利用蛋白質-配體交互作用與化合物結構為基礎之虛擬藥物篩選群集分析

學生：葛振寧　　　　　　　　　　　　　　　　　指導教授：楊進木 博士

國立交通大學生物資訊研究所碩士班

摘　要

我們發展了一個針對虛擬藥物篩選後處理(post analysis)的兩階段階層式分群分析法。此方法利用蛋白質-配體交互作用與化合物結構做為兩階段分析的主要原則。在第一階段，篩選出的候選化合物與目標蛋白質之蛋白質-配體三維結構與交互作用資訊將轉換成一維的實數表示，並採用階層式分群法針對候選化合物做第一階段的分群。在第二階段中，我們以 atom-pair 一維結構分析轉換法，淬取第一階段之分群的分子拓樸結構資訊。每一個經過交互作用分群後的群集將再進一步根據結構相似度做細分。兩階段（交互作用與藥物結構）階層式分群分析用在虛擬藥物篩選結果之組織化與視覺化分析，可以提升分析的速率與命中率，節省時間與經費，並且有助於未來實驗測試藥物的挑選與進一步分析。本方法以一組具有五種不同分子藥物目標的資料做驗證，包含胸腺嘧啶激酶(thymidine kinase)抑制劑，二氫葉酸還原酶(dihydrofolate reductase)抑制劑，雌激素受體(estrogen receptor)促進劑，雌激素受體抑制劑與神經胺酸酶(neuraminidase)抑制劑。經過在這些重要的分子藥物目標的分群分析測試後，本方法可以提供訂定分群界線之可能參考值，並能幫助研究人員有效的從虛擬篩選後產生的大量資料中找出具代表性的測試候選藥物，減少時間與金錢的花費。除了上述五個重要藥物目標之外，我們的方法也實際應用到幽門螺旋桿菌之莽草酸激酶(*Helicobacter pylori* shikimate kinase, HpSK)的抑制劑篩選分析。在對 CMC 藥物資料庫的虛擬藥物篩選後，我們由前 300 名的可能藥物分子中，經兩階段階層式分群分析後選出 23 種具代表性的藥物結構。經過合作實驗室的酵素抑制性測試後發現五個實際測試的結構中有一個具有莽草酸激酶之抑制性。此結果證明我們的方法不僅對虛擬藥物篩選與分析有效，並且確實有助於提升先導藥物開發流程的篩選速度與命中率。

Cluster analysis of Structure-based Virtual Screening by Using Protein-ligand Interactions and Compound Structures

Student : Cheng-Neng Ko
Adviser : Dr. Jinn-Moon Yang

Institute of Bioinformatics
National Chiao Tung University

ABSTRACT

We developed a cluster analysis method for post analysis of structure-based virtual screening. The analysis was composed of two stages based on protein-ligand interactions and compound structures, respectively. The first stage was to generate a protein-ligand interaction cluster by translating 3D structural binding information from a protein-ligand complex into a 1D real number representation, and using hierarchical clustering method to preliminarily cluster our screening results. In the second stage, we extracted molecular topology by atom-pair representation of each compound to re-grouping the clusters derived from the first stage. Each interaction cluster could be further divided into sub-clusters according to their topological similarities. The two-staged cluster analysis could be used to organize, analyze, and visualize the data of virtual screening and mining the representative candidates for future biological test. We validated this method on data sets having five classes: thymidine kinase inhibitors, dihydrofolate reductase inhibitors, estrogen receptor agonist, estrogen receptor antagonists and neuraminidase inhibitors. Our method on these pharmaceutical interest targets provided a suggestion of cluster threshold and helped to mining diversely representative structures from large number of virtual screening data. Our method also has been applied on the practical inhibitor screening analysis for *Helicobacter pylori* shikimate kinase (HpSK). After virtual screening in CMC database, we selected compounds from top 300 and selected 23 representative candidates. Five of 23 representative candidates were tested *in vivo*, and one of the five candidates, furosemide, was identified being able to inhibit HpSK by cooperated laboratory of Dr. Wen-Ching Wang.

Acknowledgements

我深深感謝我的指導教授, 楊進木教授, 他的建議總是那麼一針見血, 對於對的事情的執著, 還有看事物的洞察力, 讓我學習到很多做研究的方法. 另外要特別感謝陳彥甫學長, 因為有他大力的指導與幫忙, 我的論文才能如期完成. 還有感謝陳俊辰學長對於我的論文的修正提供了很多建議. 感謝陪伴我一起渡過漫漫長夜的夥伴們, 楊丁, 小強, 我會記住我們一起奮鬥的時光, 楊丁常常提供我寫程式上的技巧, 還有常常跟我抬槓的章維, 感謝藥物設計組的同學們的互相合作, 以及所有實驗室的同伴們. 還有我女友在這兩年的鼓勵與陪伴. 最後要感謝我母親長久以來無私的付出, 有您的付出我才能夠在學業上持續的努力.

