CHAPTER 1

INTRODUCTION

Statement and Significance of the Problem

Acquired immune deficiency syndrome (AIDS) is caused by infection of Human immunodeficiency virus (HIV). The HIV infection is still world's serious health problem because approximately 34 million people currently live with HIV [1] and the complete displacement of HIV infection is not successful. The unsuccessful treatments are the result of side effect [2] and HIV mutations [2, 3]. The high rates of mutation and viral replication of HIV propel the virus to escape eventually from the recognition by cytotoxic T lymphocytes (CTLs) [4, 5]. Moreover, intracellular viral replication leads to cell death and thus a decreasing total number of circulating helper T lymphocytes (CD4+ T cells) [6]. Fortunately, antiretroviral (ARV) drugs are currently available for halting the rapid decrease of CD4+ T cells in infected patients

Developing the ARV drugs are mostly classified by the steps during the retrovirus life cycle than the direct binding of the drug to the viral particles. Therefore, understanding HIV life cycle is inevitably crucial for ARV drug development. The entry of HIV into the host cell, as a first step of HIV life cycle, is initiated by a binding between the viral envelope protein, gp120, and its primary receptor, CD4, on the surface of T helper lymphocytes or macrophages [7]. Recently,

this entry step has been more targeted for designing the drugs. Examples of agents that hinder the entry of HIV are maraviroc, enfuvirtide [8], and ibalizumab [9]. However, because of the potential drug resistance problems and serious side effects from previous ARV drugs, further development is still needed in order to search the new agents. The problem of drug resistance is caused mainly by virus mutation [2, 3], especially the mutation on HIV surface, i.e. the area surrounding the CD4 binding site on HIV gp120. In addition, this problem is believed to cause the evasion of effective neutralization by the host antibodies, enabling the virus to still retain the infectivity for CD4+ T cells [10].

One of microbicide agents, Designed Ankyrin Repeat Protein (DARPin) technology [11-13], has been developed to apply for antiviral treatment. DARPins specific for human CD4 that inhibit the HIV entry were reported by Schwizer et al. [14], demonstrating antiviral property *in vitro*. They are potent and highly specific to CD4 to block HIV entry and able to act against a wide range of virus strains. Moreover, the basis T cell functions are not disturbed, although these DARPins can block binding of CD4 to MHCII. Subsequently, Pugach et al. [15] studied the binding efficiency of CD4-specific DARPin 57.2 to CD4+ cells (T helper cells, DCs, and monocytes) *in vivo*. They found that DARPin 57.2 has no inhibitory effect on SHIV-infected rhesus macaques. This result may be due to the weak interaction between CD4 and DARPin in an animal model, hence, insight into the interaction between them may provide a clue for improving the binding activity of DARPins.

To understand the mechanism of CD4-DARPin interaction, in this study, the three dimensional (3D) complex structures of CD4-DARPin and hot spots were

identified by incorporation of bio-information, computational simulation, and computer programming. Firstly, any CD4-DARPin complexes were generated using protein-protein docking tools, which are ZDOCK and RDOCK protocols in Discovery Studio (DS) 2.5. The information of CD4-gp120 complex and CD4-MHCII complex were used to fitter the feasible CD4-DARPin complexes. After getting the complex structures, the key binding residues (hot spots) in the complexes were identified. Although several computational approaches have been developed to find the hot spots of protein-protein complex, they have not been successful in all cases. The reason is that the biological properties responding for hot spot have not been fully understood. Therefore, the new approach for clarifying the roles of crucial CD4's residues in CD4-DARPin complex based on counting interaction pair was presented in this work. The interaction pairs were further performed by a histogram analysis in five criteria. Then the histogram values were carried out to determine "considered CD4 residue" and "key CD4 residue". Finally, our key residue predictions were validated by checking propensity of hot spot and comparing with other software which having different method.

Theory

HIV infection

Acquired immunodeficiency syndrome or Acquired immune deficiency syndrome (AIDS) is one of world serious health. In 2011, there were approximately 34 million HIV-infected patients in the world. The worldwide prevalence of HIV

infection was 0.8%, whereas Thailand showed 1.2% as shown in Fig. 1 [1]. Even though there were a lot of died AIDS patients, 1.7 million people in 2011, the death rate was decrease 24% comparing in 2005 due to anti-retroviral treatment (ART).

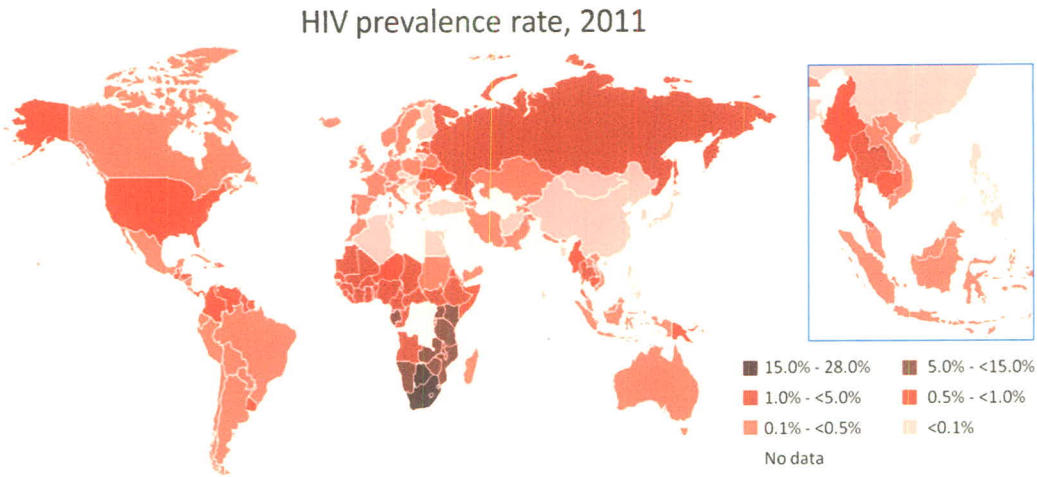

Figure 1 The estimated number of people living with HIV by UNAIDS in 2011 [1]. (Figure from http://www.unaids.org/en/dataanalysis/datatools/aidsinfo/)

AIDS is the end-stage disease of human immunodeficiency virus (HIV) infection. The result of HIV infection is a failure of immune response, allowing opportunistic infections and cancers. The collapse of immune system is the result of the CD8+ T cell control evasion by HIV and the decreasing number of CD4+ T cells.

T cell responding

The CD8+ cytotoxic T lymphocyte (CTL) is effector cell responding to the virus attack by directly killing virus-infected cells (T helper cells, monocytes, dendritic cells) [4, 16] or producing cytokines to enhance antiviral immunity and suppress infection [4]. Therefore, the number of virus is controlled initially by CTL response

[4, 17, 18]. The evasion of HIV to CD8+ cells has been described in two ways. First, down-regulation of MHC class I to present HIV peptide to CTL by HIV Nef protein [4, 19]. Second, the viral sequences that specify immunogenic epitopes are mutated [4].

Studies on T helper cell proliferation suggested that HIV-1 cannot induce T helper cell responses strongly [17, 20, 21]. In spite of the weak response of T helper cells to HIV, the role of CD4+ T helper cells plays critical maintenance of CTL responses in chronic viral infection [22, 23]. The loss of CD4+ T cells is related to a high level of replication, reactivation of persistent viral infection, and a high incidence of virus-associated cancers. Therefore, CD4+ T cells have high relevance for HIV infection [6]. The initial key of infection and decreasing number of helper T cells is CD4 molecule, which is a receptor for HIV entry during the HIV life cycle.

