CHAPTER 2

LITERATURE REVIEW

Literature review

The vast majority of proteins binding to specifically other proteins are frequently performing their functions including for example cellular communication, gene regulation, and immune response. Moreover, the binding protein aims to disrupt the other protein-protein interaction that can block the various functions. Therefore, studies on protein-protein interface create understanding clearly protein functions and designing molecule to block or alter protein-protein interactions. Many studies of protein interfaces suggested that the free energy of interface region is not homogenous distribution; instead, a small set of critical residues called "hot spot" at the interface contribute significantly to the binding free energy [57, 58]. Moreover, Bogan & Thorn [58] found that the surrounded residues of hot spots are energetically unimportant contacts in all of the different protein interfaces. Experimentally, alanine scanning mutagenesis has been used to detect and analyze the hot spots [57, 58, 68, 79, 98-100]. Hot spots are typically defined as the change in binding free energy which $\Delta\Delta G \geq 2$ kcal/mol upon mutating it to an alanine. However, because of limitations in experimental information, the computational methods have been needed to identify hot spots.

Several computational methods have been developed for identifying the binding site in protein-protein complex. The number of method can be divided into three classes: (1) using sequence information, (2) estimating energetic contribution of alanine-scanning mutation, and (3) using information of structure. The first class used the sequence as the starting point, for example, Bock and Gough (2001) [101] used a Support Vector Machine (SVM) learning system with database of known protein interaction. Shulman et al. (2008) [102] detected the interacted residues based on multiple alignments of protein-protein interfaces which recognize the spatially conserved physic-chemical interaction

The second class is determinant the changed energy ($\Delta\Delta G_{bind}$) of alanine substitution. For instance, Kortemme et al. (2002) [84] analyzed the binding site in the method of determining a free energy function. They investigated Lennard Jones interaction, solvation interaction, and hydrogen bonding in term of free energy to calculate in 19 complexes with known crystal structures and experimentally measured changes in binding energy on alanine mutagenesis. The hot spots were residues showing a change in the binding free energy ($\Delta\Delta G_{bind}$) by less or more than 1 kcal/mol when replaced by alanine. The results shown that 84% of the cases a small effect of alanine replacement on the binding energy ($\Delta\Delta G_{bind} \geq 1$ kcal/mol) was correctly predicted, whereas 69% of hot spots were identified. In many methods, the initial step was performed with Molecular dynamic (MD). For example, Huo et al. (2002) [103] applied MM-PBSA (Molecular Mechanics-Poisson_Boltzmann surface area) with human Growth Hormone (hGH) to assess the computational alanine scanning method. They extracted the experimental $\Delta\Delta G_{binding}$ with the average unsigned error ~ 1 kcal/mol [104] for alanine mutation. The results showed that the

alanine mutant complex in a separate trajectory had the rational agreement with experimental data. Rajamani et al. (2004) [105] performed MD simulations in explicit solvent in 11 differently individually crystallized (unbound) proteins that conformation of anchoring side chains similar to bond complex. They found that the conformation of anchoring side chains in the absence of their interacting partners is similar to in the bound complex. Moreover, these anchores may perform as hot spot. Gonalez et al. (2006) [106] used energetic effect of a training database of 339 mutants and a blind test of 625 mutants of protein-protein complex. The results of calculated $\Delta\Delta G$ compared to the experimental $\Delta\Delta G$ showed that the correlation and standard deviation of training and blind set are 0.81 and 0.75 kcal/mol as well as 0.80 and 0.84 kcal/mol respectively. However, because of the high computational cost (parameter selection, dataprocessing, etc.), it is difficult for users not familiar with this field.

