

# 國立交通大學

## 資訊管理研究所 博士論文

A Study in Mining and Post Screening Methods for Compounds  
Used in Various Biochemical Applications



資料探勘與篩選後分析方法於  
多方面生化應用化合物之研究

研究生： Daniel L. Clinciu

指導教授： 羅濟群 教授 楊進木 教授

中華民國一百年六月

A Study in Mining and Post Screening Methods for Compounds Used in Various  
Biochemical Applications

資料探勘與篩選後分析方法於  
多方面生化應用化合物之研究

Student: Daniel L. Clinciu

Advisor: Dr. Chi-Chun Lo

Co-Advisor: Dr. Jinn-Moon Yang

研究生：Daniel L. Clinciu

指導教授：羅濟群教授 楊進木教授



A Thesis Submitted to the Institute of Information Management at  
National Chiao Tung University in partial Fulfillment of the Requirements  
for the Degree of PhD in Information Management

June, 2011

Hsinchu, Taiwan (Republic of China)

# **A Study in Mining and Post Screening Methods for Compounds Used in Various Biochemical Applications**

Student: Daniel L. Clinciu

Advisor: Dr. Chi-Chun Lo

Co-advisor: Dr. Jinn-Moon Yang

Institute of Information Management, Institute of Bioinformatics

## **Abstract**

A phenomenal increase in the quality of human life is due to tremendous advancements and use of computer-aided methods in medicine and various biotechnological applications. Such technologies rely on the increasing availability of biochemical data and structural information which are highly significant for current advances. The solved crystal structures of 3D compounds stored in databases contribute greatly in bioinformatics as they are employed in studies and development of numerous lead compounds used in drug design and other industrial applications. However, screening and retrieving compounds for various applications presents a challenge for in retrieving and analyzing prospect targets. Therefore, a constant improvement of methods and tools is necessary for the proper classification, query, retrieval and analysis of available compounds data. With advances in computer technology, information management and data mining the developments of accurate, rapid and efficient algorithms enable studies in biotechnology to have significant improvements. However, mining appropriate candidates for various purposes by virtually screening thousands of docked protein-compound complexes is one of the biggest challenges. One of the main issues in virtual screening comes from an insufficient description of ligand binding mechanisms which results in the development of imprecise scoring functions.

In aiming to provide solutions to this issue we studied various docking algorithms and post screening methods used in mining and investigating specific compounds. Comparing different virtual screening and post screening analyses we observed that interaction profiles (e.g. van der Waals, hydrogen bonding) are highly relevant in the overall performance of compound mining. Moreover, this study concluded that a method which uses two combined stages of cluster analysis can be more efficient than one-stage clustering methods in selecting appropriate candidates for drug design and other biotechnological applications. Our study of interaction profiles also provided evidence of the possibility of mining novel compounds for potential uses in cosmetics, industry and agriculture in addition to pharmaceuticals using similar virtual screening and post screening analysis.

The above findings and observations contributed to the development of our method, Two Stage Combinative Clustering (TSCC) where we combine virtual screening and two stages of cluster analyses (interaction and physico-chemical). The methodology of TSCC has contributed to combinatorial computation approaches used to identify tetracycline derivatives for inhibiting Dengue virus neuraminidases and inhibitors for flaviviruses.

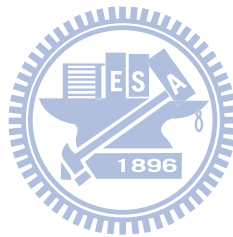
TSCC, similar to other post screening analysis methods starts with the virtual screening of compounds obtained from various databases e.g., Available Chemical Directory (ACD) or Comprehensive Medical Chemistry (CMC) using GEMDOCK. Top ranking compounds are then clustered based on their protein-ligand binding interactions and grouped into clusters with distinct binding interactions. Compounds are also clustered based on physico-chemical features using atom composition and are grouped in similar structure clusters. Compounds with lowest energy from each interaction cluster are selected as representatives while active compounds and similar to active compounds are chosen as representatives from each structure cluster. Lastly, final representatives from both interaction and structure clustering are chosen based on energy and structure similarity respectively and can be verified through bioassays for proper function and application. TSCC's novel feature is the use of two clustering stages to better filter and accurately retrieve the final representative compounds. Another key feature is to represent interactions at the atomic-level for including measures of interactions strength, enabling better descriptions of protein-ligand interactions to achieve a more specific analysis of virtual screening. The proposed two-stage clustering method enhanced our post-screening analysis by revealing more accurate performances than a one-stage clustering in visualizing and mining compound candidates and improving the virtual screening enrichment while being used successfully to identify novel inhibitors and functions of some proteins.

**Keywords:** cluster analysis, data mining, docking, GEMDOCK, lead compound, post screening analysis protein-ligand interaction profiles, target, compound database, virtual screening.



## **Acknowledgement**

I want to thank God for all the great things and for this opportunity I was given to study at such a great and prestigious university. Special thanks go to Dr. Jinn-Moon Yang, Dr. Chi-Chun Lo and Dr. Simon J. T. Mao for their great help, direction and support in this research and my overall study. I'm also thankful to my colleagues in the Bioinformatics laboratory, especially Marcco C.N. Ko, Piki, C.W. Huang and Shen-Rong. I'm also thankful to the Ministry of Education in Taiwan for making possible international programs at universities in Taiwan where people from around the world can come and study and in the same time contribute to the wonderful world of science and research so that the quality of human life and all important aspects of humanity can benefit. Lastly, I want to emphasize that what we do for others is the most important thing because together with others our contribution and investment in the future of this world will make it a better place for us and future generations of humans.



## Contents

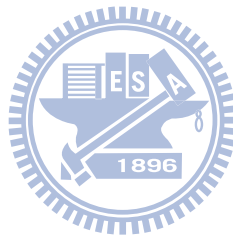
Cover page.....	i
Abstract .....	iii
Acknowledgement.....	v
List of Figures .....	vii
List of Tables .....	ix
<b>Chapter 1. Introduction .....</b>	<b>1</b>
1.1 Background .....	1
1.2 Motivation .....	4
1.3 Organization of Thesis .....	5
<b>Chapter 2. Related Studies and Review of Related Methods .....</b>	<b>7</b>
2.1 The Emergence of Post Screening Analysis.....	7
2.1.1 Interaction-Based Accuracy Classification (IBAC).....	8
2.1.2 Structural Interaction Fingerprint (SIFt).....	9
2.1.3 Visualized Cluster Analysis of Protein-Ligand Interaction.....	9
2.1.4 A New Hierarchical Clustering Approach for Large Compound Libraries..	10
2.2 The Use of Protein-Ligand Interaction Profiles in the discovery of Molecular Mechanisms and Lead Compounds .....	12
<b>CHAPTER 3. The Relevance of Protein-Ligand Interaction Profiles in Computer-Aided         Lead Compound Discovery, Functions and Applications.....</b>	<b>13</b>
3.1 INTRODUCTION.....	13
3.2 The Significance of Protein-Ligand Interaction Profiles in Methods of Retrieval and Analysis of Compounds.....	15

3.2.1	Post Screening Analysis.....	16
3.2.2	SIFt (Structural Interaction Fingerprint) .....	17
3.2.3	VISCANA (Visualized Cluster Analysis of Protein-Ligand Interaction).....	19
3.2.4	iGEMDOCK: A Graphical Environment for Recognizing Pharmacological Interactions and Virtual Screening.....	20
3.3	Summary.....	22

## **CHAPTER 4. TSCC: A Two- Stage Combinative Clustering for Virtual Screening**

	<b>Using Protein-Ligand Interactions and Physico-Chemical Features.....</b>	<b>24</b>
4.1	Introduction.....	24
4.2	Materials and Methods.....	26
	4.2.1 Preparation of Target Protein and Compound Databases.....	28
	4.2.2 Preparation of Virtual Screening Result for Cluster Analysis.....	29
	4.2.3 Testing and Verifying Datasets.....	30
4.3	Results.....	33
4.4	Verifying the TSCC method using $\beta$ -lactoglobulin .....	44
	4.4.1 Introduction .....	44
	4.4.2 Materials and methods .....	45
	4.4.3 Molecular Docking and Post Screening Analysis.....	46
4.5	Results.....	46
	4.5.1 Virtual Screening results.....	46
	4.5.2 Cluster Analysis Results.....	49
4.6	Discussion.....	52

4.7 Summary.....	53
<b>CHAPTER 5. Conclusion.....</b>	<b>55</b>
5.1 Summary.....	55
5.2 Future Works.....	56
<b>APPENDIX A.....</b>	<b>57</b>
<b>APPENDIX B.....</b>	<b>61</b>
<b>REFERENCES.....</b>	<b>62</b>





## List of Figures

Figure 1. Crystal structure of $\beta$ -lactoglobulin complexed with vitamin D-3. ....	2
Figure 2. The overall research process in investigating of interaction profiles and their role in identifying suitable methods for lead compounds retrieval and their applications. ....	6
Figure 3. The biased compound ranking in virtual screening (molecular docking). Ergocalciferol (purple color) has a lower binding energy than Riboflavin (yellow) and Celestine blue (blue) and it is ranked higher by the docking program. However, riboflavin is the active compound known for binding the to cavity of the target protein. ....	8
Figure 4. View of protein-ligand binding interactions in Betalactoglobulin .....	13
Figure 5. SIFt, VISCANA and <i>i</i> GEMDOCK; three post screening analyses based on binding interactions were investigated in our study of interaction profiles and in the designing of Two Stage Combinative Clustering (TSCC). ....	17
Figure 6. The concept of SIFT. 3D binding site of protein with an inhibitor (ligand) revealed as a sequence of positions in the binding site in contact with the ligand and their location in the structure of the protein (loop and $\beta$ ). ....	18
Figure 7. a) The overall approach of VISCANA (from VS to the selection of representatives)...19	
Figure 8. The virtual screening and post screening analysis processes in <i>i</i> GEMDOCK.....	21
Figure 9. The linear energy functions of the pairwise atoms for the steric interactions and Hydrogen bonds in GEMDOCK (bold line) with a standard Lennard-Jones potential (thin line). Figure obtained from a previous study by Yang <i>et al.</i> [26]. ....	25
Figure 10. Overall process of TSCC in our first study (a) First stage clustering using P-L interactions generated via GEMDOCK. (b) Second stage clustering of first stage results using physico-chemical features. (Figure obtained from our previous study [48]). ....	26
Figure 11. Designing a reference threshold of P-L interaction and atom-pair descriptors. ....	27
Figure 12. Cluster analysis of hDHFR. ....	39
Figure 13. (a) Overlay of all 53 docked poses of known active compounds in the vicinity of the target protein Thimidine Kinase (PDB id: 1kim). (b) Hierarchical clustering of 53 TK docked poses' protein-ligand interactions. Each docked pose is one line in the heat map, the red being the lowest P-L interaction energy and the green being the highest. The left side of the heat map shows the hierarchical clustering results of TK. The hot spots identified from known overlapping active compounds are shown at the top. (c) Overlay of docked poses of the cluster with most number of known active compounds and important h-bonds between protein and ligand. (d) Overlay of docked poses of the cluster with most number of unknown compounds and important	

h-bonds between protein and ligand. The blue frames in the heat map were the major interaction difference among clusters c and d. ....	40
Figure 14. The hierarchical clustering dendrogram for the 61 known compound structures showing the three major clusters. ....	41
Figure 15. The detail of hDHFR binding interactions of new drugs and old drugs on the verifying dataset. ....	42
Figure 16. The process and results of second stage cluster analysis on hDHFR testing dataset...	43
Figure 17. The overall approach of TSCC in our second study using <i>iGEMDOCK</i> and interaction clustering and atomic composition clustering for two-stage combinative clustering..	45
Figure 18. Active compounds used in the validation of the TSCC method.....	46
Figure 19. Conserved residues (LEU 39 and VAL 41) showing interaction through hydrogen bonding between $\beta$ -LG cavity and the three active compounds.....	48
Figure 20. The dendrogram showing the occurrence of important residues between the three active compounds and $\beta$ -LG cavity and also the top VS ranking compounds and $\beta$ -LG cavity...	49
Figure 21. Clustering analysis results. The ranking results from TSCC, two methods of one-stage clustering (Rank-IC and Rank-AC) and Virtual Screening for the three active compounds and four unknowns are shown in the four separate columns. ....	50
Figure 22. The three active compounds and four highest VS ranking unknown compounds; their structures, molecular weight and atom composition. Atom similarity: adenosine triphosphate and unknown mfc00013358 (brown circles), adenosine triphosphate and mfc00010114 (light blue circles), adenosine triphosphate and mfc00012401 (dark blue circles) and adenosine triphosphate and mfc00013358 (purple circles).....	51

## List of Tables

Table 1. Popular docking tools and evolutionary algorithms currently used in VS. ....	15
Table 2. The RMSD between docked poses and crystal ligands. ....	34
Table 3. T-test of distance between similar and non-similar binding mode generated by converting the docked pose into protein-ligand interaction profile ( $\alpha=0.01$ ). ....	35
Table 4. T-test of distance between similar and non-similar structure generated by atom-pair representation ( $\alpha=0.01$ ). ....	36
Table 5. 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 ( $\alpha=0.01$ ). ....	37
Table 6. Virtual screening results and ranking of the three active compounds (riboflavin: 575, adenosine triphosphate: 591 and calcitriol: 816). The shaded area (bottom of table) shows the highest ranking compounds (1 – 4) based on interaction energies generated by the docking program.....	47



# Chapter 1

## Introduction

### 1.1 Background

The transition of many preliminary biochemical studies from the wet to virtual laboratories propagated by computer-aided methods and an increase in technology has brought new insights and perspectives. Specifically, significant progress in development of novel compounds for pharmaceuticals, agriculture, cosmetics, nutrition and other industries has been mediated by computational techniques and approaches in preliminary steps. In this transition process many principles from other disciplines were adopted into the field of biotechnology. Applications of information management to aid with compound database management and of data mining to successfully retrieve and mine compounds from databases [1] are just a few examples of the constantly used, researched and developed applications in the field of biotechnology. Data mining, especially, has been given a lot of attention lately because of the rapid increase in number of virtual compounds available in databases. Mining of compounds from databases involves a series of steps but nowadays it can be done much faster and easier using combined methods of virtual screening and post screening analysis. Virtual screening (VS) [2, 3] is the first step towards the retrieval of prospect compounds. It is important to note that in a virtual setting the key to research and studies of biochemical compounds is the relevance of their crystal structures [4, 5] for practical applications in preliminary results which will be further confirmed by bioassays [6 – 10]. A crystal structure is composed of a pattern, a set of atoms arranged in a particular way (Fig. 1a) and a lattice exhibiting long-range order and symmetry (Fig. 1b). Patterns are located upon the points of a lattice (Fig. 1b), which is an array of points repeating periodically in three dimensions. The points can be thought of as forming identical tiny boxes, called unit cells, that fill the space of the lattice. The lengths of the edges of a unit cell and the angles between them are called the lattice parameters. The symmetrical properties of the crystal are embodied in its space group. The crystal structure of a compound and its symmetry play a role in determining many of its physical properties, such as cleavage, electronic band structure, and optical transparency. Various computer generated tools and programs are developed to “visualize and interpret” the characteristics of crystal structures and their

interaction with other crystal structures in specific studies of protein-protein complexes or protein-ligand complexes.

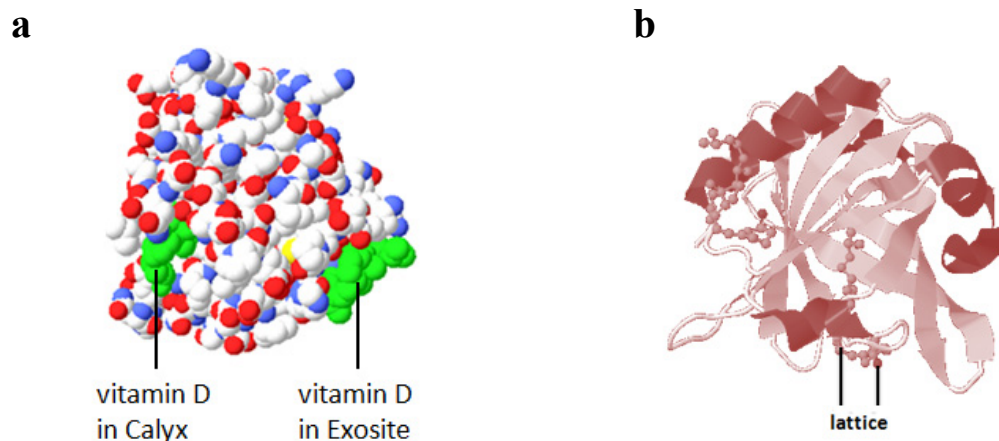


Figure 1. Crystal structure of  $\beta$ -lactoglobulin ( $\beta$ -LG) complexed with vitamin D-3. a) Space-filling model showing specific ligand binding sites (calyx and exosite) and b) Ribbon-lattice crystal structure of  $\beta$ -lactoglobulin. Crystal structures of compounds can provide many clues of binding sites and interactions between various proteins and or ligands.

Virtual screening of molecular libraries to mine compounds with an available crystal structures has emerged as a practical and inexpensive method in the discovery of novel lead compounds especially for drug design and discovery. This current increase in use of VS accounts for the following valid reasons: its enrichment and speed, the reduced cost and time of studies when using VS, increasing numbers of compounds with crystal structures and the advent of structural proteomics technologies. Computational techniques in VS involve two essential elements: efficient molecular docking (a technique to predict the preferred orientation of one molecule to a second when bound to each other to form a stable complex) and a reliable scoring method [11]. VS scoring methods must discriminate between non-native docked conformations and correct binding states of compounds during molecular docking phase to distinguish active compounds (usually a small number) from non-active compounds (an extremely large number) during the post-docking analysis. Scoring methods use three main classes of scoring functions that calculate the free binding energy: knowledge-based [12], physics-based [13] and empirical-based [14] scoring functions.

Inconsistencies in performance of scoring functions result in inadequate prediction of true binding affinity of a ligand to a receptor, thus, combining various scoring methods in VS may improve performance than in the average individual scoring functions. Similar inconsistencies have been noticed in information retrieval (IR) and Charifson *et al.* [15] proposed a study in which they used an interaction-based consensus approach to combine scoring functions which revealed enrichment in discrimination between active and inactive enzyme inhibitors. Later studies by Bissantz *et al.*, Stahl and Rarey and Verdonk *et al.* [3, 11, 16] showed consensus scores which further improved VS enrichment. Although researchers attempt to bring out the benefit of combining methods with consensus scoring, the remaining issue for VS users rather than researchers is when and how these scoring functions should be combined. Furthermore, certain VS methods can identify important interactions or binding-site hot spots obtained from known active ligands and target proteins [17]. Because most docking programs [18-20] use energy-based scoring methods which are often biased towards selection of high molecular weight compounds and charged polar compounds they have problems identifying key features (e.g. hot-spots) essential to target protein responses. Thus, methods for post-screening analysis employing clustering to identify key features through docked compounds and understanding binding mechanisms are of great use in bioinformatics. As VS encounters increasingly large databases, post screening analysis is an essential step in drug design and discovery.

The first attempt at a post screening analysis was done by Kroemer *et al* [21] in their work “Interactions-Based Accuracy Classification (IBAC)”, an approach which aimed to determine the best way to assess correctness of docking conformations. Their study showed that RMSD values alone are insufficient to predict correct poses; therefore, binding modes should be closely inspected for specific interactions when assessing pose prediction accuracy. Through this study the relevance of interaction profiles emerged as the basis in studies of interaction and bindings among protein-protein and protein-ligand complexes.

Amari *et al* and Deng *et al* [22, 23] followed the lead of IBAC and developed post screening analysis methods called Visualized Cluster Analysis of Protein–Ligand Interaction (VISCANA) and Structural Interaction Fingerprint (SIFt) respectively. Deng *et al* made a pioneering attempt by developing the first method based on binding interactions in order to facilitate the visualization, organization, analysis and data mining of virtually screened

compounds which all other post screening analysis employed. Amari *et al* devised a different approach for post screening analysis, a method based on the ab Initio Fragment Molecular Orbital Method (FMO) [24] to be used for analysis of virtual ligand screening also using the binding interactions generated from VS.

Attempts to cluster large numbers of compounds from VS by Bocker *et al* [25] resulted in a post screening method for clustering large datasets of compounds in a high dimensional space. The key feature of NIPALSTREE is its ability to handle more than 800 000 data points in high-dimensional descriptor space in less than an hour computation time.

The above studies implemented post screening analysis in an attempt to enrich the screening results of various docking tools (e.g. GOLD, AUTODOCK, GEMDOCK,) [19, 20, 26] and to facilitate the visualization, organization, analysis and data mining of virtually screened compounds. However, there are two main issues with all post screening analyses including the ones mentioned: 1) if a docking tool is used for VS, which post screening analysis should it be joined with for the most overall efficiency and accuracy and 2) if a post screening analysis method was decided (IBAC, SIFt or VISCANA) [21 – 23] which docking tool or VS method is most suitable prior to a particular post screening analysis. In addition, the ideal combination of docking tool and post screening analysis should successfully obtain novel compounds following the screening of compound databases and post analysis of selection lists obtained from VS.

This research investigates the potential drawbacks of VS and of various post screening analyses in computer aided novel compound mining and discovery. It further investigates the role of interaction profiles and proposes an algorithm that specifically optimizes a docking tool (GEMDOCK) for screening database compounds which is combined with a new, two-stage clustering method for post screening analysis in an attempt to join VS and post screening analysis for faster and more efficient compounds mining and analysis.

## 1.2 Motivation

The importance of efficient data mining and analysis of potential lead compounds to be used in various industries is of high relevance in biotechnology. Compounds can be obtained from databases through virtual screening and post screening analysis and contribute greatly in

many applications (novel compounds for drug design and industrial uses). The availability of compounds found in databases enable studies to be conducted at much cheaper costs and faster paces than previously done in “wet” or traditional laboratory settings where the use of natural compounds and live specimen was a concern for many reasons (proper disposal of hazardous materials and the constant need of live cell cultures and animals). Thus, in the virtual laboratory settings, the basis for investigating biologically active compounds is the use of high resolution x-ray structures of protein-ligand or protein-protein complexes from which a crystal structure is developed and to which all its known natural properties assigned. New functions and roles of existing compounds are always discovered; therefore compound databases must constantly be updated. Developments of high-throughput X-ray crystallography and advances in genomics [1-9] are constantly increasing the number of crystal structures available in protein databases [27, 28] leading to multiple therapeutic and industrial targets. Although the great number of available structures may present difficulties when retrieving compounds, the growing number of available methods aided by computer technology and principles from various disciplines (information management, data mining, consensus scoring) are rapidly evolving new and improved techniques to aid such studies.