HIV life cycle

There are seven major steps of HIV life cycle, as shown in Fig. 2. The first step is binding and fusion, gp120 on HIV binds to CD4 and co-receptor (CCR5 or CXCR4) on CD4+ cells. Then, virus fuses into the host cell. Secondly, HIV genome and its protein are released into the host cell. Reverse transcription, third step, subsequently occurs when a single-stranded HIV RNA is converted into double-stranded DNA by HIV reverse transcriptase. The fourth step, called integration, proceeds when an HIV DNA enters into the nucleus and is integrated with host's genomic DNA by HIV integrase enzyme. This HIV DNA can remain inactive for

several years. The next step, HIV DNA is then transcribed and translated by the host machineries to yield viral mRNA and viral proteins when the host cell receives a signal to become active. In the viral assembly step, the new HIV proteins and viral RNA genome move to host cell membrane and bud out from the host cell to generate an immature HIV. The last step, the HIV proteins are then cleaved by HIV protease and rearranged within HIV particle to become a mature HIV [24]. In this study, the binding step which gp120 on HIV binds to CD4 was focused.

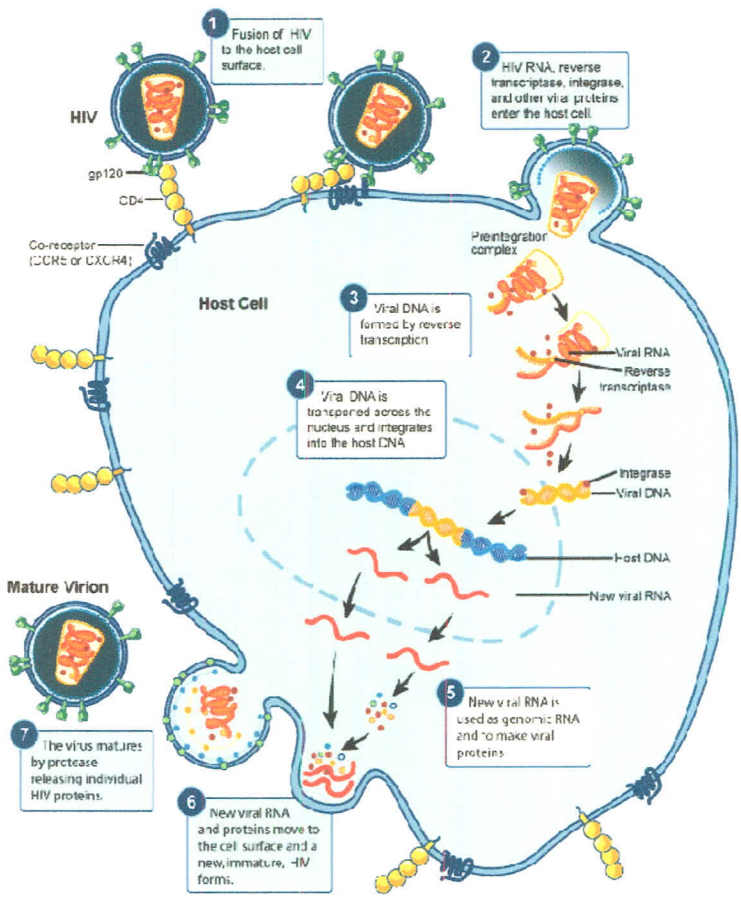

Figure 2 Schematic representation of HIV life cycle within a CD4+ cell. The figure shows key steps and viral target of HIV proliferation. (Figure from http://www.niaid.nih.gov/topics/HIVAIDS/Understanding/Biology/pages/hivreplicationcycle.aspx)

CD4 binding

CD4 is a 55-kDa glycoprotein that expresses on monocytes, dendritic cells, and predominant T helper lymphocytes. It consists of a large extracellular region (residues 1-1372), a transmembrane region (residues 373-393), and an intracellular tail (residues 394-433). The extracellular area is divided into four tandem domains (D1-D4), which has sequence and structure similar to immunolobulins. D1 and D3 resemble Ig variable (IgV) domains and D2 and D4 are Ig constant (IgC)-like domain. The D1 composes of nine β-strand viz. strands A, C, C', C", F, and G forming one surface and strands B, D, and E forming the other [25] as shown in Fig. 3A [26].

The natural ligand of CD4 is histocompatibility complex class II (MHCII), which is a critical co-stimulatory signal in CD4+ cell immune activation. The binding area of CD4 to MHCII is D1. Moebius et al. [27] use site-directed mutagenesis to find interaction between MHC-II and CD4. They found that Lys35, Phe43, Lys46, and Arg59 are amino acid residues on CD4 that bind to MHC-II. Theses residues reside on the CC′C″ and D regions. Moreover, X-ray crystallography analysis by Wang et al. [28] (Fig. 3B) suggested that amino acid on CD4, Phe43, is the first interaction; the second major CD4-MHC-II interaction is Lys46-Leu44; the third interaction site is Ser60 and Asp63. In addition, other hydrophilic residues might include Lys35, Lys46, and ARg56 [28].

The binding area of CD4 to MHCII is a same area for HIV gp120 interaction [29, 30]. The D1 of CD4 is a high affinity binding site for HIV gp120. The key residues of CD4 for binding gp120 discovered from the mutagenesis studies are Lys29, Lys35, Phe43, Leu44, Lys46, Gly47 and Arg59 [31, 32], which also compiled from the

mutagenesis studies by Ryu et.al [33]. Theses residues reside on the CC′C″ and D regions. Furthermore, these binding residues were relative with the crystal structure of CD4-gp120 complex which was successfully solved by Zhou et al. [34] as shown in Fig. 3B.

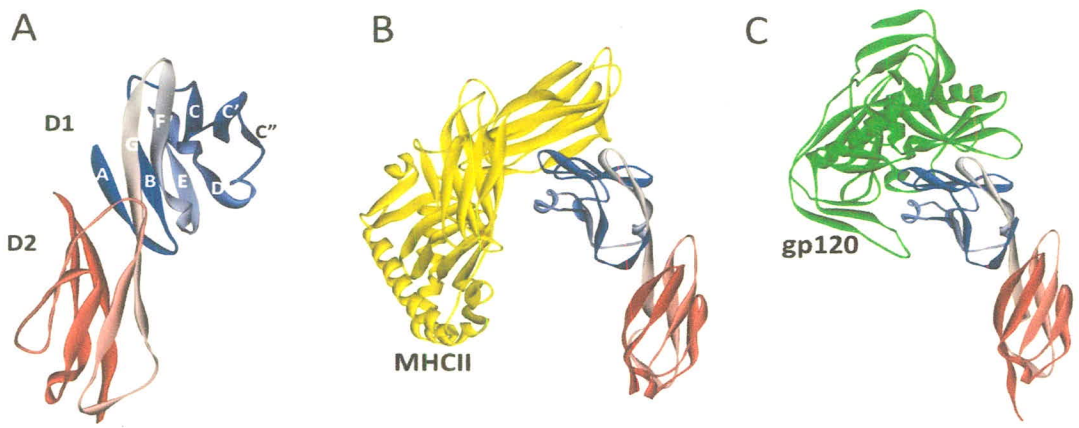

Figure 3 The x-ray structures of CD4 (A) [26], CD4-MHCII complex (B) [28], and CD4-gp120 complex (C) [34]. The blue and red gradient colors correspond to the N- to C-terminus direction is CD4. The yellow color is MHCII and the green is gp120.

