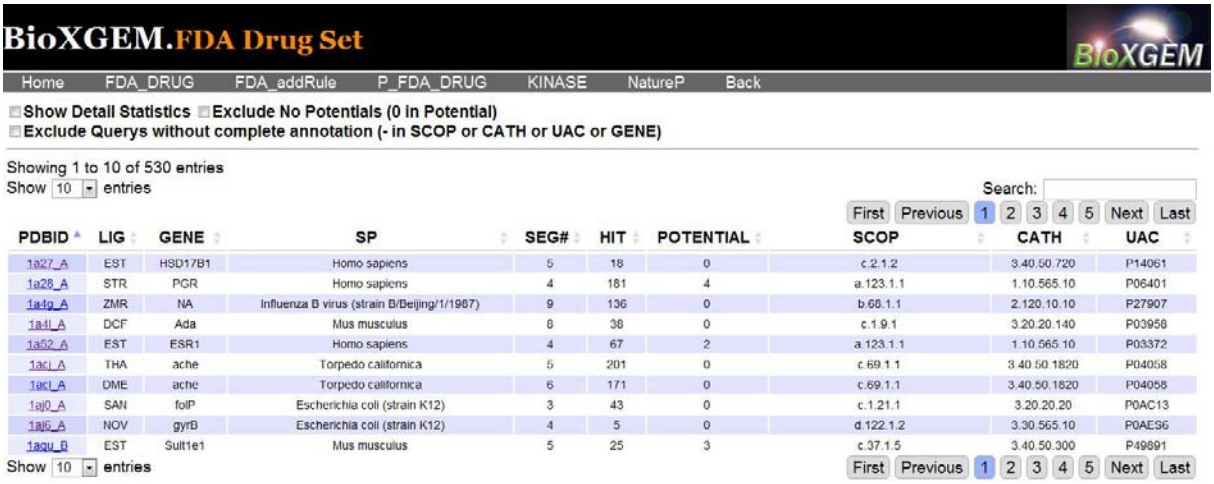
需要結晶結構是 SRP 研究的一大限制，目前已被解序的蛋白質序列如 UniProt 提供的數據為 11,916,373 條蛋白質序列，而 Protein Data Bank 目前收錄了 61,522 個蛋白質結晶結構，由此可見 SRP 研究的範疇較受侷限。尤其，目前我們需要有結晶化合物的蛋白質結晶結構來作為搜尋模板，更加地限制了我們能探討的範圍。對於需要具有結化合物的蛋白質結構的問題，未來可能使用目前已知預測結合位置的軟體做預測後在對預測的結合位建置 SRP。

其次是在將蛋白質結構片段轉換成字元集序列時，承接了 3D-BLAST 對結構編碼的特性，如上章節所討論二級結構組成比例的問題。由於編碼後沒有考慮到蛋白質序列的資訊，所以當搜尋片段單一二級結構(α-helix、β-strand)比例過高時，表示整個片段是一段α-helix 或是β-strand，這樣的結構在於整個蛋白質結構資料庫中不具鑒別度。

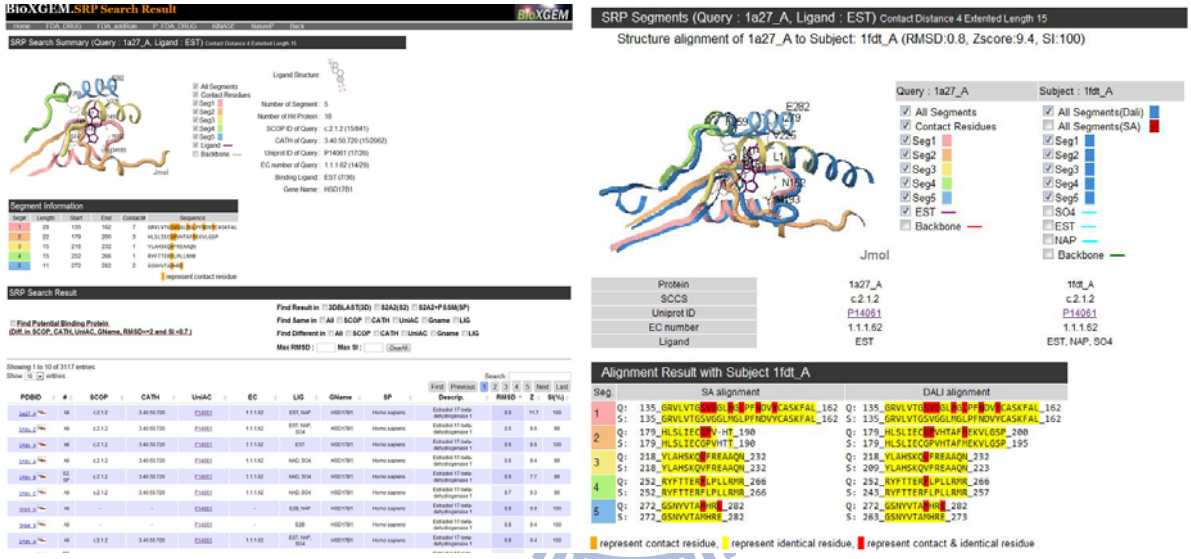
另一個則是結構比對軟體 DALI 的限制。一較過往的結構比對是以蛋白質整體對蛋白質整體的結構比對，我們使用一組空間上不連續的蛋白質片段對蛋白質整體做結構比對，導致某些情形下 DALI 無法做出正確合理的比對。對於這樣的問題，目前研究領域