CONTENTS

Abstract (in Chinese)... I

Abstract II

Acknowledgements III

Contents.. IV

List of Tables VI

List of Figures VII

Chapter 1. Introduction 01

1.1 Motivations and Purposes 01

1.2 Related Works 02

1.3 Application 03

Chapter 2. Materials and Methods 05

2.1 Preparation of the Target Protein and Ligand Database.......................... 06

2.2 Preparation of Virtual Screening Result for Cluster analysis 09

2.3 Generation of Descriptors 16

2.4 Reference Threshold for Protein-ligand Interaction and Atom-pair Descriptor 19

2.5 Method of Cluster analysis... 21

Chapter 3. Results 23

3.1 Molecular Recognition and Setting of Pharmacophore Consensus........ 24

3.2 Significance of Descriptor on Verifying Dataset 30

3.2.1 Significance of Protein-ligand Interaction Descriptor 30

3.2.2 Significance of Atom-pair Descriptor 31

3.3 Calculating Reference Threshold by Verifying Dataset........................ 32

3.4 Cluster analysis of Molecular Docking Result on Verifying Dataset 33

3.4.1 Cluster analysis of Molecular Docking on hDHFR (human dihydrofolate reductase) 33

3.4.2 Cluster analysis of Molecular Docking on TK (thymidine kinase) . 35

3.4.3 Cluster analysis of Molecular Docking on NA, ERα(3ert), and ERα(1gwr) 36

3.4.4 Cluster analysis of Compound Structures on Verifying Dataset...... 37

3.5 Cluster analysis of Virtual Screening Results on Testing Dataset 37

3.5.1 First Stage Cluster analysis on hDHFR Dataset 37

3.5.2 Second Stage Cluster analysis on hDHFR Dataset.......................... 39

Chapter 4. Applications: Using Two Stages Cluster Method for Post-analysis on the Results of Virtual Screening of Shikimate Kinase of *Helicobacter pylori*. 41

4.1 Preparations of the Target Protein and Compound set........................... 41

4.2 Molecular Recognition and Setting of Pharmacophore Consensus on the Shikimate Kinase.. 42

4.3 Virtual Screening for the Shikimate Kinase.. 43

4.4 Two Stage Cluster analysis of Result of Virtual Screening for Selecting Representative Compounds... 43

Chapter 5. Conclusions 45

5.1 Major Contributions and Future Works... 45

References 96

1896 E S

List of Tables

Table 1. Ten atom types used on atom pair descriptors 46

Table 2. Atom Types of GEMDOCK. 47

Table 3. Atom Formal Charge of GEMDOCK. 47

Table 4. Parameters Used for Docking on Verify Dataset. 47

Table 5. Parameters Used for Docking on Testing Dataset. 48

Table 6. The Ligand Preferences Calculated from Known Active Compounds Used for Virtual Screening on TK, ER, hDHFR, NA, and HpSK. 49

Table 7. The Pharmacophore consensus calculated by Superimposing Known Active Compounds Used for Molecular Docking on TK, ER, hDHFR, NA, and HpSK. 52

Table 8. The RMSD between docked poses and crystal ligands. 53

Table 9. T-test of distance between similar and non-similar binding mode generated by converting the docked pose into protein-ligand interaction profile (α=0.01). 54

Table 10. T-test of distance between similar and non-similar structure generated by atom-pair representation (α=0.01) 55

Table 11. T-test of distance between similar and non-similar compounds on each target protein. Descriptor was generated by converting the docked pose into protein-ligand interaction profile (α=0.01).. 56

List of Figures

Figure 1. Main steps of our two-stage cluster method for analyzing the result of virtual screening. **57**

Figure 2. Overall process of the two stages hierarchical clustering analysis.. .... **58**

Figure 3. The linear energy function of the pair-wise atoms for the steric interactions and hydrogen bonds in GEMDOCK (bold line) with a standard Lennard-Jones potential (light line). **59**

Figure 4. Definition of atom-pair representation. **60**

Figure 5. Definition of protein-ligand interaction. **61**

Figure 6. Ten TK (thymidine kinase) active compound structures. **62**

Figure 7 . Eleven ERα (estrogen receptor) antagonist structures. **63**

Figure 8. Ten ERα (estrogen receptor) agonist structures. **64**

Figure 9. Ten hDHFR (human dihydrofolate reductase) active compound structures. **65**

Figure 10. Twenty NA (neuraminidase) active compound structures. **66**

Figure 11. Overlapping 10 X-ray ligand structures on TK (1kim). Four important residues of the pharmacological consensus were identified and marked. The dash lines indicated the hydrogen binding. The phenolic ring of Y172 formed π - π stacking with the ligands. **67**

Figure 12. Overlapping 11 docked poses of ERα antagonists on ERα (3ert). Five important residues of the pharmacological consensus were identified and marked. The dash lines indicated the hydrogen binding. **68**

Figure 13. Overlapping 10 docked poses of ERα agonists on ERα (1gwr). Three important residues of the pharmacological consensus were identified and marked. The dash lines indicated the hydrogen binding. **69**

Figure 14. Overlapping 10 X-ray structures of ligands of hDHFR (1hfr). Four important residues of the pharmacological consensus were identified and marked. The dash lines indicate the hydrogen binding. 70

