Computational Study of Protein Recognition Using Molecular Dynamics, Free Energy Calculation and Sequence Analysis

by

Wei Wang

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA SAN FRANCISCO

[redacted]

Date University Librarian

Degree Conferred: ....

Acknowledgement

I thank my parents for their love, advice and support.

Preface

Four years of graduate school at UCSF has been a valuable experience for me. During this period, I have learned how to do research and become scientifically mature. I am grateful to many people who have given me a lot of help along the way.

The first person I would like to thank is my advisor Peter Kollman. Peter is a wonderful mentor. I appreciate very much that he gave me the freedom and supported me to pursue whatever most interested me. I know that I can always turn to him for great advice and insights. It has been a real pleasure for me to work with Peter in the past few years.

I owe special thanks to Tack Kuntz for being on my thesis committee and the chair of my orals committee. I also would like to thank him for allowing me to attend his group meeting, which has been very educational for me. I have learned not only how to design drugs, but also how to develop my career.

I would like to give my sincere thanks to Ken Dill for his guidance and full support in my preparation for the oral exam and writing my thesis.

I am also deeply grateful to Tom Ferrin, who hosted my first rotation at UCSF and has been supportive to me all these years.

Among other faculty members, I would like to specially thank Dick Shafer, Hao Li, Wendell Lim, Fred Cohen, Dennis Deen, and David Agard for their advice and support. Julie Ransom takes good care of each biophysics student and I am not an exception. I appreciate her help very much.

I am also in debt to many colleagues and friends at UCSF. Yong Duan, Lu Wang, Jian Wang and Kechuan Tu taught me how to use AMBER and also gave me a lot of help when I just joined the group. Conversations with them at the beginning of my scientific career were very beneficial to me. Conrad Huang's help made life much easier for me in my first a few months at UCSF. Junmei Wang is my resort for force field parameters. Xiaoqin Zou and I had many stimulating discussions about mean force and continuum models. Taisung Lee provided me many insights on quantum calculations. I also would like to thank Piotr Cieplak, Bernd Kuhn, Geoff Skillman, Hue San Chan, Connie Oshiro, Shijie Chen, Seth Hopkins, Jed Pitera, Cara Reyes, Rachel Brem, Michelle Lamb, Malin Young, Shuanghong Huo, Ray Luo, and Xinshan Kang for their help at various stages of my research.

Computational study of protein recognition using molecular dynamics, free energy calculation and sequence analysis

by

Wei Wang

Peter Kollman

Peter A. Kollman

Thesis Advisor and Chair of the Thesis Committee

Abstract

Protein-protein and protein-ligand interactions are central in many biological processes. Therefore, it is important to understand the molecular basis of these interactions as well as the general physical principles of protein recognition, which is the aim of this dissertation. I have performed quantitative and qualitative studies of HIV-1 protease dimer stability (protein-protein), interactions between Sem-5 SH3 domain and its ligands (protein-peptide), and HIV drug resistance (protein-ligand) using molecular dynamics, free energy calculation and sequence analysis. The computational simulations provided atomic and dynamic insights of protein interaction processes. Along the way, I proposed a protocol to identify critical residues for binding or folding by combining free energy calculation and sequence analysis, which is useful for predicting protein function and designing resistance-evading drugs for any target. I also improved a new method, the Linear Interaction Energy method, to calculate absolute binding free energy, which is useful in drug design. In summary, my research contributes to the ongoing scientific community efforts to understand how proteins interact with other molecules.

Table of contents

List of Figures and Tables

Chapter 2.

Chapter 3.

Appendix

Chapter 4.

Supporting Material

Chapter 5.

Chapter 7.

Life is one of the greatest miracles in the universe. It is beautiful, elegant and complex. To give a similarly elegant physical explanation of this fantastic phenomenon has been my dream. This dream drove me from physics to biophysics about four years ago. In the past four years (1996-2000), biology has advanced incredibly quickly. The genomes of human and many other species have been sequenced or are about to be sequenced. DNA microarray technology becomes more mature and has been widely used for all kinds of purposes, such as identification of unknown genes. The process of solving protein structures has been accelerating year by year and the resolution of such structures has become higher and higher. These advancements of technologies help to define modern biology as a multidisciplinary area, which requires a new generation of biologists who must understand, in addition to biology and chemistry, computer science, physics and mathematics to analyze and then derive biological insights from the overwhelming amount of biological data generated from these new technologies. This leap of biology also provides an opportunity for computer scientists, physicists, mathematicians and engineers to contribute to deciphering the nature of life. In this dissertation, I have tried my best to exploit my physics background to study biology.

In the post-genome era, it is important to know the functions of each protein in the cell. Protein functions have to be defined in the biological context, i.e. in which biological process it participates, with what partners it interacts and where it becomes active. Therefore, it is critical to understand the principles of protein recognition, which is the focus of this dissertation.

With the advent of more and more computer power, computer simulations have played a more and more significant roles in studying biological problems. In Chapter 2, I review techniques based on molecular dynamics to calculate binding free energies between protein-protein, protein-peptide and protein-ligand. Molecular dynamics and free energy calculation methods have become more efficient and more accurate in the past few years.

In the following three chapters, I have applied and improved these techniques to study interactions between protein-protein (the dimer stability of the HIV-1 protease, Chapter 3), protein-peptide (the Sem-5 SH3 domain interacts with its ligands, Chapter 4) and protein-ligand (the HIV-1 protease interacting with its inhibitors, Chapter 5).

One "classic" protein-protein complex is the HIV-1 protease (HIV PR) which is a dimer consisting of two identical monomers. Craik and coworkers found that several defective monomer mutants can reduce the HIV-1 protease activity and, therefore, reduce the infectivity of the HIV virus. This "dominant negative inhibition" is due to the preference of heterodimer formation between defective and wild type monomer. We have carried out molecular dynamics (MD) on different HIV PR dimers and calculated their binding free energies using the molecular mechanics/Poisson Boltzmann Solvent Area (MM/PBSA) method. The MD simulations showed that both catalytic Asp's are ionic in ligand-free protease and the flap region is less likely to open for the defective heterodimer. Free energy calculations revealed that van der Waals interaction was the dominant factor for binding affinity difference between different dimers. In order to predict new mutation to favor formation of defective heterodimers, a method called the Virtual Mutagenesis (VM) method was developed to scan residues on the dimer interface and suggest mutations according to free energy calculations. This VM method can be applied to study any other protein-protein complexes (Chapter 3).

The Src Homology 3 (SH3) domains recognize certain poly-proline peptide motifs and play crucial roles in signal transduction. Lim and coworkers found that SH3 domains recognize N-substituted residues instead of only proline. The Sem-5 SH3 domain has been studied for several years and it is a good example of protein-peptide interactions. Our MM/PBSA calculations account for the SH3 domain site preferences and reveal that the van der Waals interaction energy between the protein and the peptide is the dominant factor in the preferences. We then focus on site –1 of the ligand and show that the MM/PBSA method with two different charge models, RESP and AM1-BCC, can predict affinities of N-substituted peptoids at site –1. AM1-BCC charges can be calculated more efficiently; thus, our work enables more general use of MM/PBSA in drug design. We also introduce an empirical parameter, the VC value, defined as the product of the van der Waals energy and fraction of sequence conservation, to identify critical residues of the SH3 domain for binding and in this way, we elucidate the site preference of the protein. Critical residues identified by the VC value are consistent with findings from previous experiments and this analysis also suggests one residue, N190, as a possible residue critical for N-substituted recognition (Chapter 4).

Drug resistance has sharply limited the effectiveness of HIV-1 protease inhibitors in AIDS therapy. It is critically important to understand the basis of this resistance for designing new drugs. We have evaluated the free energy contribution of each residue in the HIV protease in binding to one of its substrates and to the 5 FDA approved protease drugs. Analysis of these free energy profiles and the variability at each position suggests: (1) drug resistance mutations are likely to occur at not well conserved residues if they interact more favorably with drugs than with the substrate; (2) resistance-evading drugs should have a similar free energy profile as the substrate and interact most favorably with well conserved residues. This method can assist in designing resistance-evading drugs for any target. We also proposed an empirical parameter, FV (free energy/variability) value, to predict drug resistance mutations (Chapter 5).

A combination of free energy calculation and sequence analysis has been shown to be useful in studying protein-ligand interactions (Chapter 4 and Chapter 5). In Chapter 6, I have discussed philosophy behind this combination and proposed two more applications. One involves the identification of critical residues in the SH3 domain for folding stability and the other involves the identification of residues crucial for binding between the human growth hormone and its receptor. Because many mutagenesis experiments have been done, it will be interesting to compare predictions from the FC (free energy/conservation) value to the experimental data.

Chapter 7 is devoted to discussing calculations of absolute binding free energies for protein-ligand complexes, which is very important for drug design. It has been a challenge to calculate absolute binding free energies. Åqvist and coworkers proposed a Linear Interaction Energy (LIE) method a few years ago to meet this challenge. In this method there are two empirical parameters. Åqvist found a set of values could give good calculated absolute binding free energies for several protein-ligand complexes. However, other groups found that different set of values were needed for different systems. We have investigated this issue and found a correlation between the hydrophobicity of the binding site and the values of the parameters. This correlation was applied to study biotin and its inhibitors and better results than using fixed values of the parameters were obtained.

In Chapter 8, I have summarized the whole dissertation and discussed the future directions and long term goals of computer simulations of protein recognition. The questions that I am interested in finding answers to are: Are there any common features of proteins that bind to a same protein partner? Is diffusion the only factor that determines a protein's binding partner? How do genes regulate each other? and How can one integrate our knowledge of gene expression and protein interactions? I am curious and eager to find out the answers.

Chapter 2

Computer simulations on protein-protein, protein-peptide and protein-ligand interactions

This chapter is part of a review paper submitted to Annual Review of Biophysics and Biomolecular Structure (Wei Wang, Oreola Donnii, Carolina M. Reyes and Peter A. Kollman, "Biomolecular simulations: Recent developments in force fields, simulations of enzyme catalysis and protein-ligand, protein-protein and protein-nucleic acid non-covalent interactions", Annual Review of Biophysics and Biomolecular Structure, 2000, accepted.)

Introduction

Interactions between proteins and their substrates play central roles in many biological processes, such as signal transduction, enzyme cooperativity, and metabolic reactions. With more and more complex structures solved, structure based computational modeling has become a powerful tool to understand and predict binding. In this section, we focus on noncovalent binding and only discuss different methods to estimate absolute or relative binding free energies for protein-protein or protein-ligand interactions.

Molecular dynamics (MD) and Monte Carlo (MC) methods have provided dynamic and atomic insights to understand complicated biological systems. Free Energy Perturbation (FEP) and Thermodynamic Integration (TI) methods have been successfully applied to predict the binding strength of a complex (7, 31, 66). Nonetheless, these methods are computationally intensive. Thus, many techniques, such as the λ-dynamics and the Chemical Monte Carlo/Molecular Dynamics (CMC/MD) method, have been developed to improve their efficiencies, and many other less rigorous methods have also been in development to estimate binding free energies quickly but with reasonable accuracy (1). Among them, we review the Linear Interaction Energy (LIE) method, the Molecular Mechanics/Poisson Boltzmann Surface Area (MM/PBSA) method, the Pictorial Representation of Free Energy Components (PROFEC), the One-Window Free Energy Grid (OWFEG) method and their applications to studying protein-protein or protein-ligand binding.

Free energy perturbation (FEP) and thermodynamic integration (TI) methods

Free energy perturbation (FEP) and thermodynamic integration (TI) methods are the most rigorous methods among those currently available for calculating free energies.

In this section, we focus on the applications of these two methods to protein-ligand or protein-protein complexes.

Suppose that one wishes to calculate the binding free energy difference between two ligands bound to the same protein, the thermodynamic cycle is shown in Scheme 1.

Scheme 1. Thermodynamic cycle for calculating relative binding free energies between two ligands bound to the same protein.

$$\begin{tikzcd}
L1 \arrow[d, "\Delta G_{solv}"'] & + & P \arrow[r, "\Delta G_b^1"] & C1 \arrow[d, "\Delta G_P"'] \\
L2 & + & P \arrow[r, "\Delta G_b^2"] & C2
\end{tikzcd}$$

Thus,

$$\Delta\Delta G = \Delta G_b^1 - \Delta G_b^2 = \Delta G_{solv} - \Delta G_P \quad (3)$$

where ΔG_b^1 and ΔG_b^2 are binding free energies for ligand 1 and 2 respectively, and ΔG_{solv} and ΔG_P are nonphysical transmutation free energy from ligand 1 to ligand 2 in free and bound state. If ligand 1 and 2 are similar to each other, ΔG_{solv} and ΔG_P usually are easier to calculate than ΔG_b^1 and ΔG_b^2 because the mutation from ligand 1 to ligand 2 is assumed to cause only localized changes. FEP or TI is used to calculate ΔG_{solv} and ΔG_P. Equation (4) is used in FEP calculations.

$$\Delta G = -RT\sum_{i=1}^{N-1}\ln\left\langle \exp\left(-\frac{H(\lambda_{i+1})-H(\lambda_i)}{RT}\right)\right\rangle_{\lambda_i} \tag{4}$$

where ΔG is the free energy difference between two states, A and B, λ_i varies from 0 (state A) to 1 (state B), $H(\lambda_i)$ represents Hamiltonian of the system at λ_i and $\langle\rangle_{\lambda_i}$ indicates an ensemble average. With the TI method, one calculates the average of derivatives of Hamiltonian at each λ, H(λ), and then uses numerical integration over λ to calculate the free energy difference between two states (Equation (5)), where λ has the same meaning as in FEP.

$$\Delta G = \int_0^1\left\langle\frac{\partial H(\lambda)}{\partial \lambda}\right\rangle d\lambda \tag{5}$$

In this section, we focus on the progress and applications in the last five years in the field of protein-ligand or protein-protein interactions. Older papers can be found in previous reviews (30, 42).

Free energy calculations combined with molecular dynamics (MD) or Monte Carlo (MC) methods can provide rationale and insights for experimental observations and can suggest new experiments.

Jorgensen and coworkers have been applying MC and free energy calculations to study hydration free energies of organic molecules and binding free energies for ligand-protein complexes. Essex *et al.* successfully applied MC and FEP to calculate accurate relative binding free energies for trypsin-benzamidine complexes (20). Their simulation was able to predict the strongest inhibitor among the four trypsin inhibitors. They also showed that changes of hydrogen bonding could not rationalize the calculated free energies. Instead, the relative binding affinities were justified in terms of bulk-solvation arguments whereby the more polar inhibitors had weaker binding affinities. This study is an excellent example of combining MC and FEP to study macromolecular systems.

Rastelli *et al.* exploited FEP simulations to rationalize the binding differences between a benzocinnolinone carboxylic acid inhibitor of aldose reductase and its methoxylated analogs in four selected substitution sites (55). They were able to reproduce the experimental trend of binding. The four substitution sites were at the interface between protein residues and water. Thus, the perturbation involved only partial desolvation. This work sheds light on how to design new inhibitors targeting sites in the protein-water interface.

Fox *et al.* (22) calculated the relative binding free energies between two transition state analog substrates of the catalytic antibody 17E8 using TI. The two substrates differ only in one side chain. The substitution of the -CH2- group to –S- leads to a 0.9-1.3 kcal/mol less favorable binding free energy. Their calculations showed that this preference for the –CH2- group over the –S- group was mainly due to the more favorable solvation free energy in the unbound form of the substrates. Free energy component analysis of the van der Waals and electrostatic contributions to the binding free energy indicated that these two terms contributed equally in solvent, whereas in the antibody, the van der Waals term clearly dominated. Several residues with large contributions to the binding were reported and new site-specific mutagenesis experiments were suggested to test the calculated results.

Combined with MD and other simulation methods, free energy calculations can help determine some properties of the biological systems, such as binding mode, protonation state of certain residues, and the flexibility of certain parts of the molecule.

Several MD and FEP simulations have been performed to study carbohydrate-protein complexes (39, 47, 74). Zacharias *et al.* (74) applied FEP simulations to study the differential binding between arabinose and fucose with arabinose binding protein (ABP). Liang *et al.* (39) reported MD and FEP simulation results on binding of mannose versus galactose with a mannose binding protein (MBP). Recently, Pathiaseril and Woods (47) examined binding between analogs of the wild type trisaccharide epitope of *Salmonella* serotype B and a fragment of the monoclonal anti- *Salmonella* antibody Se155-4. All of these simulations obtained free energy results consistent with experimental data. In the study of Pathiaseril & Woods (47), they were able to reproduce the relative Nuclear Overhauser Effect (NOE) intensities for the wild type ligand in solution and intermolecular hydrogen bond patterns in the complex from their MD trajectories. Their simulations showed that the free oligosaccharide oscillated around well-defined average glycosidic torsion angles and the bound conformation was encompassed within those observed for the free ligand. Their free energy calculations also suggested that HIS-97 was diprotonated in the antibody. The insights obtained from these theoretical studies can be employed in the design of new ligands with higher binding affinity.

Sotriffer *et al.* (62) showed a good example of how theoretical work could be carried out without direct experimental data and they were able to provide useful information for designing new ligands. Starting from an antibody structure obtained by homology modeling (14), Sotriffer *et al.* performed extensive docking searches to identify two pockets, S1 and S2, in the antibody IgE LB4 as the most probable binding site for three dinitrophenyl (DNP) amino acids (DNP-alanine, DNP-glycine, and DNP-serine). MD and FEP simulations were carried out for complexation on both pockets. A closed thermodynamic cycle was formed by transmutation between the three DNP amino acids (DNP-Ser -> DNP-Ala -> DNP-Gly -> DNP-Ser), and the FEP calculations were validated by small closure errors of this cycle. The experimental free energy differences could only be reproduced for ligands binding to S1 site. Analysis of the MD trajectory showed that the S1 complexes were characterized by a uniform binding mode, whereas ligand binding in the S2 site exhibited considerable variability. The authors concluded that the S1 site was expected to be the "real" binding site of those DNP amino acids. These theoretical predictions can be examined by crystallography or NMR experiments.

HIV-1 protease has been a therapeutic target for five FDA approved AIDS drugs. Many free energy calculations have been performed on different inhibitors binding with the protease (10, 21, 52-54, 56, 65). Rao & Murcko have calculated relative binding free energies between HIV protease inhibitor A74704 and its diester analog (53). The diester analog inhibitor missed two hydrogen bonds with the protease active site but its binding affinity was only ten fold weaker. They observed that Gly27 and Gly27′ loops were flexible and, thus, the hydrogen bonds between the inhibitor P1 and P1′ NH groups and the carbonyls of Gly27 and Gly27′ of the enzyme were weaker than those hydrogen bonds formed between the inhibitor and the flap water. Therefore, the net gain of binding due to hydrogen bond formation between the inhibitor and flexible parts of the enzyme was offset by the desolvation penalty of the polar hydrogen bonding groups and was unlikely to significantly increase binding strength. They pointed out that hydrophobic interactions with the enzyme and hydrogen bonding interactions with the two catalytic aspartates in the active site were crucial for potent inhibitors. Rick *et al.* studied the drug resistant mutant I84V of the HIV-1 protease binding with three potent inhibitors, KNI-

272, Indinavir and Saquinavir (57). They applied TI to calculate relative binding free energies between the wild type enzyme and the I84V mutant. Because HIV protease is a homodimer, the perturbation involves I84V and I84′V mutations. They found that the free energy contribution from each side chain was correlated with the other side chain. The free energy from I84′V was more variable among the three inhibitors due to the different P1′ group of the three inhibitors and therefore, to different cavity sizes in the mutant complex. They observed that the cavity size, measured either in cavity volume or surface area, correlated very well with the measured free energy changes with slopes in the range of that found for protein stability. This similarity is perhaps due to the peptidic nature of the inhibitors. McCarrick & Kollman carried out FEP simulations on haloperidol thioketal (THK) and three of its derivatives bound with HIV protease (43). Their simulations predicted tighter binding THK derivatives than the present THK compound.

FEP and TI have been widely exploited to calculate relative binding free energy for similar organic systems. Progress has been made in calculating absolute binding free energies for protein-ligand and DNA-ligand complexes as well (26, 44). Recently, Helms and Wade (26) reported the calculated absolute binding free energy for camphor binding to P450cam from *Pseudonomas putida.* By mutating the camphor into six water molecules in the binding site, they were able to reproduce the absolute binding free energy within 3 kJ/mol (<1.0 kcal/mol) of the experimental value.

It is well known that the most severe limitation in free energy calculations is sampling conformational space (5). It is not just a matter of sampling longer, but also sampling in the correct region of conformation space. In order to achieve "good" sampling, long range electrostatic interactions and molecular polarization have to be treated appropriately. In their study of organic cations bound to a cyclophane host, Eriksson *et al.* showed that using a non-additive force field, which is necessary for considering polarization, and the Particle Mesh Ewald (PME) method, to consider long range electrostatic interactions, can improve the calculated relative free energy of association of an imminium (IM) and a guanidinium (GU) binding to the host from -2.3 kcal/mol to –4.0 kcal/mol, compared to a measured value of –3.7 kcal/mol (15). Recently, Ota et al. proposed a Non-Boltzmann Thermodynamic Integration (NBTI) method, which is a combination of TI and umbrella sampling (umbrella sampling attempts to overcome the sampling problem by modifying the potential function so that the unfavorable states are sampled sufficiently) (2, 37), to enhance sampling conformational space for macromolecular systems (45, 46) and applied this method to calculate the relative binding free energy between benzamidine (BZD) and benzylamine (BZA) associated with trypsin. The calculated free energy value using NBTI (2.2 kcal/mol) was much closer to the measured value 2.6 kcal/mol than the value 0.8 kcal/mol obtained using conventional TI. This result is very encouraging.

Erion & Reddy have reported a new method that uses both QM and FEP methods for calculating relative changes in the hydration free energies between two similar molecules (17). Recently they applied this method in designing inhibitors for adenosine deaminase and cytidine deaminase (18). They showed that heteroaromatic hydration was controlled by a multitude of molecular factors. Their calculation of relative inhibitor potencies for adenosine deaminase agrees well with the experimental data (19).

As we mentioned above, FEP and TI are most rigorous methods and in principle can be used to calculate any free energy difference. Recent progress in developing and applying these methods to study complex macromolecular systems is promising. Combined with other simulation methods, such as homology modeling and docking, FEP and TI will become more powerful tools in understanding biological problems.

"Multimolecule" free energy calculation methods

FEP and TI methods are intrinsically "pair-wise" methods, i.e. each FEP/TI simulation has to be performed to obtain free energy difference between two states/molecules. It is more computationally efficient if the free energy differences between several states/molecules can be calculated in one simulation. Such "multimolecule" free energy calculation methods have been developed (23, 24, 33, 40, 50). They are specifically useful in calculating relative binding free energies for several similar ligands.

Brooks and coworkers have developed a new approach called λ-dynamics to evaluate relative hydration free energies or binding free energies between several molecules in a single run of simulation (23, 24, 33). In this λ-dynamics approach, they treated λ in Equation (3) as a set of variables λ_j $\{j=1, n\}$ and each molecule was assigned a λ_j. $\{\lambda_j=0; j=1,n\}$ and $\{\lambda_j=1; j=1,n\}$ corresponded to start and end states respectively. An extended Hamiltonian of the whole system $H_{extended}\{\lambda_j, j=1,n\}$was a combination of the n molecules' Hamiltonians, a kinetic energy term associated with a set of fictitious masses and an umbrella potential, the potential function used in umbrella sampling. In order to optimize $H_{extended}\{\lambda_j, j=1,n\}$ along a pathway from start to end state, the n molecules competed with each other. When the simulations reached equilibrium, different molecules had different λ values. The Weighted Histogram Analysis Method (WHAM) was then employed to generate free energy contours. This approach was successfully demonstrated in calculating the hydration free energies of several small organic molecules (CH_3CH_3, CH_3OH, CH_3SH and CH_3CN) and identifying the best binder to trypsin among benzamidine and three of its paraderivatives. The results obtained from λ-dynamics approach were consistent with experimental data and conventional FEP calculations (6, 16, 23, 24, 33, 50, 64).

Recently, Eriksson *et al.* calculated binding free energies of TIBO-like HIV-1 reverse transcriptase (RT) inhibitors (16). In their study (16), the adaptive chemical Monte Carlo/molecular dynamics (CMC/MD), another "multimolecule" free energy calculation method, was exploited to rank 13 different TIBO derivatives with respect to their relative free energies. The CMC/MD method was developed by Pitera and Kollman and was able to rank binding affinities for several ligands in a single MD simulation (50). The MD was used to sample conformations of each ligand, and the MC was used to sample "chemical space" of all ligands (6, 50, 64). A MD run started from the complex of the receptor and one of the several ligands. After a certain period of MD simulation, a mutation from the ligand to any ligand under consideration occurred. The Metropolis criteria was used to determine whether or not this mutation was accepted. At the end of the simulation, free energy differences between ligands could be obtained by analyzing the populations of each ligand, that is ligands chosen more often by MC were assumed to bind more tightly to RT than those ligands chosen less often by MC in the whole simulation. The calculated values were consistent with measured ones and some results were also confirmed by the Poisson-Boltzmann/solvent accessibility (PB/SA) method and

FEP/TI methods (16). One new derivative, suggested by the program PROFEC (Pictorial Representation of Free Energy Components, see below) (51), was predicted to bind 1-2 kcal/mol better than the starting ligand, R86183 (8Cl-TIBO).

PROFEC and OWFEG

In drug design, the question often asked is "What changes can be made to improve the binding constant?". Recently two methods have been developed to suggest promising changes to improve the binding (38, 51). In their study of trypsin and its inhibitors (51), Radmer & Kollman have calculated the approximate free energy at each grid point (a probe was put at that point and a single window FEP was performed) surrounding an interesting region of one of the trypsin inhibitor, benzamidine. Free energies of all grid points were then displayed as contour surfaces around the inhibitor. This PROFEC (Pictorial Representation of Free Energy Components) method could quantitatively suggest relatively more favorable regions for molecular change and was shown promising to rank 9 trypsin inhibitors. Recently, Lee & Kollman (38) showed the strength of combining FEP and PROFEC methods to predict more potent inhibitors of thymidylate synthase. Thymidylate synthase (TS) is an enzyme that catalyzes dTMP synthesis for DNA synthesis. Inhibition of TS can block dTMP synthesis and therefore implies chemotherapeutic use to combat cancer. Jones et al. designed and synthesized 31 inhibitors of TS, most of which had low water solubility (28). Lee & Kollman predicted new stronger inhibitors modified from one of the Jones et al inhibitors using PROFEC and confirmed the prediction by TI calculations. Their simulations provided guidelines for designing new potent inhibitors of TS with better solubility.

OWFEG (49) has made two modifications of PROFEC. First, each grid point underwent translation and rotation along with the atom of the ligand to which it was closest. Thus, flexible regions of the ligand could be explored. Second, three probes with neutral, positive and negative charges were used instead of only a neutral probe to examine the desirability of introducing charged groups along the grid. This feature provided hints as to what type of charges should be placed at that grid point. In two test systems, quinoline and bis-pyrimidine, and FKBP-12 FK506 protein-ligand complex, the qualitative results derived from OWFEG showed excellent agreement with the standard TI simulations (49).

Linear Interaction Energy method

The Linear Interaction Energy (LIE) method was originally proposed by Åqvist *et al.* to estimate the absolute binding free energies. The LIE method is based on linear response assumptions, that is, the solvent polarization responses to changes in the electrostatic field exerted by the solute is linear and characterized by a single dielectric constant (4). It divides the interaction between the ligand and its environment into electrostatic and van der Waals parts. The binding free energy is estimated as

$$\begin{aligned}\Delta G_{\text{bind}} &= \Delta G^{\text{el}}_{\text{bind}} + \Delta G^{\text{vdw}}_{\text{bind}} \\ &\approx \alpha < V^{\text{el}}_{\text{bound}} - V^{\text{el}}_{\text{free}} > + \beta < V^{\text{vdw}}_{\text{bound}} - V^{\text{vdw}}_{\text{free}} > \qquad (6)\end{aligned}$$

where $V^{\text{el}}_{\text{bound}}$ and $V^{\text{vdw}}_{\text{bound}}$ are the electrostatic and van der Waals interaction energies between the ligand and the solvated protein from a MD trajectory with ligand bound to protein and $V^{\text{el}}_{\text{free}}$ and $V^{\text{vdw}}_{\text{free}}$ are electrostatic and van der Waals interaction energies between the ligand and the water from an MD trajectory with the ligand in water, < > denotes an ensemble average, and α and β are two empirical parameters.

Åqvist and coworkers have applied this method to calculate absolute binding free energies of several protein-ligand complexes. They found that $\alpha = 0.5$ and $\beta = 0.16$ gave calculated binding free energies in good agreement with experimental data. In the calibration set, four inhibitors bound to endothiapepsin, this set of parameters gave a mean unsigned error of 0.39 kcal/mol and 0.59 kcal/mol for calculated absolute and relative binding free energies respectively. The absolute binding free energy for the fifth inhibitor to endothiapepsin was predicted as –9.70 kcal/mol compared with the observed value –9.84 kcal/mol (4). This LIE method was also successfully applied to calculate absolute binding free energies of HIV protease inhibitors and two charged trypsin benzamidine inhibitors (3, 25). In these two studies, an additional correction term for long-range electrostatic contribution to the binding free energy was included. The calculated and observed absolute binding free energies agree well with each other using the same values of α and β. Paulsen & Ornstein, however, found that $\alpha = 0.5$ and $\beta = 1.043$ resulted in a good estimate of the binding free energies of 11 substrates binding to cytochrome P450cam (48). The difference between the two sets of parameters was rationalized as perhaps owing to different force fields, GROMOS and CVFF respectively, used in the two studies (67). Wang *et al.* (70) applied this method to calculate binding free energies of 14 compounds binding to avidin using the Cornell *et al.* force field (11). Their results showed that $\alpha = 0.5$ and $\beta = 1.0$ gave reasonable estimates of the binding free energies with respect to the corresponding experimental results.

These studies raise an interesting question: can one set of α and β be used in different protein-ligand complexes to give reasonable estimates of binding free energies? Although Wang *et al.* used the Cornell *et al.* force field (11), they found values of α and β similar to those of Åqvist *et al.* for the trypsin-benzamidine complex (67). This suggests that the use of different force fields can not explain the difference in α and β found in different simulations. Wang *et al.* (70) further examined this issue in seven different complex systems and found a relationship between the value of β and hydrophobicity of the ligand and the binding site of the receptor; that is, the more hydrophobic groups buried after binding, the more favorable the binding, and the larger the value of β. Different β values were determined for different inhibitors bound to avidin according to this relationship. Calculated absolute binding free energies were improved compared with those from using a fixed β value (70).

Jorgensen and coworkers extended this method further for calculating hydration and binding free energies. They added another term to equation (6), which is proportional to solvent accessible surface area change upon binding. Monte Carlo (MC) simulations were used to obtain the ensemble. The values of these coefficients were calibrated in a test set and then were transferred to predict other ligands bound with the same protein. They succeeded in calculating binding free energies for sulfonamide inhibitors with human thrombin (29) and FKBP12 inhibitors (36).

The LIE method is useful for estimating absolute binding free energies for protein-ligand systems. This method is more computationally efficient than the FEP/TI method.

MM/PBSA

Recent computational advances in parallel computing, force fields and more accurate treatment of electrostatic interactions have enabled multinanosecond MD simulations of highly charged macromolecules such as nucleic acids (8, 12). Analyses of dynamics alone, however, do not sufficiently describe macromolecular recognition and complex formation. Conventional free energy calculations, as described above, have also been applied to protein-nucleic acid complexes. More recently, a hybrid method combining molecular mechanics and continuum solvent calculations has increased in popularity to analyze the free energies of binding and relative free energies of different conformations (32, 61, 63, 71-73).

The method takes solute configurations or snapshots from a MD trajectory with explicit solvent. The solvent molecules are removed to obtain the molecular mechanics energy (E_{MM}) of the solute. This is computed for each snapshot with the same molecular mechanics potential as in the simulation, but with no cut-offs to incorporate all the nonbonded interactions(63). The conformational entropy of the solute, $T\Delta S$, including rotational and vibrational contributions, is estimated from normal mode analyses.

$$\Delta G = E_{MM} - T\Delta S + \Delta G_{solvation} \qquad (7)$$

$$\Delta G_{solvation} = \Delta G_{PB} + \Delta G_{nonpolar} \qquad (8)$$

The free energy of solvation, $\Delta G_{solvation}$, is approximated as the sum of electrostatic and nonpolar contributions. The electrostatic solvation term is calculated with the Poisson-Boltzmann (PB) approach, whereas the nonpolar term as a surface-area (SA) dependent term, hence the name MM-PBSA.

A finite difference solution to the Poisson-Boltzmann equation is calculated using the Delphi II program(59, 60):

$$\nabla\varepsilon(r)\nabla\phi(r) - \kappa'\phi(r) = -4\pi\rho(r) \qquad (9)$$

where $\phi(r)$ is the electrostatic potential, $\varepsilon(r)$ is the dielectric function, $\rho(r)$ is the charge density and κ' is related to the Debye-Huckel inverse length. In the Delphi program, the solute is mapped on to a cubic lattice grid. Values for the electrostatic potential, charge density, dielectric constant, and ionic strength are assigned to each grid point(37). The derivatives of the PB equation are calculated with a finite difference formula and iteratively computed to convergence. The electrostatic component of the solvation free energy is the change of electrostatic energy from transferring the solute from a low dielectric (vacuum) to high dielectric medium using the same grid and solute dielectric.

$$\Delta G_{PB} = \frac{1}{2}\sum_i q_i\left(\phi_i^{80} - \phi_i^{1}\right) \qquad (10)$$

The nonpolar solvation term is approximated as linearly dependent to the solvent accessible surface area:

$$\Delta G_{\text{nonpolar}} = \gamma(SASA) + \beta \qquad (11)$$

where $\gamma = 0.00542$ and $\beta = 0.92$ kcal/mol(63). The surface area is computed with Sanner's MSMS software (58) using a water-sized probe. The MM energies and solvation free energies are computed for each snapshot of the solute and then averaged to compute the difference in free energies. The free energy difference can be computed for absolute binding or the relative binding of different mutants.

Chong *et al.* (9) applied this method to study dianionic hapten binding to a germ line and mature forms of the 48G7 antibody Fab fragments. Reasonable absolute and good relative binding free energies compared with experimental data were obtained.

Their calculations indicated that van der Waals interaction energies and nonpolar contributions to solvation energy were almost identical for both antigens. The >10,000 folds tighter binding of the matured antibody than that of the germ line was due to the gain of more favorable electrostatic interactions over the desolvation penalty through optimizing the binding site geometry. This work sheds light on understanding the process of antibody maturation.

Contrary to the electrostatic discrimination between two antigens, Wang *et al.* (69) observed that in the complexes of Sem-5 SH3 domain and its ligands; van der Waals interactions were primarily responsible for significant binding affinity differences between $C\alpha$- and N- substituted ligands. This shows that binding is dominated by different interactions in different complex systems. In their study, they were also able to identify several critical residues for binding by considering van der Waals energy and conservation of each residue (69).

Another application of this MM/PBSA method to study biotin and its derivatives binding with avidin/streptavidin was carried out by Kuhn & Kollman (35). They were able to reproduce relative binding free energies of 9-methylbiotin compounds with a very good correlation to experimental values. A so-called "computational fluorine scanning" technique, that is, substituting hydrogen by fluorine at different sites of the biotin in a single trajectory obtained from one compound, and then the binding free energy for the substituent was calculated upon the "substituted trajectory", was shown to work well for ranking the nine compounds. This makes free energy calculations more efficient.

Donini *et al.* (13) has used a single trajectory of a ligand binding to a matrix metalloprotease to calculate the binding free energy of five other analog inhibitors. The relative binding free energy of the neutral inhibitors and charged inhibitors were correctly ranked within their series, but the neutral inhibitors were calculated to bind more strongly than the charged inhibitors, relative to experiment.

The precedent of the "computational fluorine scanning" is the "computational Alanine scanning" technique used by Massova & Kollman (41) in their study of p53-MDM2 interactions. Mutating 20 amino acids to alanine in the native trajectory allowed them to quickly compare calculated binding free energy with measured ones. Further examination of W23A mutation by PROFEC led to the conclusion that an additional methyl group on the aromatic rings of W23 might substantially improve binding.

The above work was followed by Huo *et al.* (27) who compared results obtained from the "computational" and from the experimental Alanine scanning on human growth hormone/human growth hormone receptor (HGH/HGHr). Twelve residues were mutated; in all cases but two (R43A and R216A) the calculated $\Delta\Delta G_{bind}$ were in reasonable agreement with the experimental ones. Significant conformational change of mutating Arg43 or Arg216 to Ala caused the overestimation of the $\Delta\Delta G_{bind}$.

Wang & Kollman (68) also applied this MM/PBSA method in studying dimer stability of the HIV protease. They were able to reproduce the relative ranking order of different dimers in agree with experiments. A rapid screening method, which identified cavities on the dimer interface and suggested favorable van der Waals contacts would be created if the cavity was filled with a larger side chain, was exploited to suggest new possible mutations that might enhance binding. Conformational search and minimization were performed for mutation to larger side chains on the dimer interface. Several new stronger associated heterodimers were suggested in their study by this so-called "Virtual Mutagenesis" method.

Recently, Kuhn & Kollman (34) compared MM/PBSA and LIE methods in calculating binding free energies for diverse avidin and streptavidin ligands. Their calculations were able to reproduce experimental ΔG_{bind} with a correlation coefficient of $r^2=0.92$, which was much better that the results obtained from the LIE method with fixed parameters ($\alpha=0.5$ and $\beta=1$) and whose r^2 is 0.55. Although the β value can be adjusted based on hydrophobicity of the binding site (see above), the MM/PBSA method does not introduce any empirical parameters on a protein-by-protein basis.

Summary

In this chapter, we reviewed recent studies of protein-protein, protein-peptide and protein-ligand interactions using different free energy calculation methods. The same methods have also been applied to study hydration of or interactions between other organic molecules, which we have not addressed here.

FEP and TI methods are the most rigorous but require extensive computer resources. With the rapid increase of computer power, we can expect wider application of these methods in the future. Multimolecule free energy calculation methods are promising based on recent studies. They are certainly worth further investigation. The LIE method has a unique advantage because it allows the calculation of absolute binding free energies. With appropriate empirical parameters, this method is useful for studying specific complex systems. Replacing explicit water molecules with a solvent continuum can accelerate MD simulations and enable binding free energies to be calculated directly. Thus, the MM/PBSA is a promising direction for evaluating binding affinities. Combined with other modeling tools, free energy calculation methods will be used in a broader range of research, from evaluating stability of folding structures to the design of new drugs.

REFERENCES

1. Ajay, Murcko MA. 1995. Computational Methods to Predict Binding Free Energy in Ligand-Receptor Complexes. *Journal of Medicinal Chemistry* 38: 4953
2. Allen MP, Tildesley DJ. 1987. *Computer simulation of liquids*. New York: Oxford University Press Inc.
3. Aqvist J. 1996. Calculation of Absolute Binding Free Energies for Charged Ligands and Effects of Long-Range Electrostatic Interactions. *Journal of Computational Chemistry* 17: 1587
4. Aqvist J, Medina C, Samuelsson JE. 1994. New Method for Predicting Binding Affinity in Computer-Aided Drug Design. *Protein Engineering* 7: 385
5. Bash PA, Singh UC, Langridge R, Kollman PA. 1987. Free energy calculations by computer simulation. *Science* 236: 564
6. Bennett CH. 1976. Efficient estimation of free energy differences from Monte Carlo data. *Journal of computational Physics* 22: 245
7. Beveridge DL, Dicapua, F M. 1989. Free energy via molecular simulations: application to chemical and biochemical system. *Annu. Rev. Biophys. Biophys. Chem.* 18: 431
8. Cheatham TE, Brooks BR. 1998. Recent advances in molecular dynamics simulation towards the realistic representation of biomolecules in solution. *Theoretical Chemistry Accounts* 99: 279
9. Chong LT, Duan Y, Wang L, Massova I, Kollman PA. 1999. Molecular dynamics and free-energy calculations applied to affinity maturation in antibody 48G7. *Proceedings of the National Academy of Sciences of the United States of America* 96: 14330
10. Cieplak P, Kollman PA. 1993. Peptide Mimetics as Enzyme Inhibitors - Use of Free Energy Perturbation Calculations to Evaluate Isosteric Replacement for Amide Bonds in a Potent HIV Protease Inhibitor. *Journal of Computer-Aided Molecular Design* 7: 291
11. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, et al. 1995. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* 117: 5179
12. Darden TA, Toukmaji A, Pedersen LG. 1997. Long-range electrostatic effects in biomolecular simulations. *Journal De Chimie Physique Et De Physico-Chimie Biologique* 94: 1346
13. Donini O, Kollman PA. 2000. Calculation and prediction of binding free energies for the matrix metalloproteinases. *Journal of Medicinal Chemistry* 43: 4180
14. Droupadi PR, Varga JM, Linthicum DS. 1994. Mechanism of Allergenic Cross-Reactions .5. Evidence for Participation of Aromatic Residues in the Ligand

Binding Site of Two Multi-Specific Ige Monoclonal Antibodies. *Molecular Immunology* 31: 537

15. Eriksson MAL, Morgantini PY, Kollman PA. 1999. Binding of organic cations to a cyclophane host as studied with molecular dynamics simulations and free energy calculations. *Journal of Physical Chemistry B* 103: 4474

16. Eriksson MAL, Pitera J, Kollman PA. 1999. Prediction of the binding free energies of new TIBO-like HIV-1 reverse transcriptase inhibitors using a combination of PROFEC, PB/SA, CMC/MD, and free energy calculations. *Journal of Medicinal Chemistry* 42: 868

17. Erion MD, Reddy MR. 1995. Calculation of Relative Free Energy Differences for the Covalent Hydration of Organic Compounds - a Combined Quantum Mechanical and Free Energy Perturbation Study. *Journal of Computational Chemistry* 16: 1513

18. Erion MD, Reddy MR. 1998. Calculation of relative hydration free energy differences for heteroaromatic compounds: Use in the design of adenosine deaminase and cytidine deaminase inhibitors. *Journal of the American Chemical Society* 120: 3295

19. Erion MD, van Poelje PD, Reddy MR. 2000. Computer-assisted scanning of ligand interactions: Analysis of the fructose 1,6-bisphosphatase-AMP complex using free energy calculations. *Journal of the American Chemical Society* 122: 6114

20. Essex JW, Severance DL, TiradoRives J, Jorgensen WL. 1997. Monte Carlo simulations for proteins: Binding affinities for trypsin-benzamidine complexes via free-energy perturbations. *Journal of Physical Chemistry B* 101: 9663

21. Ferguson DM, Radmer RJ, Kollman PA. 1991. Determination of the Relative Binding Free Energies of Peptide Inhibitors to the HIV-1 Protease. *Journal of Medicinal Chemistry* 34: 2654

22. Fox T, Scanlan TS, Kollman PA. 1997. Ligand binding in the catalytic antibody 17E8. A free energy perturbation calculation study. *Journal of the American Chemical Society* 119: 11571

23. Guo Z, Brooks CL, Kong X. 1998. Efficient and flexible algorithm for free energy calculations using the lambda-dynamics approach. *Journal of Physical Chemistry B* 102: 2032

24. Guo ZY, Brooks CL. 1998. Rapid screening of binding affinities: Application of the lambda-dynamics method to a trypsin-inhibitor system. *Journal of the American Chemical Society* 120: 1920

25. Hansson T, Aqvist J. 1995. Estimation of Binding Free Energies for HIV Proteinase Inhibitors by Molecular Dynamics Simulations. *Protein Engineering* 8: 1137

26. Helms V, Wade RC. 1998. Computational alchemy to calculate absolute protein-ligand binding free energy. *Journal of the American Chemical Society* 120: 2710

27. Huo S, Massova I, Kollman PA. 2000.: manuscript in preparation

28. Jones TR, Varney MD, Webber SE, Lewis KK, Marzoni GP, et al. 1996. Structure-Based Design of Lipophilic Quinazoline Inhibitors of Thymidylate Synthase. *Journal of Medicinal Chemistry* 39: 904

29. JonesHertzog DK, Jorgensen WL. 1997. Binding affinities for sulfonamide inhibitors with human thrombin using Monte Carlo simulations with a linear response method. *Journal of Medicinal Chemistry* 40: 1539

30. Jorgensen WL. 1989. Free-energy calculations: a breakthrough for modeling organic chemistry in solution. *Accounts of Chemical Research* 22: 184

31. Kollman P. 1993. Free Energy Calculations - Applications to Chemical and Biochemical Phenomena. *Chemical Reviews* 93: 2395

32. Kollman PA, Massova I, Reyes CM, Kuhn B, Huo S, et al. 2000. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accounts of Chemical Research* in press

33. Kong XJ, Brooks CL. 1996. Lambda-Dynamics - a New Approach to Free Energy Calculations. *Journal of Chemical Physics* 105: 2414

34. Kuhn B, Kollman PA. 2000. Binding of a diverse set of ligands to avidin and streptavidin: An accurate quantitative prediction of their relative affinities by a combination of molecular mechanics and continuum solvent models. *Journal of Medicinal Chemistry* 43: 3786

35. Kuhn B, Kollman PA. 2000. A ligand that is predicted to bind better to avidin than biotin: Insights from computational fluorine scanning. *Journal of the American Chemical Society* 122: 3909

36. Lamb ML, Tirado-Rives J, Jorgensen WL. 1999. Estimation of the binding affinities of FKBP12 inhibitors using a linear response method. *Bioorganic & Medicinal Chemistry* 7: 851

37. Leach AR. 1996. *Molecular modeling: Principles and Applications*: Addison Wesley Longman Limited

38. Lee TS, Kollman PA. 2000. Theoretical studies suggest a new antifolate as a more potent inhibitor of thymidylate synthase. *Journal of the American Chemical Society* 122: 4385

39. Liang G, Schmidt RK, Yu HA, Cumming DA, Brady JW. 1996. Free Energy Simulation Studies of the Binding Specificity of Mannose-Binding Protein. *Journal of Physical Chemistry* 100: 2528

40. Liu HY, Mark AE, Vangunsteren WF. 1996. Estimating the Relative Free Energy of Different Molecular States with Respect to a Single Reference State. *Journal of Physical Chemistry* 100: 9485

41. Massova I, Kollman PA. 1999. Computational alanine scanning to probe protein-protein interactions: A novel approach to evaluate binding free energies. *Journal of the American Chemical Society* 121: 8133

42. McCammon JA. 1991. Free energy from simulations. *Current Opinion in Structural Biology* 1: 196

43. McCarrick MA, Kollman PA. 1999. Predicting relative binding affinities of non-peptide HIV protease inhibitors with free energy perturbation calculations. *Journal of Computer-Aided Molecular Design* 13: 109

44. Miyamoto S, Kollman PA. 1993. What Determines the Strength of Noncovalent Association of Ligands to Proteins in Aqueous Solution. *Proceedings of the National Academy of Sciences of the United States of America* 90: 8402

45. Ota N, Brunger AT. 1997. Overcoming barriers in macromolecular simulations: non-Boltzmann thermodynamic integration. *Theoretical Chemistry Accounts* 98: 171

46. Ota N, Stroupe C, Ferreira-da-Silva JMS, Shah SA, Mares-Guia M, Brunger AT. 1999. Non-Boltzmann thermodynamic integration (NBTI) for macromolecular systems: Relative free energy of binding of trypsin to benzamidine and benzylamine. *Proteins-Structure Function and Genetics* 37: 641

47. Pathiaseril A, Woods RJ. 2000. Relative energies of binding for antibody-carbohydrate-antigen complexes computed from free-energy simulations. *Journal of the American Chemical Society* 122: 331

48. Paulsen MD, Ornstein RL. 1996. Binding Free Energy Calculations for P450cam-Substrate Complexes. *Protein Engineering* 9: 567

49. Pearlman DA. 1999. Free energy grids: A practical qualitative application of free energy perturbation to ligand design using the OWFEG method. *Journal of Medicinal Chemistry* 42: 4313

50. Pitera J, Kollman P. 1998. Designing an optimum guest for a host using multimolecule free energy calculations: Predicting the best ligand for Rebek's "tennis ball". *Journal of the American Chemical Society* 120: 7557

51. Radmer RJ, Kollman PA. 1998. The application of three approximate free energy calculations methods to structure based ligand design: Trypsin and its complex with inhibitors. *Journal of Computer-Aided Molecular Design* 12: 215

52. Rao BG, Murcko MA. 1994. Reversed Stereochemical Preference in Binding of Ro 31-8959 to HIV-1 Proteinase - a Free Energy Perturbation Analysis. *Journal of Computational Chemistry* 15: 1241

53. Rao BG, Murcko MA. 1996. Free Energy Perturbation Studies on Binding of a-74704 and Its Diester Analog to HIV-1 Protease. *Protein Engineering* 9: 767

54. Rao BG, Tilton RF, Singh UC. 1992. Free Energy Perturbation Studies on Inhibitor Binding to HIV-1 Proteinase. *Journal of the American Chemical Society* 114: 4447

55. Rastelli G, Costantino L, Vianello P, Barlocco D. 1998. Free energy perturbation studies on binding of the inhibitor 5,6-dihydrobenzo[h]cinnolin-3(2H)one-2-acetic acid and its methoxylated analogs to aldose reductase. *Tetrahedron* 54: 9415

56. Reddy MR, Viswanadhan VN, Weinstein JN. 1991. Relative Differences in the Binding Free Energies of Human Immunodeficiency Virus-1 Protease Inhibitors - a Thermodynamic Cycle-Perturbation Approach. *Proceedings of the National Academy of Sciences of the United States of America* 88: 10287

57. Rick SW, Topol IA, Erickson JW, Burt SK. 1998. Molecular mechanisms of resistance: Free energy calculations of mutation effects on inhibitor binding to HIV-1 protease. *Protein Science* 7: 1750

58. Sanner MF, Olson AJ, Spehner JC. 1996. Reduced Surface - an Efficient Way to Compute Molecular Surfaces. *Biopolymers* 38: 305

59. Sharp KA, Honig B. 1990. Calculating Total Electrostatic Energies with the Nonlinear Poisson-Boltzmann Equation. *Journal of Physical Chemistry* 94: 7684

60. Sharp KA, Honig B. 1990. Electrostatic Interactions in Macromolecules - Theory and Applications. *Annual Review of Biophysics and Biophysical Chemistry* 19: 301

61. Smith KC, Honig B. 1994. Evaluation of the Conformational Free Energies of Loops in Proteins. *Proteins-Structure Function and Genetics* 18: 119

62. Sotriffer CA, Flader W, Cooper A, Rode BM, Linthicum DS, et al. 1999. Ligand binding by antibody IgE lb4: Assessment of binding site preferences using microcalorimetry, docking, and free energy simulations. *Biophysical Journal* 76: 2966

63. Srinivasan J, Cheatham TE, Cieplak P, Kollman PA, Case DA. 1998. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate - DNA helices. *Journal of the American Chemical Society* 120: 9401

64. Tidor B. 1993. Simulated Annealing on Free Energy Surfaces by a Combined Molecular Dynamics and Monte-Carlo Approach. *Journal of Physical Chemistry* 97: 1069

65. Tropsha A, Hermans J. 1992. Application of Free Energy Simulations to the Binding of a Transition-State-Analogue Inhibitor to HIV Protease. *Protein Engineering* 5: 29

66. van Gunsteren WF. 1989. Methods for calculation of free energies and binding constants: successes and problems. In *Computer Simulation of Biomolecular Systems*, ed. WF van Gunsteren, PK Weiner, pp. 27. Leiden: ESCOM

67. Wang J, Dixon R, Kollman PA. 1999. Ranking ligand binding affinities with avidin: A molecular dynamics-based interaction energy study. *Proteins-Structure Function and Genetics* 34: 69

68. Wang W, Kollman PA. 2000. Free energy calculations on dimer stability of the HIV protease using molecular dynamics and continuum solvent model. *J. Mol. Bol.* 303: 567

69. Wang W, Lim WA, Jakalian A, Wang J, Wang JM, et al. 2000. An analysis of the interactions between the Sem-5 SH3 domain and its ligands using molecular dynamics, free energy calculations and sequence analysis. *Journal of the American Chemical Society* in press

70. Wang W, Wang J, Kollman PA. 1999. What determines the van der Waals coefficient beta in the LIE (linear interaction energy) method to estimate binding free energies using molecular dynamics simulations? *Proteins-Structure Function and Genetics* 34: 395

71. Yang AS, Hitz B, Honig B. 1996. Free Energy Determinants of Secondary Structure Formation .3. Beta-Turns and Their Role in Protein Folding. *Journal of Molecular Biology* 259: 873

72. Yang AS, Honig B. 1995. Free Energy Determinants of Secondary Structure Formation .1. Alpha-Helices. *Journal of Molecular Biology* 252: 351

73. Yang AS, Honig B. 1995. Free Energy Determinants of Secondary Structure Formation .2. Antiparallel Beta-Sheets. *Journal of Molecular Biology* 252: 366

74. Zacharias M, Straatsma TP, McCammon JA, Quiocho FA. 1993. Inversion of Receptor Binding Preferences by Mutagenesis - Free Energy Thermodynamic Integration Studies on Sugar Binding to L-Arabinose Binding Proteins. *Biochemistry* 32: 7428

Chapter 3. Free energy calculations on dimer stability of the HIV protease using molecular dynamics and continuum solvent model

This chapter is a reprint of a published paper (Wang, W. and Kollman, P. A., Journal of Molecular Biology, 2000, 303, 4, 567-582).