上已有能夠改善的方法[4, 24-31]，未來可能設法利用文獻提供的方法協助得到更可靠的結構比對結果。

A



B **C**



圖十二、SRP 網站示意圖。(A)FDA-530 資料集資訊總覽頁面 (B)SRP 資訊及搜尋結果主頁面 (C)SRP 搜尋結果細部資訊頁面

肆、結語

一、總結

為了探索化合物其潛在具結合能力的蛋白質，我們提出了 Space-Related Pharmamotif 的概念，藉由描繪出數段集中在蛋白質與配體交互作用環境周圍的蛋白質片段，並結合實驗室先前開發的 3D-BLAST，將三級結構片段轉換成帶有結構資訊的一級結構字元集序列，並以此對蛋白質結構資料庫做快速地搜尋擁有相似結合環境的蛋白質。同時，我們對於 187 個已知擁有蛋白質結晶結構的 FDA 核准藥物，共 530 個蛋白質－藥物分子結晶結構建構 SRP 並對蛋白質結構資料庫(PDB)進行搜尋，以瞭解已知藥物潛在可能結合的蛋白質，進而幫助研究探討其舊藥新用或是副作用的可能性。最後，我們將 530 個 SRP 資訊以及搜尋結果建置於一易於使用者查詢的網站，提供觀察以及分析已知藥物與蛋白質結合環境的資料平台。在本研究中，對於 SRP 這樣的新概念，我們歸納下列一些初步的研究結論：

1. SRP 有能力搜尋擁有相似結合環境的蛋白質。
2. SRP 能辨別結合位的些微變異，以「binding-site family」的概念將蛋白質根據結合環境的差異做分類。
3. SRP 能夠協助舊藥新用的開發以及副作用的研究。

二、主要貢獻與未來研究

近年來，隨著蛋白質結構大量地被結晶，我們藉由利用已知的蛋白質－配體結晶結構建構 SRP 來探討蛋白質－配體結合環境。我們蒐集了 187 個美國食品藥品監督管理局(FDA)核准的藥物共 530 個結晶結構，將其建構 SRP 並對蛋白質結構資料庫進行搜尋，同時提供一網站給予使用者查詢已知藥物對蛋白質結合環境的分析平台，希望能幫助了解對於已知功能藥物其潛在具結合能力的蛋白質，進而提供研究舊藥新用或是副作用的可能。

在未來研究中，我們期望能夠藉由 SRP 的概念將蛋白質以結合環境做分類，對於分類後的蛋白質群，我們能夠藉由多重結構疊合的方式得到具保留性的重要片段組，並能夠將結構轉換成蛋白質序列上的樣板(pattern)，同時能夠對於現有蛋白質結構資料庫建

立用以描述結合環境的序列樣板資料庫，未來研究者將只要提供蛋白質序列資訊，我們即可根據建立的序列樣板資料庫找出可能與該蛋白質擁有相似結合環境的潛在蛋白質。

表六、FDA 藥物資料集

HET group id[a]	Ligand Name	PDB Count[b]	Heavy Atom	Molecular weight
ADN	adenosine	59	19	267.24
PAR	paromomycin	37	42	615.63
017	darunavir	24	38	547.66
EST	conjugated estrogens	22	25	372.41
NMY	neomycin	16	42	614.64
IMN	indomethacin	16	25	357.79
T44	levothyroxine	16	24	776.87
SAL	salicyclic acid	16	10	138.12
MK1	indinavir	14	45	613.79
AZM	acetazolamide	14	13	222.25
STR	progesterone	13	23	314.46
NVP	nevirapine	13	20	266.30
ROC	saquinavir	12	49	670.84
DM1	daunorubicin	12	38	527.52
STI	imatinib	11	37	493.60
TOP	trimethoprim	11	21	290.32
CLM	chloramphenicol	11	20	323.13
DR7	atazanavir	10	51	704.86
1UN	nelfinavir	10	40	567.78
DIF	diclofenac	10	19	296.15
VAN	vancomycin	9	101	1449.25
SVR	suramin	9	86	1297.28
PQN	phytonadione	9	33	450.70
_T3	liothyronine	9	23	650.97
ZOL	zoledronate	9	16	272.09
CFF	caffeine	9	14	194.19
FK5	tacrolimus	8	57	804.02
AB1	lopinavir	8	46	628.80
PDN	prednisone	8	26	358.43
ZMR	zanamivir	8	23	332.31
G39	oseltamivir	8	22	312.40
IBP	ibuprofen	8	15	206.28
152	l-carnitine	8	11	161.20
TA1	paclitaxel	7	62	853.91
TAC	tetracycline	7	32	444.43
PNT	pentamidine	7	25	340.42
MCO	captopril	7	14	217.29