Figure 15. Overlapping 20 crystal structures of ligands of NA (1mwe). Four important residues of the pharmacological consensus were identified and marked. The dash lines indicated the hydrogen binding. 71

Figure 16. The result of molecular recognition on (a)hDHFR, (b)TK, (c)NA, (d)ER α (1gwr), (e) ER α (3ert) 72

Figure 17. The result of designing a reference threshold of protein-ligand interaction and atom-pair descriptors. 73

Figure 18. (a) Overlay of all 61 docked poses of known active compounds in the vicinity of the target protein hDHFR (PDB id: 1hfr). (b) Hierarchical clustering of protein-ligand interaction of 61 docked poses on hDHFR (PDB id: 1hfr). Each docked pose is represented as one line in the heat map in the middle of the figure, and the green being the lowest protein-ligand interaction energy and the red being the highest energy. The left side of the heat map shows the hierarchical clustering results on the hDHFR, including the dendrogram. Docked poses in the heat map are rearranged according to the order given by hierarchical clustering marked by the black bar 'c' in the right side of the heat map. The amino acids identified for description were also shown in the top side of the heat map. (c) Overlay of docked poses of the cluster with most number of known active compounds, and shown the important interaction between protein and ligand. (d)(e) Overlay of docked poses of the cluster with most number of unknown compounds, and shown the important interaction between protein and ligand. (f)(g) Overlay of docked poses of the sub-cluster within hDHFR active compounds, and shown the important interaction between protein and ligand. The blue frames in the heat map were the major interaction that different between cluster f and g....75

Figure 19. The detail of difference of binding interactions between new drugs and old drugs of hDHFR on verifying dataset. 76

Figure 20. (a) Overlay of all 53 docked poses of known active compounds in the vicinity of the target protein TK (PDB id: 1kim). (b) Hierarchical clustering of protein-ligand interaction of 53 docked poses on TK (PDB id: 1kim). Each docked pose is represented as one line in the heat map in the middle of the figure, and the red being the lowest protein-ligand interaction energy and the green being the highest energy. The left side of the heat map shows the hierarchical clustering results on the TK, including the dendrogram. Docked poses in the heat map are rearranged according to the order given by hierarchical clustering marked by the black bar ‘c’ in the right side of the heat map .The hot spots identified from overlapping known active compounds were also shown in the top side of the heat map. (c) Overlay of docked poses of the cluster with most number of known active compounds, and shown the important hydrogen bonds between protein and ligand. (d) Overlay of docked poses of the cluster with most number of unknown compounds, and shown the important hydrogen bonds between protein and ligand. The blue frames in the heat map were the major interaction that different between cluster c and d. **78**

Figure 21. Hierarchical clustering of protein-ligand interaction of 61 docked poses on NA (PDB id: 1mwe). **79**

Figure 22. Hierarchical clustering of protein-ligand interaction of 61 docked poses on ERα (PDB id: 3ert). **81**

Figure 23. Hierarchical clustering of protein-ligand interaction of 61 docked poses on ERα (PDB id: 1gwr). **82**

Figure 24. (d) The dendrogram of Hierarchical clustering of 61 known compound structures. The descriptor was calculated by atom-pair representation, using tanimoto coefficient for measuring distance between two molecules. Under the reference threshold (tanimoto coefficient = 0.55), There were three major clusters, a, b, and c. (a) In the cluster a, all 10 ERα agonists were grouped within the cluster. (b) In the cluster b, all 11 ERα antagonists were also grouped within the cluster. (c) In the cluster c, all 10 TK inhibitors and 14 NA were grouped together because the structures between TK and NA inhibitors were similar. By the observation on these three clusters, we could inspect that the atom-pair descriptor could group compounds with similar structures and divided compounds with different structures. 83

Figure 25. The first stage cluster analysis of hDHFR dataset. The compound set was combined 990 random selected compounds from ACD and 10 hDHFR inhibitors. 84

Figure 26. The detail of binding interactions within the largest cluster on hDHFR testing dataset. 85

Figure 27. The process and result of second stage cluster analysis on hDHFR testing dataset. 86

Figure 28. (a) The shikimate pathway. Our target protein was shikimate kinase. (b) The structure of 3-dehydro-shikimate. (c) The structure of shikimate, substrate of shikimate kinase. (d) The shikimate-3-phosphate, product of shikimate kinase...87

Figure 29. Comparing open and close form of shikimate kinase (open form PDB id: 1ZHU, close form PDB id: 1ZHI). (a) The induced-fit movement of the lid structure from an open to a closed form. (b)(c) The electrostatic surface of open and close form of shikimate kinase.88

Figure 30. The result of molecular recognition of shikimate on HpSK (PDB id: 1zui).......89

Figure 31. The distribution of fitness of compounds while screening on shikimate kinase. 90

Figure 32. Result of clustering top 300 rankers by first stage clustering on shikimate kinase. 91

Figure 33. Overall process and result of two-stage clustering on shikimate kinase............92

Figure 34. The structures of the 23 representative candidates on the HpSK.......................95