❀MEDock 簡介：兼談虛擬篩選在新藥開發上的應用

• 林榮信

虛擬篩選(virtual screening)目前已經是開發新藥中很重要的一種方法。顧名思義，運用虛擬篩選來開發新藥，並不是利用實驗的方法，像是高效能篩選(High-Throughput Screening, HTS)，來找出有效用的化學分子；而是利用計算的方法，在電腦上面來篩選出可能有治療效用的化學分子。虛擬篩選可以簡單地分做兩大類：一種是在不知道受體結構，而僅知某一系列化學分子似乎有著相似的活性與藥理概廓(pharmacological profile)，這樣的情形下來做虛擬篩選；另一種則是先由結構生物學(structural biology)的方法得到藥物的標靶三度空間的高解析度結構後，接下去再來做虛擬篩選。一般來說，後者的應用較爲廣泛，也比較有機會找到較新穎的藥物，而且後者可以探討藥物與受體(receptor)的可能結合模式，因此可以知道如果受體某些位置有氨基酸改變時，會有什麼樣的可能影響。此外，要繼續發展到先導藥物的最佳化(lead optimization)，知道受體的結構也十分重要。我們以下的介紹，主要是在第二類的方法。

要做到成功的藥物虛擬篩選，有三個主要的面向要考慮到：(1)要考慮到足夠大的化學空間(chemical space)；(2)要有準確的評分函數(scoring function)；(3)要有高效率的搜尋演算法(searching algorithm)。所謂的化學空間，狹義地講，就是所有要考慮的化學分子的集合，實際應用上，就是所用的化學資料庫；廣義來講，化學空間則包括了配體 (ligand) 和受體 (receptor) 較可能存在的構形 (conformation)。評分函數的設計，主要是從蛋白質結構資料庫 (Protein Databank, http://www.pdb.org) 中尋找 X-ray 解出的結構中解析度較好的配體與受體的結合體 (complex)，運用到一些能量的計算，加上諸如多變量線性回歸 (multivariate linear regression) 這樣的統計方法，以建立一個準確的數學函數。這樣的函數可以快速有效地推測未知的配體對一給定的受體，是否可以穩定地結合，甚至預測結合強度 (binding affinity) 或結合的自由能 (binding free energy)。由於配體分子相對於受體做平移 (translation)、旋轉 (rotation)、或者是構形變化 (conformational changes) 時能量的變化情形十分複雜，而尋找具有最佳評分值的結合模式 (binding mode)，則有如在高低起伏不斷的山巒尋找最深的山谷，像英語俗諺的「在稻草堆中找一根針」(Finding a needle in a hay stack) 一樣的困難，所以這時一個有效的搜尋演算法便扮演了極重要的角色。

這三項虛擬篩選的基本要素相互之間的關係可說是難解難分：如果評分函數不夠好，則不論搜尋演算法再怎麼有效，也很難定出正確的結合模式；相對地，如果評分函數雖然很好，但搜尋演算法並不怎麼有效率，這樣要定出正確的結合模式可能機會也很低；再者，即便評分函數和搜尋演算法都很不錯，但所搜尋的化學空間不夠大，或者這個化學空間或化學資料庫的化學多樣性 (chemical diversity) 不夠好，這樣要找到新的化學分子 (new chemical entity, NCE) 或甚至要找吸收 (adsorption)、分佈 (distribution)、代謝 (metabolism)、排除 (excretion) 等 (合稱 ADME) 藥動性質 (pharmacokinetic properties) 較好的、有市場發展潛力的分子的機會便不是很高。所以虛擬篩選是一個複雜的問題，而一個成功的虛擬篩選則有賴許多環節和細節可以做到緊密相扣。

在虛擬篩選中最常用到的化學資料庫有 MDL 公司所提供的 Available Chemical Directory (ACD)、美國國家癌症研究所 (National Cancer Institute) 所提供的 NCI 資料庫、Derwent 所整理的 World Drug Index (WDI) 等等。ACD 資料庫可配合 MDL 公司自己所開發出的圖形介面程式 ISIS/Base 作爲搜尋工具，因此使用上十分方便。目前一般常用的 ACD 資料庫中大約有 280,000 個

化學分子，但ACD也有另一個ACDSC資料庫，所蒐藏的化學資料庫則多達1,000,000以上，可做爲進一步搜尋時之用。NCI的資料庫事實上可以透過他們所建立的網站 (http://cactus.nci.nih.gov/ncidb2/) 來搜尋，目前大約收藏有25,000個藥物分子的資料。

