MODELING RECEPTOR REORGANIZATION AND STRAIN IN PROTEIN-LIGAND BINDING

by KRISTINA A. PARIS

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Chemistry

written under the direction of

Professor Ronald M. Levy

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

January 2011

ABSTRACT OF THE DISSERTATION

Modeling Receptor Reorganization and Strain in Protein-Ligand Binding

By KRISTINA A. PARIS

Dissertation Director:
Professor Ronald M. Levy

The key objectives of computational structure-based drug design include the prediction of the protein-ligand complex binding modes and estimation of the binding affinities. The overall affinity of a ligand for a receptor can be expressed as a balance between the strength of the interactions of a ligand to any particular binding-competent conformation of the receptor and the probability of occurrence of that conformation in the absence of the ligand. The receptor conformation probability distributions can be described by the free energy landscape of the receptor from which the strain free energy required to move from one conformation to another in the absence of a ligand may be estimated. The availability of large datasets of crystal structures in the PDB can provide information about the locations of free energy basins and their shapes. Here we utilize several methods in an effort to model the strain free energy of several receptors due to binding using the vast structural data publically available in the PDB. Clustering of 99 X-ray structures of HIV-1 reverse transcriptase at the flexible non-nucleoside inhibitor binding pocket elucidates eight discrete clusters, one of which displays a novel bound

conformation of the functionally important primer grip. The clustering results served as a guide for replica exchange molecular dynamics simulations that offer a more in-depth look at the potential reorganization of the binding pocket. Clustering of 327 available X-ray structures of HIV-1 protease reveals less discrete variability in the substrate envelope than HIV-1 reverse transcriptase but does reveal some receptor reorganization that may be due to a combination of mutations.

A linear response model for incorporation of receptor strain in modern protein-ligand binding affinity estimators is proposed. Receptor-receptor contact counts are employed as estimators for changes in receptor conformation due to binding of different ligands. Overall, the linear model produces apparent reduction in binding energy estimation errors and increases in the rank-order correlation with respect to initial values determined by the commercially available Glide 5.0 XP that does not take into account receptor reorganization. It also offers information as to the type of conformational changes, if any, that may contribute to the receptor reorganization energy. A null hypothesis test is constructed to evaluate the possibility of producing fits by chance alone. Finally, an alternative estimator approach using structurally significant intrareceptor distance descriptors, where there are less possible estimators, shows some promise for several drug targets. The model has the potential to allow for coarse-grained investigation of the conformational and energetic landscapes for binding inhibitors to flexible protein receptors.

Acknowledgements

First and foremost, I would like to thank my advisor, Professor Ronald M. Levy, who has provided invaluable guidance, patience and support throughout my years as a graduate student. His visions and passion for science will continue to affect me in the many years to come. No less I would like to extend my gratitude to Dr. Anthony Felts who I am lucky to have as both a mentor and friend. His counsel and moral support were much needed and will forever be appreciated. I would also like to thank Dr. Emilio Gallicchio for his insightful ideas and direction in the projects on which I worked and Drs. Eddy Arnold and Kalyan Das, without whom I would have been lost in the sea of research surrounding HIV RT. Much gratitude also goes to the Levy group members of past and present whose ideas and work laid the foundation for the work I have done, but especially to Omar Haq and Lauren Wickstrom for their friendship and encouragement as well as helpful conversations.

Thank you to the many friends I have discovered over the years at Rutgers, especially Mauricio Esguerra who served as my partner in crime as we attempted to balance art and science. I would also like to thank members of the Chemistry department who have assisted me in my doctoral quest.

Thanks also to my parents Joel and Laurene Paris, brother Mark, grandmother Doris (Nannie) and other family and friends for their love, patience and support. Finally, thank you to my best friend and partner Benjamin Delloiacono for his love, support and encouragement through the difficult times of graduate school and for making these past few years some of the most enjoyable.

Table of Contents

Lists of tables

List of Figures

Chapter 1

Introduction

Computational structure-based drug design strives to correctly predict protein-ligand complex binding modes and estimate binding affinities in an effort to capture the physical properties that are responsible for recognition of the drug by its target protein. The amount of available structural data is ever-growing as experimental techniques improve. This large amount of structural data along with a large amount of available inhibition data allows computer-aided structure-based ligand design to serve as an alternative strategy to experimental high-throughput screening to find novel leads in drug development. Previously, ligand binding was often approached via either Fischer's "lock-and-key" model (Fischer, 1894) or Koshland's "induced fit" hypothesis (Koshland, 1958). In the "lock-and-key" model, the free and ligand-bound proteins have the same rigid conformation whereas in the "induced fit" model, the ligand induces a complementary conformational change in the protein. The conformational selection hypothesis approaches binding from a "folding funnel" point of view where protein folding is viewed as a parallel process where an ensemble of molecules goes downhill through an energy funnel (Dill and Chan, 1997; Lazaridis and Karplus, 1997; Becker and Karplus, 1997; Martinez et al., 1998; Onuchic et al., 1997; Ravindranathan et al., 2005). Folding funnels are rugged in the vicinity of the native fold of the protein, suggesting energetically competitive and similar conformations that provide an enhanced means of interactions between the protein and either ligands or other proteins. The model of

conformational selection takes into account this rugged terrain and argues that ligand binding can shift the populations towards the weakly populated, higher energy conformations that are more suitable for binding (Ma et al., 1999). Both conformational selection and induced fit appear to play roles in ligand binding (Boehr et al., 2009; Bakan and Bahar, 2009).

Accurate modeling of hydration and the ability of a scoring function to favorably score the native structure are essential in successful computational modeling problems, including the study of protein folding, conformational equilibria, and binding. The first part of this study analyzes the ability of the OPLS-AA (Jorgenson et al., 1996; Kaminski et al., 2001) scoring function in combination with the AGBNP solvent model (Gallicchio and Levy, 2004) to predict protein loop and side chain conformations.

The second part of this study utilizes several methods in an effort to model the strain free energy of several receptors due to binding using the vast structural data publically available in the Protein Data Bank (PDB; Berman et al., 2000). The first few chapters of this section focus on utilizing available X-ray structures to create a conformational landscape for binding for two human immunodeficiency virus (HIV) enzymes: reverse transcriptase (RT) and protease (PR). Inhibition of these two enzymes allows a retardation of the progression of HIV to full-blown acquired immune deficiency syndrome (AIDS). RT primarily functions as the virus' "copy machine" as it copies the viral RNA into DNA that is incorporated in the host cell's genome. PR serves in the important role of prepping viral enzymes for production of a new virion. Both enzymes have very flexible inhibitor binding pockets. As they have been the focus of many studies, there are many X-ray structures available. These large ensembles of X-ray

structures (99 for RT and 318 for PR) are clustered using hierarchical clustering schemes to pinpoint areas of conformational fluctuation due to binding. Results offer a better understanding of the potential conformational landscape for binding and elucidate novel configurations of the binding pockets that have not yet been described or fully explored. Clustering also offers a benchmark for a more extensive computational exploration of the binding landscapes.