Structure-based method, the third class, the general method of this approach uses the information on the architecture and chemistry of protein-protein interface such as the size and shape of interface, the presence of bound water molecule, the number of hydrogen bond, and the identity residue in contact. For instance, Jones et al. (1997) [59] used patch analysis of six surface properties (solvation potential, residue interface propensity, hydrophobicity, planarity, protrusion and accessible surface area - ASA) to predict the interaction site. They defined "patch" as a central surface accessible residue and *n* nearest neighbours, where *n* was a variable. All surface patches were generated and then calculated the six parameters for each patch. The prediction algorithm consisted of three steps: (1) scoring the individual parameter, (2) calculating the combined parameter scores, (3) selection of best patches. The three patches with the highest combined scores were selected as best patches. The 59

complexes were predicted involving 28 homo-complexes,. 11 heterro-complexes, 6 large protomers, and 14 small protomers. The results shown that the 39 out of 59 complexes (66%) defining as correct prediction. Some cases were unsuccessful because the size of patch used was either too large or too small.

Gao et al. (2004) [107] made grids around the binding interface and used non-covalent-interaction probes, hydrogen bonds and hydrophobic characteristics, to explore the hot spots. The dataset of 13 complexes with 250 alanine mutations of interfacial residues had been tested. They reported an 88% (66 of 75 residues) predict correctly for hot spot residues with $\Delta\Delta G \geq 1.5$ kcal/ mol.

Tuncbag et al. (2009-2012) [77, 78, 108] determined hot spots based on pair potential atom and solvent accessibility of interface residues. The extraction of computational hot spots comprised of three stages: (1) extraction of interface residues, interface was defined as distance between any two atom having difference chain was less than the sum of their van der Waals radii plus 0.5 Å. (2) Calculation of the features, solvent accessibilities were calculated and total contact potential of interfaced residues. (3) Prediction based on empirical model, the hot spot was labeled when relative accessibility of an individual interface residue is less than 20% and its total contact potential was more than 18. They trained the algorithm by using two-class, hot spot and non-hot spot, for which both conservation and solvent accessibility data were available. Then they tested the algorithm with 25 complexes (54 hot spots and 58 non-hot spots) and found that the predicted hot spots were correctly of 70% when observing to match with the experimental hot spots.

Although there are several methods for finding interactive residues, they have not been successful in all case because of complexity of protein. In order to develop a new approach for defining the binding residues, in this study, the criteria based on atomic pair interaction between protein-protein complexes were constructed. The constructed criteria were a summation of interaction pairs, atom pairs, and amino acid pairs, therefore, a histogram analysis was used for counting the frequency of these pairs.

Purpose of the study

1. To predict the possible CD4-DARPin 23.2 complex structure, based on their docking scores by using ZDOCK and RDOCK protocols.

2. To create an algorithm template histogram-based analysis for analyzing the key residues of CD4 that interacts with DARPin 23.2 using the data from ZDock database and five criteria.

3. To validate the predicted results by checking the physicochemical properties and propensity of hot spot.

4. To validate the predicted results by comparing with HotPOINT and HSPred programs having different algorithm.

Research scope

1. The 3D structures of CD4 and DARPin 23.3 are searched and minimized the energy.

2. The ZDOCK and RDOCK protocols are required for predicting the CD4-DARPin 23.2 complex structures.

3. Each CD4-DARPin 23.2 complex that DARPin 23.2 binds onto CD4 domain 1, Schwizer experiment, are carried out to find the binding atom neighbors.

4. The five criteria for predicting key CD4 amino acid binding DARPin23.2 are set up based on pair potential interface and hydrogen bond property.

5. The top 10 highest histogram values in each criterion are selected and combined together to form decision making in six patterns.

6. The top three maximum values in all six patterns are defined as first, second, and third key CD4 residues interacting with DARPin 23.2.

7. The studies of x-ray structures and experimental mutagenesis data of CD4-gp120 and CD4-MHCII as well as key CD4 residues are used to identify hot spots.

8. The hydrophobic and hydrophilic properties as well as hot spot propensity are required to validate with our predicted hot spots.

9. The percentage of identity prediction ($PID^{predict}$) wes used to compare our prediction and two software prediction.