Free energy calculations on dimer stability of the HIV protease using molecular dynamics and continuum solvent model

Wei Wang

Graduate Group in Biophysics

University of California, San Francisco

San Francisco, CA 94143

and

Peter A. Kollman*

Department of Pharmaceutical Chemistry

University of California, San Francisco

San Francisco, CA 94143

* Author for correspondence

Tel: (415) 476-4637 (PAK)

Fax: (415) 502-1411

Email: pak@cgl.ucsf.edu

Short title: Dimer stability of the HIV protease

Keywords: HIV protease; dimer stability; molecular dynamics; continuum model; MM/PBSA; Virtual Mutagenesis method.

Grant sponsor: NIH; Grant number: GM-56531 (P. Ortiz de Montellano, P.I.), GM-56609 (E. Arnold, P.I.).

Abstract

Dimerization of HIV-1 protease (HIV PR) monomers is an essential prerequisite for viral proteolytic activity and the subsequent generation of infectious virus particles. Disrupting dimerization of the enzyme can inhibit its activity. We have calculated the relative binding free energies between different dimers of the HIV protease using molecular dynamics (MD) and a continuum model, which we call MM/PBSA. We examined the dominant negative inhibition of the HIV PR by a mutated form of the protease and found relative dimerization free energies of homo- and hetero-dimerization consistent with experimental data. We also developed a rapid screening method, which was called the Virtual Mutagenesis (VM) method to consider other mutations which might stabilize non-wild type heterodimers. Using this approach, we considered the mutations near the dimer interface which might cause dominant negative inhibition of the HIV PR. The rapid method we developed can be used in studying any ligand-protein and protein-protein interactions, in order to identify mutations which can enhance the binding affinities of the complex.

1. Introduction

Molecular dynamics (MD) has provided dynamic and atomic insights to understand complicated biological systems. Free energy calculation methods have become powerful tools to provide quantitative measurement of protein-ligand or protein-protein interactions (Kollman, 1993; Beveridge & Dicapua, 1989; van Gunsteren, 1989). The most rigorous approaches to evaluate binding free energies are free energy perturbation (FEP) and thermodynamic integration (TI) methods. Both MD and FEP/TI have been successfully applied to study many protein and nucleic acids systems. Nonetheless, these methods are computationally intensive. Thus, many semi-empirical methods have been developed to estimate binding free energies faster with reasonable accuracy (Ajay & Murcko, 1995).

A new method, MM/PBSA, was proposed last year for evaluating solvation and binding free energies of macromolecules and their complexes (Srinivason et al., 1998). When this method is used to calculate binding free energy, the binding free energy is decomposed into contributions from van der Waals and electrostatic energies, non-polar and electrostatic solvation free energies, and relative solute entropy effects (Massova & Kollman, 1999). The van der Waals and electrostatic interactions between the components of the complex are calculated using molecular mechanics (MM) with an empirical force field (Cornell et al., 1995), the non-polar part of solvation free energy is estimated by empirical methods based on solvent accessible (SA) surface and the electrostatic contribution to solvation is calculated using a continuum model and solving the Poisson-Boltzmann (PB) equation. The entropy contribution has been estimated using normal mode analysis. An ensemble of different conformations is extracted from MD trajectories and each snapshot is analyzed using this MM/PBSA method. The binding free energies are obtained from this ensemble average. This method is able to calculate free energy difference between any two states, even when the two states are quite dissimilar from each other. It is also significantly more computationally efficient than FEP or TI. In the present study, we used this MM/PBSA method in studying dimer stability of the HIV protease (PR).

The dimeric HIV-1 protease (HIV-1 PR) is crucial for the maturation of viral structural (gag) and enzymatic (pol) proteins of the AIDS virion. (Debouck et al., 1987). This aspartyl protease has been the therapeutic target for the treatment of AIDS. However, the HIV-1 virus rapidly develops drug resistant variants. Therefore, it is critically important to understand the mechanism of the HIV PR for designing inhibitors to combat this resistance. The primary structure of the HIV-1 protease indicates that each monomer of the protease contributes one catalytic aspartic acid residue at the active site of the enzyme. Either mutating one of the two catalytic aspartic acids (Kohl et al., 1988, Babe et al., 1991, Krasslich, 1991) or disrupting the dimerization of active HIV PR monomers (Zhang et al., 1991; McPhee, et al., 1996; Rozzelle, et al., 1999) has been reported to eliminate the catalytic activity of the protease and thus block the infectivity of the virus.

Craik and coworkers have shown that mixing of wild type (wt) and certain mutant protease monomers could lead to inactivation of HIV-1 virus. They concluded this upon monitoring accumulation of unprocessed polyproteins and the secretion of noninfectious virons, and inferred that this loss of activity of the protease was due to the formation of inactive heterodimers between wild type and defective monomers (Babe et al., 1991; McPhee, et al., 1996). The defective monomers used in their experiments were obtained by mutating the aspartic acid in the catalytic triad and several other residues of the PR flap region. The goal was to promote the formation of defective heterodimers and decrease the stability of the wild type and mutant PR homodimers. In their studies, they found a triple-mutation, Asp25Lys, Gly49Trp and Ile50Trp, which significantly reduced the levels of PR activity and virus infectivity (McPhee, et al., 1996). Due to the large interface between two PR monomers, this dominant-negative inhibition of the HIV PR by defective monomers may be less susceptible to the emergence of resistant mutations. It suggests a potential use of gene therapy as a treatment to AIDS.

In the present study, we examined the protonation state of the ligand free HIV protease and estimated relative binding free energies between wild type homodimer and defective heterodimer or mutant homodimer using the MM/PBSA method. Since it is not trivial to measure the binding affinities of different dimers experimentally, computer simulations can provide useful insights to understand the interactions between HIV PR subunits.

We also present here a new method, which we call the Virtual Mutagenesis (VM) method, to identify mutations on the interface of two molecules which may enhance the binding between them. We applied this method to HIV PR dimer and identified a few more potential dominant negative mutations. This Virtual Mutagenesis (VM) method is applicable to any set of interacting molecules.

2. Results and Discussion

(1) Protonation state of the ligand-free HIV-PR is dianionic

Several distances between pairs of atoms in the two catalytic aspartic acid residues were measured in the crystal structure and the 100 snapshots taken from the MD trajectory (see Figure 1 and Table 1). Since the two catalytic Asp's are critical for proper function of the HIV PR, it is important to maintain their structures during the simulation. From Table 1 and Figure 1, the distances between the two CA atoms are close to that in the crystal structure in all three protonation states. However, the distances between heavy atoms on the side chains in mono- and double- protonated states are much smaller than those in the doubly ionic state. While the distances in the doubly ionic state are maintained closest to the distances in the crystal structure, the side chains of the two Asp's came closer to each other in the mono- and double- protonated states. The reason is obvious: the anionic Asp's electrostatically repel each other. If either one of the two catalytic Asp's is protonated, the repulsion between them is greatly reduced and hydrogen bonds can also form between them directly or via nearby water molecules.

It is widely assumed that free HIV PR has a mono-protonated state and a water molecule is presumed to interact with the two catalytic Asp's in the way shown in Figure 1 (Ido et al., 1991). We investigated this assumption. One water molecule was put in the active site at the beginning of the MD simulations. The water molecule moved away in a few picoseconds of MD runs. But, other nearby water molecules moved into the active site and formed hydrogen bonds with the catalytic Asp's. It is assumed that a water molecule is crucial for the proteolystic reaction. Based on our simulations, this water molecule should be quite labile rather than fixed in the active site.

We calculated the binding free energies between the wild type HIV PR dimers with different protonation states. The binding free energies and components for different dimers are shown in Table 2. In Table 2, the dianionic state has the most favorable binding free energy. Thus, the results of the binding free energy calculations and the distance measurement of pairs of atoms in the two catalytic Asp's are consistent with the NMR data (Smith et al., 1996) which suggests that both Asp's in the active site are deprotonated. It is worth of pointing out that many experimental and theoretical works have been done to study the protonation states of the two catalytic Asp's in the presence of HIV PR inhibitors (Yamazaki, et al, 1994; Wang et al., 1996; Luo et al., 1998; Trylska et al., 1999). The present and previous studies suggest that the binding of inhibitors has significant influence on the ionic states of the HIV PR.

As mentioned in the Methods section, the effect of conformational entropy upon dimerization is not included in equations (1)-(3). The absolute values of the binding free energies thus will overestimate the strength of binding. We estimated the conformational entropy using normal mode analysis. Due to the heavy computational demand of this analysis, we only carried out a single calculation as to estimate the order of magnitude of the conformational entropy contribution to the binding free energy. In our calculation, the conformational entropy is +71.1 kcal/mol. If this value was included in our calculations, the values of binding free energies of the HIV dimer would fall into the range of -9 to -13 kcal/mol. The binding free energy measured experimentally varies with experimental conditions, such as pH (Zhang et al., 1991; Cheng et al., 1990; Grant et al., 1992; Jordan, et al., 1992). At pH7, the K_d was measured as 50 nM (Cheng et al., 1990) which corresponds to a binding free energy –10.0 kcal/mol at 298K. The order of magnitude of our results is consistent with the experimental data. It is also worth pointing out that we assume that the two monomers of the HIV PR are already fully folded before forming the dimer. Because we are interested in calculating relative binding free energies between different dimers, the free energies of folding the two monomers are likely to cancel out.

If we examine each component of the binding free energy in Table 2, we can see that the order of binding free energies is the same as the order of the van der Waals interaction energies. In another word, van der Waals interactions are dominant in the HIV PR dimer binding. Nonpolar solvation terms are similar in all protonation states, which is not unexpected. Electrostatic interaction energies, ΔG_{int}^{ele}, and difference of electrostatic contribution to the solvation energy term, ΔG_{sol}^{ele}, are quite different in the three protonation states. However, the sums of these two terms, $\Delta G_{int+sol}^{ele}$, are quite similar, suggesting that the two terms compensate each other. It helps to rationalize why considering solvation energy can greatly improve ranking ligands in drug design (Zou et al., 1999).

Using value 1 underestimated interior dielectric constant of proteins. Thus, we also calculated binding free energies using a dielectric constant of 2 (see Table 2). It is encouraging that the ranking order is same as that obtained using dielectric constant 1. Therefore, we only use value 1 to calculate binding free energies for other dimers in the rest of this paper. We noticed that the absolute values of the binding free energies are too negative using dielectric constant 2. This is probably due to the fact that the parameterization of the force field we used was carried out with dielectric constant 1. We can see in Table 2 that the sums of the electrostatic interaction energy and the electrostatic contribution to solvation in three protonation states are +54.2, +53.6, and +51.4 kcal/mol respectively for dielectric constant 2, and +115.3, +114.5, and +109.7

kcal/mol respectively for dielectric constant 1. Using dielectric constant other than 1 reduced the influence of the overall electrostatic contribution to the binding free energy.

(2) MM/PBSA can differentiate stabilities of different HIV PR dimers

In Craik and coworkers' study, they found certain mutants which could inhibit the infectivity of AIDS virons (Babe et al., 1991; McPhee, et al., 1996). Among these mutants, one triple mutation, Asp25Lys, Gly49Trp, Ile50Trp, had the most significant effect. Asp25Lys mutation caused the HIV PR loss of proteolytic activity. Gly49 and Ile50 are in the flap region of the HIV PR (see Figure 3). They were mutated to two Trp's. Trp has larger side chain group than either Gly or Ile. The purpose is to enhance formation of dimer between wild type and mutant monomers but prevent dimerization between the mutant monomers. It is also known that the flexibility of the "flap" region is crucial for the activity of the protease (Ishima, et al., 1999). The residues 1-27 and 60-99 in each monomer are defined as the "core" region and 28-59 as the "flap" region (Collins et al., 1995). Two Trp's in the "flap" region could reduce the flexibility and thus reduced the activity.

The average structure for each dimer during the 120 ps data collection period in MD simulations was calculated. The MD trajectory was superimposed with the average structure and the RMSD of heavy atoms on the backbone was calculated (Figure 4 and Table 3). The "flap" region of the wild type homodimer, WTP-WTP, is most flexible. This is shown by the ratio between the deviation of the RMSD, σ, and the RMSD in Table 3. This ratio is 0.223 for WTP-WTP and about 0.140 for other dimers. In Figure 4a, the fluctuation of the WTP-WTP is also obviously larger than others. For comparison, the RMSD, its deviation σ and the σ/RMSD ratio of the "core" regions are also listed in

Table 3 and the RMSD's are plotted in Figure 4b. Although the "core" region of the WTP-WTP dimer still has the largest σ/RMSD ratio, the difference between different dimers is not as large as the "flap" region. Thus, the triple mutations mainly influence the flexibility of the flap region of the HIV PR.

Binding free energies for different dimers are shown in Table 4. Entropy terms are not included. As mentioned above, it is assumed that entropy terms are similar for the different dimers. The triple mutant KWW monomer bound most tightly to the wild type (wt) monomer whose catalytic Asp is ionic. This dimer, WTP-KWW, is much more favorable than all other dimers. The mutant KWW homodimer, KWW-KWW, is most unfavorable, even worse than the wild type homodimer, WTP-WTP. The wild type monomer with protonated catalytic Asp binds to the mutant KWW monomer with an intermediate binding free energy (WTH-KWW), however, which is still more negative than that of the wild type homodimer. The ranking order of the different dimers is consistent with experimental data. The heterodimer formed between wild type and KWW mutant monomers is observed to have a higher melting temperature than the wild type dimer (Rozzelle et al., 1999). The defective homodimer, KWW-KWW, was not obtained in that experiment due to aggregation.

We can see that in Table 4 the wild type homodimer, WTP-WTP, has the least favorable van der Waals interaction energy compared with the other dimers. This is because the triple mutant KWW has two Trp's in the "flap" region and they provide stronger van der Waals interactions between the monomers. The total electrostatic contribution to the binding free energy, $\Delta G_{int+sol}^{ele}$, of the heterodimer with ionic Asp in the active site, WTP-KWW, is most favorable. The wild type homodimer WTP-WTP, and the defective heterodimers WTP-KWW and WTH-KWW, have similar $\Delta G_{int+sol}^{ele}$. The values are +115.3, +114.8 and +117.3 kcal/mol respectively. The nonpolar part of solvation free energy is also similar for different dimers. For WTP-WTP, WTP-KWW and WTH-KWW, the van der Waals interaction energy determines the rank order of the binding free energy. For the mutant homodimer, KWW-KWW, however, although the van der Waals is much more favorable, the $\Delta G_{int+sol}^{ele}$ term is over 20 kcal/mol less favorable compared with the other three dimers. The KWW-KWW has less favorable electrostatic energy and more unfavorable electrostatic solvation energy than the wild type homodimer. The less favorable electrostatic energy is due to the stronger repulsive interaction between the two Lys's in KWW-KWW than that between the two Asp's in WTP-WTP. The average distance between the two NZ atoms in the two Lys's in KWW-KWW is 4.61±0.31Å compared with 5.01±0.23Å between the two CG atoms in the two Asp's in WTP-WTP. The Lys in one monomer also repulsively interacts with Trp49 and Trp50 in another monomer in KWW-KWW, but there are no such repulsive interactions in WTP-WTP in which Gly49 and Ile50 are further away from the catalytic Asp in another monomer. If we examine the structure of KWW-KWW, we can see that the aromatic rings of Trp49A and Trp49B are partially buried by Trp50A and Trp50B respectively. This burial of polar groups gives a larger solvation penalty to the KWW-KWW dimerization than the WTP-WTP dimerization. This explains why the KWW-KWW dimer has more unfavorable electrostatic solvation energy than the WTP-WTP dimer.

(3) The Virtual Mutagenesis (VM) method can predict several new dominant negative mutants

Encouraged by the above results, we tried a simpler but faster approach, which we named Virtual Mutagenesis (VM) method, to estimate the relative binding free energies between different dimers. We took one snapshot that has the closest binding free energy to the average binding free energy value obtained from the MD trajectory. We can see from Table 5 that the binding free energy of the snapshot we chose is -84.5 kcal/mol, which is very close to the average value –84.3 kcal/mol calculated from the MD trajectory (Table 3). Mutations were suggested by a fast screening procedure (see below) and then made on this snapshot. For each mutation, a systematic conformation search for total 100 conformations was performed. Only those conformations with no steric clash with other atoms in the molecule were further investigated (see Methods section). Each surviving conformation was minimized with a distance dependent dielectric constant while all other residues in the molecule were fixed. The binding free energy was then calculated using MM/PBSA. The final binding free energy for each mutation is the average value for all rotamers.

We first applied this Virtual Mutagenesis (VM) method to several mutations for which experimental data were available. The screening procedure is not necessary here. No binding free energy or disassociate equilibrium constant K_d has been measured on any of those mutant dimers. However, there is experimental evidence indicating that thermal denaturation of single chain heterodimers, D25K, G49W/I50W and D25K/G49W/I50W, have a 1.5°C to 7.2°C higher thermal stability than single chain wild type HIV PR (Rozzelle et al., 1999). The accumulation of unprocessed polyproteins and the secretion of noninfectious virons display the same trend as the thermal stability. Thus, we assume that the binding affinities between different mutant and wild type monomer are in the same order as viral infectivity. With this assumption in mind, we found that the calculated binding free energies are consistent with the experimental data (Table 5).

Let us further examine some of these mutations. For the 49W mutation, it has a –1.8 kcal/mol more favorable van der Waals interaction than wild type homodimer because the Trp has a much larger side chain than Gly. The nonpolar solvation free energy difference, ΔG_{sol}^{nonpol} is –0.1 kcal/mol more favorable for the mutant heterodimer. However, the total electrostatic contribution, $\Delta G_{int+sol}^{ele}$, is +1.4 kcal/mol less favorable for the mutant heterodimer. In net, the total binding free energy is –0.5 kcal/mol more favorable for mutant heterodimer.

The 49W50W mutant is similar to 49W. The 49W50W heterodimer has larger favorable van der Waals interaction energy, -193.2 kcal/mol, compared with –191.6 kcal/mol for 49W heterodimer and –189.8 kcal/mol for wild type homodimer. Ile50 mutated to Trp provides –1.6 kcal/mol van der Waals interaction energy to binding free energy versus –1.8 kcal/mol while Gly49 is mutated to Trp. This is not unexpected because Ile has a larger side chain than Gly so that the mutation from Ile to Trp has smaller effect than that of Gly to Trp. The nonpolar solvation energy has a small but favorable contribution to the stability of the 49W50W heterodimer. The total electrostatic contribution, $\Delta G_{int+sol}^{ele}$, is unfavorable compared with 49W and wild type dimer and it cancels part of the favorable van der Waals interactions. The unfavorable $\Delta G_{int+sol}^{ele}$ terms in 49W and 49W50W are due to the unfavorable electrostatic interactions between Asp25 and 49W/50W. The aromatic ring of Trp49 is partially buried by Trp50 in the 49W50W heterodimer. This helps to explain why 49W50W has a more unfavorable $\Delta G_{int+sol}^{ele}$ term than 49W. The binding free energy of 49W50W is –0.8 kcal/mol more favorable than wild type homodimer and –0.3 kcal/mol than the 49W heterodimer. Obviously, favorable van der Waals interaction is dominant in the 49W50W heterodimer.

The total electrostatic contribution term, $\Delta G_{int+sol}^{ele}$, is not unfavorable in the 25K49W50W heterodimer case. Instead, it is –2.0 kcal/mol favorable than wild type homodimer. This is due to favorable electrostatic interactions between Lys25 in the mutant monomer and Asp25 in the wild type monomer. The van der Waals and nonpolar solvation terms have similar values as in the 49W50W heterodimer. This suggests that the mutation of Asp25 to a positive charged residue with stronger binding than the wild type homodimer is mainly due to favorable electrostatic interaction.

The above conclusion is consistent with the D25K data where the van der Waals and nonpolar solvation only are –0.3 kcal/mol and –0.1 kcal/mol more favorable than wild type homodimer but $\Delta G_{int+sol}^{ele}$ term contributes –1.0 kcal/mol. However, in the D25R heterodimer, the $\Delta G_{int+sol}^{ele}$ term is unfavorable relative to the wild type dimer and it is the van der Waals and nonpolar solvation energies that make the total binding free energy of the D25R heterodimer –1.6 kcal/mol more favorable than the wild type homodimer. This is interesting. It is worthy pointing out that the calculated binding free energies of the D25R and the D25K heterodimers have larger error bars. This is because Asp25 is in the active site and has empty space around it. Many conformations can be considered for the mutant residues and this causes the relatively larger variation. Comparing D25R and D25K, we can see D25R has more favorable van der Waals interaction than D25K. This is reasonable because the Arg side chain is larger than the Lys. The unfavorable $\Delta G_{int+sol}^{ele}$ term is unexpected. Compared with D25K, D25R has similar electrostatic interaction energy, -401.8 kcal/mol versus –401.3 kcal/mol, but more unfavorable PB solvation energy, +525.1 kcal/mol versus +523.0 kcal/mol. This unfavorable $\Delta G_{int+sol}^{ele}$ term may be due to the additional burial of NH1 or NH2 polar groups in the Arg upon dimerization.

With the above encouraging results, we can try to predict some new mutations which may enhance binding of defective heterodimers. Since the interface between the HIV PR monomers is large and the interface between the two monomers is well packed, it is difficult to determine which residue can be mutated if one just visualizes the structure of the HIV PR. We exploited the simple method which is described in the Methods session to scan all possible residues which are close to interface but still have enough surrounding space to allow larger side chain replacement. Each residue which satisfies the scanning criteria is evaluated by the Virtual Mutagenesis method.

In Table 6, Contact Neighbor Atom Number (CNAN) and Total Neighbor Atom Number (TNAN) of all residues whose Cα-Cβ vector points to the interface are listed. CNAN counts how many contacts one residue has with another monomer and TNAN indicates how crowded a given residue is. In the present study, we used two distance cutoffs, 3Å and 6Å. The Shell Contact Neighbor Atom Number Ratio (SCNANR) was calculated for each residue. SCNANR shows how many contacts one residue can make with another monomer between a 3Å and 6Å shell around it. In order to find mutation to enhance binding, one wants to identify residues which have small TNAN with a 3Å distance cutoff and a large SCNANR. A small TNAN in the 3Å distance cutoff means the residue has empty surrounding space so that larger side chain replacement is possible. A large SCNANR shows that a larger side chain has the potential to have more contacts or stronger van der Waals interactions with another monomer.

In this study, we chose 20 as the threshold for TNAN using the 3 Å distance cutoff and 20% as the cutoff for SCNANR. The Virtual Mutagenesis calculations were only performed on those residues whose TNAN using 3Å distance cutoff was less than 20 and for which SCNANR was larger than 20%. Eleven residues which satisfy these two criteria are printed as bold in Table 6. Five of them, Gln2, Thr4, Trp6, Gln7 and Gly94 are in the "core" region. Gln2 is too close to the N-terminal of the chain. Therefore no mutation was made. Trp6 was not mutated because no natural amino acid with a larger side chain exists. Gly27 is in the catalytic triad. The remaining five, Gly48, Gly49, Ile50, Gly51 and Gly52, are in the flap region. Gly27 has ϕ and φ angles in the right side of the Ramachandran map. Thus, we did not mutate Gly27 either.

Since we are interested in finding new mutant monomers which can inhibit virial infectivity, we first mutated Asp25 to Lys. This D25K mutation can also reduce binding affinities between defective homodimers (see Table 4). Residues identified by our scanning method were then mutated as well. The binding free energies calculated by the Virtual Mutagenesis method are listed in Table 7. Mutations with more favorable binding free energies than D25K are printed in bold and those with binding free energies between the wild type homodimer and the D25K heterodimer are printed in italics.

The most interesting mutations are those in the "flap" region. As we discussed above, compared with mutations in other regions, these mutations can reduce the flexibility of the "flap" region and, thus, can further reduce the activity of the HIV PR. Since these residues are exposed to water, we mutated them to Trp so that they can have stronger van der Waals interactions with another monomer but do not get dramatic unfavorable solvation free energy penalties. 25K48W, 25K49W, 25K50W and 25K52W have more favorable binding free energies, even compared with the D25K heterodimer. Among these four mutations, 25K48W and 25K50W have the most and least favorable van der Waals interaction energy respectively. The reason is that the mutated residue, Trp, can pack well with Ile50 from another chain in the 25K48W case. This packing also explains the most favorable nonpolar solvation energy for 25K48W because the solvent accessible surface of Ile50 from another chain is reduced. This deeper burial of the hydrophobic residue favors binding. For 25K50W, the Ile is much larger than Gly in the wild type protease. Thus, the mutation to Trp from Ile does not create as many more van der Waals contacts than the wild type dimer as the mutation from Gly to Trp. The 25K49W and the 25K52W have intermediate van der Waals interaction energies, as one expects. In terms of total electrostatic contribution to binding free energy, $\Delta G_{int+sol}^{ele}$, the 25K50W has the least unfavorable value. This is also not unexpected because the hydrophobic residue Ile50 is exposed to water. If it is replaced by Trp, Trp has large aromatic ring and, therefore, a more favorable solvation energy. However, the 25K51W has a more unfavorable binding free energy than the wild type homodimer. We can see that this is due to an very unfavorable van der Waals interaction energy. It implies that there are steric clashes. Thus, we mutated Gly51 to Ala instead of Trp. The van der Waals interaction becomes more favorable than the 25K51W but still unfavorable if compared with the wild type homodimer. We examined the structure and found that this is due to the fact that Gly51 is flanked by Ile50 in the same chain and Phe53 in another chain. The $C\beta$ atom in the substituted Ala or Trp has unfavorable contacts with these two residues.

Three other mutations, 25K4Y, 25K7W and 25K94W are in the "core" region of the HIV PR. 25K7W has the least unfavorable $\Delta G_{int+sol}^{ele}$ term. Since Gln7 is on the surface of the HIV PR, mutation to Trp can provide more favorable solvation energy. So can the 25K4Y, which also has a less unfavorable $\Delta G_{int+sol}^{ele}$ term than the 25K mutation. After Thr4 and Gln7 are mutated to Tyr and Trp respectively, Tyr4 and Trp7 can reduce solvent accessible surface of some nearby hydrophobic residues, such as Leu10, Leu5 and Ile31. This deeper burial of hydrophobic residues is also favorable for binding. 25K4Y and 25K7W have more favorable van der Waals interaction energies than wild type as well, which is due to larger side chain replacement. However, the main contributions to the binding are from the $\Delta G_{int+sol}^{ele}$ term. The 25K94W heterodimer has the most favorable van der Waals interaction in these three dimers. This is because when Gly94 is mutated to Trp, the Trp can pack well with Trp6 from another chain. It is the van der Waals interaction that leads to the total binding free energy of 25K94W being more favorable than the wild type dimer.

In addition to those eleven residues which were identified by our scanning method, we also did calculations on other residues.

Craik and coworkers proposed that the L23Y mutation might enhance the binding for the defective heterodimer. On the basis of analyzing the structure of the HIV PR, the L23Y mutation may form new hydrogen bonds and therefore enhance binding (McPhee et al., 1996). We did mutations for L23Y alone and in combination with D25K and L23Y, i.e. 25K23Y. We can see in both cases, they do have more favorable electrostatic interaction energies, -394.8 kcal/mol in 25K23Y and –348.3 kcal/mol in 23Y versus –344.7 kcal/mol in wild type. However, they also have larger solvation penalties compared with wild type, +518.4 kcal/mol in 25K23Y and +472.2 kcal/mol versus +467.4 kcal/mol in wild type. Therefore, the total electrostatic contribution to the binding free energy,

$\Delta G_{int+sol}^{ele}$, is more unfavorable than the wild type dimer and so are the binding free energies.

We investigated another two double mutations, 25K5F and 25K87W as well. Leu5 and Arg87 have large SCNANR, 64.7% and 36.8% respectively. They are not close to termini either. The TNAN values in 3Å distance cutoff are 31 and 29 respectively. This means they are crowded. We can see in Table 7 that both of them have less favorable van der Waals interaction energies than wild type, i.e. –186.8 kcal/mol in 25K5F and –187.4 kcal/mol in 25K87W versus –189.8 kcal/mol in the wild type dimer. In addition, their $\Delta G_{int+sol}^{ele}$ terms are also more unfavorable. This is due to larger solvation penalty. Arg87 is exposed to water. The solvation energy is unfavorable if this charged residue is mutated to a neutral one. In 25K5F, the aromatic ring of Phe5 is totally buried which is also unfavorable.

3. Conclusion

In the present study, we investigated the protonation state of the free HIV PR and calculated the relative binding free energies of different dimers of HIV PR to wild type dimer using the MM/PBSA method. We also suggest several dominant negative inhibition mutations on the basis of binding free energies calculated using the Virtual Mutagenesis method.

We compared the average distances between several pair atoms in the two catalytic Asp's obtained from the MD trajectory with those in the crystal structure. These calculations indicated that the dianionic state had the closest structure to the crystal structure. According to the binding free energy calculations on the wild type dimers with different protonation states, the dianionic dimer structure was also suggested to be most stable. These results are consistent with NMR data on the ligand free HIV protease (Smith et al., 1996).

The heterodimer formed between the triple mutation monomer, Asp25Lys/Gly49Trp/Ile50Trp (25K49W50W), and the wild type monomer was shown experimentally to have higher thermal stability than the wild type dimer (Rozzelle et al., 1999). We calculated the binding free energies on the 25K49W50W bound to wild type monomer with deprotonated and protonated catalytic Asp. Both of these two heterodimers are more stable than the wild type homodimers. The homodimer of the triple mutations is the least stable dimer. The ranking order of dimer stability is consistent with the experimental observations.

We also developed a method called Virtual Mutagenesis to identify mutations which can enhance binding between two subunits of a macromolecule. With the assumption that local mutations will not change the overall structure of a protein, this Virtual Mutagenesis method was used to calculate binding free energies for different HIV protease dimers. The ranking order of calculated binding free energies is consistent with that of viral infectivity. Moreover, several new dominant negative mutations were suggested by this method. Four of them, 25K48W, 25K49W, 25K50W and 25K52W, are similar to the triple mutations in terms of mutated residues. However, another two, 25K4Y and 25K7W, are novel mutations and are not obvious choices if one just visualizes the HIV PR structure. These results await experimental verification.

In summary, the MM/PBSA method is able to calculate binding free energies on systems for which more rigorous methods such as free energy perturbation (FEP) and thermodynamic integration (TI) can not be efficiently applied. The Virtual Mutagenesis method can quickly identify residues on which mutations can be made to enhance the binding between protein-protein or protein-ligand. Caveats include the assumptions of similar entropy change upon dimerization of different mutatnts, additivity of free energy terms, and adequacy of sampling of conformation space. With these caveats in mind, the results obtained using the MM/PBSA and the Virtual Mutagenesis methods are promising and worthy of further development and experimental testing.

4. Methods

(1) Protonation state of the HIV PR and MM/PBSA method

All molecular dynamics simulations presented in this work were preformed using the AMBER5.0 simulation package (Pearlman et al., 1995) and the Cornell et al. force field (Cornell et al., 1995) with TIP3P water model (Jorgensen et al., 1983). The starting structure for the wild type homodimer of the HIV protease was taken from the Protein Data Bank. The PDB entry is 3hvp (Wlodawer et al., 1989). Mutations were made manually using SYBYL6.5 (Tripos Associates Inc., 1998) and MidasPlus (Ferrin et al., 1988). The molecules were solvated in a $80\times80\times80$Å^3 box of water. All systems were neutralized by adding counter ions close to the solute surface. The number of counter ions varied with different HIV PR dimer. Particle Mesh Ewald (PME) (Darden et al., 1993) was exploited to consider the long-range electrostatic interactions. All structures were minimized first using SANDER module in AMBER5.0. Molecular dynamics simulations were carried out thereafter. The temperature of the system was raised gradually from 50K to 298 K in 50 ps followed by 120 ps equilibration at 298 K. Another 120 ps MD simulation was performed for data collection and 100 snapshots were saved for the consequent analysis. The SHAKE procedure (Rychaert et al., 1977) was employed to constrain all bonds. The time step of the simulations was 2 fs. A 8.5Å cut-off was used for the nonbonded van der Waals interactions and no cutoff was used for nonbonded electrostatic interactions. The nonbonded pairs were updated every 15 steps.

The binding free energy between the two monomers of the HIV PR was calculated according to the thermodynamic cycle shown in Figure 2.

$$\Delta G_b = \Delta G_b^0 + \Delta G_{sol}^D - \Delta G_{sol}^{M1} - \Delta G_{sol}^{M2} \qquad (1)$$

where ΔG_b^0 and ΔG_b are the binding free energies in gas and in water respectively, ΔG_{sol}^{M1}, ΔG_{sol}^{M2} and ΔG_{sol}^D are solvation free energies for the monomer 1, monomer 2 and dimer of the HIV PR respectively. ΔG_b^0 is calculated from molecular mechanics (MM) interaction energies:

$$\Delta G_b^0 = \Delta G_{int}^{ele} + \Delta G_{int}^{vdw} \qquad (2)$$

where ΔG_{int}^{ele} and ΔG_{int}^{vdw} are electrostatic and van der Waals interaction energies between the two monomers in gas which were calculated using the CARNAL and ANAL modules in AMBER5.0 software suite.

The solvation energy, ΔG_{sol}, is divided into two parts, the electrostatic contributions, ΔG_{sol}^{ele}, and all other contributions, $\Delta G_{sol}^{nonpolar}$.

$$\Delta G_{sol} = \Delta G_{sol}^{ele} + \Delta G_{sol}^{nonpolar} \qquad (3)$$

The electrostatic contribution to the solvation free energy, ΔG_{sol}^{ele}, was calculated using the DelPhiII software package (Gilson et al., 1987), which solves the Poisson-Boltzmann equations numerically and calculates the electrostatic energy according to the electrostatic potential. The grid size we used was 0.5Å. Potentials at the boundaries of the finite-difference lattice were set to sum of the Debye-Huckel potentials. The interior dielectric constant was set to 1 in our primary simulations in order to be consistent with the molecular mechanics force field. Other value for the interior dielectric constant was also examined (see below). The dielectric constant of water was set to 80. The dielectric boundary was taken as the solvent accessible surface defined by a 1.4 Å probe sphere. The radii of atoms were taken from PARSE parameter set (Sitkoff et al., 1994). Partial charges were taken from Cornell et al. force field for standard amino acids. One non-standard amino acid in the 3hvp was ABA and its partial charges were calculated using *ab initio* calculations and the RESP method (Bayly et al., 1993).

As mentioned above, the value 1 was first used for the interior dielectric constant originally in MM/PBSA. Since the dielectric constants for the interior of proteins is considered to be in the range from 2 to 4, we examined the case where the interior dielectric constant had values other than 1. As shown in the Appendix, the binding free energy was calculated slightly different from equation 1.

$$\begin{aligned}\Delta G_b = {} & \Delta G_{int}^{vdw} + (\Delta G_{sol}{}^{nonpolar}{}_{D} - \Delta G_{sol}{}^{nonpolar}{}_{M1} - \Delta G_{sol}{}^{nonpolar}{}_{M2}) \\ & + (1/n)\, \Delta G_{int}^{ele} + (\Delta G_{RFE}{}^{D}{}_{n\text{-}80} - \Delta G_{RFE}{}^{M1}{}_{n\text{-}80} - \Delta G_{RFE}{}^{M2}{}_{n\text{-}80}) \\ = {} & \Delta G_{int}^{vdw} + \Delta G_{sol}^{nonpolar} + (1/n)\, \Delta G_{int}^{ele} + (\Delta G_{RFE}{}^{D}{}_{n\text{-}80} - \Delta G_{RFE}{}^{M1}{}_{n\text{-}80} - \Delta G_{RFE}{}^{M2}{}_{n\text{-}80})\end{aligned}$$

(4)

where n is the interior dielectric constant. In this study, it was set to 2. $\Delta G_{RFE}{}^{D}{}_{n\text{-}80}$, $\Delta G_{RFE}{}^{M1}{}_{n\text{-}80}$ and $\Delta G_{RFE}{}^{M2}{}_{n\text{-}80}$ are reaction field energies obtained from DelPhi for dimer, monomer 1 and monomer 2 of the HIV PR respectively with interior and exterior dielectric constants set to n and 80 respectively.

The solvent accessible surfaces (SAS) were calculated using the MSMS program (Sanner et al., 1996). The non-polar contribution to the solvation free energy, $\Delta G_{sol}^{nonpolar}$, was calculated as 0.00542×SAS+0.92 kcal/mol (Sitkoff et al., 1994).

It is worth pointing out that in Equation (1), no solute entropy contribution is included. We estimated the conformational entropy contribution (translation, rotation and vibration) to the binding free energy using normal mode analysis (Case, 1994). This is only an estimate for the order of magnitude of the entropy contribution. We assumed that the entropy contributions are similar for different HIV protease dimers. When we calculate the relative binding free energies between them, the entropy contribution is assumed to cancel. The normal mode analysis was carried out using the NMODE module in AMBER5.0. The structure used for normal mode analysis was obtained by minimizing the crystal structure of the wild type HIV PR dimer using a distance dependent dielectric constant which is proportional to 4r, where r is the distance between the atoms.

(2) Virtual Mutagenesis (VM) method

Mutations which might enhance binding between the two monomers of the HIV PR can only be made on residues which satisfy the following three qualitative criteria:

1. The vector from Cα to Cβ, $\mathbf{n}_{\alpha\beta}$, points toward the dimer interface;
2. The residue is close to the dimer interface;
3. The residue has some extra space around it and a number of atoms in another monomer are a short distance from this residue.

The idea is that the mutation will not change the HIV PR structure dramatically (criterion 1) and more favorable contacts with another monomer can be created if this residue is mutated to another residue with a larger side chain (criterion 2 and 3).

First, in order to identify those residues satisfying these criteria in the HIV PR, vector **n,** which was perpendicular to the plane defined by three atoms in the B chain of the HIV PR, N in Gly49B, Cα in Asn98B and Cα in Arg8B, was constructed. The plane defined by the three atoms was parallel to the dimer interface. $\mathbf{n}_{\alpha\beta}$ for each residue in the A chain of the HIV PR was also calculated. By examining the sign of $\mathbf{n}$ $\mathbf{n}_{\alpha\beta}$, we could determine whether $\mathbf{n}_{\alpha\beta}$ pointed to or away from the interface.

Second, Total Neighbor Atom Number (TNAN) and Contact Neighbor Atom Number (CNAN) values were calculated for two different distance cutoffs. The Total Neighbor Atom Number (TNAN) is the number of atoms within the distance cutoff, r_{thr}, of any atom of the residue being investigated. The Contact Neighbor Atom Number (CNAN) is the number of atoms in another subunit of the molecule within the distance cutoff r_{thr}. The value of TNAN reflects how crowded the residue being investigated is surrounded by other residues and the value of CNAN represents how many contacts this residue has with another subunit of the molecule. Obviously, the values of TNAN and CNAN depend on the distance cutoff, r_{thr}. In the present study, we used two distance cutoffs, 3Å and 6Å. A ratio, Shell Contact Neighbor Atom Number Ratio (SCNANR), was calculated as:

$$\mathrm{SCNANR} = (\mathrm{CNAN}_2 - \mathrm{CNAN}_1)/(\mathrm{TNAN}_2\text{-}\mathrm{TNAN}_1) \qquad (5)$$

where CNAN_1 and CNAN_2 are CNAN using 3Å and 6Å distance cutoff respectively and so are TNAN_1 and TNAN_2 for TNAN. This ratio, SCNANR, reflects how many contacts may form if the current side chain is replaced by a larger one.

A residue which satisfies the second criterion must have a relatively large CNAN value at least at the 6 Å distance cutoff if not already at the 3 Å distance cutoff. The third criterion requires residues which have small TNAN values in the 3 Å distance cutoff range and large SCNANR values.

Glycines were considered specifically. In addition to the three criteria, ϕ and φ torsion angles for each Glycine were also examined. Only those Glycines whose ϕ and φ were not unique for Glycines, i.e. were in the left half of the Ramachandran map, were considered for mutation.

Any residues satisfying the above three criteria were mutated to one of those amino acids which has larger side chain, such as Trp, Tyr etc. A systematic conformation search for 100 conformations was carried out for the mutant residue. For each conformation, a steric bump check was executed first to avoid serious steric clash between the mutant residue and any other residue in the molecule. If any atom of the built-in residue was closer than 1 Å to any atom of any other residue in the molecule, that conformation of the built-in residue was discarded. Each surviving conformation was then minimized using the SANDER module in AMBER5.0. Only the mutant residue was allowed to move. No explicit water was added. The solvent effect was considered roughly by using a distance dependent dielectric constant which was proportional to 4r, where r is the distance between atoms. MM/PBSA was used to evaluate the energy of each conformation. The final energy for each mutation was the average energy of all conformations. Multiple mutations were made one by one. For example, the triple mutation D25K/G49W/I50W was obtained by mutating Asp25 to Lys first. A conformation of Lys25 with closest binding free energy to the average value was chosen and Gly49 was mutated to Trp. The third mutation, I50W, was made on the conformation whose binding free energy is closest to the average one for the D25K/I49W mutations. Since no full MD simulations were carried out, the computational efficiency is quite high using such an approach.

5. Acknowledgement

This work was supported in part by grants GM-56531 (P. Ortiz de Montellano, P.I.) and GM-56609 (E. Arnold, P.I.). Part of this investigation was completed using the facilities in the UCSF Computer Graphics Lab, T. Ferrin, director, supported by RR-1081 from the NIH. WW is grateful for kind help and stimulating discussions from Drs. Yong Duan, Bernd Kuhn and Taisung Lee. Supercomputer time at the NSF centers (NCSA and SDSC) is gratefully acknowledged.

6. References

Ajay, & Murcko, M. A., (1995) Computational methods to predict binding free energy in ligand-receptor complexes. *J. Med. Chem.* 38, 4953-4967.

Babé, L. M., Rose, J. R. & Craik, C. S. (1995) Trans-dominant inhibitory human immunodeficiency virus type 1 protease monomers prevent protease activation and virion maturation. *Proc. Natl. Acad. Sci. USA*, 92, 10069-10073.

Bayly, C. I., Cieplak, P., Cornell, W. D. & Kollman, P. A. (1993) A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges - the RESP model. *J. Phys. Chem.* 97, 10269-10280.

Beveridge, D. L. & Dicapua, F. M. (1989) Free energy via molecular simulations: application to chemical and biochemical system. *Annu. Rev. Biophys. Biophys. Chem.* 18, 431-492.

Case, D. A. (1994) Normal mode analysis of protein dynamics. *Curr Opin. Strut. Biol.* 4, 285-290.