A linear model for incorporation of receptor strain in the modern protein-ligand binding affinity estimator Glide (Friesner et al., 2006; Friesner et al., 2004; Halgren et al., 2004) is proposed in the final part of the section on receptor reorganization. Receptor-receptor contact counts and "hand-picked" descriptors are employed as estimators for changes in receptor conformation due to binding of different ligands. Comparison with the use of random data leads to the question of the statistical significance of such a model based on culling from large data sets. The protocol set forth serves as an example for future projects for incorporation of receptor strain in ligand binding problems and also offers a discussion of potential repercussions of the use of large data sets.

References

Bakan, A.; Bahar, I. The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proc. Natl. Acad. Sci.* **2009**, *106*, 14349-14354.

Becker, O.M.; Karplus, M. The topology of multidimensional potential energy surfaces: theory and application to peptide structure and kinetics. *J. Chem. Phys.* **1997**, *106*, 1495-1517.

Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235-242.

Boehr, D.D.; Nussinov, R.; Wright, P.E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol*. **2009**, *5*, 789-96.

Dill, K.A.; Chan, H.S. From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **1997**, *4*, 10-19.

Fischer, E. Einfluss der configuration auf die wirkung der enzyme. *Ber. Dtsch. Chem.* **1894**, *27*, 2984-2993.

Friesner, R.A.; Banks, J.L.; Murphy, R.B.; Halgren, T.A.; Klicic, J.J.; Mainz, D.T.; Repasky, M.P.; Knoll, E.H.; Shelley, M.; Perry, J.K.; Shaw, D.E.; Francis, P.; Shenkin, P.S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem*. **2004**, *47*, 1739–1749.

Friesner, R.A.; Murphy, R.B.; Repasky, M.P.; Frye, L.L.; Greenwood, J.R.; Halgren, T.A.; Sanschagrin, P.C.; Mainz, D.T. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem*. **2006**, *49*, 6177-96.

Gallicchio, E.; Levy, R.M. AGBNP: an analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *J. Comp. Chem.* **2004**, *25*, 479-499.

Halgren, T.A.; Murphy, R.B.; Friesner, R.A.; Beard, H.S.; Frye, L.L.; Pollard, W.T.; Banks, J.L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem*. **2004**, *47*, 1750–1759.

Jorgensen, W.L.; Maxwell, D.S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225-11236.

Kaminski, G.A.; Friesner, R.A.; Tirado-Rives, J.; Jorgensen, W.L. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **2001**, *105*, 6474-6487.

Koshland, D.E. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. USA* **1958**, *44*, 98-104.

Lazaridis, T.; Karplus, M. "New view" of protein folding reconciled with the old through multiple unfolding simulations. *Science* **1997**, *278*, 1928-1931.

Ma, B.; Kumar, S.; Tsai, C.J.; Nussinov, R. Folding funnels and binding mechanisms. *Protein Eng.* **1999**, *12*, 713-720.

Martinez, J.C.; Pisabarro, M.T.; Serrano, L. Obligatory steps in protein folding and the conformational diversity of the transition state. *Nat. Struct. Biol.* **1998**, *5*, 721-729.

Onuchic, J.N.; Luthey-Schulten, Z.; Wolynes, P.G. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545-600.

Ravindranathan, K.P.; Gallicchio, E.; Levy, R.M. Conformational equilibria and free energy profiles for the allosteric transition of the ribose-binding protein. *J Mol Biol* **2005**, *353*, 196-210.

PART I.

THE AGBNP IMPLICIT SOLVENT MODEL

AND STRUCTURE PREDICTION

Chapter 2

Protein Loop Modeling

2.1 Introduction

A scoring function that scores the native conformation more favorably than other possible conformations is a necessary component for any effective computational approach to protein modeling (Skolnick, 2006; Lazaridis and Karplus, 2000). Recent developments have focused on structure refinement in an effort to optimize models for drug discovery and structure prediction problems: prediction of protein loops (Fiser at al. 2000; Soto et al., 2008; Jacobson et. al., 2004), prediction of protein side chains (Krivov et al., 2009; Jacobson et al., 2002), and prediction of ligand-receptor "induced fit" effects (Sherman et al., 2006). Evaluation of scoring functions is often accomplished using a decoy set, where a known native structure is combined with a set of plausibly misfolded decoy structures and the scoring function is graded on its ability to recognize the native conformation (Rhee and Pande, 2003). For small structural variations, such as those in loop modeling, it is necessary to further challenge the scoring function by performing extensive local conformational searches, thus making the protein loop prediction problem a powerful benchmarking tool for testing accuracy of scoring functions.

In this study, we look at the prediction ability the OPLS-AA all-atom force field (Jorgenson et al., 1996; Kaminski et al., 2001) and a selection of implicit solvent models: distance-dependent dielectric, Surface Generalized Born plus Non-Polar (SGBNP) (Zhang et al., 2001; Ghosh et al., 1998), and three parameterizations of Analytical

Generalized Born plus Non-Polar (AGB-γ, AGBNP and AGBNP+) (Gallicchio and Levy, 2004) in combination with the Protein Local Optimization Program (PLOP) (Jacobson et al. 2004) which employs a torsional angle search protocol. We also evaluate a version of PLOP that has been optimized for loop prediction in the crystal environment.

2.2 Procedures and Results

The procedures and results of this part of the thesis are presented below as a reprint of a paper published in the *Journal of Chemical Theory and Computation* **2008**, *4*, 855-868.

J. Chem. Theory Comput. **2008,** *4,* 855–868 **855**

JCTC Journal of Chemical Theory and Computation

Prediction of Protein Loop Conformations Using the AGBNP Implicit Solvent Model and Torsion Angle Sampling

Anthony K. Felts,[†] Emilio Gallicchio,[†] Dmitriy Chekmarev,[†] Kristina A. Paris,[†] Richard A. Friesner,[‡] and Ronald M. Levy*[,†]

Department of Chemistry and Chemical Biology and BioMaPS Institute for Quantitative Biology, Rutgers University, Piscataway, New Jersey 08854, and Department of Chemistry, Columbia University, New York, New York 10027

Received February 19, 2008

Abstract: The OPLS-AA all-atom force field and the Analytical Generalized Born plus Non-Polar (AGBNP) implicit solvent model, in conjunction with torsion angle conformational search protocols based on the Protein Local Optimization Program (PLOP), are shown to be effective in predicting the native conformations of 57 9-residue and 35 13-residue loops of a diverse series of proteins with low sequence identity. The novel nonpolar solvation free energy estimator implemented in AGBNP augmented by correction terms aimed at reducing the occurrence of ion pairing are important to achieve the best prediction accuracy. Extended versions of the previously developed PLOP-based conformational search schemes based on calculations in the crystal environment are reported that are suitable for application to loop homology modeling without the crystal environment. Our results suggest that in general the loop backbone conformation is not strongly influenced by crystal packing. The application of the temperature Replica Exchange Molecular Dynamics (T-REMD) sampling method for a few examples where PLOP sampling is insufficient are also reported. The results reported indicate that the OPLS-AA/AGBNP effective potential is suitable for high-resolution modeling of proteins in the final stages of homology modeling and/or protein crystallographic refinement.