[a] 為 PDB 對化合物的編碼

[b] 代表在本研究使用的蛋白質結構資料庫中結晶該化合物的 PDB 結構數量

表六、FDA 藥物資料集 (Continued)

HET group id[a]	Ligand Name	PDB Count[b]	Heavy Atom	Molecular weight
AIN	aspirin	7	13	180.16
ISZ	isoniazid	7	10	137.14
DM2	doxorubicin	6	39	543.52
TFP	trifluoperazine	6	28	407.50
EAA	ethacrynic acid	6	19	303.14
FLP	flurbiprofen	6	18	244.26
PRL	proflavine	6	16	209.25
ACA	aminocaproic acid	6	9	131.17
ERY	erythromycin	5	51	733.93
RIT	ritonavir	5	50	720.94
FUA	fusidic acid	5	37	516.71
478	amprenavir	5	35	505.63
KAN	kanamycin	5	33	484.50
VIA	sildenafil	5	33	474.58
D16	raltitrexed	5	32	458.49
BRL	rosiglitazone	5	25	357.43
FLU	fluorescein	5	25	332.31
AIC	ampicillin	5	24	349.41
SCM	spectinomycin	5	23	332.35
EFZ	efavirenz	5	21	315.68
GNT	galantamine	5	21	287.35
DES	diethylstilbestrol	5	20	268.35
IM2	imipenem	5	20	299.35
EDT	edetic acid	5	20	292.24
SHH	vorinostat	5	19	264.32
CER	cerulenin	5	16	223.27
MTL	mannitol	5	12	182.17
URF	fluorouracil	5	9	130.08
TPV	tipranavir	4	42	602.66
SRY	streptomycin	4	40	581.57
MRC	mupirocin	4	35	500.62
RAL	raloxifene	4	34	473.58
VDN	vardenafil	4	34	488.60
1N1	dasatinib	4	33	488.01
BAX	sorafenib	4	32	464.83
IU5	ursodeoxycholic acid	4	28	392.57
AJM	ajmaline	4	24	326.43
MOA	mycophenolic acid	4	23	320.34
P1Z	phenylbutazone	4	23	308.37

表六、FDA 藥物資料集 (Continued)

HET group id[a]	Ligand Name	PDB Count[b]	Heavy Atom	Molecular weight
AZZ	zidovudine	4	19	267.24
1PT	oxaliplatin	4	15	397.29
THA	tacrine	4	15	198.26
GBN	gabapentin	4	12	171.24
LDP	dopamine	4	11	153.18
TYL	acetaminophen	4	11	151.16
FCN	fosfomycin	4	8	138.06
PPF	foscarnet	4	7	126.01
SYB	quinupristin	3	121	1713.07
ZIT	azithromycin	3	52	748.98
DOL	dalfopristin	3	48	690.85
NOV	novobiocin	3	44	612.62
DM5	idarubicin	3	36	497.49
486	mifepristone	3	32	429.59
IRE	gefitinib	3	31	446.90
LYA	pemetrexed	3	31	427.41
LPR	lisinopril	3	29	405.49
DEX	dexamethasone	3	28	392.46
CLY	clindamycin	3	27	424.98
TMQ	trimetrexate	3	27	369.42
MER	meropenem	3	26	383.46
097	marimastat	3	23	331.41
PNN	penicillin g	3	23	334.39
CCA	cocaine	3	22	303.35
FFA	testosterone	3	21	288.42
2TN	atenolol	3	19	266.34
DME	decamethonium	3	18	258.49
GEO	gemcitabine	3	18	263.20
CP6	pyrimethamine	3	17	248.71
ML1	melatonin	3	17	232.28
EZL	ethoxzolamide	3	16	258.32
TEP	theophylline	3	13	180.16
EDR	edrophonium	3	12	166.24
LNR	norepinephrine	3	12	169.18
NCT	nicotine	3	12	162.23
308	amantadine	3	11	151.25
BHA	aminosalicylic acid	3	11	153.14
TEL	telithromycin	2	58	812.00
TXL	docetaxel	2	58	807.88

表六、FDA 藥物資料集 (Continued)

HET group id[a]	Ligand Name	PDB Count[b]	Heavy Atom	Molecular weight
TAO	troleandomycin	2	57	813.97
CTY	clarithromycin	2	52	747.95
OBN	ouabain	2	41	584.65
MIY	minocycline	2	33	457.48
MIX	mitoxantrone	2	32	444.48
TOY	tobramycin	2	32	467.51
AGG	tirofiban	2	30	440.60
AQ4	erlotinib	2	29	393.44
B49	sunitinib	2	29	398.47
CIA	tadalafil	2	29	389.40
SNL	spironolactone	2	29	416.57
AZR	aztreonam	2	28	435.43
CFX	cefoxitin	2	28	427.45
BEP	bepridil	2	27	366.54
CEL	celecoxib	2	26	381.37
CLS	cefalotin	2	26	396.44
HCY	hydrocortisone	2	26	362.46
EV1	papaverine	2	25	339.39
CPF	ciprofloxacin	2	24	331.34
CYZ	cyclothiazide	2	24	389.88
PNV	penicillin v	2	24	350.39
ZLD	linezolid	2	24	337.35
NOG	norgestrel	2	23	312.45
CXX	clomipramine	2	22	314.85
NDR	norethindrone	2	22	298.42
OIN	atropine	2	21	289.37
DSM	desipramine	2	20	266.38
NFL	niflumic acid	2	20	282.22
TAZ	tazobactam	2	20	300.29
CBL	chlorambucil	2	19	304.21
CDX	dexrazoxane	2	19	268.27
CL9	cladribine	2	19	285.69
DCF	pentostatin	2	19	268.27
DX2	triamterene	2	19	253.26
1FL	diflunisal	2	18	250.20
AHD	alendronate	2	14	249.10
ALE	epinephrine	2	13	183.20
PFL	propofol	2	13	178.27
CLW	chlorzoxazone	2	11	169.57