Cheng, Y. S. E., Yin, F. H., Foundling, S., Blomstrom, D. & Kettner, C. A. (1990) Stability and activity of human immunodeficiency virus protease – comparison of the nautral dimer with a homologous, single-chain tethered dimer. *Proc. Natl. Acad. Sci. USA*, 87, 9660-9664.

Collins, J. R., Burt, S. K. & Erickson, J. W. (1995) Flap opening in HIV-1 protease simulated by 'activated' molecular dynamics. *Nature Struct. Biol.* 2, 334-338.

Cornell, W. D, Cieplak, P., Bayly, C. I., Gould, I., Merz, K. M., Ferguson, D., Spellmeyer, D. C., Fox, T., Caldwell, J. W. & Kollman, P. A. (1995) A second generation force field for the simulation of proteins, , nucleic acids, and organic molecules. *J. Am. Chem. Soc.* 117, 5179-5197.

Darden, T. A., York, D. M. & Pedersen, L. (1993) Particle Mesh Ewald – an Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98, 10089-10092.

Debouck, C., Gorniak, J. G., Strickler, J. E., Meek, T. D., Metcalf, B. W. & Rosenberg, M. (1987) Human immunodeficiency virus protease expressed in Escherichia coli exhibits autoprocessing and specific maturation of the gag precursor. *Proc. Natl. Acad. Sci. USA*, 84, 8903-8906.

DiIanni, C. L., Davis, L. J., Holloway, M. K., Herber, W. K., and others (1990) Characterization of an active single polypeptide form of the human immunodeficiency virus type-1 protease. *J. Biol. Chem.* 265, 17348-17354.

Ferrin, T. E., Huang, C. C., Jarvis, L. E. & Langridge, R. (1988) The MIDAS display system. *J. Mol. Graphics* 6, 13-27.

Gilson, M. K., Sharp, K. A. & Honig, B. H. (1987) Calculating electrostatic interactions in biomolecules: method and error assessment. *J. Comput. Chem.* 9, 327-335.

Grant, S. K., Deckman, I. C., Culp, J. S., Minnich, M. D., Brooks, I. S., Hensley, P., Debouck, C. & Meek, T. D. (1992) Use of protein unfolding studies to determine the conformational and dimerc stabilities of HIV-1 and SIV proteases. *Biochemistry* 31, 9491-9501.

Hyland, L. J., Tomaszek, T. A. Jr. & Meek, T. D. (1991) Human immunodeficiency virus-1 protease. 2. Use of pH rate studies and solvent kinetic isotope effects to elucidate details of chemical mechanism. *Biochemistry* 30, 8454-8463.

Ido, E., Han, H. P., Kezdy, F. J. & Tang, J., (1991) Kinetic studies of human immunodeficiency virus type 1 protease and its active site hydrogen bond mutant A28S. *J. Biol. Chem.* 266, 24359-24366.

Ishima, R., Freedberg, D. I., Wang, Y. X., Louis, J. M. & Torchia, D. A. (1999) Flap opening and dimer-interface flexibility in the free and inhibitor-bound HIV protease, and their implications for function. *Structure* 7, 1047-1055.

Jordan, S. P., Zugay, J., Darke, P. L. & Kuo, L. C. (1992) Activity and dimerization of human immunodeficiency virus protease as a function of solvent composition and enzyme concentration. *J. Biol. Chem.* 267, 20028-20032.

Jorgensen, W. L., Chandrasekhar, J., Madura, J., Impey, R. W. & Klein, M. L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926-935.

Kohl, N. E., Emini, E. A., Schleif, W. A., Davis, L. J., Heimbach, J. C., Dixon, R. A., Scolnick, E. M. & Sigal, I. S. (1988) Active human immunodeficiency virus protease is required for viral infectivity. *Proc. Natl. Acad. Sci. USA,* 85, 4686-4690.

Kollman, P. A. (1993) Free energy calculations: Applications to chemical and biochemical phenomena *Chem. Rev.* 93, 2395-2417.

Krausslich, H. G. (1991) Human immunodeficiency virus proteinase dimer as component of the viral polyprotein prevents particle assembly and viral infectivity. *Proc. Natl. Acad. Sci. USA*, 88, 3213-3217.

Luo, R., Head, M. S., Moult, J., & Gilson, M. K. (1998) pK(a) shifts in small molecules and HIV protease: Electrostatics and conformation. *J. Am. Chem. Soc.* 120, 6138-6146.

Massova, I. & Kollman, P. A. (1999) Computational alanine scanning to probe protein-protein interactions: A novel approach to evaluate binding free energies. *J. Am. Chem. Soc.* 121, 8133-8143.

McPhee, F., Good, A. C., Kuntz, I. D. & Craik, C. S. (1996) Engineering human immunodeficiency virus 1 protease heterodimers as macromolecular inhibitors of viral maturation. *Proc. Natl. Acad. Sci. USA*, 93, 11477-11481.

Perlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, T. E., Debolt, S., Ferguson, D., Seibel, G. & Kollman, P. A. (1995) AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comp. Phys. Comm.* 91, 1-41.

Rick, S. W., Erickson, J. W. & Burt, S. K., (1998) Reaction path and free energy calculations of the transition between alternate conformations of HIV-1 protease. *Proteins: Struct. Funct. Genet.* 32, 7-16.

Rozzelle, J. E., Dauber, D. S., Kelly, R. & Craik, C. S. (1999) Macromolecular inhibitors of HIV-1 protease: characterization of desigined heterodimers. submitted.

Rychaert, J. P., Ciccotti, G. & Berendsen, H. J. C. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* 23, 327-341.

Sanner, M. F., Olson, A. J. & Spehner, J. (1996) Reduced surface - an efficient way to compute molecular surfaces. *Biopolymers,* 38, 305-320.

Sitkoff, D., Sharp, K. A. & Honig, B. (1994) Accurate Calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* 98, 1978-1988.

Smith, R., Brereton, I. M., Chai, R. Y. & Kent, S. B. H. (1996) Ionization states of the catalytic residues in HIV-1 protease. *Nature Struct. Biol.* 3, 946-950.

Srinivasan, J., Cheatham, T. E. III, Cieplak, P., Kollman, P. A. & Case, D. A. (1998) Continuum solvent studies of the stability of DNA, RNA, and Phosphoramidate – DNA helices. *J. Am. Chem. Soc.* 120, 9401-9409.

Trylska, J., Antosiewicz, J., Geller, M., Hodge, C. N., Klabe, R. M., Head, M. S., & Gilson, M. K. (1999) Thermodynamic linkage between the binding of protons and inhibitors to HIV-1 protease. *Protein Sci.* 8, 180-195.

van Gunsteren, W. F. (1989) Methods for calculation of free energies and binding constants: successes and problems. In: "Computer Simulation of Biomolecular Systems" (van Gunsteren, W. F. & Weiner, P. K., eds), pp. 27-59, ESCOM, Leiden.

Wang, Y. X., Freedberg, D. I., Yamazaki, I., Wingfield, P. T., Stahl, S. J., Kaufman, J. D., Kiso, Y., & Torchia, D. A. (1996) Solution NMR evidence that the HIV-1 protease catalytic Aspartyl groups have different ionization states in the complex formed with the asymmetric drug KNI-272. *Biochemistry* 35, 9945-9950.

Wlodawer, A., Miller M., Jaskolski, M., Sathyanarayana, B. K., and others, (1989) Conserved folding in retroviral proteases-crystal structure of a synthetic HIV-1 protease. *Science* 245, 616-621.

Yamazaki, T., Nicholson, L. K., Torchia, D. A., Wingfield P., and others. (1994) NMR and X-ray evidence that the HIV protease catalytic Aspartyl groups are protonated in the complex formed by the protease and a non-peptide cyclic urea-based inhibitor. *J. Am. Chem. Soc.* 116, 10791-10972.

Zhang, Z. Y., Poorman, R. A., Maggiora, L. L., Heinrikson, R. L. & Kezdy, F. J. (1991) Dissociative inhivition of dimeric dnzymes – kinetic characterization of the inhibition of HIV-1 protease by its COOH$^-$ terminal tetrapeptid. *J. Biol. Chem.* 266, 15591-15594.

Zou, X. Q., Sun, Y. X. & Kuntz, I. D. (1999) Inclusion of solvation in ligand binding free energy calculations using the generalized Born model. *J. Am. Chem. Soc.* 121, 8033-8043.

Figure 1. Fragments of the two catalytic Aspartic acids capped with ACE and NME groups. A water molecule is proposed to interact with Asp' in the way as shown in the figure (Ido et al., 1991).

Figure 2. Thermodynamic cycle for calculating binding free energies.

ΔG_b^0 and ΔG_b are binding free energies in gas and in water respectively, ΔG_{sol}^{M1}, ΔG_{sol}^{M2} and ΔG_{sol}^{D} are solvation free energies for the monomer 1, monomer 2 and dimer of the HIV PR respectively.

$$\Delta G_b = \Delta G_b^0 + \Delta G_{sol}^D - \Delta G_{sol}^{M1} - \Delta G_{sol}^{M2}$$

$$\Delta G_{sol} = \Delta G_{sol}^{ele} + \Delta G_{sol}^{nonpolar}$$

where ΔG_{sol}^{ele} was obtained from PB calculations and $\Delta G_{sol}^{nonpolar}$ was calculated from solvent accessible surface.

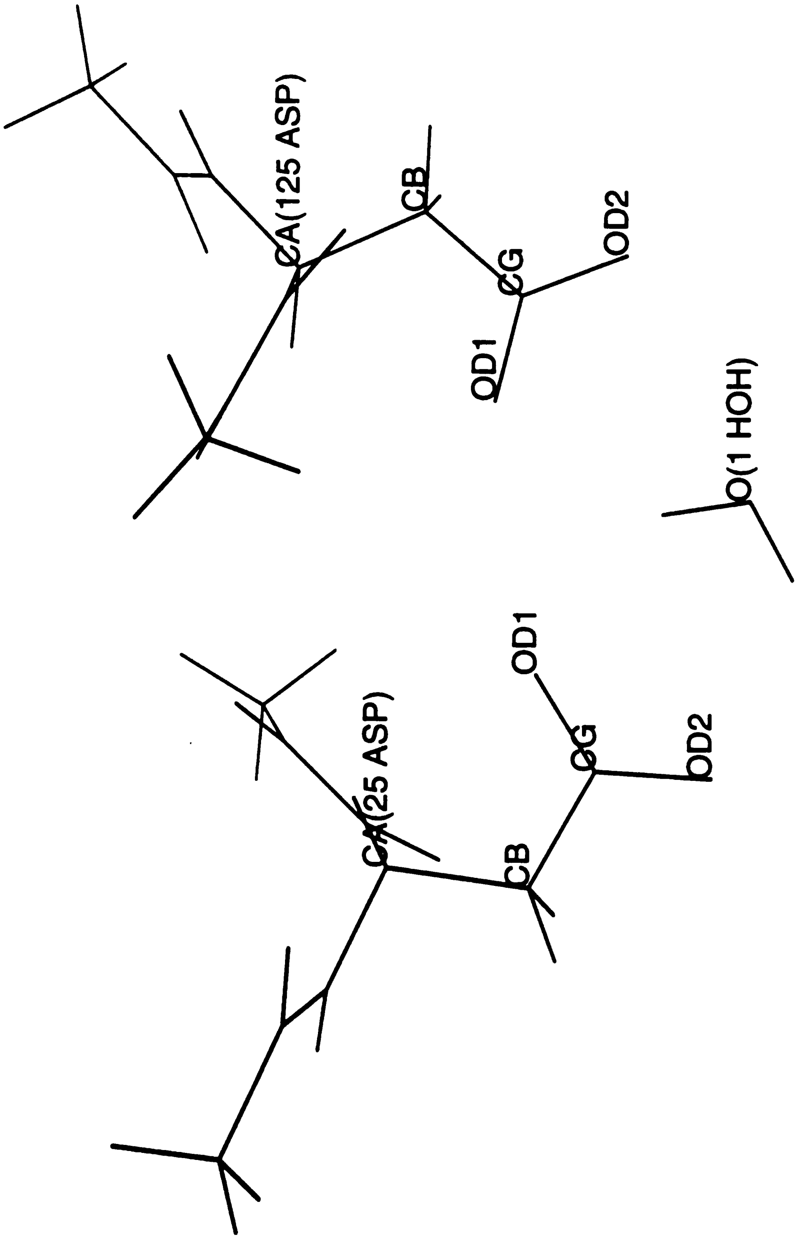
$$\Delta G_b^0 = \Delta G_{int}^{ele} + \Delta G_{int}^{vdw}$$

where ΔG_{int}^{ele} and ΔG_{int}^{vdw} were calculated from molecular mechanics energies.

Figure 3. Locations of the mutations on the HIV protease.

Figure 4a. RMSD of the "flap" region of different HIV protease dimers. Snapshot is taken every 0.120 ps.

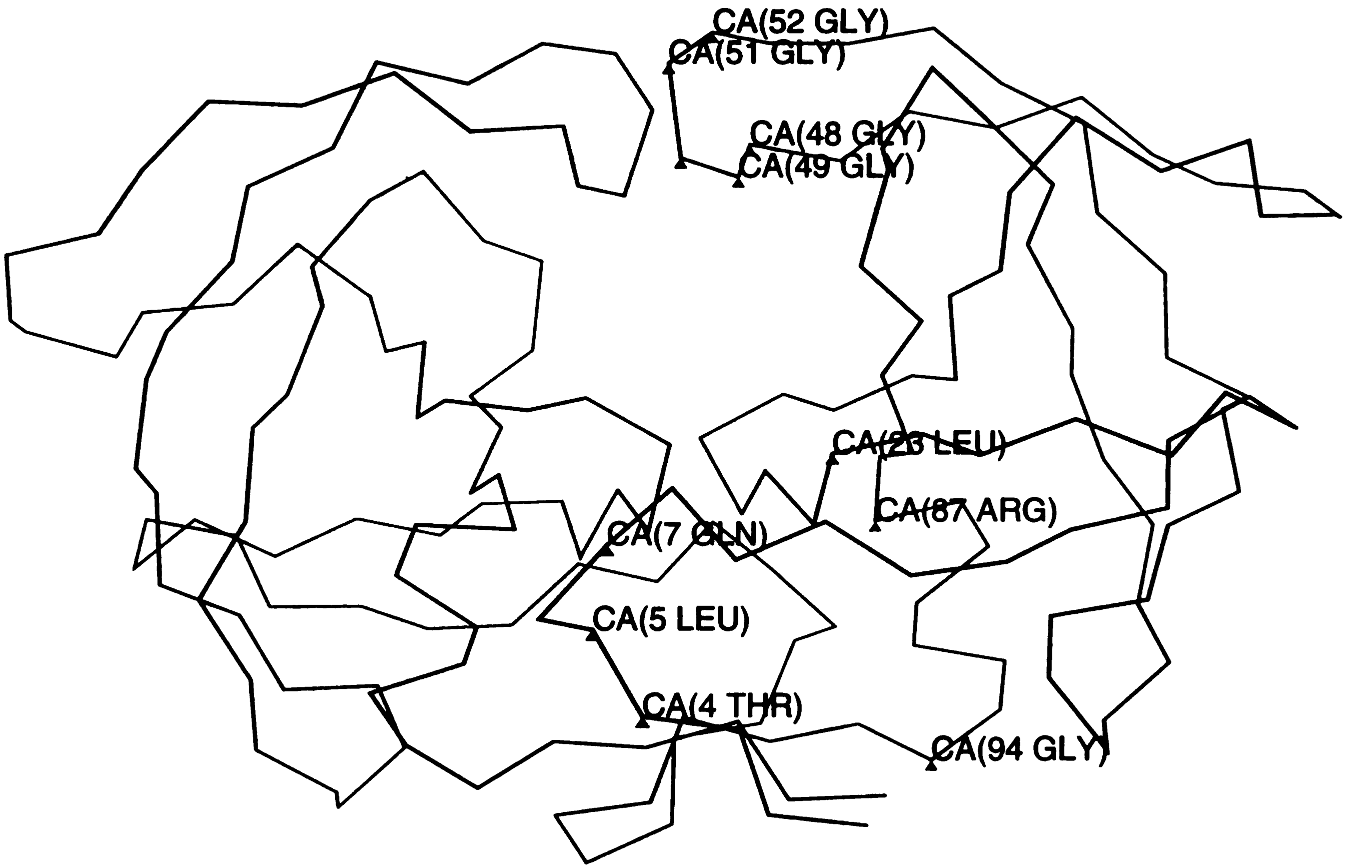
Figure 4b. RMSD of the "core" region of different HIV protease dimers. Snapshot is taken every 0.120 ps.



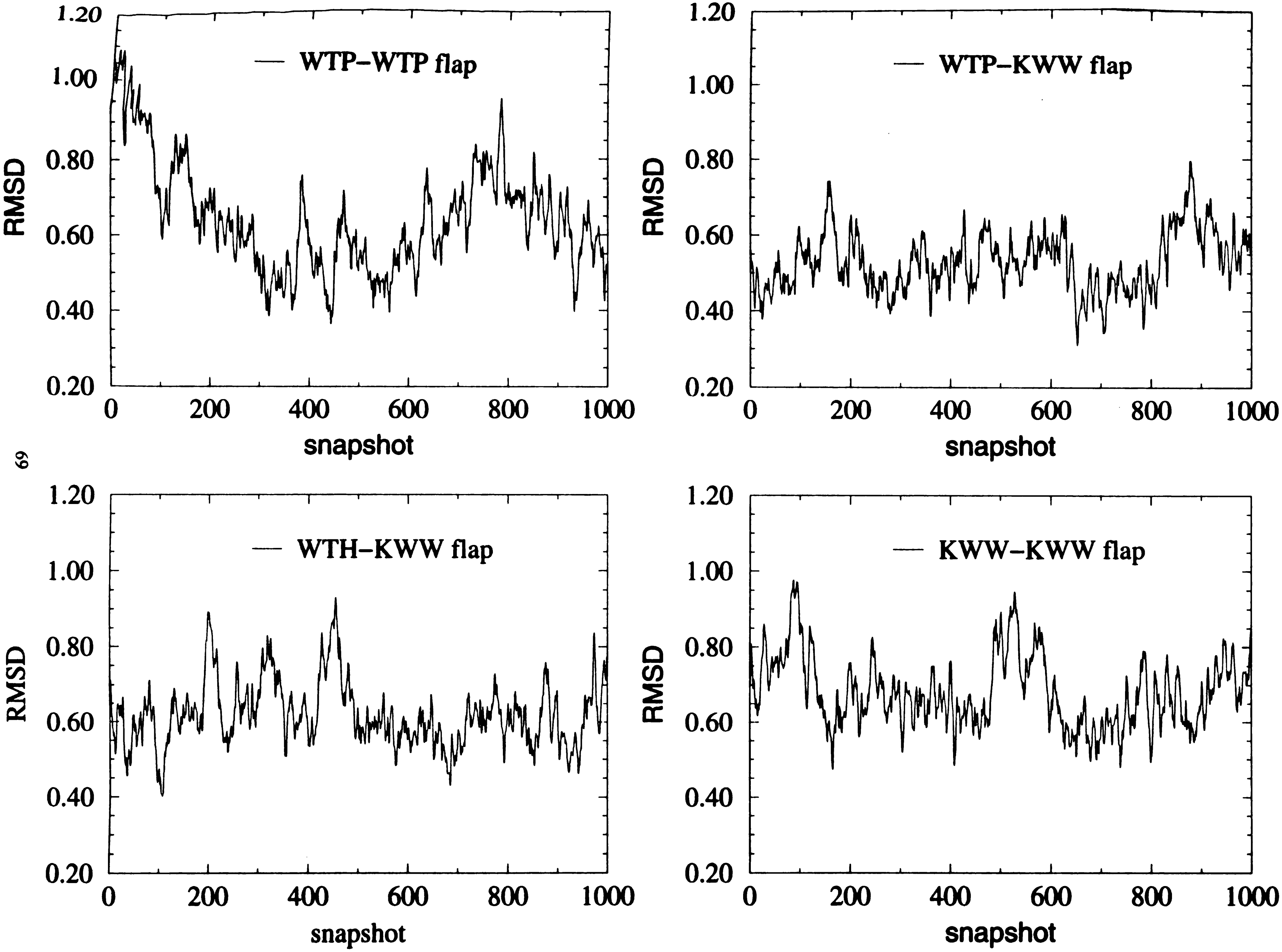
UCSF MidasPlus

Figure 2. Thermodynamic cycle for calculating binding free energies of HIV protease dimer.

$$\begin{tikzcd}
\text{In gas} & M1 \arrow[d, "\Delta G_{sol}^{M1}"'] & + & M2 \arrow[r, "\Delta G_b^0"] \arrow[d, "\Delta G_{sol}^{M2}"] & D \arrow[d, "\Delta G_{sol}^{D}"] \\
\text{In water} & M1 & + & M2 \arrow[r, "\Delta G_b"] & D
\end{tikzcd}$$



UCSF MidasPlus


WTP–WTP flap
WTP–KWW flap
WTH–KWW flap
KWW–KWW flap
RMSD
snapshot
1.20
1.00
0.80
0.60
0.40
0.20
0
200
400
600
800
1000

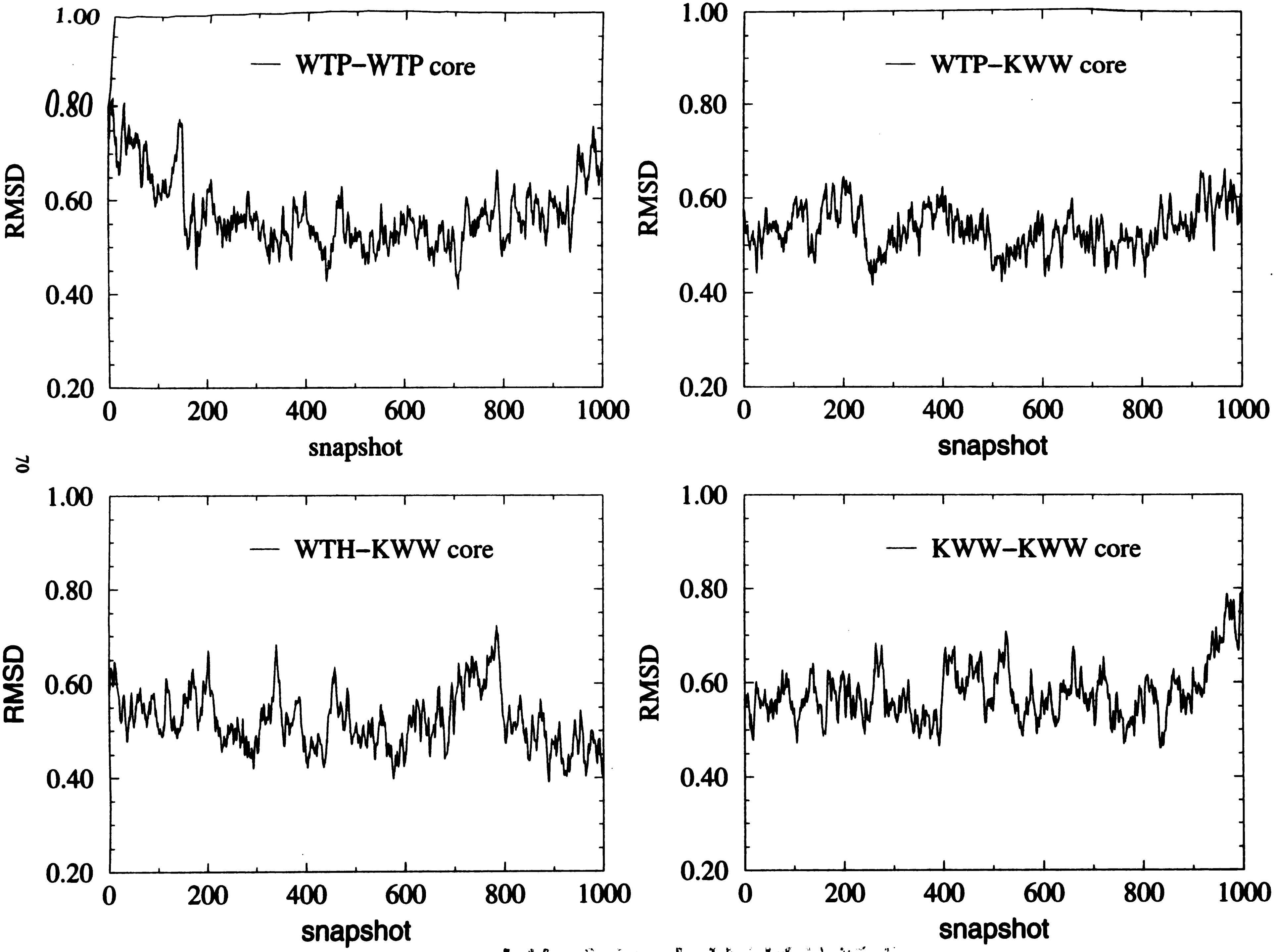

— WTP–WTP core
— WTP–KWW core
— WTH–KWW core
— KWW–KWW core
RMSD
1.00
0.80
0.60
0.40
0.20
0
200
400
600
800
1000
snapshot

Table 1. The distances between pairs of atoms in the two catalytic aspartic acid residues of the HIV protease.

Atom Pair[a]	$R_{cryt'l}$[b](Å)	PP[c]	HP	HH
		R_{MD}[d](Å)	R_{MD}[d](Å)	R_{MD}[f](Å)
CA-CA	6.71	6.55±0.23	6.32±0.17	6.29±0.20
CB-CB	7.60	7.31±0.22	6.90±0.22	6.87±0.19
CG-CG	5.28	5.01±0.23	4.50±0.24	4.41±0.22
OD1-OD1	3.01	3.68±0.29	2.58±0.14	3.14±0.25
OD2-OD2	5.81	5.12±0.29	4.71±0.34	4.48±0.28

a. The first atom is in Asp25 and the second in Asp25'; b. $R_{cryt'l}$ is measured in the crystal structure; c. Protonation states of the HIV PR: PP represents double ionic states, HP represents protonated Asp25 and deprotonated Asp25', HH represents double protonated state; d. R_{MD} is the average distance of the 100 snapshots taken from the MD trajectory.

Table 2. Influence of the protonation states of the two aspartic acids at the active site to the binding free energies.

dimers	ΔG_{int}^{vdw} (kcal/mol)	ΔG_{int}^{ele} (kcal/mol)	$\Delta G_{sol}^{nonpol\$}$ (kcal/mol)	$\varepsilon_{in}=1, \varepsilon_{out}=80$				$\varepsilon_{in}=2, \varepsilon_{out}=80$			
				$\Delta G_{sol}^{ele\dagger}$ (kcal/mol)	$\Delta G_{int+sol}^{ele\S}$ (kcal/mol)	ΔG_b^{*} (kcal/mol)	$\Delta\Delta G_b^{\P}$ (kcal/mol)	$\Delta G_{sol}^{ele\ddagger}$ (kcal/mol)	$\Delta G_{int+sol}^{ele\S}$ (kcal/mol)	ΔG_b^{**} (kcal/mol)	$\Delta\Delta G_b^{\P}$ (kcal/mol)
WTP- WTP[a]	-183.0 ±0.2	-365.1 ±4.2	-16.6 ±0.8	+480.4 ±5.1	+115.3 ±0.9	-84.3 ±1.9	0	+236.7 ±2.5	+54.2 ±0.7	-145.5 ±1.4	0
WTH- WTP[b]	-180.0 ±0.2	-332.6 ±14.4	-17.0 ±0.1	+447.1 ±12.4	+114.5 ±2.0	-82.5 ±1.8	+1.8	+219.9 ±6.1	+53.6 ±1.1	-143.4 ±0.9	+2.1
WTH- WTH[c]	-173.6 ±0.5	-307.3 ±5.5	-16.6 ±0.2	+417.0 ±6.0	+109.7 ±0.5	-80.5 ±0.1	+3.8	+205.0 ±3.0	+51.4 ±0.3	-138.8 ±0.4	+6.7

a. The wild type HIV PR dimer at double ionic state; b. The wild type HIV PR dimer at mono-ionic state; c. The wild type HIV PR dimer at double protonated state.

\$ $\Delta G_{sol}^{nonpol} = \Delta G_{sol}^{nonpol}{}_{D} - \Delta G_{sol}^{nonpol}{}_{M1} - \Delta G_{sol}^{nonpol}{}_{M2}$;

† $\Delta G_{sol}^{ele}(\varepsilon_{in}=1, \varepsilon_{out}=1) = \Delta G_{RFE}^{1\text{-}80}{}_{D} - \Delta G_{RFE}^{1\text{-}80}{}_{M1} - \Delta G_{RFE}^{1\text{-}80}{}_{M2}$;

‡ $\Delta G_{sol}^{ele}(\varepsilon_{in}=2, \varepsilon_{out}=1) = \Delta G_{RFE}^{2\text{-}80}{}_{D} - \Delta G_{RFE}^{2\text{-}80}{}_{M1} - \Delta G_{RFE}^{2\text{-}80}{}_{M2}$;

§ $\Delta G_{int+sol}^{ele} = \Delta G_{int}^{ele} + \Delta G_{sol}^{ele}$;

* $\Delta G_b = \Delta G_{int}^{vdw} + \Delta G_{int}^{ele}(\varepsilon_{in}=1, \varepsilon_{out}=1) + \Delta G_{sol}^{nonpol} + \Delta G_{sol}^{ele}(\varepsilon_{in}=1, \varepsilon_{out}=1)$;

** $\Delta G_b = \Delta G_{int}^{vdw} + (1/2)\,\Delta G_{int}^{ele}(\varepsilon_{in}=2, \varepsilon_{out}=1) + \Delta G_{sol}^{nonpol} + \Delta G_{sol}^{ele}(\varepsilon_{in}=2, \varepsilon_{out}=1)$;

¶ $\Delta\Delta G_b = \Delta G_b(\text{dimer}) - \Delta G_b(\text{WT})$.

Table 3. RMSD and its deviation σ of the "flap" region (residue 28-60 in each monomer) and the "core" region (residue 1-27 and 61-99) for different HIV PR dimers.

Dimer	Flap RMSD[a]	Flap RMSD deviation σ	Flap σ /RMSD	Core RMSD[a]	Core RMSD deviation σ	Core σ /RMSD
WTP-WTP	0.645	0.144	0.223	0.571	0.074	0.130
WTP-KWW	0.529	0.078	0.147	0.534	0.046	0.086
WTH-KWW	0.620	0.087	0.140	0.523	0.060	0.115
KWW-KWW	0.681	0.095	0.140	0.578	0.058	0.100

a. RMSD was calculated for all heavy atoms on the backbone compared with average structure obtained from the MD.

Table 4. MM/PBSA results on the binding free energies of different HIV PR dimers.

dimers	ΔG_{int}^{vdw} (kcal/mol)	ΔG_{int}^{ele} (kcal/mol)	$\Delta G_{sol}^{nonpol\$}$ (kcal/mol)	$\Delta G_{sol}^{ele\dagger}$ (kcal/mol)	$\Delta G_{int+sol}^{ele\S}$ (kcal/mol)	ΔG_b^{*} (kcal/mol)	$\Delta\Delta G_b^{\P}$ (kcal/mol)	Expt'l ranking order
WTP–WTP[a]	-183.0±0.2	-365.1±4.2	-16.6±0.8	+480.4±5.1	+115.3±0.9	-84.3±1.9	0	2
WTP-KWW[b]	-190.6±0.9	-445.4±6.6	-17.8±0.2	+560.2±4.7	+114.8±1.9	-93.6±0.8	-9.3	1
WTH-KWW[c]	-187.8±1.6	-293.7±10.6	-18.1±0.2	+411.0±8.8	+117.3±1.9	-88.6±0.5	-4.3	1
KWW-KWW[d]	-199.6±2.2	-159.0±9.0	-19.4±0.1	+298.2±5.9	+139.2±3.0	-79.8±0.9	+4.3	N/A

a. The wild type homodimer with double ionic catalytic Asp's; b. The heterodimer between the wild type monomer with ionic catalytic Asp and the triple mutation monomer; c The heterodimer between the wild type monomer with protonated catalytic Asp and the triple mutation monomer; d. The triple mutation homodimer.

$ $\Delta G_{sol}^{nonpol} = \Delta G_{sol}^{nonpol}{}_{D} - \Delta G_{sol}^{nonpol}{}_{M1} - \Delta G_{sol}^{nonpol}{}_{M2}$;

† $\Delta G_{sol}^{ele}(\varepsilon_{in}=1, \varepsilon_{out}=1) = \Delta G_{RFE}^{1\text{-}80}{}_{D} - \Delta G_{RFE}^{1\text{-}80}{}_{M1} - \Delta G_{RFE}^{1\text{-}80}{}_{M2}$;

§ $\Delta G_{int+sol}^{ele} = \Delta G_{int}^{ele} + \Delta G_{sol}^{ele}$;

* $\Delta G_b = \Delta G_{int}^{vdw} + \Delta G_{int}^{ele}(\varepsilon_{in}=1, \varepsilon_{out}=1) + \Delta G_{sol}^{nonpol} + \Delta G_{sol}^{ele}(\varepsilon_{in}=1, \varepsilon_{out}=1)$;

¶ $\Delta\Delta G_b = \Delta G_b(\text{dimer}) - \Delta G_b(\text{WT})$.

Table 5. Binding free energies of several dimers calculated using the Virtual Mutagenesis method.

Mutation	ΔG_{int}^{vdw} (kcal/mol)	ΔG_{int}^{ele} (kcal/mol)	$\Delta G_{sol}^{nonpol\$}$ (kcal/mol)	$\Delta G_{sol}^{ele\dagger}$ (kcal/mol)	$\Delta G_{int+sol}^{ele\S}$ (kcal/mol)	ΔG_b^{*} (kcal/mol)	$\Delta\Delta G_b^{\P}$ (kcal/mol)	Expt'l virial infectivity ranking order[a]
25K49W50W	-193.1 ±3.0	-425.3±1.9	-17.8±0.3	+546.0±2.8	+120.7±2.4	-90.2±1.7	-5.7	1
25R	-191.7 ±2.1	-401.8±10.0	-17.7±0.2	+525.1±9.5	+123.3±3.4	-86.1±3.5	-1.6	2
25K	-190.1 ±1.6	-401.3±15.0	-17.5±0.1	+523.0±14.4	+121.7±2.8	-85.9±3.5	-1.4	3
49W50W	-193.2 ±2.8	-350.6±1.8	-17.8±0.3	+476.3±2.11	+125.7±2.1	-85.3±1.5	-0.8	4
49W	-191.6 ±1.2	-347.2±1.3	-17.5±0.1	+471.3±2.5	+124.1±1.82	-85.0±1.1	-0.5	5
Wild Type	-189.8	-344.7	-17.4	+467.4	+122.7	-84.5	0	6

a. The smaller the ranking order, the weaker the viral infection, it is assumed, the more favorable the binding free energy for the dimer.

\$ $\Delta G_{sol}^{nonpol} = \Delta G_{sol}^{nonpol}{}_{D} - \Delta G_{sol}^{nonpol}{}_{M1} - \Delta G_{sol}^{nonpol}{}_{M2}$;

† $\Delta G_{sol}^{ele}(\varepsilon_{in}=1, \varepsilon_{out}=1) = \Delta G_{RFE}^{1\text{-}80}{}_{D} - \Delta G_{RFE}^{1\text{-}80}{}_{M1} - \Delta G_{RFE}^{1\text{-}80}{}_{M2}$;

§ $\Delta G_{int+sol}^{ele} = \Delta G_{int}^{ele} + \Delta G_{sol}^{ele}$;

* $\Delta G_b = \Delta G_{int}^{vdw} + \Delta G_{int}^{ele}(\varepsilon_{in}=1, \varepsilon_{out}=1) + \Delta G_{sol}^{nonpol} + \Delta G_{sol}^{ele}(\varepsilon_{in}=1, \varepsilon_{out}=1)$;

¶ $\Delta\Delta G_b = \Delta G_b(\text{dimer}) - \Delta G_b(\text{WT})$.

Table 6. Average Contact Neighbor Atom Number (CNAN) and Total Neighbor Atom Number (TNAN) of residues whose Cα-Cβ vector points toward the HIV PR dimer interface.

Sequence[1] Number	Residue Name	3Å distance cutoff		6Å distance cutoff		SCNANR[3] (X100%)
		CNAN[2]	TNAN[3]	CNAN[2]	RNAN[3]	
2A	**GLN**	**8**	**19**	**39**	**91**	**46.0**
4A	**THR**	**2**	**15**	**29**	**111**	**28.1**
5A	LEU	17	31	129	204	64.7
6A	**TRP**	**2**	**13**	**31**	**81**	**42.7**
7A	**GLN**	**2**	**17**	**22**	**97**	**25.0**
9A	PRO	6	26	51	192	27.1
23A[4]	LEU	3	30	24	200	12.4
24A	LEU	7	32	49	229	21.3
25A[4]	ASP	2	22	37	170	23.7
26A	THR	13	28	105	205	52.0
27A	**GLY**	**6**	**17**	**65**	**122**	**56.2**
48A	**GLY**	**4**	**18**	**28**	**100**	**29.3**
49A	**GLY**	**3**	**17**	**35**	**113**	**33.3**
50A	**ILE**	**4**	**14**	**51**	**90**	**61.8**
51A	**GLY**	**7**	**16**	**56**	**95**	**62.0**
52A	**GLY**	**2**	**16**	**26**	**100**	**28.6**
67A	ABA	1	15	6	115	5.0
69A	HIS	2	21	17	129	13.9
87A	ARG	9	29	79	219	36.8
90A	LEU	2	34	37	239	17.1
93A	ILE	3	27	21	164	13.1
94A	**GLY**	**0**	**14**	**22**	**95**	**27.2**
96A	THR	11	23	104	158	68.9
97A	LEU	24	38	163	236	70.2
99A	PHE	18	22	148	174	85.5

1. Sequence number is according to the wild type HIV PR dimer. The pdb entry is 3hvp; 2. CNAN and TNAN are Contact Neighbor Atom Number and Total Neighbor Atom Number respectively (see method); 3. SCNANR is the Shell Contact Neighbor Atom Number Ratio (see method); 4. 23A is listed due to special interests (see text).

Table 7. Binding free energies of different dimers calculated using the Virtual Mutagenesis method.

Mutation	ΔG_{int}^{vdw} (kcal/mol)	ΔG_{int}^{ele} (kcal/mol)	$\Delta G_{sol}^{nonpol\$}$ (kcal/mol)	$\Delta G_{sol}^{ele\dagger}$ (kcal/mol)	$\Delta G_{int+sol}^{ele\S}$ (kcal/mol)	ΔG_b^{*} (kcal/mol)	$\Delta\Delta G_b^{\P}$ (kcal/mol)
Wild Type	-189.8	-344.7	-17.4	+467.4	+122.7	-84.5	0
25K	-190.1 ±1.6	-401.3±15.0	-17.5±0.1	+523.0±14.4	+121.7±2.8	-85.9±3.5	-1.4
25K4Y	**-191.4 ±1.7**	**-421.9±0.5**	**-17.6±0.1**	**+542.1±3.3**	**+120.2±3.0**	**-88.8±1.8**	**-4.3**
25K5F	-186.8 ±0.0	-386.5±0.1	-17.5±0.0	+514.1±0.3	+127.6±0.0	-76.7±0.0	+7.8
25K7W	**-191.1 ±0.5**	**-423.4±0.5**	**-17.6±0.0**	**+538.0±1.6**	**+114.6±1.3**	**-94.1±1.1**	**-9.6**
25K23Y	-190.2 ±1.7	-394.8±4.3	-17.5±0.0	+518.4±4.1	+123.6±2.6	-84.1±2.9	+0.4
25K48W	**-195.7 ±0.5**	**-422.5±0.4**	**-18.5±0.0**	**+545.6±0.8**	**+123.1±0.4**	**-91.1±0.2**	**-6.6**
25K49W	**-192.4 ±0.3**	**-393.9±0.5**	**-17.6±0.1**	**+517.1±0.7**	**+123.2±0.1**	**-86.8±0.3**	**-2.3**
25K50W	**-191.6 ±2.8**	**-425.1±1.0**	**-17.7±0.3**	**+545.2±2.8**	**+120.1±2.4**	**-89.2±0.8**	**-4.7**
25K51W	-171.2 ±0.2	-420.9±0.0	-17.8±0.0	+540.0±0.5	+119.1±0.7	-69.9±0.7	+14.6
25K51A	-177.5 ±4.6	-422.5±1.3	-17.6±0.0	+540.6±1.6	+118.1±0.3	-77.0±4.3	+7.5
25K52W	**-192.4 ±0.9**	**-422.4±0.9**	**-17.7±0.0**	**+544.4±2.9**	**+122.0±2.5**	**-88.1±1.9**	**-3.6**
25K87W	-187.4 ±0.2	-428.7±0.4	-17.8±0.1	+553.0±0.2	+124.3±0.4	-80.9±0.3	+3.6
25K94W	*-192.7 ±1.8*	*-424.2±4.5*	*-17.8±0.2*	*+549.7±6.4*	*+125.5±2.7*	*-85.0±3.9*	*-0.5*
23Y	-189.7 ±2.0	-348.3±3.9	-17.4±0.0	+472.2±4.1	+123.9±1.3	-83.2±2.1	+1.3

$ $\Delta G_{sol}^{nonpol} = \Delta G_{sol}^{nonpol}{}_{D} - \Delta G_{sol}^{nonpol}{}_{M1} - \Delta G_{sol}^{nonpol}{}_{M2}$;

† $\Delta G_{sol}^{ele}(\varepsilon_{in}=1, \varepsilon_{out}=1) = \Delta G_{RFE}^{1\text{-}80}{}_{D} - \Delta G_{RFE}^{1\text{-}80}{}_{M1} - \Delta G_{RFE}^{1\text{-}80}{}_{M2}$;

§ $\Delta G_{int+sol}^{ele} = \Delta G_{int}^{ele} + \Delta G_{sol}^{ele}$;

* $\Delta G_b = \Delta G_{int}^{vdw} + \Delta G_{int}^{ele}(\varepsilon_{in}=1, \varepsilon_{out}=1) + \Delta G_{sol}^{nonpol} + \Delta G_{sol}^{ele}(\varepsilon_{in}=1, \varepsilon_{out}=1)$;

¶ $\Delta\Delta G_b = \Delta G_b(\text{dimer}) - \Delta G_b(\text{WT})$.

Appendix

In DelPhi, reaction field energy of a molecule is defined as the energy of taking the molecule from a solvent of dielectric equal to that of the interior, to that of the exterior under the condition that there is no salt present and the molecule lies entirely within the box (see DelPhi manual). For example, if the interior dielectric constant ε_{in} equals 2 and the exterior dielectric constant ε_{out} equals 1, the reaction field energy ΔG_{RFE}^{2-1} is calculated as the difference between electrostatic energies in (ε_{in}=2, ε_{out}=1) and (ε_{in}=2, ε_{out}=2) environments (Frame 1).

$$\Delta G_{RFE}^{2-1} = G_{2-1}^{ele} - G_{2-2}^{ele} \qquad (1)$$

where G_{2-1}^{ele} and G_{2-2}^{ele} are the electrostatic energies in the (ε_{in}=2, ε_{out}=1) and (ε_{in}=2, ε_{out}=2) environments respectively.

Thus:

$$\begin{aligned} G_{2-1}^{ele} &= G_{2-2}^{ele} + \Delta G_{RFE}^{2-1} \\ &= (1/2) \times G_{1-1}^{ele} + \Delta G_{RFE}^{2-1} \qquad (2) \end{aligned}$$

where G_{1-1}^{ele} is the electrostatic energy in gas. Therefore, G_{2-2}^{ele} equals half of G_{1-1}^{ele}.

The binding free energy, ΔG_b^0, of the HIV protease dimer in the (ε_{in}=2, ε_{out}=1) environment is calculated as:

$$\Delta G_b^0 = \Delta G_{int}^{vdw} + G_{2-1}^{ele}{}_{D} - G_{2-1}^{ele}{}_{M1} - G_{2-1}^{ele}{}_{M2} \qquad (3)$$

where ΔG_{int}^{vdw} is the van der Waals interaction energy between the two monomers, $G_{2-1}^{ele}{}_{M1}$, $G_{2-1}^{ele}{}_{M2}$, and $G_{2-1}^{ele}{}_{D}$ are electrostatic energies of monomer 1, monomer 2 and the dimer respectively. Substitute Equation (2) into Equation (3), we get:

$$\begin{aligned} \Delta G_b^0 = \Delta G_{int}^{vdw} &+ (1/2) \times (G_{1-1}^{ele}{}_{D} - G_{1-1}^{ele}{}_{M1} - G_{1-1}^{ele}{}_{M2}) \\ &+ (\Delta G_{RFE}^{2-1}{}_{D} - \Delta G_{RFE}^{2-1}{}_{M1} - \Delta G_{RFE}^{2-1}{}_{M2}) \end{aligned}$$

$$= \Delta G_{int}^{vdw} + (1/2) \times \Delta G_{int}^{ele} + (\Delta G_{RFE}{}^{2\text{-}1}{}_{D} - \Delta G_{RFE}{}^{2\text{-}1}{}_{M1} - \Delta G_{RFE}{}^{2\text{-}1}{}_{M2})$$

(4)

where ΔG_{int}^{ele} is the electrostatic interaction energy between the two monomers.

If one wants to calculate the binding free energy ΔG_b of the HIV PR dimer in water, one has to calculate the solvation energies ΔG_{sol}^{M1}, ΔG_{sol}^{M2} and ΔG_{sol}^{D} for the monomer 1, monomer 2 and dimer of the HIV PR (see Frame 2) respectively.

$$\Delta G_b = \Delta G_b^0 + \Delta G_{sol}^{D} - \Delta G_{sol}^{M1} - \Delta G_{sol}^{M2} \qquad (5)$$

The solvation energy can be decomposed to two parts, electrostatic contribution ΔG_{sol}^{ele} and all other contributions $\Delta G_{sol}^{nonpolar}$.

$$\Delta G_{sol} = \Delta G_{sol}^{ele} + \Delta G_{sol}^{nonpolar} \qquad (6)$$

According to the thermodynamic cycle shown in Frame 3, the electrostatic solvation energy ΔG_{sol}^{ele} of taking a molecule from gas (ε_{out}=1) to water (ε_{out}=80) is:

$$\Delta G_{sol}^{ele} = \Delta G_{RFE}{}^{2\text{-}80} - \Delta G_{RFE}{}^{2\text{-}1} \qquad (7)$$

Substitute Equation (6) and (7) into Equation (5), we get:

$$\Delta G_b = \Delta G_b^0 + (\Delta G_{sol}{}^{nonpolar}{}_{D} - \Delta G_{sol}{}^{nonpolar}{}_{M1} - \Delta G_{sol}{}^{nonpolar}{}_{M2})$$

(8)

$$+ (\Delta G_{RFE}{}^{2\text{-}80}{}_{D} - \Delta G_{RFE}{}^{2\text{-}80}{}_{M1} - \Delta G_{RFE}{}^{2\text{-}80}{}_{M2}) - (\Delta G_{RFE}{}^{2\text{-}1}{}_{D} - \Delta G_{RFE}{}^{2\text{-}1}{}_{M1} - \Delta G_{RFE}{}^{2\text{-}1}{}_{M2})$$

Substitute Equation (4) into Equation (8), we get the formula to calculate binding free energy of the HIV PR dimer whose interior dielectric constant ε_{in} equals 2.

$$\Delta G_b = \Delta G_{int}^{vdw} + (\Delta G_{sol}{}^{nonpolar}{}_{D} - \Delta G_{sol}{}^{nonpolar}{}_{M1} - \Delta G_{sol}{}^{nonpolar}{}_{M2})$$

$$+ (1/2) \times (G_{1\text{-}1}{}^{ele}{}_{D} - G_{1\text{-}1}{}^{ele}{}_{M1} - G_{1\text{-}1}{}^{ele}{}_{M2}) + (\Delta G_{RFE}{}^{2\text{-}80}{}_{D} - \Delta G_{RFE}{}^{2\text{-}80}{}_{M1} - \Delta G_{RFE}{}^{2\text{-}80}{}_{M2})$$

$$= \Delta G_{int}{}^{vdw} + \Delta G_{sol}{}^{nonpolar} + (1/2) \times \Delta G_{int}{}^{ele} + (\Delta G_{RFE}{}^{2\text{-}80}{}_{D} - \Delta G_{RFE}{}^{2\text{-}80}{}_{M1} - \Delta G_{RFE}{}^{2\text{-}80}{}_{M2}) \qquad (9)$$

where $\Delta G_{sol}{}^{nonpolar} = \Delta G_{sol}{}^{nonpolar}{}_{D} - \Delta G_{sol}{}^{nonpolar}{}_{M1} - \Delta G_{sol}{}^{nonpolar}{}_{M2}$.