1. Introduction

A necessary component for an effective computational approach to the homology modeling problem[1] for protein structure prediction[2] and crystallographic and NMR structure refinement[3,4] is a scoring function that scores more favorably the native conformation over other possible conformations.[5,6] Scoring functions aimed at fold recognition and secondary structure assignment have been evaluated on the basis of their ability to recognize the known native protein conformation among a set of plausible misfolded decoy structures.[7–12] Both physics-based[13–19] and empirical knowledge-based scoring functions[20–23] have performed reasonably well in this kind of evaluation tests.

Recent development efforts have been focused on the refinement stages of the homology modeling problem, such as the conformational prediction of protein loops[24–26] and surface side chains[27] as well as the modeling of ligand/receptor induced fit effects,[28] which are essential steps to make the model useful as a drug discovery and optimization target. These kinds of high-resolution protein structure prediction applications have generally been performed using atomistic physics-based free energy estimators.

Protein decoy scoring exercises have been useful in determining the key global features of physics-based energy functions (such as the inclusion of solvation effects)[19] necessary for recognizing the broad characteristics of native protein structures. The decoy evaluation technique, however, is in general too blunt an instrument for discriminating the ability of energy functions to recognize small structural variations within the native ensemble. For thorough testing,

* Corresponding author e-mail: ronlevy@lutece.rutgers.edu.
[†] Rutgers University.
[‡] Columbia University.

10.1021/ct800051k CCC: $40.75
Published on Web 04/26/2008

it is necessary to challenge the energy function by performing extensive local conformational searches to actively look for minima of the energy functions and measure the degree of correspondence of these with the known native conformation.

Determining the correct conformation of a loop on a protein is one of the final steps in homology model building. After secondary structures have been assigned and placed, model construction often proceeds by conformational prediction of connecting loops. In loop prediction tests, we assume that the rest of the protein frame has been folded accurately and the conformation of the loop of interest remains to be determined. Effectively, the loop is a tethered peptide whose conformations can be sampled extensively while in the presence of the energy field generated by the rest of the protein. Many different conformations of the loop can be generated and tested for false global minima which exist in the presence of the effective potential field of the protein framework. This makes the protein loop prediction problem a powerful benchmarking tool to test the accuracy of energy functions.

An accurate molecular mechanics model suitable for protein structure prediction and refinement requires a representation of the aqueous solvent environment. The polarization of the solvent favors the hydration of polar and especially charged groups that, in the absence of solvation forces, tend to form non-native intramolecular interactions. Explicit solvent models provide the most detailed and complete description of hydration phenomena.[29] However, computer simulations using explicit solvent models are computationally intensive, not only just because of the much larger number of atomic interactions that need to be considered but also, and perhaps more importantly, because of the need to average the fluctuating effects of the solvent reaction field to obtain a meaningful estimate of the solvation free energy of each protein conformation. For protein structure prediction applications effective potential models that treat the solvent implicitly have much to offer. The modeling community has developed a strong interest in a class of implicit solvent models based on the Generalized Born framework;[30–32] an approximation of the Poisson equation of continuum electrostatics.[33,34] Much of the popularity of Generalized Born (GB) models stems from their computational efficiency and ease of integration in molecular simulation computer programs.[31,35–38] Generalized Born models have been shown to be able to reproduce with good accuracy Poisson[32,39–41] and explicit solvent[42,43] results at a fraction of the computational expense.

In this work we evaluate the accuracy of the Analytical Generalized Born plus Non-Polar (AGBNP) implicit solvent model,[44] in predicting the native conformation of protein loops using the Protein Local Optimization Program (PLOP).[26] The PLOP program[26] performs loop and side chain conformational predictions based on an efficient hierarchical conformational sampling algorithm in torsional angle space, combined with a recent parametrization of the OPLS-AA force field[45,46] and a Generalized Born implicit solvation model. The AGBNP implicit solvent model is based on an analytical pairwise descreening[47] implementation of the Generalized Born model[30] and a novel nonpolar hydration free energy model which combines separate estimators for the solute−solvent van der Waals dispersion energy and the work of cavity formation.[48–50]

We previously showed[44] that the OPLS-AA/AGBNP effective potential was able to consistently score native loop conformations more favorably than non-native decoy loop conformations generated by PLOP using the OPLS-AA/SGB/NP effective potential.[26] The present work extends that work by including a larger set of loops as well as longer loops targets and by employing the OPLS-AA/AGBNP model directly in the conformational search and optimization procedure implemented in PLOP. We also evaluate various parametrizations of the AGBNP model to determine the role of the nonpolar model and of the correction terms we developed aimed at reducing the occurrence of intramolecular ion pairing, and we compare them to the distance dependent dielectric and the Surface Generalized Born (SGB/NP)[51,52] solvation models as implemented in the PLOP program.

As part of this work we have also evaluated the efficiency of the recently proposed loop conformational search schemes based on PLOP[26,53] which improves on earlier torsion angle based sampling methods.[24,25] These PLOP-based conformational search schemes have been optimized for loop conformational prediction in the crystal environment. We evaluate enhanced versions of these schemes more suitable for loop prediction calculations in the solution environment (the biologically relevant environment for most homology modeling applications). We also tested the applicability of temperature Replica Exchange Molecular Dynamics (T-REMD) to the problem of protein loop prediction, which, given its favorable scaling with respect to the number of degrees of freedom, offers an alternative route for conformational prediction of long loops and for simultaneous refinement of interacting protein elements.