表六、FDA 藥物資料集 (Continued)

HET group id[a]	Ligand Name	PDB Count[b]	Heavy Atom	Molecular weight
DX4	thioguanine	2	11	167.19
ICF	isoflurane	2	10	184.49
SC2	acetylcysteine	2	10	163.20
RPT	rifapentine	1	63	877.03
RBT	rifabutin	1	61	847.00
VLB	vinblastine	1	59	810.97
JOS	josamycin	1	58	828.00
ROX	roxithromycin	1	58	837.05
DGX	digoxin	1	55	780.94
REM	remikiren	1	44	630.84
CP0	irinotecan	1	43	586.68
117	atorvastatin	1	41	558.64
MTK	montelukast	1	41	586.18
AKN	amikacin	1	40	585.60
C41	aliskiren	1	39	551.76
G34	retapamulin	1	36	517.76
KLN	ketoconazole	1	36	531.43
MXL	latamoxef	1	36	520.47
IDB	iodipamide	1	34	1139.76
FBI	rosuvastatin	1	33	481.54
MT1	methotrexate	1	33	454.44
DXT	doxycycline	1	32	444.43
SPP	delavirdine	1	32	456.56
TTC	topotecan	1	31	421.45
115	fluvastatin	1	30	411.47
15M	bimatoprost	1	30	415.57
CE3	cefotaxime	1	30	455.47
POD	podofilox	1	30	414.41
803	lovastatin	1	29	404.54
CXN	cloxacillin	1	29	435.88
LOC	colchicine	1	29	399.44
MFX	moxifloxacin	1	29	401.43
NFN	nafcillin	1	29	414.48
715	sitagliptin	1	28	407.31
BO2	bortezomib	1	28	384.24
CTX	tamoxifen	1	28	371.51
CVI	gentian violet	1	28	372.53
QUN	quinacrine	1	28	399.96
SAS	sulfasalazine	1	28	398.39

表六、FDA 藥物資料集 (Continued)

HET group id[a]	Ligand Name	PDB Count[b]	Heavy Atom	Molecular weight
FIT	finasteride	1	27	372.54
9RA	bexarotene	1	26	348.48
LFX	levofloxacin	1	26	361.37
225	felodipine	1	25	384.25
CQA	amodiaquine	1	25	355.86
PEM	bezafibrate	1	25	361.82
5CH	etoricoxib	1	24	358.84
BL1	indapamide	1	24	365.84
CIL	cilastatin	1	24	358.45
LOR	loracarbef	1	24	349.77
TMI	bifonazole	1	24	310.39
BZ1	brinzolamide	1	23	383.51
PMZ	acepromazine	1	23	326.46
PZQ	praziquantel	1	23	312.41
CLQ	chloroquine	1	22	319.87
KLT	chlorthalidone	1	22	338.77
RFX	fluoxetine	1	22	309.33
TOR	topiramate	1	22	339.36
TPF	fluconazole	1	22	306.27
ESL	estriol	1	21	288.38
FUN	furosemide	1	21	330.74
H3P	hexachlorophene	1	21	406.90
IXX	imipramine	1	21	280.41
MOI	morphine	1	21	285.34
TIM	timolol	1	21	316.42
CFB	clofarabine	1	20	303.68
DZP	diazepam	1	20	284.74
ESR	estradiol	1	20	272.38
PNX	pentoxifylline	1	20	278.31
SCK	succinylcholine	1	20	290.40
SRE	sertraline	1	20	306.23
2PM	diphenhydramine	1	19	255.35
BFQ	ibandronate	1	19	319.23
CFV	cidofovir	1	18	279.19
GA2	ganciclovir	1	18	255.23
PE2	penciclovir	1	18	253.26
STZ	streptozocin	1	18	265.22
X0T	dyphylline	1	18	254.24
2DI	didanosine	1	17	236.23

表六、FDA 藥物資料集 (Continued)