This study investigates the significance of protein-ligand interaction profiles and compares various methods and tools used in virtual screening and post screening analysis for mining prospect compounds from databases and also expand their additional uses. It also shows the weakness of one-stage clustering methods in post screening analysis and why they have less success in identifying specific compounds and their various functions. Moreover, this study shows that a combined method of VS and a combined two-stage cluster analysis is more ideal for mining specific compounds and investigating their various functions.

### **1.3 Organization of Thesis**

This thesis is organized as follows: In chapter 2 we describe related studies and similar methods of mining and analyzing prospect compound candidates from virtual databases along with their advantages and shortcomings. In Chapter 3 we perform an in-depth study of protein-ligand interaction profiles and present novel concepts obtained from our investigations in possible future work for additional applications of virtual screening and post screening analysis such as cosmetics, nutrition, industry and agriculture. In chapter 4 we describe our core work, the



development of Two-Stage Combinative Clustering (TSCC) and its improvement over one-stage post screening analysis methods. Chapter 5 concludes our studies and includes future work prospects. In Figure 2 below, the model for this research is presented.

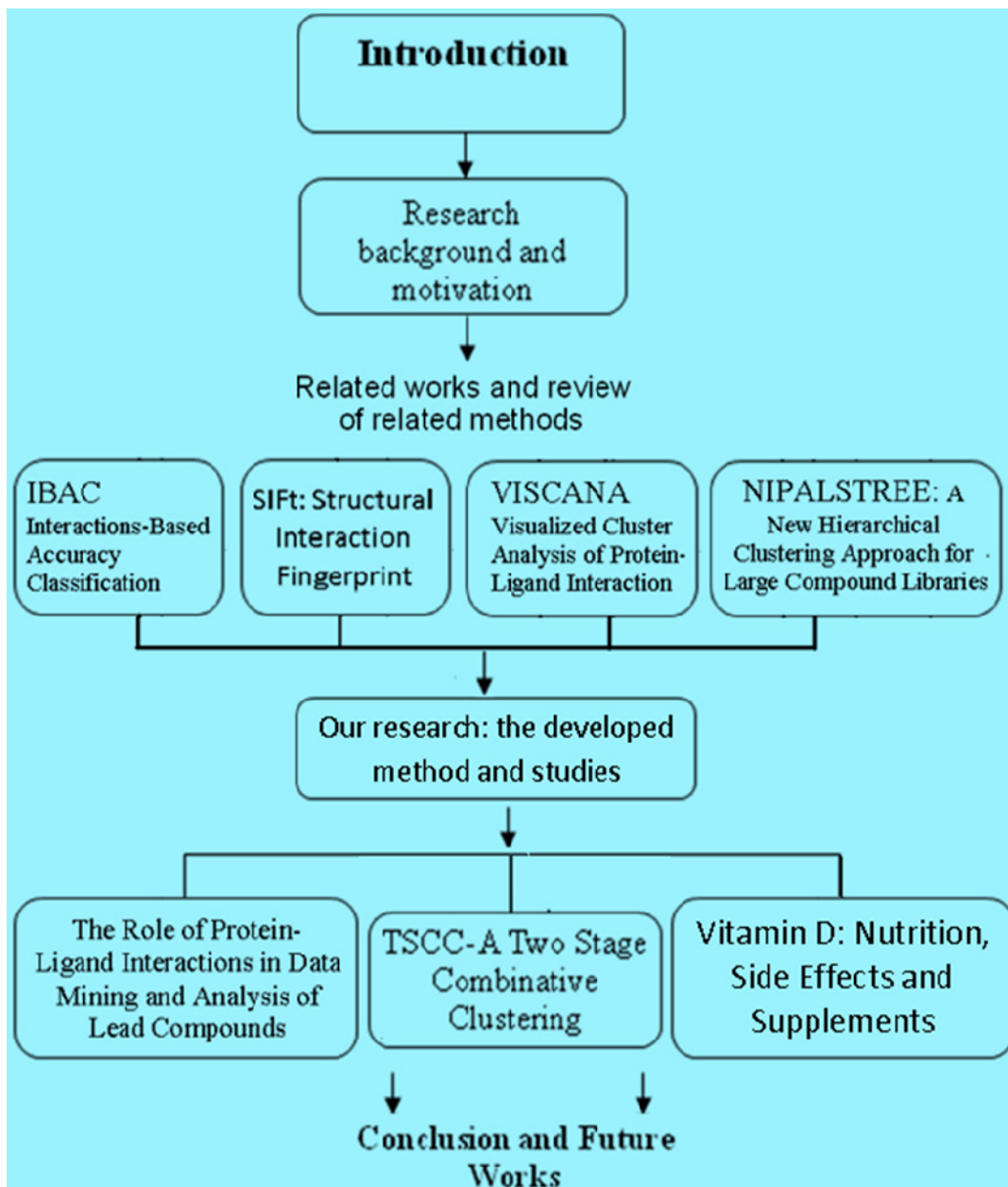


Figure 2. The overall research process in investigating of interaction profiles and their role in identifying suitable methods for lead compounds retrieval and their applications

## Chapter 2

### 2.1 Related Studies

The process of VS and post screening analysis is a common technique used in mining and analyzing compound candidates to be used in pharmaceuticals or various other applications after their retrieval from databases. The VS technique involves docking tools (e. g. DOCK, GEMDOCK or GOLD) [19, 20, 26] to screen compound databases and rank compounds according to their binding energies. Compound databases store solved crystal structures (Figure 1) of chemically significant compounds which can be used in various studies (e.g. drug design, nutrition and other industries) [6, 9, 10, 29 – 31]. VS and docking is followed by post analyses using clustering (SIFt and VISCANA) [22, 23] which aim to reduce the number of false positives obtained from VS and propagate true positives to the top of the selection list.

#### 2.1.1 The emergence of Post Screening Analysis

In the early days of computer-aided drug design, docking tools / programs were the only means of screening compounds for the possibility of drug design. Given the poor understanding of many critical factors at the time especially the incomplete knowledge of ligand binding mechanisms, VS was still a major accomplishment in moving forward a revolution in drug design and discovery with faster and more practical preliminary approaches than previously done through bioassays using biochemical methods. Traditional settings, in addition to requiring an extensive period of time to study various properties and make a drug ultimately available, had overwhelming expenses inquired through the use of conventional biochemical compounds, facilities and specimen. With the advent of computer aided drug design more of the preliminary work in drug design is done in virtual labs and when desired results are obtained, the stage requiring bioassays to confirm preliminary results is applied.

Most docking programs [19, 20, 26] use energy-based scoring methods which are often biased towards selection of high molecular weight compounds and charged polar compounds (Fig. 3). Therefore, they have problems identifying key features (e.g. hot-spots) essential to target protein responses resulting in the performance of these scoring functions to be mostly inconsistent when conducting a database search [3, 11]. The inaccuracy of various scoring methods inadequately predicting the true binding affinity of a ligand for a receptor is a major weakness for VS. Moreover, employing VS [2, 3] in computer-aided drug design usually results

in a high number of chosen compounds from which few are potential or suitable candidates. Thus, it is imperative that a post screening analysis is conducted in order to reduce the number of false positives in the selection lists generated from VS and to propagate true hits to the top of the selection lists.

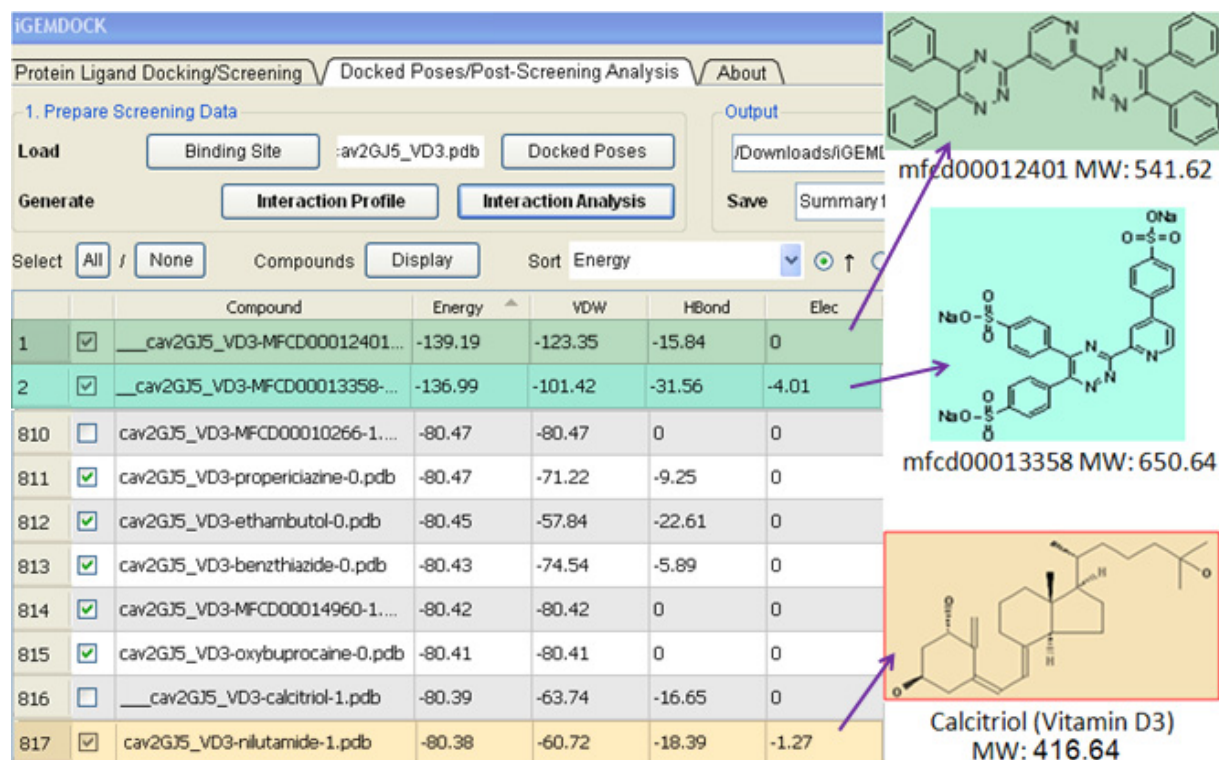


Figure 3. The biased ranking of compounds in virtual screening (molecular docking). Unknown compounds MFCD00012401 (green color) MFCD00013358 (teal green color) are ranked much higher than Vitamin D3 (ranked 816) due to their energy and molecular weight. However, only vitamin D3 is known for its ability to bind to the target protein ( $\beta$ -LG) [66, 67] among all compounds listed in this table.

### 2.1.2 Interaction-Based Accuracy Classification (IBAC)

IBAC is an approach developed by Kroemer *et al* [21] which determines the best way to assess correctness of docking conformations. It first calculates the RMS deviation of the predicted pose from the crystal structure and then it compares the predicted pose to the pose experimentally observed. In simple terms, using IBAC, Kroemer *et al* optimized the binding site definitions and docking protocols for 6 VS programs used in their studies (FlexX [32], GOLD [20], ICM [33], LigandFit [34], NWU [35, 36] and QXP [37]). They executed docking runs and

reported details of the ligand tautomeric forms and bond orders and how RMSDs from crystal structures correlated with interactions-based accuracy classifications. Kroemer *et al.* concluded that RMSD values alone lack the ability to predict correct poses and binding modes should be investigated further for specific interactions when assessing pose prediction accuracy. Through the work of Kroemer *et al.* the relevance of interaction profiles emerged as the foundation of interaction and bindings studies for protein-protein and protein-ligand complexes.

### 2.1.2 Structural Interaction Fingerprint (SIFt)

SIFt [23] uses a simple, generic and robust approach for representing and analyzing 3D protein-ligand interactions. Its key feature is the generation of an interaction fingerprint that converts 3D structural binding information into a one-dimensional (1D) binary string (Figure 9). The fingerprint representation of the interaction patterns is compact, and allows for rapid clustering and analysis of large numbers of complexes. The SIFt is calculated on a set of input 3D protein–small molecule complexes. To analyse SIFts the Tanimoto coefficient (Tc) [38] is used as the quantitative measure of bit string similarity.

This representation of interactions as fingerprints using the SIFt method enables clustering, filtering and profiling of large docking results libraries and crystal structures of the protein kinase family in complexes with various inhibitors. Although SIFt opened a broad road for post screening analysis, much of the road is still unpaved and difficult to travel in terms of methods used currently in post screening analysis.

### 2.1.3 Visualized Cluster Analysis of Protein-Ligand Interaction

VISCANA [15], a method which stands for Visualized Cluster Analysis of Protein-Ligand Interaction based on the ab Initio Fragment Molecular Orbital Method (FMO) [24] used for virtual ligand screening was proposed by Amari *et al.* They developed a cluster analysis using the dissimilarity defined as the squared Euclidean distance between interfragment interaction energies (IFIEs) of two ligands. In VISCANA a clustering method is combined with a graphical representation of the IFIEs by representing each data point with colors that quantitatively and qualitatively reflect the IFIEs. This method claims to classify structurally different ligands into functionally similar clusters according to the interaction pattern of a ligand and amino acid residues of a receptor protein. VISCANA also estimates the docking

conformation by analyzing patterns of the receptor-ligand interactions of some conformations through the docking calculations.

However, as stated by Amari *et al.* in their study, VISCANA lacks sufficient descriptions of van der Waals forces and hydrogen bond interactions which play an important role in receptor-ligand binding [39, 40]. This may account for selection of false positives instead and the failure to select true hits or active compounds. This method is aiming to increase VS enrichment; however, it doesn't provide significant improvements over SIFt or extend further uses into drug design and discovery or other possible applications.

#### **2.1.4 A New Hierarchical Clustering Approach for Large Compound Libraries: NIPALSTREE**

NIPALSTREE, is an approach by Bocker *et al* [25] for clustering large datasets of virtual compounds in a high dimensional space. It uses the first Principle Component (PC) which employs NIPALS (non-linear iterative least squares) where the data set is split at point  $i$  or  $j$  (determined points where two neighbors exceed a predefined distance threshold  $T$ ). The procedure is recursively applied on the resulting subsets until the maximal distance between cluster members exceeds a user-defined threshold. NIPALSTREE clustering employs PCA for hierarchical clustering algorithm as follows: A  $d$ -dimensional descriptor matrix is projected onto the first PC. Based on the scoring vector  $\mathbf{S}$ , the given descriptor matrix is sorted in ascending order and split at the median position, i.e., two equally large descriptor sets-from now on termed "left" and "right" submatrix  $s$  are created. This is repeated for the new subsets until the maximum distance between the entries in a submatrix underscores a predefined similarity threshold ( $\Theta$ ). In order to judge the quality of a clustering result an index is introduced to assess whether molecules interacting with the same target (receptor or receptor family) lie in the same subtree. An enrichment factor (EF) is calculated for each cluster, which gives an estimate of how well compounds that bind to the same target (or target class) are clustered in a dendrogram node  $i$  expressed in the following equation:

$$EF_{i,c} = \frac{\frac{N_{i,c}}{N_c}}{\frac{N_i}{N}}$$

$N_{i,c}$  being the number of entries in node  $i$  belonging to class  $c$ ,  $N_i$  being the total number of

entries in node  $i$ ,  $N_c$  being the total number of entries of class  $c$  in the data set, and  $N$  being the overall number of entries.  $EF > 1$  indicates that more compounds belonging to the activity class  $c$  are clustered in a tree node than expected from an equal distribution. The  $EF$  value depends on the size of the dendrogram section under consideration: On the upper dendrogram levels, where clusters are large,  $EF$  values are usually smaller, whereas  $EF$  values on the lower dendrogram level can get large without a statistical relevance. A possible way to overcome the cluster size dependency of the  $EF$  is to additionally divide it by the logarithm of the dendrogram level, assuming that at each cluster the data set is separated into equally large partitions. In this way, an adoption of the  $EF$  to the dendrogram level can be achieved.

Although NIPALSTREE is able to deal with more than 800 000 data points in high-dimensional descriptor space in less than an hour computation time it does not specify how false positives are addressed; this is a major concern for all methods performing compound retrieval and analysis. Besides a rapid clustering of compounds, NIPALSTREE cannot offer visualization and accurate data mining of compounds and it is impractical as a method of retrieval and analysis for specific compounds in either drug design or other industrial uses.

## **2.2. The Use of Protein-Ligand Interaction Profiles in the discovery of Molecular Mechanisms and Lead Compounds**

Since protein-ligand and protein-protein complexes are components of a great number of pharmaceutical [5, 41], nutritional [10] and industrial compounds [29-31] it is reasonable to employ computer-aided lead compound design and discovery methods for other applications besides pharmaceuticals. Due to its significant role and impact on the quality of human life, drug design was the main focus in early days of virtual screening and bioinformatics. However, as methods and studies in drug design reveal that VS and post screening analysis are relatively inexpensive and efficient we want to explore the other fields (nutrition, agriculture and industry) which were not given as much attention. Protein-ligand complexes of various compounds interact through similar properties [40] and necessitate similar methods of screening, retrieval and analysis of their crystal structures (Figure 1) regardless what their final application may be. Therefore, the first part of this research focuses to conduct comparative studies on features and properties of protein-ligand interaction profiles to better understand their relevance in the mining of novel compounds. Additionally, we investigate possibilities of employing interaction profiles in the mining of compounds to be used in other applications besides drug design such as

cosmetics, skin care, nutrition, safe fertilizers and pesticides, compounds for scents in perfumes and deodorants and safe detergents. Furthermore, we employ interaction profiles in investigating mechanisms of significant molecules for human health and nutrition (e.g. uptake of vitamin D in the human body by Betalactoglobulin).

Although the interest of researchers in mining novel compounds for other uses besides pharmaceuticals is minimal at the present time, as computer-aided methods continue to improve and increase in use, other industries (e.g. cosmetics, agriculture, nutrition) look to employ their benefits. Therefore, the approaches and techniques used in computer-aided drug design can be of particular interest for different biotechnological approaches. VS combined with post screening analysis are seemingly efficient in investigating transporter proteins such as  $\beta$ -lactoglobulin ( $\beta$ -LG), their mechanisms and various functions in the human body. Many compounds having various functions and mechanisms in the body are protein-ligand complexes which can be investigated based on protein-ligand interactions and physico-chemical features.





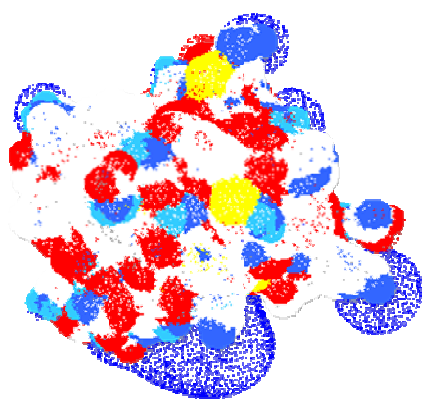
## CHAPTER 3

# The Relevance of Protein-Ligand Interaction Profiles in Computer-Aided Lead Compound Discovery, Functions and Applications

### 3.1 Introduction

Identification of protein-ligand interaction networks on a proteome scale is crucial in addressing a wide range of biological issues such as correlating molecular functions to physiological processes and designing safe and efficient target compounds which can be used in therapeutics, nutrition, cosmetics, skin care products, agriculture and industry. In order to understand the role and significance of protein-ligand interactions (Fig. 4) in various applications throughout the field of bioinformatics and biotechnology the properties and functions of a ligand [42, 43] must be well addressed. As seen previously, the ligand (vitamin D, Fig. 1) is a molecule, ion or atom which can bind to a specific location or the binding site of a protein [39, 44]. Currently, antibodies are the most commonly used ligands in biotechnology and life-science investigations, although protein scaffolds (protein regulators), nucleic acids and peptides (repeating structural units in amino acids) are also employed. Since protein-ligands complexes of various compounds are used in cosmetics, hair dyes, skin care products, fertilizers, detergents [29-31] and nutrition supplements [10], protein-ligand interaction profiles and physico-chemical features could be used in the identification of such lead compounds.

**a**



**b**

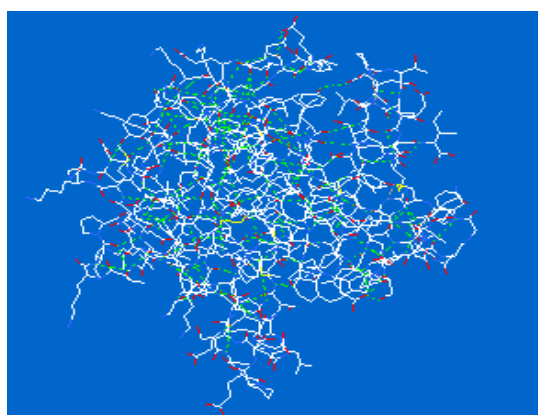


Figure 4. View of protein-ligand binding interactions in Betalactoglobulin (a transporter protein) complexed with vitamin D using Swiss PDB viewer. a) Electrostatic potential and molecular surface. b) Hydrogen bond interactions among atoms (green dotted lines).



The ligand binding site of the primary target is extracted or predicated from a 3D experimental structure or homology model of proteins [35, 45] and characterized by a geometric potential. Protein-ligand interactions occur when a ligand binds to a protein which is usually integral to the function of its cognate (assimilated or symbiotic) protein. In the binding of a ligand to a protein, the following interactions are of significance: electrostatic forces (interaction between electrically charged particles explained by Coulomb's law), van der Waals forces (the sum of the attractive or repulsive forces between molecules or parts of the same molecule) and hydrogen bonding (the attractive interaction of a hydrogen atom with an electronegative atom which can occur inter or intramolecularly) [39, 40]. Based on these interactions, evaluations are made using ligand-based approaches employed commonly in pharmacophore modeling by using physical and chemical traits of known ligands to identify novel inhibitors. Another approach, the receptor-based, identifies ligands that use structural and other features on the target receptor to identify the best inhibitor.