2. Methods

2.1. Loop Prediction Algorithms. The loop prediction algorithm implemented in the Protein Local Optimization Program (PLOP) is described in detail in ref 26. During loop buildup, a series of filters of increasing complexity is applied to eliminate unreasonable conformations as early as possible. Some of these filters detect clashes between backbone atoms and the atoms of the rest of the protein (referred to as the frame) and check that enough space is available to place the side chain of each residue. On the order of hundreds to thousands of loop conformations are generated in the loop build-up stage. To reduce the number of conformations passed to the next stages, loop conformations are clustered based on backbone rmsd using the K-means algorithm,[54] a clustering method that requires a predetermined number of clusters. The two most important parameters that control the tradeoff between accuracy and efficiency of PLOP's loop prediction algorithm are the overlap factor parameter (*ofac*), defined as the minimum permitted ratio of the interatomic distance over the sum of the Lennard-Jones radii of the atoms of interest, which controls the amount of overlap tolerated between any two atoms, and the number of clusters N_{clust}. A smaller *ofac* allows more overlap between atoms which in

effect allows for more loop conformations to be sampled which otherwise would have been eliminated due to steric clashes. The efficiency of the loop-prediction procedure is partially determined by the value of *ofac*. If *ofac* is too small, a large number of irrelavent loop conformations are generated that have to be processed in subsequent steps. On the other hand with a large *ofac* nativelike loops may be rejected due to steric clashes caused by the discreteness of the torsion library used to generate the loops. Based on the oberved value of *ofac* found in the PDB, Jacobson et al. set *ofac* to between 0.70 and 0.75.[26] The number of clusters N_{clust} needs to be sufficiently large to account for each nonredundant loop conformation. If N_{clust} is set too small, conformationally different structures could potentially be clustered together. The number of nonredundant loop conformations will depend upon how large is the conformational space available to the loop. Based on empirical evidence, Jacobson et al. set the number of clusters to four times the number of residues in the loop.[26]

The PLOP program allows sampling of loop conformations in the crystalline phase with the SGB/NP solvation model.[42,42] We performed SGB/NP prediction calculations with and without crystal symmetry in order to compare with previous literature.[26] Loop prediction calculations with all of the other implicit solvent models were conducted without crystal symmetry.

The basic loop prediction algorithm described above is often insufficient for loops with nine or more residues. For these longer loops we have adopted prediction schemes based on multiple executions of PLOP with different parameters.[26,53] These schemes are based on focusing conformational sampling in promising and progressively smaller regions of conformational space. The initial predictions with the most favorable energy scores are subjected to a series of constrained refinement calculations with PLOP in which selected loop backbone atoms are not allowed to move or move only within a given range.

The standard 9-residue loop prediction scheme is based on the procedure described in detail by Jacobson et al.[26] For loops which the standard version of loop prediction fails to find low-energy, nativelike conformations, we attempted to predict these loops with an extended version of the loop prediction algorithm. An extended version of this scheme involves using twice the number of clusters (from 36 to 72) and reduced *ofac* (overlap factor) coefficients (0.5 instead of 0.75) during the initial prediction stage. All other stages as described by Jacobson et al.[26] remain the same.

For the 13-residue loops we have adopted an alternative long loops prediction scheme developed previously for longer loops.[53] This scheme is based on the idea of refining the loop structure by sampling increasingly shorter loop segments which can be handled by PLOP's conformational search procedure. Briefly, initial predictions are produced with 3 different overlap factors (0.65, 0.70, and 0.75) and subjected to constrained refinement. The five lowest-energy nonredundant structures so obtained are passed to a series of loop prediction stages which sample progressively shorter segments obtained by fixing any possible combination up to five residues at either terminal end of the loop. The *standard sampling* and *extended sampling* variations of this sampling method differ in the number of nonredundant lowest-energy models that are processed at each stage. With extended sampling five lowest-energy models are passed from one stage to the next. With standard sampling the number of PLOP iterations is reduced by half by progressively reducing the number of models passed to later stages.

Table 1. 9-Residue and 13-Residue Loops Indicated by Their the Protein Data Bank (PDB) Designation for the Protein and R_{first} and R_{last} Are, Respectively, The First and Last Residue of the Loop[a]

PDB(R_{first} - R_{last})	PDB(R_{first} - R_{last})	PDB(R_{first} - R_{last})
1aac(58−66)	1pda(108−116)	1cnv(110−122)
1aba(69−77)	1pgs(117−125)	1d0c(A:280−292)
1amp(57−65)	1php(91−99)	1dpg(A:352−364)
1arb(90−98)	1ptf(10−18)	1dys(A:290−302)
1arb(168−176)	1ra9(142−150)	1ed8(A:67−79)
1arp(127−135)	1rhs(216−224)	1eok(A:147−159)
1aru(36−44)	1sgp(E109−E117)	1f46(A:64−76)
1btl(102−110)	1tca(170−178)	1g8f(A:72−84)
1byb(246−254)	1tca(217−225)	1gpi(A:308−320)
1cse(E95−E103)	1xif(59−67)	1h4a(X:19−31)
1csh(252−260)	1xnb(116−124)	1hnj(A:191−203)
1ede(257−265)	1xyz(A568−A576)	1hxh(A:87−99)
1fus(31−39)	1xyz(A795−A803)	1iir(A:197−209)
1fus(91−99)	1xnb(133−141)	1jp4(A:153−165)
1gpr(63−71)	1wer(942−950)	1kbl(A:793−805)
1isu(A30−A38)	2alp(139−159)	1krh(A:131−143)
1ivd(244−252)	2ayh(169−177)	1l8a(A:691−703)
1lkk(A142−A150)	2cpl(24−32)	1lki(62−74)
1lkk(A193−A201)	2eng(172−180)	1m3s(A:68−80)
1mla(194−202)	2hbg(18−26)	1mo9(A:107−119)
1mrj(92−100)	2sil(183−191)	1nln(A:26−38)
1mrk(53−61)	3pte(78−86)	1o6l(A:386−398)
1mrp(284−292)	3pte(107−115)	1ock(A:43−55)
1nfp(12−20)	3pte(215−223)	1ojq(A:167−179)
1nif(266−274)	3tgl(56−64)	1p1m(A:327−339)
1nls(131−139)	4gcr(94−102)	1qqp(2:161−173)
1noa(9−17)	1a8d(155−167)	1qs1(A:389−401)
1noa(76−84)	1ako(203−215)	1xyz(A:645−657)
1noa(99−107)	1arb(182−194)	2hlc(A:91−103)
1npk(102−110)	1bhe(121−133)	2ptd(136−148)
1onc(70−78)	1bkp(A:51−63)	

[a] A letter indicates the chain on which the loop is found.