HET group id[a]	Ligand Name	PDB Count[b]	Heavy Atom	Molecular weight
AR3	cytarabine	1	17	243.22
BHS	tetrahydrobiopterin	1	17	241.25
NIX	nalidixic acid	1	17	232.24
8MO	methoxsalen	1	16	216.19
BZM	benzyl benzoate	1	16	212.24
ETV	emtricitabine	1	16	247.25
I7A	dichlorphenamide	1	16	305.16
MIL	milrinone	1	16	211.22
3TC	lamivudine	1	15	229.26
AMR	amiloride	1	15	229.63
NBV	miglustat	1	15	219.28
CLU	clonidine	1	14	230.09
DAH	levodopa	1	14	197.19
J01	clavulanate	1	14	199.16
NFZ	nitrofurazone	1	14	198.14
TMG	thiabendazole	1	14	201.25
210	pamidronate	1	13	235.07
37T	theobromine	1	13	180.16
AZ1	azelaic acid	1	13	188.22
RIM	rimantadine	1	13	179.30
2MN	metronidazole	1	12	171.15
B40	methamphetamine	1	11	149.23
SAN	sulfanilamide	1	11	172.21
CCE	carbachol	1	10	147.20
PM6	mercaptopurine	1	10	152.18
PZA	pyrazinamide	1	9	123.11
MMZ	methimazole	1	7	114.17
X2N	posaconazole	0	51	700.78
CL6	clotrimazole	0	25	344.84
ECN	econazole	0	24	381.68
TFO	tenofovir	0	19	287.21
PFN	fenoprofen	0	18	242.27
MYT	metyrapone	0	17	226.27
142	carbidopa	0	16	226.23
MXD	minoxidil	0	15	209.25
RAS	rasagiline	0	13	171.24
1LP	tranylcypromine	0	10	133.19
2PP	valproic acid	0	10	144.21
LE1	penicillamine	0	9	149.21

表六、FDA 藥物資料集 (Continued)

HET group id[a]	Ligand Name	PDB Count[b]	Heavy Atom	Molecular weight
4PZ	fomepizole	0	6	82.10
HAE	acetohydroxamic acid	13	5	75.07
DMS	dimethyl sulfoxide	236	4	78.13
GAI	guanidine	28	4	59.07
AF3	aluminium	17	4	26.98
DHL	cysteamine	0	4	77.15
EOH	ethanol	120	3	46.07
_NO	nitric oxide	30	2	30.01
_ZN	zinc	5842	1	65.41
_FE	iron	1081	1	55.85
_LI	lithium	35	1	6.94

表七、FDA-117 資料集

PDB ID[a]	Ligand	SCOP	CATH	UniProt AC	Overall Recall	Overall Pricision
1a28_A	STR	a.123.1.1	1.10.565.10	P06401	0.30	0.93
1a4l_A	DCF	c.1.9.1	3.20.20.140	P03958	0.25	0.97
1acl_A	DME	c.69.1.1	3.40.50.1820	P04058	0.21	0.90
1b3n_A	CER	c.95.1.1	3.40.47.10	P0AAI5	0.44	0.71
1bkf_A	FK5	d.26.1.1	3.10.50.40	P62942	0.68	0.89
1bzf_A	TMQ	c.71.1.1	3.40.430.10	P00381	0.60	0.67
1cet_A	CLQ	c.2.1.5, d.162.1.1	3.40.50.720	Q27743	0.05	0.80
1cqe_A	FLP	a.93.1.2, g.3.11.1	1.10.640.10	P05979	0.21	1.00
1ctr_A	TFP	a.39.1.5	1.10.238.10	P62158	0.05	0.84
1d4f_A	ADN	c.2.1.4, c.23.12.3	3.40.50.720	P10760	0.02	0.72
1dyr_A	TOP	c.71.1.1	3.40.430.10	P16184	0.46	0.70
1dzm_A	BZM	b.60.1.1	2.40.128.20	P81245	0.21	0.81
1ei6_A	PPF	c.76.1.4	3.30.1360.110	Q51782	0.44	1.00
1eqg_A	IBP	a.93.1.2, g.3.11.1	1.10.640.10	P05979	0.18	1.00
1eta_1	T44	b.3.4.1	2.60.40.180	P02766	0.94	1.00
1fb7_A	ROC	b.50.1.1	2.40.70.10	P04585	0.57	0.89
1fcm_A	CXN	e.3.1.1	3.40.710.10	P00811	0.28	0.99
1fcn_A	LOR	e.3.1.1	3.40.710.10	P00811	0.28	0.99
1fk9_A	EFZ	c.55.3.1, e.8.1.2	3.10.10.10	P04585	0.26	0.97
1gsf_A	EAA	a.45.1.1, c.47.1.5	1.20.1050.10	P08263	0.21	0.94
1gwr_A	EST	a.123.1.1	1.10.565.10	P03372	0.19	0.97
1h61_A	PDN	c.1.4.1	3.20.20.70	P71278	0.01	1.00
1hvy_A	D16	d.117.1.1	3.30.572.10	P04818	0.89	0.86
1i1e_A	DM2	b.29.1.6, b.42.4.2, d.92.1.7, h.4.2.1	1.20.1120.10	P10844	0.32	1.00
1i2w_A	CFX	e.3.1.1	3.40.710.10	P00808	0.35	0.84
1ihi_A	IU5	c.1.7.1	3.20.20.100	P52895	0.26	0.93
1itu_A	CIL	c.1.9.7	3.20.20.140	P16444	0.03	0.44
1j3j_A	CP6	c.71.1.1	3.40.430.10	P13922	0.39	0.68
1jqe_A	QUN	c.66.1.19	3.40.50.150	P50135	0.04	1.00
1ju6_A	LYA	d.117.1.1	3.30.572.10	P04818	0.64	0.83
1jvj_A	IM2	e.3.1.1	3.40.710.10	P62593	0.36	0.86
1jzs_A	MRC	a.27.1.1, b.51.1.1, c.26.1.1	1.10.730.10	P56690	0.10	0.56
1k6r_A	MXL	e.3.1.1	3.40.710.10	P14489	0.13	0.95
1ki2_A	GA2	c.37.1.1	3.40.50.300	P03176	0.02	1.00
1ki3_A	PE2	c.37.1.1	3.40.50.300	P03176	0.03	1.00