Docking [18, 26, 32, 33, 46] is then used to identify the fit between a receptor and the potential ligand by screening a database of ligands against one or more target receptors *via* two distinct parts: docking (the search scheme to identify suitable conformations or poses) and scoring (a measure of the affinity of various poses). Scoring methods must discriminate between non-native docked conformations and correct binding states of compounds during molecular docking phase to distinguish active compounds (usually a small number) from non-active compounds (an extremely large number) during the post-docking analysis. Although there are over 60 docking programs and tools available [24], we present some of the most popular programs made publicly available (Table 1). DOCK [18], incremental construction (FlexX) [32] and evolutionary algorithms (GEMDOCK, GOLD, AutoDock) [26, 33, 46] are used to screen and downsize compound groups in order to select suitable candidates for post-screening analysis. However, inconsistencies in the performance of scoring functions results in inadequate prediction of true binding affinity of a ligand to a receptor; thus, combining various scoring methods in VS may improve performance than in the average individual scoring functions. Similar inconsistencies have been noticed in information retrieval (IR) and Charifson *et al.* [15] proposed a study in which they used an interaction-based consensus approach to combine scoring functions which revealed enrichment in discrimination between active and inactive enzyme inhibitors. Studies by Bissantz *et al.* [3], Stahl and Rarey [11] and Verdonk *et al.* [16]

showed works on consensus scores which further improved VS enrichment. However, the remaining issue for VS users rather than researchers is when and how these scoring functions should be combined in either drug design or industrial compounds design.

Docking programs	URLs	REFERENCES
DOCK	<a href="http://dock.compbio.ucsf.edu/">http://dock.compbio.ucsf.edu/</a>	18
FlexX	<a href="http://biosolveit.de/flexx/index.html?ct=1">http://biosolveit.de/flexx/index.html?ct=1</a>	32
AutoDock	<a href="http://autodock.scripps.edu/">http://autodock.scripps.edu/</a>	46
GEMDOCK	<a href="http://gemdock.life.nctu.edu.tw/dock/igemdock.php">http://gemdock.life.nctu.edu.tw/dock/igemdock.php</a>	26
GOLD	<a href="http://www.ccdc.cam.ac.uk/products/life_sciences/gold/">http://www.ccdc.cam.ac.uk/products/life_sciences/gold/</a>	33

Table 1. Popular docking tools and evolutionary algorithms currently used in VS

Furthermore, certain VS methods can identify important interactions or binding-site hot spots obtained from known active ligands and target proteins [17]. However, due to biases towards higher molecular weight and charged polar compounds [18] docking alone is not sufficient to analyse, determine and retrieve the most adequate lead compounds therefore post screening analyses are emerging as useful methods to aid with further elimination of false positive hits obtained from VS.

Methods for post-screening analysis employing clustering to identify key features obtained *via* docked compounds and the understanding of binding mechanisms are of great use in bioinformatics. Therefore, computer-aided drug and industrial target design require VS as a primary step to generate interaction and structure profiles followed by post screening analysis for adequate filtering, visualization and mining of the final candidates.

### 3.2 The Significance of Protein-Ligand Interaction Profiles in Methods of Compound Retrieval and Post Screening Analysis

Interactions between molecules (Fig. 4) are important for understanding many biological phenomena. From gene expression to enzyme reactions, the activities are dictated by molecular interactions. Because of DNA microarray success, researchers are studying the protein counterpart in greater detail [47]. Protein microarray can be used for studying a variety of

biological phenomena such as interactions of protein–ligand, protein–protein, antibody–antigen, protein–DNA, analysis of subunits in protein complexes, screening of target proteins expressed from phage library, analysis of mutant proteins, quantitative assay, discovery of diagnostic markers, analysis of protein expression profiles, development of diagnostic microarray and development of microarray-based lead screening system. The interactions of significance in analysis and retrieval of lead compounds for drug design are intermolecular interactions such as van der Waals forces, electrostatic forces and Hydrogen bonds interactions [39, 40]. Also called interaction energies, they can be obtained from virtual screening of docked compounds calculations [13]. The calculations of interaction energies are organized into data sets of interaction profiles (IPFs) and can be used as one of the criteria in a cluster analysis to further filter out and select more specific or the final target compounds. Thus, cluster analysis of various compounds with similar interaction energies will group the various compounds into separate clusters from which a representative is chosen usually based on RMSD values while undergoing what is termed a post screening analysis.

### 3.2.1 Post Screening Analysis

Methods of post screening analysis [21-23] are designed to facilitate the visualization (interpretation of binding interaction), organization (cluster and organize structures in a meaningful way), analysis (compare and profile the binding interactions of different structures) and data mining (search for structures containing key interactions or specific features) of virtually screened compounds. As mentioned earlier, binding interactions [39] (e.g. van der Waals forces, electrostatic forces and hydrogen bond interactions) of protein–ligand complexes are a critical part of mining and selecting the target representatives in post analysis methods. Descriptions of binding interactions and interaction strength measures for protein–ligand complexes are very important for better mining of appropriate candidates from selection lists generated by VS [48]. Through an in-depth study of protein–ligand interactions in various post screening analysis, we attempt to develop an integrated method of VS and post screening analysis in order to speed up the screening and analysis of compounds, generate better interaction-specific information and to obtain suitable representatives. The overall details of this study are shown in Figure 5.



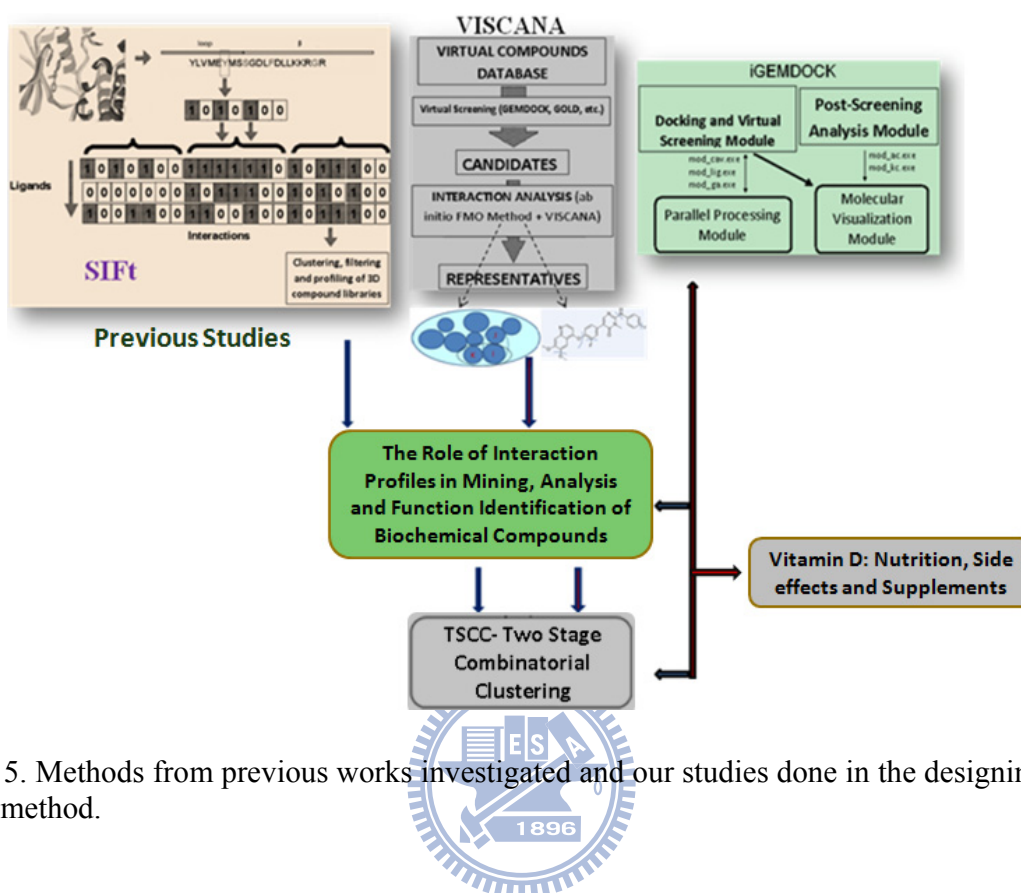


Figure 5. Methods from previous works investigated and our studies done in the designing of our TSCC method.

Bellow we investigate and compare a few pioneering methods of post screening analysis which were all originally designed to enrich virtual screening. Later in our work we will perform some comparative studies and inductive analysis which provide a foundation for expanding the use of virtual screening and post screening analysis into the mining and analysis of targets used in various other applications besides pharmaceuticals.

### 3.2.2 Structural Interaction Fingerprint (SIFt)

SIFt [23] uses a simple, generic and robust approach for representing and analyzing 3D protein- ligand interactions. Its key feature is the generation of an interaction fingerprint that converts 3D structural binding information into a one-dimensional (1D) binary string (Fig. 6). The fingerprint representation of the interaction patterns is compact, and allows for rapid clustering and analysis of large numbers of complexes. The SIFt is calculated on a set of input 3D protein–small molecule complexes. The protein structure may have been determined

experimentally by NMR or crystallography, or generated through homology modeling. The SIFt is generated by first defining the union of those residues that are in contact between the protein and the small molecule complex. The resulting panel of ligand binding site residues, which act as a mask covering all of the interactions occurring between the protein and the ligands, is then used as the common reference frame to construct the interaction fingerprints.

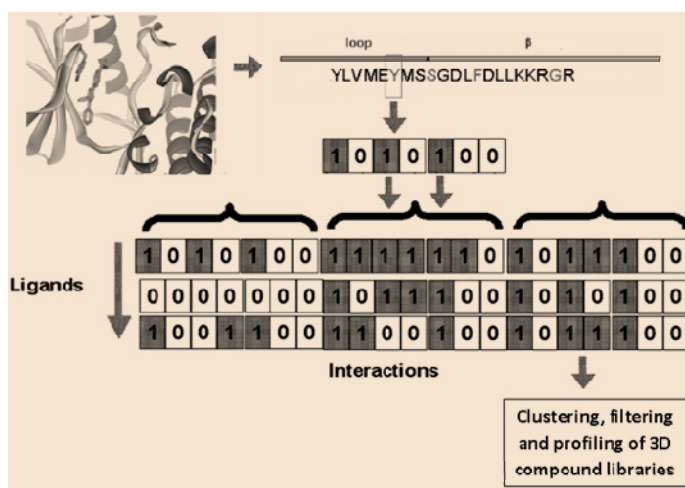


Figure 6. The 3D binding site of protein with an inhibitor (ligand) revealed as a sequence of positions in the binding site in contact with the ligand and their location in the structure of the protein (loop and  $\beta$ ). Each binding site position is represented by a bitstring. The joining of all bitstrings end-to-end for each binding site residue is repeated for all ligands and is used in the selection process.

To analyse SIFTs the Tanimoto coefficient (Tc) [38] is used as the quantitative measure of bit string similarity. The Tc between two bit strings A and B is defined as:

$$Tc(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

where  $|A \cap B|$  is the number of ON bits common in both A and B and  $|A \cup B|$  is the number of ON bits present in either A or B. Tanimoto coefficients between random bit strings with a length of 400 bits adopt a near-Gaussian distribution centered at approximately 0.33, with a sigma of about 0.03. This representation of interactions as fingerprints using the SIFt method enables clustering, filtering and profiling of large libraries of docking results as well as crystal structures of the protein kinase family in complexes with various inhibitors.

### 3.2.3 VISCANA (Visualized Cluster Analysis of Protein-Ligand Interaction)

VISCANA [22] (Fig. 7) is a method based on the ab Initio Fragment Molecular Orbital Method (FMO) [24] used for analysis of virtual ligand screening. The ab initio FMO method at the Hartree-Fock level is shown in the details following the method figure.

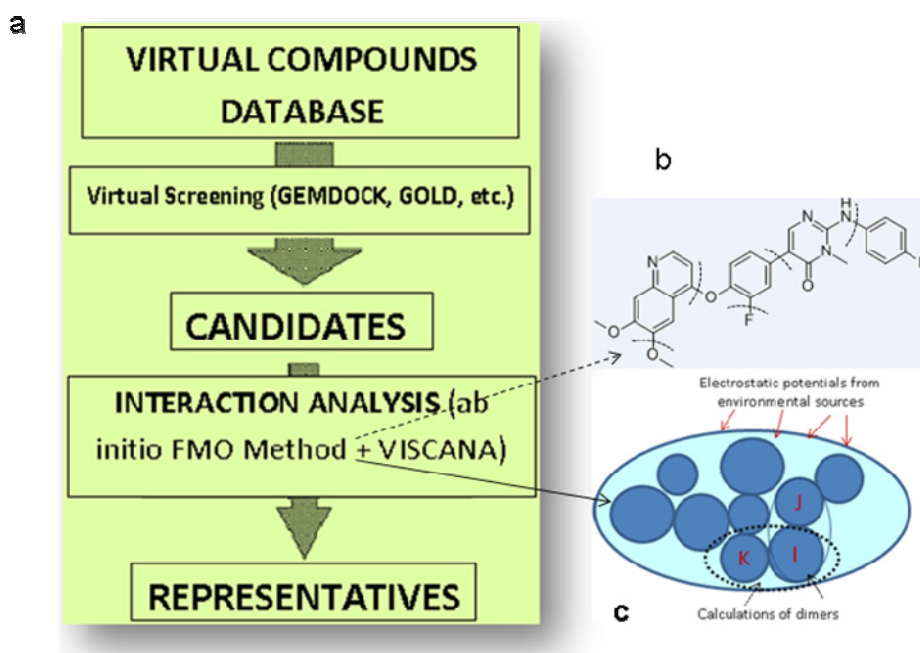


Figure 7. a) The overall approach of VISCANA (from VS to the selection of representatives). b) The fragmentation of a polypeptide at different bonds. c) Division of biomolecules into a collection of small fragments in the molecular orbital calculations (FMO method).

First, biomolecules or molecular clusters are divided into small fragments, and the ab initio MO calculations on the fragments (monomers) under the electrostatic potential from surrounding fragment pair as seen in Fig 7b and c. This is then solved repeatedly until all monomer densities become self-consistent. Finally, through the use of the total energies of the monomer  $E_I$  and the dimer  $E_{IJ}$ , the total energy of the system  $E$  is calculated by the following equation:

$$E_{\text{FMO}} = \sum_I E_I + \sum_{I < J} (E_{IJ} - E_I - E_J)$$

The FMO method has the advantage of describing the charge-transfer between a receptor and a ligand in comparison to a conventional force field method using fixed atomic charges. Based on this principle Amari *et al.* developed a cluster analysis using the dissimilarity defined as the squared Euclidean distance between interfragment interaction energies (IFIEs) of two ligands. VISCANA combines a clustering method with a graphical representation of the IFIEs by representing each data point with colors that quantitatively and qualitatively reflect the IFIEs. This method classifies structurally different ligands into functionally similar clusters according to the interaction pattern of a ligand and amino acid residues of a receptor protein. VISCANA also estimates docking conformation by analyzing patterns of the receptor-ligand interactions of some conformations through the docking calculations. VISCANA could be applied not only to the FMO method but also any molecular interaction system which can provide interaction energies or other properties of interest such as charge distribution.

### 3.2.4 iGEMDOCK: A Graphical Environment for Recognizing Pharmacological Interactions and Virtual Screening

iGEMDOCK (Fig. 8) is an extension of the original docking tool GEMDOCK developed by Yang *et al.* [26] which adds a post screening analysis method to the original docking algorithm (<http://gemdock.life.nctu.edu.tw/dock/igemdock.php>). GEMDOCK's two key functions for VS are used: 1) the searching algorithm [49] and 2) the scoring function [50] which is based on an empirical energy function:

$$E_{tot} = E_{bind} + E_{pharma} + E_{ligpre}$$

where  $E_{bind}$  is the empirical binding energy,  $E_{pharma}$  is the energy of binding site pharmacophores (hot spots), and  $E_{ligpre}$  is a penalty value if a ligand does not satisfy the ligand preferences.  $E_{pharma}$  and  $E_{ligpre}$  are especially helpful in selecting active compounds from hundreds of thousands of non-active compounds by excluding ligands that violate the characteristics of known active ligands, thereby improving the selection of true positives.



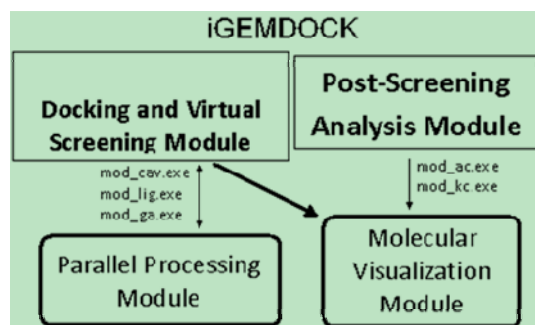


Figure 8. The virtual screening and post screening analysis processes in *iGEMDOCK*

The integration of different-stage programs of VS environments into GEMDOCK constituted the emergence of *iGEMDOCK* for docking, virtual screening and post screening analysis of database compounds using a friendly interface. In post-screening analysis *iGEMDOCK* enriches the hit rate and derives pharmacological interactions from screened compounds to provide biological insights. The pharmacological interactions represent conserved interacting residues which form binding pockets with specific physico-chemical properties expressing the essential functions of the target protein.

This new algorithm provides both virtual screening and post screening analysis as well as a more detailed and complete understanding of ligand binding mechanisms which makes the study and discovery of lead compounds much easier and less time consuming than other similar post screening analyses. *iGEMDOCK* is based on the efficiency of GEMDOCK which was able to mine various inhibitors such as aurintricarboxylic acid tetracycline derivatives which inhibit flaviviruses [6] and influenza virus neuraminidase inhibitors [8].

### 3.3 Summary

Methods of post screening analysis that enhance virtual screening enrichment and retrieve target compounds more accurately are of great use and interest in current bioinformatics. In this review we summarized and compared methods of VS and post screening analysis of lead compounds which emphasize the relevance of interaction profiles in mining suitable candidates.



SIFt (structural interaction fingerprint) is one of the pioneer methods in post screening analysis to include interaction-specific information into the real number strings. This enables the visualization, organization, analysis and retrieval of structures containing key interactions or specific features. A combination of SIFt and ChemScore (an empirical scoring function) contributed to a modest increase in the enrichment factor (EF) which was calculated based on the ability to recover known inhibitors. The enrichment increased from 37.0 EF<sup>a</sup> (SIFt) to 42.3 EF<sup>a</sup> (SIFt + ChemScore) [23].

VISCANA (Visualized Cluster Analysis of Protein-Ligand Interaction) uses a different approach through the FMO method. It has the advantage of describing the charge-transfer between a receptor and a ligand in comparison to a conventional force field method using fixed atomic charges. The difference between VISCANA and other conventional screening methods is that most methods choose the higher rank of a docking score on a point. In VISCANA a compound with a low docking score may belong to the same cluster that contains active compounds and the compound could be a suitable candidate. However, Amari *et al.* affirmed in their study VISCANA needs further development of quantum mechanical methods (the second-order Møller-Plesset perturbation theory based on the FMO method) to obtain more reliable descriptions of van der Waals interactions and hydrogen bonds which are important in determining receptor-ligand binding [22]. Other post screening studies reveal that unreliable or insufficient descriptions of important interactions account for increased numbers of false positives [48].

*i*GEMDOCK, an integration of VS and post screening methods is based on the original evolutionary docking algorithm GEMDOCK, currently one of the pioneer methods used for combining VS with visualizing, organizing, analysing and data mining of lead compounds. It has an advantage over SIFt and VISCANA primarily due to the attempt of eliminating two key issues: 1) if a docking tool is used for VS, which post screening analysis can complement it best and 2) if a post screening analysis method is decided, which docking tool or VS method is most suitable. The difference in the post screening approach of *i*GEMDOCK and other methods (VISCANA and SIFt) is the use of a module which clusters compounds based on interaction profiles and atomic compositions. Selecting representative compounds from each cluster enables the maintaining of compound diversity and reduces the number of false positives. In addition, its pharmacological scoring function can reduce the ill-effect of energy-based scoring functions

which often favor high molecular weight or highly-polar compounds. This improves the screening accuracy when the molecular weights of the active compounds are less than 400 Daltons (Da) [52]. Most notably, GEMDOCK, the earlier version of *i*GEMDOCK was used successfully to screen and identify inhibitors for influenza virus neuraminidases and flaviviruses [6, 8].

We also emphasize on the use of VS and post screening analysis in the mining of novel compounds for various other applications (e.g. industry, agriculture, cosmetics and nutritional supplements). These areas have not been getting much attention in comparison to drug design whereas certain protein-ligand complexes constitute key compounds in developing various biochemical products [29-31]. VS and post screening analysis used in computer-aided drug design reveal great potential in such applications since prospect candidates used in cosmetics and other industries may be retrieved employing interaction profiles.

Although the methods investigated in this study, SIFt, VISCANA and *i*GEMDOCK employ different techniques (structural interaction fingerprint, ab initio FMO method and interaction energy modules) they have one common feature; the use of protein-ligand interaction profiles which can be further exploited in developing new and improved methods to retrieve and analyze potential candidates for drug design and other applications. Through the development of better techniques, measures and description of interaction energies can aid methods of novel compounds retrieval and analysis, improve in accuracy and selection of active compounds. In addition, these observations point to an important aspect in the computer-aided drug design and discovery, the necessity for more than one stage of clustering in post screening analysis. From this point we proceeded with developing our new method Two-Stage Combinative Clustering (TSCC) [48] which combines our specifically optimized docking tool (GEMDOCK) with two stages of clustering for an optimized post screening analysis.

## CHAPTER 4

### **TSCC: Two-Stage Combinative Clustering for Virtual Screening Using Protein-ligand Interactions and Physical-Chemical Features**

#### **4.1 Introduction**

Continuous advancements in high-throughput X-ray crystallography and genomics [2, 28] account for increased numbers of available crystal structures enabling a more rapid development of new therapeutic targets. However, prospect ligands and proteins need to be screened in order to downsize groups [22, 23, 53] and select suitable candidates for post-screening analysis. Clustering methods based on structural similarity which are employed in post-screening analysis generally improve the scoring function performance. In developing methods for 3D compound retrieval, a detailed understanding of intermolecular interactions between proteins and their ligands is critical to structure-based inhibitor design. Various post-screening analysis methods and clustering [23, 54-56] employ RMSD values, protein-ligand interactions and computation and comparison platforms for measuring distances. Since the above methods as well as TSCC encounter challenges of specific selectivity and false positives, we aim to provide advantages to our post screening analysis method by using two combined clustering stages to rank all compounds and select final representatives more efficiently and accurately. The final representatives can be confirmed through bioassays to verify their target and the proper activity and application.

