SCREENING INTERACTIONS BETWEEN PROTEINS AND DISORDERED PEPTIDES BY A NOVEL COMPUTATIONAL METHOD

by

Weiyi Zhang

Bachelor of Science, Nankai University, 2001

Master of Science, University of Pittsburgh, 2006

Submitted to the Graduate Faculty of

The Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCE

DEPARTMENT OF PHYSICS AND ASTRONOMY

This thesis was presented

by

Weiyi Zhang

It was defended on

Apr 24, 2013

and approved by

Dr. Carlos Camacho, Associate Professor, Department of Computational & Systems Biology

Dr. Xiao-lun Wu, Professor, Department of Physics & Astronomy

Dr. Ralph Roskies, Professor, Department of Physics & Astronomy

Dr. Vladimir Savinov, Associate Professor, Department of Physics & Astronomy

Dr. David Snoke, Professor, Department of Physics & Astronomy

Dr. Daniel Zuckerman, Associate Professor, Department of Physics & Astronomy

Dissertation Advisors: Dr. Carlos Camacho

Dissertation Co-Advisors: Dr. Xiao-lun Wu

Copyright © by Weiyi Zhang

2013

SCREENING INTERACTIONS BETWEEN PROTEINS AND DISORDERED PEPTIDES BY A NOVEL COMPUTATIONAL METHOD

Weiyi Zhang, PhD

University of Pittsburgh, 2013

Concerted interactions between proteins in cells form the basis of most biological processes. Biophysicists study protein–protein association by measuring thermodynamic and kinetic properties. Naively, strong binding affinity should be preferred in protein–protein binding to conduct certain biological functions. However, evidence shows that regulatory interactions, such as those between adapter proteins and intrinsically disordered proteins, communicate via low affinity but high complementarity interactions. PDZ domains are one class of adapters that bind linear disordered peptides, which play key roles in signaling pathways. The misregulation of these signals has been implicated in the progression of human cancers. To understand the underlying mechanism of protein-peptide binding interactions and to predict new interactions, in this thesis I have developed: (a) a unique biophysical-derived model to estimate their binding free energy; (b) a novel semi-flexible structure-based method to dock disordered peptides to PDZ domains; (c) predictions of the peptide binding landscape; and, (d) an automated algorithm and web-interface to predict the likelihood that a given linear sequence of amino acids binds to a specific PDZ domain. The docking method, *PepDock,* takes a peptide sequence and a PDZ protein structure as input, and outputs docked conformations and their corresponding binding affinity estimation, including their optimal free energy pathway. We have applied *PepDock* to screen several PDZ protein domains. The results not only validated the capabilities of *PepDock* to accurately discriminate interactions, but also explored the underlying binding mechanism. Specifically, I showed that interactions followed downhill free energy pathways, reconciling a

relatively fast association mechanism of intrinsically disordered peptides. The pathways are such that initially the peptide's C-terminal motif binds non-specifically, forming a weak intermediate, whereas specific binding is achieved only by a subsequent network of contacts (7–9 residues in total). This mechanism allows peptides to quickly probe PDZ domains, rapidly releasing those that do not attain sufficient affinity during binding. Further kinetic analysis indicates that disorder enhanced the specificity of promiscuous interactions between proteins and peptides, while achieving association rates comparable to interactions between ordered proteins.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

Protein–protein interactions, which occur when two or more proteins bind together to conduct certain biological functions, are at the core of the inner working of a living cell. Many of the most important molecular processes in the cell, such as DNA replication and signal transduction, are carried out by large molecular complexes that are organized by protein–protein interactions. This subject has been studied for decades from different perspectives: biochemistry, biophysics, structural biology, bioinformatics, etc. In this chapter, the basic biophysical concepts are discussed, the topic of this thesis is presented, and key questions are proposed.

A. PHYSICS AND BIOLOGY

Biology is a natural science concerned with the study of life and living organisms. With desires to understand the origin of life, humans started biological research as early as ancient civilizations. For instance, Taoist tradition of Chinese Alchemy, which can be considered part of life science due to its emphasis on health with the ultimate goal being the elixir of life, is dated back to 4^{th} century BC [1]. Today, subdisciplines of biology are usually recognized on the basis of the scale at which organisms are studied: biochemistry examines the rudimentary chemistry of life; molecular biology studies complex interactions of systems of biological molecules; cellular biology examines the basic building blocks of all life, the cell; physiology examines the physical

and chemical functions of tissues, organs, and the organ system of an organism; and ecology examines how various organisms interact and associate with their environment.

Physics and its techniques have played a significant role in the evolution of biology. For instance, biology began to develop and grow quickly with the dramatic improvement of microscopes in the 17th century and X-ray crystallography and nuclear magnetic resonance are essential tools for structural biologists nowadays. During the 20st century, physicists and biologists work in two different ways. Physics is theory-driven and uses mathematics to represent the laws of nature, whereas biology is experimentally based and relies on words and diagrams to describe the functions. The essence of physics is to simplify phenomenon and explain it in a quantitative way, whereas molecular biology strives to tease out the smallest details [2].

Fortunately, new challenges stopped physics and biology from drifting apart and brought researchers together. Over the last decade, biologists started facing massive DNA sequences, profiles of gene expression and protein structures generated by high-throughput experimental techniques. For example, structural molecular biology concerned with how structures of molecules determine their functions and how alterations in structures affect their functions. In the last 30 years, the number of protein structures in Protein Data Base [3] has increased from 12 in 1972, to 30,000 in 2005, and to 80,000 in 2012. These new changes challenge traditional biological research methodologies while offering opportunities for physicists to contribute to the development of new theories in biology.

My interest in structural biophysics led my research to focus on understanding the mechanism of protein–protein interactions by analyzing their three-dimensional structures. One challenge in this area relates to Intrinsically Disordered Proteins (IDPs). IDPs, often referred to

as disordered proteins, are proteins characterized by lack of stable tertiary structures when the protein exists as an isolated polypeptide chain (or a subunit) under physiological conditions in vitro [4,5]. In the last 15 years, the discovery of disordered proteins challenged the traditional protein structure paradigm, which states that a specific well-defined structure is required for the correct function of a protein and the structure defines the function of the protein [5,6,7]. The disordered proteins remain functional despite the lack of a well-defined structure, but can adopt a fixed 3D structure after binding to other molecules. These fuzzy proteins are not scarce in biology. On the contrary, they play fundamental roles, and are highly prevalent and extensively involved in human diseases. For example, research showed neurodegenerative diseases such as Parkinson's disease were associated with disordered proteins [8]. In signal transduction, disordered proteins, together with scaffold proteins, are recruited to associate the correct repertory kinase and its targets into the biochemical pathway quickly and precisely.

Scaffold proteins are crucial regulators to tether multiple proteins of one pathway into complexes and localize protein components to specific areas of the cell such as plasma membrane. One example of how scaffold proteins work together with disordered proteins is that PDZ protein domains [9] associate with their target proteins by binding the linear disordered C-terminal region of the target proteins into their binding pockets. A common PDZ-containing protein, such as PSD95, has multiple PDZ domains and could bind several subunits of a particular channel. These interactions promote clustering of receptors at specific subcellular sites and help spatially organize signal channels [10,11]. Additional examples of scaffold proteins are the src homology 2 (SH2) domain [12], src homology 3 (SH3) domain [13] and pTyr-binding (PTB) domain [14].

B. OPEN QUESTIONS AND RESEARCH GOAL

Interactions involving disordered proteins are more complicated and intriguing than those between well-structured proteins because of the brisk flexibility introduced by disorder. Disordered proteins adopt fixed 3D structures in the binding grooves when binding their partners. Compared to interactions between well-structured proteins, extra free energy is required to compensate for the entropy loss caused by peptides during the transition from the state of disorder to the state of order. Is this extra free energy penalty always a useless cost? The answer is no. Nature uses disorder as a tool to adapt to different environments. Dr. Liu and Dr. Camacho [15] showed that when an individual protein binds to multiple disordered partners, which is common in signal transduction, disorder can help the protein to maximize the discrimination between different partners. This high specificity of promiscuous interactions by disorder usually accompanies relative low affinities. By contrast, it is more difficult for well-structured proteins to achieve this phenomenon than disordered proteins. In addition, compared to structured proteins, disordered proteins can associate with different partners by using their multiple underlying conformations. A list of comparisons between structured proteins and disordered proteins is presented in Table I-1.

Although several models [4,16,17,18,19] have been proposed to explain the coupling mechanism of folding and binding, there remains some uncertainty in the underlying mechanism. Here is a list of questions that the author attempts to answer:

- Do disordered peptides that bind to one class of proteins conform to certain physical characteristics, e.g. charge and hydrophobicity?
- How do disordered peptides bind to structured proteins, and what is the physical mechanism of the coupling between folding and binding?

- What are the advantages or disadvantages of protein–peptide interactions compared to protein–protein interactions, and why are they prevalent in signal transduction?
- In evolution, why does nature prefer disorder in specific biological functions?

To answer these questions, a computational framework is prerequisite to modeling and predicting the complexes formed by disordered peptides and structured proteins. Different computational methods [20,21,22,23,24,25] have been developed and are currently available for protein–peptide screening. But we found each of them had its functional or methodological limitation due to its initial design purpose, and none of them could satisfy our needs. Some of them [21,22,23] compute relative binding affinity changes of candidate peptide sequences compared to reference peptide sequences, instead of estimating the binding affinity directly due to the limitation of their models. Others [20,24,25] are based on sequence analysis by a statistical or machine learning method. More important, these data–driven methods, which start with experimental data and optimize their model terms or parameters with curve–fitting, are contrary to our research objectives. Instead of fitting model by data, we want to start building our physical model based on our understanding of protein–peptide interactions, then validate and improve our model with experimental data. So, a structure–based computational method that can quickly and accurately estimate absolute binding affinities, as well as predict the complex structures, is still missing. This motivated us to develop our own protein–peptide docking method, *PepDock*. Two key questions are considered in our methodology:

- How can we estimate the absolute binding affinity? Binding affinity estimation or binding free energy estimation is one of the most difficult questions in protein–peptide docking. Binding free energy function includes different components, e.g. electrostatic, desolvation, internal energy, entropy change, etc. Until publishing, to

my understanding, no method has incorporated a good estimation of the entropy change introduced by disorder of peptides. However, entropy change by disorder plays a very important role in binding interactions, which will be shown in later chapters, and its contribution cannot be negligible. On the contrary, some other components contribute much less and could be neglected without compromising the accuracy. So in our method, we considered balancing the computational complexity (or feasibility) and model accuracy, i.e. simplifying the binding free energy function by making some reasonable assumptions while still capturing the main contribution to the binding interaction with acceptable accuracy.

- How to sample peptide conformations? Compared to protein–protein docking, the computational complexity of protein–peptide docking is dramatically increased due to the need of sampling the peptide flexibility. However from experiments, it is known that scaffold proteins have a unique binding groove and peptides binding to them have consensus sequences. Furthermore, these consensus sequences of peptides will adopt conserved structures in the binding pockets. By employing these structural evidences, we can restrain the peptides in the known binding site and reduce the number of peptide conformations. In our method, we simplified the sampling complexity and achieve a fast docking methodology.

Bearing these questions in mind, we developed and implemented a novel structure-based computational method, *PepDock*, to predict interactions between disordered peptides and scaffold proteins. *PepDock* accepts as input the 3D structures of one scaffold protein (wild type or homology model) and the amino acid sequences of peptides. As the output, it predicts

complex structure and estimates absolute binding free energy together with the free energy landscape (Figure I-1).

As a case study, we have applied *PepDock* to PDZ domains. We successfully discriminated strong peptide binders from no-binders with 90% specificity and 70% sensitivity, respectively. In addition, *PepDock* mimicked the X-ray crystal structures of PDZ complexes that successfully capture the characteristics of contact interfaces. By analyzing the results, we determined that, sequentially, peptides start binding by anchoring their C-terminal residue into a PDZ pocket, forming the conserved binding motif by the adjacent 3 residues, and zipping the remaining 3–5 residues into the extended contact network. This observation demonstrates that the known recognition consensus sequence, usually the first 4 residues including C-terminal residue, binds to PDZ domain non-specifically and the contact by the next 3–5 residues determines the specificity. The complete procedure follows a downhill free energy pathway. Our findings highlight the induced folding/binding mechanism of disordered peptides as maximizing both the thermodynamic and kinetic specificity of promiscuous interactions, a mechanism that is likely adopted by other scaffold proteins.

C. OUTLINE OF THESIS

The content is organized in the following order. We start with the review of some basic thermodynamic and kinetic concepts in Chapter II. These are the basic building blocks of our methodology framework. In Chapter III, we present the basic procedure of protein–protein docking, and the application of our free energy scoring functions. Next, in Chapter IV, we focus on the methodology of *PepDock* and its application to PDZ–peptide interactions. Based on the

PepDock, we implemented an online PDZ–peptide interaction query and prediction web portal, which is shown in Chapter V. Finally, Chapter VI discusses the conclusion and future outlook.

Table I-1: Difference between structured proteins and intrinsic disordered proteins

	Structured protein	**Intrinsic disordered protein**
Native structure	Proteins exist with well-defined 3-D structures.	Proteins or protein regions lack specific 3-D structures and exist instead as ensembles of flexible unorganized molecules.
Structure change during binding interaction	Well-structured before and after binding interaction.	Disordered before interaction and structured or partially structured after binding interaction.
Binding affinity	High, e.g., 10^{-8} M	Immediate, e.g., 10^{-6} M. Free energy contribution is required to accomplish the disorder to order transition.
Function and specificity	Usually only interact with specific interaction partner; high binding affinity, but low specificity. Function in all biological fields.	Can interact with multiple partners (20 or more). Low binding affinity, but high specificity. Mostly found in signaling, regulation, and control functions.
Model	Lock and key model; Induced-Fit model.	Fly-Casting model; Induced-folding model.
Example	Hemoglobin; Leucine Zipper.	Nuclear receptor co-activator binding domain (NCBD), zinc fingers (linkage region), eukaryotic translation initiation factor (eIF4E).

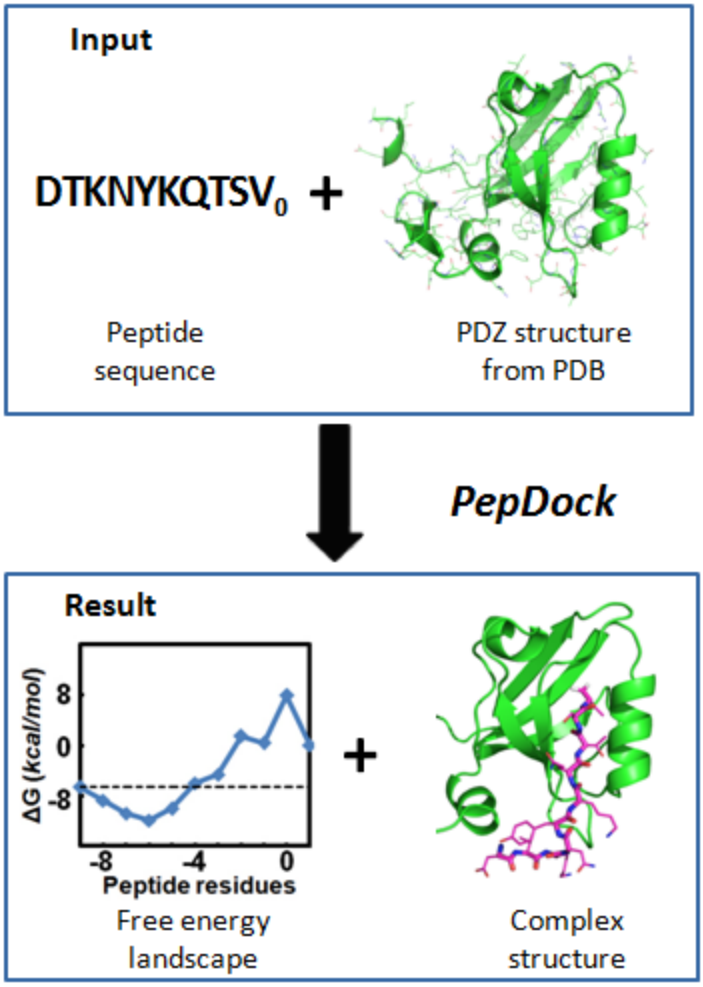


Figure I-1: ***PepDock*.** PepDock is a structure-based computational method to predict interactions between disordered peptides and scaffold protein domains. PepDock has been applied to screen interactions of PDZ protein domains. It takes as input a peptide sequence and PDZ domain structure, and outputs the complex structure prediction and binding affinity estimation including the optimal free energy pathway.

II. THERMODYNAMICS AND KINETICS OF PROTEIN ASSOCIATION

In this chapter, we review the basic concepts of biological thermodynamics and kinetics that will be used in the subsequent chapters. Firstly, we start the discussion with the derivation of the Gibbs free energy change, which determines the direction and strength of association. Secondly, we discuss molecular interactions and entropy, which contribute to free energy change. Thirdly, we introduce biological kinetics and explain the relationship between kinetics and free energy. Last, models of protein–protein association and properties of disordered peptides are presented.

A. THERMODYNAMICS

1. Gibbs free energy

The Second Law of Thermodynamics states that entropy is the essential quantity to measure the direction of the transition of an isolated macroscopic system, where the isolated system will tend towards a state of maximum entropy. However, in biophysics, free energy is the common variable to measure the direction of biological interactions. This is because, for the biological interactions in laboratory or in cell system, it is the temperature and pressure that we control at the boundaries, rather than the work or heat flow. This change in condition requires a new

thermodynamic quantity: free energy. Systems held at constant temperature and pressure tend toward their states of minimum free energy, rather than of maximum entropy [26].

Consider a process inside a test tube, which has constant pressure *p* and no interchange of particles with the surroundings. The tube is held by a heat bath with constant temperature *T*. The process inside the test tube may or may not involve chemical or phase changes. The combined system of test tube and the heat bath is isolated from its surroundings. Based on the Second Law of Thermodynamics:

$$dS_{combined\ system} = dS_{sytem} + dS_{bath} \geq 0, \quad \textbf{Equation II-1}$$

where *S* is entropy and the subscript *system* indicates the test tube. Since the combined system is isolated, the internal energy *U* follows

$$dU_{sytem} + dU_{bath} = 0. \quad \textbf{Equation II-2}$$

Use the fundamental equation,

$$dS = \frac{1}{T}dU + \frac{p}{T}dV - \frac{\mu}{T}dN \quad \textbf{Equation II-3}$$

where μ is chemical potential and *N* is number of particles. Considering the constant temperature *T*, constant pressure *p*, and no particle exchange, the entropy change is

$$dS_{bath} == \frac{1}{T}dU_{bath} + \frac{p}{T}dV_{bath}. \quad \textbf{Equation II-4}$$

Then, combine Equation II-1, Equation II-2, and Equation II-4

$$dU_{system} - TdS_{sytem} - pdV_{sytem} \leq 0. \quad \textbf{Equation II-5}$$

For enthalpy $H = U + PV$,

$$dH_{system} = dU_{system} + pdV_{system} + V_{system}dp. \quad \textbf{Equation II-6}$$

Plug Equation II-6 into Equation II-5 and consider constant *p*, then we get

$$dH_{system} - TdS_{sytem} \leq 0. \quad \textbf{Equation II-7}$$

If we define Gibbs free energy as $G(T, p, N) = H - TS$, N as the number of particles, we see that when a system is at constant temperature T and pressure p, the Gibbs free energy is at its minimum (Equation II-58).

$$dG_{system} \leq 0. \quad \textbf{Equation II-8}$$

Now, consider a protein–protein interaction: receptor protein R binding ligand protein L to form a complex protein C. The process is described as

$$R + L \leftrightarrow C. \quad \textbf{Equation II-9}$$

The molar free energy of solutions with certain concentrations of receptor R, ligand L, and complex protein C are then

$$G_R = G_R^0 + RT \ln[R], \quad \textbf{Equation II-10}$$

$$G_L = G_L^0 + RT \ln[L], \quad \textbf{Equation II-11}$$

$$G_C = G_C^0 + RTln[C], \quad \textbf{Equation II-12}$$

where G_R^0, G_L^0, and G_C^0 are the molar free energies of the standard state (by convention, one molar solution), R is the ideal gas constant, T is temperature, and "[ ]" represents concentration. The free energy change for the interaction is then

$$\begin{aligned} \Delta G &= G_C - G_R - G_L \\ &= G_C^0 - G_R^0 - G_L^0 + RT \ln\frac{[C]}{[R][L]}. \end{aligned} \quad \textbf{Equation II-13}$$

At the equilibrium $\Delta G = 0$. Taking the equilibrium concentrations as $[C]_{eq}$, $[R]_{eq}$, and $[L]_{eq}$,

$$\Delta G^0 = -RT \ln\frac{[C]_{eq}}{[R]_{eq}[L]_{eq}} = RT \ln K^{eq}, \quad \textbf{Equation II-14}$$

where $\Delta G^0 = G_C^0 - G_R^0 - G_L^0$ and K^{eq} is the equilibrium association constant for the association between R, L, and C. Another quantity, the equilibrium dissociation constant K_d^{eq}, which is used to measure the propensity of dissociation, is defined as

$$K_d^{eq} = \frac{[R]_{eq}[L]_{eq}}{[C]_{eq}} = \frac{1}{K^{eq}}. \qquad \textbf{Equation II-15}$$

Taking the exponential of this equation gives

$$K_d^{eq} = e^{\Delta G^0/RT}. \qquad \textbf{Equation II-16}$$

Equation II-16 shows the relationship between the equilibrium dissociation constant and the interaction Gibbs free energy. For folded protein–folded protein association, the typical K_d^{eq} for strong binding interaction is 10^{-8} M or 10^{-2} μM, while a typical number for folded protein–disordered peptide association is 10 μM. In the following, we will drop the subscripts "*eq*" and superscripts "^{0}" since we consider all the concentrations are equilibrium values.

2. Entropy change

In the definition of Gibbs free energy, besides enthalpy, there is another very important term, *entropy S,* which is a macroscopic quantity in terms of the multiplicity of the microscopic degrees of freedom of a system. Entropy is described in the fundamental equation of statistical mechanics:

$$S = k \ln W, \qquad \textbf{Equation II-17}$$

where W is multiplicity and k is Boltzmann's constant.

Entropy change plays an important role in biological interactions and is the key for us to understand free energy change. For example, in protein folding, proteins have a greater degree of disorder (or flexibility), in other words, greater entropy while in de-folded state but have zero entropy in the folded state, under the assumption that protein folded into one unique well-defined structure ($W = 1$). During molecule association, the entropy of the system, including receptor and ligand, will decrease because complex proteins have less degree of freedom than separated

receptors and ligands before the binding process. The change of enthalpy, together with the change of entropy, determines the direction of the biological interactions. At times, entropy change can tune the interaction to achieve its maximum performance in some specific biological cases. In section III.B.5 and IV.C.2, we will discuss entropy change in more detail.

B. KINETICS OF PROTEIN ASSOCIATION

Thermodynamics, or the analysis of free energy change, provides a way to answer the question, "Why do protein R and protein L interact to form complex protein C." But it cannot respond to the question, "How fast will the interaction occur?" To address this point, we must turn to reaction kinetics.

Consider our protein–protein association model again. Protein R and protein L interact and form complex C (Equation II-9). Based on the transition state theory, along the reaction pathway from reactant state A (R and L) to the product state B (complex C), there is an intermediate state that must have the highest free energy (Figure II-1). This leads to the diagram of energy barriers of height E^+ from the states A to B, and of height E^- from the states B to A. This transition state is the position along the reaction pathway with the highest energy [27]. This is one of the most basic ideas in relating the rate constants to the energetics of a molecule as it undergoes a reaction. According to the Arrhenius equation, the rate of transitioning over this barrier is then related to the probability of a molecule having that high energy. This can be estimated from the Boltzmann distribution,

$$k = Ce^{-E^+/RT}, \qquad \textbf{Equation II-18}$$

where is E^+called activation energy. The equation shows that the rate will be slower if the energy of the barrier is higher.

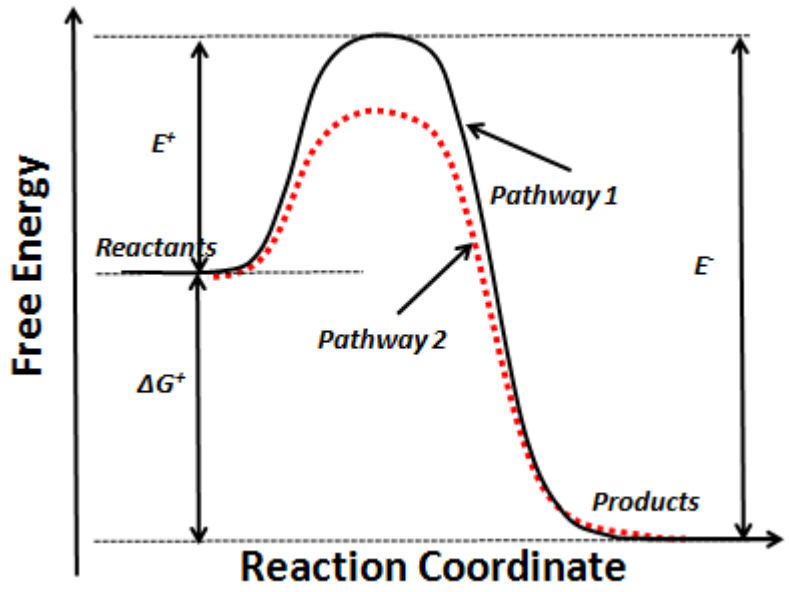


Figure II-1: An example of a reaction coordinate diagram. The reaction coordinate is a measure of the extent to which a reaction has occurred. The starting reactants are on the left, the products are on the right. The free energy is shown in a solid line. The state with the highest free energy is the transition state for the reaction. According to transition state theory, the higher is the free energy barrier, the slower the reaction.

For reactants to form a complex, molecules crossing from left to right need to overcome the energy barrier E^+, while the dissociation reaction, from right to left, needs to go over the energy barrier, E^-. By law of Arrhenius, who proposed a strong temperature dependence of reaction rates in 1889, the on-rate $k^+$and off-rate $k^-$are defined as

$$k^+ = C^+ e^{-E^+/RT}, \qquad \textbf{Equation II-19}$$

$$k^- = C^- e^{-E^-/RT}. \qquad \textbf{Equation II-20}$$

where $C^+$and C^- are constants. At equilibrium, conversion from state A to B is balanced exactly by the reverse conversion from B to A. With time derivatives,

$$\frac{d[A]}{dt} = -k^{+}[R][L] + k^{-}[C], \quad \textbf{Equation II-21}$$

$$\frac{d[B]}{dt} = k^{+}[R][L] - k^{-}[C], \quad \textbf{Equation II-22}$$

yield the following expression:

$$k^{+}[R][L] = k^{-}[C], \quad \textbf{Equation II-23}$$

$$\frac{[C]}{[R][L]} = \frac{k^{+}}{k^{-}} = K_d, \quad \textbf{Equation II-24}$$

$$\frac{[C]}{[R][L]} = \frac{C^{+}}{C^{-}} e^{-(E^{+}-E^{-})/RT} = Ce^{-\Delta G/RT}. \quad \textbf{Equation II-25}$$

where C is a constant. Equation II-25 expresses the equilibrium dissociation constant in terms of the free energy difference between the two states, $\Delta G = E^{+} - E^{-}$. This demonstrates a basic relation between kinetics and energetics, as shown in Figure II-1 Although pathway 1 and pathway 2 have the same dissociation constant K_d and the free energy difference, ΔG, pathway 1 has low on and off rate with a higher energy barrier, while pathway 2 has relatively high on and off rate with a lower energy barrier.

C. MODELS OF PROTEIN–PROTEIN RECOGNITION

Lock and key theory was first postulated by Emil Fischer in 1894 [28] to explain the specific binding interaction of an enzyme with a single substrate. The enzyme active site has a unique geometric shape that is complementary to the geometric shape of a substrate molecule. Only the correctly sized key (substrate) fits into the keyhole (active site) of the lock (enzyme).

Induced fit model was suggested by Daniel Koshland in 1958 [29] to explain protein–protein recognition, since scientists found that not all experimental evidence can be adequately

explained by the lock and key model. The induced fit model shows that receptor proteins are rather flexible structures in which the binding site continually reshapes by its interaction with the ligand substrate until the ligand is completely bound to it. This is also the point at which the final form and shape of the complex is determined.

Anchor model, which was proposed by Camacho's group in 2004 [30], states that in some protein–protein interactions, one of the interacting proteins, usually the smaller of the two, anchors a specific side chain in a structurally constrained binding groove of the other protein, providing a steric constraint that helps to stabilize a native-like bound intermediate. It has been verified that, even in the absence of their interacting partners, the anchor side chains are found in conformations similar to those observed in the bound complex. These ready-made recognition motifs correspond to surface side chains that bury the largest solvent-accessible surface area after forming the complex (>100 $Å^2$). The existence of such anchors implies that binding pathways can avoid kinetically costly structural rearrangements at the core of the binding interface, allowing for a relatively smooth recognition process. Once anchors are docked, an induced fit process further contributes to forming the final high-affinity complex. This later stage involves flexible (solvent-exposed) side chains that latch to the encounter complex in the periphery of the binding pocket. The results suggest that the evolutionary conservation of anchor side chains applies to the actual structure.

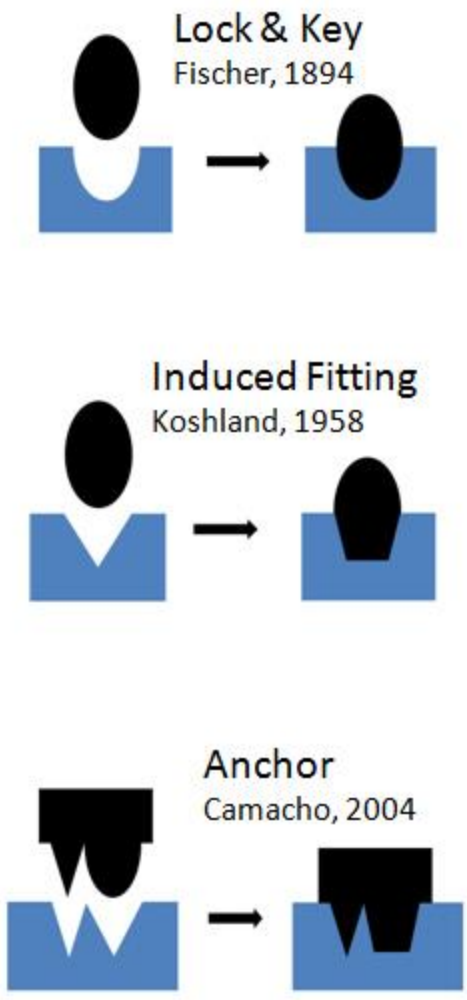


Figure II-2: Protein–protein association models. From top to bottom, the lock and key model states that interfaces of receptor and ligand proteins exactly match each other during binding interactions. The induced fitting model indicates that interfaces of receptor and ligand proteins will fit each other to achieve high affinity during binding interactions. The anchor model shows that anchor residue of ligand protein will intrude to the binding site of receptor protein first and the interface around anchor residue will induced-fit to the binding site.