[a] 前四碼代表 PDB code, 最後一碼代表 chain

表七、FDA-117 資料集(Continue)

PDB ID[a]	Ligand	SCOP	CATH	UniProt AC	Overall Recall	Overall Pricision
1klm_A	SPP	c.55.3.1, e.8.1.2	3.10.10.10	P04585	0.29	0.95
1kvl_A	CLS	e.3.1.1	3.40.710.10	P00811	0.28	0.99
1lhv_A	NOG	b.29.1.4	2.60.120.200	P04278	0.01	1.00
1m17_A	AQ4	d.144.1.7	1.10.510.10	P00533	0.07	0.84
1m2x_A	MCO	d.157.1.1	3.60.15.10	O08498	0.57	0.90
1m4d_A	TOY	d.108.1.1	3.40.630.30	P0A5N0	0.03	1.00
1m8d_A	CLW	d.174.1.1	3.90.340.10	P29477	0.72	0.85
1meh_A	MOA	c.1.5.1	3.20.20.70	P50097	0.01	0.49
1mx1_A	THA	c.69.1.1	3.40.50.1820	P23141	0.09	1.00
1n0s_A	FLU	b.60.1.1	2.40.128.20	P09464	0.02	0.45
1nd4_A	KAN	d.144.1.6	3.90.1200.10	P00552	0.14	0.83
1nhz_A	486	a.123.1.1	1.10.565.10	P04150	0.13	0.95
1nnc_A	ZMR	b.68.1.1	2.120.10.10	P03472	0.65	1.00
1nnf_A	EDT	c.94.1.1	3.40.190.10	P35755	0.04	1.00
1nx9_A	AIC	b.18.1.13, c.69.1.21	1.10.3020.10	Q8VRK8	0.80	1.00
1ohp_A	ESR	d.17.4.3	3.10.450.50	P00947	0.10	0.90
1oq5_A	CEL	b.74.1.1	3.10.200.10	P00918	0.99	0.95
1oxr_A	AIN	a.133.1.2	1.20.90.10	P60045	0.24	0.98
1p5z_B	AR3	c.37.1.1	3.40.50.300	P27707	0.06	0.98
1p62_B	GEO	c.37.1.1	3.40.50.300	P27707	0.06	1.00
1p7r_A	NCT	a.104.1.1	1.10.630.10	P00183	0.24	0.98
1p93_A	DEX	a.123.1.1	1.10.565.10	P04150	0.22	0.94
1pnv_B	VAN	c.87.1.5	3.40.50.2000	P96558	0.02	1.00
1qca_A	FUA	c.43.1.1	3.30.559.10	P00484	0.60	1.00
1qkn_A	RAL	a.123.1.1	1.10.565.10	Q62986	0.38	0.95
1qvt_A	PRL	a.4.1.9, a.121.1.1	1.10.10.60	P0A0N4	0.08	0.79
1rj6_A	AZM	b.74.1.1	3.10.200.10	Q9WVT6	0.97	0.89
1rkw_A	PNT	a.4.1.9, a.121.1.1	1.10.10.60	P0A0N4	0.22	0.97
1rxc_F	URF	c.56.2.1	3.40.50.1580	P12758	0.37	0.95
1s14_A	NOV	d.122.1.2	3.30.565.10	P20083	0.06	0.57
1s9p_A	DES	a.123.1.1	1.10.565.10	P62508	0.29	0.87
1sqn_A	NDR	a.123.1.1	1.10.565.10	P06401	0.19	0.91
1sv9_A	DIF	a.133.1.2	1.20.90.10	P59071	0.31	0.90
1t69_A	SHH	c.42.1.2	3.40.800.20	Q9BY41	0.63	1.00
1td7_A	NFL	a.133.1.2	1.20.90.10	P60045	0.30	0.99
1tlm_A	MIL	b.3.4.1	2.60.40.180	P02766	0.85	0.84

[a] 前四碼代表 PDB code, 最後一碼代表 chain

表七、FDA-117 資料集(Continue)