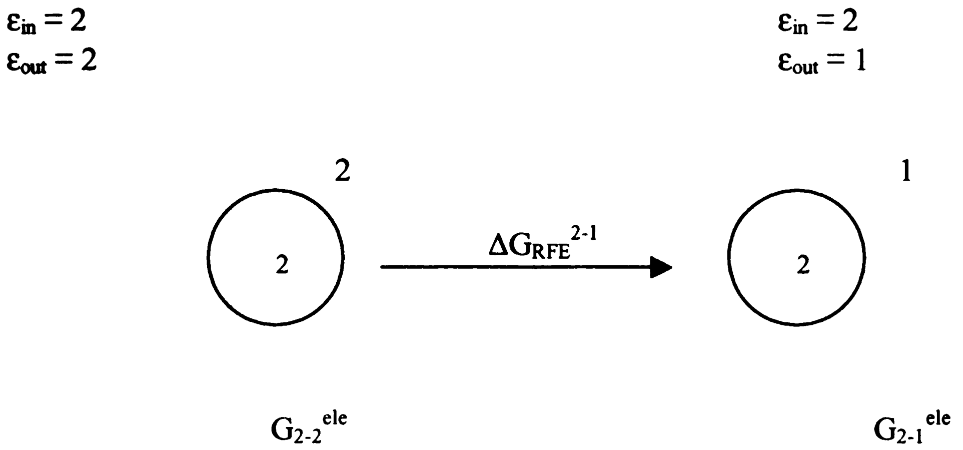
It is easy to generalize the above derivation to the case where the interior dielectric constant value equals n for any ligand-protein system. If ε_{in} equals n, Equation (9) becomes:

$$\Delta G_{b} = \Delta G_{int}{}^{vdw} + (\Delta G_{sol}{}^{nonpolar}{}_{LP} - \Delta G_{sol}{}^{nonpolar}{}_{L} - \Delta G_{sol}{}^{nonpolar}{}_{P})$$

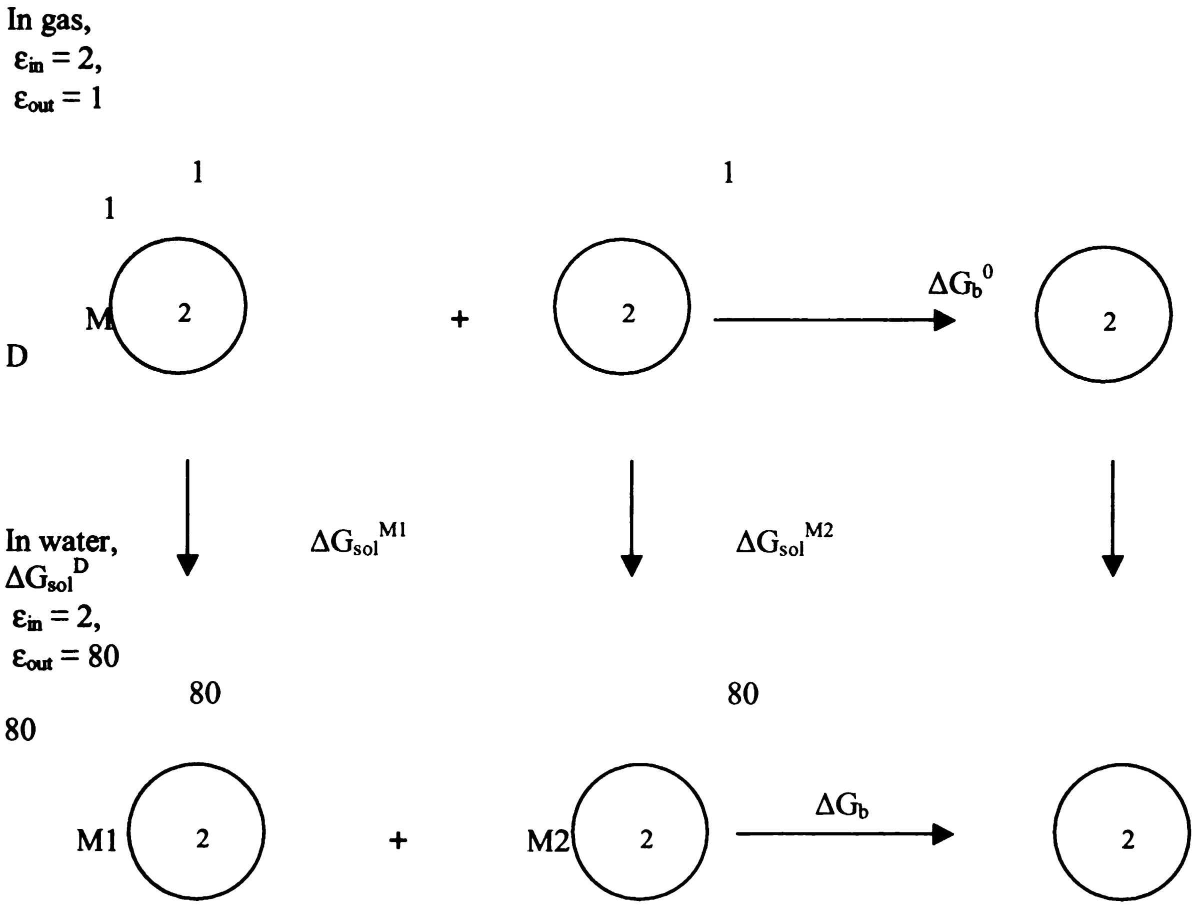
$$+ (1/n) \times (G_{1\text{-}1}{}^{ele}{}_{LP} - G_{1\text{-}1}{}^{ele}{}_{L} - G_{1\text{-}1}{}^{ele}{}_{P}) + (\Delta G_{RFE}{}^{2\text{-}80}{}_{LP} - \Delta G_{RFE}{}^{2\text{-}80}{}_{L} - \Delta G_{RFE}{}^{2\text{-}80}{}_{P})$$

$$= \Delta G_{int}{}^{vdw} + \Delta G_{sol}{}^{nonpolar} + (1/n) \times \Delta G_{int}{}^{ele} + (\Delta G_{RFE}{}^{2\text{-}80}{}_{LP} - \Delta G_{RFE}{}^{2\text{-}80}{}_{L} - \Delta G_{RFE}{}^{2\text{-}80}{}_{P}) \qquad (10)$$

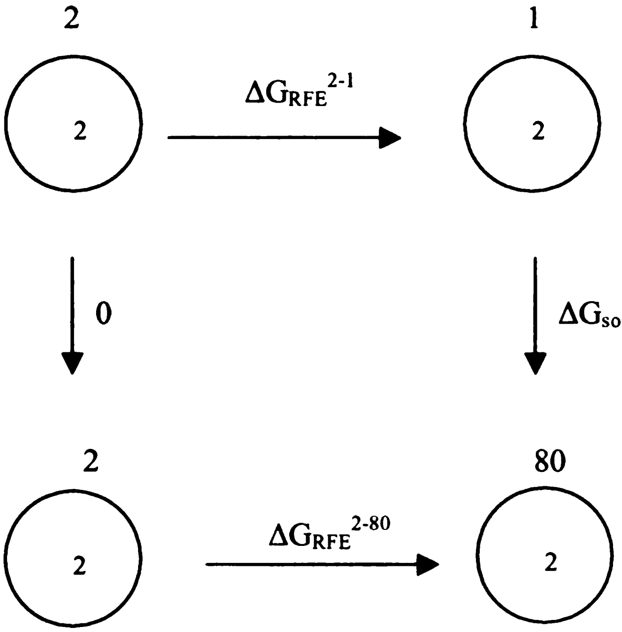
Frame 1. How the reaction field energy is calculated in DelPhi.



Frame 2 Thermodynamic cycle for calculating the binding free energy of the HIV PR dimer.



Frame 3 Thermodynamic cycle for calculating electrostatic interaction contribution to solvation free energy.



Chapter 4. An analysis of the interactions between the Sem-5 SH3 domain and its ligands using molecular dynamics, free energy calculations and sequence analysis

This chapter is a reprint of a paper accepted by Journal of the American Chemical Society. I did all the work except the calculations of the AM1-BCC charges of the peptoids, which were performed by Araz Jakalian and Christopher I. Bayly.

An analysis of the interactions between the Sem-5 SH3 domain and its ligands using molecular dynamics, free energy calculations and sequence analysis

Wei Wang,[1]

Wendell A. Lim,[1,2] Araz Jakalian,[3,4] Jian Wang,[5,6]

Junmei Wang,[6] Ray Luo,[6] Christopher I. Bayly,[3]

and

Peter A. Kollman[1,6]*

1. Graduate Group in Biophysics, University of California, San Francisco, San Francisco, CA 94143.
2. Department of Cellular and Molecular Pharmacology, Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA 94143.
3. Merck Frosst Canada & Co., 16711 TransCanada Hwy, Kirkland, Quebec H9W 5H7, Canada.
4. Current Address: Boehringer Ingelheim (Canada) Ltd., Research and Development, 2100 Rue Cunard, Laval, Quebec, H7S 2G6, Canada
5. Current Address: Ligand Pharmaceutical, 10275 Science Center Drive, San Diego, CA 92121.
6. Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143

* Author for correspondence

Tel: (415) 476-4637 (PAK)

Fax: (415) 502-1411

Email: pak@cgl.ucsf.edu

Short title: Interactions between Sem-5 SH3 domain and its ligands

Keywords: SH3 domain, binding free energy, ligand-protein interaction, molecular dynamics, continuum model.

Grant sponsor: NIH; Grant number: GM-29072 (P. A. Kollman, P.I.), GM-55040-01 (W. A. Lim, P.I.).

Abstract

The Src-homology-3 (SH3) domain of the *Caenorhabditis elegans* protein Sem-5 binds proline-rich sequences. It is reported that the SH3 domains broadly accept amide N-substituted residues instead of only recognizing prolines on the basis of side chain shape or rigidity. We have studied the interactions between Sem-5 and its ligands using molecular dynamics (MD), free energy calculations and sequence analysis. Relative binding free energies, estimated by a method called MM/PBSA, between different substitutions at site -1, 0 and $+2$ of the peptide are consistent with the experimental data. A new method to calculate atomic partial charges, AM1-BCC method, is also used in the binding free energy calculations for different N-substitutions at site -1. The results are very similar to those obtained from widely used RESP charges in the AMBER force field. AM1-BCC charges can be calculated more rapidly for any organic molecule than the RESP charges. Therefore, their use can enable a broader and more efficient use of the MM/PBSA method in drug design. Examination of each component of the free energy leads to the construction of van der Waals interaction energy profiles for each ligand as well as for wild type and mutant Sem-5 proteins. The profiles and free energy calculations indicate that the van der Waals interactions between ligands and the receptor determine whether an N- or Cα-substituted residue is favored at each site. A VC value (defined as a product of the conservation percentage of each residue and its van der Waals interaction energy with the ligand) is used to identify several residues on the receptor critical for specificity and binding affinity. This VC value may have a potential use to identify crucial residues for any ligand-protein or protein-protein systems. Mutations at two of those crucial residues, N190 and N206, are examined. One mutation, N190I, is predicted to reduce the selectivity of N-substituted residue at site -1 of the ligand and is shown to bind similarly with N- and Cα-substituted residues at that site.

1. Introduction

Molecular dynamics (MD) has provided dynamic and atomic insights to understand complicated biological systems. Free energy calculation methods have become powerful tools to provide quantitative measurement of protein-ligand or protein-protein interactions[1,2,3]. A new method, Molecular Mechanics/Poission Boltzmann Surface Area (MM/PBSA), was recently proposed for evaluating solvation and binding free energies of macromolecules and their complexes[4]. When this method is used to calculate binding free energy, the binding free energy is decomposed into contributions from van der Waals and electrostatic energies, non-polar and electrostatic solvation free energies, and relative solute entropy effects[5]. The van der Waals and electrostatic interactions between the components of the complex are calculated using molecular mechanics (MM) with an empirical force field[6], the non-polar part of solvation free energy is estimated by empirical methods based on solvent accessible (SA) surface and the electrostatic contribution to solvation is calculated using a continuum model and solving the Poisson-Boltzmann (PB) equation. The entropy contribution has been estimated using normal mode analysis[7]. An ensemble of different conformations is extracted from MD trajectories and each snapshot is analyzed using this MM/PBSA method. The binding free energies are obtained using an ensemble average. This method is able to calculate free energy differences between states even when the states are quite dissimilar from each other. It is also significantly more computationally efficient than standard free energy calculations[1].

With the human genome sequence nearing completion and the advancement of structure genomics, analyzing the amino acid sequence and structure of a protein can lead

to predictions of functions of other proteins. For example, if several critical residues for folding stability or substrate recognition are identified for one sequence whose structure is known, it could be possible to infer which residues are crucial for other sequences whose structures are unknown, which could provide useful guidance for designing new mutagenesis experiments and deducing their functions. An empirical parameter, VC value (see below), is introduced here to serve this purpose. In this paper, we combine molecular dynamics, free energy calculation, structure and sequence analysis to study interactions between Sem-5 SH3 domain and its ligands. Better understanding of the SH3 domain can lead to designing potent inhibitors or engineering its specificity.

Protein-protein interactions are essential for transmission of information in cellular signaling pathways. Specific classes of protein-protein interactions are mediated by families of small modular domains. These domains, found in diverse signaling proteins, recognize small peptide motifs in partner proteins. For example, Src homology 2 (SH2) domains bind specific phosphotyrosyl motifs while Src homology 3 (SH3) domains bind to polyproline motifs. Adaptor proteins that contain both SH2 and SH3 domains can therefore assemble multiple proteins around an activated, phosphorylated receptor[8-10]. One example is the *Caenorhabditis elegans* protein, Sem-5, which is composed solely of an SH2 and two SH3 domains. Sem-5 protein couples receptor tyrosine kinase activation to ras signaling [11-13]. The SH3 domains recognize the motif XPXXPXR, where X is any amino acid, found at the C-terminus of the exchange factor protein Sos[11,14]. Recent experimental work has focused on understanding how SH3 domains recognize the core of PXXP motif. Lim and coworkers found that SH3 domains recognize N-substituted residues instead of only prolines at site –1 and site +2 (Figure 1).

Thus, Proline is selected at these sites *in vivo* simply because it is the only natural N-substituted amino acid. In contrast, a $C\alpha$-substituted residue is required at site 0.[15] However, little is known about the energetic factors that yield this unusual backbone substitution pattern preference.

In the present study, molecular dynamics simulations are performed on Sem-5 SH3 domain complexed with ligands. Relative binding free energies between different ligands are calculated using the MM/PBSA method and the results are consistent with the measured binding affinities. We show that discrimination between N- and $C\alpha$- substituted residues at site −1, 0 and +2 are primarily due to van der Waals interactions between the SH3 domain and the ligand. N- and $C\alpha$- substituted residues are in different conformations, and this conformational heterogeneity is an essential feature of the different binding strengths. We then focus on studying different N-substitutions at site −1 of the ligand. Relative free energies of different ligands estimated by the MM/PBSA method with RESP charges correlate reasonably well with the measured ones. Free energy calculations have also been performed on these ligands using AM1-BCC charges[16,17] which can be calculated significantly faster than RESP charges. Results obtained from different charge models are very similar. Since AM1-BCC charges can be easily calculated for any organic molecule, this result suggests a more robust and general application of the MM/PBSA method in drug design. In order to identify crucial residues for binding, we construct van der Waals interaction energy profiles for the receptor and each ligand. Multiple sequence alignment is also carried out for Sem-5 SH3 domain. An empirical parameter, VC value, is implemented to identify several crucial residues on the receptor. Most of these crucial residues have also been identified in the previous experiments[18]. However, two of them, N190 and N206, were not studied before. Several mutations of these two residues are examined here. Based on the results of our free energy calculations, one mutation N190I has a very similar binding affinity with both site –1 N- and Cα- substituted ligands. Thus, the selectivity for an N-substituted residue at site –1 should be reduced for this mutant.

2. Methods

(1) MD simulations

All molecular dynamics simulations presented in this work are performed using the AMBER5.0 simulation package[19] and the Cornell et al. force field[6] with the TIP3P water model[20]. The starting structure for the wild type Sem-5 SH3 domain, which is 58 amino acid long, bound with PPPVPPR sequence is taken from the Protein Data Bank. The PDB entry is 1sem. Mutations are made manually using SYBYL6.5 (Tripos Associates Inc., 1998) and MidasPlus[21]. The molecules are solvated in a 60×60×60 Å^3 box of water. An appropriate number of counter ions are added to neutralize the system. Particle Mesh Ewald (PME)[22] is employed to calculate the long-range electrostatic interactions. All structures are minimized first using SANDER module in AMBER5.0. Molecular dynamics simulations are carried out thereafter. The temperature of the system is raised gradually from 50K to 298 K and the system is equilibrated at 298 K for 50 ps. Equilibrium is considered to be achieved after the RMSD, compared with the starting structure, reaches a plateau. Such a plateau was found within 50 ps for all the complexes. An additional 120 ps MD simulation is performed for data collection and 100 snapshots were saved for the subsequent analysis. The average backbone heavy atom RMSDs for all trajectories are around 1Å. The SHAKE procedure[23] is employed to constrain all

bonds. The time step of the simulations is 2 fs. A 8.5Å cut-off is used for the nonbonded van der Waals interactions and no cutoff for nonbonded electrostatic interactions. The nonbonded pairs are updated every 15 steps.

(2) The MM/PBSA method

The binding free energy is calculated as[24]:

$$\Delta G_b = \Delta G_{MM} + \Delta G_{sol}^{LP} - \Delta G_{sol}^{L} - \Delta G_{sol}^{P} - T\Delta S \qquad (1)$$

where ΔG_b is the binding free energies in water, ΔG_{MM} is the interaction energy between the ligand and the protein, ΔG_{sol}^{L}, ΔG_{sol}^{P} and ΔG_{sol}^{LP} are solvation free energies for the ligand, protein and complex respectively, and $-T\Delta S$ is the conformational entropy contribution to the binding. ΔG_{MM} is calculated from molecular mechanics (MM) interaction energies:

$$\Delta G_{MM} = \Delta G_{int}^{ele} + \Delta G_{int}^{vdw} \qquad (2)$$

where ΔG_{int}^{ele} and ΔG_{int}^{vdw} are electrostatic and van der Waals interaction energies between the ligand and the receptor, which are calculated using the CARNAL and ANAL modules in AMBER5.0 software suite.

The solvation energy, ΔG_{sol}, is divided into two parts, the electrostatic contributions, ΔG_{sol}^{ele}, and all other contributions, $\Delta G_{sol}^{nonpolar}$.

$$\Delta G_{sol} = \Delta G_{sol}^{ele} + \Delta G_{sol}^{nonpolar} \qquad (3)$$

The electrostatic contribution to the solvation free energy, ΔG_{sol}^{ele}, is calculated using the DelPhiII software package[25], which solves the Poisson-Boltzmann equations numerically and calculates the electrostatic energy according to the electrostatic potential. The grid size used is 0.5Å. Potentials at the boundaries of the finite-difference lattice are set to the sum of the Debye-Huckel potentials. The value of interior dielectric constant is set to 4. As shown in our previous study[26], after combining all the terms, the binding free energy is calculated as:

$$\Delta G_b = \Delta G_{int}^{vdw} + \Delta G_{sol}^{nonpolar} + (1/n)\ \Delta G_{1\text{-}1}^{ele} + (\Delta G_{RFE}{}^{LP}{}_{n\text{-}80} - \Delta G_{RFE}{}^{L}{}_{n\text{-}80} - \Delta G_{RFE}{}^{P}{}_{n\text{-}80}) \tag{4}$$

where n is the interior dielectric constant, which is 4 in this study. For comparison, free energies are also calculated using an interior dielectric constant of 1 (see Results and discussion and Supporting Material). $\Delta G_{1\text{-}1}^{ele}$ is the molecular mechanics electrostatic interaction energy between the ligand and the protein. $\Delta G_{RFE}{}^{LP}{}_{n\text{-}80}$, $\Delta G_{RFE}{}^{L}{}_{n\text{-}80}$ and $\Delta G_{RFE}{}^{P}{}_{n\text{-}80}$ are reaction field energies obtained from DelPhi for ligand, protein and complex respectively with interior and exterior dielectric constants set to n and 80 respectively.

The exterior dielectric constant is set to that of water (80). The dielectric boundary is taken as the solvent accessible surface defined by a 1.4 Å probe sphere. The radii of atoms are taken from the PARSE parameter set[27]. Partial charges are taken from Cornell et al. force field for standard amino acids. Partial charges of the non-standard amino acids were calculated using *ab initio* and RESP method[28]. AM1-BCC charges for N-substituted residues at site –1 are calculated by semi-empirical quantum method AM1 with bond charge corrections[16,17].

The solvent accessible surfaces (SAS) are calculated using the MSMS program[29]. The non-polar contribution to the solvation free energy, $\Delta G_{sol}^{nonpolar}$, is calculated as 0.00542×SAS+0.92 kcal/mol[27].

Normal mode analysis is used to estimate conformational entropy -TΔS. Because this analysis requires extensive computer time, only three snapshots are taken in this study to estimate the order of magnitude of the conformational entropy.

(3) Sequence alignment and definition of the VC value

Psi-BLAST[30] with default parameters (BLOSUM62, Expect=10, E-value threshold for inclusion in Psi-BLAST iteration=0.002, Descriptions=500, Alignments=500, composition based statistics) is used to search the SWISS-PROT database. Multiple sequence alignment is carried out on 207 sequences with scores >50 and E-value $<5\text{x}10^{-6}$ using the Pileup module in GCG software package (Version 10.1, Genetics Computer Group, Inc., 2000) with default parameters.

A parameter called the VC value (van der Waals and conservation), defined as the product of conservation percentage of an amino acid at the Sem-5 SH3 domain and its van der Waals interaction energy with the ligand, is used to identify critical residues for binding. The conservation percentage reflects how conserved the amino acid is and, therefore, it is the sum of appearance percentage (no gap included) of that specific amino acid and similar ones at a certain position. Appearance percentage reflects how often a specific amino acid appears at a certain position. For example, at position F163 of the Sem-5 SH3 domain, Tyr and Phe have 61% and 37% appearance percentage respectively in the multiple sequence alignment. Therefore, the conservation percentage of F163 is 98%.

3. Results and discussion

(1) MM/PBSA analysis accounts for observed SH3 domain site preferences

Lim and coworkers reported that in SH3 domains site –1 and site +2 favor N-substituted residues and site 0 favors a Cα-substituted residue. In their study, the representative N- and Cα-substituted residues are Sarcosine (Sar) (Figure 1(b) and Alanine (Ala) respectively[15]. We present here a computer modeling study to provide atomic and dynamic insights of how wild type and mutant ligands interact with the receptor.

A molecular dynamics (MD) simulation has been performed on the wild type peptide bound to the Sem-5 SH3 domain. The binding free energy calculated by the MM/PBSA method (see Method) is –39.3 kcal/mol (Table 1). Small errors (Table 1) and plateau RMSD compared with the crystal structure (data not shown) suggest convergence of the trajectory. Due to the considerable CPU cost for calculating the entropy contribution to the binding free energy, we only estimate the order of magnitude of the entropy contribution. We assume that the entropy contributions are similar for different ligands because all ligands in our study are just one residue different from the wild type. The entropy contribution estimated by normal mode analysis on three conformations is +29.0±1.0 kcal/mol. If this entropy term is included in the calculation, the absolute binding free energies for the wild type peptide is –10.3 kcal/mol, which is of the same order of magnitude as the measured value of –5.1 kcal/mol.

Binding free energies are also calculated for substitutions at site –1, 0 and +2 (Table 1). At site –1 and +2, Sar substitutions are more favorable than Ala substitution and at site 0 Ala is preferred over Sar. Thus, these results are able to reproduce the trend that Ala-1, Sar0 and Ala+2 bind significantly less well than the wild type sequence and the remaining mutants (Sar-1, Ala0, Pro0 and Sar+2) bind only slightly less well, qualitatively consistent with the experimental data[15]. The correlation coefficient r^2, between relative calculated and experimental binding free energies, is 0.88 (Figure 2). In summary, the MM/PBSA analysis accurately reproduces the ligand site preferences for the Sem-5 SH3 domain.

(2) van der Waals interactions between the ligand and the protein is the dominant factor for site preferences

As we mentioned above, what energetic factors determine the site preferences is not clear. One opinion is that desolvation is the determinant factor for substituting –NH with $-NCH_3$. In order to address this problem, we compare correlations between the measured binding free energies and each component of the calculated ones (Table 2). We find that van der Waals energy has the best correlation (r^2 is 0.88). There is no correlation between the measured binding free energies and the electrostatic interaction energy (Coulomb term) ($r^2 = 0.0088$) or electrostatic solvation energy (PB term) ($r^2 = 0.026$). However, these two terms compensate each other and the sum of them has a better r^2, which is 0.52. Solvent accessible surface term does not correlate well with the measure binding free energies either ($r^2 = 0.45$). It is obvious that van der Waals interactions between the ligand and the receptor is the dominant factor in site preferences.

The average van der Waals interaction energies during the trajectories between the protein and each residue of the ligand are calculated (Supporting Material). Analyzing van der Waals profiles and complex structures, following pictures of structural changes are suggested for substitutions at site 0, -1 and +2.

At site 0, discrimination between N- and Cα-substitution at site 0 is mainly due to the interaction difference between the Trp191 and residue at site 0. The average distances

between Trp191 CH2 atom and CB in the Ala0 or CD atom in the Sar0 are 5.5Å and 8.2 Å respectively.

The mutation at site –1 causes global change of the ligand van der Waals energy profile. In addition to the pair of residues at site –1, the pairs at site –2 and site +2 also have differences of more than 0.5 kcal/mol. Analysis of the profile of ligand and protein as well as the Phi angle of the ligand residues suggest the following picture of conformational changes due to mutation. For Sar-1, since Sar has a smaller side chain than Pro in the wild type and thus less attraction to N206, N206 moves toward Pro+2. The distance between N206 CG and Pro+2 CD is 4.7 Å and 4.3 Å in wild type and Sar-1 respectively. The Phi angle of Pro+2 is larger than that of the wild type, which means it moves into the pocket formed by F163, N206 and Y207. In order to maximize the interactions, Sar at site –1 moves toward N206. This makes the peptide more "helical" and Val0 inserts deeper into the pocket formed by F165, W191, P204 and Y207 (more favorable van der Waals interactions for Val0). Since Sar-1 drags Pro-2 and Arg-3 along with it, Pro-2 moves towards N190 and makes more contacts with N190. However, Arg-3 has less favorable van der Waals interactions with Gln168 and Glu172 as Arg-3 moves a little away from these residues. For the Ala-1 mutant, Pro+2 also intends to move toward N206 and Y207 (larger Phi angle). However, since there is no N-substituted group in Ala at site –1, N206 is more flexible. The interactions between Pro+2 and N206/Y207 are similar as in the wild type, but weaker than in Sar-1. The side chain of Ala-1 also keeps N190 from moving closer to Pro-2 to compensate some interactions as in Sar-1.

If the Pro at site +2 is mutated to Sar (Sar+2) or Ala (Ala+2), the primary difference is from Pro at site +3. The reason is that the new residue (Sar or Ala) has to adjust its conformation to have optimal interactions with both N206/Y207 and F163. Therefore, Sar+2 moves toward N206/Y207 and it brings Pro+3 closer to F163. Pro+3 even has more favorable van der Waals interaction energy than the wild type. However, in Ala+2 this adjustment is in the opposite direction, towards F163, which pushes Pro+3 even further away from the receptor. This introduces the major difference between Sar+2 and Ala+2 (Figure 1).

(3) Conformational changes of ligands are important for site preferences.

In order to address the importance of conformational changes of ligands for site preferences, binding free energies for substitutions at site −1, 0 and +2 with Sar and Ala, respectively, were also calculated using only the trajectory obtained for the wild type complex (Table 3). The underlying assumption is that the single mutation does not induce significant conformational change of the complex just like in the computational alanine scanning simulations[5]. This "alanine" scanning approach can only be used if the mutated residue is smaller than the wild type, which is the case for Pro->Sar or Pro->Ala mutations. From Table 3 we can see that the calculated difference between Sar and Ala substitution is small. The van der Waals interaction energies are very similar at site -1 and just slightly different for site 0 and site +2. The calculated ΔG correlate rather poorly with experiment ($r^2 = 0.34$). This suggests that using the wild type trajectory to estimate the ΔG of mutants is a poor approximation because we have neglected the subtle conformational changes that occur when a residue is substituted. The success of "computational alanine scanning" in the MDM2-p53 protein-protein complex[5] is likely due to relatively rigid backbone structure, the p53 remaining α-helical upon Ala mutations. However, in Sem-5 SH3 domain complexes, the ligands are more flexible and the assumption of backbone rigidity is less accurate.

We also calculated binding free energies using an interior dielectric constant of 1 instead of 4 (see Supporting Material). If we use a single trajectory, the results also correlate poorly with the experimental data, just as we found using a value of 4. However, using separate trajectories the calculations are consistent with the experimental measurements, just as was found with an interior dielectric of 4 ($r^2 = 0.88$).

In summary, substitution dependant conformational flexibility must be taken into account to accurately calculate the observed differences in binding.

(4) MM/PBSA method with different charge models, RESP and AM1-BCC, can reasonably reproduce relative binding free energies of N-substituted SH3 peptoids at site –1

We next focus on the site –1 and examine different N-substituted peptoids binding to the wild type receptor. Nguyen *et al* used 12-residue peptoids YEVPPPVXPRRR (X is a synthesized non-natural residue) in their study of mutations at site –1[15]. Since no crystal structure is available for any entire peptoid, we mutate the residue at site –1 in the shorter ligand PPPVPPR whose crystal structure has been solved. We assume that the relative binding free energies of different peptoids do not have significant changes in the longer or shorter peptoid. In Table 4, we present the results of using separate trajectories on different site –1 peptoids and using MM/PBSA to calculate their free energies of binding. Our calculated ΔG's correlate reasonably well with the measured values, with a correlation coefficient (r^2) of 0.78 (N2C excluded, see below).

The largest outlier is N2C. It is worth pointing out that N2C is the only charged residue at site -1 in our calculation. If N2C is included, the correlation coefficient r^2 is 0.60. The van der Waals interaction energy for N2C is not much less favorable than the wild type. The sum of its Coulomb term and electrostatic contribution to solvation (PB term) is much less favorable compared to other ligands. However, this term is not unfavorable enough.

The stereoisomers of NSF and NRF have similar solvation penalties. The difference between their binding affinities is due to their different binding patterns with the protein. The phenyl ring of NSF has close contacts with P204, Y207 and F165 while its methyl group points toward N206. However, NRF's phenyl ring packs with N190 and its methyl group points away from the protein. This is reflected in the profiles as that F165, P204, N206 and Y207 have more favorable van der Waals interactions with NSF than NRF.

NIP and NMC are similar. Both of them interact with N206/Y207 through a methyl group and meanwhile have favorable interactions with N190/W191. NMC has a longer side chain. It packs better with W191 than NIP. However, N190 is pushed a little further away from the peptide by NMC as well. The total van der Waals energies of NMC and NIP are similar. Their slight different binding affinities are due to different electrostatic contributions. This suggests that the van der Waals interactions dominate in the binding and electrostatic interactions determine the selectivity and "fine tune" the binding strength as well.

NBN packs perfectly with both N190/W191 and N206/Y207. That is why it has such a favorable van der Waals interaction energy. However, its solvation penalty is larger too. This is probably due to the burial of the polar phenyl ring.

In the previous MM/PBSA calculations, an interior dielectric constant 1 has been used to be consistent with the molecular mechanics force field[24]. Nonetheless, the dielectric constant inside a protein is considered to be in the range of 2 to 4. In order to make our model more realistic, we used a value of 4 in this study. Our calculations on site –1 N-substituted peptoids show that the results obtained using dielectric constant 4 correlate noticeably better with experimental data (r^2 = 0.78 N2C excluded) than those using value 1 (r^2 = 0.21 N2C excluded) (see Supporting Material).

In the above calculations, all atomic charges are calculated using RESP module in AMBER. This procedure requires a significant effort for each charge determination. Recently, Bayly and coworkers have developed a new algorithm to calculate partial charges for atoms, the AM1-BCC charges[16,17]. These are calculated to emulate a HF/6-31G* electrostatic potential around the molecule, as are the RESP charges, only at a fraction of the computational effort. We recalculated the binding free energies for all site-1 peptoids using AM1-BCC charges (Table 5). We can see that the results are reasonably well correlated with experimental data (r^2=0.64 N2C excluded) and those obtained from RESP charges (r^2=0.86 N2C included) (Figure 3). This suggests one can combine RESP charge for the protein and AM1-BCC charge for the ligand in applying the MM/PBSA method in drug design.

(5) The VC parameter allows identification of SH3 residues critical for binding and specificity

A. Combination of energetic and evolutionary information can be useful in identifying critical residues for binding

In this section, we focus on studying critical residues in the SH3 domain. As discussed above, the electrostatic solvation penalty compensates the Coulomb energy and van der Waals interactions dominate the site preferences. Here, we combine the evolution conservation information with the molecular mechanics energy to evaluate the significance of each residue for binding or stability. A parameter called the VC value (van der Waals and conservation) is calculated for each residue as the multiple of its van der Waals interaction energy with the ligand and its conservation percentage in the multiple sequence alignment. It is worth noting that we combine appearance percentages of similar residues together, such as Q and N, D and E, Y and F (see Methods). We observe that critical residues have larger VC values than unimportant ones.

B. Critical residues identified by VC value are consistent with findings in the previous experiments

First, van der Waals interaction energies between several critical residues of the protein and the ligand are calculated (Supporting Material). The residues with higher than 1 kcal/mol van der Waals interaction energies (absolute value) can be roughly divided into four groups (with some overlaps). The first group includes F165, Q168, E169, E172 and W191, which interact with Arg-3. The second group consists of N190, W191, P204 and N206. They have strong interactions with Pro-1. F165, W191 and P204 also form the third group that interacts with Val0. N206 constitutes the fourth group along with F163 and Y207 that interacts with Pro+2.

In the first group, F165, Q168, E169, E172 and W191 have 37%, 10%, 22%, 44% and 95% appearance percentage, respectively, in our 207 sequences obtained from Psi-BLAST search in the SWISS-PROT database (Supporting Material). Their van der Waals interaction energies are –1.8, -3.0, -1.8, +1.2 and –7.2 kcal/mol respectively. W191 has the most favorable van der Waals interaction energy and is also well conserved. It is not surprising that no mutant examined experimentally at position W191 can bind with the poly-proline peptide[18]. At the F165 position, Phe and Tyr have 37% and 61% appearance percentages respectively. F165 forms the hydrophobic core of the binding pocket with Pro204, Trp191 and Tyr207. This implies that F165 is more crucial for stabilizing the receptor rather than for binding ligands. This is a possible explanation for the fact that no mutant (e.g. F165V, F165S, F165A and F165G) but F165L can bind to the peptide. This is presumably because only Leu among those examined residues can still stabilize the hydrophobic core. This tentative explanation will require experimental measurement of the stabilities of F165 mutants to be definitive. Q168 is on the surface and has not been studied experimentally either. Various residues appear at this position in different species, Asn, Gln and Glu in Crk, Grb2 and Src proteins respectively. This implies that Q168 is tolerant to mutations. E172 forms crucial hydrogen bonds with Arg-3 to keep Arg-3 in the right position to interact with W191 and Q168. Glu also appears at this position in other species, e.g. Grb2 proteins, to perform the same function. E169 seems to assist E172 to fix Arg-3 but not as crucial as E172 because its 22% appearance percentage is relatively low. In the previous study, double mutations at E169/E172 are shown to have a significant effect on binding[18].

Each residue in the second group, N190, W191, P204 and N206, has more than 2 kcal/mol favorable van der Waals interaction energy. W191, P204 and N206 are well conserved (100% and 73% appearance percentage respectively for P204 and N206) while N190 is not (16% appearance percentage). N190 is part of the binding pocket and interacts strongly with the site −1 residue. However different species have different residues at this position as well. For example, Crk, Grb2 and Src proteins have Gln, Asn and Asp residues respectively at this position. Our speculation is that this residue may be responsible for substrate specificity. As mentioned above, P204 is part of the hydrophobic core of the binding pocket and it was shown in the previous experiments to be crucial for the stability of the Sem-5 structure[18]. N206 interacts strongly with residues at site −1 and +2. It forms a hydrogen bond with the peptide backbone (Supporting Material). It may also be important for keeping crucial residue Y207 (see below) in the right position. This may explain why this residue is well conserved in different species. No mutations have been studied for Sem-5 N190 or N206.

In the fourth group, F163, N206 and Y207 have strong favorable van der Waals interaction energies, -5.0, -3.9 and −6.1 kcal/mol respectively. Their appearance percentages are 27%, 73% and 84% respectively. At the F163 position, although Phe is not the dominant residue (27% appearance percentage), Tyr, which also has a phenyl ring, has the highest appearance percentage, 63%. F163 forms one edge of the binding pocket and interacts with Pro at site +2. In the previous study, F163V mutant shows no binding with the peptide but F163A does[18]. It is worth pointing out that in the sequence alignment, Ala has a 4% conservation percentage at the F163 position. Thus, position 163 is critical for selectivity of the Sem-5 protein. However, it may be tolerant for a Tyr or Ala mutation with weaker binding. Ala has a smaller side chain and it allows the peptide to move closer to the pocket. Therefore, the lost interactions between the peptide and the protein due to F163A mutation may be recovered to some extent. However, Val keeps the peptide from approaching closer to the receptor and, thus, the lost interaction can not be recovered. At the Y207 position, Phe has a 12% appearance percentage, which may explain why only Y207F mutation does not disrupt the binding with the peptide as found previously[18].

It is also worth pointing out that several residues, L162, D164, D187, D188 and I202, shown to be unimportant for binding in reference 18 appear to have much weaker van der Waals interactions with the ligand (<0.5 kcal/mol).

In summary, residues with strong van der Waals interaction energy and well conserved, such as W191 and Y207, play significant roles in binding affinity and specificity. These are "hot spots" which are not tolerant to mutation. Residues with strong van der Waals interaction energy and diversified in different species determine specificity. F163 is an example here. If it is mutated to residues appearing in other species, specificity will be reduced. The binding probably will not be completely disrupted, i.e. the substrate probably still can bind, but with a weaker binding affinity. Residues with moderate van der Waals interaction energies but well conserved are crucial for stabilizing the protein, e.g. F165 and P204. Only those mutations which still can stabilize the protein are tolerated at these positions. Residues with weak van der Waals interaction energy and varied in different species usually are not important. From our studies, these observations appear in other systems as well (W. Wang and P. A. Kollman, unpublished data). One caveat is that charged residues forming strong hydrogen bonds with the ligand should be examined case by case, such as E172.

From Table 6, we can see that the VC value can identify (>=1.0) F163, F165, E172, W191, P204, N206 and Y207 as crucial residues (Table 6). Two residues, Q168 and E169, with strong van der Waals interaction energies but low conservation, and L162, and D164, with high conservation but weak van der Waals interaction energies, have low VC values (Table 6). This suggests the advantage of using VC value over only using van der Waals or conservation information. The significance of E172 may be underestimated as we point out above. Evaluating the significance of residues for the binding interactions in this way is being tested for other systems (W. Wang and P. A. Kollman, unpublished data) and the preliminary results are encouraging. We hope that this VC value can serve as a guide for mutagenesis experiments in the future. We can predict that the counterpart residues of F163, F165, E172, W191, P204, N206, and Y207 in other SH3 domains are most crucial for binding.

C. N190 is the possible residue critical for N-substituted recognition

The VC value profile leads us to engineer the receptor for tighter binding with site −1 Cα-substituted ligands. Since Ala-1 has more favorable van der Waals interactions with Q168 than Sar-1 and Y207 is well conserved, the only choices of mutation are N190 and N206, if we want to avoid mutating Y207, which might introduce significant conformational changes. Since N206 forms hydrogen bond with the peptide backbone, we do not want to disrupt it either. We first tried mutating N190 and N206 to the similar residue Gln. From Table 7, we can see that N190Q and N190Q/N206Q as calculated to bind more tightly with Ala substituted ligands than the wild type receptor, which is due to more favorable van der Waals interactions. However, these two mutants are also calculated to bind more tightly with Sar substituted ligands. Because Gln is similar to Asn, this may suggest that Gln also can maintain this selectivity. As a support of this interpretation, we observe that Gln occupies the N190 position in many sequences, especially in Crk proteins in the multiple sequence alignment. We next calculated two hydrophobic residues, Leu and Ile, at the 190 position. The N190L mutant binds more favorably with Sar-1 than Ala-1 as well. However, the N190I mutant has a very similar binding affinity with Sar-1 and Ala-1. This is because that I190 takes a conformation where the short branch of the side chain interacts with Ala-1 and the longer branch can have some favorable interactions with Pro-2 and Arg-3. In all other mutations including N190L, Ala-1 keeps residue 190 away from the peptide. These simulations also suggest that having Asn or Gln at position 190 is crucial for the specificity of N-substituted residues at site –1. If a suitable hydrophobic residue is in position 190 (which can be a non-natural amino acid), the selectivity for a N-substituted residue of site -1 can be reduced or even reversed. This observation is consistent with the fact that no hydrophobic residue appears at 190 and 206 positions in our sequence alignment. This suggestion awaits experimental examination.

4. Conclusions

We have presented here a combination of molecular dynamics, free energy calculations and sequence alignment to study interactions between Sem-5 SH3 domain and its ligands. These analyses shed lights on understanding SH3 domain-ligand interactions. We have shown that subtle conformational changes of the ligands due to whether they have N- or Cα-substituted residues is crucial for reproducing the relative binding free energies since the calculated ΔG's obtained from separate trajectories correlate much better with measured ones than those from a single trajectory. These conformational changes can also be seen from the Phi angle profile of different ligands.

It is also interesting that our results are consistent with experiment by using, in each of the separate trajectories, the ensemble of ligand conformations that exist in the complex. This suggests that these bound conformations are at least representative of the free ligand ensemble, which could not be assumed for small peptides. Perhaps the rigidity of the proline residues enables this to be a good approximation.

In this study, we also test different interior dielectric constants and charge models while applying the MM/PBSA method to estimate the ΔG of binding for flexible ligands. Different dielectric constants have been examined in both simulations where one considers different residues at various sites (site -1, 0, and +2) and where one considers many residues at the –1 site. Although results at different sites are similar for $\varepsilon=1$ and $\varepsilon=4$, $\varepsilon=4$ gives noticeably better results at site –1 than $\varepsilon=1$. A new charge model called AM1-BCC[16, 17], which can be calculated for any organic molecules much more efficiently than the RESP[28] charges that are used in the protein force field, has been shown to give comparable results to the RESP charges in the MM/PBSA calculations. It will thus be possible to use RESP charges for proteins and AM1-BCC charges for ligands, which will make the MM/PBSA method more efficient and generally applicable in structure based ligand design.

Discrimination of N- and Cα-substituted residues at different sites in the ligand is shown to be primarily due to van der Waals interactions between the ligand and the receptor. By calculating van der Waals contribution of each residue to binding and analyzing conservation at each position, we are able to identify several important residues of the receptor, most of which had been shown to be crucial for binding in prior experiments. Our analysis also suggested that mutation at N190 may reduce the selectivity for N- over Cα-substituted residues at site –1 and our free energy calculations further suggest that a specific mutation N190I may bind both type of peptides equally well. This prediction awaits experimental testing.

It was pointed out to us by a reviewer that in protein folding studies, common folding nucleus for several protein families were identified by looking at the number of contacts that certain amino acids make and how conserved they are[31,32]. This suggests that conserved residues with good intra- or inter-molecular packing are crucial for folding or binding. The VC value proposed in this study is the first quantitative parameter to combine energetic and evolutionary information. It thus might be useful for studying protein folding as well.

5. Acknowledgement

This work was supported in part by grant GM-29072 to Peter A. Kollman and GM55040-01 to Wendell A. Lim. Part of this investigation was completed using the facilities in the UCSF Computer Graphics Lab, T. Ferrin, director, supported by grant RR-1081 from the NIH. Supercomputer time at the NSF centers (NCSA and SDSC) is gratefully acknowledged.

6. References

(1) Kollman, P. *Chem Rev* **1993**, *93*, 2395-2417.

(2) Beveridge, D. L., Dicapua, F M *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431-492.

(3) van Gunsteren, W. F. In *Computer Simulation of Biomolecular Systems*; van Gunsteren, W. F., Weiner, P. K., Eds.; ESCOM: Leiden, 1989; pp 27-59.

(4) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. *J Amer Chem Soc* **1998**, *120*, 9401-9409.

(5) Massova, I.; Kollman, P. A. *J Amer Chem Soc* **1999**, *121*, 8133-8143.

(6) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J Amer Chem Soc* **1995**, *117*, 5179-5197.

(7) Case, D. A. *Curr Opin Struct Biol* **1994**, *4*, 285-290.

(8) Cohen, G. B.; Ren, R.; Baltimore, D. *Cell* **1995**, *80*, 237-248.

(9) Kuriyan, J.; Cowburn, D. *Annu. Rev. Biophys. Biomol. Struct.* **1997**, *26*, 259-288.

(10) Pawson, T.; Scott, J. D. *Science* **1997**, *278*, 2075-2080.

(11) Clark, S. G.; Stern, M. J.; Horvitz, H. R. *Nature* **1992**, *356*, 340-344.

(12) Simon, M. A.; Dodson, G. S.; Rubin, G. M. *Cell* **1993**, *73*, 169-177.

(13) Olivier, J. P.; et al. *Cell* **1993**, *73*, 179-191.

(14) Lim, W. A.; Fox, R. O.; Richards, F. M. *Protein Sci* **1994**, *3*, 1261-1266.

(15) Nguyen, J. T.; Turck, C. W.; Cohen, F. E.; Zuckermann, R. N.; Lim, W. A. *Science* **1998**, *282*, 2088-2092.

(16) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *J Comput Chem* **2000**, *21*, 132-146.

(17) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. *manuscript in preparation* **2000**.

(18) Lim, W. A.; Richards, F. M. *Nature Struct. Biol.* **1994**, *1*, 221-225.

(19) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; Debolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. A. *Comp. Phys. Comm.* **1995**, *91*, 1-41.

(20) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926-935.

(21) Ferrin, T. E.; Huang, C. C.; Jarvis, L. E.; Langridge, R. *J. Mol. Graphics* **1988**, *6*, 13-27.

(22) Darden, T. A.; York, D. M.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089-10092.

(23) Rychaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327-341.

(24) Kollman, P. A.; Massova, I.; Reyes, C. M.; Kuhn, B.; Huo, S.; Chong, L. T.; Lee, M. R.; Lee, T. S.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. *Account Chem Res* **2000**, *in press*.

(25) Gilson, M. K.; Sharp, K. A.; Honig, B. H. *J. Comput. Chem.* **1987**, *9*, 327-335.

(26) Wang, W.; Kollman, P. A. *J. Mol. Bol.* **2000**, *in press*.

(27) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 1978-1988.

(28) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J Phys Chem* **1993**, *97*, 10269-10280.

(29) Sanner, M. F.; Olson, A. J.; Spehner, J. C. *Biopolymers* **1996**, *38*, 305-320.

(30) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J. H.; Zhang, Z.; Miller, W.; Lipman, D. J. *Nucl Acid Res* **1997**, *25*, 3389-3402.