D. THERMODYNAMICS OF DISORDERED PEPTIDES

Protein folding can be described as a transition from unfolded state *U* to folded state *F*, while the protein stability depends on the free energy difference $\Delta G = G^U - G^F$. The process is expressed by the following equation:

$$U \leftrightarrow F \qquad \textbf{Equation II-26}$$

$$-RT \ln K_d = \Delta G = \Delta H - T\Delta S, \qquad \textbf{Equation II-27}$$

where *R* represents the gas constant; *T*, the temperature; *K*, the equilibrium constant; Δ*G*, the free energy change between folded and unfolded; Δ*H*, the enthalpy change; and Δ*S*, the entropy change from folded to unfolded states. The enthalpy change ΔH represents the binding interaction, which includes electrostatic interactions, solvation effects, hydrogen bonds, and van der Waals potentials. Entropy change ΔS, corresponding to the flexibility of the protein, is positive when proteins change from folded state to unfolded state. The proteins become more stable when the free energy difference $\Delta G = G^U - G^F$ is increasing, i.e., the free energy of the unfolded state, G^U, is relatively higher than the folded state G^F.

For native ordered proteins that have $\Delta G > 0$ kcal/mol (Figure II-3 A), enthalpy change is sufficient to compensate for the entropy loss, i.e., the intra-molecular interaction is strong enough to form the protein into a folded structure. For native disordered proteins with $\Delta G < 0$ kcal/mol (Figure II-3 B), intra-molecular interaction is not enough to compensate for the entropy change. Note that protein stability is relative. Native ordered proteins can transform to disordered state by increasing the temperature and native disordered proteins can fall into a structure when decreasing the temperature.

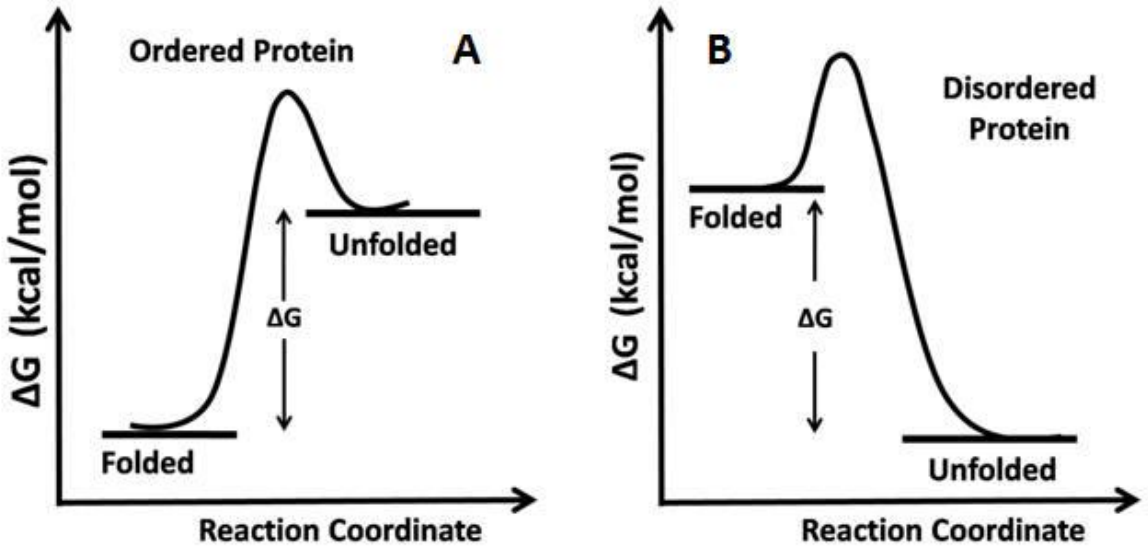


Figure II-3: Transition from order to disorder for native well-structured proteins (A) and disordered proteins (B). Because native folded proteins (ordered proteins) have folded state with lower free energy than unfolded state, they spontaneously fold into a defined 3-D structure. Transition from order to disorder requires free energy compensation. Disordered proteins, which have unfolded states with lower free energy, natively exist instead with a lack of specific 3-D structures. For disordered proteins, transition from unfolded state to folded state needs free energy contribution to compensate for the entropy loss.

III. PROTEIN–PROTEIN DOCKING AND BINDING FREE ENERGY ESTIMATION

In chapter III, we discuss how thermodynamics is applied to our protein–protein docking study. Our docking program, *SmoothDock*, which includes rigid body docking, free energy scoring function and refinement, was implemented by Dr. Camacho and his colleagues in 2004 [31,32]. It aims to predict the complex structure of structured protein–protein association. In section A, we introduce each component of our docking program. In section B, we focus on the most important and interesting part: free energy scoring function decomposition and implementation. We describe the manner by which the model and algorithm estimate the absolute binding affinity of protein–protein association and protein–disordered peptide association. In section C, the author has applied the *SmoothDock* scoring function to the experiment of Capri T45 to discriminate natural and designed protein complexes and obtained very good performance.

A. FOLDED PROTEIN–PROTEIN DOCKING

Current protein docking methods generally consist of a rigid body search that generates a large number of docked conformations with favorable surface complementarity, followed by the re-ranking of the conformations using a potential approximating free energy function [33,34,35,36,37].

1. Rigid body docking

The most widely used rigid body search is based on the Fast Fourier transform (FFT) correlation approach, introduced by Katchalski-Katzir and associates [38] in 1992. This approach provides an efficient way to predict the structure of a possible complex between molecules of known structures by systematically exploring the space of docked conformations and enables one to perform large-scale docking studies [39].

The FFT correlation approach relies on the well-established correlation and Fourier transformation techniques used in the field of pattern recognition. The algorithm requires only that the 3D structure of the molecules under consideration be known. It begins with a geometric description of the protein and the ligand molecules, derived from their known atomic coordinates. The two molecules denoted by *a* and *b*, are projected onto a three dimensional grid of *N* x *N* x *N* points, where they are represented by the discrete functions

$$a_{l,m,n} = \begin{cases} 1, & \text{on the surface of the molecule} \\ \rho, & \text{inside the molecule} \\ 0, & \text{outside the molecule} \end{cases} \qquad \textbf{Equation III-1}$$

$$b_{l,m,n} = \begin{cases} 1, & \text{on the surface of the molecule} \\ \delta, & \text{inside the molecule} \\ 0, & \text{outside the molecule} \end{cases} \qquad \textbf{Equation III-2}$$

The surface is defined here as a boundary layer of finite width between the inside and the outside of the molecule. The parameters ρ and δ describe the value of the points inside the molecules, and all points outside are set to zero. Matching of surfaces is accomplished by calculating correlation functions. The correlation between the discrete functions *a* and *b* is defined as

$$c_{\propto,\beta,\gamma} = \sum_{i=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} a_{l,m,n} \cdot b_{l+\alpha,m+\beta,n+\gamma}. \qquad \textbf{Equation III-3}$$

where α, β, and γ are the number of grid steps by which molecule b is shifted with respect to the molecule a in each dimension. If the shift $\{\alpha, \beta, \gamma\}$ is such that there is no contact between the two molecules, the correlation value is zero. If there is a contact between the surfaces, then the contribution to the correlation value is positive. By assigning large negative values to ρ in molecule a, and small non-negative values to δ in molecule b, we can predict the penetration which is physical forbidden.

A direct calculation of the correlation between the functions a and b is rather lengthy, since it involves N^3 multiplications and additions for each of the N^3 possible relative shifts $\{\alpha, \beta, \gamma\}$, resulting in an order of N^6 computing steps. Therefore the Fourier transformation is applied here to calculate the correlation function much more rapidly. The discrete Fourier transform (DFT) of a function $x_{l,m,n}$ is defined in Equation III-4. The application of this transformation to both sides of Equation III-3 yields Equation III-5.

$$X_{o,p,q} = \sum_{i=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} \exp\left[-\frac{2\pi i(ol + pm + qn)}{N}\right] \cdot x_{l,m,n} \qquad \textbf{Equation III-4}$$

$$C_{o,p,q} = A^*_{o,p,q} \cdot B_{o,p,q} \qquad \textbf{Equation III-5}$$

where $o, p, q = \{1, \ldots, N\}$ and $i = \sqrt{-1}$. C and B are the DFT of the functions c and b, respectively, and A^* is the complex conjugate of the DFT of function a. The inverse Fourier transform (IFT), which is defined as

$$c_{\propto,\beta,\gamma} = \frac{1}{N^3}\sum_{o=1}^{N}\sum_{p=1}^{N}\sum_{q=1}^{N} \exp\left[\frac{2\pi i(o\alpha + p\beta + q\gamma)}{N}\right] \cdot C_{o,p,q}\,, \qquad \textbf{Equation III-6}$$

is used to obtain the desired correlation. The Fourier transformation can be performed with the fast Fourier transform algorithm [40], which requires less than the order of $N^3 \ln(N^3)$ steps. Thus,

the overall procedure leading to Equation III-6 is significant faster than the direct calculation of function *c* according to Equation III-3.

In practice, we use the FFT correlation approach with a 10° Euler angle increment, and default values of 1 Å grid-step and 4 Å surface layer to sample approximately 10^{10} putative conformations, of which the top scoring 20,000 were retained for filtering by free energy scoring function [35]. The FFT method can explore vast numbers of docked conformations, evaluating a simple function that describes the geometric fit or surface complementarity of each structure, possibly allowing for some overlap. The approach is very successful when docking bound (co-crystallized) protein conformations. However, the situation is very different when docking unbound (independently crystallized) conformations of the component proteins. Due to the incorrect conformations of some key side chains in the binding site, all near-native structures may have relatively poor surface complementarity, and hence, the higher ranked conformations are frequently false positives, i.e., structures with good score but high root mean square deviation (RMSD) [33].

2. Empirical free energy scoring function

The free energy of association is often dominated by desolvation and/or electrostatic contributions. Consequently, the free energy scoring function, including desolvation and electrostatic, is used to filter the false positives generated by rigid body docking, and to capture the complexes whose binding mechanism is governed by any combination of the two. Performance and accuracy are both important concerns when designing the scoring function, as it is used to evaluate a large number of docking conformations. Simple and lightweight functions are always preferred.

We use the following scoring functions to estimate the binding free energy of folded protein–protein association (Equation II-7).

$$\Delta G = \Delta E_{elec} + \Delta G_{ACE} - T\Delta S_{trv}, \quad \textbf{Equation III-7}$$

where $\Delta E_{elec}$is electrostatic potential, ΔG_{ACE} is atomic contact energy which captures desolvation free energy, T is temperature, and ΔS_{trv} is association entropy loss. Free energy decomposition is discussed in the section III.B.

3. Refinement

Based on the observations that the native-binding site is expected to exhibit a free energy attractor with the greatest breadth of all the local minima on the free energy landscapes of partially solvated receptor–ligand complexes, and that the attractor is most relevant within distance separations of around a nanometer, or 10 Å, a hierarchical clustering method is used to select and rank the docked complexes that have the largest number of neighbors within a certain fixed cluster radius, 9 Å RMSD [31,41,42].

2000 docked conformations filtered by the free energy scoring function are clustered and based on the number of structures that a ligand has within a (default) cluster radius of 9 Å RMSD. The largest cluster is selected and its cluster center is ranked first. Next, the members of this cluster are removed from the matrix, and the next largest cluster is selected and ranked second, and so on. After clustering, the ranked complexes are subjected to a straightforward (300-step and fixed backbone) van der Waals minimization using CHARMM [43] to remove potential side chain clashes.

The robustness of our docking method was tested on sets of 2000 docked conformations generated for 48 pairs of interacting proteins [44]. The results showed that in 31 cases, the top 10

predictions include at least one near-native complex, with an average RMSD of 5 Å from the native structure.

B. BINDING FREE ENERGY FUNCTION OF FOLDED PROTEIN-FOLDED PROTEIN ASSOCIATION

The binding interaction free energy of one receptor protein and one ligand protein association to a complex is expressed by the form:

$$\Delta G = G^C - G^R - G^L, \qquad \textbf{Equation III-8}$$

where G^C,G^R, and G^L denote the free energies of the complex, the free receptor, and the free ligand, respectively. In a general case, we calculate the binding free energy by the form [33,45]

$$\Delta G = \Delta E_{elec} + \Delta E_{vdw} + \Delta G_{des} + \Delta E_{int} - T\,\Delta S_{sc} - T\,\Delta S_{trv}, \qquad \textbf{Equation III-9}$$

where ΔE_{elec} and ΔE_{vdw} denote the changes in the electrostatic and van der Waals energy, respectively; ΔG_{des} is the desolvation free energy, ΔE_{int} is the internal energy change due to flexible deformations (including bond stretching, angle bending and torsional energy terms), and ΔS_{sc} is the loss of side-chain entropy upon binding. The last term, ΔS_{trv} accounts for translational, rotational, and vibrational entropy change upon binding [46,47]. Since ΔS_{trv} is a weak function of the size and shape of the interacting proteins [48,49], it will be considered constant. The above free energy expression can be substantially simplified when used for docking or scoring. Since G^R and G^L are constant, i.e., they do not depend of the conformation of the complex in an arbitrary reference state, $\Delta G = G^C$.

One important factor in the implementation of free energy scoring function is the complexity. Since protein docking requires filtering or sampling millions of plausible complex

structures, the more sophisticated, and perhaps more accurate, methods in the literature are computationally expensive and are not suitable for free energy screening, e.g., free energy perturbation [50], Poisson–Boltzman [51], atomic continuum electrostatic [52], and generalized-Born solvation [53].

1. Internal energy

Internal energy E_{int} includes three different types of intra-molecular forces that describe bond stretching, angle bending, and bond torsion (Equation III-10). $k^{stretch}$ is the stretching force constant with a typical number 500 kcal/mol for an amino acid. r is the actual bond length in the molecule and r^0 is the natural bond length. k^{bend} is the angle bending force constant with a typical value 50 kcal/mol for an amino acid. θ is the actual bond angle in the molecule and θ^0 is the natural bond angle. $k^{torsion}$ is the barrier to free rotation for the natural bond with a typical value 5 kcal/mol for an amino acid, n is the periodicity of the rotation, and φ is torsion angle. Because internal energy change ΔE_{int} is small compared to the other terms in the binding free energy expression in protein–protein association [45,48,54], we neglect it to simplify the calculation.

$$E_{int} = \sum_{(i,j)\in bonds} \frac{1}{2} k_{ij}^{stretch}\left(r_{ij} - r_{ij}^0\right)^2 + \sum_{(\alpha)\in angles} \frac{1}{2} k_{\alpha}^{bend}(\theta_\alpha - \theta_\alpha^0)^2 + \sum_{(\beta)\in torsions} k_{\beta}^{torsion}(1 + \cos(n_\beta \varphi_\beta - \varphi_\beta^0) \quad \textbf{Equation III-10}$$

2. Electrostatic interaction

The electrostatic interaction E_{elec} is obtained by a simple Coulombic potential with the distance dependent dielectric of *4r*:

$$\Delta E_{elec} = \frac{1}{4\pi\varepsilon\varepsilon_0}\sum_{i<j}\frac{q_i q_j}{r} \qquad \textbf{Equation III-11}$$

$$\varepsilon = 4r$$

where charge pair $\{q_1, q_2\}$ have a distance of r and ε_0 describes the vacuum permittivity. Implicit solvation is described by a simple distance-dependent dielectric model ε. For more detail about our implicit solvation model, see section IV.C.1.

3. Desolvation interaction

The desolvation free energy change ΔG_{des} accounts for hydrophobic interactions. The expression $\Delta G_{des} - T\,\Delta S_{sc}$ is modeled by the atomic contact energy (ACE) term ΔG_{ACE}, an empirical knowledge-based contact potential [55]. In ACE, the local interactions between two molecules are given by

$$\Delta G_{ACE} = \sum_i \sum_j e_{ij} \qquad \textbf{Equation III-12}$$

where the sum is taken over all atom pairs that are less than 6 Å apart. The term e_{ij} denotes the atomic contact energy of between atoms i and j, and is defined as the effective free energy change when a solute–solute bond between i and j is replaced by a solute–solvent bond.

Although the atomic contact energies were estimated by a statistical analysis of atom pairing frequencies in high-resolution protein structures rather than in complexes, the function has been used to calculate the contribution of $\Delta G_{des} - T\,\Delta S_{sc}$ to the binding free energy in a number of applications by our group and other groups [33,35,56,57,58,59,60]. By including ACE and electrostatic energy, the binding free energies, calculated for nine protease-inhibitor complexes, were typically within 10% of the experimentally measured values [55].

With the above simplifications, the free energy function is reduced to the form:

$$\Delta G = \Delta E_{elec} + \Delta G_{ACE} + \Delta E_{vdw} - T\,\Delta S_{trv}. \qquad \textbf{Equation III-13}$$

4. Van der Waals interaction

The van der Waals interaction includes an attractive portion and a repulsive portion. The attractive portion stems from induced dipole-induced dipole interactions, i.e., fluctuations of the charge distribution in one atom or molecule induce charge fluctuations in a neighboring atom. These charge fluctuations lead to an attractive electrostatic interaction. The repulsive portion results from the *Pauli exclusion principle*, a quantum mechanical effect that results in unfavorable energies for interpenetrating electron clouds of two approaching atoms. The van der Waals interaction is usually approximated by Lennard–Jones potential energy function, which is often referred to as 6–12 potential (Equation III-14). At large distances, the energy approaches zero. At the intermediate distances, the energy is negative, which leads to attractive forces. When the distance between the atoms is further reduced, the repulsive forces grow rapidly and give highly positive energies.

$$E_{vdw} = \frac{A}{r^{12}} - \frac{B}{r^{6}} \qquad \textbf{Equation III-14}$$

The function is often further simplified by assuming van der Waals cancellation. According to this assumption, the solute–solute interfaces and solute–solvent interfaces are equally well packed, and hence, the intermolecular van der Waals interactions in the bound state are balanced by solute–solvent interactions in the free state [48,49,61,62,63,64], reducing the binding free energy to

$$\Delta G = \Delta E_{elec} + \Delta G_{ACE} - T\,\Delta S_{trv}. \qquad \textbf{Equation III-15}$$

We assume van der Waals cancellation as a first-order approximation when evaluating the binding free energy of docked conformations in the rigid-body analysis. This approximation is necessary, if no energy minimization is performed, because the docked conformations are not completely free of steric conflicts, resulting in wildly varying ΔE_{vdw}values. Since the correlation between ΔE_{vdw} and the RMSD is close to zero in the rigid-body analysis, the van der Waals term is not much more than some high frequency noise. However, the minor overlaps can be easily removed by the minimization, and ΔE_{vdw} becomes an important part of the free energy function in the further discrimination algorithm.

5. Entropy change

Our free energy function includes the association entropy change (ΔS_{trv}).

$$\Delta S_{trv} = \Delta S_{trans} + \Delta S_{rot} + \Delta S_{vib} \qquad \textbf{Equation III-16}$$

The molecules in solution have degrees of freedom representing overall movements of translation, rotation, and internal vibrations. The entropy and the free energy associated with these degrees of freedom can be calculated with high precision for simple molecules, e.g., polycyclic aromatic hydrocarbons, in the gas phase. The calculation can be extended under certain conditions to larger molecules and to proteins. It yields the price, the association entropy

penalty $-T\ \Delta S_{trv,}$ that must be paid for degrees of freedom lost when two molecules associate to form a stable complex such as an antigen–antibody complex or an enzyme-inhibitor complex, where their movements are highly constrained. This price, a reduction of entropy, which depends on the residual mobility of the components in the complex, has been empirically estimated to be 15 kcal/mol [65,66]. This free energy cost must be paid by favorable interactions between the molecules and by the increased entropy of the solvent.

The translational, rotational, and vibrational entropy of a protein changes weakly as a function of its size and shape. To illustrate this idea, let's consider a simplified model of the protein-ligand association (Equation II-9). We assume that the ligand molecule is much smaller compared to the protein. Under this assumption, the protein term and the complex term will cancel each other, and the free energy change due to translational entropy change will be only relevant to the ligand with an approximate form of:

$$\Delta G_{trs} = -RTln\left[\left(\frac{2\pi mkT}{h^2}\right)^{\frac{3}{2}} V\right], \qquad \textbf{Equation III-17}$$

where R is the ideal gas constant, T is the temperature, m is the molecular weight of the ligand, V is the volume and h is Planck's constant. For a 10-residue long ligand with an average weight of 1300 u, and at temperature of 300K, the standard state translational free energy is 10 kcal/mol. For a ten-fold bigger ligand with 100 residues and weight of 13000 u, the translational free energy increases by 2 kcal/mol. We see that translational entropy change is relatively insensitive to the mass of ligands. Of course, the approximations involved in this derivation are difficult to validate, and experience tell us that size as well as different intrinsic flexibilities of protein structures bring about protein specific terms that so far have been impossible to estimate quantitatively. Hence, to simplify protein–ligand binding and considering that protein domains

are overall structurally conserved, we consider the association entropy change as a constant for ligands with similar size. The complete derivation is shown in Chapter 4 of [27] and Chapter 11 of [26].

6. Binding free energy

In this thesis, we use Equation III-15 to estimate the binding association between folded proteins. Please note, in practice, our folded protein–folded protein docking program using Equation III-18 as scoring function by dropping $-T\,\Delta S_{trv}$, which is a constant (15 kcal/mol).

$$\Delta G = \Delta E_{elec} + \Delta G_{ACE}, \qquad \textbf{Equation III-18}$$

7. *FastContact*: A free energy scoring web server

FastContact is a well-established, freely available tool to estimate empirical binding free energy of folded protein–protein interactions [32,59], which is developed by Dr. Camacho's group. *FastContact* takes into account intermolecular Coulombic electrostatic potential (ΔE_{elec}) and an empirical desolvation contact free energy (ΔG_{ACE}) as we mentioned before. Users submit two proteins in PDB format, and the output is emailed back to the user in three files: one output file, and the two processed proteins. Besides the electrostatic and desolvation free energy, the server reports residue contact free energies that rapidly highlight the hotspots of the interaction and evaluates the van der Waals interaction using CHARMM. Response time is ~1 min. The server has been successfully tested and validated, scoring refined complex structures and blind sets of docking decoys, as well as proven useful predicting protein interactions. *FastContact* offers

unique capabilities from biophysical insights to scoring and identifying important contacts. *FastContact* is available at http://structure.pitt.edu/servers/fastcontact/.

C. APPLICATION OF FREE ENERGY SCORING FUNCTION

1. Capri Target 45

CAPRI is a community-wide experiment to assess the capacity of protein-docking methods to predict protein–protein interactions [58,68,69]. The Hendrick Kim group at the European Bioinformatics Institute (EBI) hosts the CAPRI experiment. In each round, one or more protein–protein complex targets is released and the participant groups submit their blind structure predictions before the deadline based on the known structure of the component proteins and their own docking methods. After the submission deadline, the native complex structure results will be published and the performance of each participant will be ranked by several criteria, such as fraction of native residue–residue contact, the RMSD values of the ligands after superimposing the receptors of the prediction, and the native complex structures. Since first round in 2001, CAPRI has already been a powerful driver for the community of computational biologists who develop docking algorithms. These targets, 52 targets as of Apr. 2011, can be used as a benchmark data set, complementary to Weng's docking benchmark data set [70].

Recently Fleishman et al. have developed a computational method for de novo design of protein binders [71]. This method has successfully produced two proteins that bind to a sterically hindered and, therefore, challenging surface on Spanish Influenza Hemagglutinin (SC1918/H1 HA; hereafter referred to as HA) and, following in vitro evolution 2–4 mutations in the periphery

of each of these interfaces, improved binding to low nano molar dissociation constants. Though encouraging, 71 other designed proteins that were predicted to bind did not experimentally interact with HA, as determined by yeast cell-surface display screening experiments [72], which highlights the limitations in the understanding of protein-binding energetic and their repercussions for the ability to design novel protein functions.

Capri Target 45 is hosted to test the current understanding of interface energetic. Structures of 87 designed complexes that have very favorable computed binding energies, but most do not appear to be formed in experiments, and 120 naturally occurring complexes, from ZDock2.0 and ZDock 3.0 dataset [73,74], are provided. 28 Participants are asked to identify energetic contributions and structural features that distinguish between the two sets.

2. Results and discussion

All 207 protein complexes were first processed through 20x3 energy minimization using ABNR (adopted basis Newton–Raphson) steps and the CHARMM-19 potential with polar hydrogen only, distance-dependent dielectrics ε = 4*r*, and fixed backbone. Then, each protein complex was evaluated by two free energy scoring functions, respectively:

$$\Delta G = \Delta E_{elec} + \Delta G_{ACE}, \quad \textbf{Equation III-19}$$

$$\Delta G = \Delta E_{elec} + \Delta G_{ACE} + \Delta E_{vdw} + \Delta E_{int}, \quad \textbf{Equation III-20}$$

where Equation III-19 is referred as *FastContact* and Equation III-20 is referred as *SmoothDock*. ΔE_{elec} is electrostatic potential and ΔG_{ACE} is desolvation contact free energy. ΔE_{vdw} and ΔE_{int} are the change in van der Waals and internal energy upon binding.

FastContact uses a biophysical meaningful threshold −21.62 kcal/mol. This threshold corresponds to 10^{-5} M (−6.62 kcal/mol) adding in the −15 kcal/mol entropy loss, successfully

screened 69 natural complexes out of 120, with 58% sensitivity and 60 designed complexes out of 87, with specificity 69%. *SmoothDock* discriminated the target with 58% sensitivity (70 out of 120) and specificity 89% (77 out of 87) by using an empirical threshold −79 kcal/mol. The energy scores by *FastContact* and by *SmoothDock* are shown in Figure III-1. It is clearly shown that *SmoothDock* has the best performance to discriminate the designed complexes from the natural complexes. The performance of *SmoothDock* is also represented in ROC curve in Figure III-2 with 77% AUC (area under curve). The sensitivity, specificity, true positive rate, and false positive rate are defined in the following formulas:

$$Sensitivity = \frac{number\ of\ TP}{number\ of\ TP + number\ of FN} = true\ positive\ rate \qquad \textbf{Equation III-21}$$

$$specificity = \frac{number\ of\ TN}{number\ of\ TN + number\ of FP} \qquad \textbf{Equation III-22}$$

$$false\ positive\ rate = \frac{number\ of\ FP}{number\ of\ FP + number\ of TN} \qquad \textbf{Equation III-23}$$

where *TP*, *TN*, *FP*, *FN* are true positives, true negatives, false positives, and false negatives, respectively.

It is known that predictions of empirical scoring functions depend on the underlying molecular modeling technique [75]. This is particularly true for predictions based in co-crystal structures, which capture the optimal complementarity of intermolecular forces, relative to model protein complexes that lack a consistent force field to minimize internal energies. *FastContact* [31,33,58,60], one of the first free energy based scoring functions used to predict protein interactions, reflected this dichotomy. Indeed, on the one hand, it showed almost identical sensitivity and specificity rates when discriminating complex structures in the PDB regardless of whether one accounts for changes in van der Waals (ΔE_{vdw}) and/or internal (ΔE_{int}) energies,

predicted sensitivity rates for ZDock datasets are 58% (69 true positives out of 120 total) and 58% (70 true positives out of 120) for *FastContact* and *SmoothDock*, respectively. On the other hand, the simultaneous discrimination of both co-crystal and designed model structures showed a 20% increase, from a *FastContact* prediction of 69% to 89% in specificity when accounting for ΔE_{vdw} and ΔE_{int}, reflecting the shortcomings of refining backbone rearrangements. It is important to stress that our predictions do not involve any prior knowledge of protein–protein interactions, nor have we made any attempt to incorporate features of the Rosetta scoring function in our analysis.

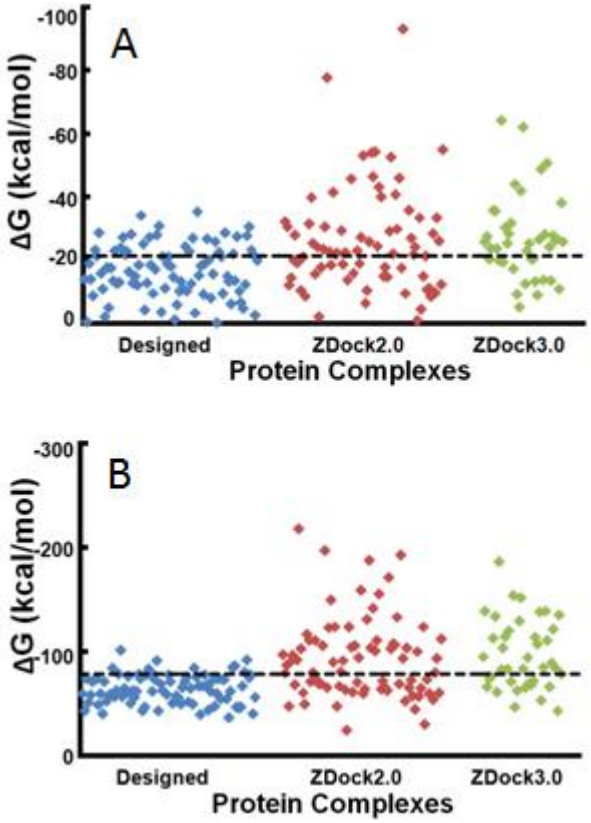


Figure III-1: Free energy scores of Capri Target 45 by FastContact (A) and SmoothDock (B). *Designed* are designed protein complexes. ZDock2.0 and ZDock3.0 are native protein complexes. FastContact provided a discrimination result with 57.5% specificity and 69% sensitivity with a physical −21.62 kcal/mol (-10^{-5} M) threshold. SmoothDock showed discrimination with 58% specificity and 89% sensitivity with an empirical threshold at −79 kcal/mol.

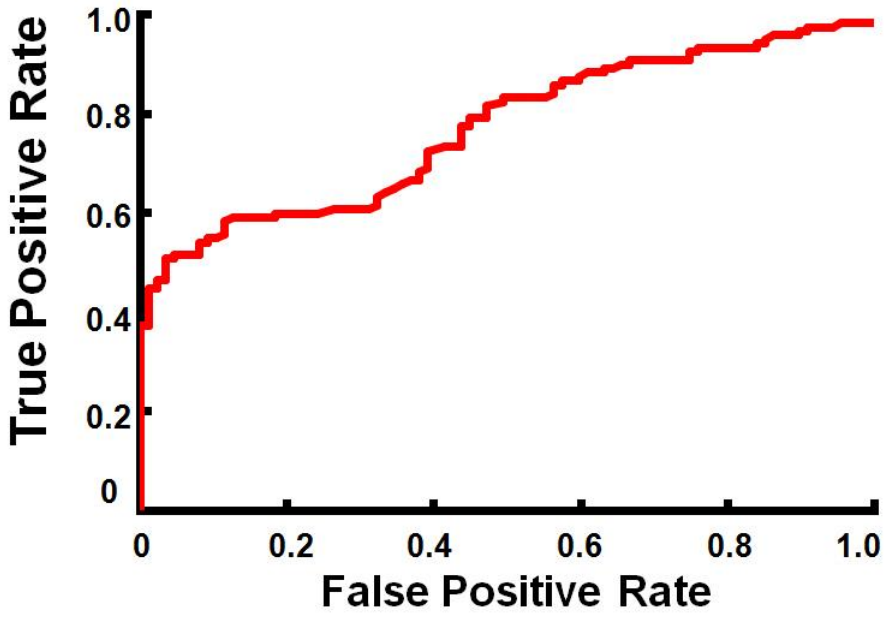


Figure III-2: ROC curve of discrimination results of Capri Target 45 by SmoothDock free energy function. *X*-axis is false positive rate and *y*-axis is true positive rate. Each point showed corresponding sensitivity and specificity by different free energy discrimination threshold. The total area below the ROC curve is 77%.