We also investigated if a technique based on replica exchange molecular dynamics importance sampling could predict loop conformations. We selected 9-residue loops which were not successfully predicted by the standard sampling algorithm built around PLOP to see if importance sampling would succeed. This subset of the 9-residue loops (Table 3) was investigated with the temperature replica exchange sampling method (T-REMD)[55−57] as implemented in the IMPACT software package.[57] The lowest-energy loop configuration obtained in the third stage of PLOP optimization was chosen as a starting point for the corresponding T-REMD run. Each loop was minimized in the field of the surrounding immobilized protein frame. T-REMD was based on constant temperature MD, and exchanges between replicas were attempted every 500 steps. During T-REMD simulations, the protein frame conformation was fixed. The OPLS-AA force field was employed to model the intramolecular potential, while the solvent was treated implicitly by the AGBNP+ effective potential model (see below). We used 12 replicas at 270, 298, 329, 363, 401, 442, 488, 539, 595,

Table 2. Summary of the Loop Conformational Predictions Results with the Standard and Enhanced Sampling Procedures[a]

	9-residue						
	SGB/NP-X	SGB/NP	ddd	AGB-γ	AGBNP	AGBNP+	13-residue AGBNP+
E	8	11(10)	19	6	4(3)	2	2
S	5(5)	7(14)	4(7)	4(7)	4(9)	5(10)	5(14)
M	2	3(4)	3	1	0	1	1(2)
E+*S*+*M*	15	21	26	11	8	8	8
⟨rmsd⟩	1.44	1.91	2.31	1.10	1.04	1.00	1.87
median rmsd	0.58	0.60	1.27	0.52	0.52	0.58	0.67

[a] SGB/NP-X: SGB/NP with crystal symmetry; ddd: distance-dependent dielectric; *E*: number of energy errors (results listed for both enhanced and (standard) sampling); *S*: number of sampling errors (results listed for both enhanced and (standard) sampling); *M*: number of marginal errors (results listed for both enhanced and (standard) sampling). The values listed were obtained with enhanced sampling; the values in parentheses were obtained with standard sampling. ⟨rmsd⟩: average rmsd (in Å) of the lowest-energy loops.

Table 3. Summary of the Loop Conformational Predictions Results with the OPLS-AA/AGBNP+ Force Field and T-REMD Conformational Sampling, Compared to the Corresponding Predictions with the PLOP-Based Standard Sampling Procedure

PDB(R_{first} - R_{last})	PLOP rmsd (Å)	T-REMD rmsd (Å)
1npk(102−110)	3.60	4.30
1onc(70−78)	7.43	2.06
1fus(31−39)	6.03	1.78
1byb(246−254)	4.00	4.95
1noa(99−107)	5.67	3.94
1wer(942−950)	4.29	1.34

657, 725, and 800 K. The T-REMD simulation length varied from 15 to 35 ns, and the data collected over the last 5 ns of the T-REMD trajectories were used for final analysis.

2.2. The Energy Functions. The energy functions we used to score the predicted loops are composed of the all-atom force field, OPLS-AA,[45,46] and an implicit solvent model. The particular version of OPLS-AA[46] we used has improved torsional parameters based on fits to high-level LMP2 quantum chemical calculations of the torsion interactions of small peptides. These fits led to improvements in the accuracy of the φ, ψ, and side chain χ torsion energies for amino acids.[27]

The implicit solvent models we investigated in this study are the simple distance-dependent dielectric and two generalized Born solvation models, the Surface Generalized Born (SGB) [42,42] and Analytical Generalized Born (AGB).[44] It is assumed in the distance-dependent dielectric model that the interaction energy between partial charges in a heterogeneous dielectric environment follows a simple Coulomb law. The Coulomb energy term is given by

$$E_{\mathrm{Coul}} = \frac{q_i q_j}{\varepsilon r_{ij}} \tag{1}$$

where r_{ij} is the interatomic distance between atoms i and j, and ε is the dielectric constant. In the distance-dependent dielectric model, ε is no longer constant but proportional to the interatomic distance as such

$$\varepsilon = r_{ij} \tag{2}$$

While the distance-dependent dielectric is known to be a poor model for solvation, we use the results generated with it to benchmark the improvements in loop prediction that can be obtained with more accurate physical models.

2.2.1. SGB/NP Implicit Solvent Model. The SGB model is the surface implementation[42,51] of the generalized Born model.[30] The generalized Born equation

$$G_{\mathrm{GB}} = -\frac{1}{2}\left(\frac{1}{\varepsilon_{\mathrm{in}}} - \frac{1}{\varepsilon_{\mathrm{w}}}\right)\sum_{ij}\frac{q_i q_j}{f_{ij}(r_{ij})} \tag{3}$$

where q_i is the charge of atom i and r_{ij} is the distance between atoms i and j, gives the electrostatic component of the free energy of transfer of a molecule with interior dielectric $\varepsilon_{\mathrm{in}}$ from vacuum to a continuum medium of dielectric constant ε_{w}, by interpolating between the two extreme cases that can be solved analytically: the one in which the atoms are infinitely separated and the other in which the atoms are completely overlapped. The interpolation function f_{ij} in eq 3 is defined as

$$f_{ij} = \left[r_{ij}^2 + B_i B_j \exp\left(-r_{ij}^2/4B_iB_j\right)\right]^{\frac{1}{2}} \tag{4}$$

where B_i is the Born radius of atom i defined as the effective radius that reproduces through the Born equation

$$G^i_{\mathrm{single}} = -\frac{1}{2}\left(\frac{1}{\varepsilon_{\mathrm{in}}} - \frac{1}{\varepsilon_{\mathrm{w}}}\right)\frac{q_i^2}{B_i} \tag{5}$$

the electrostatic free energy of the molecule when only the charge of atom i is present in the molecular cavity. The G^i_{single} are evaluated numerically by integrating the interaction between atom i and the charge induced on the solute−solvent boundary surface, S, by the Coulomb field of this atom

$$G^i_{\mathrm{single}} = -\frac{1}{8\pi}\left(\frac{1}{\varepsilon_{\mathrm{in}}} - \frac{1}{\varepsilon_{\mathrm{w}}}\right)\int_S \frac{q_i^2}{|\mathbf{r} - \mathbf{r}_i|^4}(\mathbf{r} - \mathbf{r}_i)\cdot\mathbf{n}(\mathbf{r})d^2\mathbf{r} \tag{6}$$

where $\mathbf{n}(\mathbf{r})$ is the normal to the surface, S, at $\mathbf{r}$. The atomic radii that define the solute−solvent dielectric boundary are set to the van der Waals radii based on the Lennard-Jones σ parameters. The Born radii for eq 4 are calculated using eqs 5 and 6. In this work, we set $\varepsilon_{\mathrm{in}} = 1$ and $\varepsilon_{\mathrm{w}} = 80$. The SGB implementation used in this work includes further correction terms that bring the SGB reaction field energy into closer agreement with exact PB results.[51] Coupled with the SGB model is a function describing the nonpolar interactions between the solute and solvent which is based on two terms: the van der Waals interaction between solute and solvent and the work to form the cavity in the solvent. The full

solvation model is referred to as SGB/NP. Exact details of the nonpolar function in SGB/NP can be found in ref 52.

2.2.2. AGBNP Implicit Solvent Model. The analytical generalized Born (AGB) implicit solvent model differs from SGB in the way that the Born radii are calculated. AGB is based on a novel pairwise descreening implementation[44] of the generalized Born model.[58] The combination of AGB with a recently proposed nonpolar hydration free energy estimator described below is referred to as AGBNP.[44] AGB employs a parameter-free and conformation-dependent analytical scheme to obtain the pairwise descreening scaling coefficients used in the computation of the Born radii used in the generalized Born equation, eq 3. The agreement between the AGB Born radii and exact numerical calculations was found to be excellent.[44] The AGBNP nonpolar model consists of an estimator for the solute−solvent van der Waals interaction energy in addition to an analytical surface area component corresponding to the work of cavity formation.[44] Because AGBNP is fully analytical with first derivatives it is well suited for energy minimization as well as MD sampling. A detailed description of the AGBNP model and its implementation is provided in ref 44.