如果希望所搜尋到的化學分子能夠有不錯的溶解度 (solubility) 和穿透率 (permeability)，我們也可以參考Merck藥廠的Lipinski在1997年所建議的rule of five作爲資料庫篩選的準則。所謂的rule of five是：分子的hydrogen bond donors的數目不應該大於5，分子的hydrogen bond acceptor的數目不應該大於10，分子量應該小於500，以及cLogP值應該小於5。不過，也有許多實驗室在從事虛擬篩選時並不嚴格遵守rule of five，而將溶解度與穿透率的問題留到要做先導化合物最佳化時再來處理。

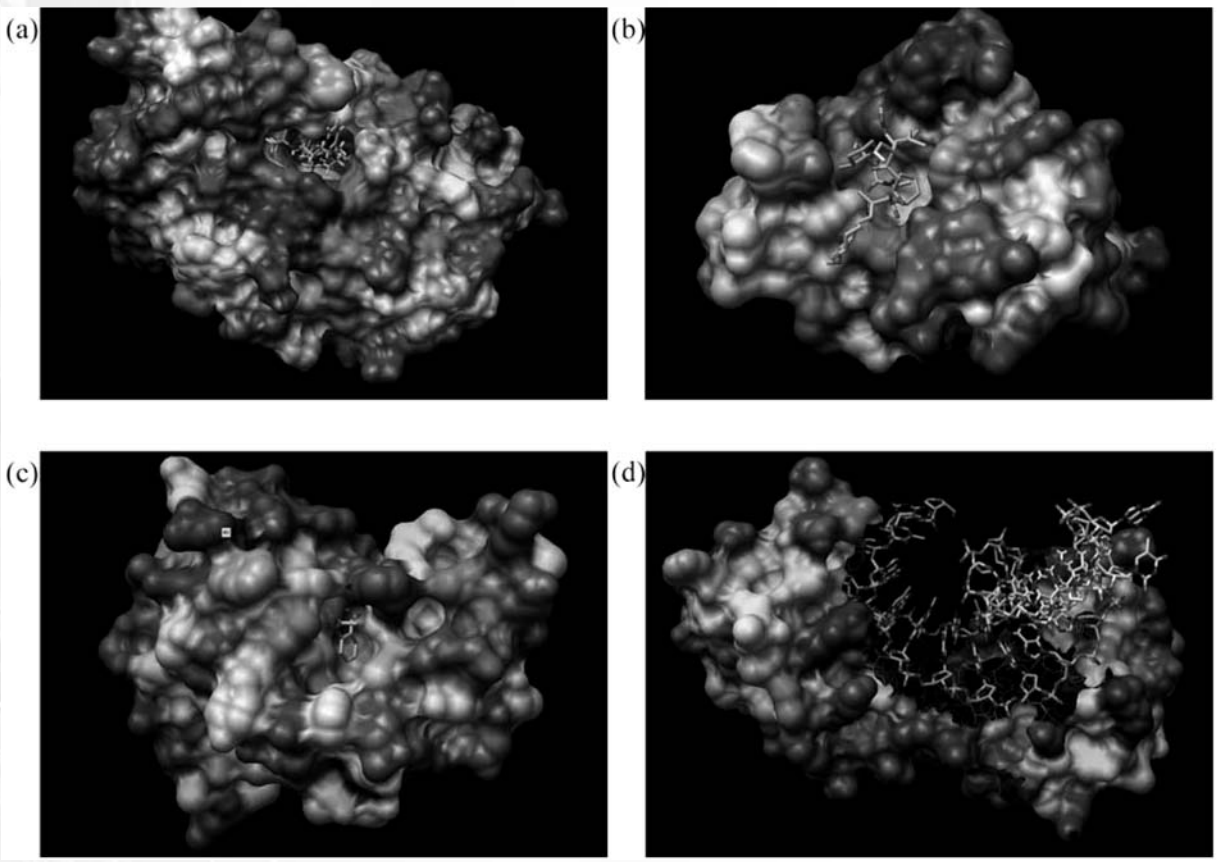
許多化學資料庫中所儲存的只是各個化學分子的結構式，有時也稱之爲二維表示式，這對我們眞正要從事藥物得虛擬篩選，常常是不太足夠的，我們需要各分子三維的構形。此外，大多數的分子都有可以旋轉的單鍵，因此可以存在許多不同的構形，也就是說，三維的構形並不是唯一的。所以我們常常需要用到其他程式，來產生各分子的三維構形以及可能存在的構形。這種二維至三維的轉換程式，著名的有CONCORD。此外，在虛擬篩選這個領域，許多程式都對檔案的格式有一定的選擇性，如果目前手邊的檔案格式並不適用於所要用的程式，則要找到轉換程式來更改檔案格式。一般來說，大部份的虛擬篩選程式都支援在美國St. Louis的Tripos公司所發展的mol2格式，如果分子的檔案是用他們所發展的Sybyl這個程式來準備，常常可以省去許多麻煩。當然，如果熟悉一些程式語言，如C、Perl、Fortran等等，也可以自行開發檔案轉換程式或甚至許多工具的整合平台。