HIV/AIDS treatment

Antiretroviral (ARV) drugs for treating HIV have been approved for more than twenty years. Currently, HIV drug therapy employs a combination of at least three drugs, called Highly Active Antiretroviral Therapy (HAART), to control HIV-1. However, HAART is not totally successfully effective in all cases because of the side effects [35] and a high viral mutation problem [2, 3]. The problematic HIV mutation on HIV surface, i.e. gp120 and gp41, results in both an evasion from host neutralizing

antibodies (NAbs) [36, 37] and the drug therapy [38-40]. The mutant HIV escapes the NAbs and the drug, and still retains the infectivity for other CD4+ cells. Therefore, new anti-retroviral drugs have been developed in order to reduce the side effects and replace the older viral resistant drug.

Currently, the available drugs do not directly kill HIV particle, they hinder the growth of virus by blocking HIV replication or HIV life cycle. The first class of anti-HIV drugs was the nucleoside reverse transcriptase inhibitors blocking the conversion of viral RNA into DNA. Subsequently, protease inhibitors hindering HIV maturation were developed [41]. The newer class of ARV, which started 10 years ago [42], is entry inhibitors and the newest class is integrase inhibitor blocking the insertion of HIV genome into the host genome. As of 2009, there were 28 approved ARV drugs approved by the United States Food and Drug Administration (FDA), where only two drugs, Enfuvirtide and Maraviroc, are approved entry inhibitors [43, 44]. Enfuvirtide binds to gp41 leading to an interference of fusion, while Maraviroc binds to CCR5 and prevents its interaction with gp120 [45]. Moreover, the non-approved agents targeting the binding of CD4 to inhibit its gp120 interaction are considered to be the new therapeutic agents, such as TNX-335 [9], PRO-542, 2F5, and 2G12 [42]. In addition, there are also recent discoveries on the new agents that bind to CD4 such as CD4-specifec DARPins.

Ankyrin repeat proteins

Structure of ankyrin

The ankyrin repeat (AR) is a stack of 33 amino-acid sequence motif that consists of two antiparallel alpha-helices separated by short loop, and can be defined as helix-turn-helix motif (Figs. 4, 5). Each repeat is connected together by intervening beta-hairpin or long loop. Usually a hairpin like β-sheet structure consists of the last three residues of the previous AR and the first four amino acids of the next AR [11, 12, 46]. The most common number of repeats per protein is two repeats, as analyzed by SMART database, and three motifs when analyzed with PFAM database [12].

The stability of the helix is due to the fact that the interrepeat interface mainly consists of hydrophobic interactions. The nonpolar residues at the intrarepeat positions 8, 11, 19, and 20 and another interrepeat positions 8, 11, 12, 19, 22, 23, and 24 form the hydrophobic core (Figs. 4B, 5D). These positions have low solvent accessibility [11, 13]. Additionally, the hydrophobic core of a stack is sealed by terminal repeats (capping repeats) that are constant regions called N- and C-terminal repeats (Fig. 4D) [11, 13]. The β-hairpin is stabilized by its main chain (position 29 and 31 of the previous AR and 1-4 of the next AR) through a hydrogen network and also hydrophobic interactions of Ala28, Val30, and Ala32 [46].

Consecutive repeats stack together to form an L-shaped domain, resembling a cupped hand representing the β-hairpin as fingers and the helices as the palm of a hand (Figs. 5A, B) [11, 12, 46]. To maintain this characteristic topology, some residues are well-conserved. Specifically, Binz et al. [11] found that the Thr-Pro-Leu-

His motif at positions 6 through 9 are highly prevalent in an AR canonical sequence. It forms a tight turn and begins the first helix (Fig. 4C). The hydroxyl group of Tyr forms a hydrogen bond with the imidazole ring of His. Moreover, His9 establishes H-bond with Ala32 and the next repeat (randomized position 5). At the central piece, positions 19-24, of the second helix, which is the Val-X (hydrophilic)-Leu/Val-Leu-Leu motif, form intra- and inter-AR hydrophobic networks to stabilize the structure. Furthermore, glycine residues are conserved at position 13 and 25, which are the end of first- and second-helix, providing flexibility for a loop to link both helices [11, 12]. The non-conserved residues located at the groove of AR are defined as the binding residues or surface-exposed residues for contacting targets protein (Fig. 5C). When AR units stack together to form an elongated molecule, the surface residue form a large solvent accessible surface (SAS) to its specific partners, as shown in Fig. 6 [46]. Binz et al. specified theses residues as positions 2, 3, 5, 13, 14, 26, and 33 (Fig. 4A) [11].

DARPin

By nature, ARs can be improved in their properties such as high expression, stabilities, and solubility, by replacing each residue with the corresponding consensus amino acid. The consensus-based protein design is an approach as a result of diversification and selection during protein evolution, has been successfully made a template or building block for engineering and design studies [11, 12]. In consensus design (Fig. 7), the conserved residues are important in maintaining the structure, while the non-conserved residues or solvent exposed residues are important in binding

the targets. Moreover, the solvent exposed residue is the "weakness", which advances to generate novel binding specificity AR. Consequently, the natural ankyrin repeat proteins that are derived typically to improve highly specific and high-affinity target protein binding are called Designed Ankyrin Repeat Protein (DARPin) [11, 12].

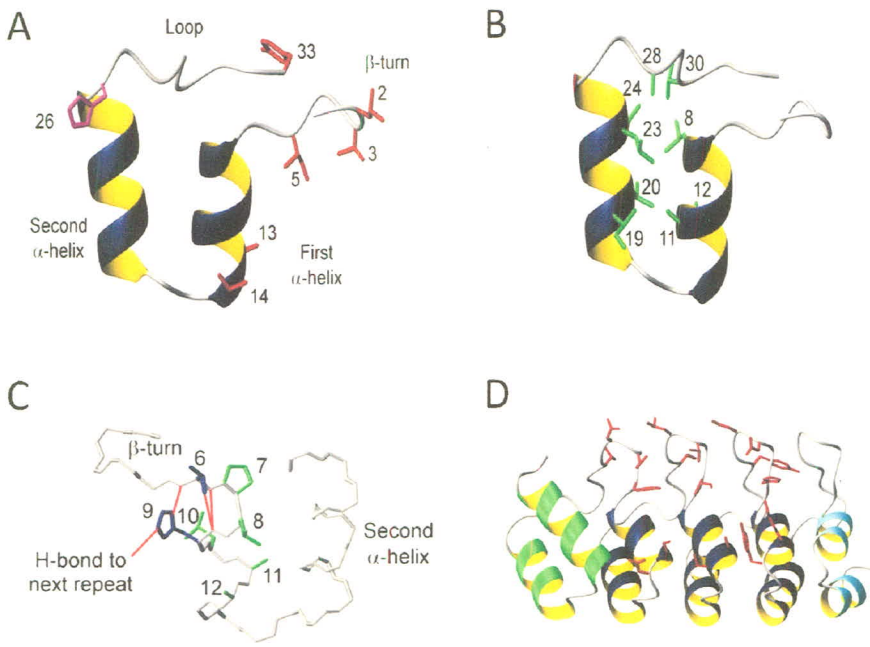

Figure 4 Structural features of the ankyrin repeat represented in one repeat (A, B, and C) and whole structure (D). The potential interaction residues of the internal AR module are displayed in red of (A). These residues are located in the β-turn and the concave surface of the L-shaped repeat. Hydrophobic framework residues and alanine residues are colored in green in (B). Ala11 and Ala12 are important for the overall shape of ARs. A rotated view of this internal AR module (C) shows more clearly the Thr-Pro-Leu-His-Leu-Ala-Ala motif (residue 6–12) of the first a-helix with its characteristic H-bond pattern. Hydrophobic residues and alanine residues are colored in green, Thr6 and His9 are colored in blue and H-bonds are colored in red. A large potential interaction surfaces show in red stick mode of (D). The N and C-terminal capping repeats and the internal repeat modules are colored in green, light blue and dark blue, respectively [11].