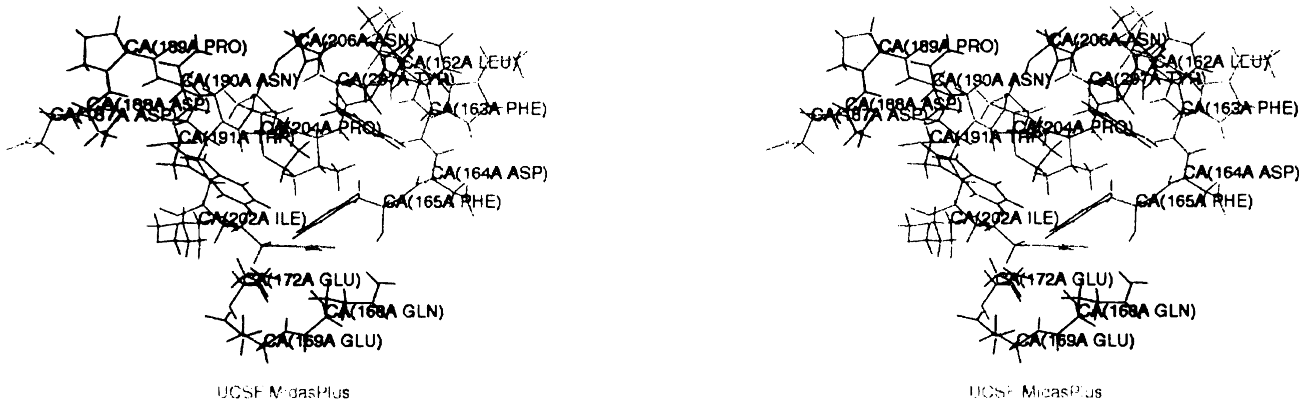
(31) Ptitsyn, O. B. *J Mol Biol* **1998**, *278*, 655-666.

(32) Ptitsyn, O. B.; Ting, K. L. H. *J Mol Biol* **1999**, *291*, 671-682.

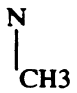
Figure legends.

Figure 1. (a) Binding sites of the Sem-5 SH3 domain and its ligands; (b) Side chains of N-substituted peptoids at site -1 of the ligand.

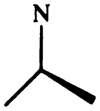
Figure 2 For site -1, 0 and +2, correlation between measured binding free energy and calculated free energy using RESP charges.


CA(189A PRO)
CA(206A ASN)
CA(162A LEU)
CA(190A ASN)
CA(207A TYR)
CA(188A ASP)
CA(187A ASP)
CA(163A PHE)
CA(204A PRO)
CA(191A THR)
CA(164A ASP)
CA(165A PHE)
CA(202A ILE)
CA(172A GLU)
CA(168A GLN)
CA(169A GLU)
UCSF MidasPlus
UCSF MidasPlus

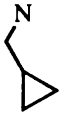
SAR



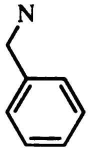
NIP



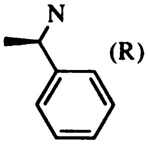
NMC



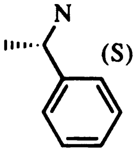
NBN



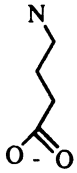
NRF



NSF



N2C



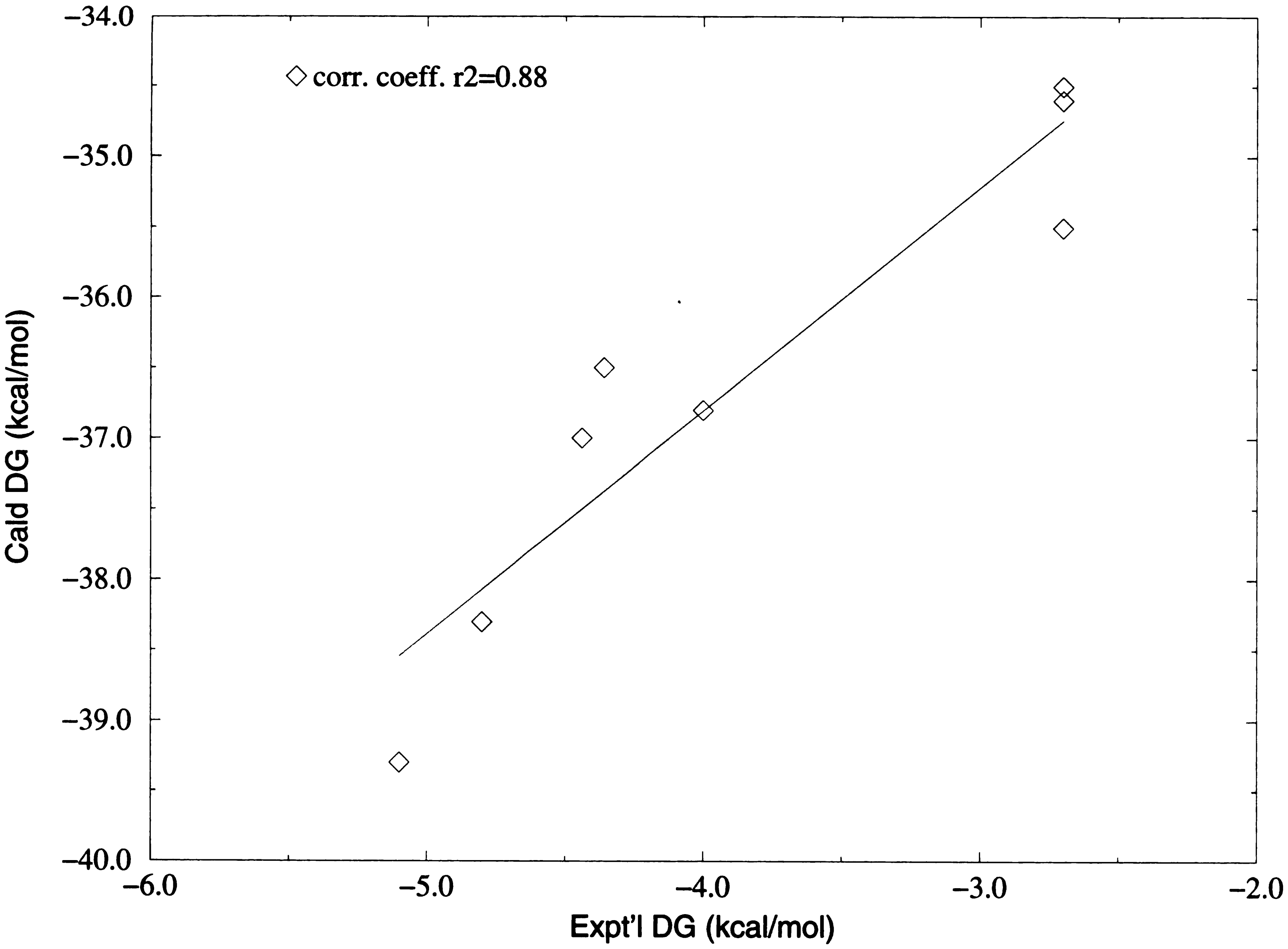

◇ corr. coeff. r2=0.88
Cald DG (kcal/mol)
−34.0
−35.0
−36.0
−37.0
−38.0
−39.0
−40.0
−6.0
−5.0
−4.0
−3.0
−2.0
Expt'l DG (kcal/mol)

Table 1. Binding free energies of Sem5 SH3 domain with its ligands (mutations at different sites)

ligand	Expt'l ΔG_b (kcal/mol)	ΔG_{int}^{vdw} (kcal/mol)	ΔG_{int}^{ele} (kcal/mol)	ΔG^{nonpol} (kcal/mol)	ΔG_{sol}^{ele} (kcal/mol)	$\Delta G_{int+sol}^{ele\S}$ (kcal/mol)	ΔG_b^{**} (kcal/mol)	$\Delta\Delta G_b^{**}$ (kcal/mol)
WT	-5.1	-37.4±0.4	-156.4±2.3	-3.7±0.0	+41.0±0.6	+1.9±0.0	-39.3±0.4	-
SAR-1[a]	-4.4	-35.1±0.5	-152.1±6.7	-3.6±0.0	+39.7±1.1	+1.7±0.6	-37.0±0.0	-1.5
ALA-1[a]	>-2.7	-33.4±0.7	-157.0±3.4	-3.6±0.0	+40.8±1.3	+1.6±0.4	-35.5±0.3	0.0
SAR0[b]	>-2.7	-32.2±0.7	-156.1±2.9	-3.6±0.0	+40.2±0.7	+1.2±0.0	-34.6±0.7	+2.2
ALA0[b]	-4.0	-34.3±0.6	-155.8±0.9	-3.7±0.0	+40.2±0.3	+1.3±0.1	-36.8±0.6	0.0
PRO0[b]	-4.8	-36.3±1.1	-156.4±2.3	-3.7±0.1	+40.7±0.2	+1.6±0.4	-38.3±0.8	-1.5
SAR+2[c]	-4.4	-34.4±0.0	-163.3±1.8	-3.6±0.0	+42.3±0.5	+1.5±0.0	-36.5±0.0	-2.0
ALA+2[c]	>-2.7	-32.2±0.1	-159.1±2.1	-3.4±0.0	+40.9±0.4	+1.1±0.1	-34.5±0.2	0.0

§ $\Delta G_{int+sol}^{ele} = \Delta G_{int}^{ele} + \Delta G_{sol}^{ele}$ ** $\Delta G_b = \Delta G_{int}^{vdw} + \Delta G_{int}^{ele} (\varepsilon_{in}=1, \varepsilon_{out}=1)/4 + \Delta G^{nonpol} + \Delta G_{sol}^{ele}$ *** $\Delta\Delta G_b$ is calculated for each site

a. SAR-1 and ALA-1 refer to mutating Pro at site –1 to Sar and Ala respectively; b. SAR0, ALA0 and PRO0 refer to mutating Val at site 0 to Sar, Ala and Pro respectively; c. SAR+2 and ALA+2 refer to mutating Pro at site +2 to Sar and Ala respectively.

Table 2. Correlation coefficients r^2 between measured binding free energies and different components of calculated ones.

component	Correlation coefficient r^2
van der Waals	0.88
electrostatic (Coloumb term)	0.0088
solvation penalty (PB term)	0.026
electrostatic + solvation penalty	0.52
SA	0.45

Table 3. Binding free energies of Sem5 SH3 domain with its ligands obtained from computational mutagenesis using the wild type peptide trajectory

Ligand	Expt'l ΔG_b (kcal/mol)	ΔG_{int}^{vdw} (kcal/mol)	ΔG_{int}^{ele} (kcal/mol)	ΔG^{nonpol} (kcal/mol)	ΔG_{sol}^{ele} (kcal/mol)	$\Delta G_{int+sol}^{ele\S}$ (kcal/mol)	ΔG_b^{**} (kcal/mol)	$\Delta\Delta G_b^{***}$ (kcal/mol)
WT	-5.1	-37.4±0.4	-156.4±2.3	-3.7±0.0	+41.0±0.6	+1.9±0.0	-39.3±0.4	-
SAR-1[a]	-4.4	-33.9±0.5	-155.8±2.1	-3.6±0.0	+40.8±0.4	+1.8±0.1	-35.7±0.4	-0.2
ALA-1[a]	>-2.7	-33.9±0.4	-156.4±2.1	-3.7±0.0	+41.1±0.6	+2.0±0.0	-35.5±0.5	0.0
SAR0[b]	>-2.7	-35.1±0.3	-155.5±2.3	-3.7±0.0	+40.5±0.6	+1.6±0.0	-37.2±0.3	+0.3
ALA0[b]	-4.0	-35.6±0.3	-155.9±2.3	-3.7±0.0	+40.8±0.6	+1.8±0.0	-37.5±0.3	0.0
SAR+2[c]	-4.4	-35.1±0.1	-156.6±2.4	-3.6±0.0	+41.1±0.5	+2.0±0.1	-36.7±0.2	-0.8
ALA+2[c]	>-2.7	-34.5±0.1	-156.2±2.4	-3.7±0.0	+41.3±0.6	+2.2±0.0	-35.9±0.1	0.0

§ $\Delta G_{int+sol}^{ele} = \Delta G_{int}^{ele} + \Delta G_{sol}^{ele}$ ** $\Delta G_b = \Delta G_{int}^{vdw} + \Delta G_{int}^{ele}(\varepsilon_{in}=1, \varepsilon_{out}=1)/4 + \Delta G^{nonpol} + \Delta G_{sol}^{ele}$ *** $\Delta\Delta G_b$ is calculated for each site

a. SAR-1 and ALA-1 refer to mutating Pro at site –1 to Sar and Ala respectively; b. SAR0, ALA0 and PRO0 refer to mutating Val at site 0 to Sar, Ala and Pro respectively; c. SAR+2 and ALA+2 refer to mutating Pro at site +2 to Sar and Ala respectively

Table 4. Binding free energies of Sem-5 SH3 domain with site –1 mutant ligands

ligand	Expt'l ΔG_b (kcal/mol)	ΔG_{int}^{vdw} (kcal/mol)	ΔG_{int}^{ele} (kcal/mol)	ΔG^{nonpol} (kcal/mol)	ΔG_{sol}^{ele} (kcal/mol)	$\Delta G_{int+sol}^{ele§}$ (kcal/mol)	ΔG_b^{**} (kcal/mol)	$\Delta\Delta G_b^{***}$ (kcal/mol)
ALA	>-2.7	-33.4±0.7	-157.0±3.4	-3.6±0.0	+40.8±1.3	+1.6±0.4	-35.5±0.3	0.0
N2C	-3.48	-38.3±0.7	-115.6±2.5	-4.0±0.0	+32.4±0.4	+3.5±0.4	-38.7±0.5	-3.2
NRF	-4.32	-34.3±1.5	-135.6±5.2	-4.4±0.6	+34.6±1.1	+0.7±0.2	-38.0±1.0	-2.5
SAR	-5.45	-35.1±0.5	-152.1±6.7	-3.6±0.0	+39.7±1.1	+1.7±0.6	-37.0±0.0	-1.5
NSF	-5.82	-37.4±1.2	-162.7±0.4	-3.8±0.1	+41.6±0.5	+0.9±0.4	-40.3±0.8	-4.8
WT	-5.89	-37.4±0.4	-156.4±2.3	-3.7±0.0	+41.0±0.6	+1.9±0.0	-39.2±0.4	-3.7
NMC	-5.97	-37.7±0.7	-156.8±2.5	-3.7±0.0	+40.6±0.5	+1.4±0.1	-40.0±0.6	-4.5
NIP	-6.27	-37.7±0.2	-155.5±3.0	-3.9±0.0	+41.2±0.5	+2.3±0.2	-39.3±0.1	-3.8
NBN	-6.32	-41.1±1.1	-157.0±4.3	-4.0±0.1	+42.2±0.1	+2.9±0.0	-42.2±1.2	-6.7

§ $\Delta G_{int+sol}^{ele} = \Delta G_{int}^{ele} + \Delta G_{sol}^{ele}$ ** $\Delta G_b = \Delta G_{int}^{vdw} + \Delta G_{int}^{ele}(\varepsilon_{in}=1, \varepsilon_{out}=1)/4 + \Delta G^{nonpol} + \Delta G_{sol}^{ele}$ *** $\Delta\Delta G_b$ is relative to ALA

Table 5. Binding free energies of Sem-5 SH3 domain with site –1 mutant ligands using AM1-BCC charges

ligand	Expt'l ΔG_b (kcal/mol)	ΔG_{int}^{vdw} (kcal/mol)	ΔG_{int}^{ele} (kcal/mol)	ΔG^{nonpol} (kcal/mol)	ΔG_{sol}^{ele} (kcal/mol)	$\Delta G_{int+sol}^{ele§}$ (kcal/mol)	ΔG_b^{**} (kcal/mol)	$\Delta\Delta G_b^{***}$ (kcal/mol)
ALA	>-2.7	-33.4±0.7	-157.0±3.4	-3.6±0.0	+40.8±1.3	+1.6±0.4	-35.5±0.3	0.0
N2C	-3.48	-38.3±0.7	-112.3±2.6	-4.0±0.0	+32.1±0.5	+4.0±0.2	-38.2±0.6	-2.7
NRF	-4.32	-34.3±1.5	-135.2±4.5	-4.4±0.6	+35.8±1.0	+2.0±0.1	-36.7±0.9	-1.2
SAR	-5.45	-35.1±0.5	-147.7±7.6	-3.6±0.0	+39.6±1.2	+2.7±0.7	-36.0±0.1	-0.5
NSF	-5.82	-37.4±1.2	-157.2±0.0	-3.8±0.1	+41.8±0.5	+2.5±0.4	-38.7±0.8	-3.2
WT	-5.89	-37.4±0.4	-156.4±2.3	-3.7±0.0	+41.0±0.6	+1.9±0.0	-39.2±0.4	-3.7
NMC	-5.97	-37.7±0.7	-156.3±2.2	-3.7±0.0	+41.0±0.5	+1.9±0.1	-39.5±0.7	-4.0
NIP	-6.27	-37.7±0.2	-147.7±3.4	-3.9±0.0	+39.9±0.6	+3.0±0.3	-38.6±0.0	-3.1
NBN	-6.32	-41.1±1.1	-158.3±4.0	-4.0±0.1	+43.8±1.0	+4.3±0.1	-40.9±1.2	-5.4

§ $\Delta G_{int+sol}^{ele} = \Delta G_{int}^{ele} + \Delta G_{sol}^{ele}$ ** $\Delta G_b = \Delta G_{int}^{vdw} + \Delta G_{int}^{ele}(\varepsilon_{in}=1, \varepsilon_{out}=1)/4 + \Delta G^{nonpol} + \Delta G_{sol}^{ele}$ *** $\Delta\Delta G_b$ is relative to ALA

Table 6. Critical residues for binding have larger VC values than unimportant ones.

Residue	Substitution Sensitivity (Expt'l data)	VC value	van der Waals (kcal/mol)	Conservation Percentage (%)
F163	Yes	4.6	-5.0	91
F165	Yes	1.8	-1.8	98
E169/E172	Yes	0.6/1.0	-1.8/+1.2	33/80
W191	Yes	6.8	-7.2	95
P204	Yes	2.4	-2.4	100
N206	Not studied	2.9	-3.9	73
Y207	Yes	5.9	-6.1	96
L162	No	0.3	-0.4	67
D164	No	0.7	-0.9	80
Q168	Not studied	0.5	-3.0	16
D187/D188	No	0.0/0.1	-0.02/-0.2	12/54
N190	Not studied	0.7	-2.8	24
I202	No	0.1	-0.3	38

Table 7. Sar-1 and Ala-1 ligands interact with mutant proteins

ligand	Expt'l ΔG_b (kcal/mol)	ΔG_{int}^{vdw} (kcal/mol)	ΔG_{int}^{ele} (kcal/mol)	ΔG^{nonpol} (kcal/mol)	ΔG_{sol}^{ele} (kcal/mol)	$\Delta G_{int+sol}^{ele§}$ (kcal/mol)	ΔG_b^{**} (kcal/mol)	$\Delta\Delta G_b^{***}$ (kcal/mol)
ALA-1/ N190Q	N/A	-36.0±0.4	-143.1±0.7	-3.8±0.0	+38.8±0.3	+3.0±0.1	-36.8±0.5	0.0
SAR-1/ N190Q	N/A	-37.1±0.5	-170.8±3.6	-4.0±0.0	+44.9±0.6	+2.2±0.3	-38.9±0.2	-2.1
ALA-1/ N190Q/ N206Q	N/A	-34.3±0.9	-158.5±3.7	-3.7±0.0	+41.4±0.8	+1.7±0.1	-36.2±1.0	0.0
SAR-1/ N190Q/ N206Q	N/A	-37.3±0.1	-182.1±3.0	-3.8±0.0	+46.9±0.8	+1.4±0.0	-39.6±0.2	-3.4
ALA-1/ N190I	N/A	-37.2±1.0	-140.5±1.8	-3.8±0.1	+37.5±0.7	+2.4±0.2	-38.6±0.8	0.0
SAR-1/ N190I	N/A	-36.4±0.1	-147.5±1.6	-3.8±0.0	+38.4±0.3	+1.6±0.1	-38.6±0.1	0.0
ALA-1/ N190L	N/A	-32.9±1.0	-175.4±2.3	-3.7±0.1	+45.1±0.5	+1.2±0.0	-35.4±1.0	0.0
SAR-1/ N190L	N/A	-39.2±0.6	-161.0±2.2	-4.1±0.0	+43.4±0.2	+3.1±0.7	-40.2±0.1	-4.8

§ $\Delta G_{int+sol}^{ele} = \Delta G_{int}^{ele} + \Delta G_{sol}^{ele}$ ** $\Delta G_b = \Delta G_{int}^{vdw} + \Delta G_{int}^{ele}(\varepsilon_{in}=1, \varepsilon_{out}=1)/4 + \Delta G^{nonpol} + \Delta G_{sol}^{ele}$ *** $\Delta\Delta G_b$ is calculated for each type of mutant receptor

Supporting Material

Table 1. Binding free energies of Sem5 SH3 domain with its ligands (mutations at different sites) using different interior dielectric constants

ligands	Expt'l ΔG_b (kcal/mol)	ΔG_{int}^{vdw} (kcal/mol)	ΔG_{int}^{ele} (kcal/mol)	ΔG^{nonpol} (kcal/mol)	ε_{in}=1, ε_{out}=80			ε_{in}=4, ε_{out}=80		
					ΔG_{sol}^{ele} (kcal/mol)	$\Delta G_{int+sol}^{ele§}$ (kcal/mol)	ΔG_b^{*} (kcal/mol)	ΔG_{sol}^{ele} (kcal/mol)	$\Delta G_{int+sol}^{ele§}$ (kcal/mol)	ΔG_b^{**} (kcal/mol)
WT	-5.1	-37.4±0.4	-156.4 ±2.3	-3.7±0.0	+171.5±2.3	+15.1±0.0	-26.1±0.4	+41.0±0.6	+1.9±0.0	-39.3±0.4
SAR-1[a]	-4.4	-35.1±0.5	-152.1 ±6.7	-3.6±0.0	+165.8±4.7	+13.7±2.0	-25.0±1.4	+39.7±1.1	+1.7±0.6	-37.0±0.0
ALA-1[a]	>-2.7	-33.4±0.7	-157.0 ±3.4	-3.6±0.0	+170.6±5.4	+13.6±1.9	-23.5±1.3	+40.8±1.3	+1.6±0.4	-35.5±0.3
SAR0[b]	>-2.7	-32.2±0.7	-156.1 ±2.9	-3.6±0.0	+168.2±2.9	+12.1±0.0	-23.7±0.7	+40.2±0.7	+1.2±0.0	-34.6±0.7
ALA0[b]	-4.0	-34.3±0.6	-155.8 ±0.9	-3.7±0.0	+168.5±1.5	+12.7±0.6	-25.3±0.1	+40.2±0.3	+1.3±0.1	-36.8±0.6
PRO0[b]	-4.8	-36.3±1.1	-156.4 ±2.3	-3.7±0.1	+170.8±0.6	+14.3±1.7	-25.6±0.6	+40.7±0.2	+1.6±0.4	-38.3±0.8
SAR+2[c]	-4.4	-34.4±0.0	-163.3 ±1.8	-3.6±0.0	+176.8±2.0	+13.6±0.2	-24.5±0.3	+42.3±0.5	+1.5±0.0	-36.5±0.0
ALA+2[c]	>-2.7	-32.2±0.1	-159.1 ±2.1	-3.4±0.0	+171.5±1.6	+12.4±0.5	-23.2±0.6	+40.9±0.4	+1.1±0.1	-34.5±0.2

Table 2. Binding free energies of Sem5 SH3 domain with its ligands obtained from computational mutagenesis using wild type peptide trajectory with different interior dielectric constant

ligand	Expt'l ΔG_b (kcal/mol)	ΔG_{int}^{vdw} (kcal/mol)	ΔG_{int}^{ele} (kcal/mol)	ΔG^{nonpol} (kcal/mol)	$\varepsilon_{in}=1$, $\varepsilon_{out}=80$			$\varepsilon_{in}=4$, $\varepsilon_{out}=80$		
					ΔG_{sol}^{ele} (kcal/mol)	$\Delta G_{int+sol}^{ele\S}$ (kcal/mol)	ΔG_b^{*} (kcal/mol)	ΔG_{sol}^{ele} (kcal/mol)	$\Delta G_{int+sol}^{ele\S}$ (kcal/mol)	ΔG_b^{**} (kcal/mol)
WT	-5.1	-37.4±0.4	-156.4 ±2.3	-3.7±0.0	+171.5±2.3	+15.1±0.0	-26.1±0.4	+41.0±0.6	+1.9±0.0	-39.3±0.4
SAR-1	-4.4	-33.9±0.5	-155.8 ±2.1	-3.6±0.0	+168.3±3.9	+12.4±1.8	-25.1±2.3	+40.8±0.4	+1.8±0.1	-35.7±0.4
ALA-1	>-2.7	-33.9±0.4	-156.4 ±2.1	-3.7±0.0	+172.0±2.2	+15.6±0.1	-21.9±0.5	+41.1±0.6	+2.0±0.0	-35.5±0.5
SAR0	>-2.7	-35.1±0.3	-155.5 ±2.3	-3.7±0.0	+169.3±2.3	+13.8±0.0	-25.1±0.2	+40.5±0.6	+1.6±0.0	-37.2±0.3
ALA0	-4.0	-35.6±0.3	-155.9 ±2.3	-3.7±0.0	+170.4±2.4	+14.5±0.0	-24.7±0.3	+40.8±0.6	+1.8±0.0	-37.5±0.3
SAR+2	-4.4	-35.1±0.1	-156.6 ±2.4	-3.6±0.0	+170.0±4.0	+13.4±1.6	-25.3±1.8	+41.1±0.5	+2.0±0.1	-36.7±0.2
ALA+2	>-2.7	-34.5±0.1	-156.2 ±2.4	-3.7±0.0	+172.7±2.4	+16.5±0.1	-21.6±0.2	+41.3±0.6	+2.2±0.0	-35.9±0.1

Table 3. Binding free energies of Sem-5 SH3 domain with site –1 mutant ligands using different interior dielectric constants

ligand	Expt'l ΔG_b (kcal/mol)	ΔG_{int}^{vdw} (kcal/mol)	ΔG_{int}^{ele} (kcal/mol)	ΔG^{nonpol} (kcal/mol)	$\varepsilon_{in}=1$, $\varepsilon_{out}=80$			$\varepsilon_{in}=4$, $\varepsilon_{out}=80$		
					ΔG_{sol}^{ele} (kcal/mol)	$\Delta G_{int+sol}^{ele§}$ (kcal/mol)	ΔG_b^{*} (kcal/mol)	ΔG_{sol}^{ele} (kcal/mol)	$\Delta G_{int+sol}^{ele§}$ (kcal/mol)	ΔG_b^{**} (kcal/mol)
ALA	>-2.7	-33.4±0.7	-157.0±3.4	-3.6±0.0	+170.6±5.4	+13.6±1.9	-23.5±1.3	+40.8±1.3	+1.6±0.4	-35.5±0.3
N2C	-3.48	-38.3±0.7	-115.6±2.5	-4.0±0.0	+136.2±1.7	+20.6±0.8	-21.6±0.0	+32.4±0.4	+3.5±0.4	-38.7±0.5
NRF	-4.32	-34.3±1.5	-135.6±5.2	-4.4±0.6	+147.8±5.4	+12.2±0.1	-26.5±0.7	+34.6±1.1	+0.7±0.2	-38.0±1.0
SAR	-5.45	-35.1±0.5	-152.1±6.7	-3.6±0.0	+165.8±4.7	+13.7±2.0	-25.0±1.4	+39.7±1.1	+1.7±0.6	-37.0±0.0
NSF	-5.82	-37.4±1.2	-162.7±0.4	-3.8±0.1	+178.8±2.5	+16.1±2.2	-25.2±0.9	+41.6±0.5	+0.9±0.4	-40.3±0.8
WT	-5.89	-37.4±0.4	-156.4±2.3	-3.7±0.0	+171.5±2.3	+15.1±0.0	-26.1±0.4	+41.0±0.6	+1.9±0.0	-39.2±0.4
NMC	-5.97	-37.7±0.7	-156.8±2.5	-3.7±0.0	+170.6±2.3	+13.8±0.2	-27.6±0.6	+40.6±0.5	+1.4±0.1	-40.0±0.6
NIP	-6.27	-37.7±0.2	-155.5±3.0	-3.9±0.0	+172.2±2.3	+16.7±0.7	-24.9±0.4	+41.2±0.5	+2.3±0.2	-39.3±0.1
NBN	-6.32	-41.1±1.1	-157.0±4.3	-4.0±0.1	+177.0±4.7	+19.9±0.4	-25.2±0.8	+42.2±1.1	+2.9±0.0	-42.2±1.2

Table 4. Binding free energies of Sem-5 SH3 domain with site –1 mutant ligands using AM1-BCC charges using different interior dielectric constants

ligand	Expt'l ΔG_b (kcal/mol)	ΔG_{int}^{vdw} (kcal/mol)	ΔG_{int}^{ele} (kcal/mol)	ΔG^{nonpol} (kcal/mol)	$\varepsilon_{in}=1$, $\varepsilon_{out}=80$			$\varepsilon_{in}=4$, $\varepsilon_{out}=80$		
					ΔG_{sol}^{ele} (kcal/mol)	$\Delta G_{int+sol}^{ele\S}$ (kcal/mol)	ΔG_b^{*} (kcal/mol)	ΔG_{sol}^{ele} (kcal/mol)	$\Delta G_{int+sol}^{ele\S}$ (kcal/mol)	ΔG_b^{**} (kcal/mol)
ALA	>-2.7	-33.4±0.7	-157.0±3.4	-3.6±0.0	+170.6±5.4	+13.6±1.9	-23.5±1.3	+40.8±1.3	+1.6±0.4	-35.5±0.3
N2C	-3.48	-38.3±0.7	-112.3±2.6	-4.0±0.0	+134.8±1.9	+22.5±0.6	-19.7±0.1	+32.1±0.5	+4.0±0.2	-38.2±0.6
NRF	-4.32	-34.3±1.5	-135.2±4.5	-4.4±0.6	+150.3±4.3	+15.1±0.2	-23.6±1.1	+35.8±1.0	+2.0±0.1	-36.7±0.9
SAR	-5.45	-35.1±0.5	-147.7±7.6	-3.6±0.0	+165.2±5.3	+17.5±2.3	-21.2±1.7	+39.6±1.2	+2.7±0.7	-36.0±0.1
NSF	-5.82	-37.4±1.2	-157.2±0.0	-3.8±0.1	+175.6±1.7	+18.4±1.6	-22.8±0.3	+41.8±0.5	+2.5±0.4	-38.7±0.8
WT	-5.89	-37.4±0.4	-156.4±2.3	-3.7±0.0	+171.5±2.3	+15.1±0.0	-26.1±0.4	+41.0±0.6	+1.9±0.0	-39.2±0.4
NMC	-5.97	-37.7±0.7	-156.3±2.2	-3.7±0.0	+172.1±2.1	+15.8±0.1	-25.6±0.6	+41.0±0.5	+1.9±0.1	-39.5±0.7
NIP	-6.27	-37.7±0.2	-147.7±3.4	-3.9±0.0	+166.8±2.5	+19.1±0.9	-22.4±0.7	+39.9±0.6	+3.0±0.3	-38.6±0.0
NBN	-6.32	-41.1±1.1	-158.3±4.0	-4.0±0.1	+183.7±4.3	+25.5±0.3	-19.6±0.9	+43.8±1.0	+4.3±0.1	-40.9±1.2

Table 5. Sar-1 and Ala-1 ligands interact with mutant proteins using different interior dielectric constants

ligands	Expt'l ΔG_b (kcal/mol)	ΔG_{int}^{vdw} (kcal/mol)	ΔG_{int}^{ele} (kcal/mol)	ΔG^{nonpol} (kcal/mol)	$\varepsilon_{in}=1$, $\varepsilon_{out}=80$			$\varepsilon_{in}=4$, $\varepsilon_{out}=80$		
					ΔG_{sol}^{ele} (kcal/mol)	$\Delta G_{int+sol}^{ele\S}$ (kcal/mol)	ΔG_b^{*} (kcal/mol)	ΔG_{sol}^{ele} (kcal/mol)	$\Delta G_{int+sol}^{ele\S}$ (kcal/mol)	ΔG_b^{**} (kcal/mol)
ALA-1/ N190Q	N/A	-36.0±0.4	-143.1 ±0.7	-3.8±0.0	+162.2±1.0	+19.0±0.3	-20.8±0.8	+38.8±0.3	+3.0±0.1	-36.8±0.5
SAR-1/ N190Q	N/A	-37.1±0.5	-170.8 ±3.6	-40.0±0.0	-188.2±2.3	+17.4±1.3	-23.6±0.8	+44.9±0.6	+2.2±0.3	-38.9±0.2
ALA-1/ N190Q/ N206Q	N/A	-34.3±0.9	-158.5 ±3.7	-3.7±0.0	+173.2±3.4	+14.7±0.4	-23.2±1.3	+41.4±0.8	+1.7±0.1	-36.2±1.0
SAR-1/ N190Q/ N206Q	N/A	-37.3±0.1	-182.1 ±3.0	-3.8±0.0	+197.2±3.3	+15.1±0.3	-25.9±0.5	+46.9±0.8	+1.4±0.0	-39.6±0.2
ALA-1/ N190I	N/A	-37.2±1.0	-140.5 ±1.8	-3.8±0.1	+157.3±2.9	+16.8±1.1	-24.2±0.1	+37.5±0.7	+2.4±0.2	-38.6±0.8
SAR-1/ N190I	N/A	-36.4±0.1	-147.5 ±1.6	-3.8±0.0	+161.1±1.1	+13.7±0.5	-26.5±0.5	+38.4±0.8	+1.6±0.1	-38.6±0.1
ALA-1/ N190L	N/A	-32.9±1.0	-175.4 ±2.3	-3.7±0.1	+189.3±2.3	+14.0±0.1	-22.6±1.1	+45.1±0.5	+1.2±0.0	-35.4±1.0
SAR-1/ N190L	N/A	-39.2±0.6	-161.0 ±2.2	-4.0±0.0	+182.2±0.6	+21.2±2.8	-22.1±2.1	+43.4±0.2	+3.1±0.7	-40.2±0.1

Table 6. van der Waals interaction energies (kcal/mol) between each residue of the peptide/peptoid and the receptor.

	WT	SAR +2	SAR 0	SAR –1	ALA +2	ALA 0	ALA –1
Site +4	-0.9±0.1	-1.4±0.1	-1.1±0.2	-0.8±0.2	-0.8±0.0	-1.2±0.2	-1.0±0.2
Site +3	-5.2±0.1	-5.9±0.1	-5.3±0.3	-4.5±0.6	-4.7±0.1	-5.6±0.1	-5.1±0.1
Site +2	-6.0±0.2	-3.8±0.4	-5.6±0.5	-6.9±0.0	-3.9±0.2	-5.9±0.3	-5.9±0.4
Site +1	-2.2±0.1	-2.0±0.1	-1.8±0.1	-2.0±0.3	-2.2±0.0	-1.9±0.0	-1.8±0.0
Site 0	-4.8±0.1	-5.3±0.1	-2.4±0.2	-5.7±0.2	-5.5±0.2	-3.4±0.1	-6.0±0.0
Site -1	-8.0±0.2	-7.6±0.1	-7.1±0.0	-5.4±0.0	-7.0±0.0	-7.3±0.2	-4.4±0.3
Site -2	-2.1±0.2	-1.8±0.1	-1.6±0.2	-2.8±0.2	-1.4±0.1	-1.6±0.0	-1.9±0.1
Site -3	-8.2±0.2	-6.5±0.0	-7.2±0.1	-7.1±0.2	-6.7±0.3	-7.4±0.1	-7.2±0.1

Table 7. van der Waals interaction energies (kcal/mol) between each residue of the peptide/peptoid and the mutant receptors.

	WT	SAR –1	ALA –1	ALA-1 N190Q	SAR-1 N190Q	ALA-1 N190Q N206Q	SAR-1 N190Q N206Q	ALA-1 N190I	SAR-1 N190I	ALA-1 N190L	SAR-1 N190L
Site +4	-0.9±0.1	-0.8±0.2	-1.0±0.2	-1.2±0.1	-1.4±0.0	-1.0±0.0	-0.8±0.1	-2.0±0.2	-1.2±0.0	-1.2±0.1	-1.1±0.0
Site +3	-5.2±0.1	-4.5±0.6	-5.1±0.1	-4.9±0.2	-5.9±0.2	-5.6±0.2	-4.2±0.0	-5.2±0.1	-5.5±0.1	-5.0±0.1	-5.2±0.2
Site +2	-6.0±0.2	-6.9±0.0	-5.9±0.4	-6.3±0.2	-6.2±0.9	-6.1±0.8	-7.1±0.2	-8.0±0.0	-6.1±0.8	-7.3±0.6	-7.0±0.1
Site +1	-2.2±0.1	-2.0±0.3	-1.8±0.0	-1.3±0.1	-1.8±0.1	-2.5±0.6	-3.2±0.1	-1.9±0.1	-1.7±0.1	-1.6±0.1	-1.3±0.1
Site 0	-4.8±0.1	-5.7±0.2	-6.0±0.0	-6.2±0.2	-5.8±0.4	-6.6±0.0	-6.5±0.1	-5.4±0.1	-5.7±0.2	-5.5±0.0	-5.4±0.0
Site -1	-8.0±0.2	-5.4±0.0	-4.4±0.3	-5.7±0.1	-4.8±0.0	-3.8±0.0	-6.4±0.1	-3.3±0.4	-4.4±0.1	-3.7±0.5	-4.2±0.1
Site -2	-2.1±0.2	-2.8±0.2	-1.9±0.1	-3.0±0.1	-2.6±0.1	-1.6±0.0	-2.8±0.1	-1.9±0.2	-2.5±0.2	-2.2±0.0	-3.8±0.2
Site -3	-8.2±0.2	-7.1±0.2	-7.2±0.1	-7.3±0.1	-8.7±0.2	-7.2±0.1	-6.3±0.1	-9.4±0.3	-9.2±0.3	-6.4±0.3	-11.2±0.4

Table 8. van der Waals interaction energies (kcal/mol) between residues on the receptor and the ligand

Residue	WT	SAR-1	ALA-1	SAR 0	ALA 0	SAR +2	ALA +2
L162	-0.4±0.0	-0.3±0.0	-0.3±0.0	-0.4±0.0	-0.4±0.0	-0.3±0.0	-0.3±0.0
F163	-5.0±0.1	-4.8±0.5	-4.9±0.1	-5.4±0.2	-5.6±0.2	-5.4±0.1	-4.9±0.0
D164	-0.9±0.1	-0.5±0.0	-0.8±0.1	-0.6±0.2	-0.8±0.0	-0.5±0.1	-0.6±0.0
F165	-1.8±0.0	-1.9±0.1	-1.9±0.1	-1.3±0.0	-1.3±0.0	-1.7±0.1	-1.8±0.0
N166	-0.4±0.0	-0.5±0.1	-0.6±0.0	-0.5±0.1	-0.4±0.0	-0.4±0.0	-0.5±0.1
P167	-0.6±0.0	-0.7±0.0	-0.7±0.0	-0.8±0.1	-0.6±0.3	-0.6±0.0	-0.5±0.0
Q168	-3.0±0.4	-1.9±0.1	-2.8±0.4	-2.2±0.1	-2.7±0.3	-1.9±0.0	-1.8±0.0
E169	-1.8±0.0	-2.1±0.0	-2.2±0.0	-1.8±0.5	-1.2±0.2	-1.8±0.1	-1.8±0.0
E172	+1.2±0.2	+1.8±0.0	+1.8±0.0	+1.5±0.0	+1.7±0.0	+2.1±0.0	+1.9±0.6
N190	-2.8±0.2	-3.2±0.0	-2.0±0.1	-2.2±0.2	-2.3±0.1	-2.5±0.1	-2.1±0.0
W191	-7.2±0.2	-6.7±0.1	-6.4±0.0	-5.7±0.3	-6.8±0.1	-7.3±0.1	-7.0±0.1
P204	-2.4±0.0	-2.2±0.1	-2.5±0.2	-1.8±0.0	-2.2±0.0	-2.4±0.0	-2.5±0.1
S205	-0.5±0.0	-0.3±0.0	-0.3±0.0	-0.4±0.0	-0.4±0.0	-0.4±0.0	-0.4±0.0
N206	-3.9±0.1	-3.5±0.0	-2.8±0.1	-3.7±0.0	-3.9±0.3	-3.4±0.0	-3.1±0.1
Y207	-6.1±0.3	-6.4±0.1	-5.5±0.2	-5.4±0.2	-5.6±0.3	-5.6±0.1	-4.9±0.2

Table 9. van der Waals interaction energies (kcal/mol) between residues on the mutant receptors and the ligand.

Residue	ALA –1	SAR-1	ALA –1	SAR-1	ALA-1	SAR-1	ALA-1	SAR-1
	N190Q	N190Q	N190Q	N190Q	N190I	N190I	N190L	N190L
			N206Q	N206Q				
L162	-0.5±0.1	-0.5±0.1	-0.3±0.0	-0.3±0.1	-1.3±0.0	-0.6±0.2	-0.6±0.2	-0.4±0.0
F163	-5.5±0.2	-5.8±0.0	-4.9±0.4	-4.7±0.1	-6.2±0.0	-5.7±0.0	-5.6±0.1	-5.4±0.1
D164	-0.4±0.0	-1.0±0.2	-0.4±0.1	-0.4±0.0	-0.6±0.0	-0.5±0.1	-0.3±0.0	-0.3±0.0
F165	-1.8±0.0	-1.8±0.2	-1.5±0.1	-1.6±0.2	-2.0±0.0	-2.0±0.0	-1.6±0.0	-1.3±0.1
N166	-0.5±0.0	-0.5±0.1	-0.3±0.0	-0.3±0.0	-0.5±0.0	-0.5±0.1	-0.4±0.0	-0.2±0.1
P167	-0.5±0.0	-0.5±0.0	-0.6±0.0	-0.6±0.0	-0.5±0.0	-0.7±0.1	-0.6±0.0	-0.3±0.1
Q168	-3.2±0.1	-2.8±0.1	-2.1±0.0	-1.3±0.4	-1.9±0.0	-2.2±0.0	-1.0±0.3	-1.1±0.2
E169	-0.4±0.0	-2.6±0.1	-1.8±0.4	-2.4±0.1	-3.0±0.0	-2.9±0.2	-1.8±0.3	-2.8±0.1
E172	+1.5±0.1	+1.5±0.2	+1.7±0.1	+1.6±0.4	+1.6±0.0	+1.8±0.2	+1.7±0.0	+1.1±0.5
N190	-3.2±0.0	-2.9±0.1	-3.2±0.2	-3.8±0.0	-1.6±0.1	-3.2±0.1	-2.2±0.1	-3.7±0.2
W191	-6.7±0.0	-6.7±0.1	-6.7±0.3	-7.3±0.3	-6.6±0.2	-6.1±0.2	-6.7±0.5	-8.3±0.1
P204	-2.9±0.1	-2.2±0.0	-2.2±0.1	-2.5±0.6	-2.4±0.2	-2.4±0.1	-2.2±0.1	-2.4±0.0
S205	-0.4±0.1	-0.3±0.0	-0.3±0.0	-0.3±0.0	-0.2±0.0	-0.3±0.0	-0.3±0.0	-0.3±0.0
N206	-3.5±0.1	-3.5±0.1	-4.1±0.1	-5.0±0.1	-2.7±0.2	-3.5±0.2	-3.3±0.1	-3.8±0.0
Y207	-6.0±0.2	-5.6±0.2	-5.7±0.3	-6.2±0.0	-6.7±0.1	-5.6±0.1	-6.0±0.3	-6.0±0.0

Table 10. Hydrogen bond occupancy (%) between the ligands and the Sem5 protein in the MD trajectory.

H donor	H receptor	WT	Sar+2	Ala+2	Sar0	Ala0	Sar-1	Ala- 1	Ala-1 N190Q	Ala-1 N190Q/N206Q
Trp191 NE1	Pro -2 O	72	66	61	38	69	94	61	38	25
Asn206 ND2	Pro +1 O	50	60	63	43	54	94	56	78	36
Tyr207 OH	X +2 O	88	80	93	85	92	100	67		54
Arg-3 NE	Glu172 OE2	96	95	93	94	95	48	92	74	90
Arg-3 NH2	Glu172 OE1	78	25	68	88	82	99	15	22	13
Arg-3 NH2	Glu172 OE2	37	86	47	23	33	100	90	79	91
Ala-1 N	Gln190 OE1									59
Trp191 NE1	X -1 O								36	

Table 11. Occupancies of residues appear at each position (total number of sequences is 207).

Residue Position	VC value	van der Waals (kcal/mol)	Non-Gap Seq. No.	Most frequent residue		Second most frequent residue	
				Name	No.(%)	Name	No.(%)
L162	0.3 (0.7)	-0.4(1.1%)	175	L	118(67%)	K	15(8%)

F163	4.6
	(12.2)
		-5.0(13.4%)	181	Y
				F
					116(64%)
					49(27%)
						A	7(4%)

D164	0.7 (1.9)	-0.9(2.4%)	183	D	**147(80%)**	N	9(5%)

F165	1.8
	(4.7)
		-1.8(4.8%)	183	Y
				F
					112(61%)
					67(37%)
						L/W/A/I	1(0.5%)

N166	0.04
	(0.1)
		-0.4(1.1%)	193	E
				D
					47(24%)
					34(18%)
						Q
						N
							18(9%)
							4(2%)

P167	0.1
	(0.02)
		-0.6(1.6%)	188	A	81(43%)	P
						G
							27(14%)
							26(14%)

Q168	0.5
	(1.3)
		-3.0(8.0%)	186	R
				K
					47(25%)
					18(10%)
						Q
						N
							19(10%)
							11(6%)

E169	0.6
	(1.6)
		-1.8(4.8%)	189	E
				D
					21(11%)
					42(22%)
						T	32(17%)

E172	1.0
	(2.6)
		+1.2(3.2%)	189	E
				D
					84(44%)
					68(36%)
						G	18(10%)

D187	0.0
	(0.0)
		-0.02(0.05%)	170	S
				T
					34(20%)
					32(19%)
						D
						E
							12(7%)
							8(5%)

D188	0.1
	(0.3)
		-0.2(0.5%)	152	D
				E
					55(37%)
					26(17%)
						N
						Q
							20(13%)
							4(3%)

N190	0.7
	(1.8)
		-2.8(7.5%)	191	D
				E
					55(29%)
					25(13%)
						G
						N
						Q
							40(21%)
							30(16%)
							15(8%)

W191	6.8 (18.3)	-7.2(19.3%)	195	W	**185(95%)**	I	3(2%)

I202	0.1
	(0.3)
		-0.3(0.8%)	186	Y
				F
				W
					53(28%)
					19(10%)
					31(17)
						L
						I
						M
						V
							35(19%)
							17(9%)
							12(6%)
							7(4%)

P204	2.4 (6.4)	-2.4(6.4%)	187	P	**187(100%)**		
S205	0.3 (0.7)	-0.5(1.3%)	187	S	95(51%)	A	36(19%)
N206	2.9 (7.6)	-3.9(10.4%)	175	N	**127(73%)**	S	16(9%)

Y207	5.9
	(15.6)
		-6.1(16.3%)	172	Y
				F
					144(84%)
					21(12%)
						H
						L
							2(1%)
							2(1%)

Chapter 5. Computational study on the molecular basis of the HIV-1 protease drug resistance

Computational study on the molecular basis of the HIV-1 protease drug resistance

Wei Wang,

Graduate Group in Biophysics

University of California, San Francisco

San Francisco, CA 94143

and

Peter A. Kollman*

Department of Pharmaceutical Chemistry

University of California, San Francisco

San Francisco, CA 94143

* Author for correspondence

Tel: (415) 476-4637 (PAK)

Fax: (415) 502-1411

Email: pak@cgl.ucsf.edu

Short title: Drug resistance of the HIV protease

Keywords: HIV protease; drug resistance; molecular dynamics; MM/PBSA; FV value.
Grant sponsor: NIH; Grant number: GM-56531 (P. Ortiz de Montellano, P.I.).