分子嵌合的演算法與配體受體的結合模式(binding mode)的預測效率有著極爲密切的關係。目前有許多學術及商業軟體都可以做分子嵌合的預測。著名的程式包括有：Dock10, AutoDock[11], Ludi[12], GOLD[13], Glide[14], FlexX[15]/FlexE[16]/FlexS[17], ICM[18]等等。在這裡我們先稍爲介紹一下與我們後來開發的MEDock較爲相關的AutoDock。AutoDock是在美國聖地牙哥The Scripps Research Institute的Art Olsen的實驗室所開發出來的軟體，筆者所引的他們1998年的這篇文章，至今已被引用超過800次。這個程式所用的評分函數是利用經驗方法，先從Protein Databank中挑出30個ligand-receptor complexes，由這些complexes實驗上決定出的Ki值，換算出結合自由能值，再利用多變數線性迴歸 (multivariate linear regression) 來定出中間的一些待決定參數。因此其評分函數可以直接來估計某一對ligand receptor之間的結合強度 (binding affinity)，而不是只是給一個無物理化學意義的數字。其程式中允許使用者利用基因演算法 (genetic algorithm)、模擬退火法 (simulated annealing)、以及拉馬克基因演算法 (Larmarkian genetic algorithm) 來搜尋最佳的結合模式。根據他們在1998年的比較，利用拉馬克基因演算法的搜尋效率最高。不過一般來說，他們也發現在配體分子的可旋轉單鍵太多 (大於14個) 時，這幾種方法的效率都非常不理想。所以在做較大的配體分子的嵌合預測時，要考慮是否可將一些不必要旋轉的單鍵固定，這樣會較容易得到一致與收斂的結果。