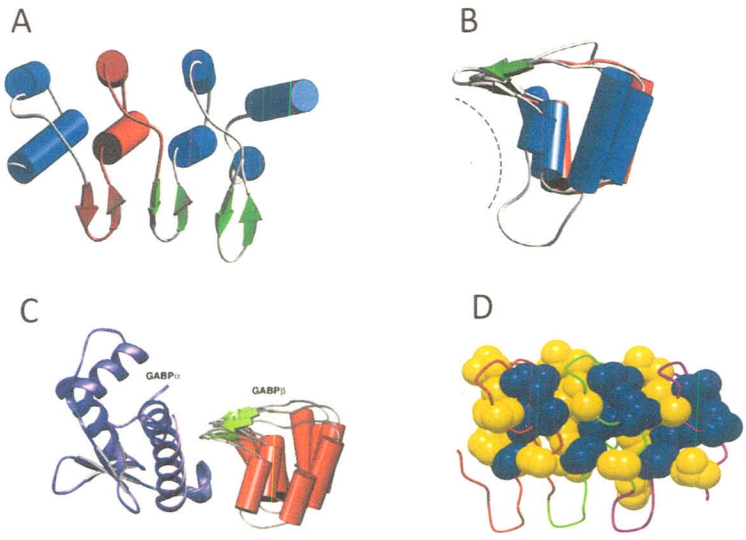

Figure 5 Structure of ankyrin repeat consisted of alpha-helices (cylinders) and beta-hairpins (arrows) is shown in top view (A) and side view (B). A single AR is highlighted in red. The side view (B) of AR is shown as L-shape. The continuous beta-sheet projects away from the helical stack to form the ankyrin groove, which is indicated by the dotted arc. The complex of AR (red) and its target (purple) (C) clearly illustrates the use of the ankyrin groove as a binding surface. (D) is assembly of the AR domain. The backbone of three consecutive ANK motifs is shown as red, green and purple coils. Non-polar side-chain atoms, shown in a space-filling representation, are located on both the N-terminal face (olive green) and C-terminal face (blue) of the helical pairs. These form complementary surfaces that associate and constitute the core interactions that form the ANK-repeat stack [46].

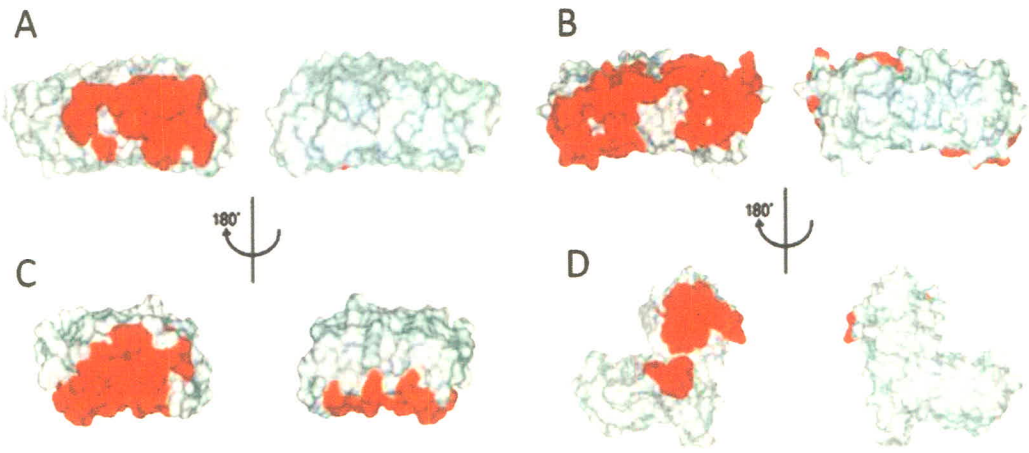

Figure 6 Locations of the interaction interface in four different AR co-crystal structures. The residues involved in the interaction with protein partners are colored red in the surface representations. The beta-hairpin/loop region is pointing directly towards reader (left) and away from the reader (right). (A), (B), (C), and (D) are interface between GABP-β and GABP-α (PDBid: 1AWC), I_k-B_α and NF_k-B (PDBid: 1NFI), p16 and Cdk6 (PDBid: 1BI7), and 53BP2 and p53 (PDBid: 1YCS) respectively [13].

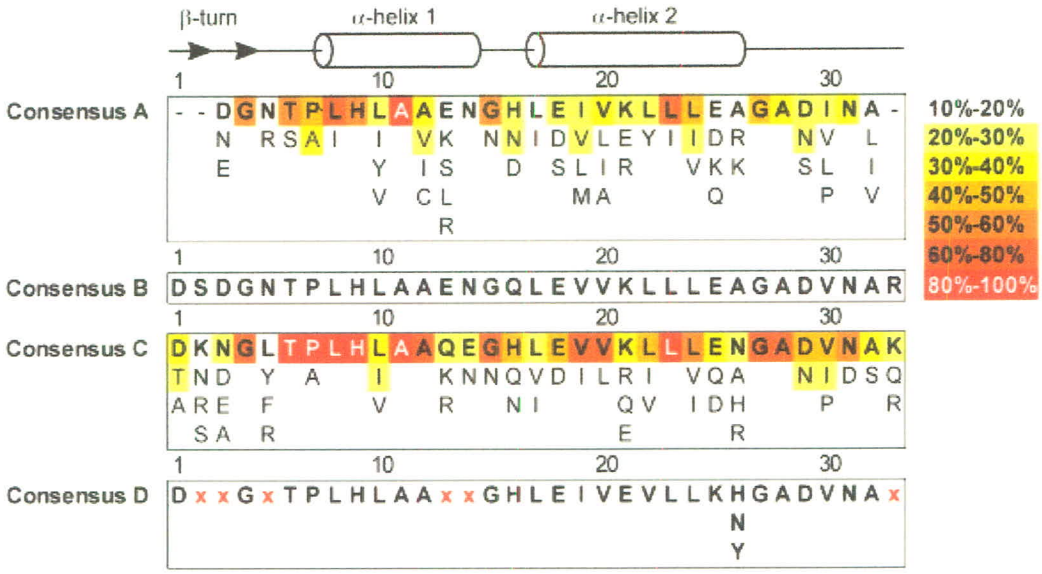

Figure 7 The AR consensus from Binz et al. (2003). The amino acid frequency color code is indicated in the panel. For orientation, the secondary structure elements are indicated above the sequences. Consensus A was derived from an alignment of 229 ARs of the SMART database. The sequence of consensus B was derived from consensus A, where lacking or non-conserved (cut-off #30%) residues were substituted by residues resulting from an alignment of repeats of AR proteins with known structure. Consensus C was derived from the BLAST search with consensus B. Structure-based considerations led from consensus C to consensus D, the final sequence of the designed AR module. In consensus D, the potential target interaction residues are highlighted in red [11].