IV. PREDICTING THE INTERACTIONS BETWEEN PROTEINS AND DISORDERED PEPTIDES

In this chapter, we explain *PepDock*, a novel docking method to predict protein-peptide interactions. We introduce protein-peptide interactions and the design of *PepDock* methodology. *PepDock* has been applied to screening the interactions involving PDZ domains. The discrimination results validate the capabilities of *PepDock* to estimate binding affinity and predict complex structure accurately and robustly. More important, these results help us to explore the mechanism behind PDZ–peptide association, which will be shown in results and discussion. *PepDock* development and application to PDZ–peptide screening are completed by the author and directed by the dissertation advisor.

A. OVERVIEW

1. Interactions between adapter proteins and disordered peptides

In addition to the biological interactions between two structured proteins, there is another class of interactions that involve one structured protein and one intrinsic disordered protein. In fact, the occurrence of unstructured regions of significant size (>50 residues) is surprisingly common in functional proteins [76,77]. The intrinsically disordered proteins are commonly observed in

crucial areas, such as transcriptional regulation, translation, and cellular signal transduction, of which their functional roles have been recognized only recently [4].

In signal transduction, the assembly of proteins into biochemical pathways or networks is typically by the association of autophosphorylated receptor tyrosine kinases with cytoplasmic proteins containing specialized protein modules that mediate formation of signaling complexes [10]. The kinases normally have broad substrates and may be used in many biological interactions. One mechanism to organize the correct repertories of enzymes into individual pathways quickly and precisely is achieved by recruitment of scaffolding proteins, which can localize signaling molecules with certain disordered peptides region to the site of reaction.

The scaffold proteins, also referred as adapter proteins, are usually well structured with conserved binding sites. When disordered peptides from signaling proteins bind to the pocket, they fold into an ordered structure. The procedure of folding and binding begins with a non-specific intermediate that evolves to the fully bound/folded state without dissociation from its target [17]. This mechanism resembles the so called “fly-casting” effect [78], which suggests that non-specific interactions of unstructured regions can enhance the binding rate by having a greater capture radius. Disordered proteins have two features that provide important functional advantages for signaling [19,76,79]. First, disordered regions can bind their targets with high specificity and low affinity. They tune the binding affinity to maximize the specificity of promiscuous interactions [15]. Second, intrinsic disorder promotes binding diversity by enabling proteins to interact with numerous partners.

Typical peptide-binding domains in the signal transduction are SH2, SH3, PDZ, PTB, WW, and 14-3-3 domains (Figure IV-1). One important characteristic among these interactions is that each domain has a conserved binding motif and binds relatively structureless linear

peptides. For example, SH2 domains bind specific phosphotyrosyl residues on activated receptors, SH3 domains bind to poly-proline motifs on a separate set of target proteins, and class I PDZ domain binds peptide with hydrophobic anchor residue (Table IV-1). Hence, a natural question to ask is, "How and why does nature regulate signal transduction through these disordered motifs?"

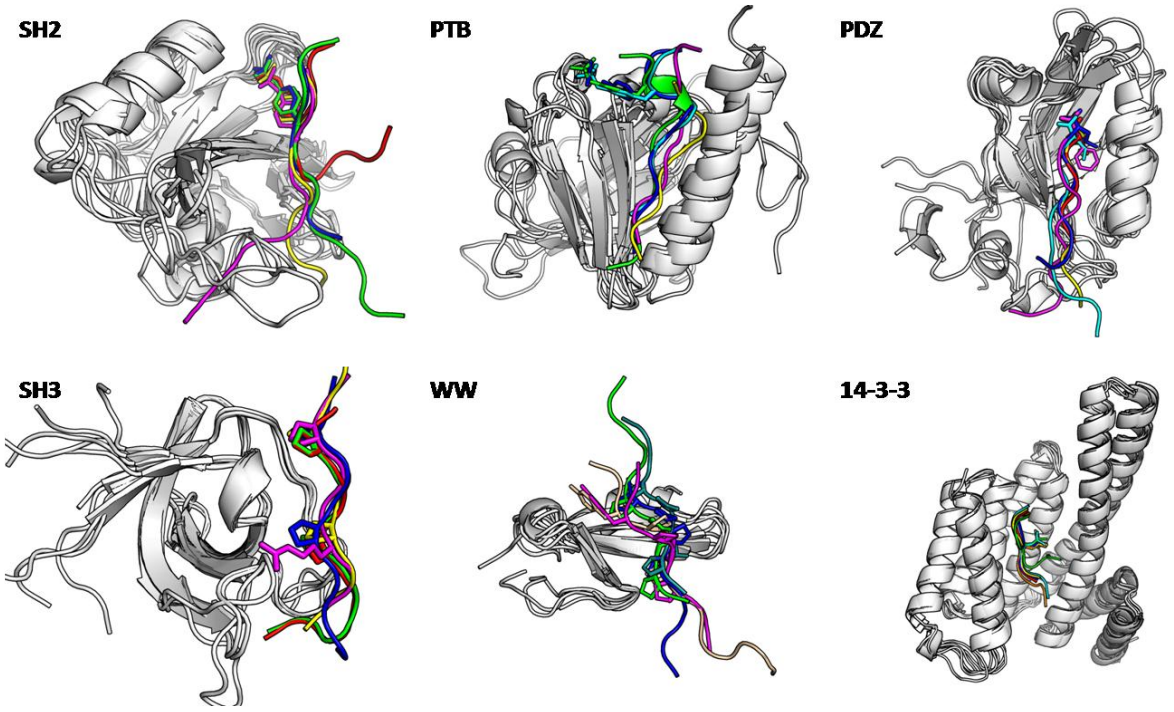


Figure IV-1: Protein modules for the assembly of signaling complexes. Several modular domains (in white) have been identified that recognize disordered peptide motif (in color) with specific sequences on their target acceptor proteins. Different complex structures of same protein module are overlapped by protein modules structure. From top left are SH2 domain (PDBID: 1JYR, 1LCJ, 1SPS, 2CI9 and 3MAZ), PTB domain (PDBID: 1IRS, 1UEF, 2G35, 2YT2 and 3ML4), PDZ domain (PDBID: 1BE9, 1N7F, 2H2B, 3CBX and 3DIW), SH3 domain (PDBID: 1ABO, 1BBZ, 1N5Z, 1W7O and 2AK5), WW domain (PDBID: 1EG4, 2HO2, 2JO9, 2OEI and 2RLY), and 14-3-3 (PDBID: 2BR9, 2C1J, 2C74, 2NPM, 2V7D and 2WH0). All protein modules have conserved binding grooves and are identified to bind disordered peptides with sequence consensus.

Table IV-1: Target peptide sequence consensus of selected protein modules.

Protein Modules	Target Peptide Sequence Consensus
SH2	-Yp-X-X-hy-
PTB	-hy-X-N-P-X-Yp-
PDZ	-E-S/T-D-V-COOH
SH3	-P-X-X-P-X-
WW	-P-P-X-Y-
14-3-3	-R-S-X-Sp-X-P

X: amino acid.
Xp: phosphorylated residue.
Yp: phosphotyrosine.
hy: hydrophobic residue.
COOH: carboxyl-terminus.

2. Interactions of PDZ domains

PDZ domains, an acronym combining the first letters of three proteins – Postsynaptic Density protein (PSD95), Drosophila Disc Large Tumor Suppressor (Dlg1) and Zonula Occludens-1 (ZO-1), are 80–90 residue long molecular scaffolds, which are typically found as tandem arrays in signaling proteins [10,80,81]. They have been shown to mediate protein–protein interactions at the plasma membrane, participating in processes such as cell polarity, motor transduction, ion transport, among others [81]. In general, PDZ domains bind to a short region of the C-terminus of other specific proteins. With close to 180 human proteins containing PDZ modules, these promiscuous domains have emerged as a critical modulator of cell regulation, yet the specificity and extent of their multiple interactions is still poorly understood.

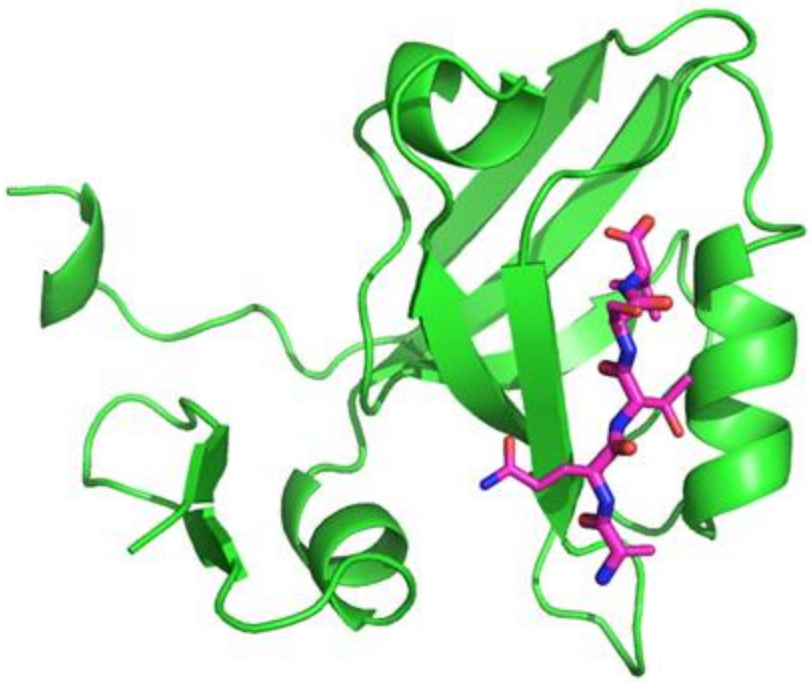


Figure IV-2: Structure of PDZ domain. Ribbon diagram of PDZ3 of PSD-95 that is complexed with a C-terminal peptide from CRIPT (PDBID: 1BE9). The structure demonstrates the 5 β-strands and 2 α-helices (green) with the peptide (cyan) binding as a β-strand between the helix and the strand.

A significant advancement in our understanding of PDZ adapters was revealed by crystallographic studies that demonstrated that the recognition motif includes the C-terminal of target proteins [9,82,83,84]. A classic example is MacKinnon and collaborators' structure of the third PDZ domain of the postsynaptic density protein 95 (PSD95-3) bound to the C-terminal of a cytoskeleton protein (CRIPT) [9]. PSD95 is a member of the membrane-associated guanylate kinase family that contains three PDZ domains, whose binding targets include the NR2 subunits of the NMDA-type glutamate receptor [85] and Shaker-type K channels [86]. Based on the co-crystal of CRIPT and PSD95-3, it was shown that a four-residue long CRIPT strand binds to PDZ to form an anti-parallel beta sheet. The strand is anchored [30] by the position "0" C-terminal hydrophobic residue side chain (by convention, residues of PDZ ligands are numbered starting with the last C-terminal residue as occupying the position "0", penultimate residue as

occupying the position −1 and so forth, moving along the ligand sequence from C- to N-terminus) projecting into the PDZ binding groove and is further stabilized by two backbone hydrogen bonds to the COOH-terminal group. Viable PDZ binding interactions have been further characterized by either a serine/threonine or a hydrophobic (Φ) reside at position "*−2*". These two C-terminal consensus motifs, S/T-X-Φ_0 and Φ-X-Φ_0, are referred to as Class I and II, respectively [11]. A third class of PDZ domains, referred to nNOS, prefers negatively charged amino acids at the "*−2*" position [87]. The above notwithstanding, it is important to emphasize that the consensus motif (*0* to *−3*) is not enough to secure binding and several other residues beyond *−3* are needed to form a stable complex [88,89].

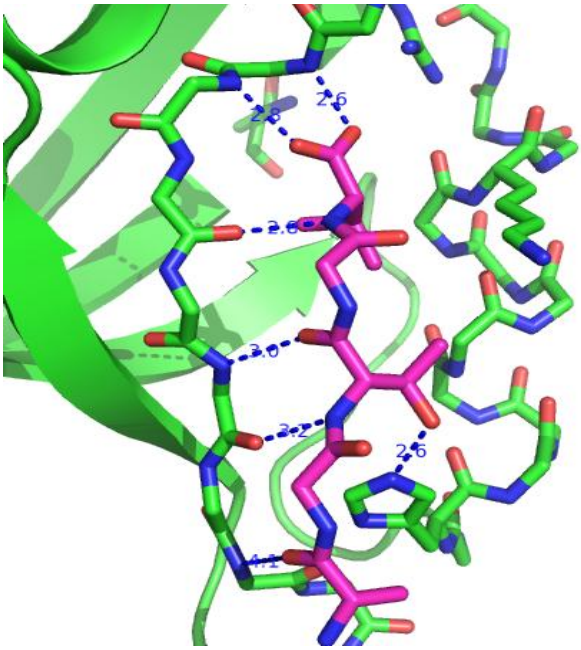


Figure IV-3: Conserved binding site of PDZ3 domain of PSD-95, a class I PDZ domain. PDZ domain binds peptide strongly through backbonebackbone hydrogen bonds. Dashed lines indicate the hydrogen bonds between peptide carboxylate group and PDZ backbone, between peptide backbone and PDZ backbone, and peptide side chain with PDZ side chain.

3. Screen the interactions between PDZ domain and disordered peptides

Despite the modular nature of the PDZ interactions, detection of their binding partners has remained elusive. Traditional in vivo and in vitro assays are hampered by the relatively weak binding constants (~μM) [90], which are common in signal transduction, and the intrinsic disorder of the C-terminal peptides. To overcome some of these problems, novel proteomics techniques of PDZ-peptide interactions have been developed, resulting in a number of validated sequence specific targets of PDZ domains. Kurakin et al. [88] used a semi-quantitative ELISA assay to screen the relative affinities of both 126 natural and 95 phage selected artificial peptides that included the C-terminal consensus motifs against PSD95. This experiment confirmed the aforementioned recognition pattern of amino acids in the window "0" to "−3", and showed that other amino acids in the window "−4" to "−7" also play a role in binding. MacBeath's group used microarrays and quantitative fluorescence polarization to study the binding selectivity of 157 mouse PDZ domains against 217 genome-encoded peptides [91]. Madsen et al. used a similar technique to assess the affinity of PICK-1 PDZ with its partners [92]. Pei's group developed a new methodology to synthesize and screen peptide libraries containing free C-termini and applied the method to identify consensus recognition motifs of PDZ domains [93].

Complementary to proteomics efforts, researchers have used computational methods to gain biophysical insights into the binding of peptides [22,23,94,95] to specific scaffolds. Based on MD simulations, Basdevant and co-workers [94] concluded that PDZ domain interactions are characterized by favorable non-polar contributions and negligible electrostatics, a result that is not easy to reconcile with at least some PDZ domains for which Lys and Arg salt bridges appear to be important for binding [88]. Niv and Weinstein [22] developed a simulated-annealing procedure to dock flexible peptides. They tested their methods in 5–8 residue long peptides

binding to known structures, obtaining excellent docking to bound PDZ domains. Complexes were scored using the CHARMM interaction energy, including the internal energy of the peptide, but not its entropy. Wang's group [23] attempted to further include entropic effects using MD simulations. Binding free energies of 15 SH3 binding peptides did not compare well with experiments, though the method showed some success correlating experimental and predicted changes in free energy from closely related sequences. No structural validation was presented for the predictions. Missing is a structure-based approach capable of predicting novel PDZ-ligand complexes and their specificities. We note that traditional docking methods can predict docked models [22,96]. However, the lack of accurate estimates of changes in entropy, van der Waals interaction, and internal energy upon binding has so far limited the successful prediction of de novo physical interactions.

A different computational approach for predicting PDZ binding peptides involves machine-learning techniques [20,97]. In particular, MacBeath's group used their experimental interaction data to construct a position-specific scoring matrix to predict PDZ–peptide interaction, resulting in optimal sensitivity–specificity ratios of 70–80% [20]. These methods, however, are limited by the training data set.

B. *PEPDOCK*: AN NOVEL COMPUTATIONAL METHOD TO PREDICT INTERACTIONS BETWEEN PROTEINS AND DISORDERED PEPTIDES

Although there are several computational methods to screen PDZ–peptide interactions, there are still several concerns: (1) there is no good scoring function that can give a reasonable estimation of binding free energy. Most groups use interaction free energy change or specific scoring

function and compare their score to the benchmark to screen binding affinities. (2) Screening a new peptide may require much time. Missing is a structure-based approach to predict novel PDZ-ligand complexes and their specificities. Here, we present a structure-based computational method to predict interactions of disordered peptides with applications to PDZ adapter proteins. The method, referred to as *PepDock*, uses a selected known PDZ complex structure as template for the C-terminal recognition motif, and a novel semi-flexible docking approach and scoring function to minimize the full binding free energy under the assumption that peptides are fully disordered in the unbound state. Backbone flexibility is accounted for by developing a comprehensive library of peptide backbones extracted from molecular dynamics (MD) simulations, and side chain conformations are sampled by using Dumbrack's rotamer library [98]. Strain is maintained below a self-consistent optimal threshold to prevent unrealistic conformations, while van der Waals cancelation [30,61] is also enforced by a self-consistent threshold [59]. Each docked conformation is scored based on an atomic empirical free energy function that includes electrostatics, desolvation, and full entropy loss [60,99,100] upon association.

1. The methodology of *PepDock*

The biggest differences between flexible peptide docking, which is docking flexible peptides to structured receptor proteins, and regular protein–protein docking are that a conformation ensemble of peptide is needed to describe the dynamics of the peptide, and searching the optimal complex of the peptide ensemble with receptor protein requires huge computational work compared to regular protein–protein docking. In general, the scheme of flexible peptide docking can be divided into four stages: preprocessing, rigid-docking, refinement, and scoring [101]. In

the preprocessing stage, the flexibility of peptide is sampled by Molecular Dynamic (MD) or Monte Carlo (MC) simulation and the generated conformation ensemble will be docked to the receptor protein in the next step. When the binding site of receptor protein is unknown, the computational complexity and consuming time will increase dramatically with the increasing size of conformational ensemble comparing. However, for peptides binding to adapter proteins, the task is simpler. With crystal structures of the complexes, it is possible to restrain the docking in the known binding sites and reduce the conformation sampling space of flexible peptides.

The advantages of *PepDock* are these: (1) It samples the peptide dynamic of one given amino acid sequence by sampling its backbone and side chain conformations separately. The peptide backbone model is extracted from a shared peptide backbone library and side chain conformation is iterated by using Dumbrack's rotamer library. This design allows us to use one predefined peptide backbone library to approximately model the backbone of any peptides binding to one group of PDZ domains with consensus sequence. (2) One selective known PDZ–peptide complex is used as the structural template to predict new PDZ–peptide complexes. The PDZ from the template complex and PDZ in the prediction are not necessarily identical, but must have certain structural and binding similarity. By structural analysis of PDZ complexes from the Protein Database, we have grouped 55 different PDZ domains into several groups and use the center PDZ complex of each group to predict the interactions of PDZs from the same group. For example, PSD95-3 domain complex structure (1BE9) is used as the template to predict PDZ interactions of ZO1-PDZ1 domain, SAP97-PDZ3 domain, GRIP1-PDZ6 domain, etc. These two design characteristics dramatically simplify computational complexity and enable *PepDock* to predict one PDZ interaction in 30 minutes. (3) *PepDock* models entropy changes upon binding by adding residual contribution along the peptide with the simplified conditions that peptides are

completely disordered before association and bind to PDZ domain following the "Anchor Model". Residual entropy changes are inferred from experimental and computational analysis [99,100] (See IV C.2 for detail). By incorporating entropy change into its scoring function, *PepDock* provides an estimation of absolute binding affinities with successful performance statistics.

In design, *PepDock* includes three components: *preparing*, *docking*, and *searching* (Figure IV-5). For one adapter protein domain, preparing will sample the peptide backbone by using MD or MC simulation. Docking will dock the peptide backbone models into the binding site of the adapter protein domain, iterate side chain conformation of 20 amino acids on each peptide residual position, evaluate the individual residual free energy contribution, and save it into the scoring function database. For any input target peptide sequence, searching will search the lowest free energy combination from the scoring function database and then assemble the complex according to the combination. Please note that preparing and docking, which consumes most computational time, are required to run only once before predicting any peptide sequence binding to one specific PDZ protein domain.

Preparing I: generate the peptide backbone models. Based on the characteristics of PDZ protein domain, choosing 10-residue long typical binding peptide sequence alone (not with the PDZ) as the input, we used 10 ns Molecular Dynamic (MD) simulation to generate peptide dynamic sampling snapshots, extract the backbone structures from peptide snapshots, cluster and select group center structures as the peptide backbone models in the library (Figure IV-4).

Preparing II: dock the peptide backbone models within the library onto the target PDZ based on the template complex structure. Target PDZ domain structure is aligned to PDZ

structure in the template complex and peptide backbone models are overlaid to the backbone between *0* to *−4* of the C-terminal peptide in the template.

Preparing III: establish van der Waals (vdW) thresholds. Docked backbone complexes are energy minimized as described below. The resulting distribution of the vdW is used to set self-consistent energy threshold feasible complexes, above which docked complexes are eliminated and the backbone model is considered as infeasible. This approach circumvents the problem of optimizing backbone conformations by considering only complexes that do not build strain or clashes in the complex.

Docking: sampling amino acid side chain conformations and generating scoring function database. For each residual position on the docked peptide backbone, we iterate all 20 amino acids in all rotamer conformations [31]. Each rotamer is minimized locally on the docked backbone complex and its contribution to binding is estimated by the modified *FastContact* [34] free energy scoring function. The score is saved into a residual scoring database, which has all feasible backbones (~300), all amino acids (20) and their rotamers (~10 per amino acid) for 10 residue positions, i.e., a total of around 600,000 free energies.

Searching I: search for the rotamer combination of target peptide sequence on one feasible backbone model. Complex structure of the target peptide sequence on one feasible backbone model is predicted by searching for the rotamer with the lowest binding free energy at every position along the peptide. The structure is energy minimized and checked by the vdW threshold, side chain clashing, and hydrogen-bond competition. If any check fails, the complex is discarded and the next lowest free energy model is built and checked, and so on.

Searching II: search for complex structure with the lowest free energy. Repeat step Searching I for all feasible peptide backbone models. The final predicted complex structure is the

one with the lowest binding free energy. Plot free energy landscape based on rotamer combinations in the residual scoring matrix. Note that the minimum binding free energy can occur for any number of residues up to a maximum of ten.

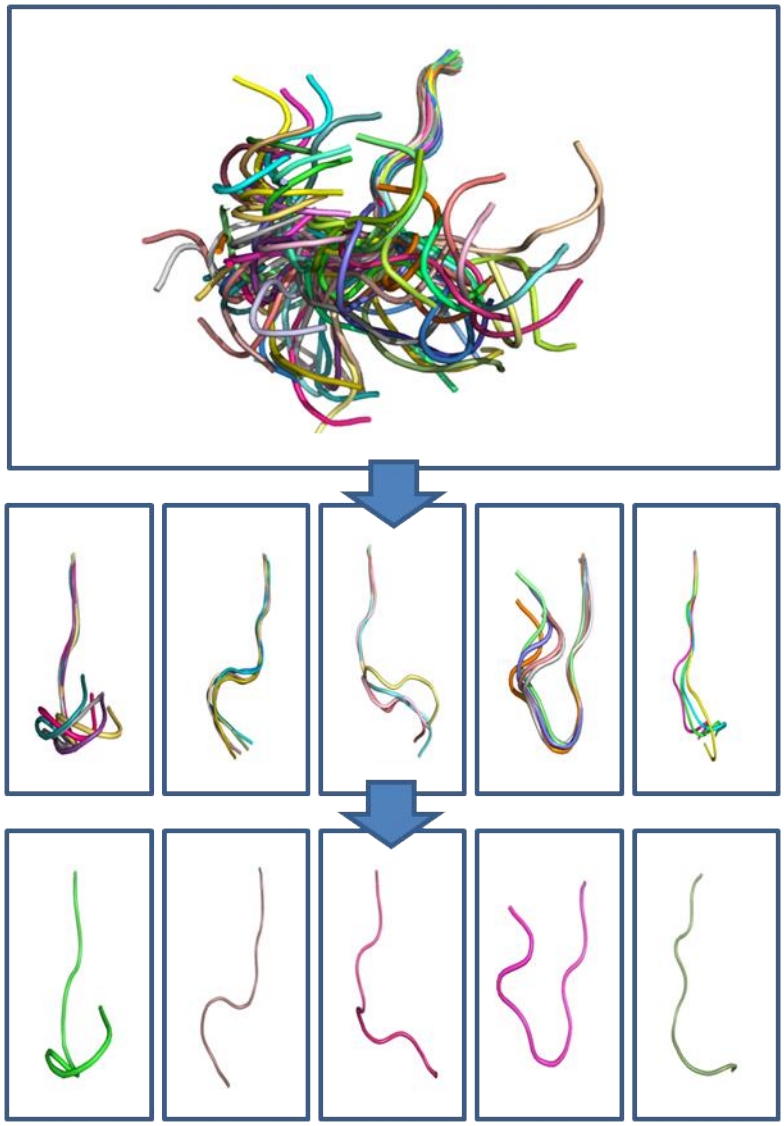


Figure IV-4: Peptide backbone model library. The peptide backbone models are extracted from 8 ns of total 10ns MD simulation of peptide alone with C-terminal motif ("0" to "−4") backbone restrained. The snapshots are grouped into different groups by an iterative clustering algorithm with 1.8 A RMSD threshold. The center snapshots of each cluster are selected as the backbone models. We repeat above steps with by three strong binding peptide sequences against interested PDZ domains and generate a backbone library with ~500 models.

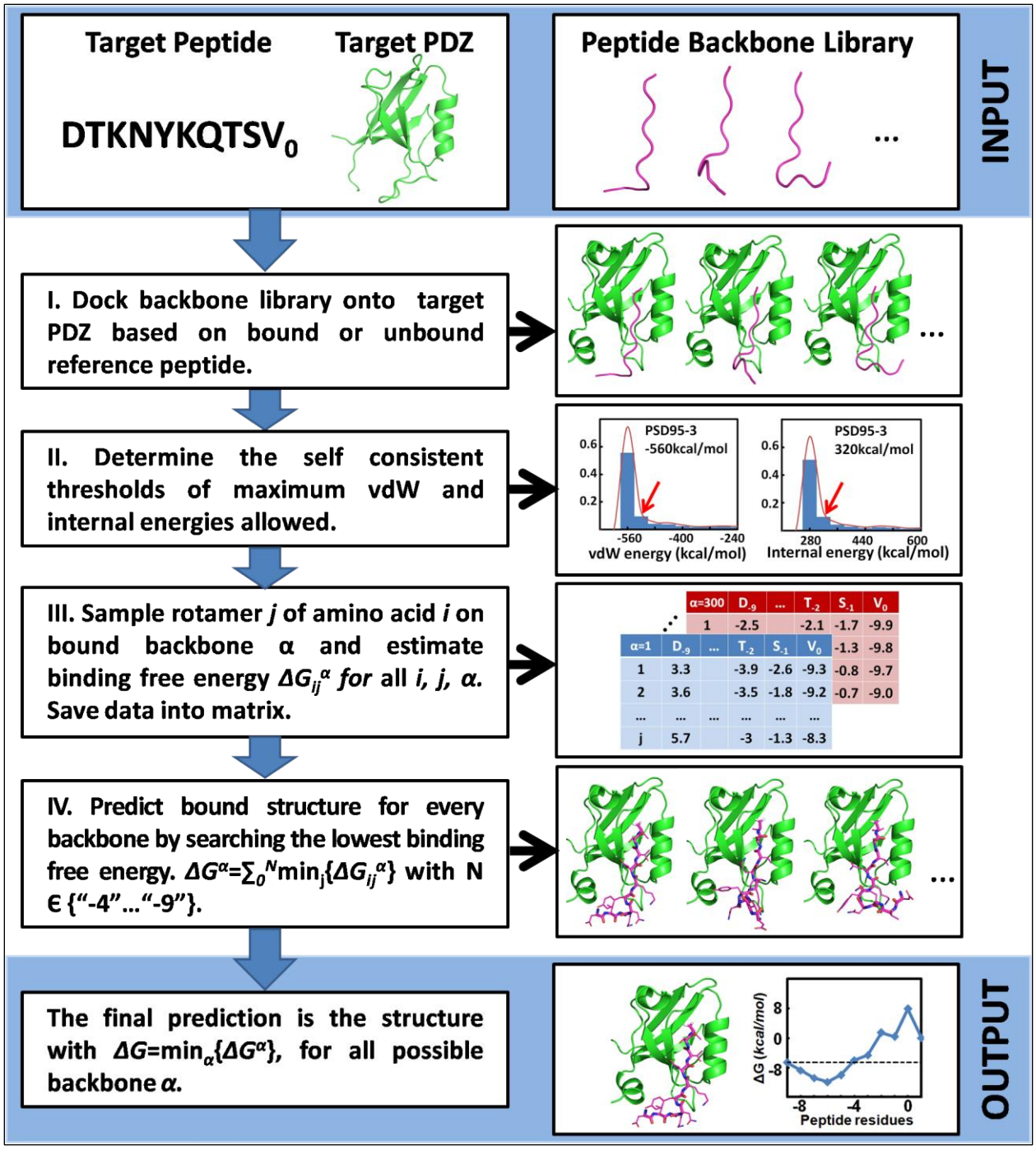


Figure IV-5: Flow chart of *PepDock* methodology. **I**. Dock peptide backbone models into the protein domain and determine the vdW energy threshold. **II**. Sample amino acid side chain conformations on the docked peptide backbones and estimate the binding free energy contributions by scoring function. **III**. For one target peptide sequence, search the optimal combination of side chain conformations with the lowest free energy on each backbone models. **IV**. Search the most optimal result in all docked complex models.