Abstract

Drug resistance has sharply limited the effectiveness of HIV-1 protease inhibitors in AIDS therapy. It is critically important to understand the basis of this resistance for designing new drugs. We have evaluated the free energy contribution of each residue in the HIV protease in binding to one of its substrates and to the 5 FDA approved protease drugs. Analysis of these free energy profiles and the variability at each position suggests: (1) drug resistance mutations are likely to occur at not well conserved residues if they interact more favorably with drugs than with the substrate; (2) resistance-evading drugs should have a similar free energy profile as the substrate and interact most favorably with well conserved residues. This method can assist in designing resistance-evading drugs for any target.

Introduction

One of most challenging problems in AIDS therapy is that the HIV virus develops drug resistant variants rapidly due to the low fidelity of its reverse transcriptase and the high replication rate[1-4]. Extensive research in the past decade has been dedicated to designing resistance-evading drugs for the HIV protease, which is critical for the maturation of viral structural (gag) and enzymatic (pol) proteins. The HIV protease is an aspartyl protease and is composed of two symmetric monomers. Many crystal structures of the HIV protease and its complexes with inhibitors have been solved and extensive clinical resistance data have been accumulated for the 5 FDA approved drugs. This provides the ground for understanding the molecular basis of drug resistance. Here we show that resistance mutations to the 5 FDA approved HIV protease drugs only occur at functionally unimportant positions, which also interact more favorably with drugs than with the substrate. A combination of conservation analysis and free energy calculations on each protease residue suggests that more potent protease drugs should interact more favorably with well conserved residues, i.e. those functionally important residues, especially with Ala28 and Asp29. This strategy can be exploited to design resistance-evading drugs for any target. We also propose an empirical parameter, the FV (free energy/variability) value, to identify resistance mutations for any HIV protease inhibitors, which can be easily extended to identify critical residues for other protein-protein and protein-ligand interactions.

Methods

(1) MD simulations

All molecular dynamics simulations presented in this work are performed using the AMBER5.0 simulation package[5] and the Cornell et al. force field[6] with the TIP3P water model[7]. The starting structures of protease-drug complexes are taken from the Protein Data Bank. The PDB entries are 1hxb (Saquinavir), 1hpv (Amprenavir), 1hxw (Ritonavir), 1hsg (Indinavir), 1ohr (Nelfinavir), and 1hvi (A77003). There are two conformations for Saquinavir in the crystal structure and the first is used in our simulation. The structure of the substrate (Ace-Ser-Gln-Asn-Tyr-Pro-Ile-Val) was modified from an inhibitor JG365 (Ace-Ser-Leu-Asn-Phe-PSI(CH(OH)-CH2N)-Pro-Ile-Val-OME) complexed with the protease (PDB entry is 7hvp)[8]. This substrate covers the whole binding site of the protease. The molecules are solvated in a 80×80×80 $Å^3$ box of water. An appropriate number of counter ions are added to neutralize the system. Particle Mesh Ewald (PME)[9] is employed to calculate the long-range electrostatic interactions. All structures are minimized first using the SANDER module in AMBER5.0. Molecular dynamics simulations are carried out thereafter. The temperature of the system is raised gradually from 50K to 298 K in 50 ps and equilibrated at 298 K for another 120 ps. An additional 120 ps of MD simulation is performed for data collection and 100 snapshots are saved for the subsequent analysis. The deviations are estimated by the difference between the first and second half of the trajectories. The SHAKE procedure[10] is employed to constrain all bonds. The time step of the simulations is 2 fs. A 8.5Å cut-off is used for the nonbonded van der Waals interactions and no cutoff for nonbonded electrostatic interactions. The nonbonded pairs are updated every 15 steps.

(2) The MM/PBSA method

The binding free energy is calculated as[11]:

$$\Delta G_b = \Delta G_{MM} + \Delta G_{sol}^{LP} - \Delta G_{sol}^{L} - \Delta G_{sol}^{P} - T\Delta S \qquad (1)$$

where ΔG_b is the binding free energies in water, ΔG_{MM} is the interaction energy between the ligand and the protein, ΔG_{sol}^{L}, ΔG_{sol}^{P} and ΔG_{sol}^{LP} are solvation free energies for the ligand, protein and complex respectively, and $-T\Delta S$ is the conformational entropy contribution to the binding. ΔG_{MM} is calculated from molecular mechanics (MM) interaction energies:

$$\Delta G_{MM} = \Delta G_{int}^{ele} + \Delta G_{int}^{vdw} \qquad (2)$$

where ΔG_{int}^{ele} and ΔG_{int}^{vdw} are electrostatic and van der Waals interaction energies between the ligand and the receptor, which are calculated using the CARNAL and ANAL modules in AMBER5.0 software suite.

The solvation energy, ΔG_{sol}, is divided into two parts, the electrostatic contributions, ΔG_{sol}^{ele}, and all other contributions, $\Delta G_{sol}^{nonpolar}$.

$$\Delta G_{sol} = \Delta G_{sol}^{ele} + \Delta G_{sol}^{nonpolar} \qquad (3)$$

The electrostatic contribution to the solvation free energy, ΔG_{sol}^{ele}, is calculated using the DelPhiII software package[12], which solves the Poisson-Boltzmann equations numerically and calculates the electrostatic energy according to the electrostatic potential. The grid size used is 0.5Å. Potentials at the boundaries of the finite-difference lattice are set to the sum of the Debye-Huckel potentials. The value of interior dielectric constant is set to 1. As shown in our previous study[13], after combining all the terms, the binding free energy is calculated as:

$$\Delta G_b = \Delta G_{int}^{vdw} + \Delta G_{sol}^{nonpolar} + (1/n)\, \Delta G_{1\text{-}1}^{ele} + (\Delta G_{RFE}{}^{LP}{}_{n\text{-}80} - \Delta G_{RFE}{}^{L}{}_{n\text{-}80} - \Delta G_{RFE}{}^{P}{}_{n\text{-}80})$$

(4)

where n is the interior dielectric constant, which is 1 in this study. $\Delta G_{1\text{-}1}^{ele}$ is the molecular mechanics electrostatic interaction energy between the ligand and the protein. $\Delta G_{RFE}{}^{LP}{}_{n\text{-}80}$, $\Delta G_{RFE}{}^{L}{}_{n\text{-}80}$ and $\Delta G_{RFE}{}^{P}{}_{n\text{-}80}$ are reaction field energies obtained from DelPhi for ligand, protein and complex respectively with interior and exterior dielectric constants set to n and 80 respectively.

The dielectric constant of water is set to 80. The dielectric boundary is taken as the solvent accessible surface defined by a 1.4 Å probe sphere. The radii of atoms are taken from the PARSE parameter set[14]. Partial charges are taken from Cornell et al. force field for standard amino acids.

The solvent accessible surfaces (SAS) are calculated using the MSMS program[15]. The non-polar contribution to the solvation free energy, $\Delta G_{sol}{}^{nonpolar}$, is calculated as $0.00542 \times SAS + 0.92$ kcal/mol[14].

Normal mode analysis is used to estimate conformational entropy $-T\Delta S$. Because this analysis requires extensive computer time, only three snapshots are taken in this study to estimate the order of magnitude of the conformational entropy.

(3) Psi-BLAST and FV value

Psi-BLAST[16] with default parameters (BLOSUM62, Expect=10, E-value threshold for inclusion in Psi-BLAST iteration=0.002, Descriptions=500, Alignments=500, composition based statistics) is used to search the SWISS-PROT database. Multiple sequence alignment is carried out on 80 sequences with scores >64 and E-value $<1\text{x}10^{-10}$ using the Pileup module in GCG software package (Version 10.1, Genetics Computer Group, Inc., 2000) with default parameters. These 80 sequences include HIV, SIV and FIV proteases.

In order to identify critical residues for binding, we defined an empirical parameter called the FV value. The FV value is defined as a product of one residue's contribution to binding free energy ΔG_{res} and variability at that position V_i. ΔG_{res} is estimated as:

$$\Delta G_{res} = E_{vdw} + E_{ele} + \Delta G_{res}^{sol} \qquad (5)$$

where E_{vdw} and E_{ele} are van der Waals and electrostatic interaction energies between the residue and the whole ligand respectively. ΔG_{res}^{sol} is the contribution of solvation penalty by that residue. It is calculated as:

$$\Delta G_{res}^{sol} = \Delta G^{sol} - \Delta G_0^{sol} \qquad (6)$$

where ΔG^{sol} and ΔG_0^{sol} are the solvation energies calculated from Equation (3) with normal partial charges and zero charges on that specific residue respectively.

The variability V_i is calculated as:

$$V_i = \Sigma_j\, (1-P_{ij}/P_{ii}) * W_j \qquad (7)$$

where W_j is the weight of the jth sequence. W_j is calculated for each sequence in the alignment based on sequence identity. If n sequences are >80% identical to each other, each sequence has 1/n weight. Next, the sum of all sequences in the alignment is normalized to 1. This weight prevents over-presenting very similar sequences in the Psi-BLAST search results.

P_{ij} in Equation (7) represents how likely the amino acid a_i in the ith sequence can be mutated to the amino acid a_j in the jth sequence and is calculated as:

$$P_{ij} = 2^{(2*Mij)} \qquad (8)$$

where M_{ij} is the element of BLOSUM62 for a_i and a_j. BLOSUM62 is chosen to be consistent with the matrix used in Psi-BLAST search. M_{ij} for gap is assigned a penalty score of -4 in the BLOSUM62 matrix.

Results and discussion

(1) Mutations are not tolerated by functionally important residues.

Drug resistance mutants of the HIV protease significantly reduce inhibitor binding without severally deteriorating its own function. This determines that those functionally critical residues, such as the catalytic triad Asp25, Thr26 and Gly27, are not tolerant to any mutations. The variability of each position in the HIV protease is shown in Figure 1. Low variability means that the residue in the HIV protease is well conserved in the sequence alignment. From this figure, it is clear that no drug resistance mutations have ever been observed for positions with variability lower than 0.25 (P9, D25, T26, G27, A28, D29, G49, G51, G86 and R87). These conserved residues are either crucial for catalyzing polypeptide cleavage, e.g. Asp25, or stabilizing the structure of the protease dimer, e.g. Arg87. These residues apparently mutate very little or not at all under the drug selection pressure. Therefore, drug resistance mutations only can occur at those residues unimportant for viral activity but important for drug binding.

(2) More favorable interactions with Leu23, Ala28, Gly49, Arg87 and more importantly, with Asp29 could enable the 5 FDA approved drugs to be less sensitive to viral resistance.

In order to identify drug resistant mutations, first we have identified residues responsible for binding with ligands by calculating the van der Waals interaction energy between each residue in the HIV protease and the substrate; and secondly, we have evaluated each of these residues' contribution to the binding free energy ΔG_{res} and calculated the difference of ΔG_{res}, $\Delta\Delta G_{res}$, between the substrate and drug. We have then analyzed the $\Delta\Delta G_{res}$ and the variability of each residue to suggest which mutations may cause drug resistance.

As the first step, we modeled the complex of the wild type HIV protease and one of its gag cleavage sequences, SQNYPIV, based on the complex structure of one linear peptide inhibitor, JG365[8], because there was no crystal structure of substrate-protease complex available and JG365 is reasonably similar to the substrate. This substrate covers the whole binding site of the protease. We optimized the substrate complex in water using molecular dynamics until equilibrium was achieved (all heavy atoms RMSD became flat at ~1.5 Å). We note that Schiffer and coworkers[17] recently have solved a complex structure between an inactive HIV-1 protease (D25N) and a long substrate peptide, KARVLAEAMS, which is different from the substrate we are studying. The structure has been deposited to the PDB database (on hold and pdb entry 1f7a). It would be interesting to compare our modeled structure with this crystal structure after it is released to public access.

Residues in the HIV protease are considered to be in or close to the binding site if they have a van der Waals interaction energy with the substrate that is more negative than –0.5 kcal/mol (Figure 2). Residues with the most frequent drug resistance mutations have relatively more favorably van der Waals energies. The exceptions are L24, G73 and L90. This implies that resistances caused by mutations at these three positions might be due to changes of conformation or stability and, therefore, the proteolytic kinetics of the HIV protease. All residues which have van der Waals interaction energies with the substrate more favorable than –0.5 kcal/mol as well as the three known resistance residues (L24, G73, and L90) which do not have such a favorable van der Waals interaction energy were evaluated for their contributions to the binding free energy in this study. This set of residues includes all major drug resistance mutations.

Before we evaluate ΔG_{res} for each residue, we estimate the binding free energies of the substrate and the 5 FDA approved drugs, i.e. Ritonavir, Saquinavir, Amprenavir, Indinavir and Nelfinavir, using the MM/PBSA method[11] (Table 1). It is obvious that all inhibitors bind more tightly than the substrate. The substrate is the largest ligand and, thus, it has the most favorable van der Waals interactions with the protease. However, it also has the least favorable electrostatic contribution $\Delta G_{int+sol}{}^{ele}$ to the binding free energy. This suggests that it is important for potent drugs to have optimal electrostatic interaction with the protease but less desovlation penalty. It is worth pointing out that the $T\Delta S$ for substrate or inhibitor binding is not included in Table 1 because we assume that they are similar in magnitude for the inhibitors and the substrate. This assumption seems reasonable, given Kuhn and Kollman's calculated $T\Delta S$ for various ligands binding to avidin[18].

Studying drug resistance requires evaluation of each residue's contribution to the binding, which is an experimentally difficult but computationally feasible task. Here we combine molecular mechanics energies (van der Waals and electrostatic energy) and desolvation penalty (by solving Poisson-Boltzmann equation) to estimate a single residue's contribution to the binding. $\Delta\Delta G_{res}$'s between the substrate and drugs for selected residues are plotted in Figure 3. It should be noted that the HIV-1 protease is a dimer. A single mutation of its gene is a double mutation in the protein. Figure 3 plots $\Delta\Delta G_{res}$'s of the double mutations as well as single mutation on each monomer. Among residues with <0.25 variability (Figure 1), it shows that all drugs interact much more favorably with Asp25 and Gly27, slightly more favorably with Leu23, but slightly less favorably with Ala28, Gly49 and Arg87, and much less favorably with Asp29 than the substrate. All these residues are well conserved (Figure 1) and appear to be functionally important; thus, mutations do not tend to occur at these positions.

The above analysis suggests that these 5 FDA approved drugs can be altered to become more powerful to combat HIV drug resistance if their interactions with Leu23, Ala 28, Gly49, Arg87 and more importantly, with Asp29 are improved. Leu23 and Ala28 are hydrophobic and in the center of the binding site. More favorably interactions with them can be achieved by adding some non-polar groups in the drugs at P1 and P1'. Interactions with Gly49, Arg87 and Arg29 only can be improved if drugs can designed to make more favorable electrostatic interactions with them but less desolvation penalty, which is difficult but not impossible. One speculation is to add some polar or even charged groups at P3 and P3'.

In order to further illustrate how to improve the 5 FDA approved drugs, we compared contribution to binding for every residue in each drug (Figure 5). A residue in a drug is defined based on chemical groups and as similar as possible to a natural amino acid. Investigation of van der Waals and electrostatic contribution to binding can provide clues to improve these drugs. For example, the second residue of Saquinavir is Asn. It has the least favorable free energy and electrostatic contribution (PB + Coulomb term) and second least favorable van der Waals energy. This Asn is close to Gly49A, which is well conserved. This suggests that a residue at this position with more favorable interactions with Gly49A and less desolvation penalty can help Saquinavir to combat resistance.

(3) FV value can identify drug resistant mutations.

We have plotted $\Delta\Delta G_{res}$ and variability in Figure 4. Most drug resistance mutations are in the region of $\Delta\Delta G_{res}$ between 0~-3.0 kcal/mol. 7 of them have variability between 0.65-0.85.

We propose an empirical parameter, FV value, to quantitatively identify drug resistant mutations (Table 2). The FV value is defined as the product of ΔG_{res} and variability of that residue. The purpose to define this parameter is to include free energy and evolution information into one value. A mutation is considered drug resistant if it causes >10 fold change of K_i of drugs. Usually it is assumed that homodimers of the HIV protease are formed and, thus, double mutations should be considered. A threshold of -1.0 (=2×1.4×0.35) is used for the FV value of double mutations for identifying resistance, which corresponds to 1.4 kcal/mol (10 fold change of K_i) and variability greater than 0.35. The accuracy of identifying resistance mutations by the FV value varies among drugs, but average accuracy is 76%, which is quite good.

The FV value did not find G48 resistant for Saquinavir and Ritonavir. From previous studies[13,19,20], mutating Gly48 to other hydrophobic residues favors formation of heterodimer of the HIV protease. We can see at least one G48 interacts more favorably with drugs than with the substrate. This implies that possibly heterodimers of the HIV protease are formed under the selection pressure of Saquinavir and Ritonavir. Another residue on which the FV value is not informative is M46. This may be due to the fact that M46 is on the surface of the protein and at the boundary of the interior and exterior region when solving Poisson-Boltzmann equation; thus, more error may be introduced in the PB calculations.

Multiple mutations tend to be found *in vivo*, but in this study we really only compare residues for which *in vitro* single mutation experimental data are available. It is obvious that the same analysis can be applied to study multiple mutations.

Conclusion

We have shown here that drug resistant mutations only can occur at functionally unimportant positions. Therefore, resistance-evading drugs should interact strongly with these conserved residues. We have analyzed the 5 FDA approved drugs and suggest that improving interactions between these drugs and residues Gly27, Ala28, and more important, with Asp29 and Arg87 in the protease may possibly enhance their abilities to combat drug resistance. An empirical parameter, FV value, was exploited to identify drug resistant mutations and it can be useful in studying other protein-protein or protein-ligand interactions as well.

Acknowledgement

This work was supported in part by grant GM-56531 (P. Ortiz de Montellano, P.I). Part of this investigation was completed using the facilities in the UCSF Computer Graphics Lab, T. Ferrin, director, supported by grant RR-1081 from the NIH. Supercomputer time at the NSF centers (NCSA and SDSC) is gratefully acknowledged.

References

1. Wlodawer, A. & Erickson, J.W. Structure-Based Inhibitors of Hiv-1 Protease. *Annual Review of Biochemistry* **62**, 543-585 (1993).
2. Erickson, J.W. & Burt, S.K. Structural Mechanisms of Hiv Drug Resistance. *Annual Review of Pharmacology and Toxicology* **36**, 545-571 (1996).

3. Wlodawer, A. & Vondrasek, J. Inhibitors of HIV-1 protease: A major success of structure-assisted drug design. *Annual Review of Biophysics and Biomolecular Structure* **27**, 249-284 (1998).

4. Erickson, J.W. The Not-So-Great Escape. *Nature Structural Biology* **2**, 523-529 (1995).

5. Pearlman, D.A. et al. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comp. Phys. Comm.* **91**, 1-41 (1995).

6. Cornell, W.D. et al. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* **117**, 5179-5197 (1995).

7. Jorgensen, W.L., Chandrasekhar, J., Madura, J., Impey, R.W. & Klein, M.L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926-935 (1983).

8. Swain, A.L. et al. X-Ray Crystallographic Structure of a Complex between a Synthetic Protease of Human Immunodeficiency Virus-1 and a Substrate-Based Hydroxyethylamine Inhibitor. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 8805-8809 (1990).

9. Darden, T.A., York, D.M. & Pedersen, L. Particle Mesh Ewald - an Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089-92 (1993).

10. Rychaert, J.P., Ciccotti, G. & Berendsen, H.J.C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes.. *J. Comput. Phys.* **23**, 327-341 (1977).

11. Kollman, P.A. et al. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accounts of Chemical Research* **in press**(2000).

12. Gilson, M.K., Sharp, K.A. & Honig, B.H. Calculating electrostatic interactions in biomolecules: method and error assessment. *J. Comput. Chem.* **9**, 327-335 (1987).

13. Wang, W. & Kollman, P.A. Free energy calculations on dimer stability of the HIV protease using molecular dynamics and continuum solvent model. *J. Mol. Bol.* **303**, 567-582 (2000).

14. Sitkoff, D., Sharp, K.A. & Honig, B. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* **98**, 1978-88 (1994).

15. Sanner, M.F., Olson, A.J. & Spehner, J.C. Reduced Surface - an Efficient Way to Compute Molecular Surfaces. *Biopolymers* **38**, 305-320 (1996).

16. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-3402 (1997).

17. Prabu-Jeyabalan, M., Nalivaika, E. & Schiffer, C.A. How does a symmetric dimer recognize an asymmetric substrate? A substrate complex of HIV-1 protease. *Journal of Molecular Biology* **301**, 1207-1220 (2000).

18. Kuhn, B. & Kollman, P.A. Binding of a diverse set of ligands to avidin and streptavidin: An accurate quantitative prediction of their relative affinities by a

combination of molecular mechanics and continuum solvent models. *Journal of Medicinal Chemistry* **43**, 3786-3791 (2000).

19. Babe, L.M., Pichuantes, S. & Craik, C.S. Inhibition of Hiv Protease Activity by Heterodimer Formation. *Biochemistry* **30**, 106-111 (1991).
20. McPhee, F., Good, A.C., Kuntz, I.D. & Craik, C.S. Engineering Human Immunodeficiency Virus 1 Protease Heterodimers as Macromolecular Inhibitors of Viral Maturation. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 11477-11481 (1996).
21. Maschera, B. et al. Human immunodeficiency virus - Mutations in the viral protease that confer resistance to saquinavir increase the dissociation rate constant of the protease-saquinavir complex. *Journal of Biological Chemistry* **271**, 33231-33235 (1996).
22. Gulnik, S.V. et al. Kinetic Characterization and Cross-Resistance Patterns of Hiv-1 Protease Mutants Selected under Drug Pressure. *Biochemistry* **34**, 9282-9287 (1995).
23. Partaledis, J.A. et al. In Vitro Selection and Characterization of Human Immunodeficiency Virus Type 1 (Hiv-1) Isolates with Reduced Sensitivity to Hydroxyethylamino Sulfonamide Inhibitors of Hiv-1 Aspartyl Protease. *Journal of Virology* **69**, 5228-5235 (1995).
24. Ermolieff, J., Lin, X.L. & Tang, J. Kinetic properties of saquinavir-resistant mutants of human immunodeficiency virus type 1 protease and their implications in drug resistance in vivo. *Biochemistry* **36**, 12364-12370 (1997).

25. Klabe, R.M., Bacheler, L.T., Ala, P.J., EricksonViitanen, S. & Meek, J.L. Resistance to HIV protease inhibitors: A comparison of enzyme inhibition and antiviral potency. *Biochemistry* **37**, 8735-8742 (1998).

26. Jacobsen, H. et al. Characterization of Human Immunodeficiency Virus Type 1 Mutants with Decreased Sensitivity to Proteinase Inhibitor Ro 31-8959. *Virology* **206**, 527-534 (1995).

Figure legends:

Figure 1. Variability at each position of the HIV protease. Single mutations on any red-labeled residues can cause resistance to at least one drug and residues labeled magenta cause resistance while occuring with other mutations according to the Stanford HIV database http://hivdb.stanford.edu/hiv/Notes.pl maintained by Robert Shafer.

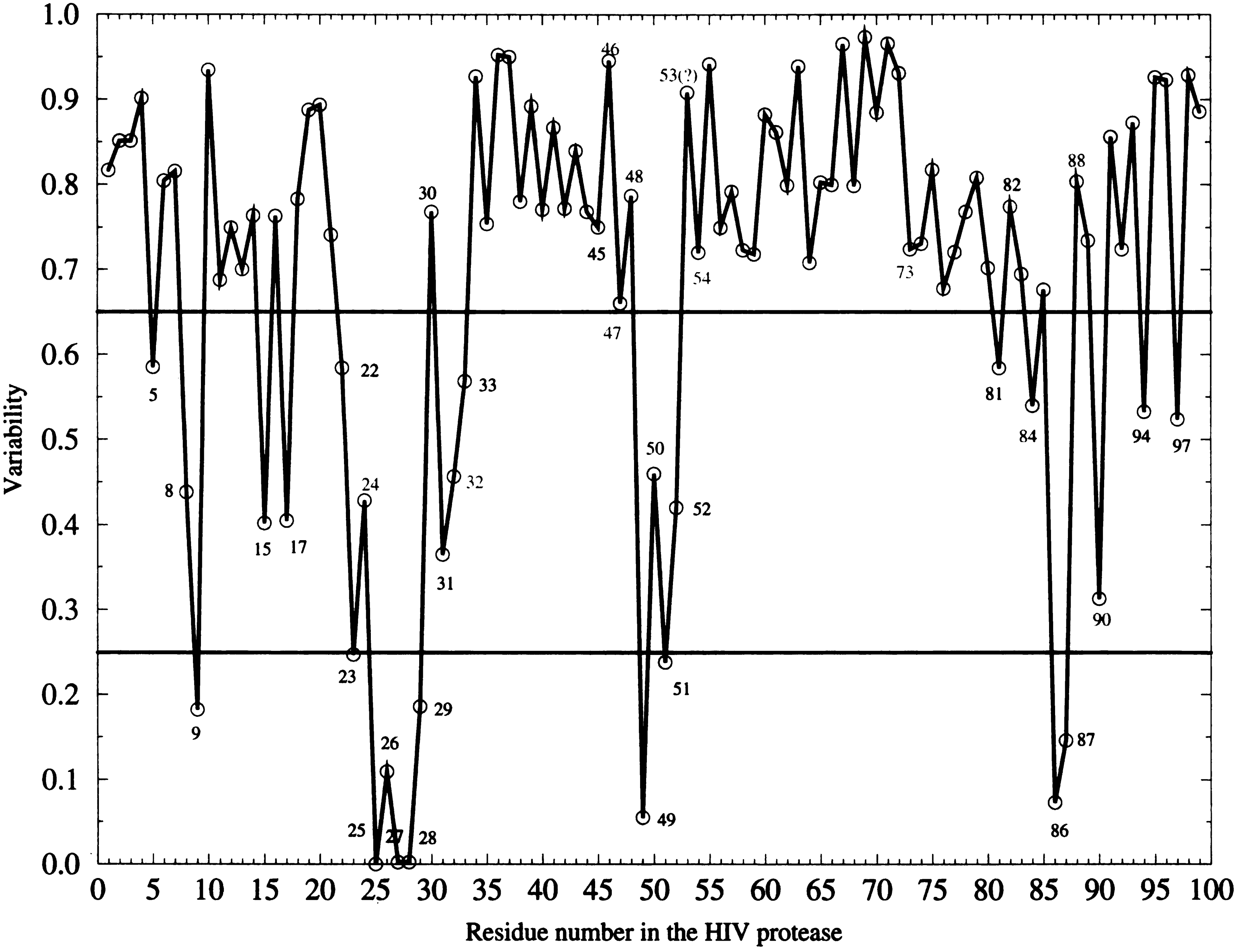
Figure 2. van der Waals interaction energy between each residue in the HIV protease and the substrate.

Figure 3. Free energy difference between each residue's contribution to the binding with the substrate and drugs.

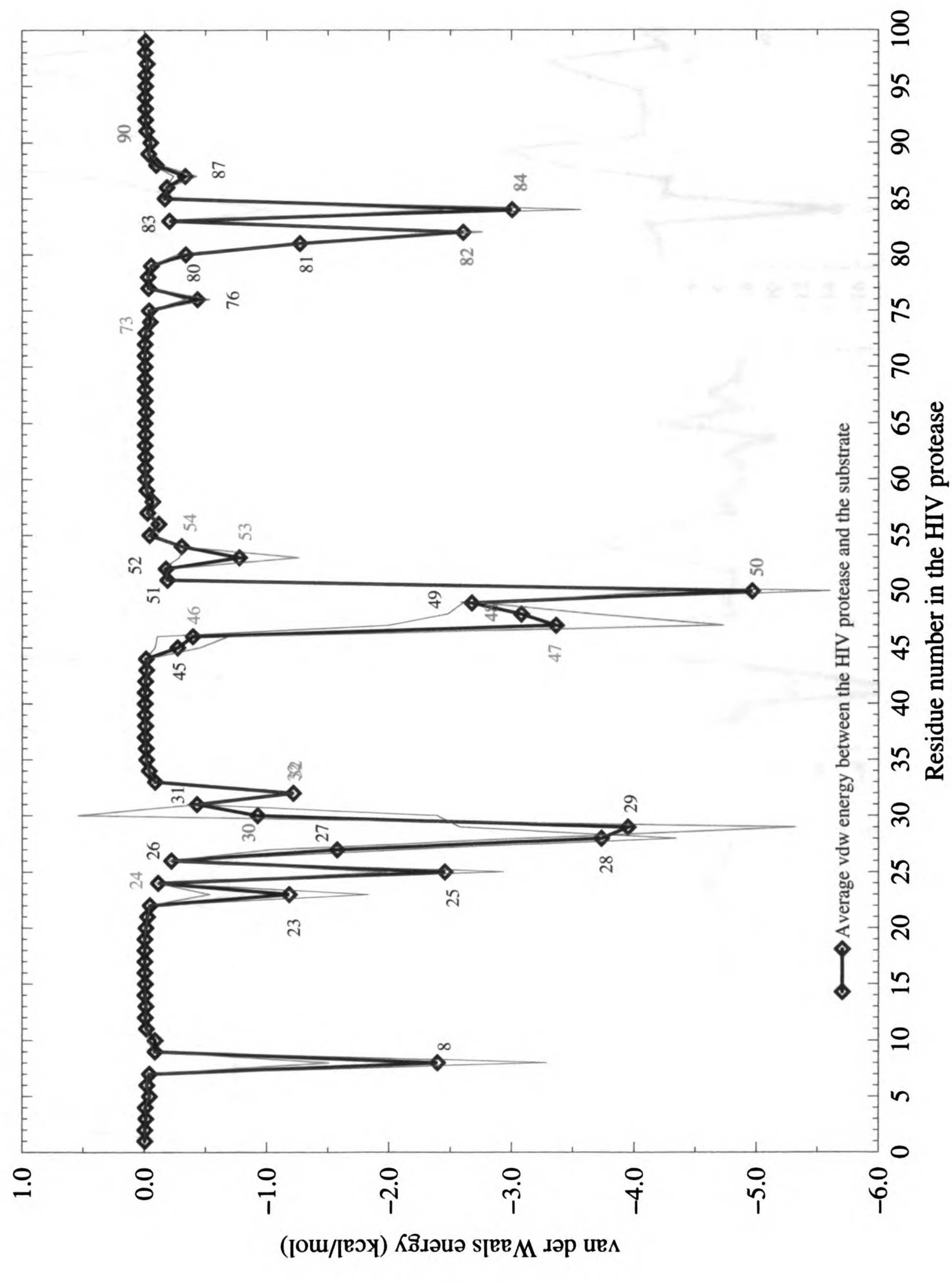
Figure 4. 2-D plot of variability at each position and free energy difference between each residue's contribution to the binding with the substrate and drugs.

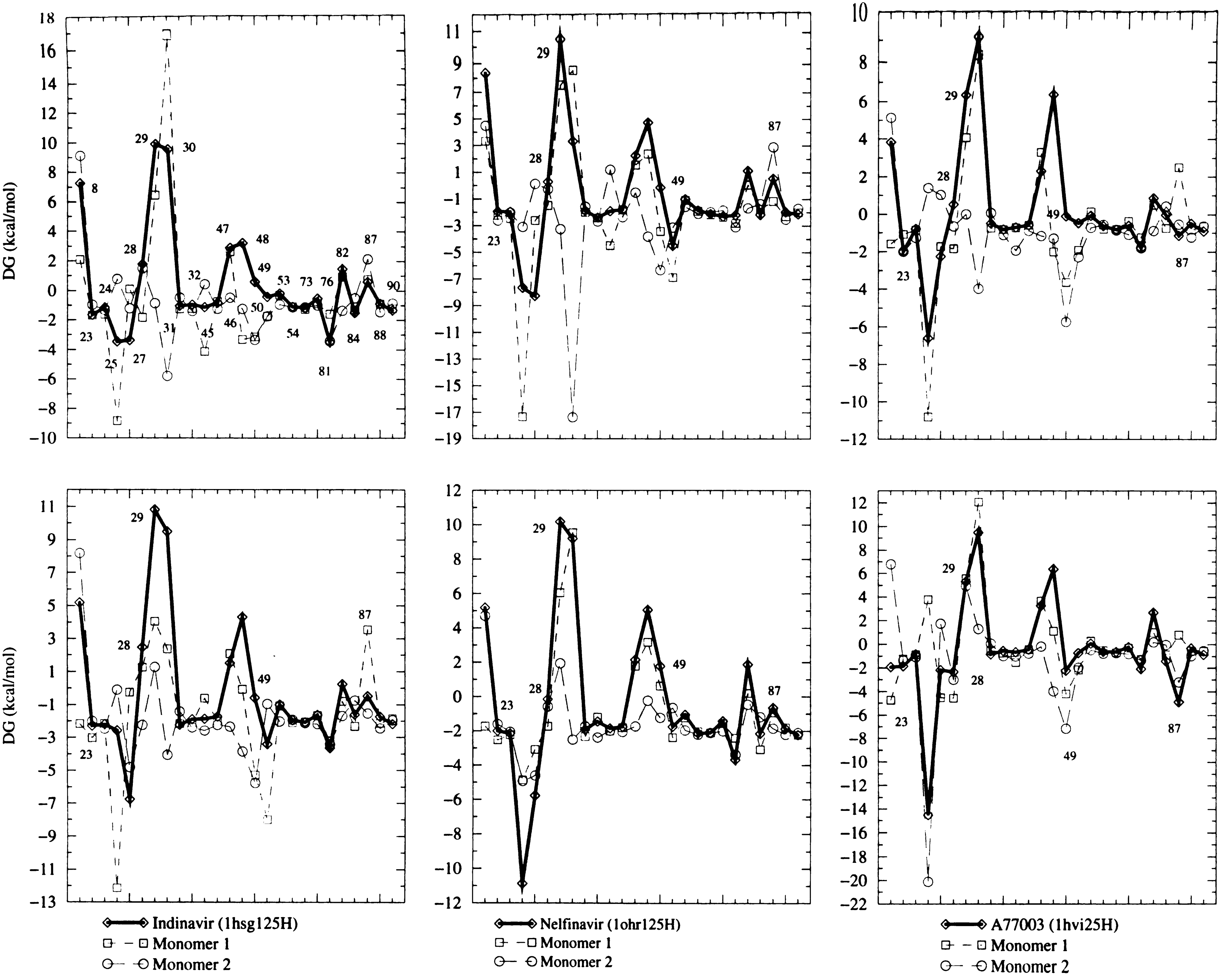
Figure 5. Definition of residues in the drugs.

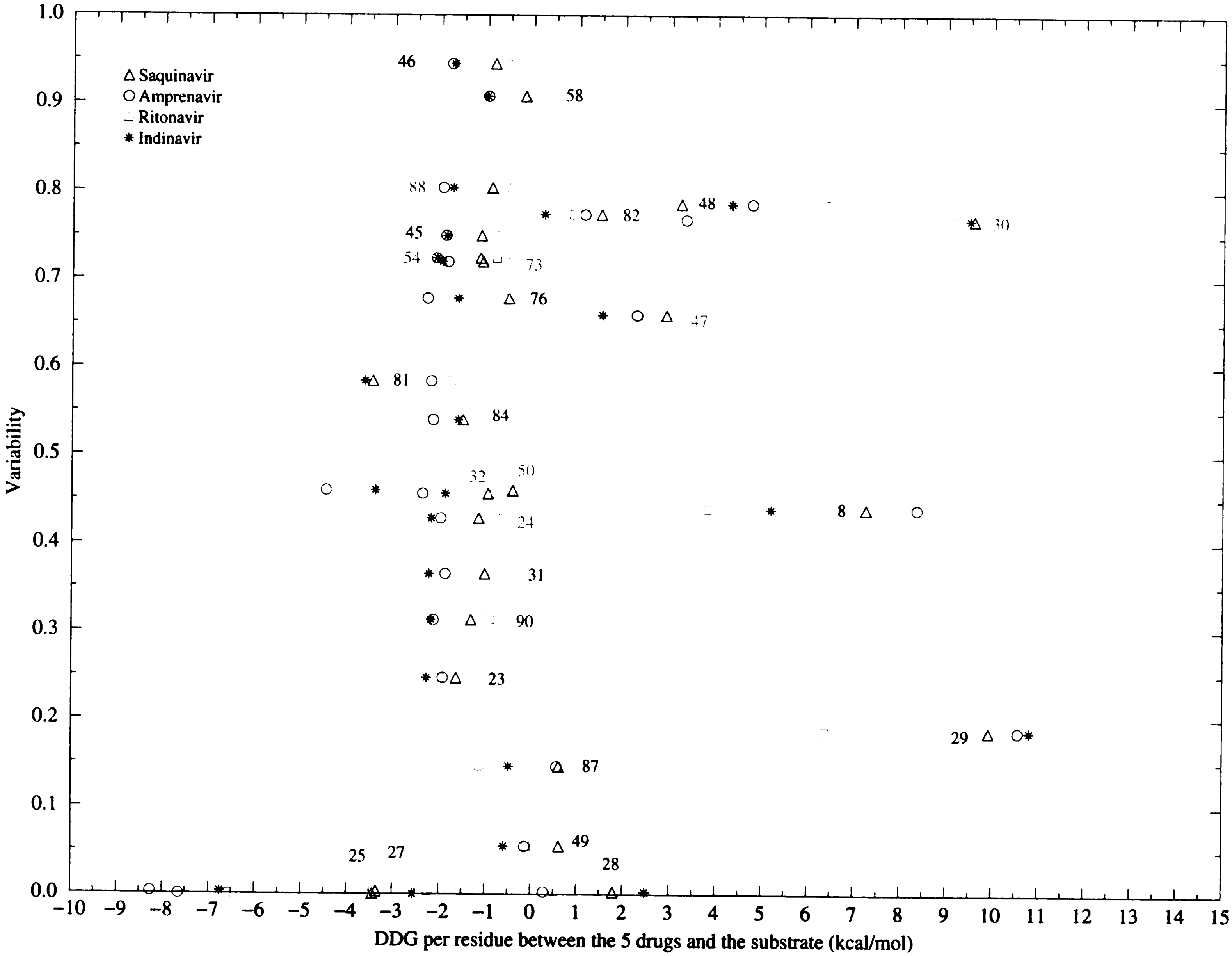
Figure 6. (a) Free energy, (b) electrostatic contribution (Coulomb + PB) and (c) van der Waals energy of each residue in the drugs.



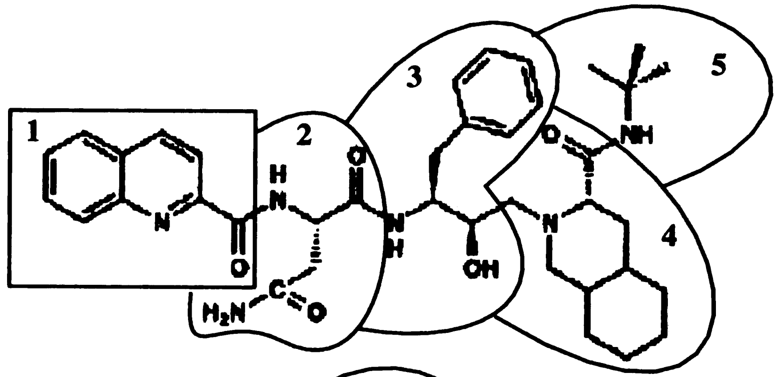
Variability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
Residue number in the HIV protease
5
8
9
15
17
22
23
24
25
26
27
28
29
30
31
32
33
45
46
47
48
49
50
51
52
53(?)
54
73
81
82
84
86
87
88
90
94
97


van der Waals energy (kcal/mol)
1.0
0.0
-1.0
-2.0
-3.0
-4.0
-5.0
-6.0
Residue number in the HIV protease
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
Average vdw energy between the HIV protease and the substrate
8
23
24
25
26
27
28
29
30
31
32
45
46
47
49
50
51
52
53
54
73
76
80
81
82
83
84
87
90

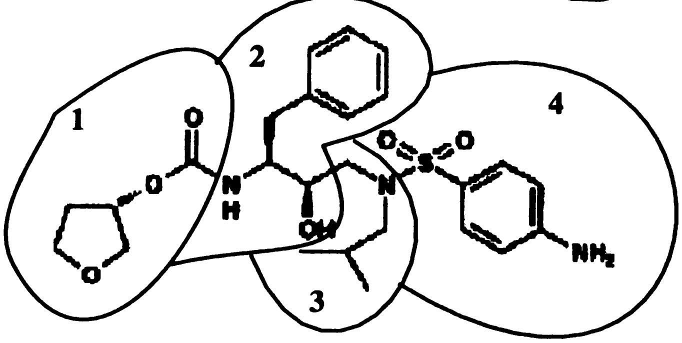

DG (kcal/mol)
18
16
14
12
10
8
6
4
2
0
−2
−4
−6
−8
−10
8
23
24
25
27
28
29
30
31
32
45
46
47
48
49
50
53
54
73
76
81
82
84
87
88
90
11
9
7
5
3
1
−1
−3
−5
−7
−9
−11
−13
−15
−17
−19
23
28
29
49
87
10
8
6
4
2
0
−2
−4
−6
−8
−10
−12
23
28
29
49
87
DG (kcal/mol)
11
9
7
5
3
1
−1
−3
−5
−7
−9
−11
−13
23
28
29
49
87
12
10
8
6
4
2
0
−2
−4
−6
−8
−10
−12
23
28
29
49
87
12
10
8
6
4
2
0
−2
−4
−6
−8
−10
−12
−14
−16
−18
−20
−22
23
28
29
49
87
Indinavir (1hsg125H)
Monomer 1
Monomer 2
Nelfinavir (1ohr125H)
Monomer 1
Monomer 2
A77003 (1hvi25H)
Monomer 1
Monomer 2


△ Saquinavir
○ Amprenavir
Ritonavir
✱ Indinavir
Variability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−10
−9
−8
−7
−6
−5
−4
−3
−2
−1
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
DDG per residue between the 5 drugs and the substrate (kcal/mol)
46
58
88
82
48
30
45
54
73
76
47
81
84
32
50
24
8
31
90
23
29
87
25
27
49
28

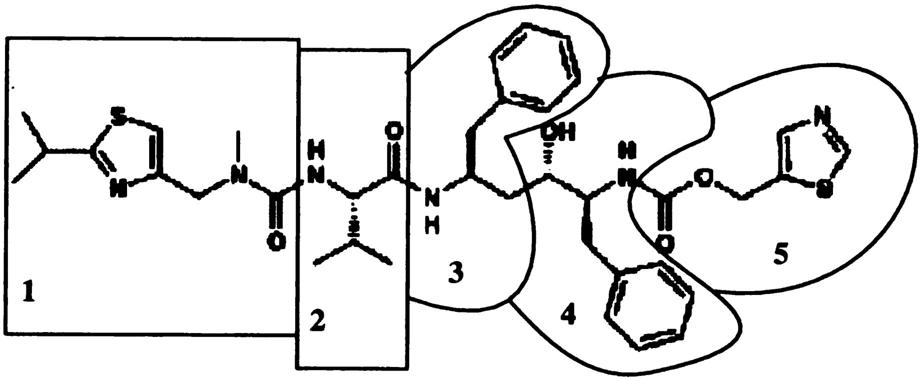
Saquinavir



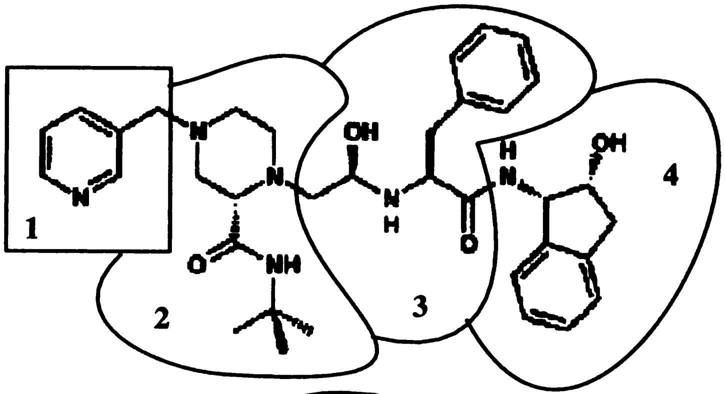
Amprenavir



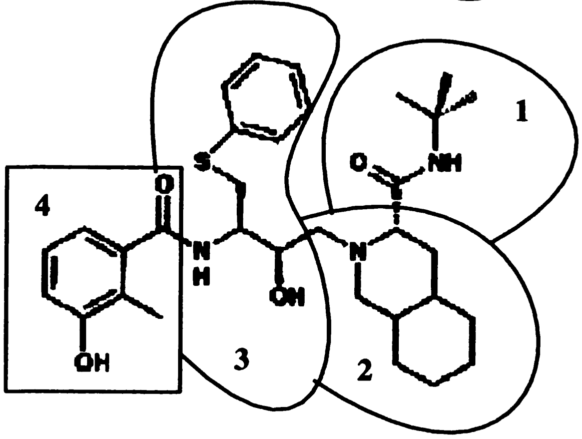
Ritonavir

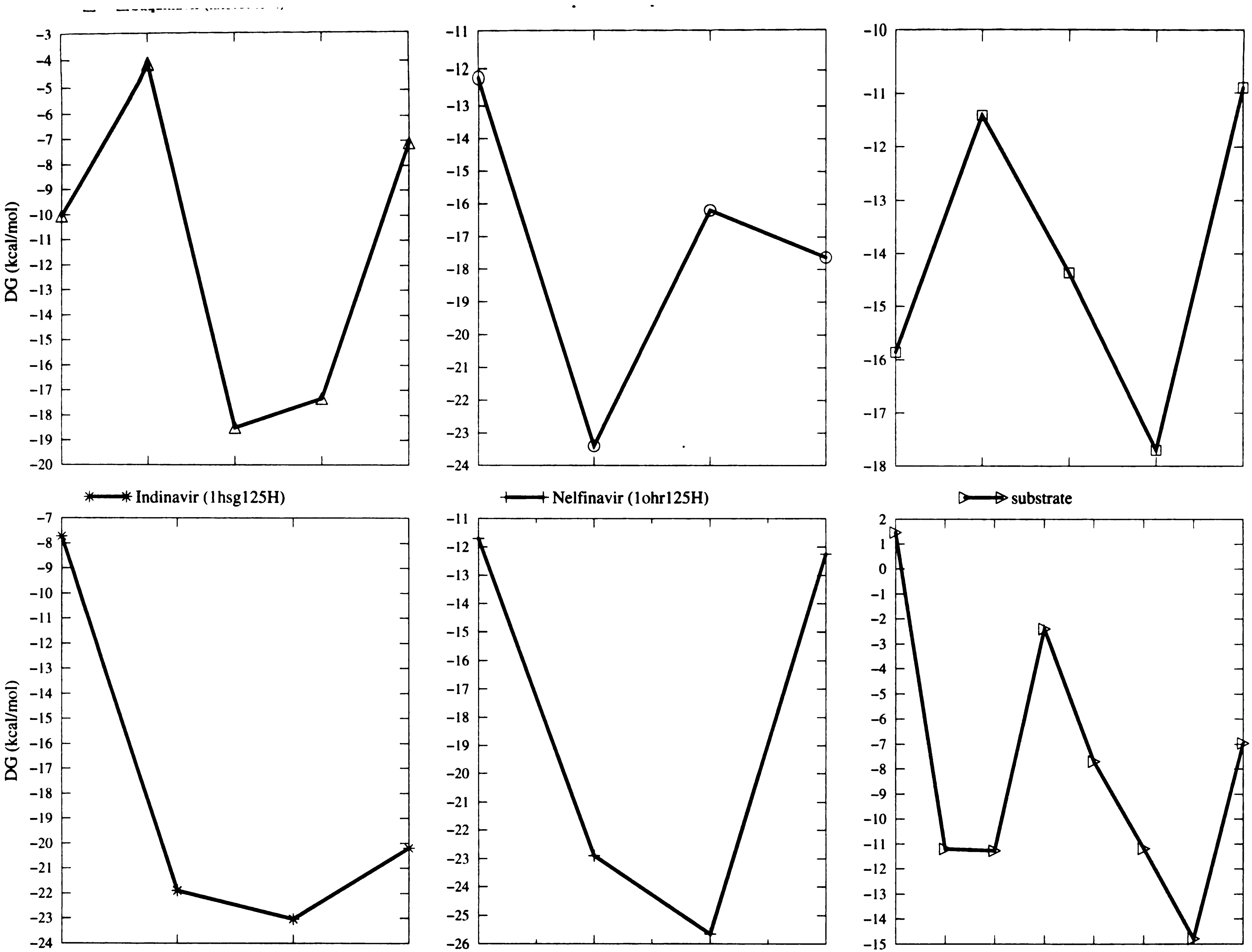


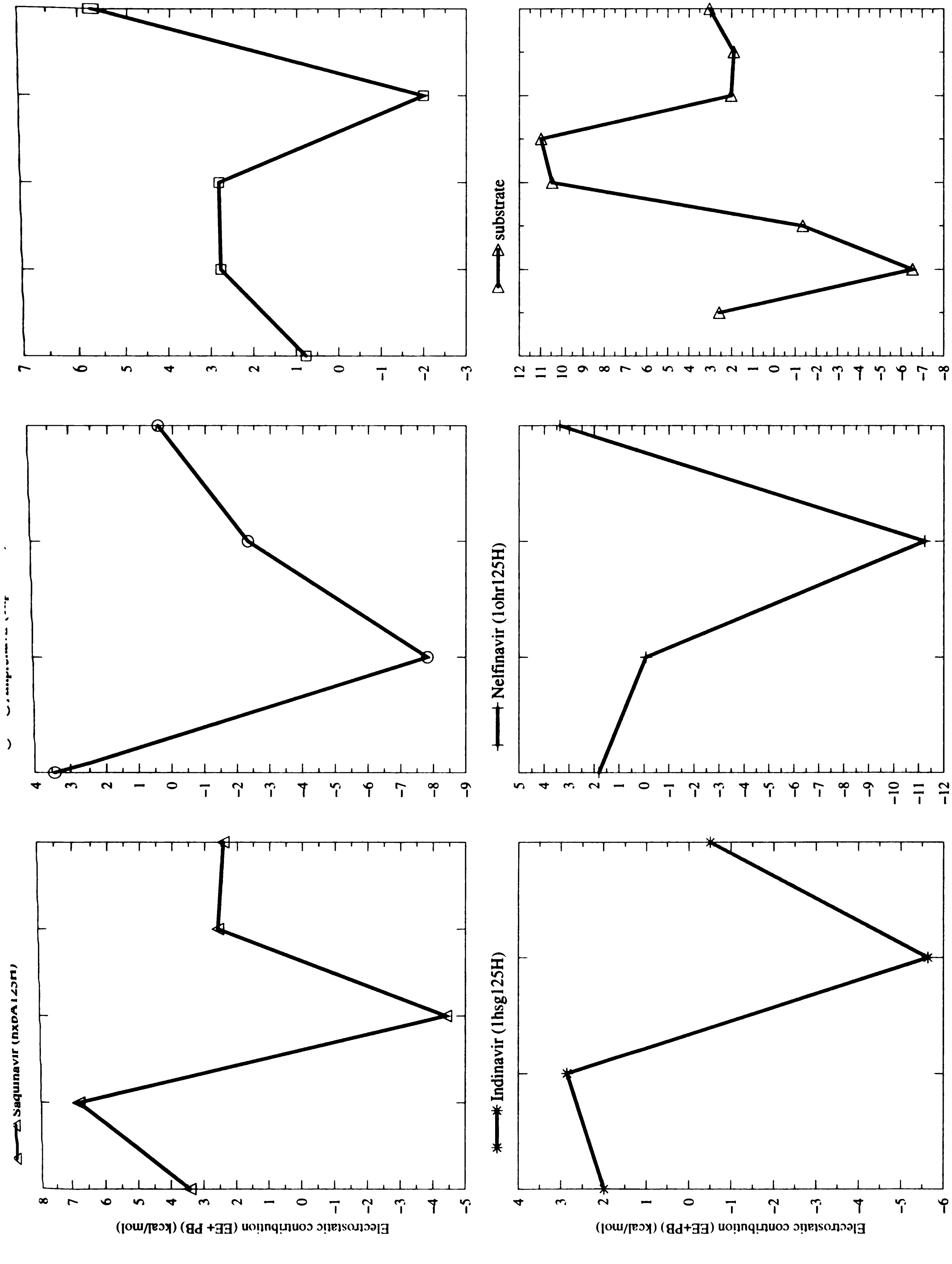
Indinavir



Nelfinavir




DG (kcal/mol)
-3
-4
-5
-6
-7
-8
-9
-10
-11
-12
-13
-14
-15
-16
-17
-18
-19
-20
-11
-12
-13
-14
-15
-16
-17
-18
-19
-20
-21
-22
-23
-24
-10
-11
-12
-13
-14
-15
-16
-17
-18
Indinavir (1hsg125H)
Nelfinavir (1ohr125H)
substrate
DG (kcal/mol)
-7
-8
-9
-10
-11
-12
-13
-14
-15
-16
-17
-18
-19
-20
-21
-22
-23
-24
-11
-12
-13
-14
-15
-16
-17
-18
-19
-20
-21
-22
-23
-24
-25
-26
2
1
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
-10
-11
-12
-13
-14
-15