Although similar methods (IBAC, SIFt and VISCANA) [21-23] have used visualization and clustering of compounds to enrich VS, they have not identified novel compounds for any practical applications (drug design or industrial purposes). In addition, with the use of such methods one main issue remains unsolved: which combination of VS and post screening analysis is the most efficient. Our goal is to provide a more efficient method for post screening analysis to identify novel compounds, their possible functions and practical applications. Thus, we employ the empirical energy function from GEMDOCK [26] and the basic premise of SIFt [23] to encode additional interaction-specific information into the real number strings, hydrogen bonds, van der Waal and electrostatic forces. By representing interactions at the atomic-level as opposed to the residue level and including measures of interactions strength, protein-ligand interactions can be described better and a more precise analysis of virtual screening can be obtained.

TSCC is accomplished by joining two clustering stages; one of protein-ligand interactions (e.g. hydrogen bonds, electrostatic interactions, and van der Waals forces) with another of physico-chemical features (e.g. atom composition). We employed our docking tool, GEMDOCK, to generate protein-ligand interactions and used the Accelrys Cerius QSAR module for obtaining physico-chemical features for the compounds. Based on normalized feature profiles, hierarchical and K-mean [57] clustering methods were used to cluster and select compound candidates. Since clustering based upon similarity requires a quantitative measure (descriptor) of the similarity between two molecules, 2D and 3D methods were used to generate a descriptor such as the atom pair descriptor (i.e. compound topological similarity) [58].

To handle the vast results from virtual screening and use more specific information for protein-ligand binding, we utilize the empirical energy function from GEMDOCK [26] specifically optimized for virtual screening of ligands. GEMDOCK uses piecewise linear potential (PLP), a simple scoring function (Fig. 9), comparable to similar scoring functions for estimating binding affinities [60, 61]. Our previous works showed a comparison of GEMDOCK and other docking methods for 100 protein-ligand complexes and two virtual screening targets [49-50]. In addition, GEMDOCK has been successfully applied to identify inhibitors and binding sites for some targets [6, 8]. Here, we utilize the PLP of GEMDOCK to generate the protein-ligand interaction profiles.

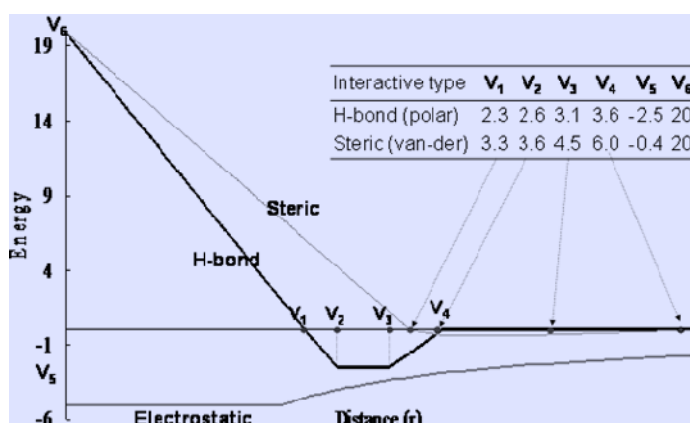


Figure 9. The linear energy functions of the pairwise atoms for the steric interactions and Hydrogen bonds in GEMDOCK (bold line) with a standard Lennard-Jones potential (thin line) Yang *et al.* [26].

To demonstrate the efficiency of our method we successfully applied its combinative two-stage concept in two separate post screening analysis studies. In the first study (sections 4.2, 4.3) two compound sets (testing and verifying) were designed to determine if the protein-ligand interaction descriptor is suitable for identifying compounds with similar binding modes. The two sets were also used to determine if the compound structure descriptor is suitable to identify similar structure compounds and to evaluate the database enrichment potential and the property of compounds in the same cluster by docking a diverse set of compounds spiked with known active compounds into the same target protein.

## 4.2 Materials and Methods

### *The Two-Stage Combinative Clustering (TSCC) Methodology*

The overview of TSCC concept in our first study is shown in Figure 10. We first calculated the atom-based protein-ligand interactions by converting every docked pose into a one dimensional real number string in order to visualize and analyze large data obtained from virtual screening using Yang *et al* [26].

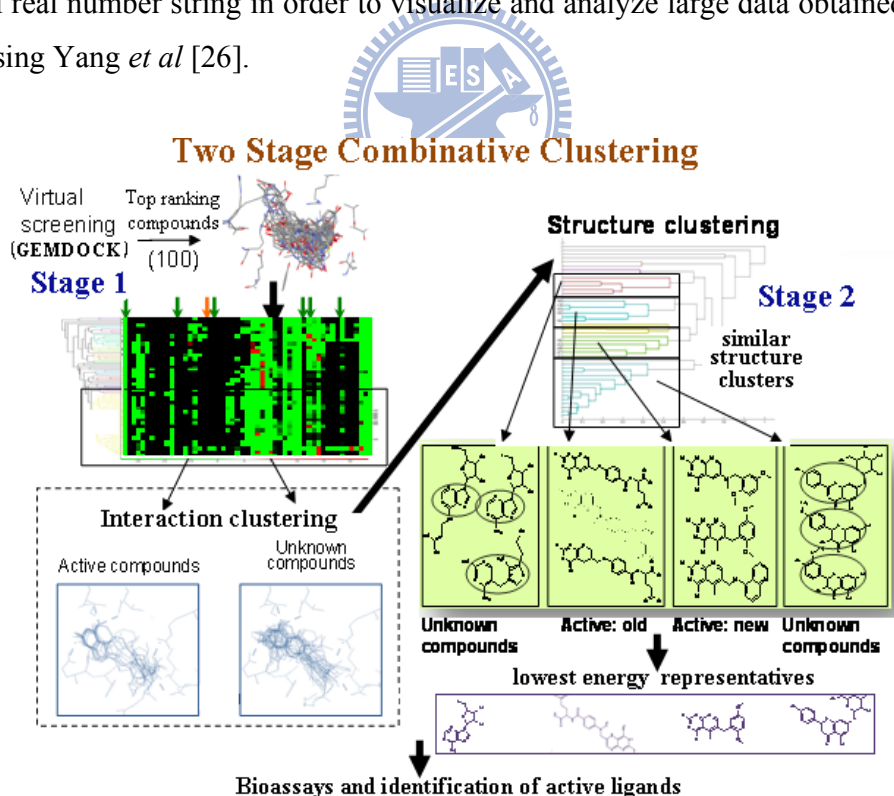


Figure 10. Overall process of TSCC in our first study (a) First stage clustering using P-L interactions generated via GEMDOCK. (b) Second stage clustering of first stage results using physico-chemical features. (Figure obtained from our published study [48]).

Due to protein-ligand interactions representation, we were able to evaluate the distance of binding modes between two docked poses and to carry out hierarchical clustering analysis. Compounds with a similar binding mode were visualized and grouped into clusters [59]. In our structure based clustering section, each structure was represented by a one dimension atom-pair descriptor, an approach proposed by Carhart *et al* [58]. After analyzing the distance between active and non-active compounds, a reference threshold was decided for demarcating similar compounds (Fig. 11).

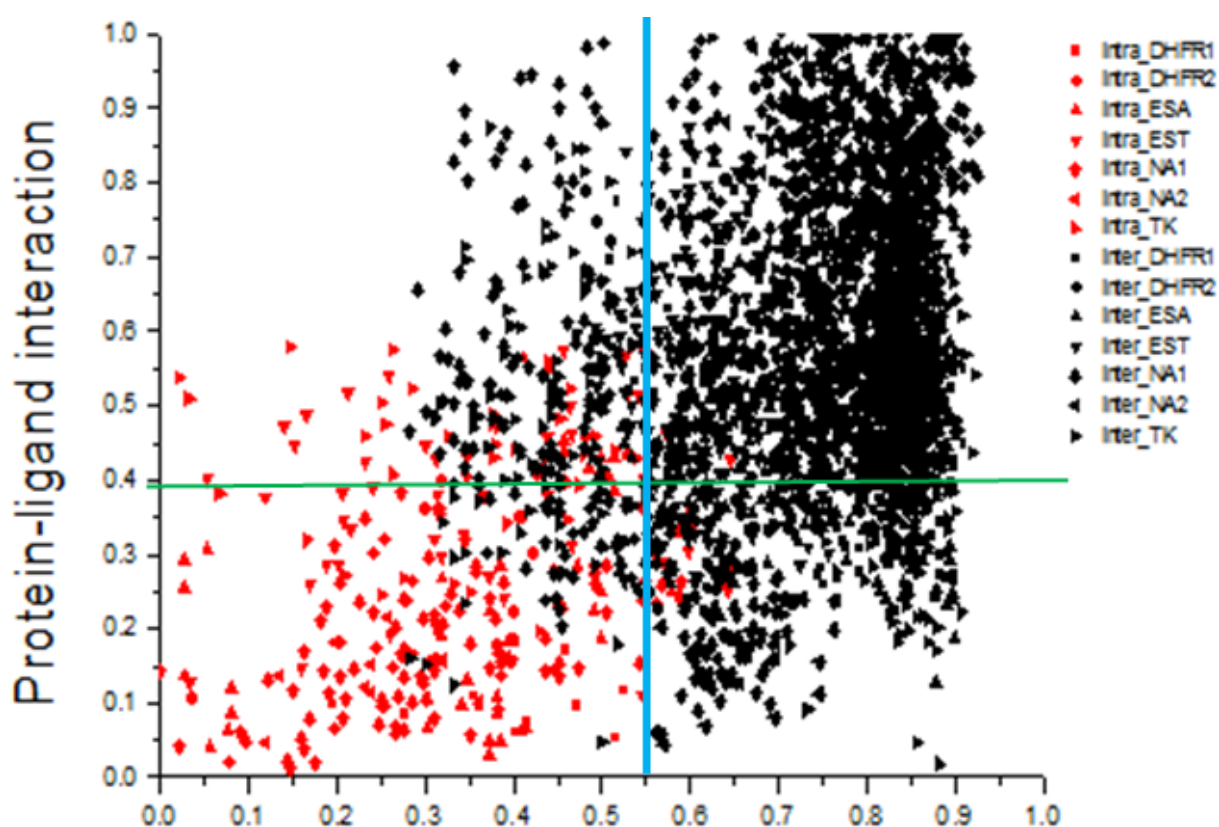


Figure 11. Designing the reference thresholds for protein-ligand interaction and atom-pair descriptor (Figure obtained from our published study [48]). The complementation between atom-pair descriptor and the protein-ligand interaction descriptor is also shown in this figure. The distance threshold of atom-pair descriptor obtained was 0.55 (tanimoto coefficient) and the threshold of distance of protein-ligand interaction descriptor was 0.39 (correlation coefficient).

We generated two sets of structure-based virtual screening results: 1) to verify if the protein-ligand interaction descriptor is suitable for identifying compounds with similar binding mode and 2) to evaluate the database enrichment potential and the property of compounds in the same cluster by docking a diverse set of compounds spiked with known inhibitors into the same target protein.

#### 4.2.1 Preparation of Target Protein and Compound Databases

The Ligand binding site was defined as a collection of amino acids using a cutoff radius of 10 Å from each atom on the bound ligand, since most studies in lead discovery use a cutoff radius between 8 to 12 Å. Structure files were stored as a PDB format for GEMDOCK analysis.

##### Compound databases

We constructed two compound sets for screening against each target protein: thymidine kinase (TK) PDB id: 1kim, estrogen receptor alpha-agonist (ER $\alpha$ ) PDB id: 3ert, estrogen receptor alpha-antagonist (ER $\alpha$ ) PDB id: 1gwr, human dihydrofolate reductase (hDHFR) PDB id: 1hfr, tern n9 influenza virus neuraminidase (NA) PDB id: 1mwe. The structures used were obtained from the database Comprehensive Medicinal Chemistry (CMC) and American Chemical Directory (ACD) and compounds with molecular weights between 200 and 800 D were chosen only based on the similar size of our active compounds. The active compounds (61 compounds, Appendix 1 a) were listed as the following: 1) TK: 10, 2) ER  $\alpha$  antagonists: 11, 3) ER  $\alpha$  agonists: 10, 4) hDHFR: 10, and 5) NA: 20. The two crystal structures of human estrogen receptors alpha have been intensively studied for their different functions (agonist 1GWR promotes coactivator binding while antagonist 3ERT blocks it) and ability to bind on the same site of the protein. The agonists play an important role in regulation of gene expression and prevention of osteoporosis while the antagonists have been used as treatment of hormone-dependent breast cancer [60, 61].



The testing dataset contained 990 randomly selected compounds combined with known active compounds for each target protein using a method from Bissantz *et al* [3]. This is a small scale public set of compounds used by studies in lead compound discovery. All compound structures were converted to mol formats and their hydrogen atoms removed using CORINA3.0 in order to be virtually screened by GEMDOCK. The active compound, target proteins and compound sets are available on our web at <http://gemdock.life.nctu.edu.tw/dock/download.php>.

#### 4.2.2 Preparation of Virtual Screening Result for Cluster Analysis

GEMDOCK was substantially modified, in preparation for the docking of different complex poses and to predict the binding affinity for each compound in the dataset using two key functions: 1) The searching algorithm [49] and 2) The scoring function which is based on an empirical energy function [50].

##### GEMDOCK scoring function

The energy function can be expressed by the following terms and equations:

$$E_{tot} = E_{bind} + E_{pharma} + E_{ligpre} \quad (1)$$

where  $E_{bind}$  is the empirical binding energy,  $E_{pharma}$  is the energy of binding site pharmacophores (hot spots), and  $E_{ligpre}$  is a penalty value if a ligand does not satisfy the ligand preferences.  $E_{pharma}$  and  $E_{ligpre}$  (see Mining pharmacological consensus subsection) help select active compounds by improving the number of true positives. The values of  $E_{pharma}$  and  $E_{ligpre}$  are set to zero if active compounds are not available. Thus, the empirical-binding energy ( $E_{bind}$ ) is given as:

$$E_{bind} = E_{inter} + E_{intra} + E_{penal} \quad (2)$$

where  $E_{inter}$  and  $E_{intra}$  are the intermolecular and intramolecular energies, respectively, and  $E_{penal}$  is a large penalty value if the ligand is out of the range of the search box. For this study,  $E_{penal}$  was set to 10,000. The intermolecular energy is defined as:

$$E_{inter} = \sum_{i=1}^{lig} \sum_{j=1}^{pro} \left[ F(r_{ij}^{B_{ij}}) + 332.0 \frac{q_i q_j}{4r_{ij}^2} \right] \quad (3)$$

where  $r_{ij}$  is the distance between the atoms  $i$  and  $j$ ;  $q_i$  and  $q_j$  are the formal charges and 332.0 is a factor that converts the electrostatic energy into kilocalories per mole. The *lig* and *pro* denote the numbers of the heavy atoms in the ligand and receptor, respectively.  $F(r_{ij}^{B_{ij}})$  is a simple atomic



pairwise potential function (Fig. 9) as defined in our previous study [5] where  $r_{ij}^{B_{ij}}$  is the distance between atoms  $i$  and  $j$  with interaction type  $B_{ij}$  formed by pair-wise heavy atoms between ligands and proteins,  $B_{ij}$  is either a hydrogen bond or a steric state. We used the atom formal charge to calculate the electrostatic energy [3], which is set to 5 or  $-5$ , respectively. The intramolecular energy of a ligand is:

$$E_{intra} = \sum_{i=1}^{lig} \sum_{j=i+2}^{lig} \left[ F(r_{ij}^{B_{ij}}) + 332.0 \frac{q_i q_j}{4r_{ij}^2} \right] + \sum_{k=1}^{dihed} A [1 - \cos(m\theta_k - \theta_0)] \quad (4)$$

where  $F(r_{ij}^{B_{ij}})$  is defined as in Equation 3 except the value is set to 1000 when  $r_{ij}^{B_{ij}} < 2.0 \text{ \AA}$ , and *dihed* is the number of rotatable bonds in a ligand. We followed the work of Gehlhaar *et al* [9] to set the values of  $A$ ,  $m$ , and  $\theta_0$ . For the  $sp^3$ - $sp^3$  bond,  $A = 3.0$ ,  $m = 3$ , and  $\theta_0 = \pi$ ; for the  $sp^3$ - $sp^2$  bond,  $A = 1.5$ ,  $m = 6$ , and  $\theta_0 = 0$ . When known active ligands are available, GEMDOCK uses a pharmacophore-based scoring function (Equation 1). If known active compounds are not available  $LP_{elec}$  and  $LP_{hb}$  are set to zero and GEMDOCK uses a purely empirical-based scoring function (Equation 2). After all of the protein-ligand interactions were calculated, the atom interaction-profile weight of the target protein representing the pharmacological consensus of a particular interaction was given as:

$$Q_j^k = \frac{f_j^k}{N} \quad (5)$$

where  $N$  is the number of known active compounds and  $f_j^k$  is the total interaction number of an atom  $j$  (in a protein) interacting with an atom of known active ligands with the interaction type  $k$  (e.g., hydrogen bonding or hydrogen-charged interactions). An atom  $j$  in the reference protein was considered a hot-spot atom when  $Q_j^k$  was more than 0.5.

#### 4.2.3 Testing and Verifying Datasets

The lowest energy conformation was retained for generating the representative docked pose of each compound.

## Generation of Descriptors (Protein-Ligand interaction descriptors)

We converted 3D docked conformations (poses) into a one dimension real number string by calculating the energy between each atom present on protein and ligand. The interaction energy of each atom  $j$  on a protein is defined as:

$$E_j = \sum_{i=1}^{lig} \left[ F(r_{ij}^{B_{ij}}) + 332.0 \frac{q_i q_j}{4r_{ij}^2} \right] \quad (6)$$

where  $r_{ij}^{B_{ij}}$  is the distance between atoms  $i$  and  $j$  with interaction type  $B_{ij}$  formed by pair-wise heavy atoms between ligands and proteins,  $B_{ij}$  is either a hydrogen bond or a steric state. These two potentials are calculated by the same function, although from different parameters;  $V_1, \dots, V_6$ .  $q_i$  and  $q_j$  are the formal charges and 332.0 is a factor that converts the electrostatic energy into kilocalories per mole. The *lig* and *pro* denote the number of heavy atoms on the ligand.  $F(r_{ij}^{B_{ij}})$  is a simple atomic pair-wise potential function.

### Atom pair descriptors

Atom-pair descriptors are 2D topological descriptors counting the distance between two atoms as the shortest path of bonds [58]. The procedure for preparing atom pair descriptors:

- 1) Structure files in mol format
- 2) Remove hydrogen atoms
- 3) Convert to mol2 format via CORINA3.0
- 4) Calculate atom pair descriptors via AP generator (distance bins: 15)
- 5) Store in binary coding form.

A total of 825 (55 x 15) atom pair descriptors were generated for each molecular structure by removing all columns with zero values.

### Reference Threshold for Protein-Ligand Interaction and Atom-Pair Descriptor

To design a reference threshold of protein-ligand interaction, a verifying dataset was used in establishing a reference threshold of distance by determining a maximum discrimination that

exists between similar and non-similar binding modes. The equation is as follows:

$$\max\left(\left(\frac{C_{intra-d<t}}{C_{intra}} + \frac{C_{inter-d>t}}{C_{inter}}\right)/2\right) \quad (7)$$

Where  $t$  is the reference threshold,  $C_{intra-d<t}$  is the number of intra active compound pairs with the distance  $<$  threshold and  $C_{inter}$  is the number of compound pairs between active and non-active compounds.

### The Cluster Analysis Method

First, we used a protein-ligand interaction descriptor for clustering compounds with similar binding modes and applied the correlation coefficients as similarity measurements. The following formula was used:

$$D_{xy}^{corr} = 1 - \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)S_x S_y} \quad (8)$$

where  $D_{xy}^{corr}$  is the correlation distance between docked pose  $X$  and  $Y$ .  $S_x$  is the standard deviation of  $X$ .  $X_i$  is the  $i$ th value of  $X$ .  $n$  is the number of descriptors. We applied the standard UPGMA clustering method for calculating the distance between two clusters while constructing the dendrogram. The formula is defined as:

$$D_{rs}^{clu} = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D_{risj}^{corr} \quad (9)$$

The reference threshold was calculated from the verifying dataset using equation (2) to determine the number of clusters.

Second, we applied the AP descriptor for clustering compounds within each clustering stage and applied the tanimoto coefficients as similarity measurements. Formula is as follows:

$$D_{xy}^{tani} = \frac{|X \cap Y|}{|X \cup Y|} \quad (10)$$

where  $D_{xy}^{\text{tani}}$  is the tanimoto distance between  $X$  and  $Y$ .  $|X \cap Y|$  is the number of ON bits common in both  $X$  and  $Y$ , and the  $|X \cup Y|$  is the number of ON bits present in either  $X$  or  $Y$ . This equation is similar to equation (4);  $D_{xy}^{\text{corr}}$  by  $D_{xy}^{\text{tani}}$ . The dendrogram graph was plotted for visualizing the binding mode of multi docked poses by the protein-ligand interactions.