Template complex structure: We use the structure of the complex of PSD95-3 domain with CRIPT peptide as the template to predict PDZ–peptide interactions (PDB id, 1BE9), not only because PDZ95-3 is the most studied PDZ domain, but also it is the best template to represent the same class of PDZ domains in the PDB database (see discussion for more detail). Also, if the target adapter protein has a crystal structure in complex with some peptide in PDBs, this complex structure, referred to as *bound template,* will be used as the template for the target protein. Otherwise, the representative PDZ–peptide complex in each PDZ group from our structural alignment study will be chosen as the template, as *unbound template*. For example, we use PDB 2H2B as the bound template for screening the interactions of target ZO1-PDZ1 domain, and also use PDB 1BE9 as the unbound template for ZO1-PDZ1 domain.

Semiflexible peptide backbone models: Based on the structural similarity of the bound C-terminal motif of PDZ binding peptides, we assume that the C-terminal recognition motif resembles the peptide in the template complex structure. Due to the high similarity of PSD95-3 to other PDZ domains, we use the bound C-terminal CRIPT peptide [8] between residues "−4" and *"0"* as a template for our docking studies. Based on this assumption, we assembled a backbone library of 15 residue long peptides from equilibrium snapshots of MD simulations in explicit solvent [44], where the backbones of the 5 residues at the C-terminal end (except C-α) were restrained as in CRIPT (forming an anti-parallel beta sheet with PSD95-3). The force constant of the harmonic constraint was set to 2.4 kcal/mol/Å^2. We ran 10 ns MD on three different sequences that have the strongest affinities to PSD95-3 twice, with and without the extra hydrogen at the amino terminal, keeping 2000 snapshots from the last 8 ns of each simulation (MD protocol is as in [30]). Finally, clustering the backbones of the last 10 residues (after superposition of the 5 restrained C-terminal residues) of each MD snapshot using a 1.8 Å

radius resulted in 559 representative backbone cluster centers. We also tested extracting the backbone library from MDs with peptide residues from "−3"-to-"0" restrained. However, these backbones were consistently clashing or moving away from the PDZ, resulting in poor complexes.

Energy minimization: I use CHARMm27b3 [43] to compute energy minimizations after performing 60 (3 rounds of 20 steps) adopted basis Newton–Raphson (ABNR) steps of fixed backbone energy minimization. The distance threshold of non-covalent interactions is 15 Å, and the long-range electrostatic screening is calculated with a distance-dependent dielectric of $\varepsilon = 4r$.

C. BINDING FREE ENERGY FUNCTION OF FOLDED PROTEIN-DISORDERED PEPTIDE ASSOCIATION

Compared to the association of two folded proteins, which is assumed no conformational change occurred upon the formation of a complex, the association of flexible peptide and protein involves a disorder/order transition in the peptide fragment. The bound peptide segment has a unique conformation, but the unbound peptide fragment has multiple conformations (Figure IV-6). If peptide is fully disordered, one needs to account further for ΔS^{pep}, the folding entropy loss of the flexible peptide:

$$\Delta G = \Delta E_{elec} + \Delta E_{vdw} + \Delta G_{des} + \Delta E_{int} - T\,\Delta S_{trv} - T\,\Delta S^{pro} - T\,\Delta S^{pep}, \quad \textbf{Equation IV-1}$$

$$\Delta S^{pro} = \Delta S^{pro}_{surface}, \quad \textbf{Equation IV-2}$$

$$\Delta S^{pep} = \Delta S^{pep}_{bb} + \Delta S^{pep}_{buried} + \Delta S^{pep}_{surface}, \quad \textbf{Equation IV-3}$$

where ΔS^{pro} represents the protein entropy change, which includes only surface side chain entropy change since it is folded. ΔS^{pep} is the entropy change of the flexible peptide upon binding, which includes backbone (ΔS_{bb}^{pep}), surface side chain, $\Delta S_{surface}^{pep}$, and buried side chain ΔS_{buried}^{pep} entropy change. Freire, Amzel, and collaborators carefully estimated these values for each residue [99,100]. Hence, a fully consistent binding free energy for docking disordered peptides can be written as

$$\Delta G = \Delta E_{elec} + \Delta E_{vdw} + \Delta G_{des} + \Delta E_{int} - T\,\Delta S_{trv} - T\,\Delta S_{surface}^{pro} - T\left(\Delta S_{bb}^{pep} + \Delta S_{buried}^{pep} + \Delta S_{surface}^{pep}\right). \quad \textbf{Equation IV-4}$$

And with the assumption of negligible internal energy and Van der Waals cancelation, which are used in the folded protein–protein association, it can be simplified to

$$\Delta G = \Delta E_{elec} + \Delta G_{ACE} - T\Delta S_{trv} - T\left(\Delta S_{bb}^{pep} + \Delta S_{buried}^{pep}\right), \quad \textbf{Equation IV-5}$$

$$\Delta G_{ACE} = \Delta G_{des} - T\,\Delta S_{surface}^{pro} - T\,\Delta S_{surface}^{pep}, \quad \textbf{Equation IV-6}$$

where ΔG_{ACE} is the atomic contact energy term, as we mentioned before.

In principle, a complete binding free energy function should also contain a term reflecting the internal energy change associated with the disorder-to-order transition of the peptide. However, for simplicity, we take the unfolded state to be highly extended, and, because the bound state is also extended peptides, the internal free energy difference will be small compared to other terms in the expression.

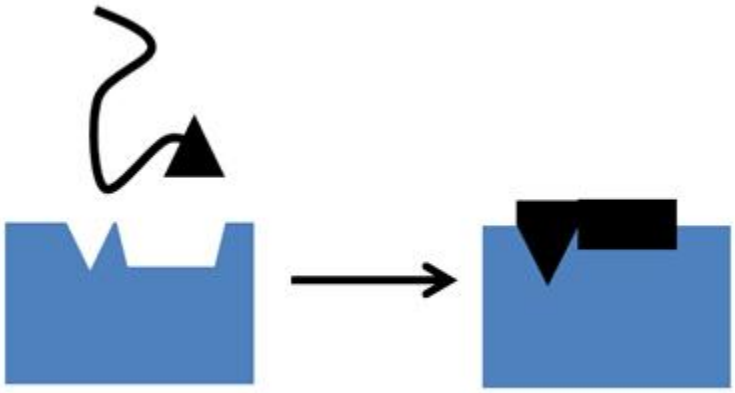


Figure IV-6: Folded protein–disordered peptide association. Before binding, the peptide is flexible without unique 3-D structure; while, after binding, the peptide folds into a specific structure in the binding groove of its interaction partner protein.

1. Implicit solvent model

Electrostatic interactions are obtained by a simple Coulombic potential with the distance dependent dielectric of 4r (Equation III-11). Protein molecules exist in solution surrounded by water molecules that promote their shapes and stabilities. These water molecules decrease the potential of mean force of a salt bridge bond, a hydrogen bond, as well as modulate solvation forces. Recent studies show that introducing a solvation factor to describe the role of water molecular in protein–DNA interactions can significantly improve the accuracy of binding free energy estimation [102,103]. To incorporate the solvation effect and minimum distance in the salt bridge bond interactions, we apply following factor in our free energy calculation:

$$\Delta E'_{elec} = (1 - \lambda_w) * \Delta E_{elec}, \qquad \text{with}$$

$$\lambda_w = \begin{cases} 0.4, & \text{when } \Delta E_{elec} \leq -4 \text{ kcal/mol} \\ 0, & \text{when } \Delta E_{elec} > -4 \text{ kcal/mol} \end{cases}$$

Equation IV-7

where λ_w is the solvation factor, ΔE_{elec} is the coulumbic potential, and $\Delta E'_{elec}$ is the solvated electrostatic potential. The threshold of λ_w, -4 kcal/mol, is choosen as the energy corresponding to the maximum salt bridge bond distance 4 Å on Figure IV-8. To determine the optimal value of λ_w, we iteratively tested the statistical performance of our scoring function in the discrimination of 126 human peptides binding against PSD95-3 domain (see section IV D.2 for more detail) with different λ_w value. Our results show that $\lambda_w = 0.4$ with criteria $\Delta E_{elec} \leq -4$ kcal/mol corresponds to the maximum sum of the sensitivity (Equation III-21) and specificity (Equation III-22) of discrimination and makes sure $\Delta E'_{elec}$ is a smooth function. In addition, $\lambda_w = 0.4$ is consistent with the factor in DNA-protein interaction [102,103].

One acid group (GLU, ASP) residue can make only one salt bridge contact with a basic group (HIS, LYS, ARG) residue. To eliminate double counting effects when two basic group residues, e.g., LYS and ARG, are approaching one acid residue, e.g., GLU, we have applied the following rule when iterating all atoms to calculate electrostatic energy (Figure IV-7):

- Only one basic residue can form a salt bridge bond with one acid residue at one time.
- If two basic residues approach one acid residue in a certain distance, 4 Å between heavy atoms, the basic residue with stronger interaction is kept and the other one is neglected.

After introducing the solvation factor and salt bridge contact rule into our scoring function, we observed that our peptide–protein docking methodology has improved its free energy estimation accuracy and its discrimination performance with higher specificity and sensitivity (91% and 74%) than before (95% and 60%) (Figure IV-10, Figure IV-11).

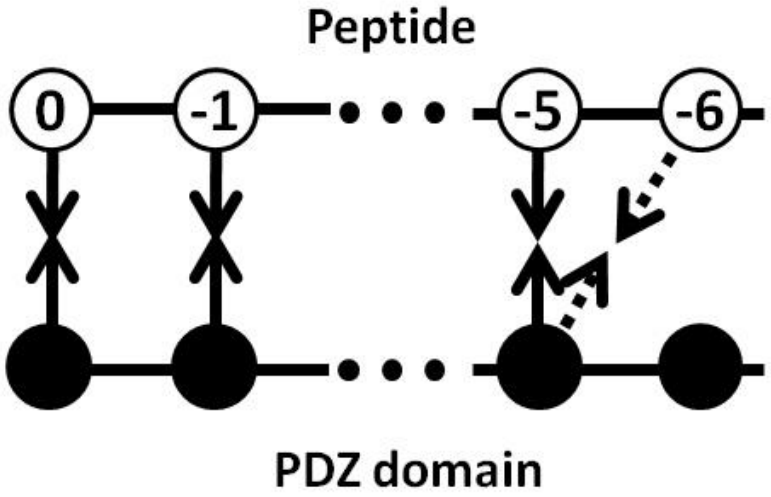


Figure IV-7: Eliminate double counting of salt-bridge bonds. When two basic group side chains try to approach one acid group, the scoring function will compare the strength of these two possible salt bridge interactions and only count the one with larger free energy contribution and ignore the other one. This consideration highly improves the accuracy of the scoring function.

Sensitivity testing of dielectric parameters is conducted to access the effect of the dielectric model parameter on the performance of free energy scoring function. Distance dependent dielectric model, $\varepsilon = \alpha r^{\beta}, \alpha = 4$ and $\beta = 1$, is an empirical function derived from experimental data. We performed a sensitivity analysis of α and β. Based upon one factor at a time methods of local sensitivity analysis, we varied α (4 ± 0.2) and β (1 ± 0.1) on each a time, keeping the other fixed. The average free energy landscapes of top 11 strong binders and bottom 20 weak binder peptides with different parameter values are shown in Figure III-7. Results show that four free energy scoring functions with different parameter perturbations can clearly discriminate the strong and weak binders and the maximum binding free energy variance comparing to standard function is less than 1.6 kcal/mol. This testing proves that distance dependent dielectric model with α = 4.0 and β = 1.0 is robust in free energy estimation.

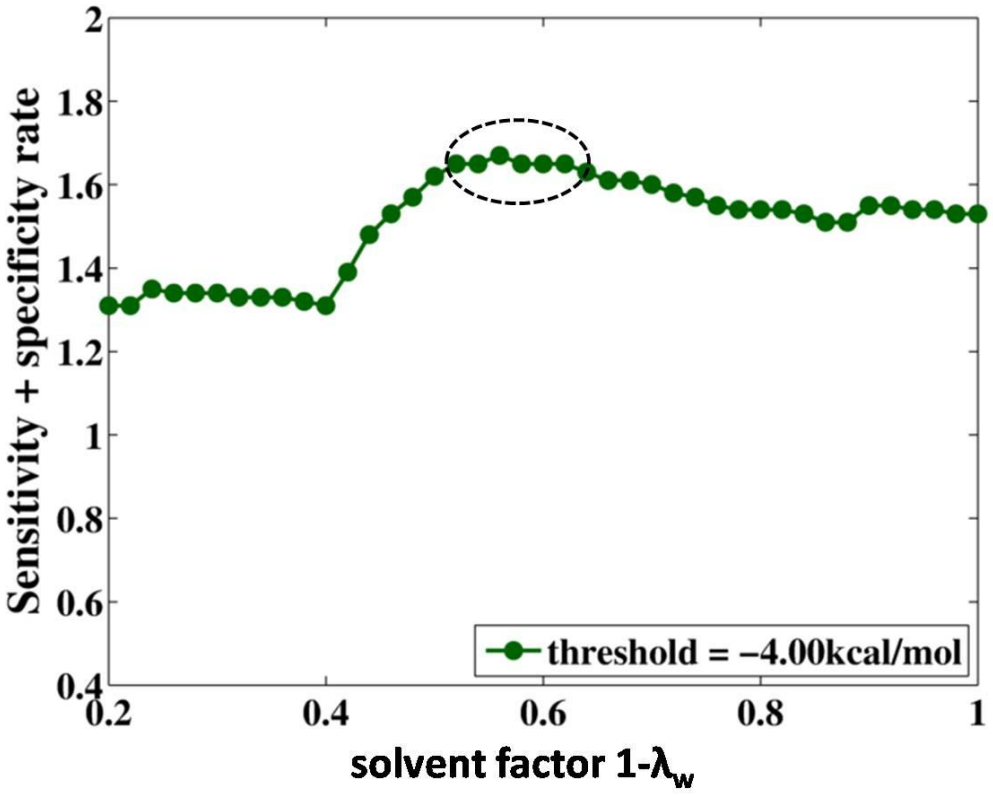


Figure IV-8: Searching optimal solvent factor. To determine the optimal value of solvent factor, we iteratively calculate the specificity and sensitivity of PSD95-3 domain binding to 126 human peptides (see Chapter IV for more detail) with different values of λ_w and threshold. We change λ_w from 0 to 1 while fixing the threshold equal to −4.0 kcal/mol, and find the sum of sensitivity and specificity reach its maximum and the electrostatic energy conforms a smooth function around $(1 - \lambda_w) = 0.6$ and threshold = −4.00 kcal/mol. The result $1 - \lambda_w = 0.6$ is consistent with the value of solvent factor in the study of DNA-protein interactions [102,103].

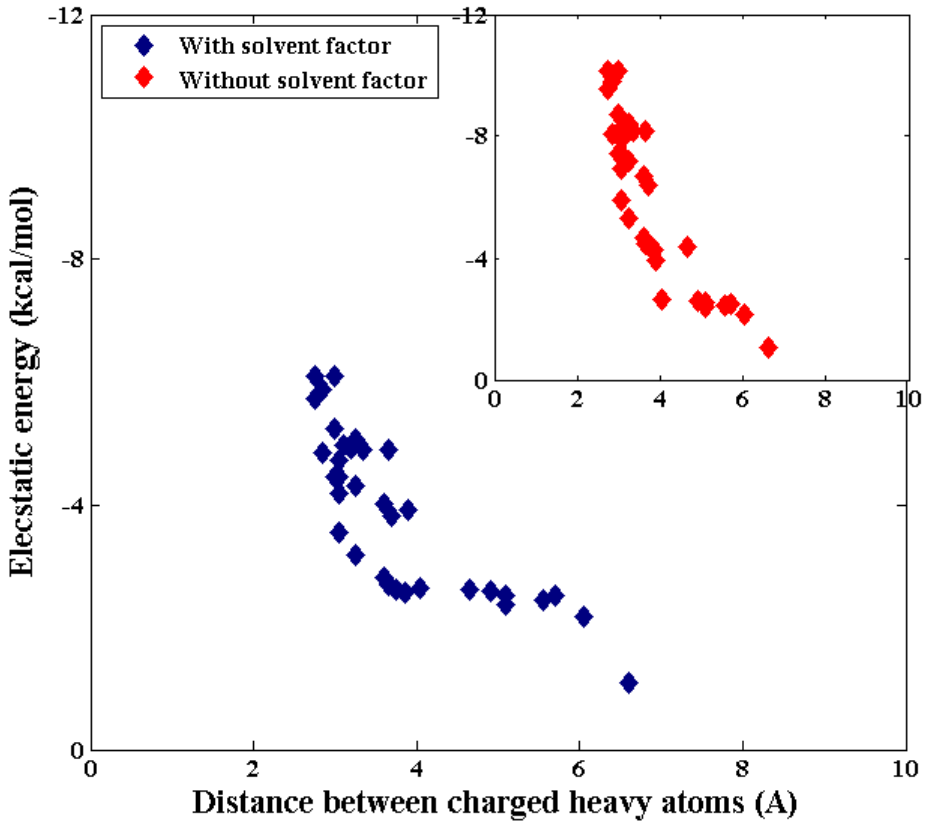


Figure IV-9: Comparison of electrostatic energy function with/without solvent factor. Examples of electrostatic energy calculation between acid group and basic group in the discrimination testing of PSD95-3domain binding to 126 human peptides (see Chapter IV for more detail) with and without solvent factor are shown in blue and red color respectively. The electrostatic energy function with solvent factor, $(1 - \lambda_w) = 0.6$ and threshold = −4.00 kcal/mol, conforms a smooth function versus the distance between heavy atoms in the acid and basic amino groups.

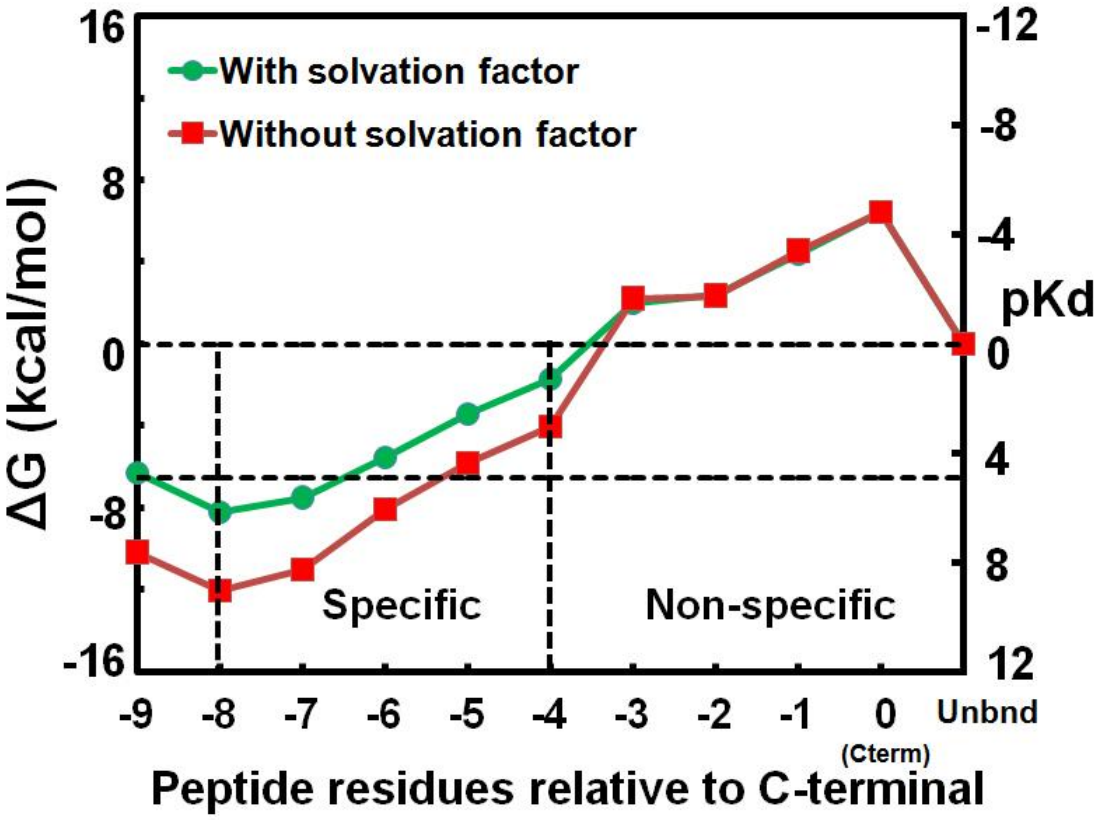


Figure IV-10: Comparison of free energy estimation before and after improvement. The average binding free energy changes of top 11 peptides binding to PSD95-3 domain calculated before improvement (in green), and after improvement (in red) show that applying solvation factor to salt-bridge bond and eliminating salt-bridge double counting improve the accuracy of free energy estimation.

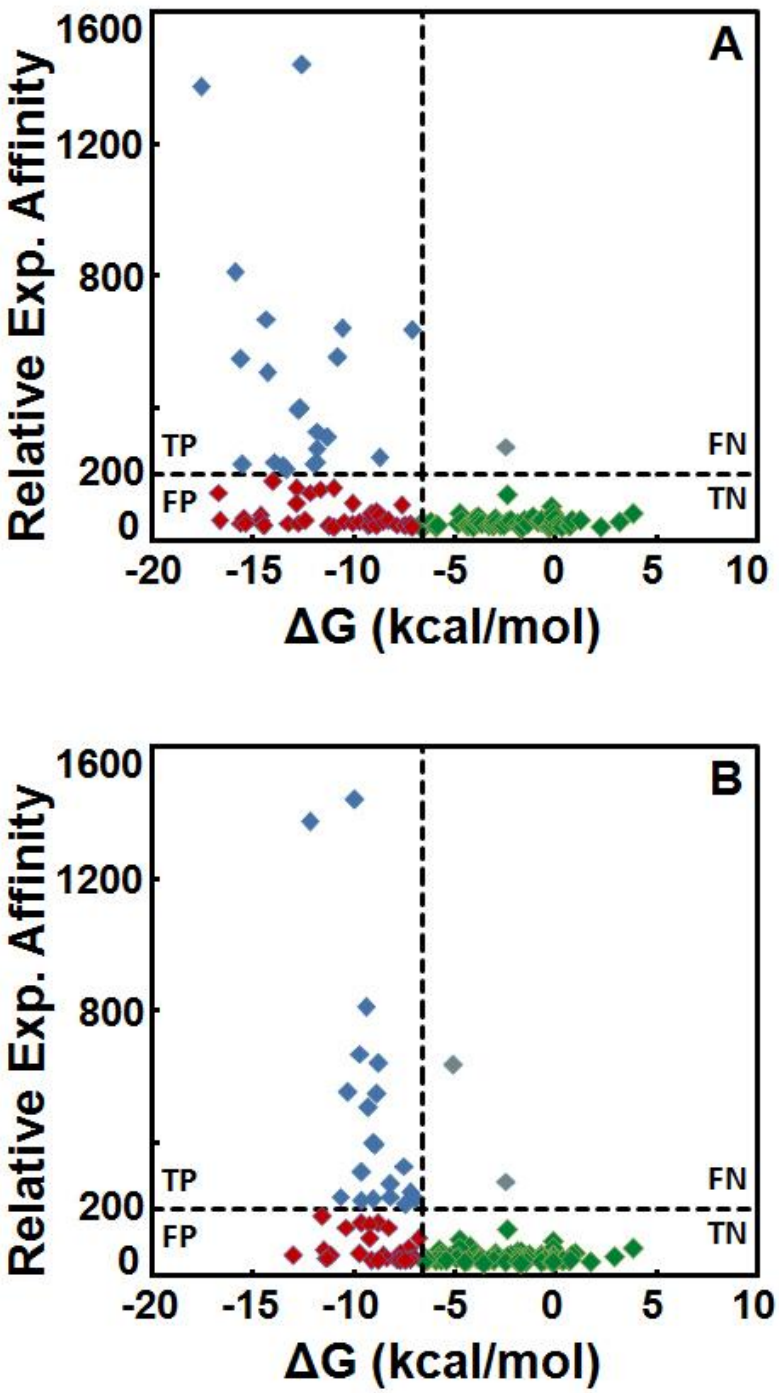


Figure IV-11: Comparison of free energy estimation before and after improvement. Scatter plots of experimental affinity versus calculated binding free energy of human peptides binding to PSD95-3 domain before improvement (A) and after improvement (B) show that applying solvation factor to salt-bridge bond and eliminating double counting not only improve the discrimination performance with higher specificity and sensitivity, but also improve the accuracy of free energy estimation.

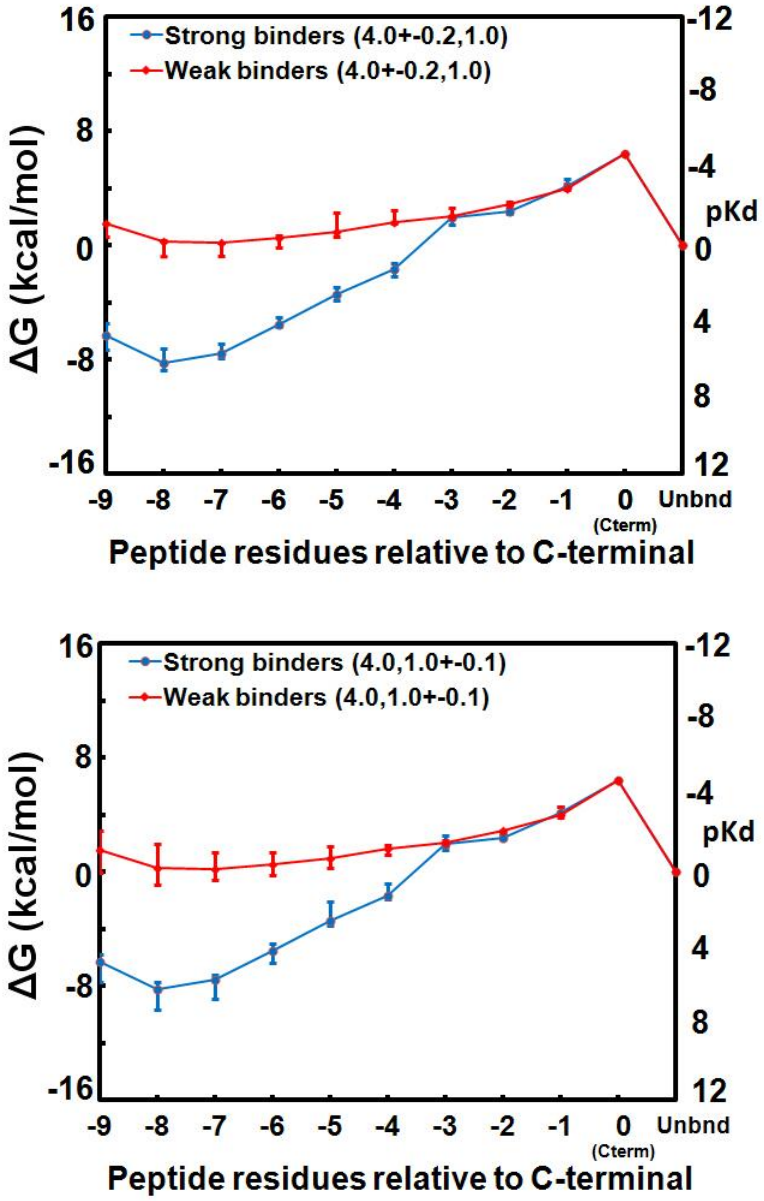


Figure IV-12: Sensitivity analysis of dielectric parameter on the performance of free energy scoring function. Plots show the average free energy landscape of top 11 strong binder (blue color) and bottom 20 weak binders (red color) with distance dependent dielectric parameter α and β $(4 \pm 0.2, 1.0)$ in top plot and $(4.0, 1.0 \pm 0.1)$ in bottom plot. Top error bars are $\alpha = 4.0 + 0.2$ or $\beta = 1.0 + 0.1$ while bottom error bars are $\alpha = 4.0 - 0.2$ or $\beta = 1.0 - 0.1$. Results show that free energy scoring function with parameter perturbations can clearly discriminate the strong and weak binders, and the maximum energy variance is less than 1.6 kcal/mol.

2. Entropy change

In both folded protein–folded protein association and folded protein–disordered peptide association, we assume that folded protein is a rigid protein that has only the side chain change on the contact surface upon binding. Therefore, we need to consider only $\Delta S^{x}_{surface}$, which is already included and calculated by the atomic contact potential (ACE).

$$\Delta G_{ACE} = \Delta G_{des} - T\,\Delta S^{receptor}_{surface} - T\,\Delta S^{ligand}_{surface}, \qquad \textbf{Equation IV-8}$$

$$\Delta G_{ACE} = \Delta G_{des} - T\,\Delta S^{pro}_{surface} - T\,\Delta S^{pep}_{surface}, \qquad \textbf{Equation IV-9}$$

For association of disordered peptides against folded receptor protein, the peptide will experience a transition from totally disordered before binding to folded in a unique conformation after binding. Therefore, two additional terms $-T\,\Delta S^{pep}_{bb}$ and $-T\,\Delta S^{pep}_{buried}$, together with $-T\,\Delta S^{pep}_{surface}$, are needed to account for the degree of freedom lost upon binding. The term $-T\,\Delta S^{pep}_{bb}$ represents backbone flexibility loss, and $-T\,\Delta S^{pep}_{buried}$ represents the entropy change upon peptide side chain changes from buried state to exposed state. Since $-T\,\Delta S^{pep}_{surface}$ has already been included in the ΔG_{ACE}, the other two terms are added to our scoring function:

$$\Delta G = \Delta E_{elec} + \Delta G_{ACE} - T\,\Delta S_{trv} - T(\Delta S^{pep}_{bb} + \Delta S^{pep}_{buried}), \qquad \textbf{Equation IV-10}$$

The magnitude of the conformational entropy change experienced by the peptide backbone (ΔS^{pep}_{bb}) and side chain (ΔS^{pep}_{buried} and $\Delta S^{pep}_{surface}$) upon protein folding was investigated experimentally and by computational analysis (Table IV-2) [99,100]. In [100], Lee et al., calculated the energy profiles of amino acid side chains as a function of the dihedral angles. With these energy profiles, they directly estimated the probability distribution of different conformers and therefore the conformational entropy of side chains of amino acids. D'Aquino et al. extended

this method in measuring backbone entropy of amino acids [99]. They first used experimental microcalorimetric analysis to measure the backbone entropy change between ALA and GLY. Then they followed the energy profile method to estimate the backbone entropy from the calculation of the Boltzmann weighted probability of the different conformers available to the amino acids. Computational estimates had been validated with experimental results.