PDB ID[a]	Ligand	SCOP	CATH	UniProt AC	Overall Recall	Overall Pricision
1u65_A	CP0	c.69.1.1	3.40.50.1820	P04058	0.22	0.90
1uae_A	FCN	d.68.2.2	3.65.10.10	P0A749	0.30	1.00
1usq_A	CLM	b.2.3.6	2.60.40.1570	P24093	0.79	1.00
1uwj_A	BAX	d.144.1.7	1.10.510.10	P15056	0.31	0.79
1v3q_E	2DI	c.56.2.1	3.40.50.1580	P00491	0.20	0.82
1w6r_A	GNT	c.69.1.1	3.40.50.1820	P04058	0.22	0.90
1x70_A	715	b.70.3.1, c.69.1.24	2.140.10.30	P27487	0.98	0.99
1xos_A	VIA	a.211.1.2	1.10.1300.10	Q07343	0.97	0.70
1xot_A	VDN	a.211.1.2	1.10.1300.10	Q07343	0.98	0.69
1xzx_X	T3	a.123.1.1	1.10.565.10	P10828	0.08	0.93
1z9y_A	FUN	b.74.1.1	3.10.200.10	P00918	0.99	0.94
1zgy_A	BRL	a.123.1.1	1.10.565.10	P37231	0.14	0.89
1zzq_A	MTL	d.174.1.1	3.90.340.10	P29476	0.96	0.94
2a3r_A	LDP	c.37.1.5	3.40.50.300	P50224	0.02	0.86
2a7q_A	CFB	c.37.1.1	3.40.50.300	P27707	0.07	0.96
2ack_A	EDR	c.69.1.1	3.40.50.1820	P04058	0.25	0.90
2aot_A	2PM	c.66.1.19	3.40.50.150	P50135	0.04	1.00
2b0q_A	NMY	d.144.1.6	3.30.200.20	P0A3Y5	0.11	1.00
2bxf_A	DZP	a.126.1.1	1.10.246.10	P02768	0.78	0.98
2bxq_A	P1Z	a.126.1.1	1.10.246.10	P02768	0.83	1.00
2f38_A	15M	c.1.7.1	3.20.20.100	P42330	0.26	0.94
2fum_A	MIX	d.144.1.7	1.10.510.10	P0A5S4	0.36	0.83
2i2z_A	SAL	a.126.1.1	1.10.246.10	P02768	0.66	0.99
2ij7_B	TPF	a.104.1.1	1.10.630.10	P0A514	0.05	1.00
2inq_A	MT1	c.71.1.1	3.40.430.10	P0ABQ4	0.62	0.63
2ity_A	IRE	d.144.1.7	1.10.510.10	P00533	0.11	0.78
2jj8_A	AZZ	c.37.1.1	3.40.50.300	Q9XZT6	0.05	0.99
2nni_A	MTK	a.104.1.1	1.10.630.10	P10632	0.06	0.96
2nnj_A	225	a.104.1.1	1.10.630.10	P10632	0.16	0.98
2o7o_A	DXT	a.4.1.9, a.121.1.1	1.10.10.60	P0ACT4	0.07	0.76
2otf_A	2TN	a.133.1.2	1.20.90.10	P59071	0.07	0.83
2pl0_A	STI	d.144.1.7	1.10.510.10	P06239	0.16	0.79
2pou_A	I7A	b.74.1.1	3.10.200.10	P00918	0.99	0.94
2qwk_A	G39	b.68.1.1	2.120.10.10	P03472	0.65	1.00
2trt_A	TAC	a.4.1.9, a.121.1.1	1.10.10.60	P0ACT4	0.07	0.77
2v3d_A	NBV	b.71.1.2, c.1.8.3	2.60.40.1180	P04062	0.08	1.00

[a] 前四碼代表 PDB code, 最後一碼代表 chain

表七、FDA-117 資料集(Continue)

PDB ID[a]	Ligand	SCOP	CATH	UniProt AC	Overall Recall	Overall Pricision
2zd8_A	MER	e.3.1.1	3.40.710.10	P0AD64	0.30	0.81
3bl1_A	BL1	b.74.1.1	3.10.200.10	P00918	0.99	0.94
3caj_A	EZL	b.74.1.1	3.10.200.10	P00918	0.99	0.95
3cfl_A	5CH	c.94.1.2	3.40.190.10	P24627	0.10	1.00
3d2t_A	1FL	b.3.4.1	2.60.40.180	P02766	0.75	0.82
3hvt_A	NVP	c.55.3.1, e.8.1.2	3.10.10.10	P03366	0.19	0.96
3pah_A	ALE	d.178.1.1	1.10.800.10	P00439	0.97	0.97
3znc_A	BZ1	b.74.1.1	3.10.200.10	Q64444	0.90	0.87
4cox_A	IMN	a.93.1.2, g.3.11.1	1.10.640.10	Q05769	0.20	0.94
4pah_A	LNR	d.178.1.1	1.10.800.10	P00439	0.91	0.97