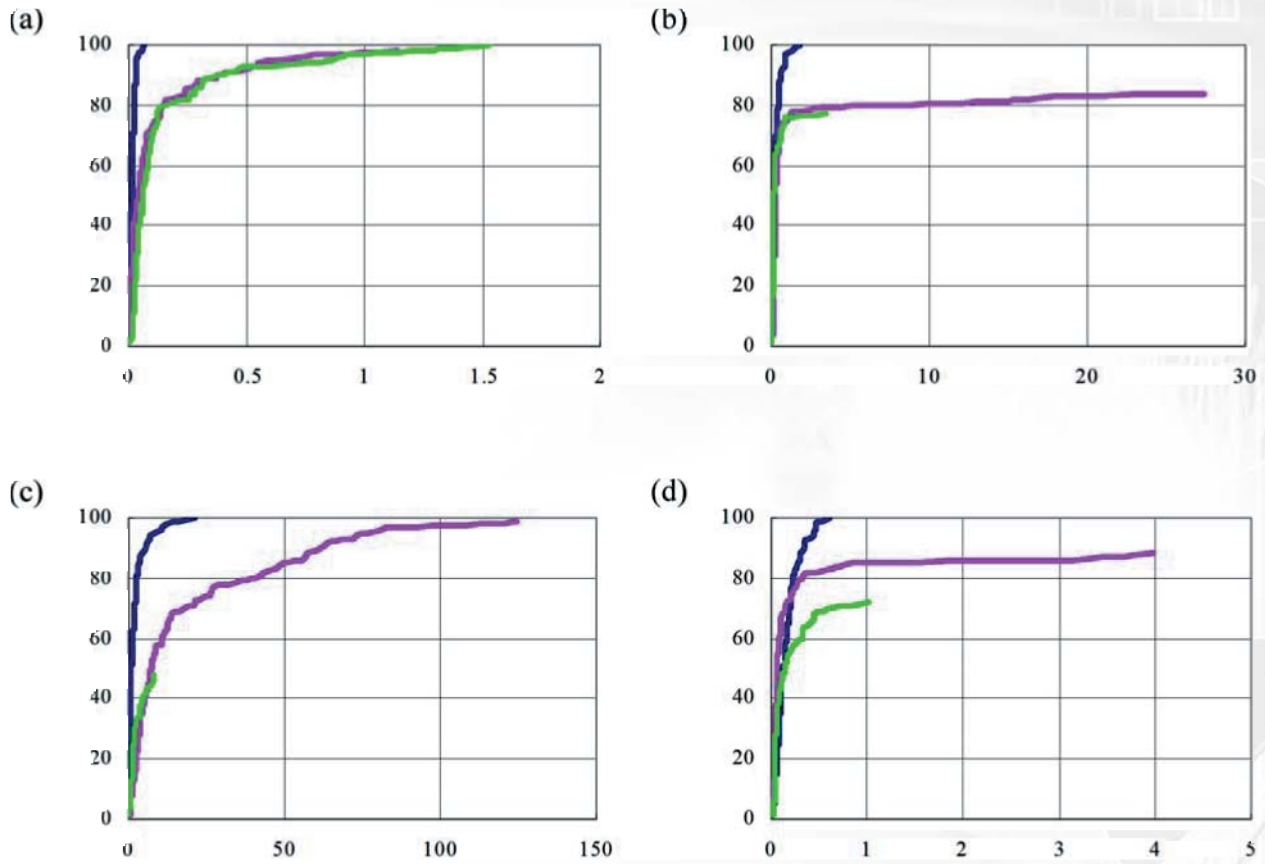
鑑於Docking這類方法在新藥開發中的重要性，因此筆者一直覺得非常值得投入並提升這類方法的效率與準確性，同時也應該讓許多wet lab實驗室的人員可以很容易地使用。回國後，剛好與本校資工系有些合作的機會，於是便提議可以合作開發的一個新的網站，也得到資工系歐陽彥正教授的支持，覺得這是提高台灣大學國際知名度很好的做法。這個網站後來就叫做MEDock19 (http://bioinfo.mc.ntu.edu.tw/medock)。

如上面所提到，設計這個網站的原因之一是由於大多數的分子嵌合的軟體都需要許多複雜的安裝及使用程序，有時候甚至需要使用者相當程度地熟稔UNIX作業系統；而這些都對生物醫學背景的人有時有點困難。目前可以用來預測分子嵌合結合模式的網站可說是非常罕見的，因此可以幫助更多的人來做這樣的預測。除了網站的建立，相當程度地簡化使用的複雜度之外，MEDock的核心，事實上是一個新發展的搜尋演算法，其用到了高斯函數 (Gaussian functions) 非常良好的性質，也就是它們在訊息理論 (information theory) 中在一定條件下，滿足所謂的熵最大化原理 (maximum entropy)。我們跟AutoDock的拉馬克基因演算法做了非常嚴格的比較，所用的系統如圖二所示。這些系統的配體大小差異非常大，同時受體結合位置的凹槽形狀也十分不同。這樣的系統挑選，是用來證明MEDock的搜尋演算法適用於很不同的情況。如圖三所示，MDock要比參數最佳化之後LGA更有效率與更可靠。

在圖三中，綠色線是AutoDock中使用LGA預設參數的結果，粉紅線是LGA的參數最佳化後的結果，藍色線是MEDock的結果。橫軸是所需的評分函數計算次數(單位是千萬次)。縱軸是在100次的測試中成功的次數。這個圖的解讀方式是，如果允許用到橫軸某個數量的評分函數計算，則100次中會成功搜尋到正確結果的次數就是曲線上對映到的點。從圖三可以看到，只要允許計算的時間足夠長，MEDock可以做到100次中有100次都可以成功地搜尋到正確結果。相對地，對LGA來說，其要做到100次中有100次都可以成功地搜尋到正確結果的評分函數計算次