Then, how do we model the entropy change of the peptide during association? First, we assume that, before binding, the peptide is completely disordered or flexible and has positive entropy. After binding, part of peptide is folded into a unique conformation and the other part remains disordered. During this process, the entropy decreases and there must be some favorable interaction energies (either electrostatic or desolvatioin or both) to compensate for the entropy penalty. The penalty is estimated as the sum of conformation entropy change of residues which lost flexibility:

$$\Delta S^{pep} = \sum_i \Delta S^i = \sum_i (\Delta S^i_{bb} + \Delta S^i_{buried} + \Delta S^i_{surface}), \quad \textbf{Equation IV-11}$$

where ΔS^i is the entropy change of peptide residue i and $\Delta S^i_{bb}, \Delta S^i_{buried}, \Delta S^i_{surface}$ are components of peptide residue i. Then the free energy scoring function is in the form:

$$\Delta G = \Delta E_{elec} + \Delta G_{ACE} - T\Delta S_{trv} - T\sum_i (\Delta S^i_{bb} + \Delta S^i_{buried}) \quad \textbf{Equation IV-12}$$

For example, suppose a 10-residue long peptide binds to a PDZ domain protein (receptor). Before binding, the peptide is highly disordered in the solution and after binding, the first four residues binds tightly to the protein, while the other six residues remain flexible in the solution,

as shown in Figure IV-13**.** The conformation entropy change of the peptide is the sum of the conformation entropy change for the first four residues.

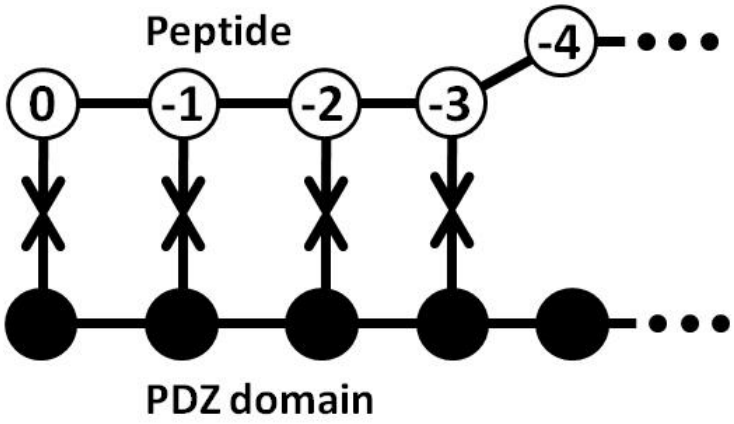


Figure IV-13: Example of entropy calculation of 10-residue disordered peptide binding to PDZ protein domain. Binding interaction starts from peptide residue at position *0*, the next three residues ("−1" to "−3") have bound to PDZ domain sequentially to the PDZ and the following six residues ("−4" to "−9") are still free. The entropy loss of during the binding interaction is the sum of translation/rotation/vibration entropy change (−15 kcal/mol) and residual configurational entropy change (from "0" to "−3"). We assume that trans/rot/vib entropy change is applied when peptide residue "0*"* binds to the PDZ protein domain.

Table IV-2: Conformational entropies change of amino acids [99,100]

Amino Acid	$\Delta S_{bu->ex}$(cal/K·mol)	$\Delta S_{ex->u}$(cal/K·mol)	ΔS_{bb}(cal/K·mol)
ALA	0.00	0.00	4.10
ARG	7.11	-0.84	3.40
ASN	3.29	2.24	3.40
ASP	2.00	2.16	3.40
CYS	3.55	0.61	3.40
GLN	5.02	2.12	3.40
GLU	3.53	2.27	3.40
GLY	0.00	0.00	6.50
HIS	3.44	0.79	3.40
ILE	1.74	0.67	2.18

LEU	1.63	0.25	3.40
LYS	5.86	1.02	3.40
MET	4.55	0.58	3.40
PHE	1.40	2.89	3.40
PRO	0	0	0
SER	3.68	0.55	3.40
THR	3.31	0.48	3.40
TRP	2.74	1.15	3.40
TYR	2.28	3.12	3.40
VAL	0.12	1.29	2.18

3. Binding free energy function

We use free energy scoring function to estimate the residual binding affinity contribution. A fully consistent binding free energy for docking disordered peptides can be written as

$$\Delta G = \sum_{i=0}^{N} \Delta G_i - T\,\Delta S_{trv}$$

$$= \sum_{i}^{N} \left(\Delta E^i_{elec} + \Delta G^i_{ACE} - T\,\Delta S^i_{bb} - T\,\Delta S^i_{buried}\right) - T\,\Delta S_{trv} \quad \textbf{Equation IV-13}$$

If PDZ structure is fixed, items in Equation IV-13 are addable, and one can easily evaluate the binding free energy per residue ΔG_i, with the caveat that $-T\,\Delta S_{trv}$ is applied only once upon forming the encounter complex. We note that internal and vdW energies are assumed to cancel

between unbound and bound state, i.e., complexes do not build strain, and solute–solute and solute–solvent vdW compensate.

D. SCREEN PDZ–PEPTIDE INTERACTIONS BY *PEPDOCK*

To test the performance of free energy scoring function and capability of discriminating strong binders from non-binders, we computed the binding affinity of 126 human peptides and 95 artificial peptides against 2 PDZ domains sequences and compared our discrimination results to the published experimental assays data [88]. By using a typical PDZ–peptide binding affinity, 10^{-5} M (−6.62 kcal/mol) as threshold, we obtained robust prediction rates. In addition, we validated *PepDock* by predicting the PDZ–peptide interactions and complex structures on five peptides bound to four different PDZ domains PSD95-3 [9], GRIP1-6 [83], ZO1-1 [82], and TIP-1 [84], where our top ranked models accurately predict crystal structures.

1. Analysis of PDZ domains in the Protein Data Bank (PDB)

There are close to 180 human PDZ proteins, including unbound structures in the PDB database. Can *PepDock* screen all of them? For PDZ screening, we generated a peptide backbone library from MD simulation of peptide alone, constraining the 5-residue C-terminal motif similar to the crystal peptide backbone, i.e., in the simulation, the backbone (position "0*"* to "−4") of peptide is mimicking the CRIPT from PDB 1BE9. When screening a target PDZ, this library will work adequately if target PDZ is similar to PSD95-3 and CRIPT. It will not fit for target PDZ, such as

DVL-2 (PDB 3CBX), whose PDZ and peptide are much different from PSD95-3 and CRIPT. To find out which groups of PDZs can share the peptide backbone library, we did an exhaustive analysis of the bound and unbound PDZ domains in the PDB database and the results showed a high degree of structural similarity among PDZs and bound peptides (Figure IV-14, Table IV-3). Indeed, clustering [38] the full set of the bound peptides with at least 5-residue long (37 X-ray and 14 NMR) shows that the largest cluster is the group of CRIPT peptides, which is bound to PSD95-3, including GRIP1-6, ZO1-1, and 9 more PDZ domains. This observation suggests PSD95-3 complex structure is the best candidate to model peptides bound to structurally similar PDZs and the library can be shared to screen the PDZs of same cluster, such as SAP97-3 in 2I0I and ZO1-1 in 2H2B. For PDZs, such as DVL-2 in 3CBX and MAGI-1 in 1V1T, from other groups, *PepDock* needs a new backbone library. In addition, for PDZs that have only unbound structure in PDBs, they could be modeled by using another PDZ complex as template.

Table IV-3: Cluster of PDZ domains from Protein Database.

Cluster centers	Bound complex structures	Unbound PDZ structures
1BE9	1BE9, 1TP5, 1N7F, 1Q3P, 2H2C, 2I0I, 2H2B, 1TP3, 2EJY, 2I0L, 2OQS, 2PNT	2VRF, 1QAV, 1BFE, 1TQ3, 2FNE, 2I1N, 1PDR, 2AWU, 2FE5, 2FCF, 2JIL, 2JIK, 2JIN, 2QG1, 2AWX
1I92	1I92, 1GQ5, 1GQ4, 1KW4	
1V1T	1V1T, 2EGN, 1W9O, 1W9Q	
2I04	2I04, 2OPG, 2EGN	
1YBO	2I04, 2OPG, 2EGN	

1RZX	2I04, 2OPG, 2EGN	

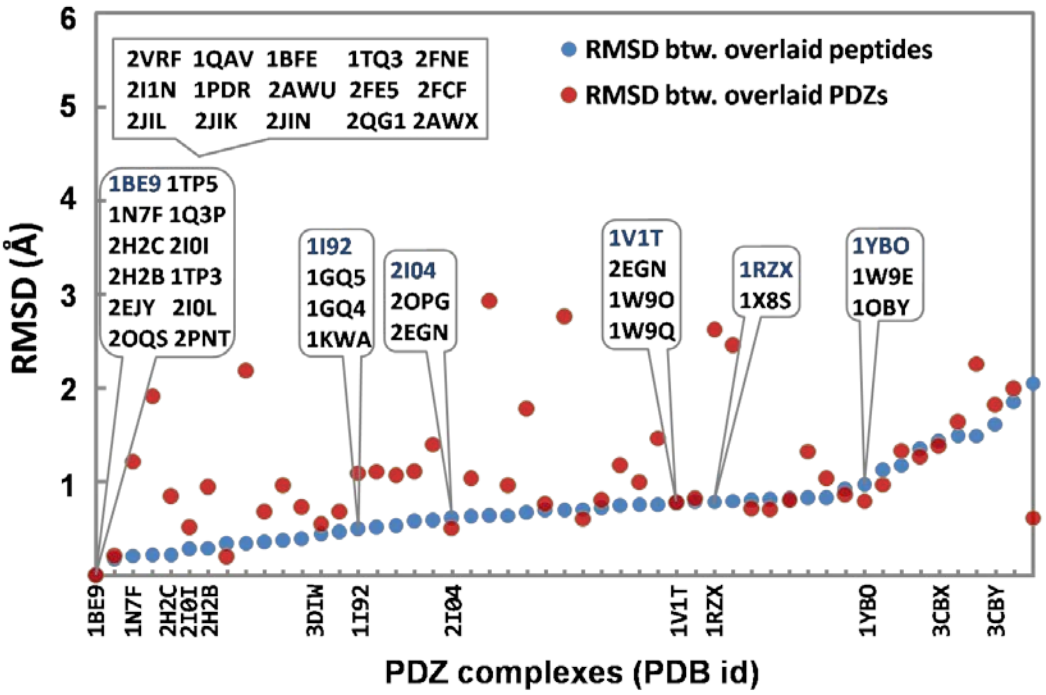


Figure IV-14: Structural analysis of PDZ domain in PDB. 51 structures of PDZ complex containing at least 5-residue long peptides are found in the PDB database. The structural similarity between bound peptides from complexes and CRIPT-peptide from PSD95-3 complex (PDB 1BE9), and similarity between PDZs to PSD95-3 PDZ domain are shown in the figure: (a) RMSDs between CRIPT and bound peptides after overlaid by fitting first 5 C-terminal residues to the CRIPT (blue symbols), (b) RMSDs between CRIPT and bound peptides after overlaid by fitting core motifs of PDZs to PSD95-3 (red symbols), (c) RMSDs between PSD95-3 and PDZs after overlaid by fitting their core motifs to PSD95-3 (green symbols). The correlation coefficient between a and b/c are 0.88/0.30. 28 of 51 PDZ complexes are clustered into six groups by using pairwise RMSD between overlaid peptides with 0.4 Å radius. Clusters are shown in round boxes and the center complex is shown as the first one (blue font). 15 unbound PDZ structures, which have RMSD to PSD95-3 less than 0.65 Å, are shown in the square box. Our work concludes that PDZs with unbound structures are predictable when their RMSD to PSD95-3 are less than 1.27 Å.

2. Screening of human peptides interacting with PDZ domains

We tested the capabilities of the method to discriminate binding from non-binding peptides by screening the binding affinities of two independent experimental assays of 126 natural and 95 artificial peptides interacting with PSD95-3 and SAP97-PDZ3 domains, individually [88].

Based on the recognition motif of CRIPT bound to PSD95-3, we followed the procedure described in Methodology to minimize the binding free energy for the full set of 126 human peptides in Ref. [88]. Figure IV-15 shows the cumulative binding free energy as a function of the number of bound residues from the C-terminal "0" to residue "−9" for the top 11 and bottom 20 (experimentally ranked) interacting peptides. Note that, in principle, the binding order of each residue could be arbitrary (see next section for detail). However, Figure IV-15 shows that the sequential binding of each residue permits a downhill binding pathway. The binding free energy landscapes revealed several insights into the binding mechanism of disordered peptides to PDZs:

a. The contribution to the binding free energy of the C-terminal residue ($\Delta G_0 \sim -9.5$ kcal/mol) was stronger than any other residue. It results 6 kcal/mol under the assumption that first binding residue needs to compensate the −15 kcal/mol association entropy change. Below, we argue that only by anchoring the C-terminal first, a disordered peptide can partially compensate for the estimated 15 kcal/mol entropy loss upon association;

b. Binding of the recognition motif between "0" and "−3" is non-specific, whereas residues between "−4" to "−8" determine peptide specificity. There is no obvious difference between free energy pathways of strong binder and weak binder peptides. Beginning from "−4" position, strong binders lower the free energy while weak binders remain flat.

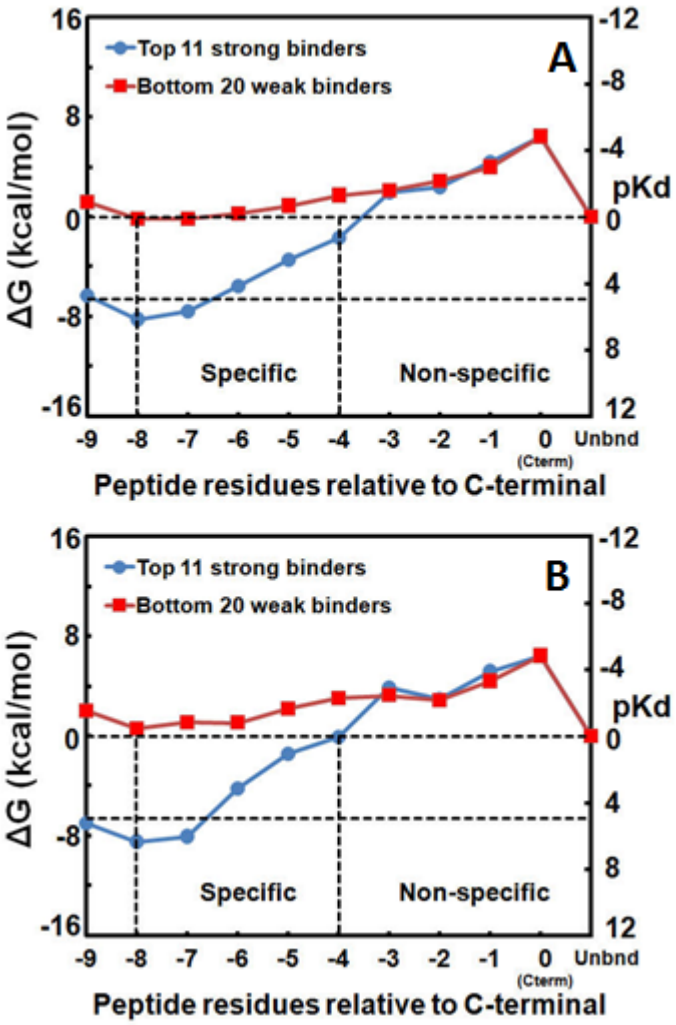


Figure IV-15: Specific and non-specific binding landscapes of PDZ–peptide interactions. Cumulative average of the lowest predicted binding free energies for the strongest and weakest binding peptides to PSD95-3(A) and SAP97-3 among the full set of 126 natural peptides experimentally ranked in [88]. Plots are shown as a function of the number of bound residues starting from the C-terminal Val_0 that contributes about −10 kcal/mol to the binding free energy, compensating for most of the 15 kcal/mol entropy loss upon association. The average binding free energies of strongest and weakest complexes are about −8 kcal/mol and 0 kcal/mol respectively. The landscapes demonstrate that peptides bind non-specifically between residues "0" to "−3", while specificity was determined by the remaining residues at the amino end of the peptides. The fact that the lowest free energies are achieved following a downhill binding pathway strongly suggests an induced folding zipping mechanism.

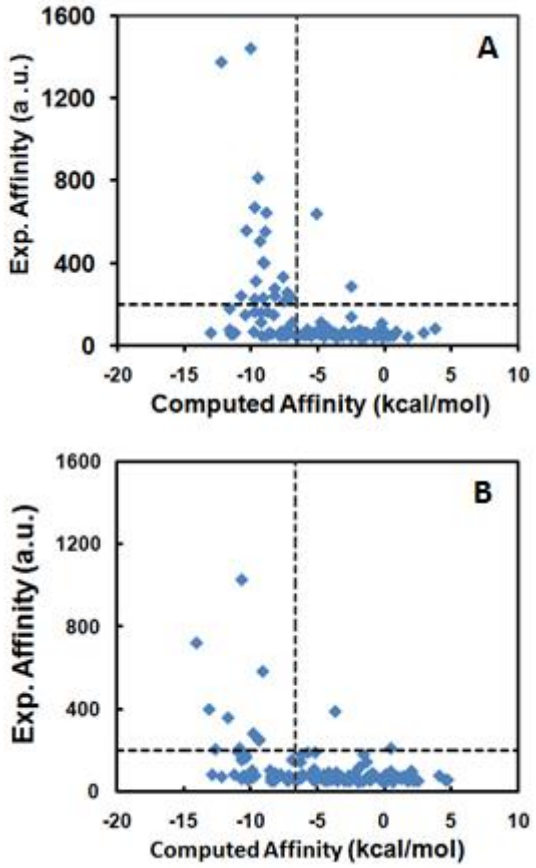


Figure IV-16: Scatter plot of 126 human peptides binding to PSD95-3 (A) and SAP97-PDZ3 (B) domains. *Y* axis is relative experimental binding affinity with arbitrary unit and *X* axis is computed binding free energy estimation. Both vertical line (−6.62 kcal/mol) and horizontal line (200 a.u.) correspond to 15 μM binding affinity.

c. Different peptides minimize the binding free energy using 7, 8, or 9 residues. None of the sequences studied here were found to reach a lower minimum using 10 residues.

d. A key group that yields the most dramatic difference between binding and non-binding peptides is the side chain at position "−4". For this position, 9 of the top 10 peptides have Lys or Arg (one has Thr) forming a salt bridge with Glu331 of PSD95-3, while the bottom 20 peptides have mostly Asp or Glu acids. Residues from "−5" to "−8" are highly variable and in average contribute about −2 kcal/mol (per residue) to the binding energy.

e. The landscapes suggest that the PDZ–peptide binding follows a downhill pathway mechanism in which peptides with high specificity undergo induced folding by sequentially "zipping" each residue into the binding pocket of PDZ domain while minimizing the binding free energy after anchoring the C-terminal residue first.

Figure IV-16 shows the correlation of computed binding free energy to relative experimental affinity for all 126 human peptides screened in Ref. [88]. Consistent with PDZ affinity data [90] of around 10^{-6} M or better, we defined a thermodynamic threshold K_d^T of 10^{-5}M (i.e., $\Delta G = -6.8\,\mathrm{kcal/mol}$, equivalent to a relative experimental affinity of 200 in Ref. [88]) to distinguish between strong and weak binding peptides, obtaining sensitivityspecificity rates of 91–74%. This threshold corresponds to the middle point between strong (~10 nM) and weak (~ 10 mM) binding pathways in Figure IV-15. Independently, we noted that $K_d^T = 10^{-5}$ M is also the lower specificity threshold for μM concentration of protein (see Fig. 3 in Ref. [18] for an exact relation), achieving the largest possible (>10 fold) differential in complex formation relative to strong binding peptides. Finally, the same affinity gap was observed between the lowest and second lowest predicted complex of strong binding peptides (between 2 and 3 kcal/mol), suggesting that specific peptides lead to well defined binding modes.

Based on K_d^T, *PepDock* correctly predicts 20 experimentally ranked strong binders as true positive (*TP*) sequences; 2 false negatives were detected (*FN* = 2) (Figure IV-16). It is important to emphasize that error bars in the scoring function can be as much as 2 kcal/mol, while the assumption that peptides are fully disordered might also further modulate our free energy estimates. Nonetheless, the sharp contrast between *TP*s in Figure IV-23B and true negatives (*TN* = 77) in Figure IV-23C is enough to clearly distinguish between strong and weak binding peptides. Strikingly, the profile of several landscapes among the thermodynamic false positives (*FP* = 27) is quite different from *TP*s (Figure IV-23C). In the next section, we will explore the kinetic implications of these landscapes.

Table IV-4: Results of screening strong/weak peptides by *PepDock*

PDZ Domain	Template Complex	Template PDZ	PDZ Structure	Sensitivity	Specificity	Correlation
PSD95-3	1BE9(B)	PSD95-3	1BE9	91% (20/22)	74% (77/104)	N/A
SAP97-3	2I0I(B)	SAP97-3	2I0I	83% (10/12)	75% (85/114)	N/A
Syntrophin	2PDZ(B)	Syntrophin	2PDZ	93% (13/14)	67% (2/3)	0.65

Is the −6.82 kcal/mol free energy threshold arbitrary, or can the free energy scoring function of *PepDock* provide a good estimation of binding affinity? To answer this question, we plotted the sensitivity and specificity change with free energy threshold range from −15 kcal/mol to 5 kcal/mol (Figure IV-17). Sensitivity curve increased smoothly with change of threshold and had a value above 80% when it reached −7 kcal/mol. The sum of specificity and sensitivity reached its maximum around the free energy −6.82 kcal/mol. This observation confirms that *PepDock* obtained its best performance of discrimination when using a thermodynamically

meaningful threshold. Together with the result of free energy landscape pathway, we can say that *PepDock* can provide a reliable estimation of the interaction binding affinity.

We repeated the discrimination test on the same set of human peptides against SAP97-PDZ3 domain and observed the similar results as PSD95-3. The free energy landscape shows that strong binders followed a downhill pathway, started discriminating from weak binder from position "−4" and reach the global minimum, −8 kcal/mol at position "−8". Weak binders kept flat and never fell below −2 kcal/mol. Based on the same thermodynamic threshold −6.82 kcal/mol, *PepDock* discriminates 10 strong binders out of 12 and 85 non-binders out of 114, with 83% sensitivity and 75% specificity, respectively. Comparing PSD95-3 and SAP97-PDZ3, we observed SAP97-PDZ3 was more selective than PSD95-3 and only have 12 peptides with relative affinity above 200. In addition, by plotting the sensitivity and specificity versus thermodynamic free energy threshold, we found that *PepDock* reached its maximum discrimination around −6.8 kcal/mol (10^{-5} M), which is consistent with experimental evidence [90].

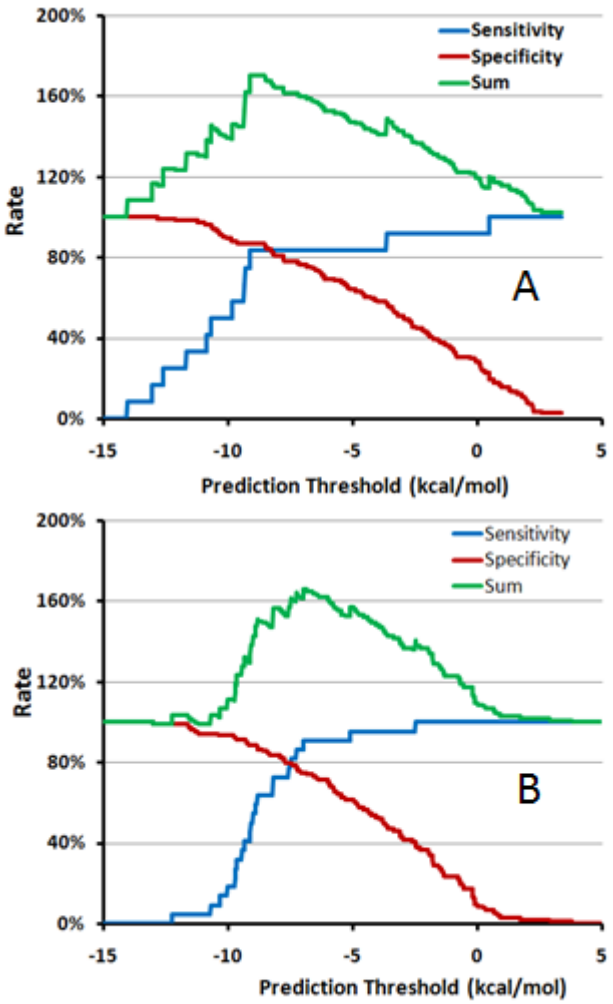


Figure IV-17: Sensitivity curve of screening strong/non-binding peptides by *PepDock*. The experimental data array of 128 native human peptides binding against PSD95-3 (A) and SAP97-PDZ3 domain (B) are re-computed and screened by *PepDock*. The sensitivity and specificity curve shows that *PepDock* has strong ability to discriminate strong and non-binders. Please note: the total performance (sum of specificity and sensitivity) reached its maximum around −6.70 kcal/mo*l*, which is consistent with our physical threshold −6.62 kcal/mol. This observation strongly supports the robustness and accuracy of *PepDock* free energy scoring function.

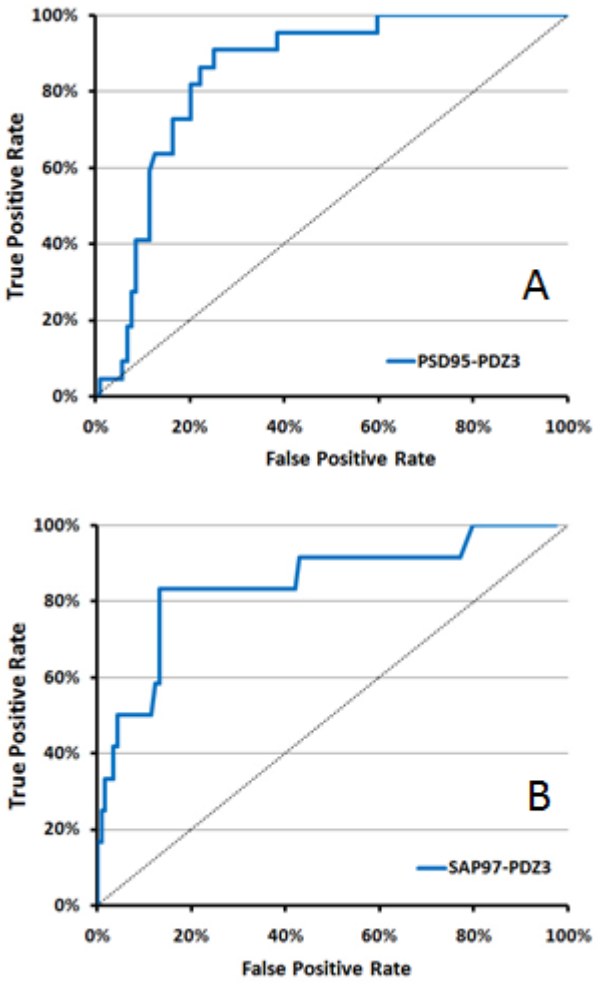


Figure IV-18: ROC curve of screening strong/non-binding peptides by *PepDock*. The experimental data array of 128 native human peptides binding against PSD95-3 (**A**) and SAP97-PDZ3 domain (**B**) are re-computed and screened by *PepDock*. As shown in the plots, *PepDock* showed strong discrimination ability with area under the curve, 85% for PSD95-3 and 82% for SAP97-PDZ3, respectively.

3. Screening of artificial peptides interacting with PDZ

The robustness of *PepDock* was further confirmed by screening a whole new database of 95 artificial peptides that were experimentally validated using Phage ELISA [88]. Contrary to natural peptides, this dataset was phage selected to include mostly true positives. Since there is no quantitative mapping between the ELISA readings and binding affinities for this assay, we assumed a thermodynamic threshold equivalent to a tenth of the experimental scale. This dataset again shows that the minimum binding free energies are consistent with downhill zipping pathways, strongly suggesting that this is a general mechanism for binding disordered peptides. The sensitivity–specificity rates using the same thresholds as for natural peptides are: for the thermodynamic threshold,$K_d^T = 10^{-5}$μM, 80–64%; and, for both the thermodynamic and kinetic threshold $K_d^T = 1$M combined, 68–91%. The consistency of the performance obtained for both natural and artificial peptides, i.e., average rates of 80–80% or a combined 160%, provides a strong support for *PepDock* as a tool to design artificial peptides to bind specific PDZ domains.

4. Predicting the complex structures

We probed the robustness of our method by predicting the complex structure and absolute affinities between 7 peptides and 5 different PDZ domains (Figure IV-19 and Table IV-5). Among these complexes, the backbone of four of these peptides, bound to PDS95-3, GRIP1-6 and two for ZO1-1, are within a 0.4 Å RMSD of the overlapped CRIPT peptide; one peptide bound to TIP-1 is 0.44 Å RMSD away from CRIPT; and, two peptides bound to DVL-2 are very different from the CRIPT template structure (> 1.6 Å RMSD). We docked the backbone library onto the target PDZ by using both the complex peptide as bound reference and the CRIPT

peptide, after the target PDZ is overlapped into the PSD95-3 co-crystal, as unbound reference (see section IV.B.1). It is important to emphasize that even though we used a bound complex peptide as reference to pre-dock the backbones, the library was actually generated independently and no crystal backbones are included. Figure IV-19A shows that peptides docked to PDZ domains similar to CRIPT/PSD95-3 not only formed strong complexes, but also had landscapes consistent with that of TP sequences in Figure IV-23.

The PDB of CRIPT and PSD95-3 shows five C-terminal residues, but the amino end and several side chains are not resolved in the crystal structure. Figure IV-19B shows the predicted structure of the full CRIPT complex overlapped with the crystal, including the electron density map (EDM) within 1 Å of the model. The predicted structure recovers bound motifs, including some significant overlap with EDM of missing groups. However, contrary to other peptides, contacts made by $N_{-6}Y_{-5}$ add almost no binding free energy in a region of the landscape where the affinity is still above the binding threshold (light blue path in Figure IV-19A). Since these residues have no preference to form the on-pathway contacts relative to, say, no contact at all (unbound), the rate of folding into the right backbone configuration is necessarily slower than downhill contacts. Hence, we speculate that this feature on the CRIPT binding landscape might have contributed towards the poor resolution of the remaining residues in the crystal, despite the fact that overall CRIPT has been shown to be a good binder [9,88,89].

The predicted complexes of GRIP1-6 [9] for both bound and unbound reference peptides capture the main features of the complex with the exception of Tyr-3 (Figure IV-19C), which finds a hydrophobic pocket that also buries an unmatched hydrogen bond. Energy-wise, the difference between the two rotamers is minimal. The problem lay in the subtle balance between hydrophobic and polar contacts of the extra OH group. For ZO1-1 [7], we docked two peptides

using both bound and unbound PDZ and peptides. As shown in Figure IV-19D/E, all four models recovered the hydrogen bonds and strong crystal contacts with a backbone RMSD of 1.53 Å or less. Interestingly, the docked structures correctly modeled the aromatic side chains of Trp_{-1}, but again, the energetic balance of Tyr_{-1} is shifted between two rotamers. TIP-1 [10] is probably at the boundary of what one should model using CRIPT as template. Nevertheless, despite some visible backbone differences between bound and unbound models, the predicted contacts were still in good agreement with the crystals (Figure IV-19F). To a large extent, the tolerance to backbone misfits was due to the pairwise nature of the scoring function that de-emphasizes the precise orientation of side chains and hydrogen bonds. Note that large backbone-RMSDs differences observed at the amino end residues $A_{-8}T_{-7}$ and $Q_{-9}L_{-8}A_{-7}$ of GRIP1-6 and TIP-1, respectively, are due to the fact that these residues do not contact the PDZs, and have minimal binding energies (Figure IV-19A). Without energetic constraints, the method cannot pin down a structure.

