Chapter 1

Introduction

1.1 Motivation and purposes

Molecular docking is an important technique in structure-based drug design[1], providing the prediction of a ligand conformation and orientation relative to the active site of a target protein. In recent years, the interest in molecular docking increases with an increasing number of protein crystal structures available and many works make progress in this area.

A computational docking process identifies the lead compounds by optimizing the energy of intermolecular interactions. Using a computer method to solve the protein-ligand docking problem involves two important elements: an efficient searching algorithm and a good scoring function. A good scoring function is capable of predicting correctly binding modes from a large number of potential solutions simply and rapidly. Various scoring functions have been developed for calculating the free energy of binding, embracing knowledge-based[2,3], empirical, physics-based[4], and solvent-based scoring functions[5]. Generally, these scoring functions are often complicated and demonstrate a rugged funnel shape[6].

There are many different approaches to solving the docking problem. Early docking methods treat both the ligand and protein as rigid, such as DOCK[1]. However, such a method is not able to make allowance for different ligand binding conformations and activities. Hence, the later approaches treat ligands as flexible to overcome this problem. With flexible ligand conformations considered, the size of conformation searching space becomes enormous and highly efficient search procedures are required. Stochastic sampling techniques include genetic algorithms[4,7-9], simulated annealing[10], evolutionary programming[4] and other algorithms[11]. In order to handle flexibility problems on protein, these docking methods usually make allowance for limited variation on proteins like side-chain flexibility or small motions of binding site loops[12-14]. Most of previous docking methods were evaluated on small test sets which are no

more than 20 complexes except GOLD[9] and FlexX[15].

In our previous works, GEMDOCK has been validated on 100 protein-ligand complexes with successful rate of 79%[16]. In this study we further evaluated GEMDOCK on large diverse test set which is included 305 protein-ligand complexes[17] and make use of the ability of GEMDCOK for virtual screening. Virtual screening encompasses many different computational techniques together and the aim of virtual screening is to reduce the size of a virtual compound collection to a possibly manageable size[18]. To apply a molecular docking tool to virtual screening the scoring function must have the ability to identify correctly docking conformations and prioritize potential hits. In order to achieve such aim, we developed pharmacophore-based scoring functions[19,20] which involved ligand preferences and binding-site pharmacophore and a novel concept of consensus ranking scores which combined different orders from different method for improving the screening performance.

1.2 Related Works

Molecular docking techniques are important in structure-based drug design and there are many docking tools available in this field now. These methods developed base on different computational algorithms and scoring functions. Generally, the workflow of a virtual screening run against a specific target is described as follow, database and target preparation, molecular docking, post-analysis[18]. First, database and target preparation is the initial stage for virtual screening. When a specific target is decided, we have to recognition its binding site region and prepare the ligand set to be screened. The source of compound databases includes physically available collection and combinatorial chemistry database. The well-known database of available compounds is Available Chemicals Directory (ACD). This database contains over twenty hundred thousand available chemical compounds which could be used for screening. A molecular docking tool includes two important elements: an efficient searching technique and a powerful scoring function. In early stages docking programs treated ligand and protein as rigid, such as DOCK[1]. DOCK used incremental build searching and physical-based scoring function for solving docking problems, but it didn't consider the relations of different binding modes and binding affinities. Thus efficient searching algorithms were made use of solving the problem of

ligand flexibility. Genetic algorithms[4,7-9] and other algorithms[10,11] are several powerful stochastic sampling techniques and adopted by many famous docking programs like GOLD, LigandFit, Glide. Taking GOLD as example, it uses genetic algorithm for flexible lignad docking and includes rotational flexibility for selected receptor hydrogens along with full ligand flexibility. Another aspect of the docking problem is that the flexibility of protein target has to discuss. Most of docking tools currently make the assumption that crystal structure of protein target is fixed. This is usually incorrect in real problem but necessary for docking approximation because accurately sampling the flexibility of binding site would increase complexity and computational cost. There are some programs attempting to take protein flexibility into account in certain degree, such as Slide, ICM. Slide and ICM allow the motion and relaxation of the side chain in binding site for replying the docking ligand conformation. After generating multiple ligand conformations and orientations, there should be a scoring function for deciding correct solutions for target. Scoring functions can be roughly grouped as physical-based, empirical and knowledge-based scoring functions[18]. Physical-based scoring functions are developed based on atomic force fields, like Amber and CHARMM. These force fields are generally accurate at estimation of binding free energies with adopting free energy perturbation or thermodynamic integration methods. Empirical-based scoring functions base on physical-chemical properties like hydrogen-bond counts, electrostatic interaction counts or other terms and estimate the binding free energy via additive approximation. Our program, GEMDOCK, just uses empirical-based scoring function and it takes hydrogen-bond, electrostatic and van der Waal interactions into consideration when docking ligands into protein target. A knowledge-based scoring function is derived from statistic mechanism of protein-ligand crystal complexes. The binding free energy is given by a sum of potential of mean forces which calculate interatomic interactions by their distances and contact frequencies with statistical mechanism method. The Knowledge-based scoring function is fast to compute and accurate on standard test set. Comparing to empirical-based scoring functions, knowledge-based ones need structure data to develop statistic model and are easily limited by a paucity of suitable information.

A molecular docking screen on a specific target yields enormous data of docking solutions. Consequently, an accurate and effective analysis method is needed eagerly. Consensus scoring is now on most publicly adopted strategy. The procedure of consensus scoring is to combine the screening results of different methods through their scorings. Generally, the combination result of three or four scorings can reach an acceptable performance. The meaning of consensus scoring is

that combination of different scoring functions can reduces the bias in individual ones. Consensus scoring is work on virtual screening analysis and has ability to improve the enrichment and hit rates.

1.3 Thesis Overview

This thesis is organized as follow. Chapter 2 introduces the scoring function, mining pharmacological consensus and the search algorithm of GEMDOCK. The scoring function consists of a simple empirical binding score and a pharmacophore-based score to reduce the number of false positives. GEMDOCK evolves the binding site pharmacological consensus and ligand preferences from both known active ligands and the target protein to improve screening accuracy. The search algorithm of GEMDOCK is a generic evolutionary method[16]. The core idea of our evolutionary approach was to design multiple operators that cooperate using the family competition model, which is similar to a local search procedure and the detail is described in this chapter.

Chapter 3 consists of validation and application of GEMDOCK. We evaluate GEMDOCK on CCDC/Astex test set which is a set with 305 divergence protein-ligand complexes. We test GEMDOCK on this test set for benchmark with popular docking tool, GOLD and analyze the characteristics of GEMDOCK. Then we apply GEMDOCK to virtual screening on human α-thrombin. In this screening test we prepare the test set for thrombin from databases and we mining the pharmacological preference of thrombin. Finally we analyze the accuracy of screening and make comparison with GOLD.

Chapter 4 discusses the application of data fusion in virtual screening. Data fusion is widely used in information retrieval domain and we utilize the technique for improving the performance of virtual screening. The analysis carries out on three pharmacological interest targets: thymidine kinase, dihydrofolate reductase and estrogen receptor. We develop this unique combination for scoring methods in virtual screening and analyze the advantages and disadvantages of this method.

Chapter 5 is conclusion. We summarize our works and contributions in this thesis. We drew

our conclusion for validation of GEMDOCK, application on virtual screening and application of data fusion. In addition we also show the future work on this study at last.