CD4-specific DARPin

Schweizer et al. [14] successfully developed human-CD4-specific DARPins, which are very potent and highly specific inhibitors for HIV entry *in vitro*. The six candidate DARPins bound to CD4 at domain 1 to interfere the interaction between gp120 and CD4. Moreover, CD4 binding to MHCII was also meddled. In addition, their efficiency prevented HIV infection in various virus strains while basic cellular

functions of memory CD4 T cells were not disturbed. CD4+ T cells retained its proliferation and inducement of dendritic cell (DC) maturation after being treated by CD4-specific DARPins, though any cytotoxic effects were not revealed.

Pugach et al. [15] proved Schweizer's *in vitro* experiment by examined one of the six candidates of CD4-specific DARPins, DARPin 57.2, in both *in vitro* and *in vivo*. PBMCs and lymph nodes were isolated from rhesus macaques which were treated with DARPin 57.2. The result showed that DARPin 57.2 could block SIV infection in blood cells, which is the same blocking level as in the CD4 T cell in lymph nodes. Other examination, after DARPin 57.2 was injected in SIV-infected macaques, the free DARPin 57.2 and bound CD4-specificDARPin 57.2 in circulation were detected. DARPin-bound CD4+ cells were detected in the peripheral blood (T cells, monocytes, dendritic cells) as early as 30 minutes, decreasing within 6 hours and almost undetecTable within 24 hours. It was also detected on CD4+ cells in lymph nodes within 30 minutes. The free DARPins in the plasma were detected but they were rapidly cleared from circulation. In addition, the amounts of bound and unbound DARPins were dependent on the amount of DARPin injected. It was demonstrated that the CD4-specific DARPins can bind its target cells rapidly and selectively *in vivo*.

Protein-protein docking

The protein-protein docking is a computational method to predict complex structure of two unbound proteins. The "docking problem" divides into two parts: developing a scoring function and a searching method. The scoring function is

discriminative value of correct or near-correct docked orientation and incorrect ones. The simplest approach for scoring function is shape complementarity concerning surface shape of protein. In addition, the electrostatic and hydrophobic energy may also be added. The searching method is an algorithm for rapidly finding all possible orientations. The most popular method is the Fast Fourier Transform (FFT), while others are Monte Carlo sampling, genetic algorithms, and geometric hashing [47, 48].

Protein-protein docking is often performed in two stages: the initial stage and refinement stage, as shown in Fig. 8. In the initial stage, proteins are treated as rigid bodies and then a search is performed in a six-dimensional (6D) space (three rotational and three translational degrees of freedom). The scoring function is calculated all orientation and the top scores are identified for a set of candidate complexes. In the refinement stage, side-chain residues of candidate complexes from the initial stage are flexed. Subsequently, they are rescored and re-ranked by using minimization and a more detailed energy function. Further filtering steps, using biological information to restrict the possible complexes, may be used to help choosing the correct one from false positive. For example, the residues known to be binding site, as CDR of antibody, or interfering site [48].

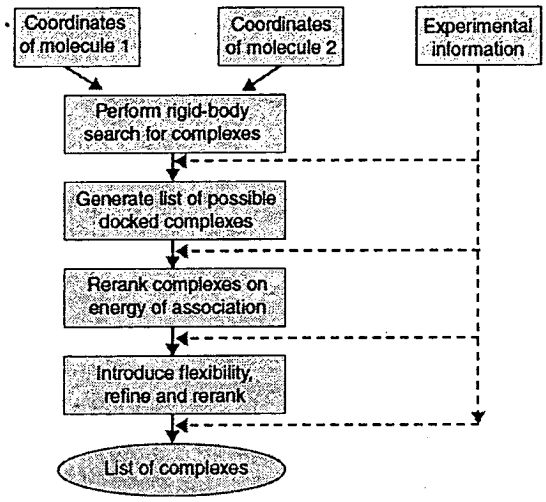

Figure 8 Overview of the protein-protein docking stage, which consists of initial and refinement stage. The filter stage may be used [48].

There are several docking tools, which have different scoring functions and searching methods, for example, AutoDOCK, BiGGER, ClusPro, DOCK, DOT, FTDOCK, GRAMM, HADDOCK, HEX, ICM, PatchDock, RossettaDcok, and ZDOCK.

ZDOCK protocol

ZDOCK is an initial-stage docking program and focuses on rigid-body and non-bond docking. It uses FFT to find the 6D structure of protein complex, which search space is partitioned in the Cartesian (three rotations and three translations) coordinate system. The rotation and translation searches are performed on the ligand with respect to the receptor, which is fixed as the origin. Note that in protein-protein docking, the liagand is smaller protein and receptor is larger one. The ZDock scoring is optimized by shape complementarity parameters (electrostatics and desolvation free energy may be added) to become a scoring function [49] used following equation:

$$\text{ZDock score} = \alpha PSC + DE + \beta ELEC \quad (1)$$

, where α and β are default values as 0.01 and 0.06, respectively. The high value of ZDock score shows a good complex.

Pairwise shape complementarity (PSC) is scoring function which counts the number of atom pairs between ligand and receptor within distance cutoff. The cutoff is defined as a parameter D plus the radius of receptor atom. It is composed of a favorable term and penalty term. The two discrete functions R_{SC} and L_{SC} (SC defines for shape complementarity) are used to describe the geometric characteristics of receptor (R) and ligand (L). A $NXNXN$ grid with each grid point (l,m,n = 1,2,...N) is assigned on both R and L and the scoring value is calculated following equations:

$$\mathrm{Re}[R_{PSC}(l,m,n)] = \begin{cases} \text{number of receptor atoms within (D+receptor atom radius)} & \text{open space} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$$\mathrm{Re}[L_{PSC}(l,m,n)] = \begin{cases} 1 & \text{if this grid is the nearest grid of a ligand atom} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$\mathrm{Im}[R_{PSC}(l,m,n)] = \mathrm{Im}[L_{PSC}(l,m,n)] = \begin{cases} 3 & \text{solvent excludeing surface of the protein} \\ 9 & \text{protein core} \\ 0 & \text{open space} \end{cases} \quad (4)$$

$$S_{PSC}(o,p,q) = \mathrm{Re}[\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} R_{PSC}(l,m,n)\square L_{PSC}(l+o,m+p,n+q)] \quad (5)$$

, where, S_{PSC} is the final PSC scoring computed by correlation function between R_{PSC} and L_{PSC}; o,p and q are the number of a grid points which L is translated with respect to R; *Re[]* and *Im[]* are the real and imaginary parts of a complex function.

For favorable term, all atom pairs within cutoff are assigned with the same score; score of 1 for each atom pair without atom types, as calculated in equation (2) and (3). In equation (4), the penalty component is linearly proportional of a combination of overlapping grid points between receptor and ligand. By assigning, core-core, surface-core, and surface-surface are scores as -81, -27, and -9, respectively. The final PSC, equation (5), computes both favorable and penalty components of PSC. The positive PSC score indicates a good shape complementarity [50, 51].

Desolvation (DE) energy is energy required to break the molecule-solvent interactions at the binding site when molecules bind in an aqueous environment. The DE is estimated based on the Atom Contact Energy (ACE). The ACE is defined as the free energy of replacing a protein-atom/water contact with protein-atom/protein contact. The ACE scores are derived from observed protein-atom/protein-atom contacts in 90 crystal structures for all pairs of 18 atom types. The DE in ZDock score is the summation of the ACE scores of all pairwise atom of two proteins within distance cutoff at 6 Å [51].