For completeness, we also attempted to dock two artificial peptides bound to DVL-2 [104]. In this case, the peptide backbone and core PDZ domains were very different from CRIPT/PSD95-3. Not surprisingly, predicted models did not fit the crystal. This negative exercise confirms our initial assumption that target PDZs should resemble the template structure. Next, we generated a new library of peptide backbone models from MD simulation of the artificial peptide WKWYGWF and used the backbone models as the input to predict the complex structure. No doubt, the prediction shows consistent side chain contact between peptide and PDZ domain (Figure IV-20). This test supported our conclusion that, for DVL-2 or other PDZs, which are away from 1BE9, a new PDZ template and a new peptide backbone library are necessary for *PepDock* to predict and much likely to lead to a reliable prediction.

Table IV-5: Top ranked prediction model of complex structures based on bound/unbound PDZ and bound/unbound peptide

Crystal Structure	PDZ	PDZ Structure	Peptide Sequence	Template Complex	Template PDZ	ΔG (kcal/mol)	BB RMSD (Å)	Side Chain Contacts
1BE9	PSD95-3	1BE9	KQTSV	1BE9(B)	PSD95-3	–8.89	0.62	3/3
2IOI	SAP95-3	2I0I	RRETQV	2I0I(B)	SAP97-3	–9.66	1.12	2/3
1N7F	GRIP1-6	1N7F	ATVRTYSC	1N7F(B)	GRIP1-6	–9.32	3.34	3/3
2H2B	ZO1-1	2H2B	WRRTTYL	2H2B(B)	ZO1-1	–9.59	0.91	5/5
2H2C	ZO1-1	2H2C	WRRTTWV	2H2C(B)	ZO1-1	–14.96	1.25	5/5
3CBX	DVL2-1	3CBX	WKWYGWF	3CBX(B)	DVL2-1	–11.91	0.52	5/5
3DIW	TIP1	3DIW	QLAWFDTDL	3DIW(B)	TIP1	–5.64	8.99	4/4
1N7F	GRIP1-6	1N7E	ATVRTYSC	1BE9(UB)	PSD95-3	–5.42	2.45	3/3
2H2B	ZO1-1	2H2C	WRRTTYL	2H2C(UB)	ZO1-1	–11.92	1.28	5/5
2H2C	ZO1-1	2H2B	WRRTTWV	2H2B(UB)	ZO1-1	–14.00	1.11	5/5
2H2B	ZO1-1	2H3M	WRRTTYL	1BE9(UB)	PSD95-3	–7.66	1.81	4/5
2H2C	ZO1-1	2H3M	WRRTTWV	1BE9(UB)	PSD95-3	–11.04	1.90	4/5
3DIW	TIP1	3DJ1	QLAWFDTDL	1BE9(UB)	PSD95-3	–3.32	8.48	3/4

* Predictions that use unbound CRIPT as peptide template and apo-PDZ structure (if available) are highlighted by yellow color.

† Binding free energy landscapes are shown in Figure IV-19.

§ RMSDs are with respect to residues resolved in crystal.

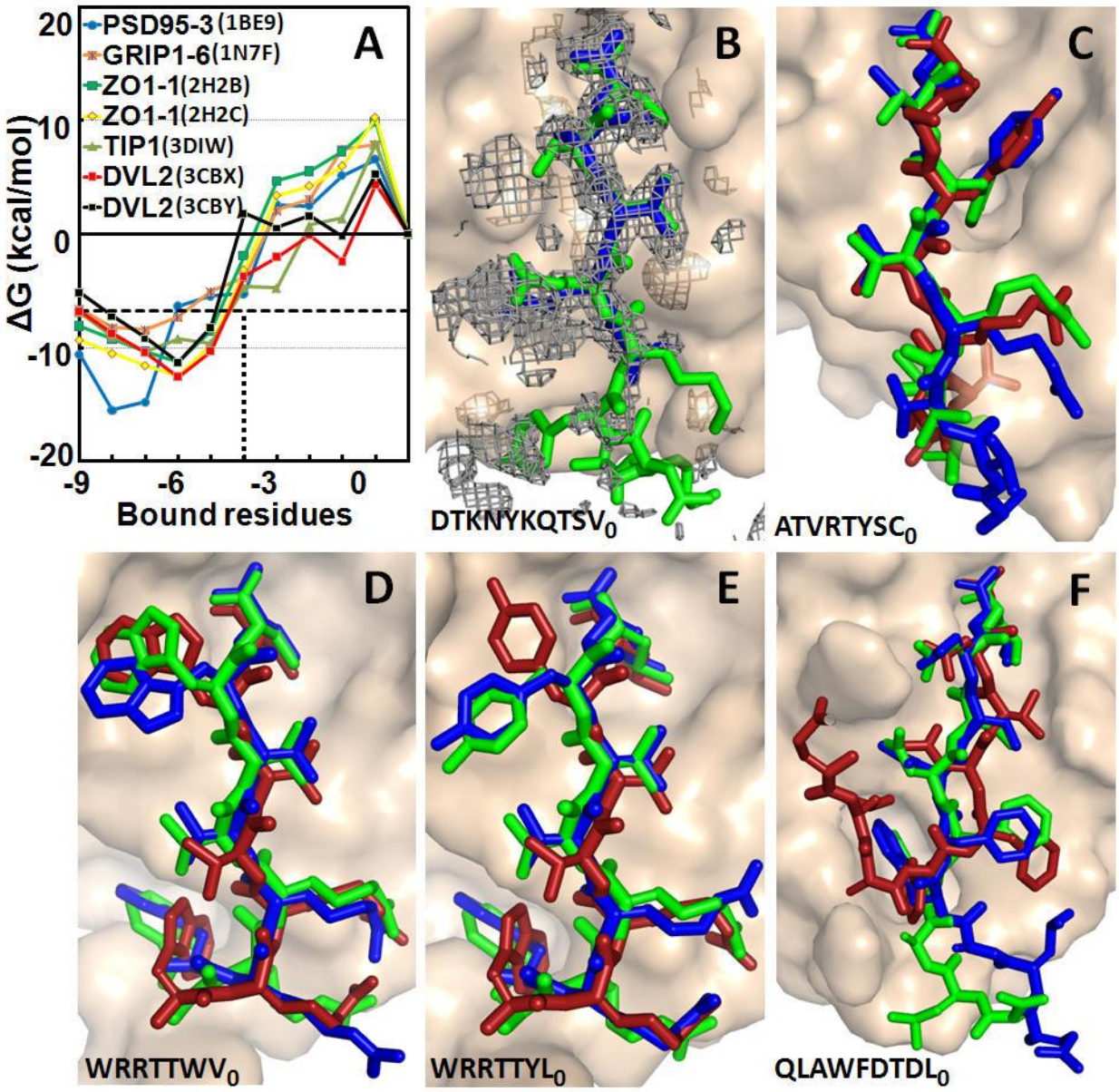


Figure IV-19: Prediction of PDZ–peptide interactions and their complex structures using PSD95-3 as template. (**A**) Binding landscapes of five known PDZ–peptides docked onto four different domains. All predicted models show the downhill landscape that characterizes strong binding peptides (see Figure IV-15). Top ranked predictions based on bound and unbound reference peptide are shown in green and red sticks, respectively; crystal structures are shown in blue. (**B**) CRIPT docked to PSD95-3, also shown is the electron density map. (**C**) ATVRTYSC docked to GRIP-6. Both bound and unbound prediction capture the main features of the complex. (**D**) WRRTTWV and (**E**) WRRTTYL peptides docked to ZO1-1. All four models recover the crystal contacts. (**F**) QLAWFDTDL docked to TIP-1. The bound and unbound structures recover the main contacts of bound motif ("0 to −5") in the crystal. Note that the models and crystals deviate at the amino end residues $A_{-8}T_{-7}$ and $Q_{-9}L_{-8}A_{-7}$ of GRIP-6 and TIP-1, respectively, which not only do not contact PDZ but also have positive binding energies.

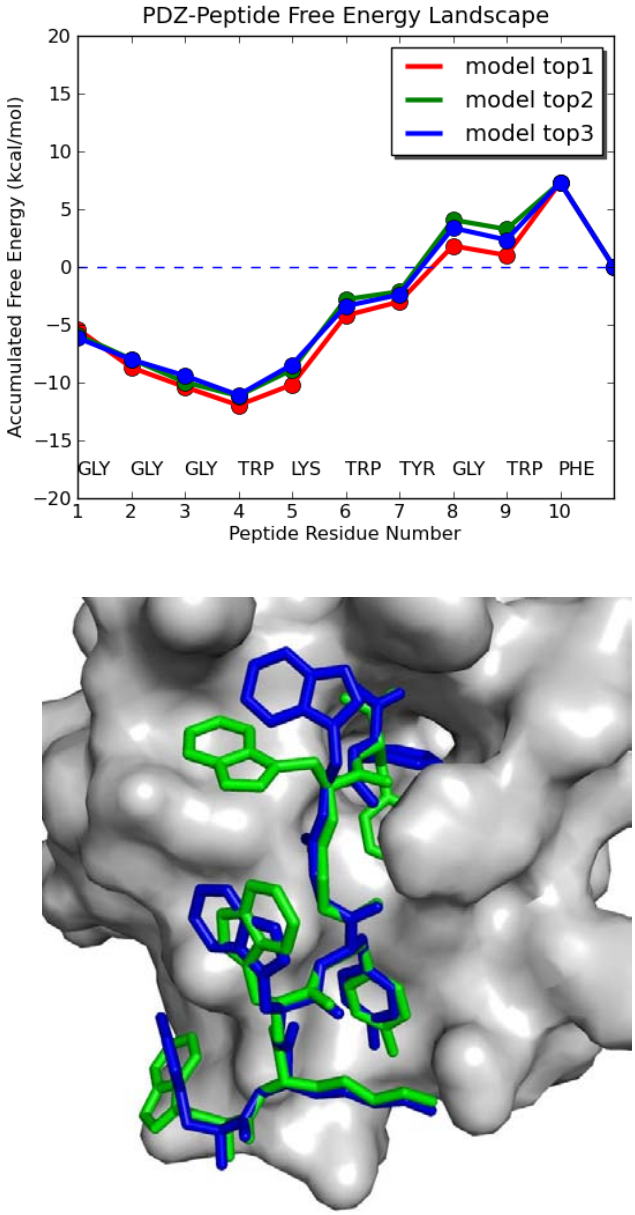


Figure IV-20: Prediction of the interaction between WKWYGWF peptide and DVL2-PDZ domain. We generated a new library of peptide backbone models since the structure WKWYGWF peptide and DVL2-PDZ domains appear structural different from CRIPT peptide and PSD95-3 PDZ domain. We use new backbone model as the input to predict the complex structure and binding affinity. The top three prediction models show downhill free energy pathway with the lowest free energy lower than −10 kcal/mol (in top figure). Predicted complex structure (in blue color) recovered the strong side chain contacts between peptide and PDZ domain with the backbone RMSD 0.52 Å. Structure of peptide from crystallography is shown in green color. This result supported our conclusion that, for DVL-2 or other PDZs, which are away from 1BE9, a new peptide backbone library and a new PDZ template are necessary for *PepDock* to predict and lead to a reliable prediction.

E. MORE DISCUSSION ABOUT DOCKING AND BINDING MODELS

1. Novel approach to dock disordered peptides

Consistent with the notion that binding is mostly determined by non-covalent interactions, our main assumption is that bound peptides do not build strain upon binding. Hence, we developed a backbone library extracted from equilibrium MD simulations in explicit solvent (see section IV.B.1 above). Then, by simply eliminating docked conformations that build strain or clashes above some feasible thresholds, an idea reminiscent of the constrained vdW minimization used in protein–protein docking [59], we circumvented the challenging problem of optimizing the backbone and vdW energies. The binding affinity is estimated based on a free energy scoring function that incorporates entropy loss upon association and folding entropy loss per residue [99,100]. Collectively, these terms yield a meaningful thermodynamic decomposition of the full binding free energy of fully disordered peptides.

2. Docking disordered peptides into PDZ domains

Our approach is sufficiently general to screen any peptide sequence, and therefore, to discover novel binding patterns. Based solely on the complex structure of one 5-residue long peptide to PSD95-3, and a thermodynamic threshold of $K_d^T = 10^{-5}$, the method successfully discriminates strong from poor binders of PSD95-3 with sensitivity and specificity rates of 91% and 74%, respectively. This threshold is consistent with experiments showing a 10^{-6}–10^{-7} M affinity for cognate peptides [90]. The robustness of the method is also reflected in the accurate all atom

docked conformations predicted for several PDZ targets (Figure IV-19), providing strong support for in silico-screening of protein-PDZ interactions.

3. On the fast association of PDZ-peptide interaction

Kinetics is also important in signaling. Indeed, a seminal study by Kiel and Serrano [105] demonstrated that k_{on} of Ras–Raf interactions play an important role in MAPK signal transduction, independently of K_d. Figure IV-21 sketches two binding pathways resulting in two different kinetic mechanisms: In (A), peptides fold before forming the high affinity complex, and, in (B), they undergo induced folding [17,106]. A third possibility is to consider that peptides actually fold, i.e., they are not disordered. The latter, however, would lead to either specific interactions that are not consistent with PDZ-peptide promiscuity, or misfolded peptides that would slow down binding by requiring extra free energy to first unfold in order to refold upon binding.

The efficiency of PDZ–peptide interactions is reflected in association rates on the order of 10 s^{-1}, i.e., comparable to interactions between folded/ordered proteins, and off rates on the order of 10 s^{-1} [90]. These rates ruled out mechanism A, which entails a slow rate of association due to the high (entropic/folding) transition state barrier, and corresponding slow dissociation rate. Indeed, from the point of view of an efficient signal, a slow on rate is highly inefficient since binding would require multiple attempts, while a slow off rate not only would slow down the resetting of the signal, but also would hinder it if multiple PDZs were targeting the peptide. As shown in Figure IV-15, the largest contribution to the binding free energy of the C-terminal strongly suggests that this residue docks first, such that it lowers the transition state the most (as in mechanism B). These anchors are minimally hindered by the rest of the peptide, explaining

their optimal solvent/receptor accessibility to form the encounter complex. This, of course, is fully consistent with the highly conserved structures and sequences of the C-terminal recognition motif, and the fact that all the other contacts are rather superficial. Hence, we suggest that the fast association rates of disordered peptides to PDZs are triggered by the well-defined anchoring, or burying, of the energetically critical hydrophobic C-terminal, a mechanism that has also been shown to describe the initial recognition step of stable proteins [30].

A downhill-induced folding mechanism suggests non-specific screening of PDZ-peptide interactions. After anchoring the C-terminal, peptides can bind/fold by following either non-sequential binding pathways or a sequential "zipping" pathway (Figure IV-22). The fact that all PDZ complex structures show an anti-parallel beta sheet next to the C-terminal strongly suggests that docking the C-terminal is followed by the zipping of the beta sheet. This is consistent with the lowering of the binding free energy by the consensus motif (i.e., S/T-X-Φ_0) between "0" and "−3", regardless of whether or not the peptide specifically interacts with PDZ (Figure IV-15). Hence, we conclude that PDZs screen peptides non-specifically, but quickly detach from those that do not attain sufficient affinity (~10^{-5} s, for $k_{on} \sim 10^7\,\mathrm{M}^{-1}\mathrm{s}^{-1}$).

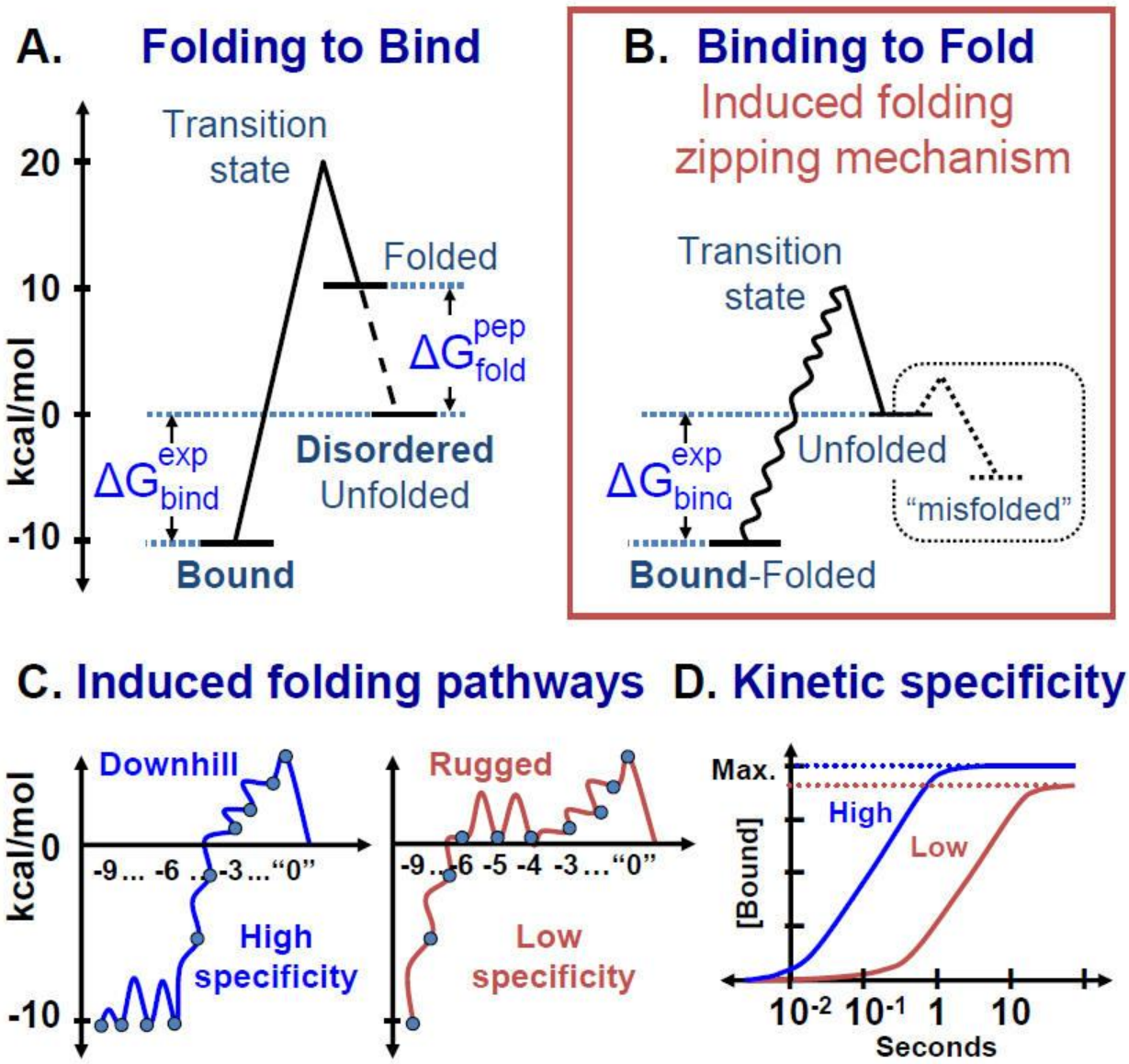


Figure IV-21: Induced folding "zipping" mechanism and kinetic specificity of promiscuous interactions. Sketches of the binding transition of a disordered peptide that (**A**) folds before binding and (**B**) folds during binding. (**C**) Sketch of high and low specificity folding landscapes mimicking those found for true and false positives in Figure IV-23. (**D**) Kinetic specificity resulting from landscapes in C: k_{on} of C-terminal is assumed to be 10^7 M−1s−1; baseline binding rate between residue "i" to "i − 1" is assumed to be 10^8 s^{-1}, folding rates are further scaled by barriers and ΔG_i that are drawn to scale in each landscape in C. Induced folding mechanism result in kinetic specificity whereby the contacts closer to the C-terminal will bind faster than energetically favorable random contacts further removed from "0". We also note that rates obtained by this model are consistent with experiments $k_{on} \sim 10^{-7}$ $M^{-1}s^{-1}$ and $k_{off} \sim 10$ s^{-1} [90].

Kinetic specificity of downhill pathways: It is also clear that non-sequential pathways would trigger higher entropic barriers from constraining multiple residues without an enthalpic compensation (Figure IV-22B), while sequential pathways entail smaller barriers, binding by one residue at a time, immediately compensating for folding entropy (Figure IV-22A). In what follows, we explore the kinetic implications of downhill pathways observed in binding landscapes of true positives relative to false positive sequences whose sequential pathways are not downhill (Figure IV-23).

A simple model mimicking these landscapes (Figure IV-23C) demonstrates that *FP*s reaching the same minimum binding free energy as *TP* peptides, but with a rugged landscape at residue "−4", bound significantly much slower than *TP* sequences that have the rugged spot after reaching the thermodynamic binding threshold K_d^T (Figure IV-23D). The extra barriers between residues "−4" and "−6" solely determine this kinetic specificity, whereas the difference in the maximum amount of bound PDZ in Figure IV-23D is due to the thermodynamic contribution associated with the three extra low free energy states in the high specificity landscape. The origin of the kinetic barriers is that a flat step ($\Delta G_i = 0$) at, say, residues $i = -5, -6$ implies that almost every other configuration of these residues is equally or more favorable than the contacts required by the pathways leading to thermodynamic stability. A rough estimate of 3 kcal/mol (a factor of 0.006), as depicted in Figure IV-23C, leads to the kinetic discrimination between *TP* and *FP* pathways in Figure IV-23D.

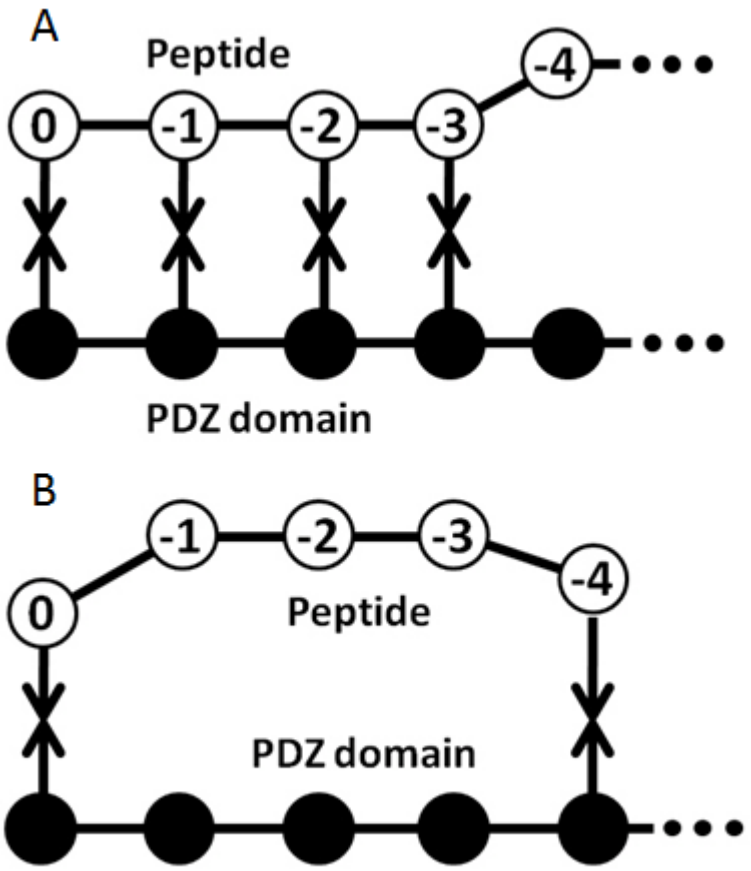


Figure IV-22: Comparison between sequential binding and non-sequential binding. All binding interactions start from peptide residue at position "0". In sequential binding scenario (**A**), residue at position "−4" binds to PDZ domain after residues ("−1" to "−3"). In non-sequential binding (**B**), residue at position "−4" binds to PDZ while peptide residues "−1" to "−3" are still partially flexible. Comparing two scenarios, it is obvious that residue "−4" need to compensate more entropy loss, which leads to higher free energy barrier in non-sequential binding than in sequential binding. So, we conclude that sequential binding is the most efficient way for disordered peptide binding to PDZ domain.

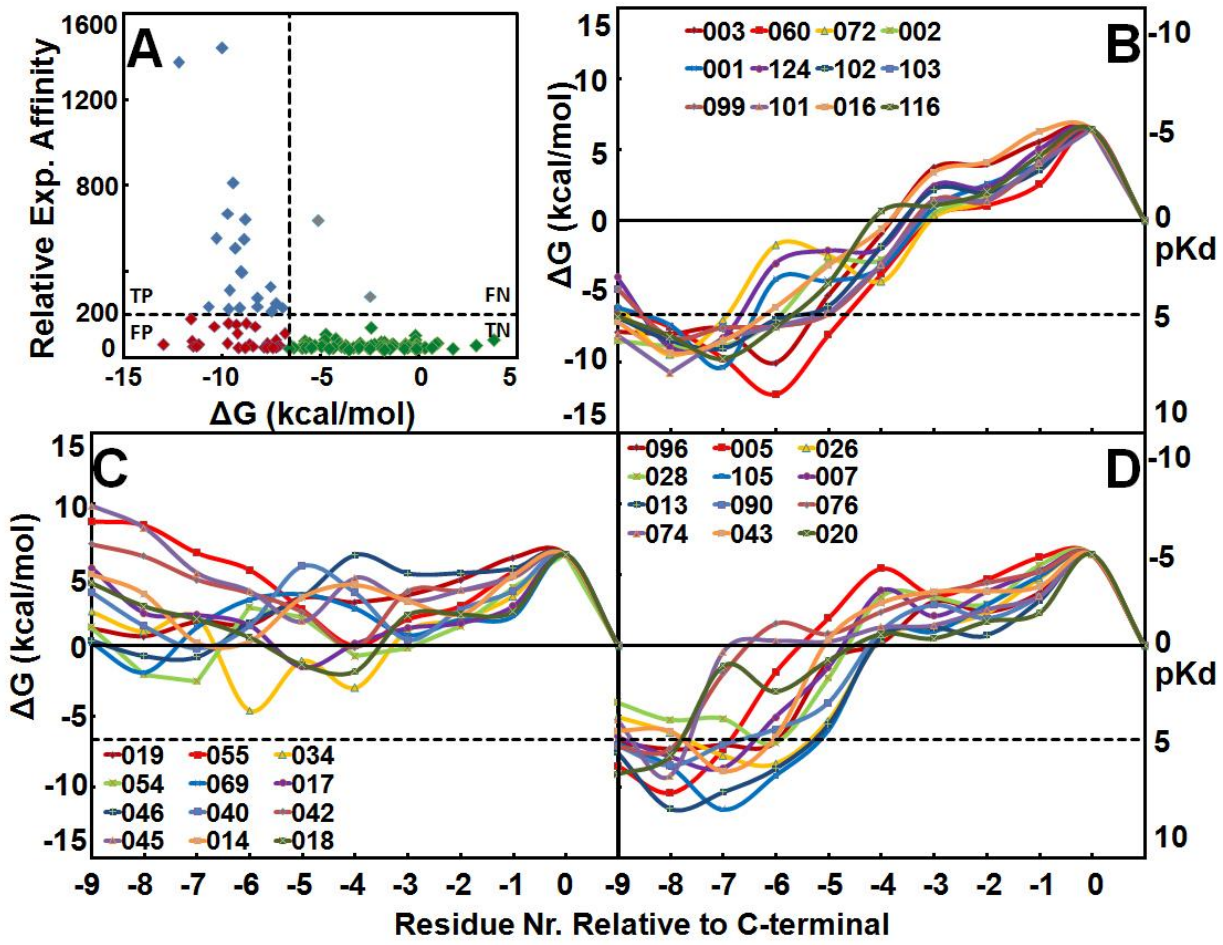


Figure IV-23: Thermodynamic specificity of 126 natural peptides binding PSD95-3. (**A**) Correlation of relative experimental affinity and binding free energy. Binding free energy landscapes for (**B**) 11 true positives (*TP*; blue symbols in A), (**C**) 20 bottom true negatives (*TN*; green symbols), and (**D**) 16 (out of 52) of the false positives (*FP*; red symbols) sequences corresponding to the weakest experimental and strongest predicted free energies. Dashed lines correspond to the thermodynamic binding affinity thresholds $K_d^T = 10^{-5}$ M, or 200 experimental affinity [88]. Sequence numbers followed [88]. All *TP* and 63 out of 115 *TN* are correctly predicted by our computational method. Differences in *TP* and *FP* landscapes suggest the binding profile might have kinetic implications not readily captured by K_d^T.

Regardless of the details of the model, it is clear that downhill pathways lead to faster binding. In particular, landscapes, such as those of *FP* sequences (Figure IV-23B), i.e., they do not lower the binding affinity soon enough after anchoring the consensus motif, lead to a slow association rate. The same kinetic discrimination occurs with non-sequential pathways, since any advantage of locking a locally favorable residue will vanish when considering the thermodynamic and kinetic cost entailed by the entropy loss of randomly constraining the residues skipped from along the way. The latter also rationalizes the limited specificity observed on PDZ binding peptides, restricted to the 7–9 residues at the C-terminal of target proteins. We quantify this effect by re-classifying those *FP* sequences that do not reach below an empirical kinetic threshold K_d^k of 1 M (or $\Delta G^k = 0\,\mathrm{kcal/mol}$; see Figure IV-15 and Figure IV-23) by residue "-4" (after the non-specific region) as "kinetic true negatives."

F. SUMMARY

We present a novel full free energy scoring function for disordered peptides, which, in combination with a semi-flexible docking method, is used to screen the binding specificity of 221 different peptides against the third domain of PSD95. This structure-based approach can be applied to PDZs with known structure, providing an efficient alternative method to detect PDZ–peptide interactions and identify novel binding sequences. The detailed sampling of all possible binding modes strongly suggest that peptides bind non-specifically by anchoring the C-terminal end in a well-defined cavity with association rates similar to folded proteins, while specificity is determined by an extended network of contacts at the amino-end terminal. These high complementarity low affinity complexes ($K_d < 10^{-5}\,\mathrm{M}$) optimize the specificity ofdisordered

binding peptides [15], while compensating for the peptide entropy loss upon folding (~1 kcal/mol per residue [99]). Consistent with Wright and collaborators' mechanism [17], specific interactions proceed by a downhill-induced folding pathway. The ruggedness of the landscapes can also lead to kinetic specificity, a mechanism that prioritizes fast association relative to dissociation [105]. In fact, the right order of association should matter for genes whose tandem PDZ domains are known to bind promiscuously to C-terminals of proteins belonging to the same regulatory pathway [89]. The large number of true positive artificial peptides relative to natural ones [88] is also consistent with Lim and collaborators' notion [107] that adapter signaling has evolved by negative selection. Collectively, these findings strongly suggest that the downhill-induced folding mechanism described here should also apply to other adapter proteins whose specificity is associated to disordered peptides with a well-defined anchoring site.