The nonpolar solvation free energy is given by the sum of two terms: the free energy to form the cavity in solvent filled by the solute and the dispersion attraction between solute and solvent.[49,59] The nonpolar free energy is written as[44]

$$\Delta G_{\mathrm{np}} = \sum_i \left(\gamma_i A_i + \Delta G_{\mathrm{vdW}}^{(i)}\right) \tag{7}$$

where the first term is the cavity term, γ_i, is the surface tension proportionality constant for atom i, and A_i is the solvent exposed surface area of atom i. The second term is the dispersion interaction term which is given by[44]

$$\Delta G_{\mathrm{vdW}}^{(i)} = \alpha_i \frac{-16\pi\rho_w \varepsilon_{i,w}\sigma_{i,w}^6}{3(B_i + R_w)^3} \tag{8}$$

where α_i is an adjustable solute−solvent van der Waals dispersion parameter for atom i. The parameter ρ_w is the number density of water at standard conditions (0.033428/Å^3). $\varepsilon_{i,w}$ and $\sigma_{i,w}$ are the pairwise Lennard-Jones (LJ) well-depth and diameter parameters for atom i and the TIP4P water oxygen as given by the OPLS-AA force field.[45,46] ($\varepsilon_{i,w} = \sqrt{\varepsilon_i \varepsilon_{w,}}$ where ε_i is the LJ well-depth for atom i and ε_w is similarly for the TIP4P water oxygen. The ε for water hydrogens is set to zero. $\sigma_{i,w}$ is defined in a similar manner.) R_w is the radius of a water molecule (1.4 Å). By not incorporating the Lennard-Jones parameters into the dispersion parameter, α_i, atoms with different though similar ε_i's and σ_i's are assigned the same α so as to minimize the number of adjustable parameters. B_i is the Born radius of atom i. The Born radius in this equation provides a measure of how buried atom i is in the solute. The deeper the atom is in the solute, the smaller will be its contribution to the total solute−solvent dispersion interaction energy. The functional form of ΔG_{vdW} in both SGB/NP and AGBNP depends upon the Born radius since it is a measure of the degree of burial of the atom. In SGB/NP, the dependence of ΔG_{vdW} on the Born radius was chosen on an ad hoc basis. The form of eq 8 for the solute−solvent van der Waals interaction energy component has been derived on the basis of simple physical arguments.[44]

In this work we use two sets of parametrizations of α and γ to test the full nonpolar function described above relative to a simpler nonpolar function. In past implementations,[19] the total nonpolar solvation free energy is given by a term proportional to the solvent-accessible surface area, or in terms of eq 7, setting all values of α_i to zero

$$\Delta G_{\mathrm{np}} = \sum_i (\gamma_i A_i) \tag{9}$$

where γ_i is set for all atoms to 0.015 kcal/mol/Å^2. This implicit solvent model with the less-detailed nonpolar function is referred to as "AGB-γ". When we use the full nonpolar function including the dispersion term (eq 8) using the parameters set forth in the work of Gallicchio and Levy,[44] the implicit solvent model is referred to as "AGBNP".

A third parametrization aimed at implementing a correction for salt bridge interactions (which are generally overestimated by generalized Born solvent models)[56,60] is also investigated. To correct for the overstabilization of salt bridges by the generalized Born model, we used modified radii and γ_i for carboxylate oxygens. The radius of the carboxylate oxygen is decreased from 1.48 Å, as in the original AGBNP, to 1.30 Å; γ_i of the carboxylate oxygen is set to −0.313 kcal/mol/Å^2. These have the combined effect of increasing the solubility of carboxylate oxygens and decreasing the likelihood of ion pairing between the carboxylate groups on glutamate and aspartate and positively charged groups found on lysine and arginine. We have parametrized this radius and γ_i to experimental data for small molecules and to provide results which matched those generated with explicit solvent (unpublished results). The implicit solvent model that has additional descreening of ion pairing is referred to as "AGBNP+".

2.3. The Protein Loop Data Sets. We have tested the loop prediction algorithms on two sets of protein loops of known structure of nine and 13 residues in length. The first set is composed of the 57 9-residue loops listed in Table 1. This set was originally compiled by Fiser et al.[24] and by Xiang et al.[25] The 35 13-residue loop set is the same as the one investigated by Zhu et al.[53] These loops were culled from the PISCES[61] database. The proteins in these databases have been filtered using the following selection criteria: (i) low sequence identity (60% for Fiser et al.,[24] 20% for Xiang et al.,[25] and <40% for Zhu et al.),[53] (ii) complete X-ray structure available with resolution <2 Å, R < 0.25, and average temperature factor within the loop <35, (iii) 6.5 < pH < 7.5, (iv) overlap factor for any loop atom >0.7, (v) no significant loop secondary structure, (vi) no more than 4 additional loop residues on either side of the selected loop, (vii) distance between any loop atom and any ligand atom >4 Å (6.5 Å for a metal ion).[26,53] While some of the loops contain very small amounts of secondary structure, in general, they are representative of longer loops found in globular proteins. All crystallographic water molecules are removed prior to loop prediction. Hydrogen atoms are added to each structure.[26]

2.4. Characterization of the Predicted Loop Structures. The predicted loop conformation is the one that has the lowest energy among those found by the conformational search procedures described above. The accuracy of the predicted conformations is analyzed by computing their root-mean-square deviation (rmsd) with respect to the corresponding crystallographically determined native structures (the X-ray structure). The native and predicted protein loops are already in a common frame because only the conformation of the loop is varied during loop torsion angle sampling. The rmsd of the backbone atoms (N, C, and C_α) predicted and X-ray conformations are calculated in this common frame. We characterize the accuracy of the predictions based on the average and median backbone rmsd of the predictions and the number of correct predictions. Correct predictions are defined as those that fall within a chosen rmsd threshold value from the X-ray structure.

An incorrect prediction (one with an rmsd larger than the threshold, see below) is further classified as an *energy error* when the prediction has an energy significantly lower than native, and otherwise as a *sampling error*, when the predicted loop has an energy higher than the native. This classification of incorrect predictions is aimed at determining the cause of the failure of the method to produce a nativelike conformation. An energy error is indicative of the failure of the energy function to score the native conformation more favorably than non-native conformations; so that, even if the conformational search method had produced them, near-native conformations would not be recognized as good predictions. A sampling error is indicative of the conformational search procedure failing to sample conformations near the native conformation, even though the energy function scores at least some of them more favorably than non-native conformations.