Electrostatics (ELEC) is expressed as electric potential of specific partial charge-charge interactions. In ZDock score, ELEC is calculated by Gabb et al. [52] approach based on the coulombic formula.

Moreover the docked poses can be calculated in energy function by ZRANK program [53] in ZDOCK protocol. The ZRank scoring is linear combination of van

der Waals, electrostatics, and desolvation. The van der Waals and short range electrostatics energies, distance between two atoms is lesser than 5.0 Å, use parameters of CHARMm 19 polar hydrogen potential to calculate. For long range electrostatics interaction, only fully charged side chain atoms are used. The desolvation uses ACE to calculate.

RDOCK protocol

RDOCK is an algorithm for refinement and rescoring of docked pose from ZDOCK to pick out near-native structures. RDOCK consists of two-stage energy minimization; first, minimization with ionic residues in neutral state; second, minimization with ionic residues at charged state. During the two-stage energy minimization, CHARMm is used to remove any clashes from ZDOCK conformations and optimize polar and charge interactions. The scoring function of RDOCK is the sum of the CHARMm electrostatics energy and ACE desolvation energy [54].

Both ZDOCK and RDOCK protocols are algorithms in Discovery Studio (DS) software and also find in ZDOCK SERVER (http://zdock.umassmed.edu/).

Other docking software

Hex is the one of docking software that focuses on rigid-body and non-bond docking. It uses FFT-based approach for searching complex orientation, which search space is partitioned in the spherical polar fourier (SPF) (five rotations and one

translation) system. For scoring function, it calculates an excluded volume model of shape complementarity with an optional *in vacuo* electrostatic contribution [55, 56]. This docking algorithm is in Hex software and HexServer, a web server interface for Hex (http://hexserver.loria.fr/).

Key binding site residues

The protein-protein interface (PPI) is the area of two protein engagement. This area is not equally homogenous distribution, instead, a small set of crucial residues called hot spot (key binding residue) at the PPI contribute significantly to the binding free energy [57, 58]. Therefore, the PPI is composed of binding/interacting and nearby residues. There are many methods for distinguishing interface residues such as numerical value-based methods, probabilistic method and clustering process. The examples of numerical value-based method are linear regression, scoring function, support vector machine, and neural network. Probabilistic method consists of naïve Bayesian, Bayesian network, Hidden Markov model, and conditional random field. The different methods are suited for different type of input data [59].

Physical properties of hot spot [60-62]

Physicochemical features of hot spot were described by hydrophobic and hydrophilic properties. The properties of 20 amino acids were shown in Table 1. Hydrophobic residues are those side-chain consists of hydrocarbon without polar and poorly interact with water. Avoiding of these side-chains to water lead to pack

against each other called hydrophobic effect. So, the stability of 3D protein structure requires this interaction. Relatively, many studies [63-69] suggested that hydrophobic residue plays a role in core of monomer protein. For hydrophilic residues which those side chain have polar (hydroxyl group) group are able to make hydrogen bond to each other. Moreover, these residues easily interact with water, as forecasting they are dominant residue on surface of protein [63, 64]. Some of them have charge in side chain which interaction between ionic groups of opposite charge make a strong interaction called salt bridge as shown in Fig. 9. Altogether, the monomer structure is stabilized by interaction between side chains comprising hydrophobic interaction, hydrogen bond, salt bridges, including disulfide bond [70, 71].

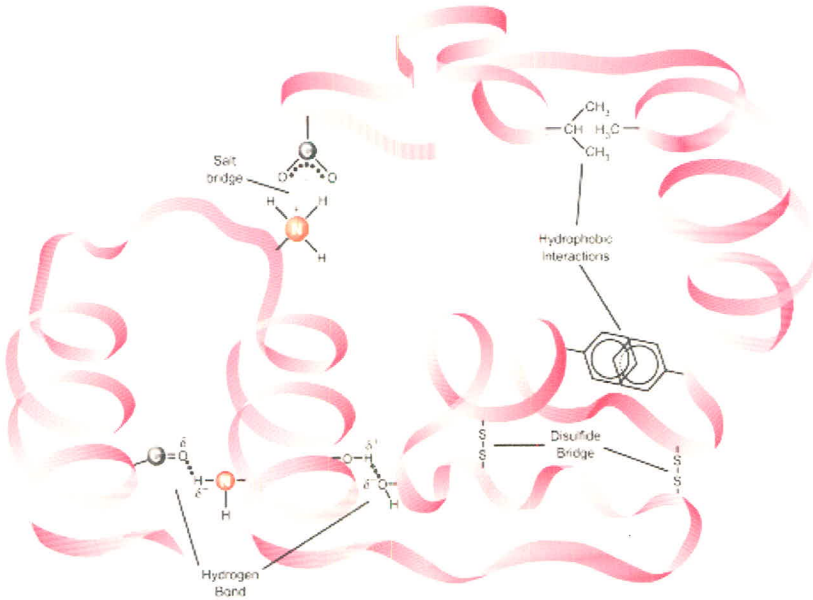

Figure 9 The schematic of intramolecular interactions (page 135 of [71]).

The interactions between two proteins are still required functional group of side chain. In protein-protein binding, hydrophobic residues are more dominate than hydrophilic residues, however, the frequency of interaction is less than core of monomer protein [63-69]. Although hydrophilic is less dominate than hydrophobic on interface, these residues are prevalent in hertodimer interface [63-67]. The detail of binding residue is described as below.

1) Hydrophobic mostly are found in protein-protein interfaces more than the exposed area. Hydrophobic amino acids tend to be increased in the interface of hydrophobic patches of 2000 to 400 $Å^2$.
2) A high degree of electrostatic and hydrogen-bonding complementarity is also observed for protein-protein interfaces. Charged groups, particularly the planar guanidium and carboxyl groups of Arg, Glu, and Asp, are also to be binding residue. Moreover, both charge and polar residue showed increased affinity for the interface [72]. These groups affect to formation of salt bridge across the intermolecular interface.
3) Aromatic side chain residues, such as Tyr, Trp, His, and Phe, are found to play in the recognition and activation of receptor.
4) The enrichment of Arg is attributed to cation-π interactions.
5) Propensity of binding residues was studies by many studies.

- Bogan & Thorn [58] used heterodimer complex and alanine mutant database to observe the binding propensity. The result showed that Trp is most highly enriched, followed by Arg, Tyr, Ile, Asp, and His.
- Glaser et al. [73] demonstrated that large protein-protein interface consists of abundant hydrophobic, whereas, polar residues are abundant in small interface.

- Ariel et al. [74] observed the residue distribution in protein-protein interface. They found that Asn, Thr, Gly, Ser, Asp, Ala, and Cys are highest propensity to be at protein-protein interfaces.
- Yan et al. [64] suggested that hydrophobic residues (Try, Phe, Met, Cys, Pro, Leu, Val, Ile) have high preferences at protein-protein interface. For hydrophilic residues accepting Arg and His were not preferred at interface when considering normalized interface propensities.
- Dong et al. [75] showed that Phe, Ile, Leu, Met, Val, Trp, Tyr, His, and Arg are flavored in interface area. The residues that did not prefer in the interface were Trp, Glu, Pro, and Ala.
- Conte et al. [68] and Bahadur et al. [69] suggested that hydrophobic residues (Phe, Trp, Leu, Ile, Val, Met) and His as well as Tyr are rich in the interface and are depleted in charged residues (Asp, Glu, Lys, but not Arg).

6) Solvent accessibility of binding residue has higher solvent accessibilities than non-interface surface residues.