From our results, we found that the minimum free energy structures of strong binding peptides revealed a downhill binding landscape that begins by anchoring the C-terminal recognition motif non-specifically, while specificity is determined by further zipping the next 3 to 5 residues into an extended network of sequence dependent contacts. These pathways are kinetically preferred since they lead to the fast recognition of their substrates. Kinetic specificity favors favorable contacts closer to the C-terminal, while complexes that form contacts further along the polypeptide chain bind much more slowly. Quantifying kinetic specificity as a steep downhill pathway, we obtained average sensitivity–specificity rates of 91–74% for natural peptides. Our findings highlight the induced folding/binding mechanism of unstructured peptides as maximizing both the thermodynamic and kinetic specificity of promiscuous interactions, a mechanism that is likely relevant to other adapter molecules as well.

V. DISCOVERY OF NEW BIOLOGICAL INTERACTIONS BY USING *PEPDOCK*

With the success of the *PepDock* application to PDZ domains, we developed an online database and prediction web portal for users to search our pre-calculated prediction results and submit new prediction jobs if the relevant prediction cannot be found in the database. Each pre-calculated record provides users with comprehensive information about peptide, PDZ domain, and prediction confidence, which can facilitate users exploring new interactions or functionalities of PDZ domain. This work is designed and implemented by the author and directed by the dissertation advisor.

A. *PEPDOCK* WEB PORTAL

PepDockWeb is a web-based tool whose aim is to facilitate the study of the specificity of PDZ domain-disordered peptide interactions, and predict new functionality and interaction partners of the adapter protein domain. To achieve this goal, *PepDockWeb* starts with anchor residue in the known binding pocket, mimics the conservative motif, and samples the peptide conformation to search the possible partner to the adapter protein domain. For a given 10-residue long peptide sequence submitted by the user, *PepDockWeb* calculates the absolute binding free energy and analyzes the free energy change upon binding for each peptide residue. A Jmol-based [108] tool allows the user to interactively visualize peptide residues in the binding pocket contacting with

the surrounding region. *PepDockWeb* includes a PDZ–peptide interaction database of pre-calculated result of 126 human protein peptides and 85 artificial peptide sequences against 11 human PDZ domains, together with confirmed or partial experimental evidences. Users can submit a new query of an arbitrary peptide to selected PDZ domain and will receive the result within an hour. A dedicated computing cluster provides the computational power of *PepDock* and one run typically takes 30 minutes of CPU time. *PepDockWeb* provides a resource to rapidly and accurately assess of PDZ–peptide interactions for the specificity of PDZ domains and the inhibitor modeling.

PepDockWeb provides the user with three different components: *Results*, *Database* and *Prediction*, and it is available at: http://smoothdock.ccbb.pitt.edu/*PepDock*/.

1. Results

The *Results* presents the validation of *PepDock* and statistics of pre-computed results, which include specificity and sensitivity testing, complex structure predictions, and correlation testing. The *Specificity and Sensitivity* uses the experimental datasets of 126 human proteins/peptides against 6 class I type PDZ domains as the reference [88], compares calculated binding affinities with relative experimental affinities. A consistent thermodynamic threshold of $\Delta G = -6.62$ kcal/mol ($K_d < 10^{-5}$M) is used to calculate the specificity and sensitivity. The results of each test are shown in scattering-plot, ROC curve and specificity/sensitivity. Five out of six (PSD95-1, PSD95-2, PSD95-3, SAP97-2 and SAP97-3) tests show strong correlation between calculated free energy and relative experimental affinity. One test (SAP97-1) failed to show the correlation in the scattering plot. And for two tests (PSD95-2, SAP97-2), the optimal threshold to use for predicting is away from the physically relevant threshold, −6.62 kcal/mol. We concluded

the reason for these failures is that the similarity of the adapter protein structure to the template structure is low and the protein structures are not good quality.

The *Structure Prediction* compares the predicted PDZ–peptide complex structure with existing crystallographic structure from Protein Database (PDB). We used *PepDock* to predict eight PDZ–peptide interactions, which have known X-ray structures (PDBID, 1BE9, 2I0L, 2H2B, 2H2C, 1N7F, 3CBX and 3DIW). Each prediction outputs the top three prediction models with complex structure and binding free energy landscape. Five of six predictions have calculated affinities passing the thermodynamic threshold, while one (3DIW) failed. One example of WRRTTYL peptide binding to ZO1-1 PDZ domain is shown in the Figure V-1.

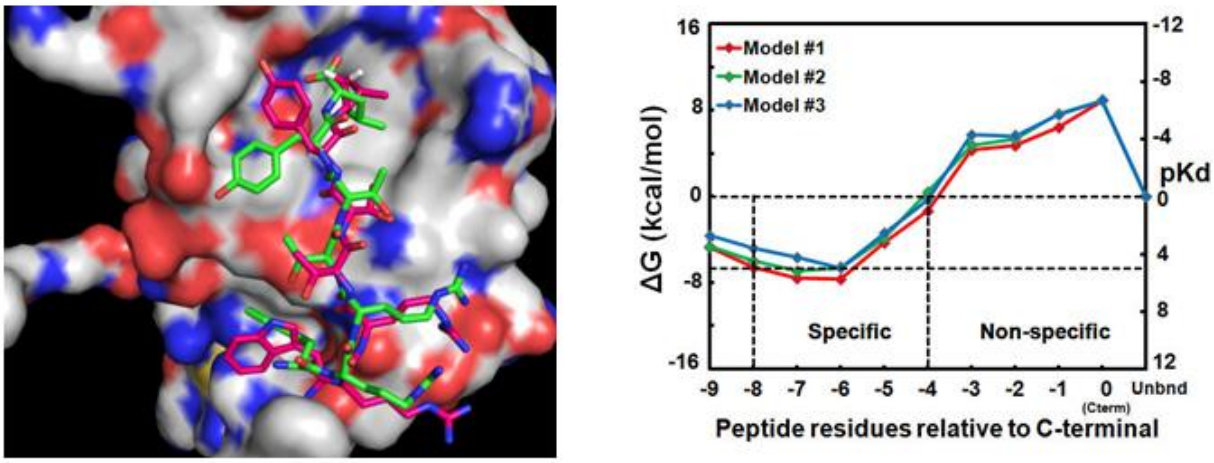


Figure V-1: Prediction results of "WRRTTYL" peptide binding to ZO1-1 PDZ domain. The overlapping of top 1 predicted complex structure model ranked by computed binding affinity and crystal structure from PDB (2H2B) shows the prediction structure captures the main contact interaction characteristics (left) with backbone RMSD 1.28 A. The free energy landscape of the top three prediction models shows the binding free energy change with the number of peptide residue bound to PDZ domain upon binding. All three models have minimum binding affinity lower than −6.62 kcal/mol threshold and confirm a downhill free energy pathway pattern.

2. Database

The *Database* contains pre-calculated results for human protein peptides and artificial peptide sequences against 11 PDZ protein domains, together with direct experimental evidence or indirectly reference literature. Peptide sequences are extracted from the published PDZ–peptide experimental dataset.

The *Database* front page (http://smoothdock.ccbb.pitt.edu/*PepDock*/DB/) lists the brief information of PDZ domains and natural/artificial peptides that are included in the databases (Figure V-2). The PDZ information contains the structure information, which is used by *PepDock*, the known binding sequence consensus, and direct link to the Swiss-Prot database. Peptides are classified into natural and artificial classes, while each human peptide has sequences, protein gene, organism, and brief function shown on the webpage. In addition, around 300 experimental instances of PDZ–peptide interactions are recorded in the database, with each record including peptide information, PDZ information, *PepDock* prediction results, and a *PubMed* reference link. Users can easily use the search toolkit on the top of the page to locate the pre-computed records in the database.

The *Database Query* page lists the queried interactions when the user submits a database query for the PDZ domain or peptide, or both. For example, a query of interactions of the PSD95-3 domain is shown as in the Figure V-3. Interactions are grouped into five categories and displayed with different colors: *Confirmed*, *Mismatched*, *High*, *Middle*, and *Low*, based on consistency between the experimental evidence and computed result. Hovering the mouse pointer over each row will illustrate the binding free energy landscape and users can go to the detailed prediction result page by clicking each row.

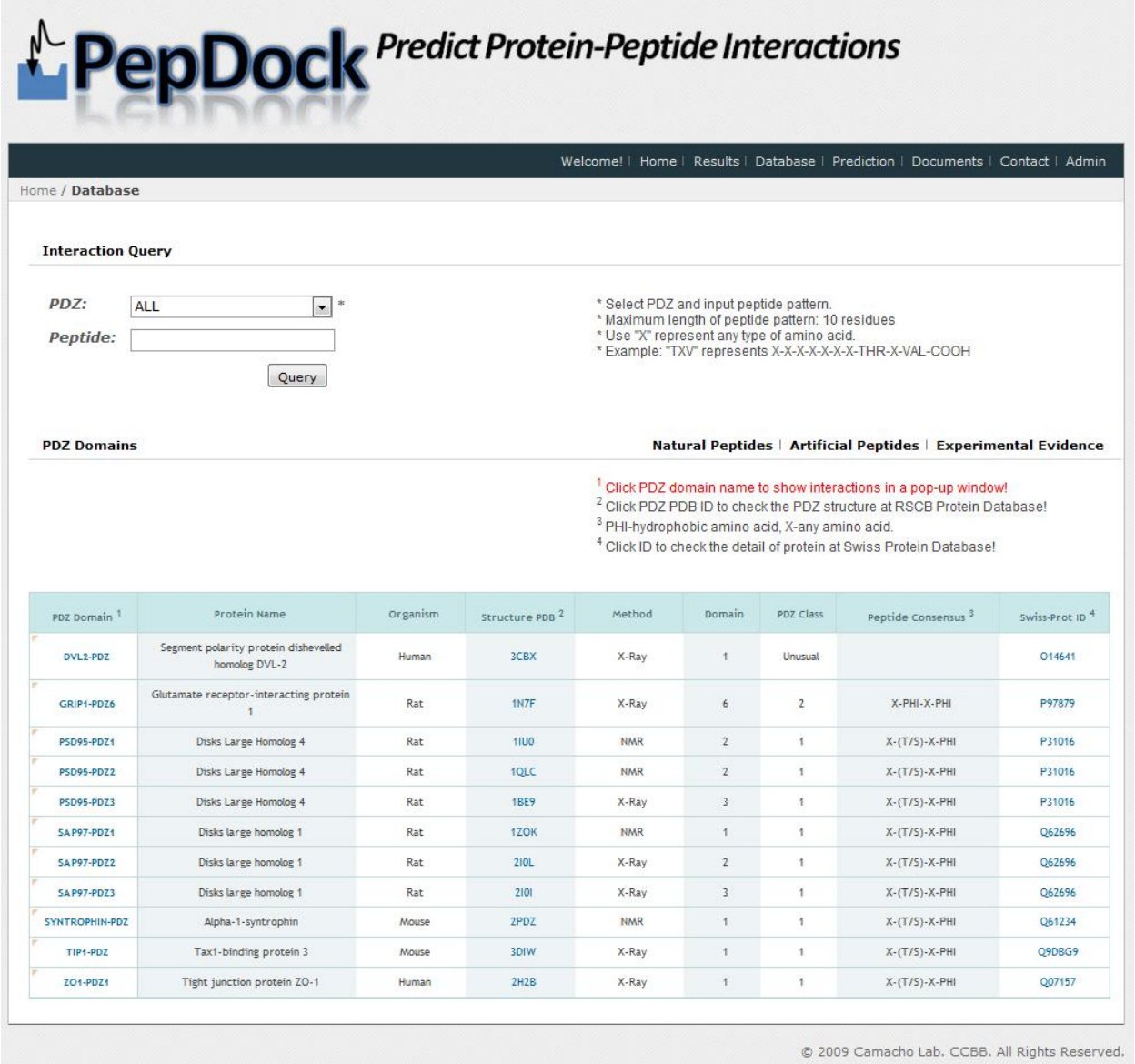


Figure V-2: Database page of *PepDockWeb* portal. The database page of *PepDock* web portal can help users check all pre-computed PDZ-peptide interaction results and relative experimental or literuature information. The database includes human and artificial peptides cross binding against 11 PDZ domains. Users can either query interactions by specify the PDZ domain and input the peptide sequence pattern in the top section or can click the PDZ domain name to browse all data records with respect to this domain. In additon, users can browse all peptide information by clicking the *Natural Peptides* or the *Artificial Peptide*s link and find desired interaction from there. The *Experimental Evidence* link will list all experimental information relevant to peptide and PDZ domains in the database.

The *Prediction Result* page represents the detailed prediction results of one PDZ–peptide interaction, which is provided with three top models ranked by binding affinity estimation. The top panel (Figure V-4) in the page displays complex structures of predictions by using *Jmol* molecule visualization plug-in [108] and the bottom panel shows the free energy landscapes of each models, residual contribution, and known functionalities of the peptide and PDZ domain (Figure V-5). With the full capability of *Jmol*, users can interactively visualize the selected residual contacts between PDZ and peptides, as well as the properties of the surrounding region in the display panel, and compare the difference between binding models. Both free energy change and residual contribution, including conformation entropy change, are both shown in the plots for user to compare and identify the key residues. A summary section includes the gene and structural information about PDZ domain in the prediction, peptide information, and whether available experimental data has confirmed the interaction. In the function prediction section, the functionality and cellular component of peptide and PDZ domains are listed, together with literature references, if available. All these information are important for users to study the functionality of PDZ domains and predict the new interaction in the signal pathway. We believe the *PepDockWeb* portal can facilitate the analysis and help user to find relative targets to PDZ domain.

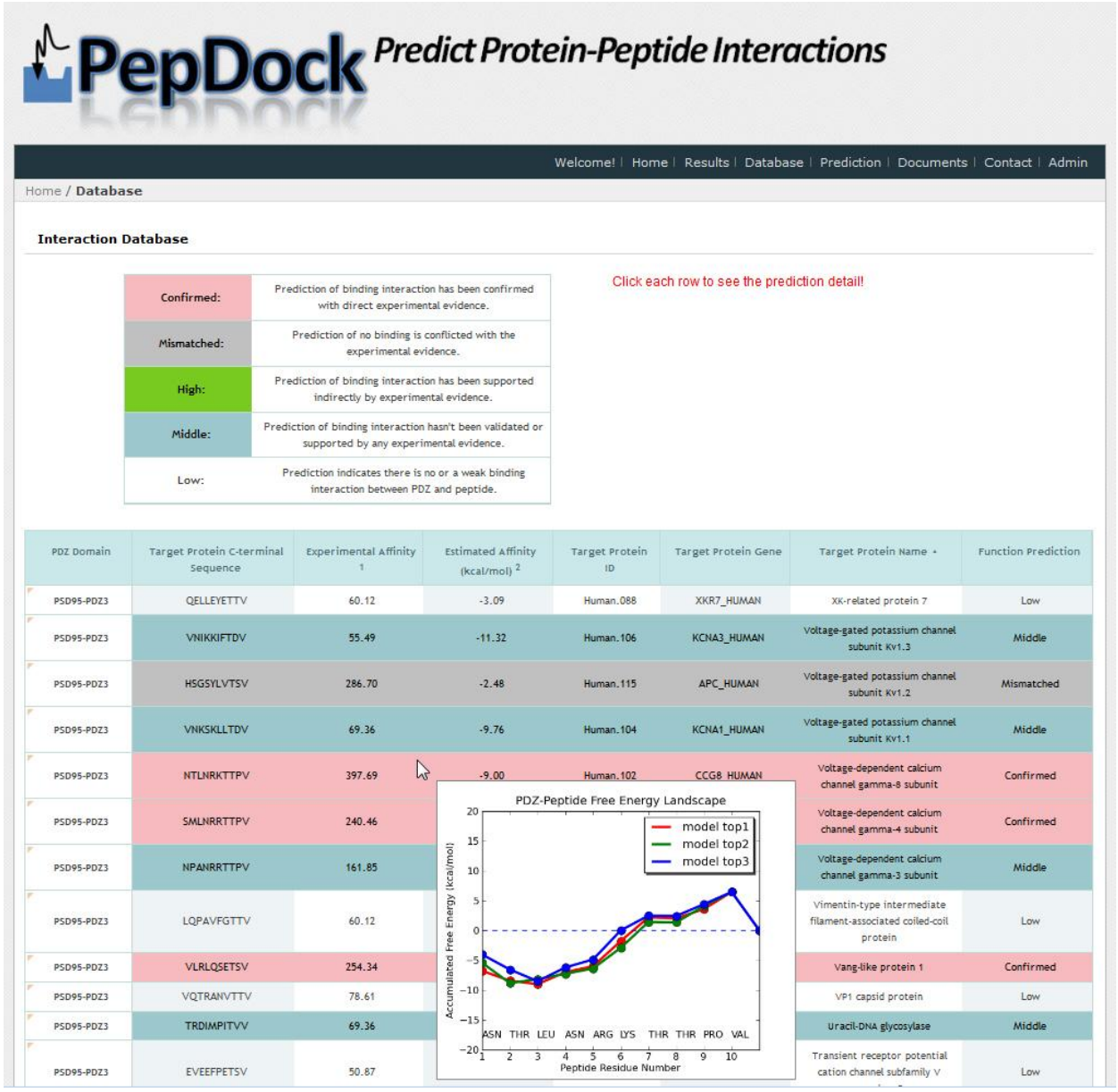


Figure V-3: Database query page of *PepDockWeb* portal. Database query page display all interaction data records which user queried from *PepDock* database front page. All interaction records are classified into five categories and displayed with different colors, i.e. Confirmed, Mismatched, High, Middle and Low. When hovering over each row data record, the free energy landscape of top three prediction results will automatically displayed and give user a quick view of the interaction. Users can go into the detailed prediction result page by clicking each row.

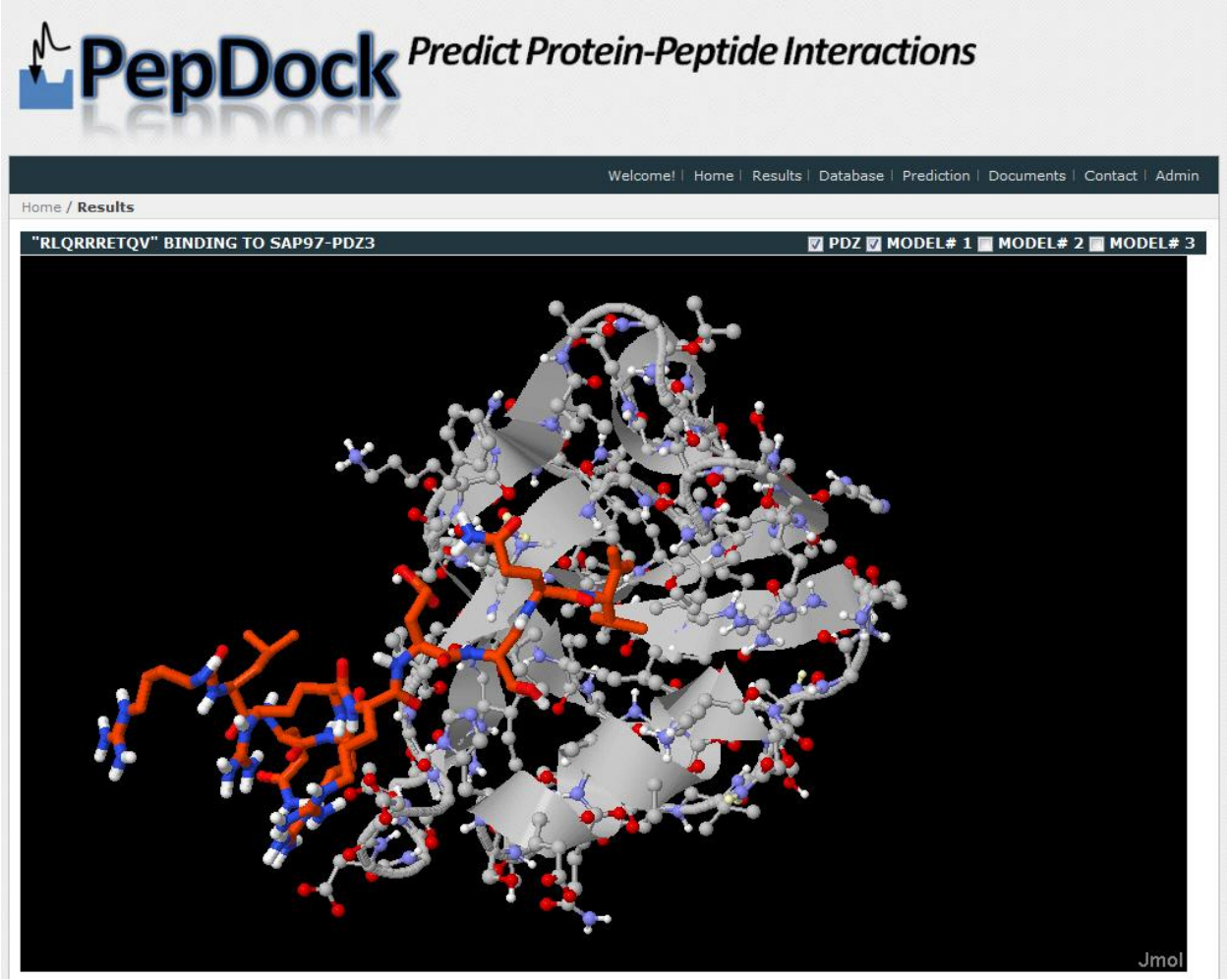
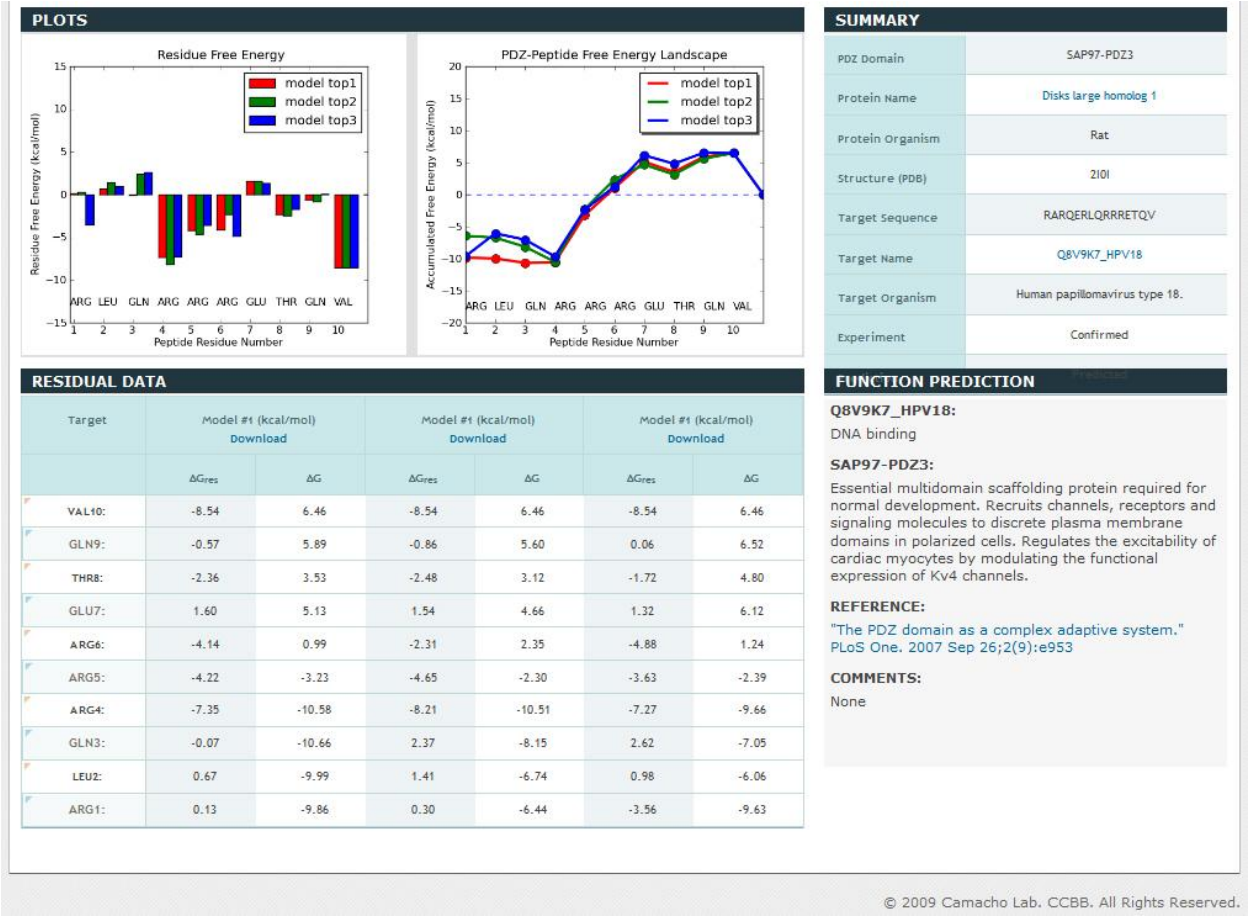


Figure V-4: Visualization panel of prediction result page of *PepDock* web portal. *Jmol* molecule visualization plug-in module is used to display the complex structures of top three prediction models ranked by computed affinity. Users can show/hide molecule models by changing the check box on the right top of the panel. In addition, the visualization panel provides full functionality of *Jmol* and user can display the structure models in different views by changing the properties though *Jmol* operations.



Target	Model #1 (kcal/mol) Download		Model #1 (kcal/mol) Download		Model #1 (kcal/mol) Download	
	ΔG_{res}	ΔG	ΔG_{res}	ΔG	ΔG_{res}	ΔG
VAL10:	-8.54	6.46	-8.54	6.46	-8.54	6.46
GLN9:	-0.57	5.89	-0.86	5.60	0.06	6.52
THR8:	-2.36	3.53	-2.48	3.12	-1.72	4.80
GLU7:	1.60	5.13	1.54	4.66	1.32	6.12
ARG6:	-4.14	0.99	-2.31	2.35	-4.88	1.24
ARG5:	-4.22	-3.23	-4.65	-2.30	-3.63	-2.39
ARG4:	-7.35	-10.58	-8.21	-10.51	-7.27	-9.66
GLN3:	-0.07	-10.66	2.37	-8.15	2.62	-7.05
LEU2:	0.67	-9.99	1.41	-6.74	0.98	-6.06
ARG1:	0.13	-9.86	0.30	-6.44	-3.56	-9.63

Figure V-5: Data panel of prediction result page of *PepDockWeb* portal. Data panel shows the detailed results of top three prediction models, including *Plots*, *Residual Data*, *Summary*, and *Function Prediction*. *Plots and Residual Data* sections show the free energy landscape and residual contribution upon binding interaction. Please notice that the peptide residue numbered *10* is the C-terminal (anchor) residue, which is usually shown as *0* before. *Summary* section summarizes information of the peptide and PDZ domain, and user can check more detail through the link to UniProt database. The *Function Prediction* section presents the biological function of PDZ domain and peptide, as well as direct or indirect literature reference, if available. All together, interaction predictions and information about functionality provide user a good start point to forecast new biological functionality involving PDZ domain and disordered peptides.

3. Prediction

The database includes pre-calculated estimations between selected native peptides against 11 PDZ domains. For those peptides that are not included or synthesized, *PepDockWeb* provides the functionality for users to input the peptide residual sequence and submit prediction jobs online (Figure V-6). Known interaction target sequence consensus to each PDZ domain is shown on the page for users to refer to. The computational process runs on a 10-node (2 CPU/node) computer cluster and normally finishes in 30 minutes, which may vary depending on the load of cluster. When it is complete, an email will be sent to the user, including the web link to retrieve the results. The prediction result is presented in the same format as we described in the prediction result page and will be kept on the server for 30 days.

B. PREDICT NEW INTERACTIONS BY USING *PEPDOCK*

The *PepDock* web portal provides an interface for users to access *PepDock* methodology and can facilitate the analysis of PDZ–peptide interactions with regard to biological functionality and suitability for drug design.

Wnt signaling pathways play critical roles in embryonic and postembryonic development and have been implicated in tumorigenesis [109,110,111,112]. In the Wnt-β-catenin pathway, secreted Wnt glycoproteins bind to seven trans-membrane Frizzled (Fz) receptors and activate intracellular Dishevelled (Dvl) proteins. Activated Dvl proteins then inhibit glycogen synthase kinase-3β (GSK-3 β); this inhibition causes destabilization of a molecular complex formed by

GSK-3β, Adenomatous Polyposis Coli (APC), axin, and β-catenin, and weakens the ability of GSK-3β to phosphorylate β-catenin. Unphosphorylated β-catenin proteins escape from ubiquination and degradation and accumulate in the cytoplasm. This accumulation leads to the translocation of β-catenin into the nucleus, where it stimulates transcription of Wnt target genes. Numerous reports address mutations of Wnt-β-catenin signaling pathway components that are involved in the development of neoplasia [113,114].

Dvl proteins that have a DIX domain, a central PDZ domain, and a DEP domain relay the Wnt signals from membrane-bound receptors to downstream components and thereby play an essential role in the Wnt signaling pathway. Of these three, the PDZ domains, which make a connection between the membrane-bound receptor and downstream components of the pathways, play an important role not only in distinguishing the canonical and non-canonical Wnt pathways but also in nuclear localization [115]. Experiments [116] showed that Dvl PDZ interacts directly with Fz receptors by recognition of an internal motif lacking a free C-terminus. This evidence revealed that the PDZ domain of human Dvl2 (Dvl2-PDZ) recognizes C termini that differ significantly from typical PDZ ligands [97]. Recently, Zhang and his co-workers have conducted a detailed study [104] to solve the crystal structures of four different Dvl2-PDZ complexes and shown that a flexible binding cleft of Dvl2-PDZ is capable of accommodating both C-terminal and internal ligands. This study also showed that a peptide ligand recognizes Dvl2-PDZ domains in cells and inhibits Wnt/β-catenin pathway. Therefore, interference with PDZ domains may be a viable therapeutic strategy for inhibiting Wnt signaling in cancers that are dependent on Dvl function. Small organic inhibitors of the Dvl2-PDZ domain might be useful in dissecting molecular mechanisms and formulating pharmaceutical agents. Because the

structure of Dvl2-PDZ domain is known, this has permitted us to use our structure-based computational method to screen potential ligands.

We used *PepDock* to screen 126 human peptide sequences [88] for potential ligands that could fit into the binding groove of Dvl2-PDZ domain. The peptide backbone library is extracted from the molecular dynamic simulation of the synthetic peptide "$WKWYGWF_{COOH}$," by using the protocol described in section IV.B.1. Then each peptide sequence is docked into the binding groove of Dvl2-PDZ domain (bound structure from PDB entry 3CBX) [104]. Docked models of each peptide sequences are ranked by binding free energy computed by *PepDock* and the top three models are saved for further analysis.