The classification of correct and incorrect predictions requires the specification of a rmsd threshold value. This choice depends on the level of prediction accuracy required by the application. We report our results based on C_α rmsd thresholds of 1.5 and 2.0 Å for the 9- and 13-residue loop sets, respectively, which have been used before to analyze the accuracy of loop prediction methods.[26,53] In addition, the classification of incorrect predictions requires the specification of an energy gap threshold value. If the difference in energies of the native and predicted conformations (where the predicted is lower in energy than the native) exceeds the energy gap threshold value, the incorrect prediction is classified as an energy error. In this work the results have been reported using an energy gap threshold value of 5 kcal/mol. The choice of this value absorbs the effects due to configurational entropy missing from our free-energy estimator as well as the acceptable level of error in the energy function. We have explored a range of rmsd and energy gap threshold parameters and confirmed that the conclusions drawn in this work are not qualitatively affected by the particular choices made here. The energy of the native conformation used in the computation of the energy gap of the predicted conformation is determined in three ways: (1) a minimization of the loop with the frame, (2) a minimization followed by an optimization of the side chains on the loop, and (3) a confined search within 2 Å rmsd from the X-ray conformation similarly as for the second stage of refinement in the loop prediction procedure. We selected the native energy as the lowest energy determined from any of these. In almost all cases this conformation differs from the X-ray structure by no more than 1 Å C_α rmsd.

A minority of incorrect predictions were not classifiable as either energy errors or sampling errors. These were typically cases that do not qualify as clear energy errors because, even though the energy of the predicted non-native conformation is lower than the native conformation, the magnitude of the energy gap is within the 5 kcal/mol margin and do not qualify as sampling errors because native conformations of reasonable low energy were sampled. In the following we label these cases as *marginal errors*. Marginal errors are effectively incorrect predictions due to subtle and not easily attributable energetic, entropic, and methodological causes.

In order to be able to compare the T-REMD predictions with those obtained from the PLOP-based prediction schemes and with the native structures, we energy-minimized the loop conformations found at the lowest target temperature of 270 K and recomputed the loop backbone rmsds with respect to the reference crystal structure. The conformation with the lowest energy was selected as the predicted conformation. The predicted conformation was then classified in terms of the energy gap and rmsd from the native conformation using the scheme described above.

3. Results

The results of the loop prediction tests are summarized in Table 2 for the standard and extended conformational sampling procedures (see Methods). Extended sampling was conducted on the loops that resulted in a sampling error with standard sampling; Table 2 includes the combined standard and extended sampling results. For the 57 9-residue loops (see Table 1) loop prediction tests were conducted with OPLS-AA and the following implicit solvent models: distance-dependent dielectric, SGB/NP, AGB-γ, AGBNP, and AGBNP+. It has been stated that the results for loop prediction with PLOP was independent of the presence of crystal symmetry.[26] However, we found that crystal symmetry significantly influenced the results with SGB/NP. In order to compare with previous results,[26] we performed loop predictions with SGB/NP both in the presence and absence of crystal symmetry. Loop prediction calculations with all of the other implicit solvation models were conducted only in the absence of crystal symmetry. Loop prediction tests for the 35 13-residue loops (see Table 1) were conducted with AGBNP+. As described in the Methods section we characterized each loop prediction as being either correct or incorrect. In turn each incorrect prediction is classified as an energy error, a sampling error, or a marginal error. Table 2 reports the total number of errors and the number of energy and sampling errors and the mean and median rmsd of the predictions from the X-ray structure.

The results in Table 2 for the 9-residue loops demonstrate that the total number of prediction errors (energy and sampling) is the lowest for the AGB implicit solvent models.

The distance-dependent dielectric model (ddd) performs the worst, followed by SGB/NP in the absence of crystal symmetry. The introduction of crystal symmetry results in a significant reduction in the number of sampling errors. (This is discussed further below.) Of the three AGB-based models, AGB-γ which mimics GB/SA is the one with the largest number of prediction errors, whereas AGBNP and AGBNP+ are equivalent in this respect. The number of energy errors, a measure of the quality of the energy model, varies greatly from one energy model to another. The fewest energy errors are found with AGBNP+, followed by in order AGBNP, AGB-γ, SGB/NP with crystal symmetry, SGB/NP, and distance-dependent dielectric. The number of sampling errors in general does not vary as greatly from one energy model to another, and their occurrence decreases significantly by using the extended sampling procedure (as shown in Table 2). This is particularly noticeable for the 13-residue loops for which two-thirds of the sampling errors with standard sampling are avoided (decrease 14 errors to five) when using extended sampling.

Comparison of the results for SGB/NP with and without crystal symmetry reveals that the inclusion of crystal symmetry has a dramatic effect on the number of sampling errors when using standard sampling; SGB/NP without crystal symmetry produces 14 sampling errors compared to five sampling errors with crystal symmetry (see Table 2). The effect of crystal symmetry on the number of sampling errors is greatly diminished when using extended sampling (Table 2). With extended sampling the number of SGB/NP sampling errors drops to seven, whereas the number of sampling errors (five) with SGB/NP with crystal symmetry is unchanged.

Table 2 also reports the mean and median rmsd of the loop predictions with respect to the X-ray structure. The mean rmsd of the 9-residue loops predictions with the AGB-based energy models is around 1 Å, which is significantly better than all the other solvation models including SGB/NP with the inclusion of crystal symmetry. The worst mean rmsd for the 9-residue loops is 2.31 Å obtained with the distance-dependent dielectric model. The median rmsd's, which are less affected by outliers corresponding to grossly incorrect predictions, are significantly smaller than the mean rmsd's. The difference between mean and median rmsd's is larger for SGB/NP-based and distance-dependent dielectric models than AGB-based solvation models due to the fact that incorrect predictions with the latter are generally closer to the X-ray structures than with the other models. The larger difference between mean and median rmsd for the 13-residue loop predictions with AGBNP+ relative to the 9-residue loop predictions reflects the fact that, expectedly, incorrect predictions with the longer loops tend to be farther away from the X-ray structure in terms of rmsd.

We repeated loop prediction calculations for six of the 9-residue protein loops classified as sampling errors with the loop prediction algorithm and using the AGBNP+ solvation model, using the T-REMD sampling procedure described in the Methods section. These loops are 1npk (residues 102−110), 1onc (70−78), 1fus (31−39), 1byb (246−254), 1noa (99−107), and 1wer (942−950) (see Table 1). We sampled these loops using temperature replica exchange molecular dynamics (T-REMD) as implemented in the IMPACT molecular mechanics package. The distribution of conformations in terms of potential energy and rmsd from the X-ray structure from the last 5 ns of the T-REMD trajectories for 1fus (31−39) is shown in Figure 4. The rmsd from the native of the lowest-energy conformations extracted from the T-REMD trajectories is reported in Table 3. For comparison, this table also reports the corresponding predictions using the standard conformational search procedure with PLOP. This table shows that in half of the cases examined (1onc, 1fus, and 1wer), T-REMD is able to produce predictions significantly closer to the X-ray structure than the PLOP-based standard sampling procedure. However, only one (1wer) of the six incorrect PLOP-based predictions results in a correct prediction with T-REMD, based on the 1.5 Å rmsd threshold value.