## 4.3 Results

### *Molecular Recognition*

#### **I. Thymidine kinase (TK)**

The significance of TK as a target in computer-aided drug design is its involvement in the phosphorylation of nucleosides or nucleoside analogs [62]. Various antiviral drugs attack the replication of the viral genome with nucleoside analogs. These analogs are activated by phosphorylation with TK and prevent DNA synthesis by the introduction of a chain-terminating nucleoside at the 3' end of the growing DNA strand. Thus, we screened against this target and choose the crystal coordinates of TK (Appendix 1a) in complex with its natural substrate (deoxythymidine). This is reasonable since the active site can accommodate a broad variety of ligands. The average RMSD of all ten docked poses was 1.39 Å. (Table 2)

#### **II and III. Estrogen receptor $\alpha$ (ER $\alpha$ antagonists and agonists)**

Estrogens contribute to the maintenance of bone tissue through a process involving bone resorption and bone formation [60] which makes for another appropriate target. The target protein structures of ER $\alpha$  (Appendix 1a) were obtained from PDB, whereas antagonists and agonists were derived from previous studies [3, 63]. We docked four antagonists into the target protein (3ert) and four agonists into another one (1gwr), and concluded their results based on the root mean square deviation (RMSD) error in the heavy atoms ligand between the docked pose and the crystal structure. The average RMSD of docked antagonists and agonists was 1.42 Å. The RMSD values of 1hj1.AOE and 1qkm.GEN were larger than 2.0 Å because the native proteins were crystal structures of Er $\beta$ -ligand complexes. (Table 2)

Table 2. The RMSD between docked poses and crystal ligands [48]

TK (1kim)		ER (3ert, 1gwr)		DHFR (1hfr)		NA (1mwe)	
Complex name	RMSD (Å)	Complex name	RMSD (Å)	Complex name	RMSD (Å)	Complex name	RMSD (Å)
<i>1e2k.TMC</i>	0.69	<i>1err.RAL<sup>a</sup></i>	1.27	<i>1boz.PRD</i>	1.13	<i>lig1l7f_BCZ</i>	0.88
<i>1e2m.HPT</i>	0.51	<i>3ert.OHT<sup>a</sup></i>	0.71	<i>1dlr.MXA</i>	0.62	<i>lig1nnc_GNA</i>	0.75
<i>1e2n.RCA</i>	1.34	<i>1hj1.AOE<sup>a</sup></i>	3.13	<i>1dls.MTX</i>	1.53	<i>lig2qwf_G20</i>	0.60
<i>1e2p.CCV</i>	0.67	<i>1uom.PTI<sup>a</sup></i>	0.81	<i>1drf.FOL</i>	1.24	<i>lig1bji_G21</i>	0.81
<i>1ki2.GA2</i>	3.04	<i>1gwr.EST<sup>b</sup></i>	0.71	<i>1hfr.MOT</i>	0.51	<i>lig1f8b_DAN</i>	0.64
<i>1ki3.PE2</i>	3.21	<i>1l2i.ETC<sup>b</sup></i>	0.52	<i>1kms.LIH</i>	1.36	<i>lig1f8c_4AM</i>	0.46
<i>1ki6.AHU</i>	0.37	<i>1qkm.GEN<sup>b</sup></i>	2.92	<i>1kmv.LII</i>	0.83	<i>lig1f8d_9AM</i>	0.59
<i>1ki7.ID2</i>	0.49	<i>3erd.DES<sup>b</sup></i>	1.32	<i>1mvs.DTM</i>	0.75	<i>lig1f8e_49A</i>	0.60
<i>1kim.THM</i>	0.41			<i>1ohj.COP</i>	1.27	<i>lig1ina_ST6</i>	0.79
<i>2ki5.AC2</i>	3.14			<i>2dhf.DZF</i>	1.12	<i>lig1ing_ST5</i>	1.03
						<i>lig1inw_AXP</i>	0.93
						<i>lig1inx_EQP</i>	0.92
						<i>lig1ivc_ST2</i>	2.09
						<i>lig1ivd_ST1</i>	1.02
						<i>lig1ive_ST3</i>	1.03
						<i>lig1mwe_SIA</i>	0.52
						<i>lig1xoe_ABX</i>	1.33
						<i>lig1xog_ABW</i>	2.42
						<i>lig2qwg_G28</i>	0.80
						<i>lig2qwh_G39</i>	0.74
<i>Average RMSD (Å)</i>	1.39	<i>Average RMSD (Å)</i>	1.42	<i>Average RMSD (Å)</i>	1.03	<i>Average RMSD (Å)</i>	0.95



<sup>a</sup> Four antagonists docked into the target protein (3ert)

<sup>b</sup> Four agonists docked into the target protein (1gwr)

#### IV. Human Dihydrofolate Reductase (hDHFR)

The inhibition of DHFR activity reduces the intracellular pool of THF resulting in inhibition of DNA synthesis and leading to cell death. Based on this mechanism, human DHFR (hDHFR) has become a major drug target in anticancer therapy. It is also a target for inhibition of bacterial, fungal, and protozoal DHFRs to treat human infectious diseases by many implicated microorganisms [61]. Therefore we screened against hDHFR and also evaluated the docking accuracy of GEMDOCK by docking 10 known active compounds (Appendix 1a) into this target

protein. Then, the RMSD values between the docked pose and the bound ligand in crystal structure were compared. The average RMSD of all ten docked active compounds was 1.03 Å (Table 2), substantially lower than 2 Å, which means GEMDOCK computations were within the range of accepted accurate values.

## V. Neuraminidase (NA)

Inhibitors of NA can protect the host from viral infection [62]. Influenza is an RNA virus that contains two major surface glycoproteins, neuraminidase (NA) and hemagglutinin (HA). It causes major respiratory infections associated with significant morbidity in the general population and mortality in elderly and high-risk patients, therefore NA could be a potential target to inhibit the influenza virus [64]. Thus, we docked 20 known active compounds (Appendix 1a) of NA into the target protein and obtained an average RMSD of 0.95 Å for all docked poses. (Table 2)

### Significance of the Descriptor (Significance of Protein-ligand Interaction Descriptor)

*Significance of known compounds in the five classes:* the results are listed in Table 3 using T-scores as the standard two sample *t*-test statistics. Using equation 7, the maximum discrimination was determined (Figure 10) in distinguishing between similar and non-similar binding modes.

Table 3. T-test of distance between similar and non-similar binding mode generated by converting the docked pose into protein-ligand interaction profile ( $\alpha=0.01$ ) [48].

Target protein	H <sub>0</sub>	Similar : Average Distance(Å)	Non-similar : Average Distance (Å)	P-value	Similar : Std <sup>a</sup> of Distance	Non-similar : Std <sup>a</sup> of Distance
DHFR	Reject	0.21	0.50	1.71E-58	0.09	0.13
ESA	Reject	0.25	0.42	7.04E-20	0.13	0.12
EST	Reject	0.31	0.48	7.94E-39	0.09	0.12
NA	Reject	0.17	0.73	0.00E+00	0.07	0.20
TK	Reject	0.19	0.47	3.89E-64	0.08	0.15

<sup>a</sup> Standard Deviation

**Significance of similar compounds:** For the purpose of post-analysis, we tested similar compounds' docking behavior (pose, interaction) on a protein receptor. There are five classes of similar compounds on each target protein. We tested to see whether the mean distance between similar compounds represented by protein-ligand interactions is different than the mean distance between non-similar compounds and recorded their results in table 4.

Table 4. T-test of distance between similar and non-similar structure generated by atom-pair representation ( $\alpha=0.01$ ), [48].

Target protein	H <sub>0</sub>	Similar : Average Distance (Å)	Non-similar: Average Distance (Å)	P-value	Similar : Std <sup>a</sup> of Distance	Non-similar : Std <sup>a</sup> of Distance
DHFR	Reject	0.42	0.63	5.84E-23	0.15	0.12
ESA	Reject	0.24	0.66	4.60E-65	0.11	0.14
EST	Reject	0.27	0.63	2.85E-56	0.14	0.14
NA	Reject	0.32	0.65	1.75E-131	0.18	0.17
TK	Reject	0.22	0.63	2.11E-93	0.09	0.19

<sup>a</sup> Standard Deviation

**Significance of an atom pair descriptor:** Similar structures were defined as active compounds and non-similar structures were defined as non-active compounds (Table 5). Active compounds of hDHFR and NA (Appendix 1) were divided into two classes because of their diverse compound structures. The maximum discrimination between similar and non-similar structures was determined by distinguishing between similar and non-similar structures.

### Calculating a reference threshold by verifying dataset

Using a verifying dataset, we calculated the distance threshold (correlation coefficient: 0.39) that had the maximum discrimination. The reference threshold of atom-pair (Tanimoto coefficient: 0.55 in Fig. 11) was calculated *via* 7 classes of structures showing the complement between atom-pair descriptor and protein-ligand interaction descriptor.

Table 5. T-test of distance between similar and non-similar compounds on each target protein. The descriptor was generated by converting the docked conformation into a protein-ligand interaction profile ( $\alpha=0.01$ ). (table obtained from our published study [48])

Target protein	Compound class	H <sub>0</sub>	Similar : Average Distance (Å)	Non-similar : Average Distance (Å)	P-value	Similar : Std <sup>a</sup> of Distance	Non-similar : Std <sup>a</sup> of Distance
DHFR	DHFR	Reject	0.21	0.50	1.71E-58	0.09	0.13
	ESA	Reject	0.52	0.58	2.73E-03	0.18	0.12
	EST	Reject	0.52	0.63	7.51E-07	0.21	0.13
	NA	Reject	0.46	0.55	5.34E-23	0.13	0.14
	TK	Reject	0.38	0.51	8.03E-11	0.16	0.13
ESA	DHFR	Pass	0.55	0.62	0.10111	0.28	0.16
	ESA	Reject	0.23	0.48	2.29E-31	0.14	0.14
	EST	Pass	0.67	0.76	0.23105	0.25	0.14
	NA	Reject	0.33	0.59	1.51E-58	0.24	0.20
	TK	Reject	0.46	0.57	0.000121	0.25	0.20
EST	DHFR	Pass	0.55	0.57	4.01E-01	0.21	0.14
	ESA	Reject	0.25	0.42	7.04E-20	0.13	0.12
	EST	Reject	0.31	0.48	7.94E-39	0.09	0.12
	NA	Reject	0.40	0.46	1.46E-09	0.15	0.15
	TK	Reject	0.28	0.43	2.17E-29	0.09	0.15
NA	DHFR	Reject	0.35	0.68	3.46E-25	0.22	0.25
	ESA	Reject	0.59	0.71	2.91E-04	0.28	0.24
	EST	Reject	0.56	0.66	2.46E-04	0.25	0.24
	NA	Reject	0.17	0.73	0.00E+00	0.07	0.20
	TK	Reject	0.48	0.60	3.46E-07	0.18	0.23
TK	DHFR	Reject	0.42	0.62	9.80E-12	0.13	0.10
	ESA	Reject	0.16	0.52	9.99E-62	0.07	0.13
	EST	Pass	0.58	0.65	6.28E-02	0.18	0.14
	NA	Reject	0.40	0.53	2.92E-53	0.11	0.15
	TK	Reject	0.19	0.47	3.89E-64	0.08	0.15

<sup>a</sup> Standard Deviation



## **Protein-ligand interaction clustering**

### ***Cluster analysis of Human Dihydrofolate Reductase Molecular Docking***

The overlays of all 61 docked poses of known active compounds in the vicinity of the target protein hDHFR are shown in Figure 12a. Using the reference threshold of protein-ligand interaction (correlation coefficient: 0.39), three major clusters can be identified and are shown in Figure 12b, clusters c, d and e. Each cluster has interaction details displayed above (e.g. cluster c with fig. c). All active compounds were grouped together (Fig. 12c). The hDHFR ligands in cluster c had hydrogen bonds (E30-OE1, E30-OE2, V115-O, I7-O in green dotted lines) and van der Waals forces shown by a blue arc (I60-CAR, F31-RING) revealing that binding interactions of each docked pose within cluster c were similar. Cluster d contained 6 TK ligands and one NA ligand and cluster e had only NA ligands, as seen in Figure 12e. Docked poses within both clusters d and e had hydrogen bonding (V115, I7-O; E30-OE1, V8-N).

When comparing the binding interaction between clusters in Figures 12c, d, e, f, and g we observe that docked compound poses are clustered into distinct clusters revealing specific binding interactions and important protein-ligand interactions residues.

### **Cluster analysis of Thymidine Kinase following Molecular Docking**

After filtering out clustered compounds, 53 docked poses were obtained including the 10 docked poses of active compounds and a total of 305 atoms were identified here (Fig. 13).

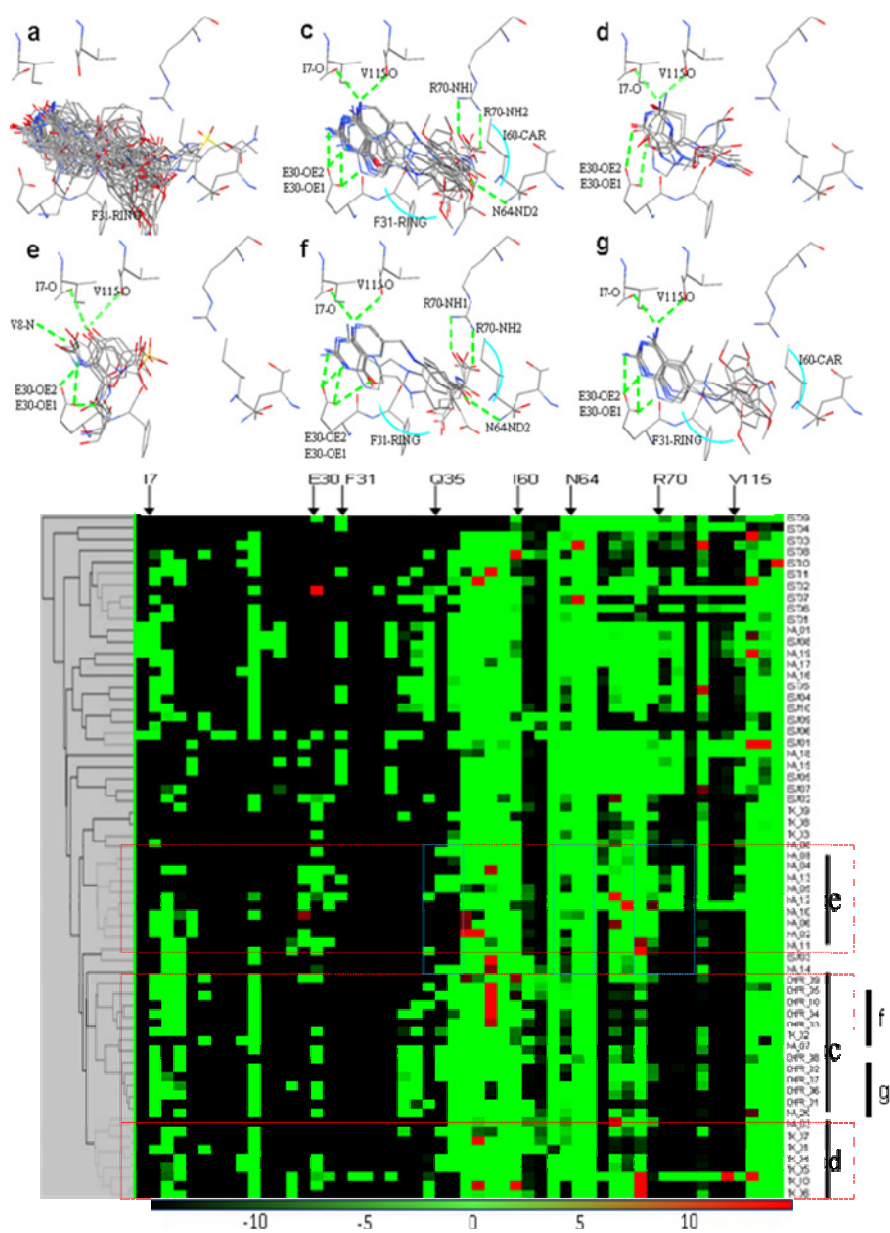


Figure 12. Cluster analysis of hDHFR (a) Overlay of all 61 docked poses of known active compounds in the vicinity of the target protein hDHFR (PDB id: 1hfr). (b) The dendrogram and hierarchical clustering results of 61 docked poses of hDHFR. Each cluster has its interaction details in the figures above (e.g. cluster c in fig c). 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 are shown in the top side of the heat map. (c, d, e) Overlay of the known active compounds and their important interactions. (f, g) Docked poses overlay of the sub-cluster within hDHFR active compounds. The differences of clusters f and g are shown by blue frames in the heat map. (Figure from our published study [48])

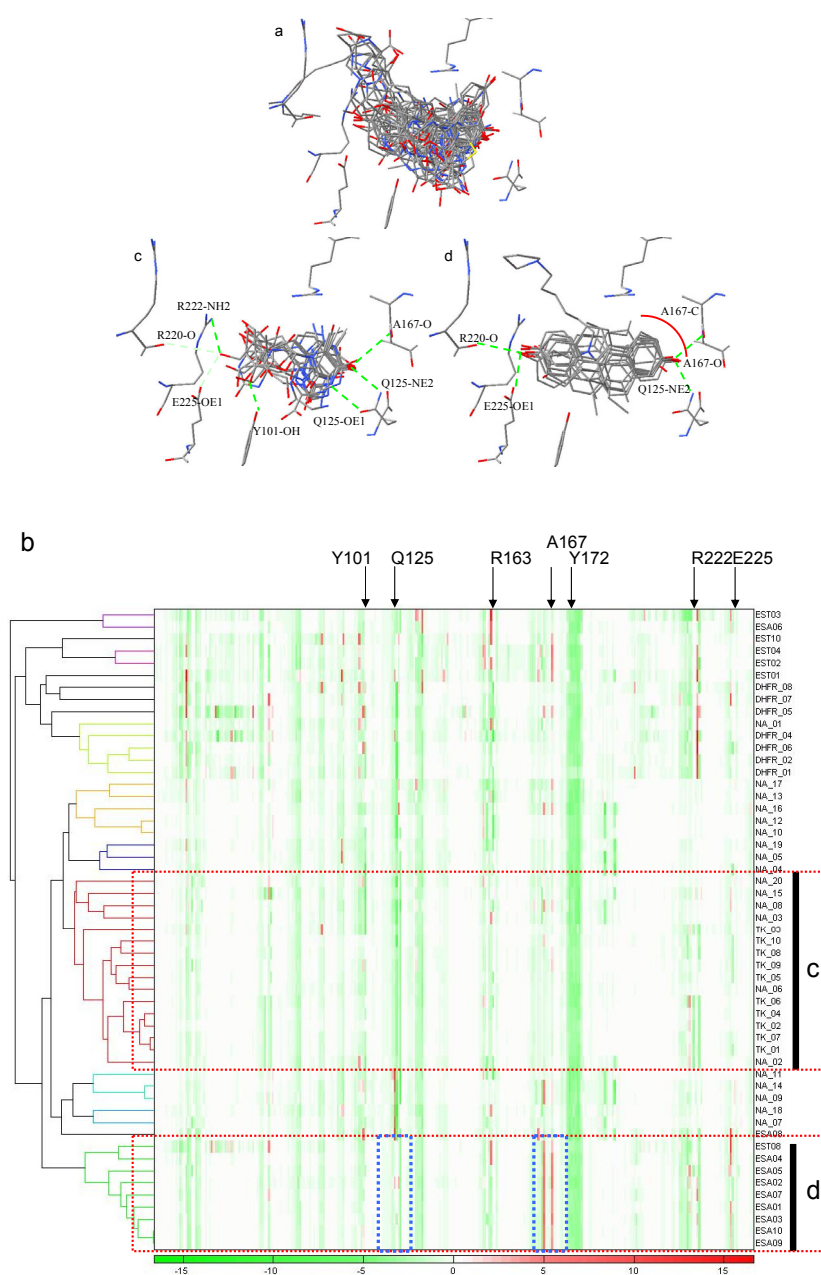


Figure 13. (a) Overlay of all 53 docked poses of known active compounds in the vicinity of the target protein Thimidine Kinase (PDB id: 1kim). (b) Hierarchical clustering of 53 TK docked poses' protein-ligand interactions (PDB id: 1kim). Each docked pose is one line in the heat map, the red being the lowest P-L interaction energy and the green being the highest. The left side of the heat map shows the hierarchical clustering results of TK. The hot spots identified from known overlapping active compounds are shown at the top. (c) Overlay of docked poses of the cluster with most number of known active compounds and important h-bonds between protein and ligand. (d) Overlay of docked poses of the cluster with most number of unknown compounds and important h-bonds between protein and ligand. The blue frames in the heat map were the major interaction differences among clusters c and d. (figure from our published study [48])

## Clustering using the atom-pair descriptor

### *Cluster analysis of compound structures for the verifying dataset*

Observing these three clusters, we deduced the atom-pair descriptor could group compounds with similar structures and sorts them from those with different structures (Fig. 14).

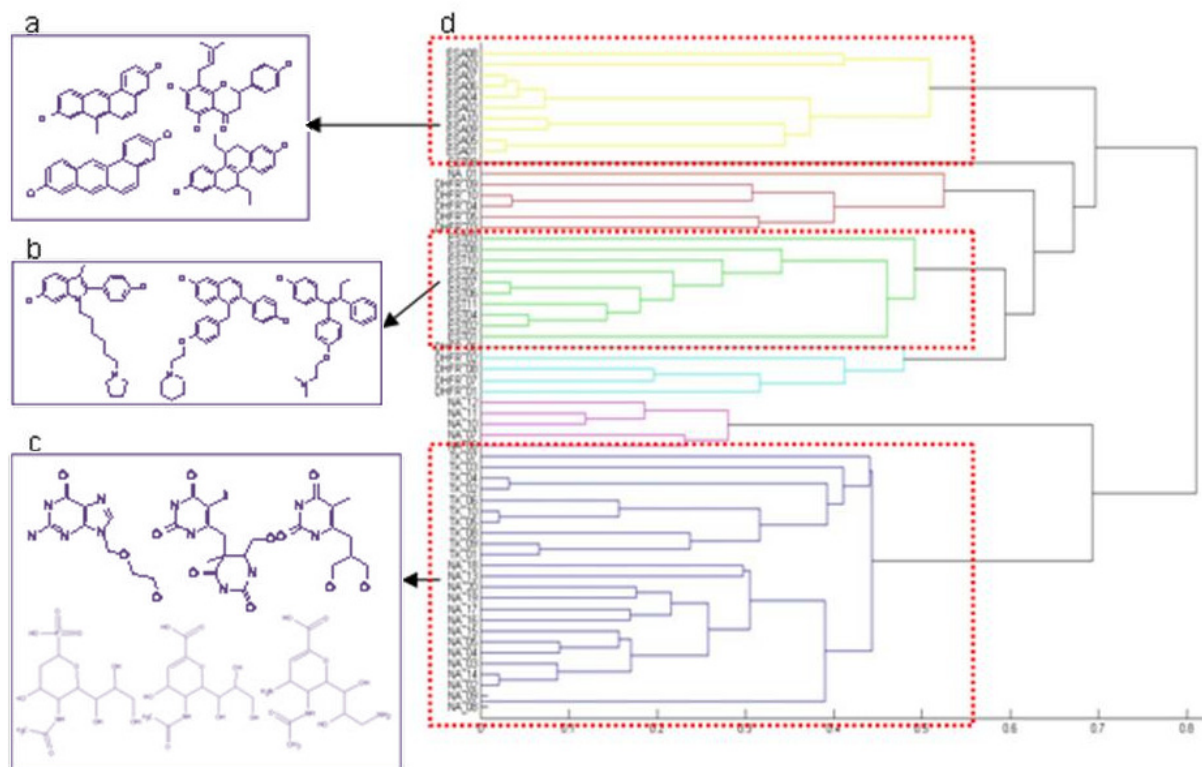


Figure 14. The hierarchical clustering dendrogram for the 61 known compound structures showing the three major clusters. (a) 10 ER $\alpha$  agonists. (b) 11 ER $\alpha$  antagonists. (c) 10 TK and 14 NA inhibitors were grouped into one cluster due to their structure similarity. The descriptor grouped only compounds with similar structures, sorting them out from those with different structures. (Figure from our previous study [48])

### **Cluster analysis of virtual screening results on the testing dataset**

#### *Analysis of the hDHFR dataset (first and second stages)*

**First stage:** We performed virtual screening for a set of 10 hDHFR inhibitors all spiked into 990 randomly selected compounds from ACD. A total of 476 involved atoms were

identified in 100 docked poses that include 10 known active compounds. Protein ligand interactions of all complexes were generated, each complex being composed of 316 real numbers. All hDHFR inhibitors were grouped together into one cluster. In Figure 15a indicated by red arrows are: F31-stacking forces, I60-van der Waals forces and NAP-stacking forces. Figures 15b and 15c shows similar hydrogen bonding (I7-O, V115-O, E30-OE1, E30-OE2, and N64-ND2) for the target protein and the 35 unknown compounds, however, the old drug (Fig. 15c) contains additional hydrogen bonds (R70-NH1, R70-NH2, and N64-ND2). We also identified and pointed out important forces on the heat map using red arrows (I60-van der Waals forces, F31-stacking forces, F34-stacking forces, NAP-stacking forces) Residues within old and new drug structures (Fig. 15a and b) are shown in yellow and the dendrogram in Figure 14b shows the exact split of these two compounds. We utilized 2D topology to select representative compounds within a cluster after protein-ligand interaction analysis was performed. A total of 45 compounds (10 active and 35 unknown compounds) were selected *via* first-stage clustering.