Saquinavir (hxbA125H)
Electrostatic contribution (EE+ PB) (kcal/mol)
8 7 6 5 4 3 2 1 0 −1 −2 −3 −4 −5
Nelfinavir (1ohr125H)
4 3 2 1 0 −1 −2 −3 −4 −5 −6 −7 −8 −9
7 6 5 4 3 2 1 0 −1 −2 −3
Indinavir (1hsg125H)
Electrostatic contribution (EE+PB) (kcal/mol)
4 3 2 1 0 −1 −2 −3 −4 −5 −6
5 4 3 2 1 0 −1 −2 −3 −4 −5 −6 −7 −8 −9 −10 −11 −12
substrate
12 11 10 9 8 7 6 5 4 3 2 1 0 −1 −2 −3 −4 −5 −6 −7 −8

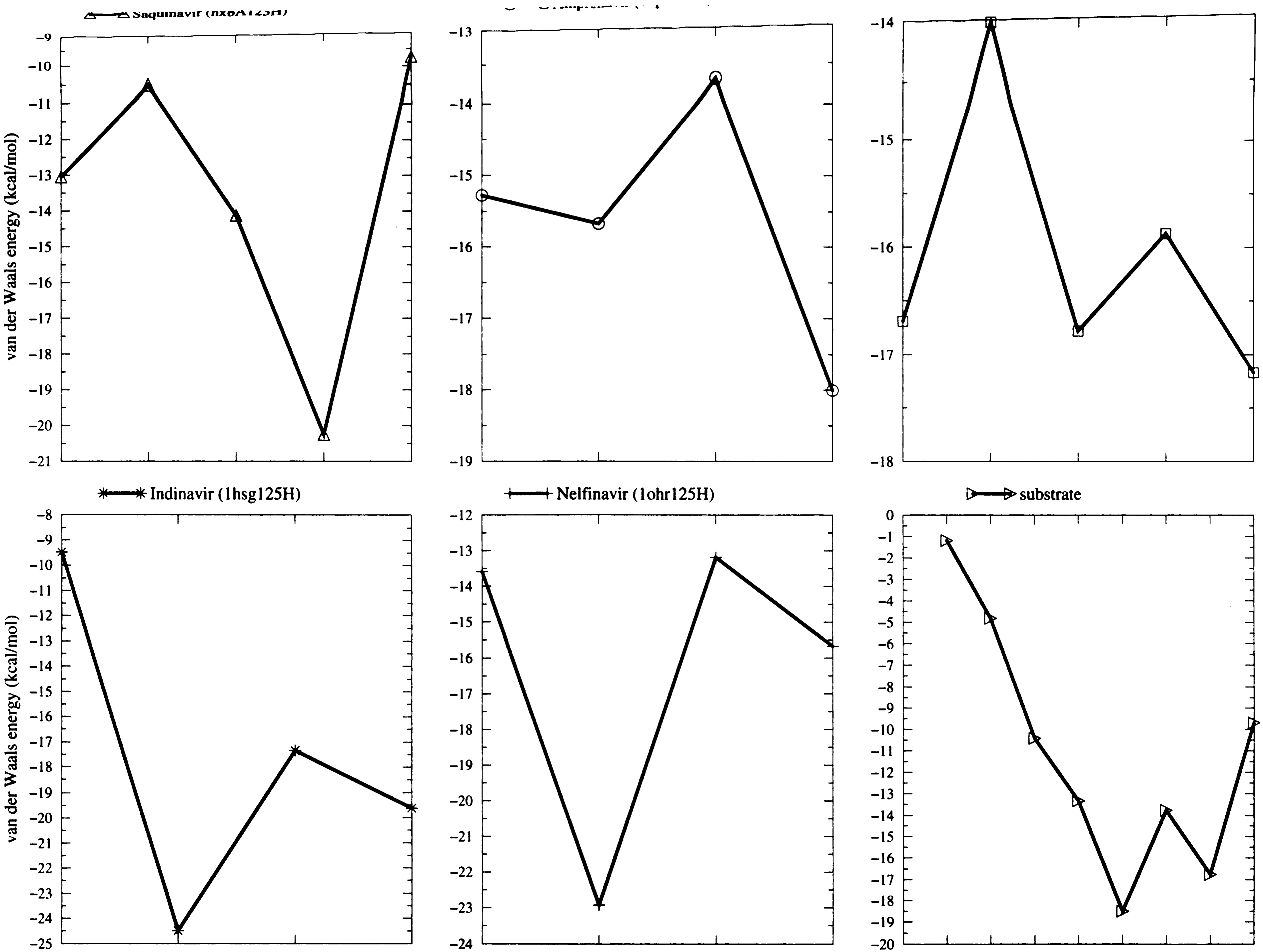

Saquinavir (nxbA125H)
van der Waals energy (kcal/mol)
-9
-10
-11
-12
-13
-14
-15
-16
-17
-18
-19
-20
-21
-13
-14
-15
-16
-17
-18
-19
-14
-15
-16
-17
-18
Indinavir (1hsg125H)
Nelfinavir (1ohr125H)
substrate
van der Waals energy (kcal/mol)
-8
-9
-10
-11
-12
-13
-14
-15
-16
-17
-18
-19
-20
-21
-22
-23
-24
-25
-12
-13
-14
-15
-16
-17
-18
-19
-20
-21
-22
-23
-24
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
-10
-11
-12
-13
-14
-15
-16
-17
-18
-19
-20

Table 1. Binding free energies of the substrate, inhibitor A77003, and 5 FDA approved drugs.

Name	Expt'l ΔG_b (kcal/mol)	ΔG_{int}^{vdw} (kcal/mol)	ΔG_{int}^{ele} (kcal/mol)	ΔG^{nonpol} (kcal/mol)	$\varepsilon_{in}=1$, $\varepsilon_{out}=80$		
					ΔG_{sol}^{ele} (kcal/mol)	$\Delta G_{int+sol}^{ele§}$ (kcal/mol)	ΔG_b^{*} (kcal/mol)
Substrate	N/A	-88.3±0.1	-80.6±1.1	-7.6±0.0	+161.1±1.4	+80.6±0.3	-15.4±0.2
Ritonavir[a] (1hxw125H)	-14.9	-80.5±1.0	-38.4±0.5	-6.9±0.1	+100.8±0.6	+62.4±0.1	-24.9±1.0
Saquinavir[a] (hxbA125H)	-14.3	-67.6±0.3	-24.6±1.9	-6.6±0.1	+72.0±2.0	+47.4±0.1	-26.8±0.1
Amprenavir[a] (1hpv125H)	-13.9	-62.6±0.5	-49.6±0.1	-5.1±0.1	+96.5±0.8	+46.9±0.9	-20.8±0.4
Indinavir[a] (1hsg125H)	-13.3	-70.9±1.8	-31.7±3.8	-6.3±0.1	+86.3±3.5	+54.6±0.3	-22.6±2.2
Nelfinavir[a] (1ohr125H)	-13.0	-65.3±2.3	-36.8±0.8	-5.7±0.0	+82.8±0.8	+45.9±1.6	-25.1±0.6
A77003 (1hvi25H)	-15.1	-87.0±0.0	-46.3±0.6	-7.2±0.0	+113.9±1.5	+67.7±0.9	-26.6±0.9

a. K_i's were measured at pH=6.5[21].

Table 2. Predictions of resistant mutations based on FV value calculations.

Residue #	Saquinavir (hxbA125H)					
	Expt'l K_i^{mutant}/K_i^{wt}	FV_{M1} value	FV_{M2} value	FV_{avg} value	FV_{double} value	Resistant?
L24	1.0->43.0(MM[a])	-0.7±0.0	-0.6±0.1	-0.6±0.0	-0.5±0.1	N
D30	0.5-1.0 (MM)	+13.0±1.7	-4.4±0.6	+4.3±0.9	+7.4±1.4	N
V32	1.6-7.3[22] 1.0-4.0(MM)	-0.6±0.2	-0.6±0.1	-0.6±0.1	-0.4±0.3	N[b] (N[c])
M46	1.0[23] 3.6[22] (M46I) 1.0-125.0(MM)	-0.7±0.2	-1.2±0.1	-0.9±0.1	-0.8±0.2	N (N)
I47	1.0[23] (I47V) 1.0-41.0(MM)	+1.7±0.7	-0.3±0.3	+0.7±0.4	+1.9±0.6	N (N)
G48	**13.5 [24] 163.6[21](G48V) 1.0-100.0(MM)**	**-2.6±1.0**	**-1.0±0.8**	**-1.8±0.7**	**+2.5±1.2**	**N (Y)**
I50	**21 (I50V)[23] 1.0-41.0(MM)**	**-0.8±0.4**	**-0.8±0.5**	**-0.8±0.3**	**-0.2±0.6**	**N (Y)**
F53	4.0-270.0(MM)	-0.3±0.3	-0.9±0.1	-0.6±0.2	-0.2±0.4	N
I54	1.0-580.0(MM)	-0.8±0.1	-0.8±0.1	-0.8±0.1	-0.8±0.1	N
G73	1.0-270.0(MM)	-0.9±0.0	-0.8±0.1	-0.9±0.0	-0.8±0.1	N
V82	1.0-2.0 (V82A/T/F) 0.7-3.7 (V82A/F/I)[22] 3.3-7.3 (V82F/A/I)[25] 1.0-8.0(MM)	+0.7±0.3	-1.0±0.2	-0.2±0.2	+1.1±0.5	N (N)
I84	5.8(I84V)[22] 10.7(I84V)[25] 12.0 (I84V)[23] 1.0-125.0(MM)	-0.7±0.4	-0.3±0.4	-0.5±0.3	-0.8±0.5	N (N)
N88	0.5-2.0(MM)	-0.7±0.1	-1.2±0.1	-0.9±0.1	-0.7±0.1	N
L90	3.0[24,26] 20.7[21] (L90M) 4.9-120.0(MM)	-0.4±0.0	-0.3±0.0	-0.3±0.0	-0.4±0.0	N (N)

a. MM means multiple mutations; b. The criterion used for predicting resistance is $FV_{double} \leq -1.0$; c. If $K_i^{mutant}/K_i^{wt} > 10$ for single mutation, that residue is considered to be resistant (in parenthesis).

Prediction accuracy 6/8=75%

Residue #	Indinavir (1hsg125H)					
	Expt'l K_i^{mutant}/K_i^{wt}	FV_{M1} value	FV_{M2} value	FV_{avg} value	FV_{double} value	Resistant
L24	?	-0.9±0.0	-1.1±0.0	-1.0±0.0	-0.9±0.1	N
D30	0.5-3.0 (MM[a])	+1.8±1.3	-3.1±0.9	-0.6±0.8	+7.3±1.1	N
V32	8.0(V32I)[22] 2.0-36.0 (MM)	-1.0±0.2	-1.1±0.2	-1.0±0.1	-0.9±0.3	N[b] (N[c])
M46	**4.3 (M46I)[22]** **2.0-47.0(MM)**	**-1.9±0.2**	**-2.1±0.1**	**-2.0±0.1**	**-1.6±0.2**	**Y (N)**
I47	3.0 (I47V) 3.0-29.0(MM)	+1.4±0.5	-1.5±0.4	-0.1±0.2	+1.0±0.7	N (N)
G48	6.3(G48V)[21] 1.0-4.0(MM)	-0.1±0.9	-3.0±1.1	-1.5±0.7	+3.4±1.1	N (N)
I50	1.0-29.0(MM)	-3.7±0.4	-0.4±0.6	-2.0±0.4	-1.6±0.6	Y
F53	?	-1.0±0.3	-1.8±0.1	-1.4±0.2	-0.9±0.4	N
I54	1.0-100(MM)	-1.5±0.1	-1.4±0.1	-1.4±0.1	-1.4±0.1	Y
G73	1.0->100.0(MM)	-1.5±0.0	-1.5±0.0	-1.5±0.0	-1.5±0.0	Y
V82	0.6-6.4 (V82A/F/I)[25] 6.9-84.7 (V82A/F/I)[22] 2.0-29.0(MM)	-0.6±0.3	-1.3±0.4	-1.0±0.3	+0.2±0.5	N (N)
I84	2.6(I84V)[25] 10.0 (I84V)[22] 2.0-47.0(MM)	-1.3±0.4	-0.4±0.4	-0.8±0.3	-0.9±0.6	N (N)
N88	0.5-36.0(MM)	-1.7±0.1	-2.0±0.1	-1.8±0.1	-1.4±0.1	Y
L90	5.8(L90M)[21] 2.0-29.0(MM)	-0.6±0.0	-0.6±0.0	-0.6±0.0	-0.7±0.0	N (N)

a. MM means multiple mutations; b. The criterion used for predicting resistance is $FV_{double} \leq -1.0$; c. If $K_i^{mutant}/K_i^{wt} > 10$ for single mutation, that residue is considered to be resistant (in parenthesis).

Prediction accuracy 6/7=86%

Residue #	Ritonavir (1hxw125H)					
	Expt'l K_i^{mutant}/K_i^{wt}	FV_{M1} value	FV_{M2} value	FV_{avg} value	FV_{double} value	Resistant
L24	2.0-203.0(MM[a])	-0.4±0.0	-0.5±0.1	-0.5±0.0	-0.3±0.1	N
D30	0.4-1.0 (MM)	+6.4±1.4	-3.0±1.0	+1.7±0.8	+7.1±1.2	N
V32	3.0-72.0(MM)	-0.4±0.2	-0.5±0.1	-0.4±0.1	-0.4±0.3	N
M46	**4(M46I)[23]** **1.0-80.0(MM)**	**-0.5±0.2**	**-0.8±0.1**	**-0.7±0.1**	**-0.5±0.2**	**N[b] (N[c])**
I47	**3(I47V)[23]** **16.0-21.0(MM)**	**+2.2±0.5**	**-0.8±0.4**	**+0.7±0.3**	**+1.5±0.7**	**N (N)**
G48	**66.7(G48V)[21]** **2.0-4.4(MM)**	**-1.6±1.0**	**-1.0±0.8**	**-1.3±0.6**	**+5.0±1.0**	**N (Y)**
I50	**10(I50V)[23]** **0.06-0.1(MM)**	**-0.9±0.3**	**-1.0±0.5**	**-1.0±0.3**	**-0.2±0.5**	**N**
F53	17.0-37.0(MM)	+0.1±0.3	-0.6±0.1	-0.3±0.2	-0.1±0.4	N
I54	3.0->203.0(MM)	-0.6±0.1	-0.4±0.1	0.5±0.1	-0.4±0.1	N
G73	1.0-164.0(MM)	-0.6±0.0	-0.6±0.0	-0.6±0.0	-0.6±0.0	N
V82	**0.8-14.7** **(V82A/F/I)[25]** **16.0-72.0 (MM)**	**+0.4±0.4**	**-0.7±0.3**	**-0.1±0.2**	**+0.7±0.5**	**N (Y)**
I84	**11.2(I84V)[25]** **20(I84V)[23]** **11.3-80.0(MM)**	**-0.4±0.4**	**+0.2±0.4**	**-0.1±0.2**	**-0.0±0.5**	**N (Y)**
N88	0.06-54.0(MM)	-0.7±0.1	-1.0±0.1	-0.8±0.1	-0.4±0.1	N
L90	**6.7(L90M)[21]** **2.0-71.0(MM)**	**-0.2±0.0**	**-0.2±0.0**	**-0.2±0.0**	**-0.3±0.0**	**N (N)**

a. MM means multiple mutations; b. The criterion used for predicting resistance is $FV_{double} \leq -1.0$; c. If $K_i^{mutant}/K_i^{wt} > 10$ for single mutation, that residue is considered to be resistant (in parenthesis).

Prediction accuracy 4/7=57%

Residue #	Amprenavir (1hpv125H)					
	Expt'l K_i^{mutant}/K_i^{wt}	FV_{M1} value	FV_{M2} value	FV_{avg} value	FV_{double} value	Resistant
L24	1.0-30.0(MM[a])	-1.1±0.0	-0.9±0.0	-1.0±0.0	-0.8±0.1	N
D30	0.4-2.0 (MM)	+6.6±1.3	-13.3±0.9	-3.4±0.6	+2.5±1.0	N
V32	0.4-82.0(MM)	-1.1±0.2	-1.2±0.2	-1.2±0.1	-1.1±0.3	Y
M46	**1.0(M46I)[23]** **0.4-270.0**	**-1.7±0.2**	**-2.2±0.1**	**-2.0±0.1**	**-1.7±0.2**	**Y[a] (N[b])**
I47	**1.0 (I47V)[23]** **1.0-200.0(MM)**	**+1.0±0.5**	**-0.4±0.2**	**+0.3±0.2**	**+1.5±0.6**	**N (N)**
G48	**3.5[21]** **1.0-2.3(MM)**	**+1.9±0.8**	**-3.0±0.9**	**-0.6±0.4**	**+3.7±1.1**	**N (N)**
I50	**83(I50V)[23]** **0.15-270(MM)**	**-3.2±0.3**	**-1.4±0.6**	**-2.3±0.2**	**-2.1±0.5**	**Y (Y)**
F53	7.0-22.0(MM)	-1.0±0.3	-1.4±0.1	-1.2±0.1	-0.9±0.4	N
I54	1.0-69.0(MM)	-1.4±0.1	-1.5±0.1	-1.4±0.0	-1.3±0.1	Y
G73	1.0-47.0(MM)	-1.6±0.0	-1.4±0.0	-1.5±0.0	-1.5±0.0	Y
V82	**0.4-3.3 (V82A/F/I)[25]** **1.0-82.0(MM)**	**+0.1±0.3**	**-1.3±0.5**	**-0.6±0.2**	**+0.9±0.6**	**N (N)**
I84	**23.0 (I84V)[23]** **2.7(I84V)[25]** **2.0-51.0(MM)**	**-0.9±0.4**	**-0.7±0.4**	**-0.8±0.2**	**-1.2±0.5**	**Y (Y)**
N88	Hypersensitivity 0.04-9.0(MM)	-1.8±0.1	-2.0±0.1	-1.9±0.0	-1.6±0.1	Y
L90	**2.7[21]** **1.0-28.0(MM)**	**-0.6±0.0**	**-0.5±0.0**	**-0.6±0.0**	**-0.7±0.0**	**Y (Y)**

a. MM means multiple mutations; b. The criterion used for predicting resistance is $FV_{double} \leq -1.0$; c. If $K_i^{mutant}/K_i^{wt} > 10$ for single mutation, that residue is considered to be resistant (in parenthesis).

Prediction accuracy 6/7=86%

Residue #	Nelfinavir(1ohr125H)					
	Expt'l K_i^{mutant}/K_i^{wt}	FV_{M1} value	FV_{M2} value	FV_{avg} value	FV_{double} value	Resistant
L24	2.0->56.0(MM[a])	-0.9±0.0	-0.9±0.0	-0.9±0.0	-0.9±0.1	N
D30	7.1-35.0 (MM)	+7.3±1.4	-1.9±0.6	+2.7±0.7	+7.1±1.3	N
V32	1.0->32.0(MM)	-0.5±0.2	-1.1±0.2	-0.8±0.1	-0.7±0.3	N
M46	1.0-130.0(MM)	-1.7±0.2	-1.9±0.1	-1.8±0.1	-1.6±0.2	Y
I47	1.0-11.0(MM)	+1.2±0.5	-1.1±0.3	+0.0±0.3	+1.4±0.6	N
G48	**1.0(G48V)[21] 2.0->98.0(MM)**	**+2.5±0.8**	**-0.2±0.8**	**+1.2±0.6**	**+4.0±1.0**	**N[b] (N[c])**
I50	1.0-5.0(MM)	-1.1±0.3	-0.3±0.5	-0.7±0.3	-0.8±0.5	N
F53	10.0-130.0(MM)	-1.1±0.3	-1.8±0.1	-1.4±0.2	-0.9±0.4	N
I54	4.0-479.0(MM)	-1.5±0.0	-1.6±0.1	-1.5±0.1	-1.5±0.1	Y
G73	3.0-130.0(MM)	-1.5±0.0	-1.5±0.0	-1.5±0.0	-1.5±0.0	Y
V82	**1.0-4.9(V82F/A/I)[25] 1.0-121.7(MM)**	**+0.2±0.5**	**-0.3±0.3**	**-0.1±0.3**	**+1.5±0.5**	**N (N)**
I84	**0.9(I84V)[25] 2.0-95.0(MM)**	**-1.7±0.4**	**-0.6±0.4**	**-1.1±0.3**	**-1.2±0.5**	**Y (N)**
N88	1.0-479.0(MM)	-1.4±0.1	-1.8±0.1	-1.6±0.1	-1.5±0.1	Y
L90	**3.5 (L90M)[21] 2.0-479.0(MM)**	**-0.7±0.0**	**-0.7±0.0**	**-0.7±0.0**	**-0.7±0.0**	**N (N)**

a. MM means multiple mutations; b. The criterion used for predicting resistance is $FV_{double} \leq -1.0$; c. If $K_i^{mutant}/K_i^{wt} > 10$ for single mutation, that residue is considered to be resistant (in parenthesis).

Prediction accuracy 3/4=75%

Average prediction accuracy of 5 drugs is 76%

Chapter 6. Philosophy and further test of the combination of free energy and evolutionary information

1. Philosophy behind the combination of free energy and evolutionary information.

The philosophy behind the combination of free energy and evolutionary information is: (1) the result of free energy calculations or sequence analysis is not unambiguous and a combination of both of them might help reduce uncertainties; (2) free energy calculations only can be performed on proteins whose structures are available and predictions of functions of proteins with no structure available can be made based on evolutionary information.

Free energy calculations obey physical principles, but it is not fully clear what laws evolution obeys. One law that it appears to follow is the Darwinian Law. If we believe that physical principles are the basis of all phenomena in nature, then the Darwinian Law should not be an exception. Of course, evolution has its own unique features: it depends on history and many stochastic processes occur in the progress. My opinion is that the origin of life is a stochastic event; once life appears, evolution is still not a deterministic process but the probability of any evolutionary event should be able to be described by physical laws, which we have not totally understood yet. It may suggest that we need a new physics, new physics concepts and probably new mathematic tools, which can explain phenomena at a much larger scale than those being explained by statistical mechanics. On the other hands, evolution is still going on and has not reached an optimum as yet. However, our goal is to understand the current biology, not necessarily the optimal biology. Therefore, evolutionary information is helpful for understanding such as protein function in the current world. Given above, the combination of free energy and evolutionary information is somewhat analogous to the combination of van der Waals, solvation free energy and other terms as an estimate of the total free energy. It is not rigorous but may well be practically useful.

Sequencing of a protein is much easier than solving its crystal structure. Thus, there are many more sequences available than crystal structures. Nowadays, structure based free energy calculations require considerable human effort and computer time; thus, one would like to fully use such results, if possible, in understanding functions of other proteins whose structures are not solved yet.

The VC (van der Waals/conservation) or FC (free energy/conservation) value (see below) are examples of simple ways to combine free energy and evolutionary information. There are surely alternative and more sophisticated ways to achieve the same goal. Therefore, further investigation of the VC and FC values is appropriate.

The VC and FC values can identify critical residues for binding or folding. If we look at free energy and conservation separately, we can find those residues specifically important for a certain protein. For example, N190 in the Sem-5 SH3 domain has a very favorable van der Waals interaction energy with the ligand (-2.8 kcal/mol) but low conservation percentage in the sequence alignment (16%). Therefore, this suggests that N190 might contribute to the specificity of the Sem-5 SH3 domain.

In summary, a combination of free energy and evolutionary information can suggest residues universally important for a protein superfamily as well as residues specifically important for a certain protein or protein family.

2. Definitions of the VC value, the FV value and the FC value and the relationship between them.

The VC value, FV value and FC value are defined as product of van der Waals energy and conservation (Chapter 4), free energy and variability (Chapter 5), and free energy and conservation respectively. The VC and FC value are used to identify critical residues for binding or folding. The VC value is used if van der Waals interaction is the dominant factor (see Chapter 4). The FC value is more general. The FV value is designed to consider drug resistance mutations where these mutations only occur at positions that are poorly conserved (see Chapter 5).

The free energy for each residue ΔG_{res} can be estimated using equations (5) and (6) in Chapter 5 and is summarized below:

$$\Delta G_{res} = E_{vdw} + E_{ele} + \Delta G^{sol} - \Delta G_0^{sol} \qquad (1)$$

where E_{vdw} and E_{ele} are van der Waals and electrostatic interaction energies between the residue and the whole ligand respectively, ΔG^{sol} and ΔG_0^{sol} are the solvation energies calculated by solving the Poisson-Boltzmann equation with normal partial charges and zero charges on the specific residue, respectively.

Conservation C_i is calculated as:

$$C_i = \Sigma_j (P_{ij}/P_{ii}) * W_j \qquad (2)$$

where W_j is the weight of the jth sequence. W_j is calculated for each sequence in the alignment based on sequence identity. If n sequences are >80% identical to each other, each sequence has 1/n weight. Next, the sum of all sequences in the alignment is normalized to 1. This weight prevents over-presenting very similar sequences in the Psi-BLAST search results.

P_{ij} in Equation (2) represents how likely the amino acid a_i in the ith sequence can be mutated to the amino acid a_j in the jth sequence and is calculated as:

$$P_{ij} = 2^{(2*M_{ij})} \qquad (3)$$

where M_{ij} is the element of BLOSUM62 for a_i and a_j. BLOSUM62 is chosen to be consistent with the matrix used in Psi-BLAST search. M_{ij} for gap is assigned a penalty score of -4 in the BLOSUM62 matrix.

Variability V_i is defined as (Chapter 5):

$$V_i = \Sigma_j (1-P_{ij}/P_{ii}) * W_j \qquad (4)$$

It is obvious that $C_i + V_i = 1$.

Entropy S can also be used to represent how variable one position is. Usually the 20 amino acids are divided into different groups based on their, such as, hydrophobicity. For instance, one classification is to divide the 20 amino acids into 6 groups: (1). I, L, V, M, A; (2). D, E, N, Q; (3). K, R; (4). F, Y, W, H; (5). T, S, C, P; (6). G. We can calculate entropy S as:

$$S = -\Sigma_j p_j \ln p_j \qquad (5)$$

where p_j is the appearance percentage of the jth type of amino acid.

The difference between variability and entropy is: variability represents how conserved the specific residue of the sequence on which we perform free energy calculations is; but the entropy reflects how many different types of amino acids appear at that position. For example, in the sequence alignment for the Sem-5 SH3 domain (Chapter 4) we are interested in position 166. At this position, Sem-5 SH3 domain has a Asn. In the multiple sequence alignment at this position, Asn is found only 2% of the time but Asp and Glu are found 18% and 24%, respectively. The entropy at this position is a medium value but the conservation for Asn is small and the variability is large. The free energy or van der Waals energy is calculated for Asn in Sem-5 SH3 domain.

Correspondingly, conservation or variability of Asn at this position is more useful than entropy if we want to combine free energy calculation and sequence analysis into a single parameter. Therefore, in Chapter 4, we used the VC value which combines van der Waals energy and conservation.

3. Further test of the combination of free energy and evolutionary information.

A combination of free energy calculation and sequence analysis has been shown to be useful in studying Sem-5 SH3 domain binding with ligands (Chapter 4) and the HIV drug resistance (Chapter 5) and, thus, it is worth of further investigations. Here I propose more applications of this method to identify critical residues for the SH3 domain folding stability and the human growth hormone binding to its receptor.

Baker and coworkers found a structurally polarized transition state for folding of the Src SH3 domain[1]. Their experiments showed that a hydrogen bond network was always formed between S47, T50 and E30 but other interactions were partially formed in the transition state. If this hydrogen bond network is disrupted, the folding rate is significantly reduced. This raises a challenge for computational modeling as it requires a precise identification of these 3 crucial residues. I suggest that it would be interesting to calculate FC value for each residue of the SH3 domain and see if we can identify those residues critical in the folding kinetics.

The human growth hormone (hGH) and its receptor (hGHbp) belong to cytokine superfamily, which controls a number of immune and growth functions in the cell. Using mutational studies and structure analysis, Wells and coworkers found that a small number of residues on the interface between the hGH and the extracellullar domain of its receptor, called the hGHbp, can account for most of the binding affinity[2]. These critical residues form small patches in where well packed hydrophobic residues are surrounded by polar and charged residues. They have substituted most residues on the hGH-hGHbp interface to Alanine and measured the changes in K_i's due to mutations.

This provides an ideal system to quantitatively test the idea of combining free energy calculation and sequence analysis to identify critical residues for binding. Preliminary results show that W104 and W169 on the hGHbp have the most favorable van der Waals interaction energies with the hGH. These two residues have been shown to be most critical for binding in Alanine scanning experiments. They are also well conserved in the sequence analysis. It will be interesting to see if there is any correlation between the FC (free energy/conservation) value and the $\Delta\Delta G_b$ of binding.

The growth hormone not only can bind with its receptor, its also can bind to prolactin receptor. One complex structure of hGH and prolactin receptor has been solved (PDB entry 1bp3). We can perform the same calculations on this structure and identify critical residues for binding between hGH and prolactin receptor. A comparison between critical residues on this complex and those on the hGH-hGHbp complex will be interesting. It can suggest how reliable our prediction is on specificity and function of another protein in the same family based on detailed studies of one complex.

It is worth pointing out that the FC value identifies residues critical for binding, i.e. these residues are not tolerant to mutations. Therefore, mutations to all types of amino acids but not just Alanine are required to evaluate if one residue is crucial for binding. The FC value analysis can thus provide a guide for mutagenesis experiments.

4. References:

1. Grantcharova, V.P., Riddle, D.S., Santiago, J.V. & Baker, D. Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nature Structural Biology* **5**, 714-720 (1998).

2. Wells, J.A. Binding in the Growth Hormone Receptor Complex. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 1-6 (1996).

Chapter 7. What Determines the van der Waals Coefficient β in the LIE (Linear Interaction Energy) Method to Estimate Binding Free Energies Using Molecular Dynamics Simulations?

This chapter is a reprint of a published paper (Wang, W., Wang, J., and Kollman, P. A., Proteins: Structure, Function and Genetics, 1999, 34, 3, 395-402.)

What Determines the van der Waals Coefficient β in the LIE (Linear Interaction Energy) Method to Estimate Binding Free Energies Using Molecular Dynamics Simulations?

Wei Wang,
Graduate Group in Biophysics,
University of California, San Francisco
San Francisco, CA 94143
and
Jian Wang and Peter A. Kollman*
Department of Pharmaceutical Chemistry
University of California, San Francisco
San Francisco, CA 94143

* Author for correspondence

Tel: (415)476-4637 (PAK)

Fax: (415)476-0688

Email: pak@cgl.ucsf.edu

Short title: What determines β in the LIE.

Keywords: molecular dynamics, ligand-protein interactions, van der Waals energies, electrostatic energies, desolvation.

Grant sponsor: NIH; Grant number: GM-29072 (PAK, P.I.), GM-56609 (E. Arnold, P.I.).

Published in Proteins: Structure, Function and Genetics, 1999, 34, 3, 395-402.

Abstract

Recently a semiempirical method has been proposed by Åqvist *et al* [1,2,3] to calculate absolute and relative binding free energies. In this method, the absolute binding free energy of a ligand is estimated as $\Delta G_{\text{bind}} = \alpha < V^{\text{el}}_{\text{bound}} - V^{\text{el}}_{\text{free}} > + \beta < V^{\text{vdw}}_{\text{bound}} - V^{\text{vdw}}_{\text{free}} >$, where $V^{\text{el}}_{\text{bound}}$ and $V^{\text{vdw}}_{\text{bound}}$ are the electrostatic and van der Waals interaction energies between the ligand and the solvated protein from an molecular dynamics (MD) trajectory with ligand bound to protein and $V^{\text{el}}_{\text{free}}$ and $V^{\text{vdw}}_{\text{free}}$ are the electrostatic and van der Waals interaction energies between the ligand and the water from an MD trajectory with the ligand in water. A set of values, $\alpha = 0.5$ and $\beta = 0.16$, was found to give results in good agreement with experimental data. Later, however, different optimal values of β were found in studies of compounds binding to P450cam[4] and avidin.[5] The present work investigates how the optimal value of β depends on the nature of binding sites for different protein-ligand interactions. By examining seven ligands interacting with five proteins, we have discovered a linear correlation between the value of β and the weighted non-polar desolvation ratio (WNDR), with a correlation coefficient of 0.96. We have also examined the ability of this correlation to predict optimal values of β for different ligands binding to a single protein. We studied twelve neutral compounds bound to avidin. In this case, the WNDR approach gave a better estimate of the absolute binding free energies than results obtained using the fixed value of β found for biotin-avidin. In terms of reproducing the relative binding free energy to biotin, the fixed β value gave better

results for compounds similar to biotin, but for compounds less similar to biotin, the WNDR approach led to better relative binding free energies.

I. Introduction

Free energy perturbation (FEP) and thermodynamic integration (TI) methods are rigorous approaches to evaluate binding free energies of ligands to a receptor. However, sampling and convergence problems associated with these approaches prevent them from being widely used in structure-based drug design.[4,5,6] Thus, development of fast and accurate methods to be used for structure based drug design is still very much needed.[4,5,6] Åqvist *et al* have recently proposed a semiempirical molecular dynamics method, termed the linear interaction energy (LIE) approximation, for the estimation of absolute binding free energies.[1,2,3] It is faster than FEP or TI because it does not sample any intermediate state between the initial and final states. The method has been applied to several different systems and the results are in good agreement with experimental data.

The LIE method is based on linear response assumptions.[1,2,3] It divides the interaction between the ligand and its environment into electrostatic and van der Waals parts. The binding free energy is estimated as

$$\Delta G_{\text{bind}} = \Delta G_{\text{bind}}^{\text{el}} + \Delta G_{\text{bind}}^{\text{vdw}}$$

$$\approx \alpha < V_{\text{bound}}^{\text{el}} - V_{\text{free}}^{\text{el}} > + \beta < V_{\text{bound}}^{\text{vdw}} - V_{\text{free}}^{\text{vdw}} > \quad (1)$$

where $V_{\text{bound}}^{\text{el}}$ and $V_{\text{bound}}^{\text{vdw}}$ are the electrostatic and van der Waals interaction energies between the ligand and the solvated protein from an molecular dynamics (MD) trajectory with ligand bound to protein and $V_{\text{free}}^{\text{el}}$ and $V_{\text{free}}^{\text{vdw}}$ are the electrostatic and van der Waals interaction energies between the ligand and the water from an MD trajectory with the ligand in water, $< >$ denotes an ensemble average, and α and β are two empirical parameters.

For several protein systems,[1,2,3] Åqvist *et al* have found that $\alpha = 0.5$ and $\beta = 0.16$ gave calculated binding free energies in good agreement with experimental data. Paulsen and Ornstein, however, found that $\alpha = 0.5$ and $\beta = 1.043$ resulted in good estimation of the binding free energies of 11 substrates binding to cytochrome P450cam.[7] The difference between the two sets of parameters was rationalized as perhaps due to different force fields, GROMOS and CVFF respectively, used in the two studies.[7] Wang *et al* [8] applied this method to calculate binding free energies of 14 compounds binding to avidin using the Cornell *et al* force field.[9] Their results showed that $\alpha = 0.5$ and $\beta = 1.0$ gave reasonable estimates of the binding free energies with respect to the corresponding experimental results.

These studies raise an interesting question: can one set of α and β be used in different protein-ligand complexes to give reasonable estimates of binding free energies? Although Wang *et al* used the Cornell *et al* force field,[9] they found similar values of α and β as did Åqvist *et al* for the trypsin-benzamidine complex.[8] This suggests that the use of different force fields can not explain the difference in α and β found in different simulations.

What leads to a common value of 0.5 for α, albeit Åqvist has shown α may be reduced from this value for ligands containing OH groups?[25] The value of $\alpha = 0.5$ came from the first order approximation of electrostatic contribution to the binding free energy.[4] This 0.5 also appears in other semiempirical methods, such as Generalized Born model (GB). It has been shown that this first order approximation is reasonable.[19] Thus, the use of $\alpha = 0.5$ has a physical justification. On the other hand, there is no similar argument to obtain β, which has been derived empirically. It is reasonable to think that the value of β is binding site dependent since it is the scaling factor for van der Waals interactions. Thus, the question is, is there a factor to describe the nature of binding sites and can one relate this factor to the value of β?

In the present work, we have applied the LIE method to a variety of protein-ligand systems and have tried to answer the above question. In our study, α has been kept as 0.5 (as discussed above) and β is adjustable. We defined a ratio factor which we termed as the weighted non-polar desolvation ratio (WNDR) (described below). We studied seven different complexes whose binding affinities were known experimentally and examined the correlation between β and WNDR. β was optimized separately for each complex to reproduce the experimental binding free energy. WNDR of these seven complexes were also calculated. For these seven complexes we have observed a linear correlation between the value of β and WNDR.

We then used this observed linear correlation to calculate the binding free energies for 12 neutral compounds bound to avidin. WNDR was calculated for each compound and used to pick a separate β value for each ligand. In this case, the WNDR approach gave a better estimate of absolute binding free energies than the results obtained using the fixed value of β found for biotin-avidin. In terms of relative free energies of binding to biotin, the fixed β gave better results for compounds similar to biotin than the WNDR approach. On the other hand, for dissimilar compounds, better relative binding free energies were obtained using the WNDR approach.

II. Methods

All calculations presented in this work were performed using the AMBER5.0[10] simulation package and the Cornell *et al* force field[9] with TIP3P[11] water model. RESP[12] charges were used for all ligand atoms. For each system a pair of simulations was performed, one with the ligand in a 20 Å sphere of waters, the other with the ligand bound to the protein with a cap of waters around the complex. The cap of waters around the complexes were filled up to 20 Å from the center of mass of the ligand. All simulations were carried out at 300K. The SHAKE[13] procedure was employed to constrain all bonds connecting hydrogen atoms. The time step of the simulations was 1.5 fs with a dual cutoff of 10 Å and 17 Å for the nonbonded interactions. The nonbonded pairs were updated every 30 steps. All atoms within 18 Å of the center of mass of the ligand were allowed to move. Atoms between 18 Å and 20 Å were restrained by a 20 kcal/mol/Å^2 harmonic force. A 100 kcal/mol/Å^2 harmonic position restrain was applied on the center of mass of the ligand in each simulation. Electrostatic and van der Waals interaction energies between the ligand and its environment, i.e. water molecules in the ligand free state or protein residues and water molecules in the ligand bound state, within the 20 Å sphere centered at the center of mass of the ligand were calculated using the CARNAL and ANAL modules of AMBER.[10]

Since adding counter ions to the system leads to slow convergence of the simulations,[5] we followed Åqvist *et al*'s approach to maintain a neutral protein system in the simulations by changing the charge state of some charged residues.[5] Specifically, the farthest charged residues from the center of mass of the ligand were turned off to keep the 20 Å sphere of the protein neutral. The protonation states of charged residues beyond 20 Å were also adjusted to neutralize the entire system.

Prior to MD simulations, a series of minimizations were carried out using a protocol in which all heavy atoms were restrained with 5,000, 1,000, 100, and 10 kcal/mol/Å^2 harmonic forces respectively. The maximum number of minimization steps was 50,000 steps and the convergence criterion for the energy gradient was 0.5 kcal/mol/Å. Data collection was performed after a 50 ps equilibration. It took 100 to 300 ps to obtain converged average energies. Our criteria for convergence was that the average energies in the two halves of the trajectory differ by less than 5 kcal/mol. The convergence was checked for every simulation (see below). Solvent accessible surfaces (SAS) were calculated using program MSMS[14] after all hydrogen atoms were removed from the PDB files.

When protein and ligand bind to each other, the solvent accessible surface (SAS) of the complex is smaller than the sum of SAS of protein and ligand before binding, because part of the protein and ligand which are exposed to water in the free states are buried upon binding. We termed the loss of the SAS due to binding as the total desolvation SAS (tdSAS). It is calculated as following:

$$\text{total desolvation SAS(tdSAS)} = \text{SAS}_{complex} - \text{SAS}_{protein} - \text{SAS}_{ligand} \qquad (2)$$

Obviously, the total desolvation solvent accessible surface of atom i (tdSAS^{i}) due to binding was calculated as:

$$\text{total desolvation SAS}^{i}\,(\text{tdSAS}^{i}) = \text{SAS}^{i}_{complex} - \text{SAS}^{i}_{protein} - \text{SAS}^{i}_{ligand} \qquad (3)$$

where i = C, N, O, N^+,O^- and S, following the atom classes defined by Eisenberg and McLachlan.[17]

The non-polar desolvation ratio (NDR) is defined as the ratio of total desolvation SAS of all nonpolar groups, carbon and sulfur atoms in this work, to the total desolvation SAS (tdSAS).

$$\text{NDR} = (\text{tdSAS}^{C} + \text{tdSAS}^{S})/\text{tdSAS} \qquad (4)$$

where tdSAS^{C} and tdSAS^{S} represent total desolvation SAS of carbon and sulfur atoms respectively. NDR roughly reflects how hydrophobic the binding site is.

In order to get a more accurate representation of the hydrophobicity of the binding site, different groups on the surface of the binding site should not be treated equally. For example, burial of charged groups is more unfavorable than burial of polar groups. So the desolvation SAS for different groups have been weighted differently.

The total weighted desolvation SAS (twdSAS) due to binding was calculated as:

$$\text{total weighted desolvation SAS(twdSAS)} = \sum_{i} (\sigma(i) \times \text{dSAS}_{i}) \qquad (5)$$

where $\sigma(i)$ is surface tension parameters taken from Eisenberg and McLachlan's work.[17] These values are listed in Table I.

The weighted nonpolar desolvation ratio (WNDR) was defined as the ratio of all nonpolar groups' weighted desolvation SAS to total weighted desolvation SAS:

$$\mathrm{WNDR} = ((\sigma(\mathrm{C}) \times \mathrm{dSAS}^{C})+(\sigma(\mathrm{S}) \times \mathrm{dSAS}^{S}))/\mathrm{twdSAS} \qquad (6)$$

It should be pointed out that WNDR can be greater than 1 since the weights for polar groups are negative.

Initially, simulations on seven ligands binding to five proteins were done. All these simulations started from crystal structures taken from Protein Data Bank. Their PDB entries are 3ptb, 1dwc, 1dwd, 1aaq, 5hvp, 1avd and 2cpp. The value of α was kept as 0.5 and the values of adjustable parameter β were optimized to fit the experimental data.

The simulations on twelve neutral compounds bound to avidin were performed in a similar way. A rationale for the use of a neutral COOH rather than the COO^- present in biotin and its analogs was presented in ref. 8. The structures for the twelve complexes were obtained using a docking algorithm.[21] The computational details were reported elsewhere.[8]

III. Results and discussion

The average ligand-solvent electrostatic and van der Waals interaction energies for ligand bound and free states are shown in Table II together with the calculated binding free energies. The convergence errors estimated by averaging over the first and second half of the simulation trajectories are also listed in the Table II. The small errors indicate that the averaged results are stable.