7) Secondary structure of interfaces seem to favor β-strands while disfavor α-helix.

8) Amino acid side chains in hot spots are located near the center of protein-protein interfaces.

Table 1 The properties of 20 standard amino acid.

residue name	symbols (1 letter)	properties	polarities	Hydropa-thies[76]	number of atom	
					D[a]	A[b]
glycine	Gly (G)	hydrophobic	nonpolar	−0.4		
alanine	Ala (A)	hydrophobic	nonpolar	1.8		
valine	Val (V)	hydrophobic	nonpolar	4.2		
leucine	Leu (L)	hydrophobic	nonpolar	3.8		
isoleucine	Ile (I)	hydrophobic	nonpolar	4.5		
phenylalanine	Phe (F)	hydrophobic	nonpolar	2.8		
thryptophan	Trp (W)	hydrophobic	nonpolar	−0.9	1	
methionine	Met (M)	hydrophobic	nonpolar	1.9		
cysteine	Cys (C)	hydrophobic	nonpolar	2.5		
proline	Pro (P)	hydrophobic	nonpolar	−1.6		
serine	Ser (S)	hydrophilic	polar	−0.8	1	1
theonine	Thr (T)	hydrophilic	polar	−0.7	1	1
tyrosine	Tyr (Y)	hydrophilic	polar	−1.3	1	1
asparagine	Asn (N)	hydrophilic	polar	−3.5	1	1
glutamine	Gln (Q)	hydrophilic	polar	−3.5		1
aspartic acid	Asp (D)	hydrophilic	acidic polar	−3.5		3
glutamic acid	Glu (E)	hydrophilic	acidic polar	−3.5		2
lysine	Lys (K)	hydrophilic	basic polar	−3.9	1	
arginine	Arg (R)	hydrophilic	basic polar	−4.5	3	
histidine	His (H)	hydrophilic	basic polar	−3.2	2	2

[a] D is hydrogen dornor.

[b] A is hydrogen acceptor.

Evaluation of web servers [60]

1) HotPoint [77, 78]

This server resolves computational hot spots in protein-protein interfaces from complex structure. The method is based on solvent accessibility or accessible surface area (ASA) and pair potentials of residues and adjusted the scoring formula according to experimental data. The experimental data were the available crystal structure having alanine scanning data obtained from ASEdb [79]. The training data set composed of 150 experimentally alanine mutated residues which were 58 hot spots and 92 non-hot spots. Note that the interface residues whose binding free energy change $\geq$ 2.0 kcal/mol are considered as hot spots and the non-hot spots are defined as energy change < 0.4 kcal/mol. For test set, the 112 residues composed of 54 hot spots and 58 non-hot spots obtained from BID [80].

The interface residues were carried out to calculate the ASA by using Naccess [81]. Then the ASAs were transformed into relative accessibility by using the following equation:

$$relCompASA_i = \frac{ASA}{\max ASA_i} \text{ X } 100 \qquad (6)$$

, where $relCompASA_i$ is the relative ASA in complex of i-th residue and maxASA_i is the maximum ASA of a residue in a tri-peptide state [82].

Moreover, the pair potential (PP) was computed on the interface residues as following equation:

$$PP_i = abs\left(\sum_{j=1}^{n} Pair(i,j)\right) \qquad \text{for } |i-j| \geq 4 \tag{7}$$

$$Pair(i,j) = \begin{cases} \text{contact potential of type } (i,j) & \text{if } d(i,j) \leq 7.0 \text{ and } |i-j| \geq 4 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

, where PP_i is pair potential of residue *i*, *Pair(i, j)* is the contact potential of residues *i* and *j*, and *d(i, j)* is the distance between two residues. The contact potentials of type *(i,j)* is the knowledge-based solvent mediated inter-residue potentials which calculating from probability of being solvent exposed and core for a residue of type (Keskin et al., 1998 [83]). The potential was extracted from frequencies of contacts between different residues in known 3D structure of proteins. There are 210 distinct potentials (all possible pairs of 20 amino acids) in R*T* unit (R is universal gas constant; *T* stands for temperature) for contacting residue.

The results showed that the relCompASA$_i$ at ≤ 20.0 can discriminate between hot spot and non-hot spot with the high precision (P), recall (R), specificity (S), accuracy (A), and F_1 score (F_1) as shown in Table 2. For PP_i, the threshold distinguishing between hot spot and non-hot spot with good above assessment is $\geq$ 18.0. Interestingly, when they combined relCompASA$_i$ $\leq$ 20.0 with PP_i $\geq$ 18.0, the precision, specific, and accuracy of combined two features in both training set and test set had higher value than single relCompASA$_i$ and PP_i. Therefore, this method uses criteria of relCompASA$_i$ ≤ 20.0 and $PP_i \geq 18.0$ to determine the hot spots. Moreover, this result from this method had better precision, recall, and F_1 than Robetta server [84] which becomes the *de facto* standard of comparison in the field.

Note that precision is the ratio of number of correctly classified hot spot residues to the number of all residues classified as hot spots; recall (R) is the proportion of number of correctly classified hot spot residues to the number of all hot spot residues; specificity (S) is the proportion of number of correctly predicted non-hot spot residues to the number of all non-hot spot residues; accuracy (A) is the ratio of number of correctly predicted residues to number of all predicted residues; and F_1 score is the balance between precision and recall.

Table 2 Performance values of various empirical prediction method of HotPoint

set	Model	P	R	S	A	F_1
training set	relCompASA ≤ 20.0	0.55	0.81	0.58	0.67	0.65
	PP ≥ 18.0	0.56	0.55	0.73	0.66	0.56
	relCompASA ≤ 20.0 + PP ≥ 18.0	0.64	0.52	0.82	0.70	0.57
test set	relCompASA ≤ 20.0	0.60	0.67	0.59	0.63	0.63
	PP ≥ 18.0	0.69	0.70	0.71	0.71	0.70
	relCompASA ≤ 20.0 + PP ≥ 18.0	0.73	0.59	0.79	0.70	0.65
	Robetta	0.63	0.57	NS[a]	NS[a]	0.60

[a] NS is not shown.

(2) HSPred [85, 86]

In this server, the hot spots were identified using a Support Vector Machine (SVM) as machine learning approach with the energy term as input features. In basic thermodynamic consideration, the hot spot is determined by estimating the binding free energy change ΔΔG upon alanine mutation using following equations:

$$\Delta G_{bind}^{(wt)} = G_{AB}^{(wt)} - G_{A}^{(wt)} - G_{B}^{(wt)} \tag{9}$$

$$\Delta G_{bind}^{(mut)} = G_{AB}^{(mut)} - G_{A}^{(mut)} - G_{B}^{(wt)} \tag{10}$$

$$\Delta\Delta G = \Delta G_{bind}^{(mut)} - \Delta G_{bind}^{(wt)} = \left[G_{AB}^{(mut)} - G_{AB}^{(wt)}\right] - \left[G_{A}^{(mut)} - G_{A}^{(wt)}\right] \tag{11}$$

, let A and B denoted the unbound monomer and AB the complex.

In this study, they aimed to calculate ΔΔG only energy of mutated complex structure ($G_{AB}^{(mut)}$) instead of estimating all of four states as shown in equation 11. They divided the region on complex structure into three regions: side-chain atoms of the mutated residue and the atom within 10.0 Å of C_β of the mutated residue in the same and different protein as shown in Fig. 10. The interaction area on complex structure was distinguished into three types: side-chain inter-molecular, environment inter-molecular, and side-chain intra-molecular. These interaction areas were directly determined energy terms contributing hot spot interactions i.e. van der Waals potentials, solvation energy, hydrogen bonds, and Coulomb electrostatics.