[a] 前四碼代表 PDB code, 最後一碼代表 chain

參考文獻

1. Adams, C.P. and V.V. Brantner, *Estimating the cost of new drug development: is it really 802 million dollars?* Health Aff (Millwood), 2006. **25**(2): p. 420-8.
2. Kennedy, T., *Managing the drug discovery/development interface.* Drug Discovery Today, 1997. **2**(10): p. 436-444.
3. Keiser, M.J., et al., *Relating protein pharmacology by ligand chemistry.* Nature Biotechnology, 2007. **25**(2): p. 197-206.
4. Durrant, J.D., et al., *A multidimensional strategy to detect polypharmacological targets in the absence of structural and sequence homology.* PLoS Computational Biology, 2010. **6**(1): p. e1000648.
5. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Research, 1997. **25**(17): p. 3389-3402.
6. Altschul, S.F., et al., *Basic Local Alignment Search Tool.* Journal of Molecular Biology, 1990. **215**(3): p. 403-410.
7. Shindyalov, I.N. and P.E. Bourne, *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.* Protein Engineering, 1998. **11**(9): p. 739-747.
8. Holm, L. and C. Sander, *Protein-Structure Comparison by Alignment of Distance Matrices.* Journal of Molecular Biology, 1993. **233**(1): p. 123-138.
9. Schreiber, S.L., *Small molecules: the missing link in the central dogma.* Nature Chemical Biology, 2005. **1**(2): p. 64-66.
10. Paolini, G.V., et al., *Global mapping of pharmacological space.* Nature Biotechnology, 2006. **24**(7): p. 805-815.
11. Vieth, M., et al., *Kinomics-structural biology and chemogenomics of kinase inhibitors and targets.* Biochimica Et Biophysica Acta-Proteins and Proteomics, 2004. **1697**(1-2): p. 243-257.
12. Izrailev, S. and M.A. Farnum, *Enzyme classification by ligand binding.* Proteins-Structure Function and Bioinformatics, 2004. **57**(4): p. 711-724.
13. Kinoshita, K. and H. Nakamura, *Protein informatics towards function identification.* Current Opinion in Structural Biology, 2003. **13**(3): p. 396-400.
14. Skolnick, J. and J.S. Fetrow, *From genes to protein structure and function: novel applications of computational approaches in the genomic era.* Trends in Biotechnology, 2000. **18**(1): p. 34-39.
15. Hvidsten, T.R., et al., *A Comprehensive Analysis of the Structure-Function Relationship in Proteins Based on Local Structure Similarity.* Plos One, 2009. **4**(7): p. -.
16. Yang, J.M. and C.H. Tung, *Protein structure database search and evolutionary classification.* Nucleic Acids Research, 2006. **34**(13): p. 3646-3659.
17. Tung, C.H., J.W. Huang, and J.M. Yang, *Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for rapid search of protein structure database.* Genome Biology, 2007. **8**(3): p. -.
18. Murzin, A.G., et al., *Scop - a Structural Classification of Proteins Database for the Investigation of Sequences and Structures.* Journal of Molecular Biology, 1995. **247**(4): p. 536-540.
19. Orengo, C.A., et al., *CATH - a hierarchic classification of protein domain structures.* Structure, 1997. **5**(8): p. 1093-1108.

20. Sigrist, C.J.A., et al., *PROSITE, a protein domain database for functional characterization and annotation.* Nucleic Acids Research, 2010. **38**: p. D161-D166.
21. Hata, K., et al., *Limited inhibitory effects of oseltamivir and zanamivir on human sialidases.* Antimicrobial Agents and Chemotherapy, 2008. **52**(10): p. 3484-3491.
22. Joensuu, H., et al., *Effect of the tyrosine kinase inhibitor STI571 in a patient with a metastatic gastrointestinal stromal tumor.* New England Journal of Medicine, 2001. **344**(14): p. 1052-1056.
23. Karaman, M.W., et al., *A quantitative analysis of kinase inhibitor selectivity.* Nature Biotechnology, 2008. **26**(1): p. 127-132.
24. Coleman, R.G. and K.A. Sharp, *Travel depth, a new shape descriptor for macromolecules: Application to ligand binding.* Journal of Molecular Biology, 2006. **362**(3): p. 441-458.
25. Nayal, M. and B. Honig, *On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites.* Proteins-Structure Function and Bioinformatics, 2006. **63**(4): p. 892-906.
26. Coleman, R.G., et al., *An intuitive approach to measuring protein surface curvature.* Proteins-Structure Function and Bioinformatics, 2005. **61**(4): p. 1068-1074.
27. Agarwal, P.K., et al., *Extreme elevation on a 2-manifold.* Discrete & Computational Geometry, 2006. **36**(4): p. 553-572.
28. Hendrix, D.K. and I.D. Kuntz, *A surface solid angle-based shape descriptor for molecular docking.* Biophysical Journal, 1997. **72**(2): p. Wp432-Wp432.
29. Liang, J., H. Edelsbrunner, and C. Woodward, *Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design.* Protein Science, 1998. **7**(9): p. 1884-1897.
30. Norel, R., H.J. Wolfson, and R. Nussinov, *Small molecule recognition: Solid angles surface representation and molecular shape complementarity.* Combinatorial Chemistry & High Throughput Screening, 1999. **2**(4): p. 223-236.
31. Watson, J.D., R.A. Laskowski, and J.M. Thornton, *Predicting protein function from sequence and structural data.* Current Opinion in Structural Biology, 2005. **15**(3): p. 275-284.