Since Åqvist *et al* and Ornstein *et al* obtained different values of β, we calculated the binding free energies on the same systems as they did, namely complexes of trypsin-benzamidine (PDB entry 3ptb)[6] and camphor-P450cam (PDB entry 2cpp).[7] We found two different values of β, 0.14 and 0.81, which gave a good fit to the experimental data respectively. The two values are far from each other and neither can give satisfactory results for both systems. We used the Cornell *et al* force field in these two systems while Åqvist *et al* used GROMOS force field for trypsin-benzamidine and Ornstein *et al* used CVFF force field for camphor-P450cam. The β values we found for these two systems are each close to what Åqvist *et al* ($\beta = 0.16$) and Ornstein *et al* ($\beta = 1.043$) obtained. This further emphasizes what we noted above - the force field is not the determining factor in why β is so different in trypsin from that in P450cam. Different force fields may give different values of interaction energies in the ligand bound and free states. However, taking the interaction energy differences between the two states may lead to some cancellation of the differences between the different force fields.

We applied this method to other five protein-ligand complex systems and the optimal values of β were also different for each. As discussed above, we keep $\alpha = 0.5$ in all our simulations (see Introduction). One fixed β could not give the results which are in good agreement with experimental values in all cases. This is in contrast to previous studies by Åqvist *et al* in which one universal value of β was good for different systems.

The dependence of β on ligand-protein system is not unreasonable, given that the nature of binding sites is different for different proteins: some binding sites have many non-polar residues while other binding sites are mostly composed by polar residues. From the previous study of biotin binding to streptavidin by Miyamoto and Kollman,[14,15] it was found that the binding is dominated by the difference between the van der Waals interaction energies in ligand bound and free state. It appears to be determined by the dispersion attraction in the bound state and the hydrophobic effect (repulsion term) in the ligand free state. On the other hand in the complex of N-acetyl tryptophan with α-chymotrypsin, the binding free energy has a larger contribution from the electrostatic term. This suggested that there might be some correlation between the value of β and the hydrophobicity of the binding site.

Using Equation (1) to calculate binding free energies, the contributions of binding free energy due to electrostatic and van der Waals interactions between the ligand and the protein are considered. The contributions of binding free energies due to desolvation, entropy loss due to reduced conformation freedom, etc, are implicitly included by empirically optimizing the value of β.

In order to study how the value of β depends on the nature of the binding site, as the first step, we examined the non-polar desolvation ratio (NDR), i.e. the ratio of desolvation SAS of non-polar groups to the total desolvation SAS (tdSAS). It is a rough representation of how hydrophobic the binding site is. We plotted the NDR versus the values of β in Figure 1. The linear correlation coefficient is 0.89.

Many previous studies have shown that there is a correlation between the change of solvent accessible surfaces and solvation and binding free energies. Empirical atomic solvation parameters have been obtained for different atom types[17,18] and these parameters have been successfully applied to studying protein stability and protein ligand binding.[19,20] Encouraged by the good correlation between the NDR and β, we used a weighted non-polar desolvation ratio (WNDR see definition in Methods) to discriminate contributions to binding from different groups. It is natural to use the corresponding solvation parameter of each group as its weight. Thus, non-polar groups have positive weights and polar and charged groups have negative weights (see Table I). In the present work, we used the solvation parameters published by Eisenberg and McLachlan.[17] We calculated the WNDR and the optimal β for the seven protein-ligand complexes. The correlation between WNDR and values of β is shown in Figure 2 and the correlation coefficient is 0.96.

It is clear that the Åqvist Equation (1) is a simplification of a more general attempt to represent ΔG_{bind} in terms of components. For example, Equation (7) can be considered as such a general approach.

$$\Delta G_{bind} = \Delta G_{el} + \Delta G_{vdw} + \Delta G_{desolv} + \Delta G_{t/r} + \Delta G_{intra} \qquad (7)$$

where ΔG_{el} and ΔG_{vdw} are contributions to the binding free energy from electrostatic and van der Waals interactions between the ligand and the protein respectively, ΔG_{desolv} is the free energy change due to the desolvation effect, i.e. the loss of solvent accessible surface due to complexation, $\Delta G_{t/r}$ is the translational/rotational free energy change upon complexation, and ΔG_{intra} includes free energy contributions from the conformational entropy loss and the changes of intramolecular energies of the ligand and the protein upon binding.

The Åqvist Equation (1) attempts to "fold in" the last three components of Equation (7) into Equation (1) by using parameters of α and β. Of those three terms that are in Equation (7) but not in Equation (1), $\Delta G_{t/r}$ stands out as not being very molecule dependent and should be roughly constant for typical ligand-protein complexes. Thus, we examined the ability of the following equation to represent the binding data in the seven ligand-protein complexes studied about.

$$\Delta G_{bind} = \alpha < V^{el}_{bound} - V^{el}_{free} > + \beta < V^{vdw}_{bound} - V^{vdw}_{free} > + \Delta G_{t/r} \quad (8)$$

We tested values of $\Delta G_{t/r}$ in the range of 7-11 kcal/mol [26,27] and derived β value for each system . Fitting of these constant values of $\Delta G_{t/r}$ still led to a good fit between the WNDR and β, with correlation coefficient r varying between 0.98 ($\Delta G_{t/r} = 7$ kcal/mol) to 0.95 ($\Delta G_{t/r} = 11$ kcal/mol), compared to the 0.96 we found before. Thus, the use of Equation (8) rather than Equation (1) appear to be equally efficient for molecular dynamics modeling of the free energies of ligand-protein complexes.

All of the above simulations started from crystal structures. In real drug design, people are interested in docking novel ligands into binding sites. In another word, this method would be more useful if it could be used in estimating binding free energies for structures obtained from docking. In addition, the linear correlation shown in Figure 2 was obtained from examining different protein-ligand systems. It is of interest to examine this relationship for different ligands binding to the same protein. We thus have applied this correlation to predict optimal values of β for twelve neutral compounds bound to avidin (compounds 2-13 in Figure 3).[8] No crystal structure of any of these complexes was available. The compounds (see Figure 3) were docked into the binding site of avidin using a docking program developed by Wang *et al.*[21] Different values of β were assigned based on WDNR of each complex structure after the MD simulation. As a comparison, a fixed β value, β = 0.87, was applied to calculate the binding free energies too (see Table III). The reason to use β = 0.87 is that it gave best estimate of binding free energy of biotin (compound 1) binding to avidin.

From Figure 4, we can see, compared with a fixed β value, nine points out of twelve are in equal or better agreement with experiment using the WDNR to determine a different β for each complex. For a fixed β value, the binding free energies were almost always underestimated. However, using the WDNR values, this systematic error was reduced in most cases. In three complexes, the binding free energies were overestimated. This implies there were random errors involved instead of systematic ones. The error may come from the correlation itself since we only had seven points in our fit set. Another possible source may be errors of the structures which were obtained from docking ligands into the binding site. The worst case is ligand four. The reason for this was discussed elsewhere.[8]

We also examined the relative binding free energies between biotin (compound 1) and other compounds binding to the avidin (Table IV and Figure 5). From Figure 3, we can see that compounds 2, 3, 6, 7 are more similar to biotin than compounds 5, 8-13. The desolvation effect of compounds 2, 3, 6, 7 should be also similar to biotin. In Table IV, a fixed β gave better results for compounds 2, 3, 6, 7 than WNDR. The reason is that errors introduced by second term in Equation (1) using a fixed β cancelled each other when we calculate relative binding free energies between similar compounds. However, for compounds 5, 8-13, WNDR gave better results than fixed β since the desolvation effect of compounds 5, 8-13 are different from biotin.

After observing the correlation between the WNDR and the value of β, we want to understand the physical meaning of this correlation. As we noted above in Equation (7), the value of β has in it implicit contributions from a number of terms.

Jorgensen and coworkers have included a specific ΔG_{desolv} term, a solvent accessible surface term, to Equation (1).[23,24] They calculated the change of solvent accessible surface due to complexation and introduced another parameter γ in addition to α and β. By optimizing the three parameters, they obtained some improvements compared with results obtained using Equation (1). However they did not discriminate the different contributions of polar and non-polar groups to the binding free energies.[23,24] It is known that burial of non-polar groups is favorable for binding, while burial of polar groups contributes unfavorably to binding free energies.

From our own experience and previous studies,[1,2,3,8,23,24] van der Waals interactions are always favorable for binding. The contribution to binding free energy of desolvation effect depends on the components of the binding site. The higher percentage of non-polar groups in the binding site, the higher percentage non-polar groups are desolvated due to binding, apparently, the more favorable contribution to the binding free energy due to the desolvation effect. Since the weights for polar groups are negative, the more polar groups buried, the smaller the denominator of Equation (6), thus, the larger the WNDR. Conversely, the smaller the WNDR, the more hydrophobic a binding site, the larger the value of β. Therefore, the correlation between WNDR and β makes sense.

IV. Conclusion

In this work, we present a correlation between the weighted desolvation non-polar ratio (WNDR) and values of β in the linear interaction energy method. β in the LIE method is a factor to describe how much van der Waals interactions and desolvation effect contribute to binding free energies. The WNDR represents the hydrophobicity of the binding sites. This correlation was found by studying different systems. It suggests that the value of β is predictable by calculating the WNDR of the system, especially for systems in which very different ligands bind to the same protein or ligands bind to different binding sites of the same protein. We applied this correlation to twelve complexes of avidin whose complex structures were obtained using a docking algorithm. The results are promising, but the further investigation is needed to examine if the correlation is found in other systems.

The correlation between the WNDR and the value of β should be useful in drug design. It is easy to calculate the WNDR for any ligand-protein complex. The value of β can be obtained from the correlation presented here and this should allow a more accurate estimate of binding free energy to be obtained from the linear interaction energy (LIE) method. The further role of LIE method in drug design is to refine the leads found by docking algorithms using database screening or *de novo* design. With the development of increased computer power, molecular dynamics can become more useful. Thus the LIE method can be expected to be more widely used in finding novel leads for ligands that bind tightly to macromolecules.

V. Acknowledgement

This research was supported in part by grants GM-29072 to P.A.K and GM-56609 (E. Arnold, P.I.). Part of this investigation was completed using the facilities in the UCSF Computer Graphics Lab, T. Ferrin, director, supported by RR-1081 from the NIH. W.W is grateful to Drs. Lu Wang and Yong Duan and Jed Pitera for their stimulating discussions, and a Regents fellowship of UCSF for support. J.W. thanks the Natural Science and Engineering council of Canada for a postdoctoral fellowship to carry out part of this research.

V. References

(1) Åqvist, J., Medina, C., Samuelsson, J. New method for predicting binding affinity in computer-aided drug design. Protein Eng. 7:385-391, 1994.

(2) Hansson, T., Åqvist, J. Estimation of binding free energies for HIV proteinase inhibitors by molecular dynamics simulations. Protein Eng. 8:1137-1144, 1995.

(3) Åqvist, J., Calculation of absolute binding free energies for charged ligands and effects of long-range electrostatic interactions. J. Comput. Chem. 17:1587-1597, 1996.

(4) Kollman, P. A. Free energy calculations: Applications to chemical and biochemical phenomena. Chem. Rev. 93: 2395-2417, 1993.

(5) Beveridge, D. L., Dicapua, F. M. Free energy via molecular simulations: application to chemical and biochemical system. Annu. Rev. Biophys. Biophys. Chem. 18: 431-492, 1989.

(6) van Gunsteren, W. F. Methods for calculation of free energies and binding constants: successes and problems. In: "Computer Simulation of Biomolecular Systems" van Gunsteren, W. F., Weiner, P. K.(eds.). ESCOM:Leiden. 1989:27-59.

(7) Paulsen, M. K., Ornstein, R Binding free energy calculations for P450cam-substrate complexes. Protein Eng. 9:567-571, 1996.

(8) Wang, J., Dixon, R., Kollman, P. A. Ranking ligand binding affinities with avidin: a molecular dynamics based interaction energy study. Proteins, accepted for publication, 1998.

(9) Cornell, W. D, Cieplak, P., Bayly, C. I., Gould, I., Merz, K. M., Ferguson, D., Spellmeyer, D. C., Fox, T., Caldwell, J. W., Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J. Am. Chem. Soc. 117:5179-5197, 1995.

(10) Perlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, T. E., Debolt, S., Ferguson, D., Seibel, G., Kollman, P. A. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. Comp. Phys. Comm. 91:1-41, 1995.

(11) Jorgensen, W. L., Chandrasekhar, J., Madura, J., Impey, R. W., Klein, M. L. Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. 79:926-935, 1983.

(12) Bayly, C. I., Cieplak, P., Cornell, W. D., Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges - the RESP model. J. Phys. Chem. 97:10269-10280, 1993.

(13) Rychaert, J. P., Ciccotti, G., Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J. Comput. Phys. 23:327-341, 1977.

(14) Sanner, M. F., Olson, A. J., Spehner, J. Reduced surface - an efficient way to compute molecular surfaces. Biopolymers 38:305-320, 1996

(15) Miyamoto, S., Kollman, P. A Absolute and relative binding free energy calculations of the interaction of biotin and its analogs with strepavidin using molecular dynamics free energy perturbation approaches. Proteins 16:226-245, 1993.

(16) Miyamoto, S., Kollman, P. A. What determines the strength of noncovalent association of ligands to proteins in aqueous solution. Proc. Natl. Acad. Sci. USA 90:8402-8406, 1993.

(17) Eisenberg, D., McLachlan, A. D. Solvation energy in protein folding and binding. Nature 319:199-203, 1986.

(18) Williams, R. L., Vila, J., Perrot, G., Scheraga, H. A. Empirical solvation models in the context of conformational energy searches - application to bovine pancreatic trypsin inhibitor. Proteins 14:110-119, 1992.

(19) Still, W. C, Tempcyk, A., Hawley, R. C., Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J. Am. Chem. Soc. 112:6127-6129, 1990.

(20) Cramer, C. J., Truhlar, D. G. General parameterized SCF model for free energies of solvation in aqueous solution. J. Am. Chem. Soc. 113:8305-8311, 1991.

(21) Wang, J., Kollman, P. A., Kuntz, I. D. Flexible ligand docking: A systematic approach. Proteins accepted for publication, 1998.

(22) Ajay, Murcko, M. A. Computational methods to predict binding free energy in ligand-receptor complexes. J. Med. Chem. 38:4953-4967, 1995.

(23) Jones-Hertzog, D. K., Jorgensen, W. L. Binding affinities for sulfonamide inhibitors with human thrombin using Monte Carlo simulations with a linear response method. J. Med. Chem. 40:1539-1549, 1997.

(24) Carlson, H. A., Jorgenson, W. L. An extended linear response method for determining free energies of hydration. J. Phys. Chem. 99:10667-10673, 1995.

(25) Marelius, J., Hansson, T., Åqvist, J. Calculation of ligand binding free energies from molecular dynamics simulations. Int. J. Quant. Chem. 69:77-88, 1998.

(26) Williams, D. H., Cox, J. P. L., Doig, A. J., Garner, M., Gerhard, U., Kaye, P. T., Lal, A. R., Nicholls, I. A., Salter, C. J., Mitchell, R. C. Toward the semiquantitative estimation of binding constants - guides for peptide peptide binding in aqueous solution. J. Am. Chem. Soc. 113:7020-7030, 1991.

(27) Searle, M. S., Williams, D. H., Gerhard, U. Partitioning of free energy contributions in the estimation of binding constants - residual motions and consequences for amide-amide hydrogen bond strengths. J. Am. Chem. Soc. 114:10697-10704, 1992.

Figure legends

Figure 1. β value versus non-polar desolvation ratio (NDR) for the seven calibration systems.

Figure 2. β value versus weighted non-polar desolvation ratio (WNDR) for the seven calibration systems.

Figure 3. Biotin analogues used for comparing the WNDR approach and the fixed β approach.

Figure 4. Observed versus calculated binding free energies for the 12 compounds binding to avidin using β predicted from the correlation obtained from Figure 2 and β=0.87 respectively.

Figure 5. Observed versus calculated binding free energies between biotin (compound 1) and other compounds binding to avidin using β predicted from the correlation obtained from Figure 2 and β=0.87 respectively.

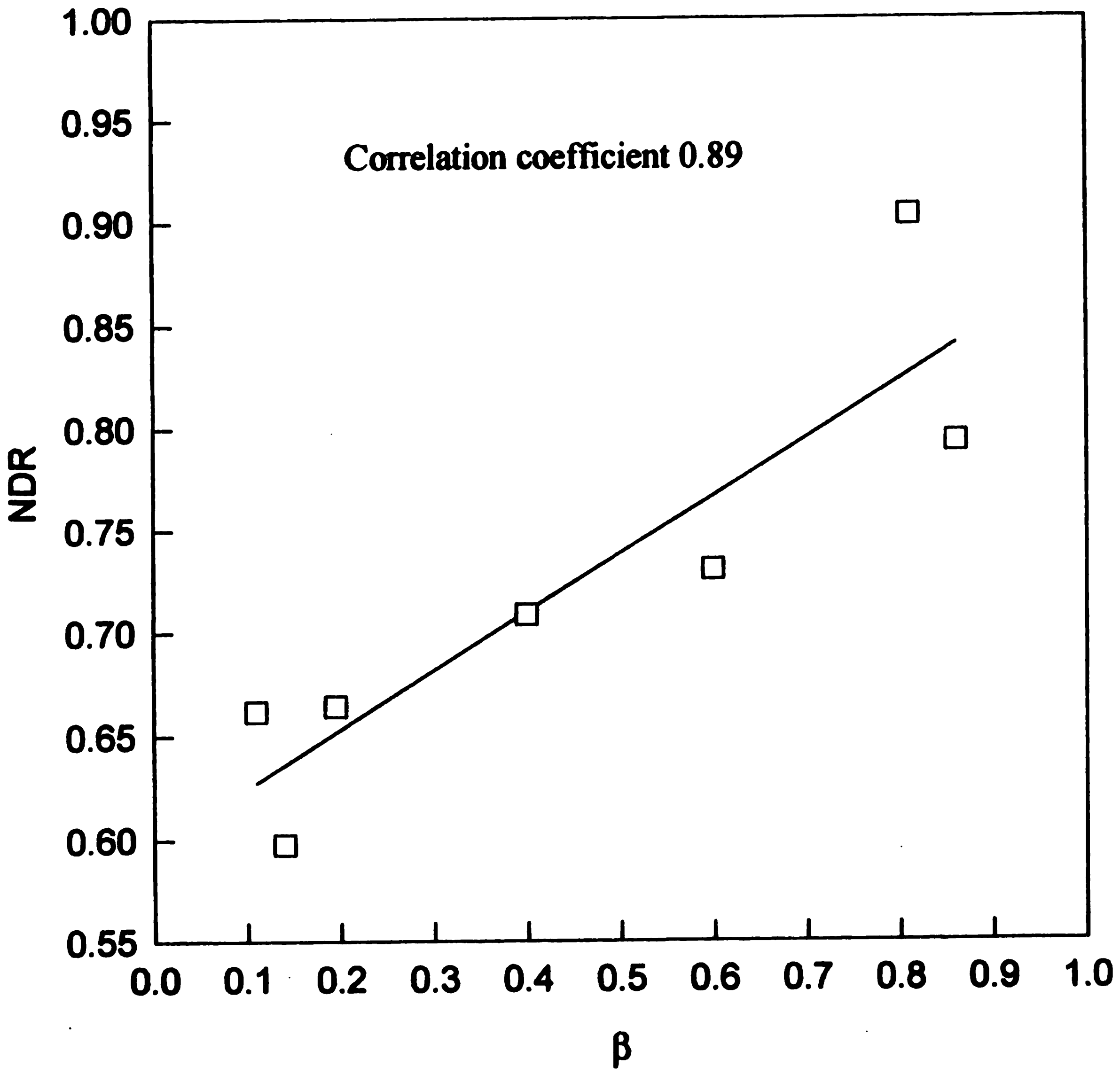


Fig. 1 β value versus non-polar desolvation ratio (NDR) for the seven calibration systems

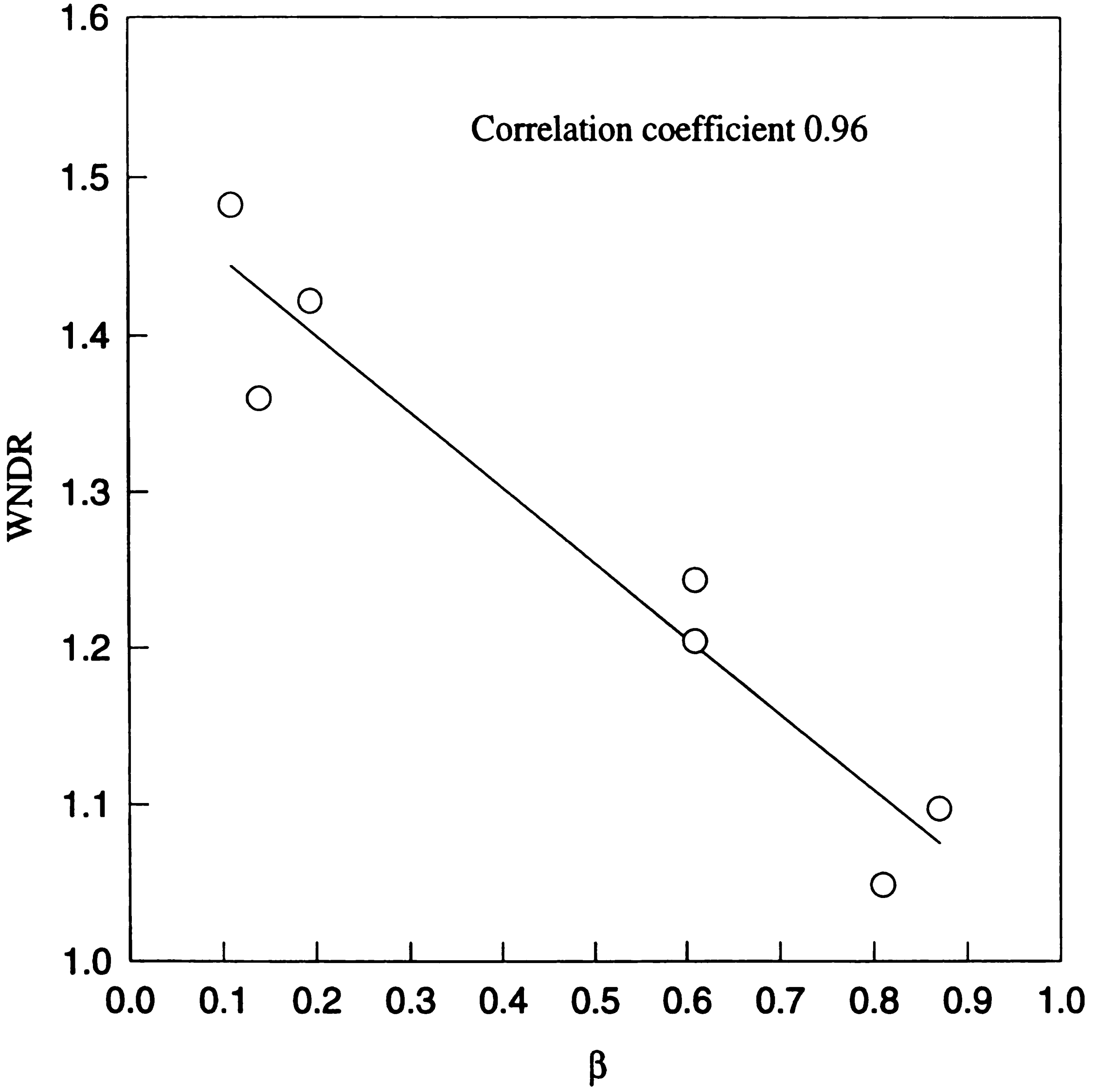
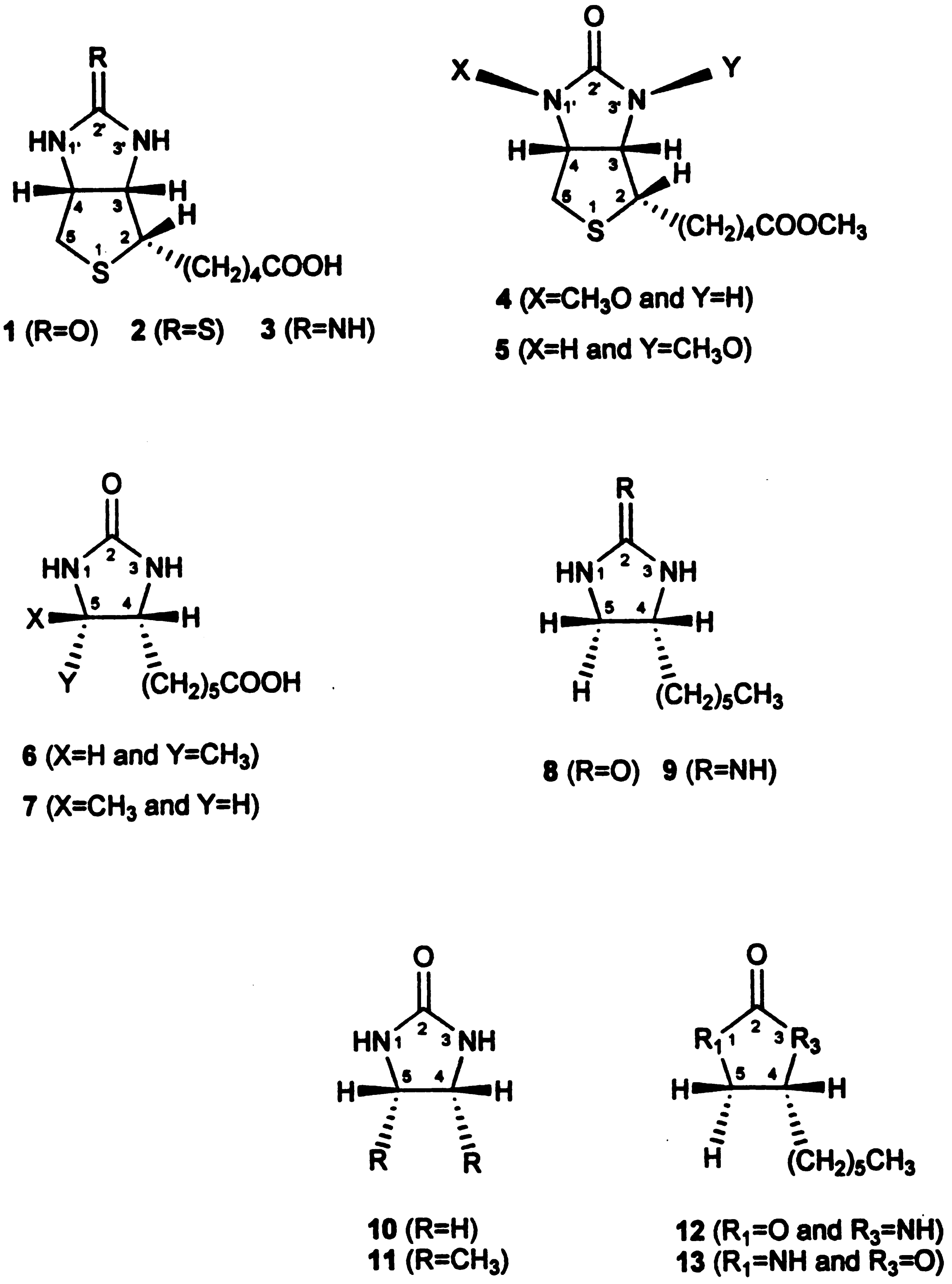


Fig.2 β value versus weighted non-polar desolvation ratio (WNDR) for the seven calibration systems



1 (R=O) 2 (R=S) 3 (R=NH)

4 (X=CH_3O and Y=H)

5 (X=H and Y=CH_3O)

6 (X=H and Y=CH_3)

7 (X=CH_3 and Y=H)

8 (R=O) 9 (R=NH)

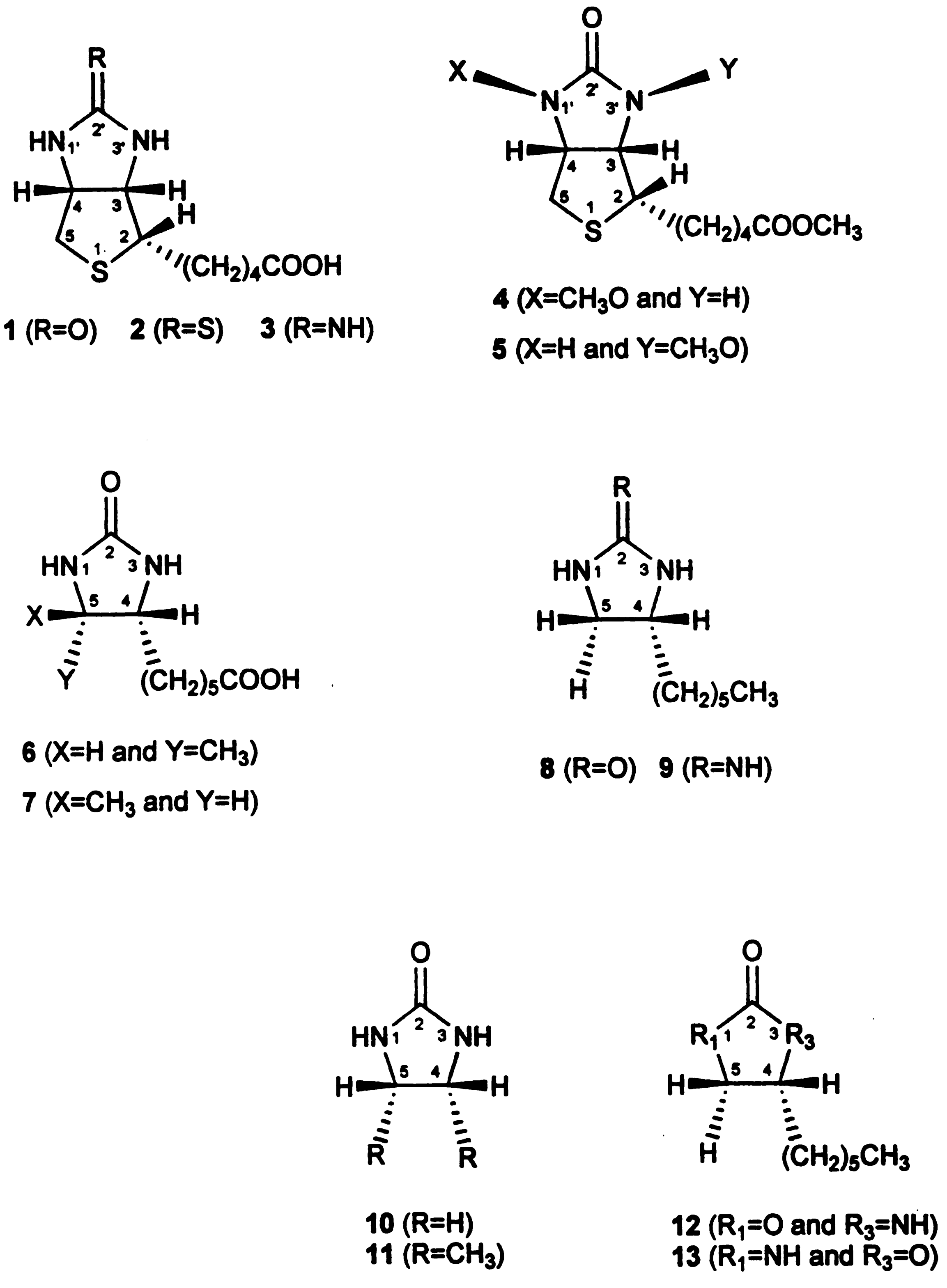
10 (R=H)

11 (R=CH_3)

12 (R_1=O and R_3=NH)

13 (R_1=NH and R_3=O)

Figure 3. Biotin analogues used for comparing the WNDR approach and the fixed β approach.



1 (R=O) 2 (R=S) 3 (R=NH)

4 (X=CH_3O and Y=H)

5 (X=H and Y=CH_3O)

6 (X=H and Y=CH_3)

7 (X=CH_3 and Y=H)

8 (R=O) 9 (R=NH)

10 (R=H)
11 (R=CH_3)

12 (R_1=O and R_3=NH)
13 (R_1=NH and R_3=O)

Figure 3. Biotin analogues used for comparing the WNDR approach and the fixed β approach.

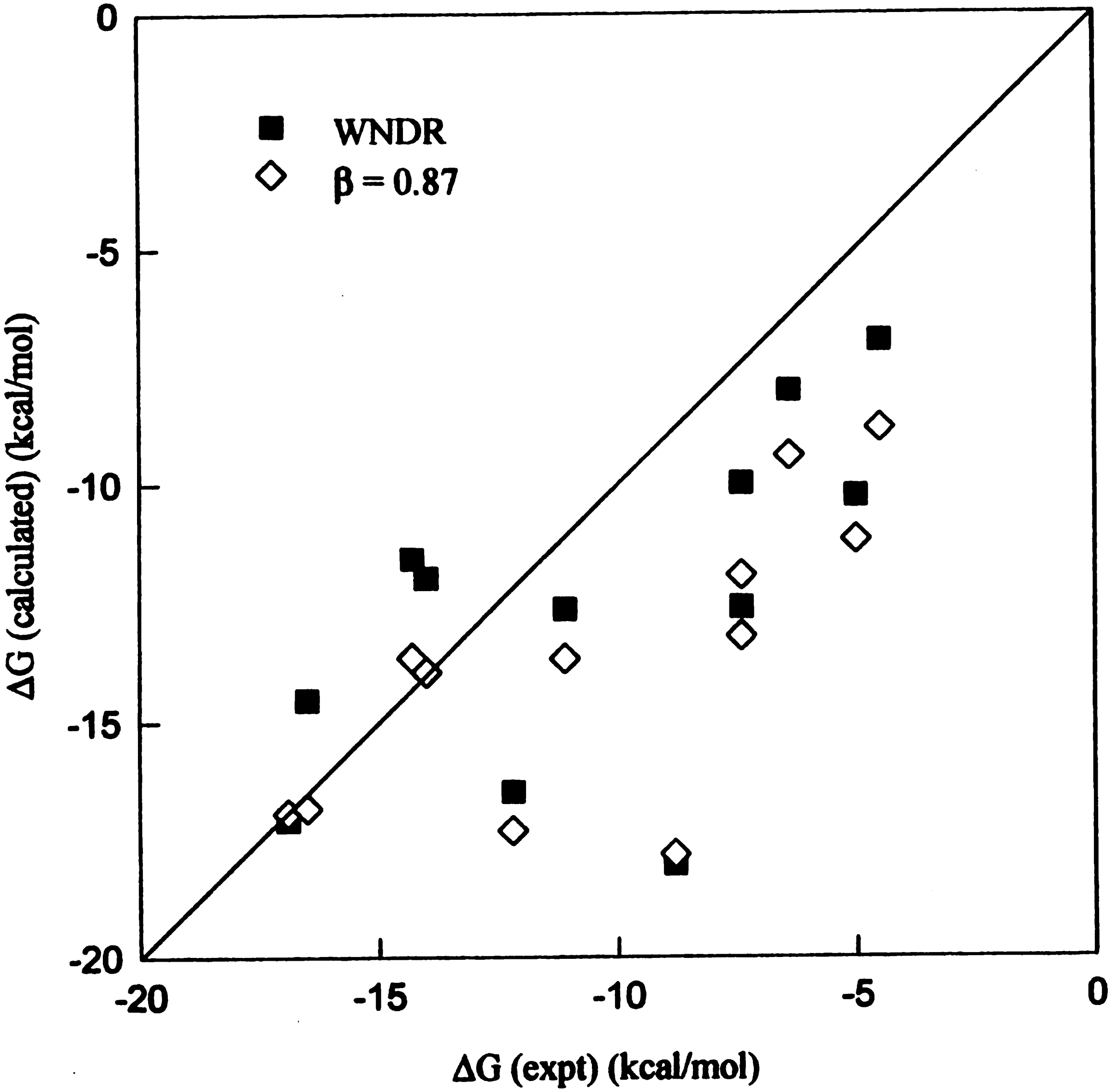


Fig. 4 Observed versus calculated binding free energies for the 12 compounds binding to avidin using β predicted from the correlation obtained from Figure 2 and β=0.87 respectively.

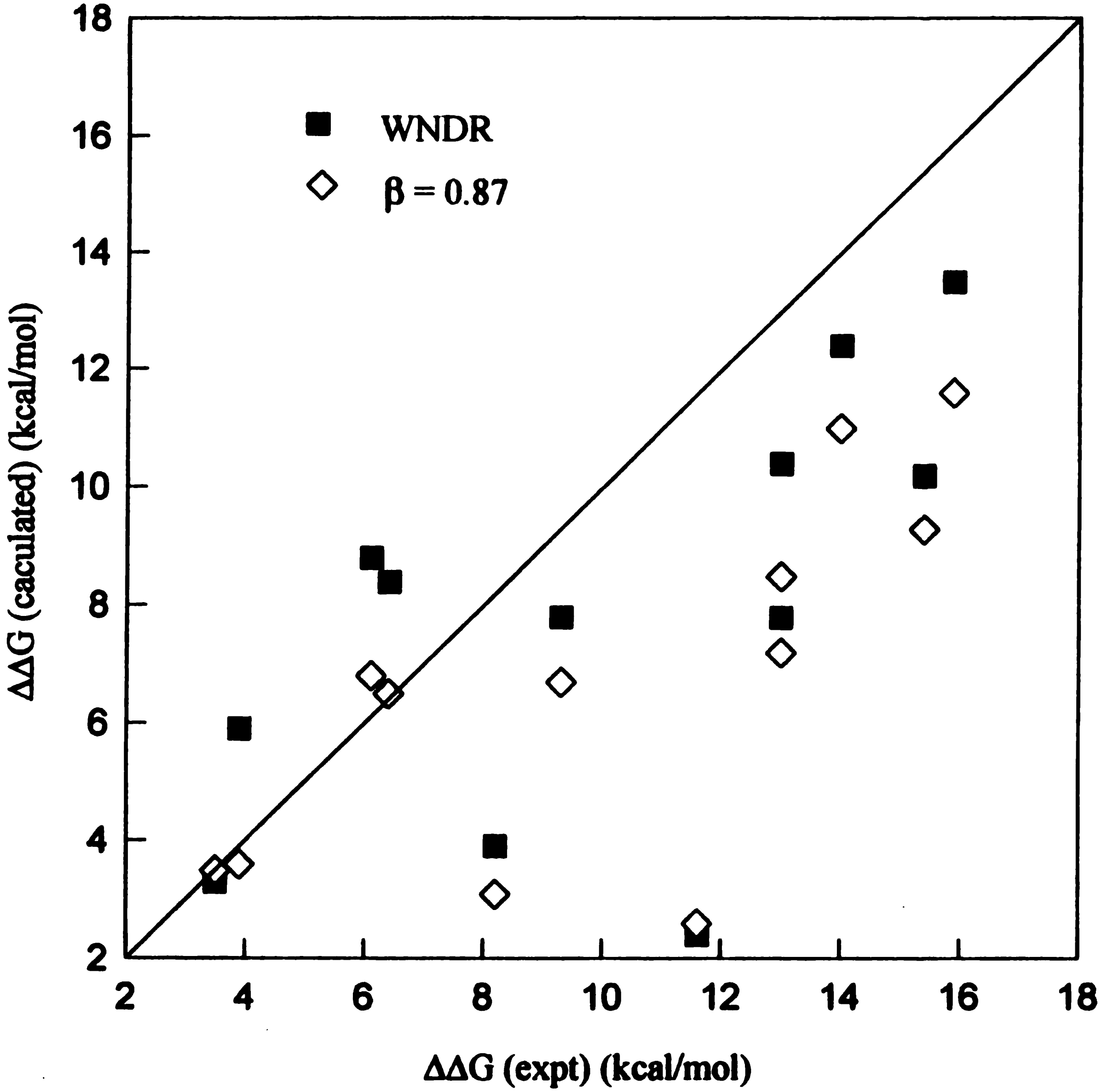


Fig. 5 Observed versus calculated relative binding free energies between biotin (compound 1) and other compounds binding to avidin using β predicted from the correlation obtained from Figure 2 and β = 0.87 respectively.

Table I Surface tension parameters taken from Eisenberg and McLachlan's work.[17]

Atom type	σ (cal/mol/Å^2)
C	16
S	21
N	-6
N^+	-50
O	-6
O^-	-24

Table II Calculated binding free energies, WNDR and β values of seven systems

Protein	PDB entry	VDW[d] (kcal/mol)	EL[e] (kcal/mol)	WNDR[f]	β	ΔG_{cald} (Kcal/mol)	ΔG_{expt} (kcal/mol)
Trypsin	3ptb						
In water[a]		-8.20±0.18	-104.44±0.92	1.361	0.14	-6.55	-6.54
In protein[b]		-20.35±0.69	-114.14±1.92				
Difference[c]		-12.15	-9.70				
HIVP	1aaq						
In water		-39.76±0.5	-111.48±1.50	1.42	0.195	-7.61	-7.60
In protein		-64.31±0.25	-117.12±1.02				
Difference		-24.55	-5.64				
HIVP	5hvp						
In water		-35.44±0.82	-252.18±2.99	1.48	0.11	-10.30	-10.30
In protein		-65.15±0.63	-266.24±0.80				
Difference		-29.71	-14.06				
Thrombin	1dwc						
In water		-23.82±0.08	-229.80±1.03	1.24	0.61	-10.76	-10.67
In protein		-45.21±0.31	-225.24±1.04				
Difference		-21.39	+4.56				
Thrombin	1dwd						
In water		-34.76±0.28	-209.07±1.98	1.20	0.61	-11.59	-11.63
In protein		-60.99±0.18	-200.25±0.70				
Difference		-26.228	+8.82				
Avidin	1avd						
In water		-19.53±0.06	-52.68±0.16	1.10	0.87	-20.37	-20.40
In protein		-37.57±0.30	-62.40±0.16				

Difference		-18.04	-9.36				
P450-cam	2cpp						
In water		-14.97±0.22	-12.87±0.24	1.05	0.81	-7.85	-7.90
In protein		-26.69±0.14	-9.58±0.06				
Difference		-11.72	+3.29				

a. Ligand in the free state, i.e. in water; b. Ligand in the bound state, i.e. ligand bound to solvated protein; c. The difference of van der Waals and electrostatic interaction energies between the ligand free and bound states; d, e. van der Waals and electrostatic interaction energies between the ligand and its environments, which are water in the free state and solvated protein in the bound state respectively; f. Weighted non-polar desolvation ratio (WNDR), see text for definition.

Table III Calculated binding free energies for 12 compounds binding to avidin with WNDR and β=0.87 approaches

#	ΔG(kcal/mol) (expt)	WNDR	β	ΔG(kcal/mol) (cald)[a]	ΔG(kcal/mol) (cald)[b]
2	-16.9	1.070	0.88	-17.09	-16.93
3	-14.3	1.130	0.76	-11.58	-13.65
4	-8.8	1.070	0.88	-18.03	-17.81
5	-12.2	1.100	0.82	-16.47	-17.29
6	-14.0	1.150	0.72	-11.98	-13.94
7	-16.5	1.145	0.73	-14.53	-16.82
8	-11.1	1.117	0.78	-12.64	-13.66
9	-7.4	1.150	0.72	-9.97	-11.91
10	-4.5	1.240	0.53	-6.93	-8.78
11	-6.4	1.170	0.67	-8.00	-9.38
12	-5.0	1.105	0.81	-10.24	-11.15
13	-7.4	1.096	0.83	-12.60	-13.19

a. Results obtained by using β which was estimated from the correlation between WNDR and β; b. Results calculated using β=0.87.

Table IV Relative binding free energies between biotin (compound 1) and other compounds binding to avidin

Relative	ΔΔG(kcal/mol) (expt)	ΔΔG(kcal/mol)[a]	ΔΔG(kcal/mol) [b]
2-1	3.5	3.3	3.5
3-1	6.1	8.8	6.8
4-1	11.6	2.4	2.6
5-1	8.2	3.9	3.1
6-1	6.4	8.4	6.5
7-1	3.9	5.9	3.6
8-1	9.3	7.8	6.7
9-1	13.0	10.4	8.5
10-1	15.9	13.5	11.6
11-1	14.0	12.4	11.0
12-1	15.4	10.2	9.3
13-1	13.0	7.8	7.2

a. Results obtained by using β which was estimated from the correlation between WNDR and β; b. Results calculated using β=0.87.

DNA sequencing technology has become mature. The genomes of human and other species have been or will be sequenced soon. A focus of research in this area is on identifying genes and their functions. DNA microarray allows genome-wide analysis of gene expression. Advancement of proteomics and structure genomics have also provided an overwhelming amount of data. All the data obtained from technologies mentioned above have to be analyzed statistically. For example, genes and proteins are assumed to have similar functions if they (genes) have similar DNA microarray expression pattern or they (proteins) have similar 3-D structures. Predictions of novel gene or protein functions are based on their similarities with genes or proteins whose functions are known. Statistical analysis of experimental observations can provide a significant amount of useful information and insights, but such an approach lacks a physical basis and sometimes the conclusions are even misleading due to the still limited and noisy dataset.

On the other hand, physicists and chemists have studied protein-ligand interaction and protein folding using methods based on physical principles. Among these methods, molecular dynamics and free energy calculations have become much more efficient and accurate than a few years ago. These methods can provide atomic, dynamic and physical insights of a few systems. In this dissertation, I have shown the success of molecular dynamics, the MM/PBSA method, and the Linear Interaction Energy method in studying protein-protein and protein-ligand interactions. The shortcoming of these physical methods is their efficiency. It still takes considerable computer power and human efforts to apply them. There is a long way to go in order to apply these methods genome-wide. However, insights and data obtained from these physical studies are very valuable. For example, we can perform molecular dynamics (MD) and free energy calculations on the

Sem-5 SH3 domain. It is wasteful if the data generated from the extensive MD and free energy calculations is only used to understand the Sem-5 SH3 domain. These insights should be extendable to other SH3 domains and we have suggested a way to do this in this dissertation.

This dissertation has proposed a heuristic way to combine physical principles (free energy calculations) and statistical principles (sequence analysis) to study protein recognition. This method succeeded in some systems (Sem-5 SH3 domain and HIV drug resistance) and is under investigation for other systems. There may be some more sophisticated and more elegant approaches for this combination, which could be even more powerful in studying biological phenomena.

Another problem for biologists nowadays is how to integrate knowledge at the gene level (e.g. DNA sequence, DNA microarray expression pattern) and the protein level (e.g. protein structure, protein interaction pairs) to understand the function of the cell. For example, when we study signal transduction pathway in yeast, there are usually some proteins in the pathway that are well studied (e.g. their structures and interaction partners are known). DNA microarray experiments can be done to cluster genes; comparative genomics, such as comparing the yeast genome with the worm genome, can be used to predict functions of some proteins; detailed computational studies can be performed on proteins whose structures are known to predict their interaction partners. How to integrate all this information to get a complete picture of the signal pathway is not clear yet.

The next question is to address the dynamics of the cell using physical principles: after a novel protein is synthesized, how does it know where it should go, with what targets can it interact, is this a random process which depends on diffusion or is it determined by properties, such as charge distribution, shape, hydrophobicity, of the protein; and how does a cell adjust gene expression levels in response to external stimuli. If we are able to explain these phenomena using physics, chemistry and mathematics, it will be possible someday to simulate how the cell lives using a computer. This will be very exciting in terms of basic science and human health.

Going one step further, one might be able to develop a theory to describe how cells live. Ideally, each cell can be described by a wave function like basic particles in quantum mechanics. Time, environment and external stimuli can be described as different operators. Each operator can induce changes in the wave function of the cell. Therefore, the probability of occurring of each biological process can be calculated by the wave function. This probability allows random process which ensures evolution. There is a long way to go from current research to build such a grand theory but this goal makes research even more challenging and exciting!