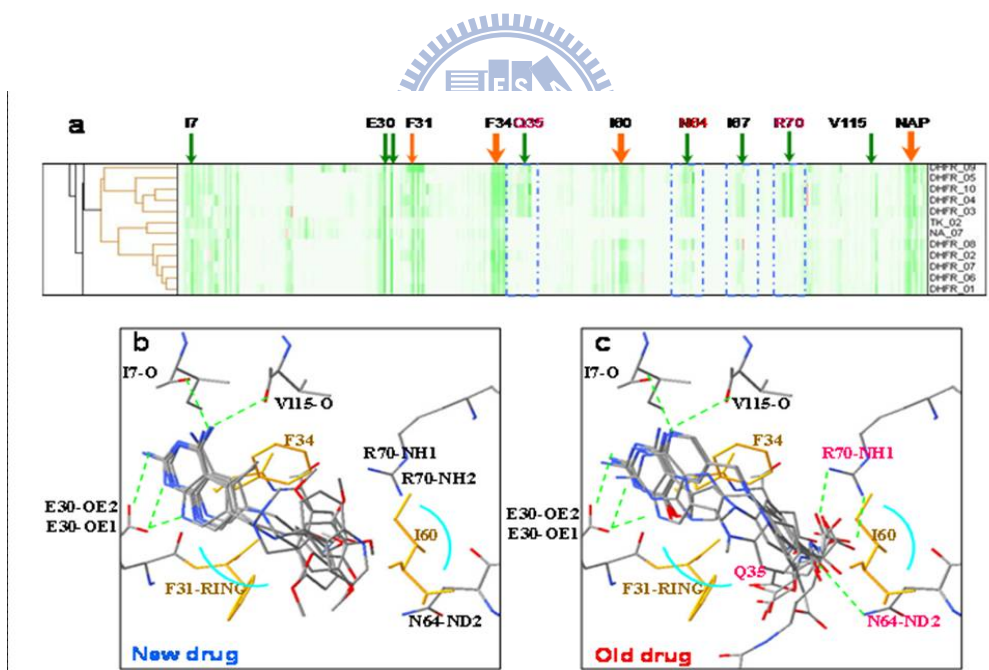


Figure 15. The detail of hDHFR binding interactions of new drugs and old drugs on the verifying dataset. (a) Important forces (red arrows) on the heat map (I60-van der Waals force, F31-stacking force, F34-stacking force, NAP-stacking force); (b), (c) The binding interactions of new and old drugs and their residues (yellow). The old drug (c) has additional hydrogen bonding with the target protein (Q35, N64, and R70). Interactions of residues (Q35, N64, R70) are seen in (b) while (N64 and R70) interactions are seen in (c). (Figure from our published study [48])



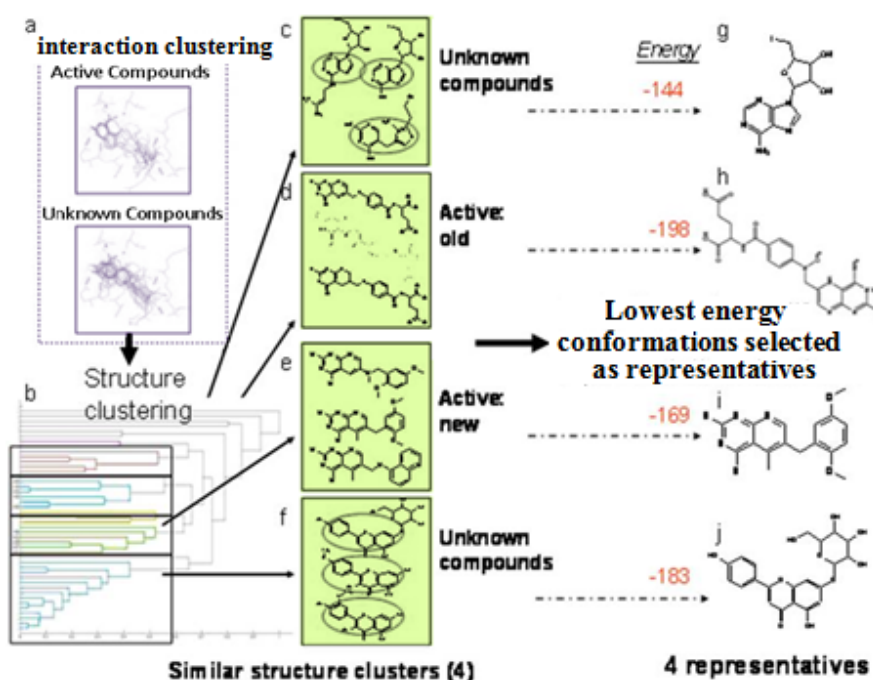


Figure 16. The process and results of second stage cluster analysis on hDHFR testing dataset. (a) The binding interactions of the largest cluster generated from first stage clustering: 45 compounds include the 10 active compounds and 35 unknown compounds. (b) The result of hierarchical clustering: there were four major clusters identified by the dendrogram (c, d, e, f). The active compounds were spliced into two clusters: (d) the old drugs and (e) the new drugs due to the difference in carboxylic acid groups. The sub-structures from clusters within the red circles in (c) and (f) had similar compounds and the lowest energy compound within each cluster was selected as a final representative (g), (h), (i) and (j). (Figure from our published study [48])

**Second Stage:** The cluster contained 45 compounds: 10 active compounds and 35 unknown compounds (Fig. 16a). A one dimension atom-pair binary string of 2D topology represented each compound. After performing hierarchical clustering four major clusters were identified in the dendrogram (Fig. 16b). The active compounds were spliced into two clusters; the old drugs (Fig. 16d) and the new drugs (Fig. 16e) due to the differences in carboxylic acid groups. The sub-structures within each cluster inside the circles (Figs. 16c and f) showed similar compounds within a cluster with the lowest energy compound from each cluster being selected as a final representative (Figs. 16g, h, i and j). The selected candidates are considered suitable for investigations by bioassays for further clues, specific functions and possible applications.

## 4.4 Verifying the TSCC method using $\beta$ -lactoglobulin

### 4.4.1 Introduction

The relevance of VS and post screening analysis in various applications besides drug design is well established [65, 66]. In this second part of our TSCC study, we aim to further test our method's two-stage clustering efficiency of ranking compounds by docking non-inhibitor type active compounds mixed in a dataset of randomly chosen compounds into a target protein. We are interested to explore the uses of TSCC in mining various other compounds besides inhibitors, thus, we use a transporter protein ( $\beta$ -lactoglobulin) as our target and three active compounds, Riboflavin, Calcitriol (activated vitamin D) and adenosine triphosphate that have various important functions in the human body but no known inhibiting functions. Additionally instead of using the results from first-stage clustering (a small number of compounds) as done in the first part of our TSCC study, we used the same number of compounds in the original dataset for both stages. We also used a specific physico-chemical feature for our second-stage clustering, an atomic composition clustering stage. In structure clustering, compounds with similar structures contain many similar atoms, thus, the atomic composition is a suitable feature in clustering chemical compounds based on their atomic composition similarity. The TSCC method using a slightly different approach from the first part of our study (Section 4.2) is shown in Figure 17.

The target used,  $\beta$ -lactoglobulin ( $\beta$ -LG) is a whey protein found in bovine, ovine and caprine milk and other species of related families. It is known for its ability to transport various important molecules in the human body and it's especially important in the uptake of vitamin D [67]. It also has various antimicrobial and cytotoxic functions when glycosylated with several sugars [68] and cholesterol modulating functions [69] and it may have more additional functions and mechanisms not yet known. These facts make  $\beta$ -LG a very important protein and studies are continuing to investigate current and additional roles of this target.

The three active compounds used, Calcitriol (vitamin D), Riboflavin (Vitamin B2) and Adenosine Triphosphate are ligands known from previous studies to bind to  $\beta$ -LG [67-69] and all are involved in significant functions of the human body. Vitamin D is involved in modulating calcium, gene expression and it is highly involved in autoimmune and infection suppression. Riboflavin plays a key role in energy metabolism and also the necessary factor for preventing

pellagra. It also protects the body from invasion by free radicals. Adenosine triphosphate (ATP) is a multifunctional nucleotide made of several sugars that inter-convert into ATP, ADP and AMP. When ATP is used in DNA synthesis, the ribose sugar is first converted to deoxyribose. In this study, we also want to see the efficiency of TSCC when the number of active compounds is rather small (three actives in this study vs. 61 actives in previous study and only one target,  $\beta$ -LG is used as opposed to five targets in sections 4.2, 4.3).

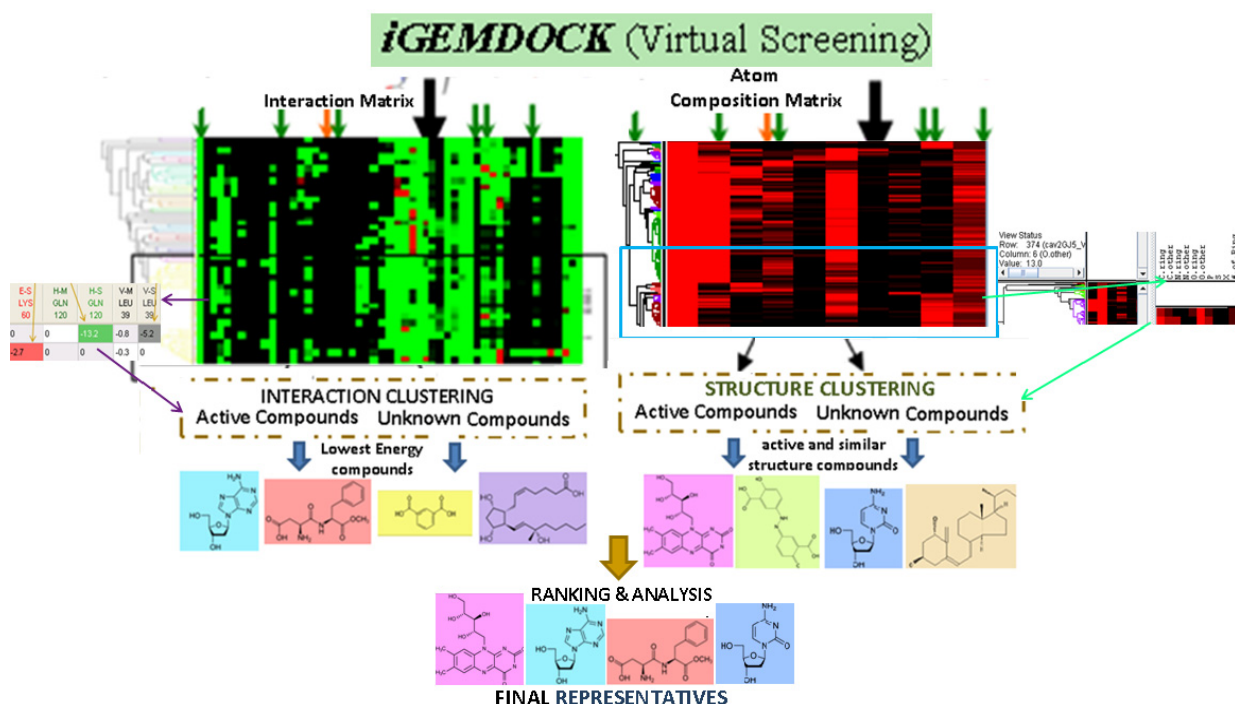


Figure 17. The overall approach of TSCC in our second study using interaction clustering and atomic composition clustering

#### 4.4.2 Materials and methods

##### Target protein, active compounds and dataset

The protein target,  $\beta$ -LG PDB id: 2gj5 (Fig. 1) was obtained from Protein Data Bank (PDB) and its cavity was prepared for molecular docking. The three active compounds (Fig. 18) were spiked into a dataset constructed from a 990 compound dataset from ACD [3] and 1306 compounds randomly selected from FDA databases, a total of 2206 compounds. The 1,306



compounds were found after virtually screening FDA database for ligands with a molecular weight between 200 and 600 Daltons similar to our active compounds (Fig 18).

#### 4.4.3 Molecular Docking and Post Screening Analysis

The method and steps for VS were the same as done in our previous study (section 4.2) except *iGEMDOCK*, an improved version of the old GEMDOCK generic algorithm specifically optimized for virtual screening was used to screen and generate profiles for each compound after compounds were docked into the cavity of the protein target. After the interactions and atomic composition profiles were generated for each compound, the compounds were clustered into similar interaction and similar atoms clusters (Fig. 19) and were ranked in order of their interaction energies generated by *iGEMDOCK*.

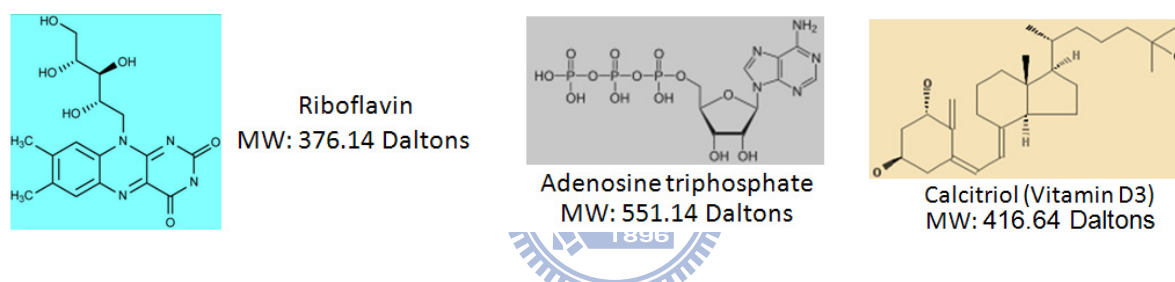


Figure 18. Active compounds used in the validation of the TSCC method.

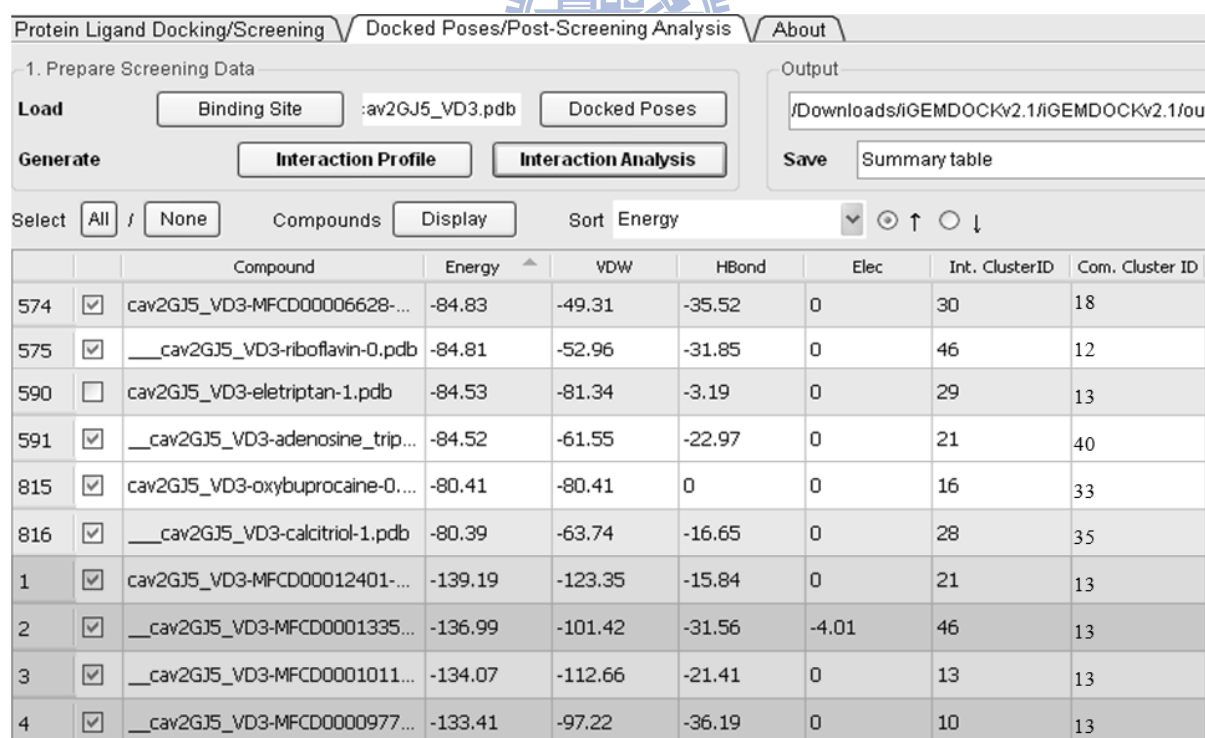
Post screening analysis in the  $\beta$ -LG study differs slightly in that it uses atom composition as a physico-chemical feature for compound structure clustering. Interaction clustering and atom composition clustering was performed on all compounds in single stages and then combined interaction and atom composition clustering was performed to compare the efficiencies of a one-stage vs. a two-stage cluster analysis. Instead of performing compound structure clustering on the interaction clustering results, we will combine the two stages clustering together choosing 50 clusters for each stage. We consider the top 20 compounds as a suitable choice to retrieve active and possible novel compounds taking in consideration the biased energy based scoring methods of VS [18-20]. We also acknowledge some discrepancy among structures of similar compounds (not always perceived as similar by the structure clustering algorithm or the opposite, considered as similar when they are not).

## 4.5 RESULTS

### 4.5.1 Virtual Screening results

After VS was applied to the 2206 compound dataset containing the three actives, VS ranked the various compounds in the dataset based on their interaction energies. The three active compounds were ranked #575 (riboflavin), #591 (adenosine triphosphate) and #816 (calcitriol or vitamin D3) out of 2206 total compounds based on their binding energies (-84.81, -84.52 and -80.39) respectively (Table 6). The 2,206 total compounds obtained from ACD and FDA databases were divided into 50 interaction clusters and 50 structure clusters in order to obtain a suitable number of compounds in a cluster (not too many or too few compounds, unless they are unique; e.g. riboflavin).

Table 6. Virtual screening results and ranking of the three active compounds (riboflavin: 575, adenosine triphosphate: 591 and calcitriol: 816). The shaded area (bottom of table) shows the highest ranking compounds (1 – 4) based on interaction energies generated by the docking program.



		Compound	Energy	VDW	HBond	Elec	Int. ClusterID	Com. Cluster ID
574	<input checked="" type="checkbox"/>	cav2GJ5_VD3-MFCD00006628-...	-84.83	-49.31	-35.52	0	30	18
575	<input checked="" type="checkbox"/>	__cav2GJ5_VD3-riboflavin-0.pdb	-84.81	-52.96	-31.85	0	46	12
590	<input type="checkbox"/>	cav2GJ5_VD3-eletriptan-1.pdb	-84.53	-81.34	-3.19	0	29	13
591	<input checked="" type="checkbox"/>	__cav2GJ5_VD3-adenosine_trip...	-84.52	-61.55	-22.97	0	21	40
815	<input checked="" type="checkbox"/>	cav2GJ5_VD3-oxybuprocaine-0....	-80.41	-80.41	0	0	16	33
816	<input checked="" type="checkbox"/>	__cav2GJ5_VD3-calcitriol-1.pdb	-80.39	-63.74	-16.65	0	28	35
1	<input checked="" type="checkbox"/>	cav2GJ5_VD3-MFCD00012401-...	-139.19	-123.35	-15.84	0	21	13
2	<input checked="" type="checkbox"/>	__cav2GJ5_VD3-MFCD0001335...	-136.99	-101.42	-31.56	-4.01	46	13
3	<input checked="" type="checkbox"/>	__cav2GJ5_VD3-MFCD0001011...	-134.07	-112.66	-21.41	0	13	13
4	<input checked="" type="checkbox"/>	__cav2GJ5_VD3-MFCD0000977...	-133.41	-97.22	-36.19	0	10	13

The top four ranking unknown compounds (Table 6 green shade) were ranked by VS based on their binding energies (-139.19, -136.99, -134.07 and -133.41). They are termed “unknowns” because at present time, their complex to  $\beta$ -LG has no known function. Therefore, they will be further investigated by TSCC and other criteria (e.g. conserved residue activity) to confirm whether they might be novel compounds able to bind  $\beta$ -LG and carry on particular functions or may just happen to be false positive hits as it is expected with many compounds obtained from VS.

The three active compounds showed hydrogen bonding activity in the LEU 39 and VAL 41 residues of the  $\beta$ -LG calyx (Fig. 19) which were also confirmed in previous bioassays as important binding residues in the calyx of the  $\beta$ -LG. These important residues can further aid in validating TSCC and either support or reject the VS and post screening analysis results.