We dock the synthetic peptide "$WKWYGWF_{COOH}$" into the PDZ domain first. This synthetic peptide has been experimentally identified as a Dvl2-PDZ partner and has a crystal complex structure. Our top-ranked docked model has a conformation similar to that found in the crystal structure with a backbone RMSD 0.52 Å and calculated binding free energy −11.91 kcal/mol (Figure IV-20). This contrast indicates that *PepDock* is able to sample and evaluate the ligands binding to Dvl2-PDZ domain. Among 126 human peptides, 92 peptides show binding abilities to Dvl2-PDZ domain, and 10 peptides have stronger binding scores than the reference, $WKWYGWF_{COOH}$ peptide. A full list of screening result is available online at the PepDock website.

In the above experiment, we screened only a limited set of human peptides with C-terminal motif that conform canonical consensus due to the limitation of resource and time. Also we did not conduct experimental fluorescence spectroscopy or Elisa analysis to further validate our prediction. But the results show *PepDock* is a reliable resource to predict new interaction and can help user design pharmaceutical candidates to inhibit PDZ domains.

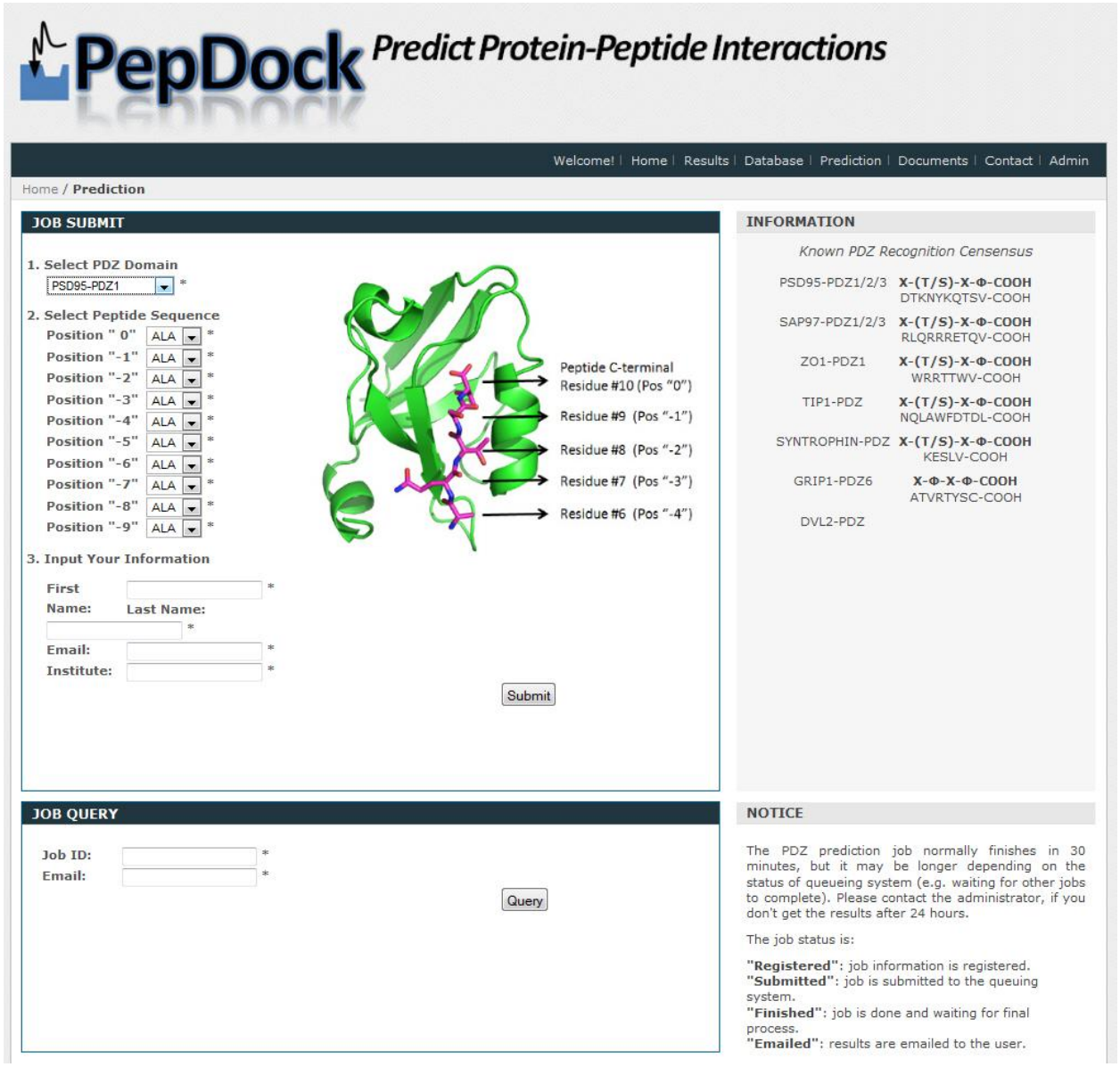


Figure V-6: Prediction of PDZ–peptide interaction by *PepDockWeb* portal. Users can submit new interaction prediction online by select a PDZ domain and input peptide residue sequence through the *PepDockWeb* portal. Currently, 11 PDZ domains are ready to be predicted. Each prediction job will be finished in 30 minutes. When it is completed, an email including the web link to retrieve the results will be sent to the user.

VI. CONCLUSION AND OUTLOOK

A. ACCOMPLISHMENT

With the continuous increase of computer technology in the last decade, molecule docking plays an increasingly important role in the biophysical field. In part, this is due to the results of creating advanced algorithms, which are available to the community through web servers, and the exploration of more structural information by experiment. For example, protein-ligand docking, which has been heavily used for new drug discovery, faces the challenge of fully exploiting the rapidly increasing protein and chemical structure libraries. Although more advanced algorithms have been applied to docking, there are still exist challenges. First, most docking methods use relative free energy or statistical scoring functions. These functions need to be trained before use to get good performance. A lightweight, accurate, and general free energy scoring function, which can estimate absolute binding affinity, is needed. Second, a docking method, which can accommodate the flexibility of the ligand, is needed to study protein recognition with disordered regions. Third, with protein–protein interface knowledge and docking methods, a methodology to predict protein's functionality and construct the protein interaction network is needed.

In Chapter III, we demonstrated the implementation of our free energy scoring functions for protein–protein interactions and protein–disordered peptide interactions respectively. The

function of protein–protein interaction, which is generic and without any training, has been used in our rigid body docking program. In addition, we showed that it could estimate absolute binding affinity and screen strong binders in Capri Target 45 and successfully discriminate native protein complexes from designed models.

Chapter VI focuses on the study of interactions involving disordered peptides. Based on the free energy scoring function model, we designed and implemented a novel structure based computational docking program, *PepDock*, to predict structured protein–disordered peptide interactions. Taking the 3-D structure of receptor protein and amino acid sequences as the input, *PepDock* can estimate the absolute binding affinity and the complex structure. To explore the binding mechanism of disordered peptides, we studied the interactions between peptides with the PDZ domain, one common scaffold protein in signal transduction, by using *PepDock*. *PepDock* successfully discriminated strong binders from non-binders with 91% sensitivity and 74% specificity, when comparing with experimental array results. In addition, *PepDock* successfully mimicked the X-ray crystallographic complex structures of seven peptides against five PDZ domains, capturing the main contact characters. By analysis of the results, we found that peptides binding PDZ domain interactions followed a downhill free energy pathway. Before association, peptides are flexible and take any conformation. When binding, first, the carboxyl termini of peptides anchor into the binding pocket, contributing the most binding influence. Then, the backbone of the next three residues forms an anti-parallel beta sheet by paired hydrogen bonds with the backbone of PDZ the domain, forming a non-specific complex. Next, the remaining residues of peptides will zip onto the surface of PDZ domain, which determines the specificities. These downhill pathways, with non-specific intermediate complexes anchored by the C-terminal motif, are thermodynamically and kinetically favorable to the scaffold protein. Disordered

peptides have an advantage of tuning the maximum binding specificity over structured binding partners [15].

Based on *PepDock*, we developed an online database query and interaction prediction system for PDZ domains. More than 100 native peptide sequences have been selected from gene databases and pre-calculated against 11 PDZ domains by *PepDock*. Each interaction is provided with the top three prediction complex structure models ranked by estimated binding affinity and by their corresponding free energy landscapes. Based on the estimations and experimental evidences, the interactions are tagged into five classes: confirmed, mismatched, high, middle, and low confidence of binding. The free energy contribution of each residue including residual conformation entropy loss is also presented for the user to identify the key residue. For native peptides and PDZ domains, common functionalities and cellular components are listed with relative experimental reference literature to help users to do the functional prediction. In case desired peptides are not included in the database, users can input the peptide sequence and submit a new prediction job online, from which results are received in 30 minutes. The *PepDockWeb* portal is a very powerful tool and can be used to study PDZ–peptide interactions. We expect that this new technology will contribute significantly to the structural biology and biophysical research community.

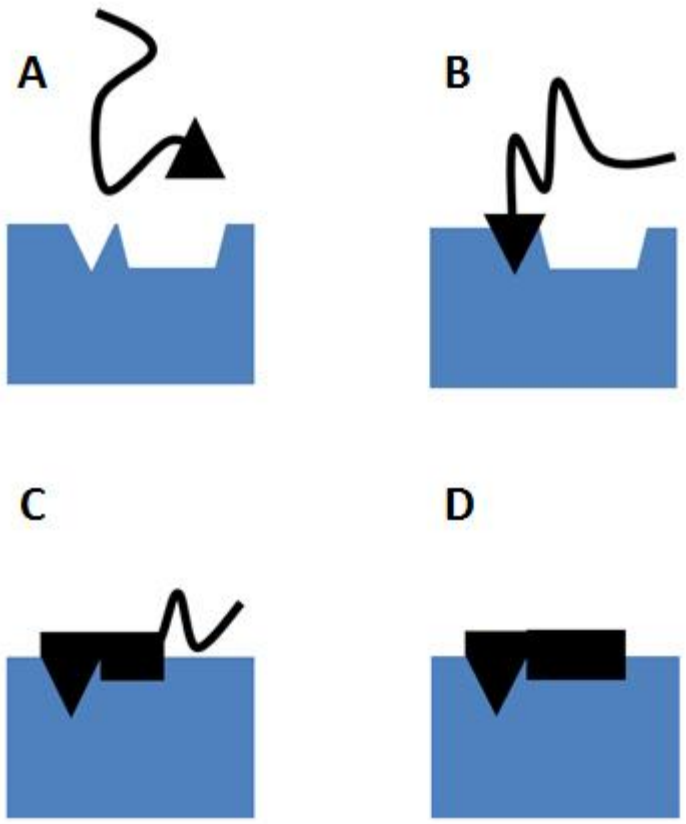


Figure VI-1: Cartoon of disordered peptide binding to PDZ domain. The interaction starts with a completely disordered (flexible) peptide approach the PDZ domain (**A**). The peptide C-terminal residue, which acts as an anchor residue, projects into the binding groove of PDZ domain (**B**). This residue usually contributes most of the binding free energy and compensates the translational/rotational/vibrational entropy loss upon binding. Then the next three to four residues sequentially bind to the PDZ domain and form an anti-beta sheet by backbone–backbone interactions, encountering a non-specific temporary intermediate complex (**C**). The specificity of interaction is determined by the next following peptide residues, which search for the most optimal free energy position on the PDZ surface and zips themselves into an extended network of sequence dependent contacts (**D**).

B. OUTLOOK

Despite the novelty and advancement described, there are further improvements and extensive studies to be done in the future. We have showed in Chapter IV that, to predict the interactions of target PDZ domains, e.g. DVL2-PDZ, which are over 1 Å away from the complex template (PDBID: 1BE9, complex of PSD95-3 PDZ domain with CRIPT peptide), a new complex template from the same PDZ domain cluster is needed to better describe the character of the target PDZ domain. The new complex template will be used to generate the peptide backbone library and as a template to dock the target PDZ domain and peptides. In this work, we clustered all PDZ–peptide complexes from PDB into different groups based on similarity and selected the center complex as the template of each group. Due to the time and computational power limit, we have not extended our prediction testing to other PDZ clusters besides PSD95-3 PDZ, but these works can be done easily by following the same procedure as described.

PDZ domain is one of the most common scaffolding proteins in signal transduction. Many other scaffolding protein domains follow similar binding patterns as we have explored in the PDZ domain study, e.g. SH2, SH3, and PTB domains. These domains have conserved binding grooves, and interaction partners follow certain residual sequence consensus. Because of the generality and portability of our free energy scoring function and docking methodology, *PepDock* can be easily applied to the study of other scaffold proteins. A very preliminary testing on SH3 domains has been conducted by our lab and obtained very good results that are consistent with experimental data. We expect more cases will be studied by following *PepDock* methodology in the future.

BIBLIOGRAPHY

1. Cooper JC (1984) Chinese alchemy : the Taoist quest for immortality. Wellingborough, Northamptonshire: Aquarian Press. 160 p. p.
2. Knight J (2002) Bridging the culture gap. Nature 419: 244-246.
3. Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, et al. (2000) The Protein Data Bank and the challenge of structural genomics. Nat Struct Biol 7 Suppl: 957-959.
4. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 6: 197-208.
5. Wikipedia Intrinsically unstructured proteins, from Wikipedia.
6. Cortese MS, Uversky VN, Dunker AK (2008) Intrinsic disorder in scaffold proteins: getting more from less. Prog Biophys Mol Biol 98: 85-106.
7. Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 293: 321-331.
8. Uversky VN, Eliezer D (2009) Biophysics of Parkinson's disease: structure and aggregation of alpha-synuclein. Curr Protein Pept Sci 10: 483-499.
9. Doyle DA, Lee A, Lewis J, Kim E, Sheng M, et al. (1996) Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. Cell 85: 1067-1076.
10. Pawson T, Scott JD (1997) Signaling through scaffold, anchoring, and adaptor proteins. Science 278: 2075-2080.
11. Songyang Z, Fanning AS, Fu C, Xu J, Marfatia SM, et al. (1997) Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. Science 275: 73-77.
12. Sadowski I, Stone JC, Pawson T (1986) A noncatalytic domain conserved among cytoplasmic protein-tyrosine kinases modifies the kinase function and transforming activity of Fujinami sarcoma virus P130gag-fps. Mol Cell Biol 6: 4396-4408.
13. Pawson T, Schlessingert J (1993) SH2 and SH3 domains. Curr Biol 3: 434-442.
14. Blaikie P, Immanuel D, Wu J, Li N, Yajnik V, et al. (1994) A region in Shc distinct from the SH2 domain can bind tyrosine-phosphorylated growth factor receptors. J Biol Chem 269: 32031-32034.
15. Liu J, Faeder JR, Camacho CJ (2009) Toward a quantitative theory of intrinsically disordered proteins and their function. Proc Natl Acad Sci U S A 106: 19819-19823.
16. Narayanan R, Ganesh OK, Edison AS, Hagen SJ (2008) Kinetics of folding and binding of an intrinsically disordered protein: the inhibitor of yeast aspartic proteinase YPrA. J Am Chem Soc 130: 11477-11485.
17. Sugase K, Dyson HJ, Wright PE (2007) Mechanism of coupled folding and binding of an intrinsically disordered protein. Nature 447: 1021-1025.

18. Liu J, Faeder JR, Camacho CJ (2009) Toward a quantitative theory of intrinsically disordered proteins and their function. Proc Natl Acad Sci U S A In press.
19. Dyson HJ, Wright PE (2002) Coupling of folding and binding for unstructured proteins. Curr Opin Struct Biol 12: 54-60.
20. Chen JR, Chang BH, Allen JE, Stiffler MA, MacBeath G (2008) Predicting PDZ domain-peptide interactions from primary sequences. Nat Biotechnol 26: 1041-1045.
21. Kaufmann K, Shen N, Mizoue L, Meiler J (2011) A physical model for PDZ-domain/peptide interactions. J Mol Model 17: 315-324.
22. Niv MY, Weinstein H (2005) A flexible docking procedure for the exploration of peptide binding selectivity to known structures and homology models of PDZ domains. J Am Chem Soc 127: 14072-14079.
23. Hou T, Chen K, McLaughlin WA, Lu B, Wang W (2006) Computational analysis and prediction of the binding motif and protein interacting partners of the Abl SH3 domain. PLoS Comput Biol 2: e1.
24. Hui S, Xing X, Bader GD (2013) Predicting PDZ domain mediated protein interactions from structure. BMC Bioinformatics 14: 27.
25. Smith CA, Kortemme T (2010) Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains. J Mol Biol 402: 460-474.
26. Dill KA, Bromberg S (2003) Molecular driving forces : statistical thermodynamics in chemistry and biology. New York: Garland Science. xx, 666 p. p.
27. Jackson MB (2006) Molecular and cellular biophysics. Cambridge: Cambridge University Press. xiii, 512 p. p.
28. Fischer E (1894) Einfluss der Configuration auf die Wirkung der Enzyme. Berichte der deutschen chemischen Gesellschaft 27: 9.
29. Koshland DE (1958) Application of a Theory of Enzyme Specificity to Protein Synthesis. Proc Natl Acad Sci U S A 44: 98-104.
30. Rajamani D, Thiel S, Vajda S, Camacho CJ (2004) Anchor residues in protein-protein interactions. Proc Natl Acad Sci U S A 101: 11287-11292.
31. Comeau SR, Gatchell DW, Vajda S, Camacho CJ (2004) ClusPro: a fully automated algorithm for protein-protein docking. Nucleic Acids Res 32: W96-99.
32. Comeau SR, Vajda S, Camacho CJ (2005) Performance of the first protein docking server ClusPro in CAPRI rounds 3-5. Proteins 60: 239-244.
33. Camacho CJ, Gatchell DW, Kimura SR, Vajda S (2000) Scoring docked conformations generated by rigid-body protein-protein docking. Proteins 40: 525-537.
34. Camacho CJ, Vajda S (2002) Protein-protein association kinetics and protein docking. Curr Opin Struct Biol 12: 36-40.
35. Comeau SR, Gatchell DW, Vajda S, Camacho CJ (2004) ClusPro: an automated docking and discrimination method for the prediction of protein complexes. Bioinformatics 20: 45-50.
36. Vajda S, Camacho CJ (2004) Protein-protein docking: is the glass half-full or half-empty? Trends Biotechnol 22: 110-116.
37. Halperin I, Ma B, Wolfson H, Nussinov R (2002) Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins 47: 409-443.
38. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, et al. (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci U S A 89: 2195-2199.

39. Vakser IA, Matar OG, Lam CF (1999) A systematic study of low-resolution recognition in protein--protein complexes. Proc Natl Acad Sci U S A 96: 8477-8482.
40. Elliott DF, Rao KR (1982) Fast transforms : algorithms, analyses, applications. New York: Academic Press. xxii, 488 p. p.
41. Camacho CJ, Kimura SR, DeLisi C, Vajda S (2000) Kinetics of desolvation-mediated protein-protein binding. Biophys J 78: 1094-1105.
42. Camacho CJ, Weng Z, Vajda S, DeLisi C (1999) Free energy landscapes of encounter complexes in protein-protein association. Biophys J 76: 1166-1178.
43. Brooks BR, E. BR, D. OB, J. SD, S. S, et al. (1983) CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. J Comp Chem 4: 187-217.
44. Chen R, Weng Z (2003) A novel shape complementarity scoring function for protein-protein docking. Proteins 51: 397-408.
45. Vajda S, Weng Z, Rosenfeld R, DeLisi C (1994) Effect of conformational flexibility and solvation on receptor-ligand binding free energies. Biochemistry 33: 13977-13988.
46. Brady GP, Sharp KA (1997) Entropy in protein folding and in protein-protein interactions. Curr Opin Struct Biol 7: 215-221.
47. Vajda S, Sippl M, Novotny J (1997) Empirical potentials and functions for protein folding and binding. Curr Opin Struct Biol 7: 222-228.
48. Horton N, Lewis M (1992) Calculation of the free energy of association for protein complexes. Protein Sci 1: 169-181.
49. Jackson RM, Sternberg MJ (1995) A continuum model for protein-protein interactions: application to the docking problem. J Mol Biol 250: 258-275.
50. Kollman P (1993) Free energy calculations: Applications to chemical and biochemical phenomena. Chemical Reviews 93: 23.
51. Honig B, Nicholls A (1995) Classical electrostatics in biology and chemistry. Science 268: 1144-1149.
52. Michael Schaefer MK (1996) A Comprehensive Analytical Treatment of Continuum Electrostatics. The Journal of Physical Chemistry A 100: 22.
53. Qiu S, Hollinger, Still (1997) The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. The Journal of Physical Chemistry A 101: 10.
54. Gilson MK, Given JA, Head MS (1997) A new class of models for computing receptor-ligand binding affinities. Chem Biol 4: 87-92.
55. Zhang C, Vasmatzis G, Cornette JL, DeLisi C (1997) Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol 267: 707-726.
56. Chen R, Li L, Weng Z (2003) ZDOCK: an initial-stage protein-docking algorithm. Proteins 52: 80-87.
57. Chen R, Weng Z (2002) Docking unbound proteins using shape complementarity, desolvation, and electrostatics. Proteins 47: 281-294.
58. Camacho CJ, Gatchell DW (2003) Successful discrimination of protein interactions. Proteins 52: 92-97.
59. Camacho CJ, Vajda S (2001) Protein docking along smooth association pathways. Proc Natl Acad Sci U S A 98: 10636-10641.
60. Camacho CJ, Zhang C (2005) FastContact: rapid estimate of contact and binding free energies. Bioinformatics 21: 2534-2536.

61. Novotny J, Bruccoleri RE, Saul FA (1989) On the attribution of binding energy in antigen-antibody complexes McPC 603, D1.3, and HyHEL-5. Biochemistry 28: 4735-4749.
62. Nauchitel V, Villaverde MC, Sussman F (1995) Solvent accessibility as a predictive tool for the free energy of inhibitor binding to the HIV-1 protease. Protein Sci 4: 1356-1364.
63. Nicholls A, Sharp KA, Honig B (1991) Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins 11: 281-296.
64. Krystek S, Stouch T, Novotny J (1993) Affinity and specificity of serine endopeptidase-protein inhibitor interactions. Empirical free energy calculations based on X-ray crystallographic structures. J Mol Biol 234: 661-679.
65. Alexei V. Finkelstein JJ (1989) The price of lost freedom: entropy of bimolecular complex formation. Protein Eng 3: 3.
66. Janin J (1995) Elusive affinities. Proteins 21: 30-39.
67. Gilson MK, Given JA, Bush BL, McCammon JA (1997) The statistical-thermodynamic basis for computation of binding affinities: a critical review. Biophys J 72: 1047-1069.
68. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, et al. (2003) CAPRI: a Critical Assessment of PRedicted Interactions. Proteins 52: 2-9.
69. Mendez R, Leplae R, De Maria L, Wodak SJ (2003) Assessment of blind predictions of protein-protein interactions: current status of docking methods. Proteins 52: 51-67.
70. Hwang H, Vreven T, Janin J, Weng Z (2010) Protein-protein docking benchmark version 4.0. Proteins 78: 3111-3114.
71. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, et al. (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science 332: 816-821.
72. Chao G, Lau WL, Hackel BJ, Sazinsky SL, Lippow SM, et al. (2006) Isolating and engineering human antibodies using yeast surface display. Nat Protoc 1: 755-768.
73. Hwang H, Pierce B, Mintseris J, Janin J, Weng Z (2008) Protein-protein docking benchmark version 3.0. Proteins 73: 705-709.
74. Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, et al. (2005) Protein-Protein Docking Benchmark 2.0: an update. Proteins 60: 214-216.
75. Camacho CJ, Ma H, Champ PC (2006) Scoring a diverse set of high-quality docked conformations: a metascore based on electrostatic and desolvation interactions. Proteins 63: 868-877.
76. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z (2002) Intrinsic disorder and protein function. Biochemistry 41: 6573-6582.
77. Uversky VN (2002) Natively unfolded proteins: a point where biology waits for physics. Protein Sci 11: 739-756.
78. Shoemaker BA, Portman JJ, Wolynes PG (2000) Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. Proc Natl Acad Sci U S A 97: 8868-8873.
79. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol 323: 573-584.
80. Cho KO, Hunt CA, Kennedy MB (1992) The rat brain postsynaptic density fraction contains a homolog of the Drosophila discs-large tumor suppressor protein. Neuron 9: 929-942.
81. Harris BZ, Lim WA (2001) Mechanism and role of PDZ domains in signaling complex assembly. J Cell Sci 114: 3219-3231.

82. Appleton BA, Zhang Y, Wu P, Yin JP, Hunziker W, et al. (2006) Comparative structural analysis of the Erbin PDZ domain and the first PDZ domain of ZO-1. Insights into determinants of PDZ domain specificity. J Biol Chem 281: 22312-22320.
83. Im YJ, Park SH, Rho SH, Lee JH, Kang GB, et al. (2003) Crystal structure of GRIP1 PDZ6-peptide complex reveals the structural basis for class II PDZ target recognition and PDZ domain-mediated multimerization. J Biol Chem 278: 8501-8507.
84. Zhang J, Yan X, Shi C, Yang X, Guo Y, et al. (2008) Structural basis of beta-catenin recognition by Tax-interacting protein-1. J Mol Biol 384: 255-263.
85. Kornau HC, Schenker LT, Kennedy MB, Seeburg PH (1995) Domain interaction between NMDA receptor subunits and the postsynaptic density protein PSD-95. Science 269: 1737-1740.
86. Kim E, Niethammer M, Rothschild A, Jan YN, Sheng M (1995) Clustering of Shaker-type K+ channels by interaction with a family of membrane-associated guanylate kinases. Nature 378: 85-88.
87. Stricker NL, Christopherson KS, Yi BA, Schatz PJ, Raab RW, et al. (1997) PDZ domain of neuronal nitric oxide synthase recognizes novel C-terminal peptide sequences. Nat Biotechnol 15: 336-342.
88. Kurakin A, Swistowski A, Wu SC, Bredesen DE (2007) The PDZ domain as a complex adaptive system. PLoS ONE 2: e953.
89. Lim IA, Hall DD, Hell JW (2002) Selectivity and promiscuity of the first and second PDZ domains of PSD-95 and synapse-associated protein 102. J Biol Chem 277: 21697-21711.
90. Gianni S, Engstrom A, Larsson M, Calosci N, Malatesta F, et al. (2005) The kinetics of PDZ domain-ligand interactions and implications for the binding mechanism. J Biol Chem 280: 34805-34812.
91. Stiffler MA, Chen JR, Grantcharova VP, Lei Y, Fuchs D, et al. (2007) PDZ domain binding selectivity is optimized across the mouse proteome. Science 317: 364-369.
92. Madsen KL, Beuming T, Niv MY, Chang CW, Dev KK, et al. (2005) Molecular determinants for the complex binding specificity of the PDZ domain in PICK1. J Biol Chem 280: 20539-20548.
93. Joo SH, Pei D (2008) Synthesis and screening of support-bound combinatorial peptide libraries with free C-termini: determination of the sequence specificity of PDZ domains. Biochemistry 47: 3061-3072.
94. Basdevant N, Weinstein H, Ceruso M (2006) Thermodynamic basis for promiscuity and selectivity in protein-protein interactions: PDZ domains, a case study. J Am Chem Soc 128: 12766-12777.
95. Gerek ZN, Keskin O, Ozkan SB (2009) Identification of specificity and promiscuity of PDZ domain interactions through their dynamic behavior. Proteins 77: 796-811.
96. Belda I, Madurga S, Llora X, Martinell M, Tarrago T, et al. (2005) ENPDA: an evolutionary structure-based de novo peptide design algorithm. J Comput Aided Mol Des 19: 585-601.
97. Tonikian R, Zhang Y, Sazinsky SL, Currell B, Yeh JH, et al. (2008) A specificity map for the PDZ domain family. PLoS Biol 6: e239.
98. Dunbrack RL, Jr., Cohen FE (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 6: 1661-1681.
99. D'Aquino JA, Gomez J, Hilser VJ, Lee KH, Amzel LM, et al. (1996) The magnitude of the backbone conformational entropy change in protein folding. Proteins 25: 143-156.

100. Lee KH, Xie D, Freire E, Amzel LM (1994) Estimation of changes in side chain configurational entropy in binding and folding: general methods and application to helix formation. Proteins 20: 68-84.
101. Andrusier N, Mashiach E, Nussinov R, Wolfson HJ (2008) Principles of flexible protein-protein docking. Proteins 73: 271-289.
102. Bueno M, Temiz NA, Camacho CJ (2010) Novel modulation factor quantifies the role of water molecules in protein interactions. Proteins 78: 3226-3234.
103. Temiz NA, Camacho CJ (2009) Experimentally based contact energies decode interactions responsible for protein-DNA affinity and the role of molecular waters at the binding interface. Nucleic Acids Res 37: 4076-4088.
104. Zhang Y, Appleton BA, Wiesmann C, Lau T, Costa M, et al. (2009) Inhibition of Wnt signaling by Dishevelled PDZ peptides. Nat Chem Biol 5: 217-219.
105. Kiel C, Serrano L (2009) Cell type-specific importance of ras-c-raf complex association rate constants for MAPK signaling. Sci Signal 2: ra38.
106. Spolar RS, Record MT, Jr. (1994) Coupling of local folding to site-specific binding of proteins to DNA. Science 263: 777-784.
107. Zarrinpar A, Park SH, Lim WA (2003) Optimization of specificity in a cellular protein interaction network by negative selection. Nature 426: 676-680.
108. Jmol: an open-source Java viewer for chemical structures in 3D.
109. Moon RT, Bowerman B, Boutros M, Perrimon N (2002) The promise and perils of Wnt signaling through beta-catenin. Science 296: 1644-1646.
110. Wodarz A, Nusse R (1998) Mechanisms of Wnt signaling in development. Annu Rev Cell Dev Biol 14: 59-88.
111. Polakis P (2000) Wnt signaling and cancer. Genes Dev 14: 1837-1851.
112. Shan J, Shi DL, Wang J, Zheng J (2005) Identification of a specific inhibitor of the dishevelled PDZ domain. Biochemistry 44: 15495-15503.
113. Polakis P (2007) The many ways of Wnt in cancer. Curr Opin Genet Dev 17: 45-51.
114. Rothbacher U, Laurent MN, Deardorff MA, Klein PS, Cho KW, et al. (2000) Dishevelled phosphorylation, subcellular localization and multimerization regulate its role in early embryogenesis. EMBO J 19: 1010-1022.
115. Itoh K, Brott BK, Bae GU, Ratcliffe MJ, Sokol SY (2005) Nuclear localization is required for Dishevelled function in Wnt/beta-catenin signaling. J Biol 4: 3.
116. Wong HC, Bourdelas A, Krauss A, Lee HJ, Shao Y, et al. (2003) Direct binding of the PDZ domain of Dishevelled to a conserved internal sequence in the C-terminal region of Frizzled. Mol Cell 12: 1251-1260.