圖二、四個運用MEDock所用來預測配體受體結合模式的系統。(a)是HIV-II protease與其抑劑 L-735,524. (PDB ID: 1HSH) (b) 是FKBP-FK506, an 各為immunophilin及immunosuppressant. (PDB ID: 1FKF) (c) 是Phospholipase A2 與aspirin. (PDB ID: 1OXR) (d) 是Tata-box binding protein (Ytbp) 與其DNA配體 Tata-box. (PDB ID: 1YTB)。



圖三、Autodock 的拉馬克基因演算法(LGA)與 MEDock 演算法的效能比較。(a)-(d)各是圖二中的四種情形。綠色線是 AutoDock 中使用 LGA 預設參數的結果，粉紅線是 LGA 的參數最佳化後的結果，藍色線是 MEDock 的結果。橫軸是所需的評分函數計算次數(單位是千萬次)。縱軸是在 100 次的測試中成功的次數。

數，都長許多，大約要 10 倍左右的時間，如(a)及(c)；有時候，即使允許給再長的計算時間，LGA 還不見得保証 100 次中有 100 次都可以成功地搜尋到正確結果，如(b)與(d)。這顯示了 MEDock 的演算法的效率遠遠地超越 LGA，並且其結果較爲可靠。

我們的 MEDock 網站，自 2005 年 7 月在 Nucleic Acids Research 發表以來，到現在 2006 年 3 月爲止，已經有超過 100 個各國的使用者，包括來自美國、德國、法國、瑞士、俄羅斯、澳大利亞、日本、印度等等國家，其中也包括像是美國國家衛生研究院(National Institute of Health, NIH)、美國食品藥物管制局(Food and Drug Administration, FDA)、法國國家科學研究院(Centre National de la Recherche Scientifique, CNRS)這些國家級的機構。

藥物虛擬篩選與新藥開發是一個高度整合性的工作，各個環節彼此之間緊緊相扣，而且不能在許多細節上疏忽，因爲越往下面的步驟去，所需要付出的時間、金錢與人力都會節節上升。在過去，一個新藥開發所需要花的時間常常是長達 10 到 15 年，而所投入的經費總共大約要數百億元，因此從事新藥開發的研究人員，無一不是憚精竭慮希望可以縮短新藥開發所需要花的時間，並增加可以眞正成功上市的機會。虛擬篩選的應用還只算是在起步的階段，其重要特徵之一是所需經費較低，找到新的有潛力的分子所需的經費，大約都在一般實驗室可以負擔範圍。目前最需要的便是更多成功運用的經驗，相信在不久的將來應該可以看到對新藥開發所需的時程的有

效縮短，並協助設計出更有效、更有專一性的藥物。

參考文獻

1. Ewing, T.J.A., Makino, S., Skillman, A.G. & Kuntz, I.D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *Journal Of Computer-Aided Molecular Design* **15**, 411-428 (2001).
2. Morris, G.M. et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal Of Computational Chemistry* **19**, 1639-1662 (1998).
3. Bohm, H.J. The Computer-Program Ludi - A New Method For The Denovo Design Of Enzyme-Inhibitors. *Journal Of Computer-Aided Molecular Design* **6**, 61-78 (1992).
4. Jones, G., Willett, P., Glen, R.C., Leach, A.R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking. *Journal Of Molecular Biology* **267**, 727-748 (1997).
5. Friesner, R.A. et al. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal Of Medicinal Chemistry* **47**, 1739-1749 (2004).
6. Rarey, M., Wefing, S. & Lengauer, T. Placement of medium-sized molecular fragments into active sites of proteins. *Journal Of Computer-Aided Molecular Design* **10**, 41-54 (1996).
7. Claussen, H., Buning, C., Rarey, M. & Lengauer, T. FlexE: Efficient molecular docking considering protein structure variations. *Journal Of Molecular Biology* **308**, 377-395 (2001).
8. Lemmen, C., Lengauer, T. & Klebe, G. FLEXS: A method for fast flexible ligand superposition. *Journal Of Medicinal Chemistry* **41**, 4502-4520 (1998).
9. Abagyan, R., Totrov, M. & Kuznetsov, D. Icm - A New Method For Protein Modeling And Design - Applications To Docking And Structure Prediction From The Distorted Native Conformation. *Journal Of Computational Chemistry* **15**, 488-506 (1994).
10. Chang, D.T.H., Oyang, Y.J. & Lin, J.H. MEDock: a web server for efficient prediction of ligand binding sites based on a novel optimization algorithm. *Nucleic Acids Research* **33**, W233-W238 (2005). ❖