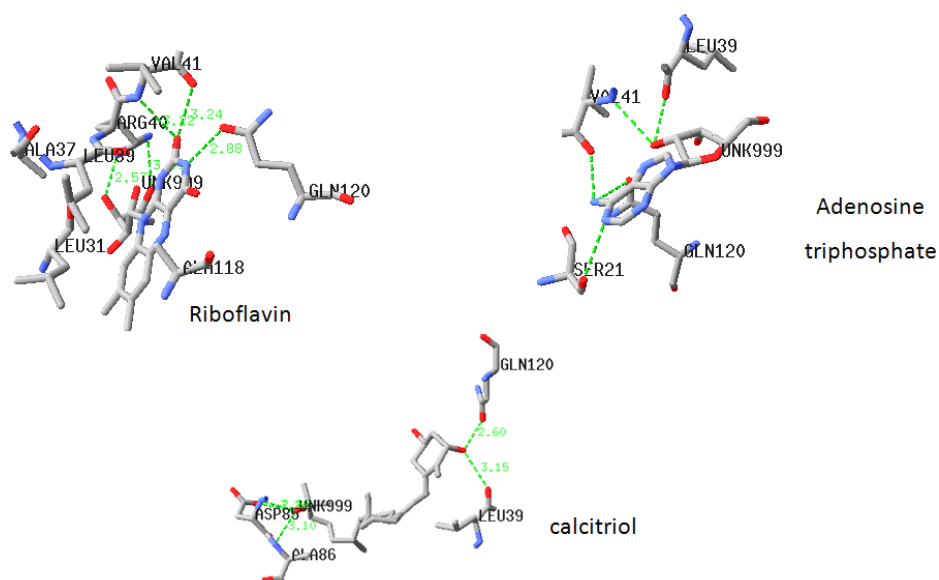


Figure 19. Conserved residues (LEU 39 and VAL 41) showing interaction through hydrogen bonding between  $\beta$ -LG cavity and the three active compounds.

The dendrogram of interaction energies shows the activity of amino acid groups in important residues of compounds being an additional source of confirming the binding of a compound to the target cavity (Fig. 20). The green color blocks indicate activity or contact between the conserved residues of the target and its docked compound. The unknown compound MFCD00012401 inside the yellow box shows no bonding activity (black blocks) in any of the conserved residues. This is evidence that it may not be able to properly bind the  $\beta$ -LG cavity and

form either a functional complex or be able to inhibit it. The other compounds in Figure 20 show fair amount of bonding activity in the conserved regions and three of them (active compounds) are known from previous experiments to bind at least some of these conserved regions (LEU 39, VAL 41). Compound MFCD00010114 is shown to have the most activity with the conserved residues and we can use the TSCC to verify whether these findings stand their ground.

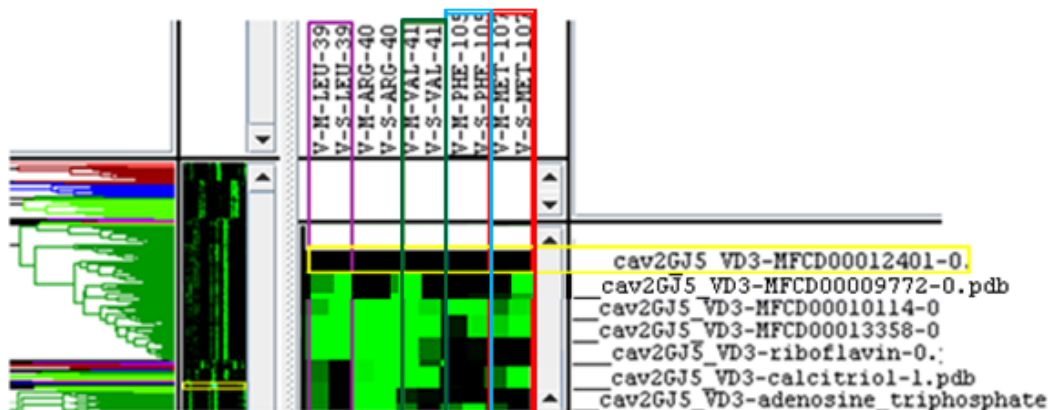


Figure 20. The molecular docking dendrogram showing the occurrence of conserved (important) residues between the three active compounds (riboflavin, vitamin D3 / calcitriol and adenosine triphosphate) and  $\beta$ -LG cavity and also the four highest ranking compounds (MFCD00012401, MFCD00009772, MFCD00010114, MFCD00013358) and  $\beta$ -LG cavity.

#### 4.5.2 Cluster Analysis Results

One-Stage cluster analyses for both interaction energy profiles and atom composition of compounds were done on the entire 2206 compound dataset from ACD and FDA databases which include the three active compounds to investigate the performance of both one-stage clustering methods and our TSCC method. We particularly paid close attention to the three active compounds and the top ranking four unknown compounds identified by molecular docking. We describe the results obtained based on the criteria used and deductive analysis.

##### Riboflavin

This is the highest ranking active compound by our TSCC method and it also ranked the highest among actives with one-stage IC. Its activity with conserved residues is most likely the best reason why it obtained a high rank by both TSCC and one-stage IC. Its atomic composition

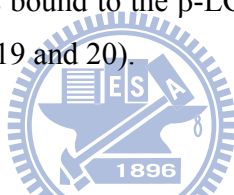
also favors it to fit the cavity of  $\beta$ -LG better than the other actives which gave it a fair ranking (#32 Fig. 21) on one-stage AC.

### Adenosine triphosphate (ATP)

Although ATP is ranked #18 by TSCC it is reasonable considering the large dataset of 2206 compounds in which almost 600 compounds had better energy rankings than ATP (#591 Energy Rank, Fig. 21). Our TSCC was able to retrieve and rank this compound (Fig. 21) accordingly based on its activity in the conserved residues (Figs. 19 and 20).

### Vitamin D3 (Calcitriol)

The low rank of this active compound (#1154 by one-stage AC Fig. 21) is expected because vitamin D3 requires significant reposition of the calyx in order for it to insert into the cavity of  $\beta$ -LG calyx [66]. We therefore have to consider more than the ranking of AC and rely on confirmed findings from bioassays [66] and consider the additional criteria such as the conserved residues once vitamin D3 is bound to the  $\beta$ -LG cavity where some good activity with the target calyx of  $\beta$ -LG is seen (Figs. 19 and 20).



AC-IC-50-CTL20110601 (5)				Two-stage	one-stage	one-stage	Virtual screening
	A	B	C	GY	GZ	HA	HB
1	I.C. Cl ID	A.C. Cl ID	#Compound	Rank-IC-AC50	Rank-IC	Rank-AC	Rank-Energy
2	21	13	cav2GJ5_VD3-MFCD00012401-0.pdb	1273	1248	112	1
3	46	13	cav2GJ5_VD3-MFCD00013358-0.pdb	4	3	113	2
4	13	13	cav2GJ5_VD3-MFCD00010114-0.pdb	1	1	114	3
5	10	13	cav2GJ5_VD3-MFCD00009772-0.pdb	2	2	115	4
6	46	12	cav2GJ5_VD3-riboflavin-0.pdb	3	4	32	575
7	21	40	cav2GJ5_VD3-adenosine_triphosphate-0.pdb	18	35	96	591
8	28	35	cav2GJ5_VD3-calcitriol-1.pdb	12	12	1154	816

Figure 21. Clustering analysis results. The ranking results from TSCC, two methods of one-stage clustering (Rank-IC and Rank-AC) and Virtual Screening for the three active compounds and four unknowns are shown in the four separate columns.

### MFCD00012401

This is the highest ranking compound based on VS total energy. However, the very low ranking (1248 out of 2206) it obtained from Interaction Clustering (IC) depicted in Figure 21

affirms the results shown in the dendrogram (Fig. 20 yellow box) which indicate no activity from this compound within conserved (important) residues of the target. Somewhat surprising, the one-stage atom composition (AC) clustering ranked compound MFCD00012401 better than the other unknowns (MFCD00013358, MFCD00010114, MFCD00009772) and better than one of the active compounds (calcitriol or vitamin D3). This can be explained by the fact that this compound has some similar atoms and some structure similarity with one of the active compounds (Fig. 22) and their molecular weight is also similar (541.62 D and 551.14 D). However, when our TSCC method was used (IC-AC50 Fig. 21) it ranked this compound quite reasonable (1273 out of 2206 compounds) according to the findings shown in the dendrogram (no activity in the conserved residues, Fig. 20) and classified it as non-active compound.

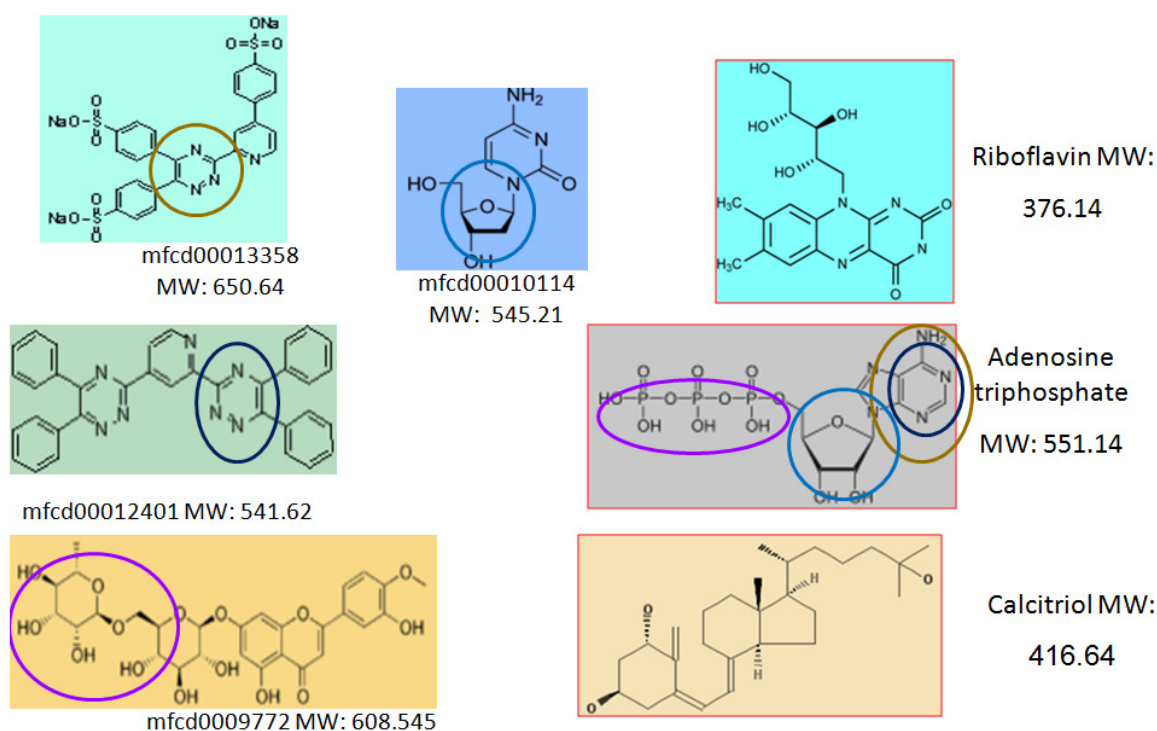


Figure 22. The three active compounds and four highest VS ranking unknown compounds; their structures, molecular weight and atom composition. Atom similarity: adenosine triphosphate and unknown mfcid00013358 (brown circles), adenosine triphosphate and mfcid00010114 (light blue circles), adenosine triphosphate and mfcid00012401 (dark blue circles) and adenosine triphosphate and mfcid00013358 (purple circles).

### **MFCD00013358**

Although it has the highest molecular weight of all compounds selected (Fig. 22) which may account for its energy rank (Table 6), this compound shows a fair amount of activity in the conserved residues (Fig. 20). It also has some atomic similarity to riboflavin and was ranked #113 by one-stage AC clustering. Our TSCC method ranked it 4 (Fig. 21), one rank lower than one-stage IC and this is likely due to the combined clustering using interaction and AC profiles.

### **MFCD00010114**

This compound was ranked 1<sup>st</sup> by both our TSCC and one-stage IC and it is not surprising considering that it had the most activity in the conserved residues (Fig 20.) Its rank by one-stage AC is also adequate since it does have some atomic similarity with riboflavin (Fig. 22) which ranks # 96 (one-stage AC clustering Fig. 20).

### **MFCD00009772**

This compound ranked #2 on both TSCC and one-stage IC, most likely because of its good activity in conserved residues (Fig. 20). Its one-stage AC rank is #115 and it shows some similarity in its atom composition and structure with both calcitriol and ATP (Fig. 22).

## **4.6 Discussion**

We demonstrated that when a one-stage clustering analysis was used on the three known actives at least one compound was not identified as a true hit (ranked within the top 20) by both IC and AC (one-stage clustering). IC ranked ATP at #35 well below the 20 top compound we selected as a suitable limit for identifying more accurate positive hits. We know from previous studies our active compounds are able to form a complex with  $\beta$ -LG and setting the limit lower than the top ranking 20 compounds would have classified ATP as a non-active compound. We therefore conclude that the top 20 ranking compounds present on the list obtained from TSCC post screening analysis is a good choice when selective the active compounds and eliminating the non-actives. TSCC successfully ranked the three active compounds used in this study within the top 20 rankings despite their low VS standings (575, 591 and 816 out of 2206 respectively). It also clearly rejected compound MFCD00012401 from the list of actives although it was ranked highest by VS. This confirms that TSCC can considerably improve VS enrichment and propagate true hits generated from molecular docking to the top of the compound list. Our work also



confirmed that virtual screening and TSCC can be successfully employed to mine and study compounds used in other applications besides drug design. To further investigate the three top ranking compounds for more clues regarding their ability to bind the  $\beta$ -LG cavity and their possible functions when complexed with  $\beta$ -LG, bioassays can be performed at this stage for unknowns MFCD00013358, MFCD00010114 and MFCD00009772.

#### 4.7 Summary

In search of an improved method for retrieving and post analysing protein-ligand complexes we developed a combinative clustering method using two clustering stages to mine and visualize compound candidates generated by virtual screening. Six classes of targets and three different data sets were used to validate and thoroughly investigate this method. Our TSCC method encodes additional interaction-specific information into the real number string, hydrogen bond, van der Waal and electrostatic forces in comparison to other post screening analyses. These interaction energies are important features of interaction profiles due to their significance in receptor-ligand binding and the efficiency of first stage clustering (protein-ligand interaction based clustering). The structure clustering stage can use various features; physico-chemical features (sections 4.2, 4.3) or atom composition ( $\beta$ -LG study, sections 4.5, 4.6). The final representatives can be retrieved either after second-stage clustering is performed on first-stage results (sections 4.1 – 4.3) or after combined interaction and compound structure (AC) clustering (sections 4.4 – 4.6) depending on the scope of the study.

VISCANA, [22] a one-stage post screening analysis uses protein-ligand interactions but lacks sufficient descriptions of van der Waals forces and hydrogen bond interactions as pointed out by Amari *et al.* in their study. Such descriptions play an important role in receptor-ligand binding and determine for the most part the success of a post screening method which uses interaction profiles. In addition, the use of a docking tool not specifically optimized for protein-ligand interactions during VS and the lack of an additional method (a second clustering stage) to eliminate additional false positive hits and to retrieve missed true hits are the downsides of this method. Furthermore, VISCANA has not identified novel compounds in any studies, likely to a combination of things; lack of a specifically optimized docking program, lack of sufficient descriptions of interaction energies and lack of additional screening (a second clustering stage).

SIFt, [23] another one-stage clustering method encodes more interaction-specific descriptions into real number using seven bits per binding-site residue to represent seven



different types of interactions which results in the encoding of a protein-ligand interaction into a binary string. SIFt represents a ligand by the interactions it forms in the binding site of a protein. This approach is ideal in post screening analysis because the more descriptions of binding interactions a method provides, an improved enrichment of VS will occur. However, Deng *et al.* does not indicate that novel compounds or inhibitors were identified using SIFt. They do show that EF (enrichment factor) varies using different docking tools for VS, confirming one of our conclusions that optimizing docking programs is also significant in the overall process of mining and visualization of biochemical compounds from databases.

Our goal was to develop a method for selecting adequate representative compounds from a database that could be suitable candidates for various applications. TSCC, with some modifications of its original version successfully identified inhibitors for influenza virus and flaviviruses [6, 8]. This is a major accomplishment in comparison to other post screening methods [21-23] which contributed mainly to visualization and enrichment of VS but were not particularly successful in novel compound discovery. In addition, through our studies of interaction profiles and  $\beta$ -LG we show that VS and post screening analysis may be successfully used in other applications besides drug design.

Employing GEMDOCK, a specifically optimized VS tool and two stages of combined cluster analysis is a much more efficient technique than done by other one-stage post screening analyses (SIFt, VISCANA) revealing why our TSCC method was able to identify inhibitors for dengue virus and flaviviruses [6, 8]. This study also shows that the overall performance of TSCC is due to sufficient descriptions of protein-ligand interactions when mining for ideal targets as well as the efficiency of *iGEMDOCK* in generating suitable interaction and atom composition matrixes which are important features in clustering biochemical compounds.

However, comparing TSCC to other clustering methods can be somewhat ambiguous since different approaches may vary in goals and purpose. Some post screening analyses are not aiming to identify novel compounds or inhibitors and NIPALSTREE [25] is one of such examples. In TSCC one of the main objectives is to select ideal representatives for novel compound design and discovery through the combination of an optimized docking tool and two clustering stages for. Overall our study confirms that a two-stage combined clustering method is superior to one-stage methods because of its two main advantages: 1) lower selection of non-active compounds and 2) an improvement in the selection of active compounds.

## Chapter 5

### Conclusion

#### 5.1 Summary

This research accomplished and demonstrated the following important aspects:

(1) It thoroughly investigated the significance of protein-ligand interaction and compound structure profiles in the mining of lead compounds using computer-aided methods and various interdisciplinary principles and showed that novel compounds can be mined and analyzed more efficiently. It revealed that the use of interaction and compound structure profiles may successfully be employed in studies of drug design and various other applications such as novel compound design for various kinds of industries [65].

(2) Post screening analysis methods (SIFt and VISCANA) were studied in detail and two main issues were pointed out: a) if a docking tool is used for VS, can a post screening analysis complement it? b) If either post screening analysis method (SIFt or VISCANA) is used, how can a docking program be optimized to complement the post screening analysis employed? Using non-optimized docking is likely to reduce the efficiency of VS since docking optimization is an important task as seen during the generation of interaction and compound structure profiles in GEMDOCK. In addition, post screening analyses which lack enough descriptions of interaction energies (VISCANA) and one-stage clustering methods in general encounter higher number of false positives and inefficient retrieval of active compounds as shown in the  $\beta$ -LG study (sections 4.5, 4.6).

3) We showed that by optimizing the docking program (GEMDOCK / *iGEMDOCK*) for VS and including two-stages of clustering in post screening analysis (TSCC) more accurate results in the mining and analysis of compounds can be obtained. We successfully retrieved the selected active compounds through TSCC (sections 4.5, 4.6) while both one-stage clustering methods (IC and AC Fig. 21) failed to rank all three active compounds within the top 20. In addition, our TSCC method was used with some modifications in other studies to successfully identify various inhibitors (influenza virus and flaviviruses inhibitors [6, 8]) revealing significant improvements over one-stage post screening methods.

## 5.2 Future works

Computer-aided methods for virtual screening and post screening analysis can be successfully used in identifying a wide range of protein-ligand complexes for a variety of biochemical applications. Thus, the following studies are of significance and particular interest in our future works:

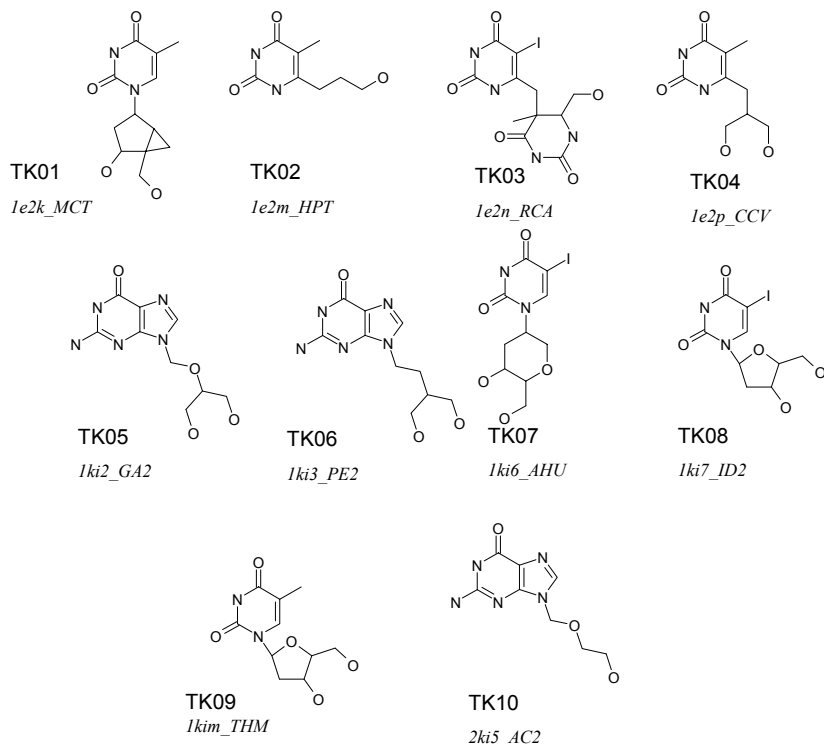
1) We want to investigate a new clustering technique, NeatMap [53], a non-clustering approach using microarray datasets instead of traditional clustered heat map for the possibility of improving accuracy and efficiency of ranking compounds and select more suitable representatives.

2) We aim to investigate targets for nutrition and skin care in future studies. These two areas deserve attention because they are an important part of human health and well being since both nutrition and skin are important defenses against pathogens of all kinds.