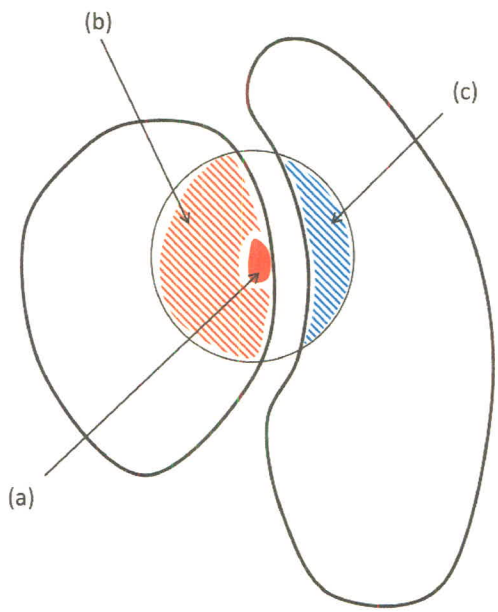

Figure 10 Theschematic overview of protein regions defined the different energy contributions. The red filled area (a) is side-chain atoms of the mutated residues; the red (b) and blue (c) striped regions are any atoms within 10 Å of the C_β of mutated residue. The interaction was distinguished into three types: side-chain intra-molecular between (a) and (b), side-chain inter-molecular between (a) and (c), environment inter-molecular between (b) and (c).

The energy calculation of van der Waals interaction between two atoms *i* and *j* used "smoothed" 6-12 Lennard-Jones potential:

$$VvdW(r) = \begin{cases} \varepsilon_{ij}\left[\left(\frac{r_{ij}}{r+r_0}\right)^{12} - 2\left(\frac{r_{ij}}{r+r_0}\right)^{6}\right] & \text{if } r < r_{ij} - r_0 \\ \varepsilon_{ij} & \text{if } r_{ij} - r_0 < r < r_{ij} + r_0 \\ \varepsilon_{ij}\left[\left(\frac{r_{ij}}{r-r_0}\right)^{12} - 2\left(\frac{r_{ij}}{r-r_0}\right)^{6}\right] & \text{if } r < r_{ij} + r_0 \end{cases} \quad (12)$$

, where, r is the distance between the two atoms, ε is the depth of potential well, and the parameters r_{ij} and ij are taken from CHARMM19 force fields [87]. The parameter $r_0 = 0.5$ Å has the widening effect in the region of maximum affinity and to reduce the potential energy at $r = 0$ to a finite value. In case of $r > 8.0$ Å which is a long distance, the V_{vdW} (r) is defined as 0. The electrostatic interactions were calculated with a screened Coulomb potential, in which the dielectric constant increases linearly with distance. The hydrogen bond energy was evaluated from a linear combination of a distance-dependent part and two angular dependent components following [84]. The desolvation energy was computed using implicit solvation model [88], which decomposes the solvation free energy into a sum of pairwise atomic interactions

In term of three interaction areas and four energy terms, therefore there were 12 input features (4 X 3) but not all of them were used to build their SVM models. The calculation of SVM_X shows a good accuracy with almost of amino acid types. However, this scoring function was not performing well in glutamic acid (Glu) and arginine (Arg). So, the scoring functions for these two amino acids were optimized. Altogether, their models composed of three linear scoring functions; SVM_X calculated for any residue except Glu and Arg while SVM_E and SVM_R computed for Glu and Arg respectively. Each linear scoring function had different weight and input feature as shown in Table 3.

They used the 20 crystal structures of complex having alanine mutational data (from ASEdb [79]) to adjust the scoring. In total mutating data consisted of 349 mutations, of which 81 hot spots. Note that they defined hot spots with experiment $\Delta\Delta G \geq 2$ kcal/mol. The results showed that the calculated $\Delta\Delta G$ at > 0 kcal/mol can discriminate between hot spot and non-hot spot with precision and recall respectively

of 56% and 65%. Moreover, in the same data set, the precision and recall of this method had better score than Robetta server (52% precision and 47% recall).

3) Other web server: Cons-PPISP, Promate, PINUP, PPI-Pred, and Meta-PPISP.

Table 3 Weight of energy terms in the scoring functions.

Feature (energy term)	**SVM_X**	**SVM_E**	**SVM_R**
Side-chain inter-molecular	–	–	–
van der Waals	0.25±0.03	–	0.79±0.04
hydrogen bond	0.16±0.04	0.63±0.04	–
electrostatics	–	–	–
desolvation	0.21±0.03	–	–
Environment inter-molecular			
van der Waals	0.13±0.02	–	–
hydrogen bond	0.18±0.03	0.69±0.03	0.50±0.04
electrostatics	–	–	–
desolvation	0.10±0.01	–	–
Side-chain intra-molecular			
van der Waals	0.26±0.06	0.60±0.04	0.49±0.04
hydrogen bond	–	–	–
electrostatics	–	–	0.47±0.06
desolvation	–	–	–
Threshold	0.43±0.05	0.54±0.07	0.32±0.07

Histogram analysis

A histogram is a graphical representation of a frequency distribution of data. The graph shows a count of points falling in various ranges by rectangles. The height of a rectangle is equal to the frequency density of interval and the total are the histogram is equal to the number of data. The benefits of histogram are; first, helps to easily see where the majority of values fall and how much variation is; second, use a tool to assist in decision making [2, 89]. For example, it is used for summation of

large data such as recording climate change [90], showing environmental (compositional) data [91], and putting down the electrocardiogram (ECG) signals [92]. In addition, histogram is applied in image processing field to get a desired image [93, 94] and in a biomolecular simulation field [95].

In this study, the frequencies of interaction pairs in each residue of protein were plotted in histogram form that relative with constructed criteria. Then, the histogram frequencies were selected and normalized to get the key residues of complex structures. This analytical procedure was validated with other software, which can find key residue to ensure that it is capable to give reliable results.

Validation

Validation is process for determining whether the analytical models demonstrate the accurate results. Analytical parameters for validation include accuracy, precision, , precision, recall, specificity, F-measure, linearity, rage, raggedness, limit of detection, limit of quantitation, and selectivity. For accuracy parameter, it is a measure of closeness between the test results and accepted reference results. Accuracy indicates the deviation of test values when comparing with true values [96]. The accuracy formulated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{13}$$

,where TP, FP, TN, FN are number of true positives, false positive, true negatives, and false negatives respectively. The meaning of true positive is that the individual

has the condition and predicts as positive result for the condition. Unlike TP, false positive is that the prediction of individual shows positive for the condition although the individual does not have the condition. For true negative, the individual does not have the condition and the prediction is negative. Lastly, false negative is that the individual has the condition but the prediction shows negative value. Accuracy has been evaluated by several suggested approaches, for instance, measuring the CRM (Certified reference material) and comparing test results with a reference method. The reference method refers to a nationally or an internationally fully validated method having similar or different measurement principles and sources of errors [97]. In both methods, recovery is defined as the ratio of the observed value to the expected value which is expressed as percentage [96].

For precision, it is the ratio of the number of true correctly true prediction to the number of both true and false prediction. The recall is the fraction of correctly true prediction among all true data. The specificity is the proportion of correctly false prediction to all false data. The F-measure is the way to check the balance between precision and recall. The formulas as showed below:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{14}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{15}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{16}$$