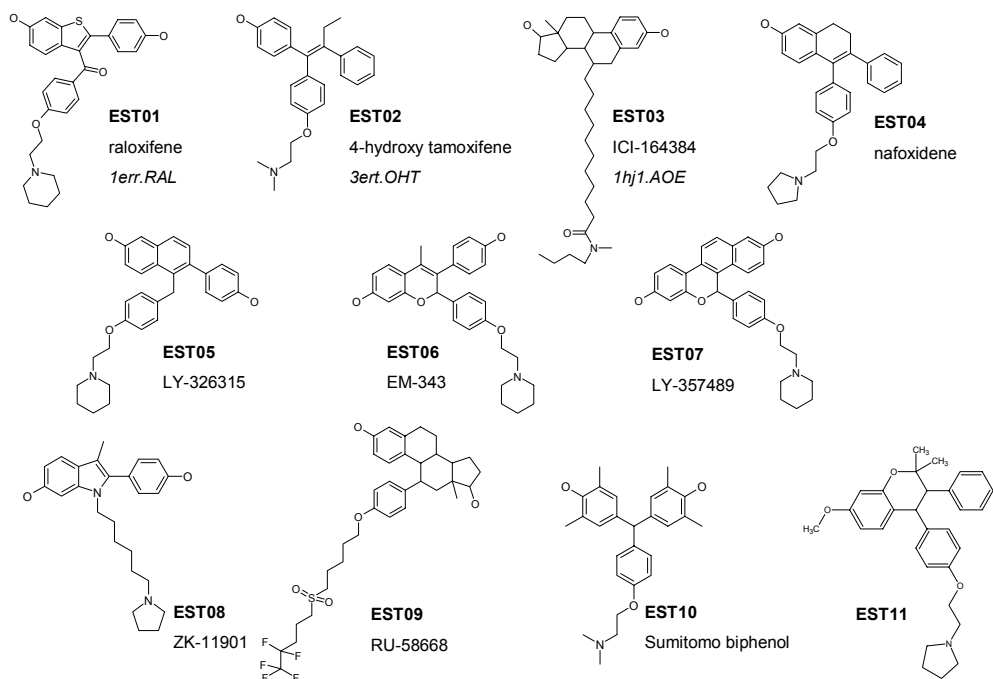
As concerns of pollution increase around the world, safer and more natural food products, fertilizers, pesticides and detergents are highly sought after. People in general are becoming more concerned with eating proper diets and maintaining a strong and healthy body therefore, new findings in nutrition, supplements and skin care are of particular interest. Therefore, computer-aided methods for providing the necessary means in identifying compounds used these areas will continue to grow and expand in the near future.

**Appendix A** (Obtained from our published study [48])

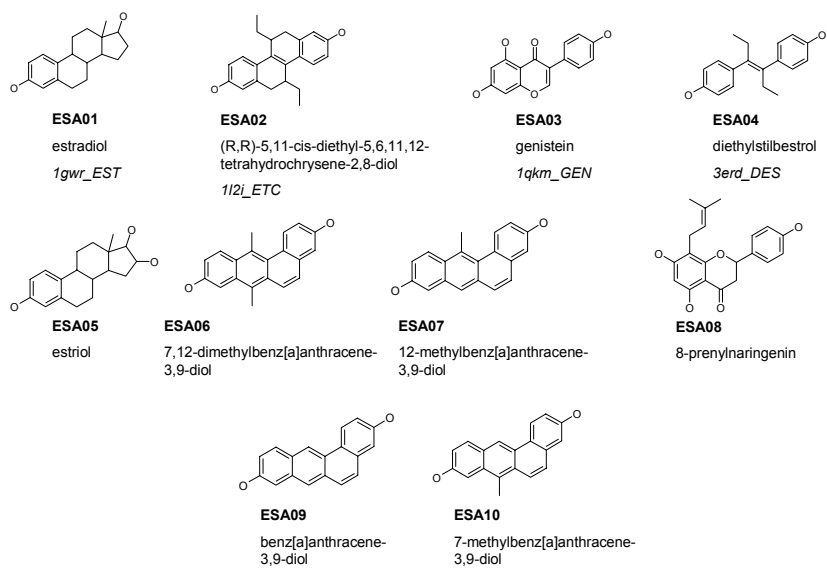
1) 61 Active compounds used in TSCC obtained from ACD and CMC public databases:



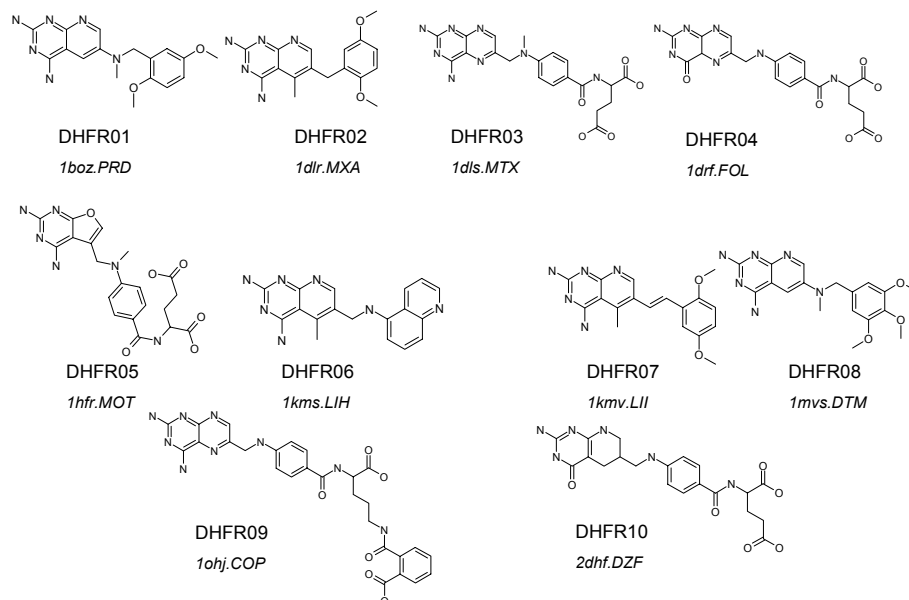
Ten TK (thymidine kinase) active compound structures.



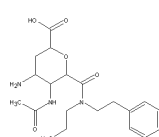
Eleven ER $\alpha$  (estrogen receptor) antagonist structures.



Ten ER $\alpha$  (estrogen receptor) agonist structures.

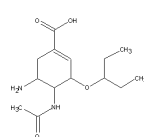


Ten hDHFR (human dihydrofolate reductase) active compound structures.



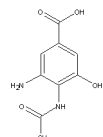
**NA01**

*lig1bjj\_G21*



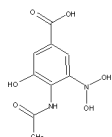
**NA20**

*lig2qwh\_G39*



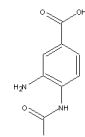
**NA10**

*lig1ivc\_ST2*



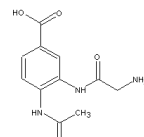
**NA11**

*lig1ivd\_ST1*



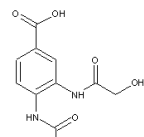
**NA12**

*liglive\_ST3*



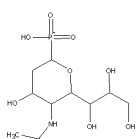
**NA06**

*lig1ina\_ST6*



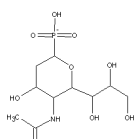
**NA07**

*ligling\_ST5*



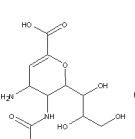
**NA08**

*liglinw\_AXP*



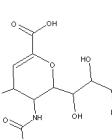
**NA09**

*lig1inx\_EQP*



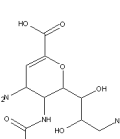
**NA03**

*lig1f8c\_4AM*



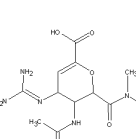
**NA04**

*lig1f8d\_9AM*



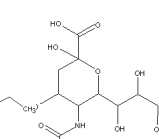
**NA05**

*lig1f8e\_49A*



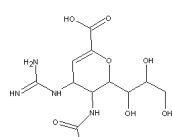
**NA18**

*lig2qwf\_G20*



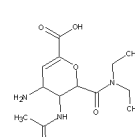
**NA14**

*lig1mwe\_SIA*



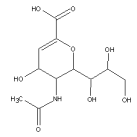
**NA15**

*lig1nnc\_GNA*



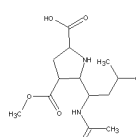
**NA19**

*lig2qwg\_G28*



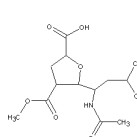
**NA02**

*lig1f8b\_DAN*



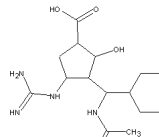
**NA16**

*lig1xoe\_ABX*



**NA17**

*lig1xog\_ABW*



**NA13**

*lig1l7f\_BCZ*



Twenty NA (neuraminidase) active compound structures.

## Appendix B

### List of Publications

**Clinciu, D. L.**, Yang, J. M., Lo, C. C., The Relevance of Interaction Profiles in Various Computer-Aided Novel Compound Design and Applications, *Journal of Current Bioinformatics*, 2011, vol 6, no 3, doi:1574-8936/11

**Clinciu, D. L.**, Chen, Y.F., Ko, C.N., Lo, C. C., and Yang, J.M., TSCC: Two-Stage Combinatorial Clustering for virtual screening using protein-ligand interactions and physicochemical features, *BMC Genomics*, 2010. doi:10.1186/1471-2164-11-S4-S26

**Clinciu, D. L.**, Chen, Y. L., Yang, M.C., Wallace, S., Yang, J. M., Mao, S. J. T., Vitamin D; Nutrition, Side Effects and Supplements, *Nova Science Publishers*, ISBN: 978-1-61728-601-8, 2010)

Wallace, S., Reed, A., **Clinciu, D. L.**, and Yu, H. C., A Comparison of the Usability of Heuristic Evaluations for Online Help, *Information Design Journal* (accepted, December 2010)

Khudaverdyan S., Dokholyan, J, Arustamyan V., Khudaverdyan, A., **Clinciu, D.L.**, Nuclear Instruments and Methods in Physics Research, *Elsevier*, 2009. A 610, 314–316



### Conference Paper

**Clinciu, D. L.**, Yang, J. M., Chen, Y.F, Ko, C. N, Lo, C. C., InCoB2010, The 9<sup>th</sup> International Conference on Bioinformatics, TSCC: Two-Stage Combinatorial Clustering for virtual screening using protein-ligand interactions and physicochemical features (Presented in Tokyo, Sep 27, 2010)



## REFERENCES

1. Frank, E., et al., Data mining in bioinformatics using Weka. *Bioinformatics*, 2004. **20**(15): p. 2479-2481.
2. Stahl, M. and T. Schulz-Gasch, Practical database screening with docking tools. Ernst Schering Res Found Workshop 2003. **42**: p. 24.
3. Bissantz, C., G. Folkers, and D. Rognan, Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *Journal of Medicinal Chemistry*, 2000. **43**(25): p. 4759-4767.
4. Joachimiak, A., High-throughput crystallography for structural genomics. *Current Opinion in Structural Biology*, 2009. **19**(5): p. 573-584.
5. Blundell, T.L., H. Jhoti, and C. Abell, High-throughput crystallography for lead discovery in drug design. *Nature Reviews Drug Discovery*, 2002. **1**(1): p. 45-54.
6. Yang, J.M., et al., Combinatorial computational approaches to identify tetracycline derivatives as flavivirus inhibitors. *PLoS ONE*, 2007. **2**(5): p. e428.
7. Chin, K.H., et al., The cAMP receptor-like protein CLP is a novel c-di-GMP receptor linking cell-cell signaling to virulence gene expression in *Xanthomonas campestris*. *J Mol Biol*, 2010. **396**(3): p. 646-62.
8. Hung, H.C., et al., Aurintricarboxylic acid inhibits influenza virus neuraminidase. *Antiviral Res*, 2009. **81**(2): p. 123-31.
9. Yang, M.C., et al., Rational design for crystallization of beta-lactoglobulin and vitamin D-3 complex: revealing a secondary binding site *Crystal Growth & Design*, 2008. **8**: p. 4268-4276.
10. Clinciu, D. L., et al. *Vitamin D: Nutrition, Side Effects and Supplements*, Nova Science Publishers, 2010. ISBN: 978-1-61728-601-8.
11. Stahl, M. and M. Rarey, Detailed analysis of scoring functions for virtual screening. *Journal of Medicinal Chemistry*, 2001. **44**(7): p. 1035-1042.
12. Pfeffer, P. and H. Gohlke, DrugScore(RNA) - Knowledge-based scoring function to predict RNA-ligand interactions. *Journal of Chemical Information and Modeling*, 2007. **47**(5): p. 1868-1876.
13. Weiner, S.J., et al., A New Force-Field for Molecular Mechanical Simulation of Nucleic-Acids and Proteins. *Journal of the American Chemical Society*, 1984. **106**(3): p. 765-784.
14. Gehlhaar, D.K., et al., Molecular recognition of the inhibitor AG-1343 BY HIV-1 Protease - Conformationally Flexible Docking by Evolutionary Programming. *Chemistry & Biology*, 1995. **2**(5): p. 317-324.
15. Charifson, P.S., et al., Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *Journal of Medicinal Chemistry*, 1999. **42**(25): p. 5100-5109.
16. Verdonk, M.L., et al., Virtual screening using protein-ligand docking: Avoiding artificial enrichment. *Journal of Chemical Information and Computer Sciences*, 2004. **44**(3): p. 793-806.
17. Fradera, X., R.M.A. Knegtel, and J. Mestres, Similarity-driven flexible ligand docking. *Proteins-Structure Function and Bioinformatics*, 2000. **40**(4): p. 623-636.
18. Ewing, T.J.A., et al., DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *Journal of Computer-Aided Molecular Design*, 2001. **15**(5): p. 411-428.

19. Ewing, T.J.A. and I.D. Kuntz, Critical evaluation of search algorithms for automated molecular docking and database screening. *Journal of Computational Chemistry*, 1997. **18**(9): p. 1175-1189.
20. Jones G, Willett P, Glen RC, Leach AR, Taylor R: Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology* **1997**; *267*: 727-748.
21. Kroemer, R.T., et al., Assessment of docking poses: Interactions-based accuracy classification (IBAC) versus crystal structure deviations. *Journal of Chemical Information and Computer Sciences*, 2004. **44**(3): p. 871-881.
22. Amari, S., et al., VISCANA: visualized cluster analysis of protein-ligand interaction based on the ab initio fragment molecular orbital method for virtual ligand screening. *Journal of Chemical Information and Modeling*, 2006. **46**(1): p. 221-30.
23. Deng, Z., C. Chuaqui, and J. Singh, Structural interaction fingerprint (SIFt): A novel method for analyzing three-dimensional protein-ligand binding interactions. *Journal of Medicinal Chemistry*, 2004. **47**: p. 337-344.
24. Nakano, T., et al., Fragment molecular orbital method: use of approximate electrostatic potential. *Chemical Physics Letters*, 2002. **351**(5-6): p. 475-480.
25. Bocker, A., G. Schneider, and A. Teckentrup, NIPALSTREE: A new hierarchical clustering approach for large compound libraries and its application to virtual screening. *Journal of Chemical Information and Modeling*, 2006. **46**(6): p. 2220-2229.
26. Yang, J.M. and C.C. Chen, GEMDOCK: A generic evolutionary method for molecular docking. *Proteins-Structure Function and Bioinformatics*, 2004. **55**(2): p. 288-304.
27. Shin, J.M. and D.H. Cho, PDB-ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures. *Nucleic Acids Research*, 2005. **33**: p. D238-D241.
28. Nuzzo, A. and A. Riva, Genephony: a knowledge management tool for genome-wide research. *Bmc Bioinformatics*, 2009. **10**.
29. Jerajani, H.R., Mizoguchi, H., Li J, The effects of a daily facial lotion containing vitamins B3 and E and provitamin B5 on the facial skin of Indian women, *Indian Journal of Dermatology*, 2010. **76** (1): p. 20-26
30. Revollo, J.R., Grimm, A. A., Imai, S, The NAD Biosynthesis Pathway Mediated by Nicotinamide, *Journal of Biological Chemistry*, 2004. **279** (4): p. 50754-50763
31. Singh, A., Casey, K.D., King, W.D., Efficacy of urease inhibitor to reduce ammonia emission from poultry house, *Journal of Applied Poultry Research*, 2009. **18** (1): 34-42
32. Rarey, M., et al., A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology*, 1996. **261**(3): p. 470-489.
33. Jones, G., et al., Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology*, 1997. **267**(3): p. 727-748.
34. Abagyan, R., M. Totrov, and D. Kuznetsov, Icm - a New Method for Protein Modeling and Design - Applications to Docking and Structure Prediction from the Distorted Native Conformation. *Journal of Computational Chemistry*, 1994. **15**(5): p. 488-506.
35. Venkatachalam, C.M., et al., LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *Journal of Molecular Graphics & Modelling*, 2003. **21**(4): p. 289-307.
36. Kuntz, I.D., Structure-Based Strategies for Drug Design and Discovery. *Science*, 1992. **257**(5073): p. 1078-1082.

37. Lorber, D.M. and B.K. Shoichet, Flexible ligand docking using conformational ensembles. *Protein Science*, 1998. **7**(4): p. 938-950.
38. Willett, P., J.M. Barnard, and G.M. Downs, Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, 1998. **38**(6): p. 983-996.
39. Leigh, D.A., Summing up ligand binding interactions. *Chemistry & Biology*, 2003. **10**(12): p. 1143-1144.
40. Chen, K., Kurgan, L., Investigation of Atomic Level Patterns in Protein—Small Ligand Interactions, *PLoS One*, 2009. **4** (2): e4473
41. Wyss, P.C., et al., Novel dihydrofolate reductase inhibitors. Structure-based versus diversity-based library design and high-throughput synthesis and screening. *J Med Chem*, 2003. **46**: p. 2304-2312.
42. deWolf, F. A., and Brett, G. M., Ligand-Binding Proteins: Their Potential for Application in Systems for Controlled Delivery and Uptake of Ligands, *Pharmacological Reviews*, 2000. **52** (2): 207-236
43. Hsieh, R.W. et al. Identification of Ligands with Bicyclic Scaffolds Provides Insights into Mechanisms of Estrogen Receptor Subtype Selectivity, *Journal of Biological Chemistry*, 2006. **281** (26): 17909-17919
44. Champness, J.N., et al., Exploring the active site of herpes simplex virus type-1 thymidine kinase by X-ray crystallography of complexes with aciclovir and other ligands. *Proteins-Structure Function and Genetics*, 1998. **32**(3): p. 350-361.
45. Yang, J.M., et al., Consensus scoring criteria for improving enrichment in virtual screening. *Journal of Chemical Information and Modeling*, 2005. **45**(4): p. 1134-1146.
46. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ., Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry* 1998; **19**: 1639-1662.
47. Mitchell, P., A perspective on protein microarrays. *Nature Biotechnology*, 2002. **20**(3): 225-229.
48. Clinciu, D. L., et al. TSCC: Two-Stage Combinatorial Clustering for virtual screening using protein-ligand interactions and physico-chemical features, *BMC Genomics*, 2010. doi:10.1186/1471-2164-11-S4-S26
49. Yang, J.M. and T.W. Shen, A pharmacophore-based evolutionary approach for screening selective estrogen receptor modulators. *Proteins*, 2005. **59**(2): p. 205-20.
50. Yang, J.M., et al., Consensus scoring criteria for improving enrichment in virtual screening. *J Chem Inf Model*, 2005. **45**(4): p. 1134-46.
51. Pearlman DA, Charifson PS. Improved scoring of ligand-protein interactions using OWFEG free energy grids. *J Med Che* 2001; **44**: 502-511.
52. Pan Y, Huang N, Cho S, MacKerell AD, Jr. Consideration of molecular weight during compound selection in virtual target-based database screening. *Journal of Chemical Information and Computer Science* 2003; **43**: 267-272.
53. Rajaram, S. and Y. Oono, NeatMap - non-clustering heat map alternatives in R. *Bmc Bioinformatics*. 2010; **11**.
54. Matter, H., Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors. *Journal of Medicinal Chemistry*, 1997. **40**(8): p. 1219-1229.

55. Ruvinsky, A.M., Role of binding entropy in the refinement of protein-ligand docking predictions: Analysis based on the use of 11 scoring functions. *Journal of Computational Chemistry*, 2007. **28**(8): p. 1364-1372.
56. Liu, Q., et al., RNACluster: An integrated tool for RNA secondary structure comparison and clustering. *Journal of Computational Chemistry*, 2008. **29**(9): p. 1517-1526.
57. Zheng, W.F. and A. Tropsha, Novel variable selection quantitative structure-property relationship approach based on the k-nearest-neighbor principle. *Journal of Chemical Information and Computer Sciences*, 2000. **40**(1): p. 185-194.
58. Carhart, R.E., D.H. Smith, and R. Venkataraghavan, Atom Pairs as Molecular -Features in Structure Activity Studies – Definitions and Applications *Journal of Chemical Information and Computer Sciences*, 1985. **25**(2): p. 64-73.
59. Dubes, R. and A.K. Jain, Clustering methodologies in exploratory data analysis. *Adv Comput*, 1980. **19**: p. 113-228.
60. Gluck, O. and Maricic, M. Raloxifene: Recent information on skeletal and non-skeletal effects. *Current Opinion in Rheumatology*, 2002. **14**(4): p. 429-432.
61. Cody, V. et al. Comparison of ternary crystal complexes of F31 variants of human dihydrofolate reductase with NADPH and a classical antitumor fuopyrimidine. *Anti-cancer Drug Design*, 1998. **13**(4): p. 8.
62. Verma, R.P. and C. Hansch, A QSAR study on influenza neuraminidase inhibitors. *Bioorganic & Medicinal Chemistry*, 2006. **14**(4): p. 982-996.
63. Yang, J.M. and T.W. Shen, A pharmacophore-based evolutionary approach for screening selective estrogen receptor modulators. *Proteins-Structure Function and Bioinformatics*, 2005. **59**(2): p. 205-220.
64. Birch, L., et al., Sensitivity of molecular docking to induced fit effects in influenza virus neuraminidase. *Journal of Computer-Aided Molecular Design*, 2002. **16**(12): p. 855-869.
65. Clinciu, D. L., et al. The Relevance of Interaction Profiles in Various Computer-Aided Novel Compound Design and Applications, *Journal of Current Bioinformatics*, 2011, vol 6, no 3, doi:1574-8936/11
66. Yang, M. C. et al. Crystal structure of a secondary vitamin D3 binding site of milk b-lactoglobulin, *Proteins*, 2008; 71:1197-1210
67. Yang, M. C. et al. Evidence for b-lactoglobulin involvement in vitamin D transport in vivo – role of the c-turn (Leu-Pro-Met) of b-lactoglobulin in vitamin D binding. *FEBS Journal*, 2009; 276: 2251–2265
68. Chevalier, F. et al. Scavenging of free radicals, antimicrobial, and cytotoxic activities of the Maillard reaction products of beta-lactoglobulin glycosylated with several sugars. *J Agric Food Chem* 2001; **49**:5031–5038.
69. Nagaoka, S. et al. Identification of novel hypocholesterolemic peptides derived from bovine milk beta-lactoglobulin. *Biochem Biophys Res Commun* 2001; **281**:11–17.