4. Discussion

4.1. Prediction Accuracy. The loop prediction procedure based on PLOP with the AGBNP+ solvation model and the extended sampling schemes we devised is very successful in predicting the conformations of the 9- and 13-residue loops we have investigated. As Table 2 shows, the successful prediction rate is 86% and 77% for 9- and 13-residue loops, respectively. We obtained a signficant reduction in the rates of successful predictions when using the SGB/NP and distance-dependent dielectric solvation models, even when we include crystal symmetry.

Although in this work we define the predicted conformation as the lowest-energy loop conformation, it is interesting to examine also how well the loop prediction procedure captures nativelike conformations within a given energy range from the minimum energy conformation found. In homology modeling, the choice of the candidate structures may not be restricted to selecting only the lowest-energy conformation. It may be desirable to investigate structures whose energies lie within some range about the minimum energy structure found in the search. For instance, a modeler may consider all those structures whose energies are within the lowest 5 kcal/mol as possible candidates to represent the native conformation. Under this scenario the prediction calculation can be considered successful if any one of the candidate conformations approximates well the native conformation. While the energy range is increased, the probability of including a nativelike conformation increases at the expense of the greater cost associated with having to carry over a larger number of candidate conformations. On average there are roughly 150 loop predictions per protein within 5 kcal/mol from the minimum energy. Figure 1 illustrates this cost/benefit analysis for the 57 9-residue loop prediction calculations (Table 2). Each point on the curves in Figure 1 was obtained by collecting for each loop target the set of predicted conformations with energies within a given energy range ΔE from the energy of the lowest-energy prediction and recording their number N as well as whether at least one native conformation (within 1.5 Å rmsd from the X-ray conformation) is contained in this set, that is

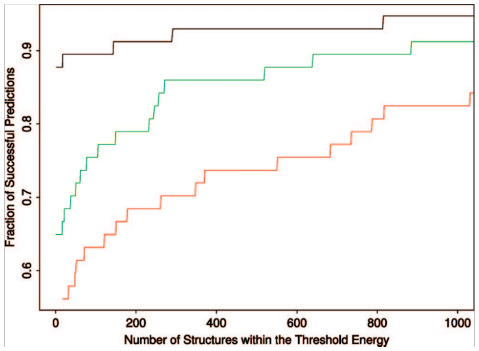

Figure 1. We plotted the ratio between the number of successfully predicted loop targets and the total number (57) of loop targets (the fraction of successful predictions) for a given threshold energy, ΔE, versus the average number of low-energy predicted conformations within this value of ΔE for the AGBNP+, SGB/NP, and distance-dependent dielectric solvation models. All loop predictions are ordered relative to their energy from the lowest-energy prediction. For an average given number of loops from the minimum (the abscissa), the fraction of proteins that have at least one nativelike loop among the top number of loops is shown above along the ordinate. The black line presents the results for AGBNP+, the green line presents the results for SGB/NP, and the red line presents the results for distance-dependent dielectric.

whether for this particular loop target and energy range the result is regarded as a successful prediction. We did this over a range of ΔE values for all 9-residue targets and solvation models. We then plotted the ratio between the number of successfully predicted loop targets and the total number (57) of loop targets (the fraction of successful predictions) for a given ΔE versus the average number of low-energy predicted conformations within this value of ΔE for the AGBNP+, SGB/NP, and distance-dependent dielectric solvation models (see Figure 1). The abscissa in this plot represents the cost, as measured by the number of conformations that one is willing to consider as possible candidates, whereas the ordinate represents the benefit, as measured by the probability of including at least one native conformation within this set of conformations. This plot can be used in two complementary ways. Given the maximum cost one is willing to sustain on the abscissa the corresponding ordinate of the curves yields for each solvation model the expected rate of success. Alternatively, given the desired rate of success in the ordinate, the curves give the required associated cost.

The minimum cost corresponds to retaining only the lowest-energy prediction ($N = 1$). This assumes that the lowest-energy loop prediction from the algorithm is the native conformation without any additional analysis. For this value of N the success rates are 86%, 77%, and 55% for the AGBNP+, SGB/NP, and distance-dependent dielectric models, respectively, see Figure 1. For all values of ΔE examined, the AGBNP+ solvation model provides the best success rate for a given cost level, followed by SGB/NP and the distance-dependent dielectric solvation models. A greater cost level

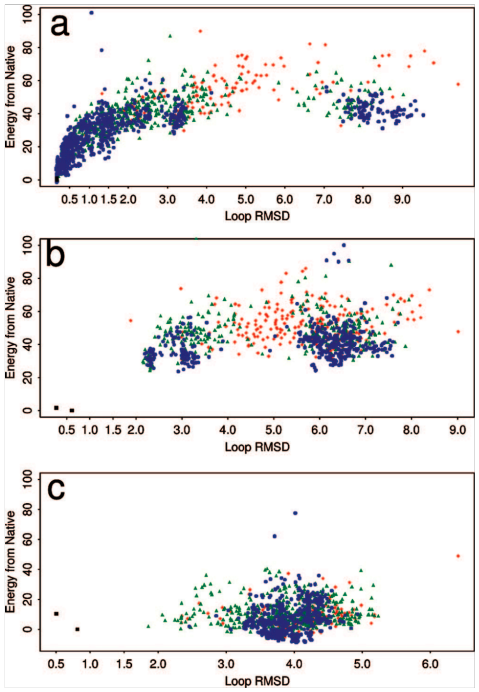

Figure 2. Energy gaps relative to the optimized native conformation (in kcal/mol) versus the rmsd (in Å) relative to the X-ray crystal conformation for three representative 9-residue loop prediction cases with the OPLS-AA/AGBNP+ potential and the standard conformational sampling algorithm: (a) 1php(91−99) (a successful prediction), (b) 1fus(31−39) (a sampling error), and (c) 3pte(215−223) (an energy error). The initial prediction results are in red, the first stage of refinement is in green, and the second stage of refinement is in blue. The native (minimized and optimized) are in black.

entails retaining more than one low-energy loop conformation which would have to be analyzed in more detail. Conversely, AGBNP+ yields a higher success rate with less cost than the other solvation models; for example, to obtain with the SGB/NP model a success rate of 86% requires considering on average 500 conformations. To obtain a similar success with distance-dependent dielectric would require consideration of over 1000 conformations on average per loop target.

It is useful to compare our results with those obtained by other groups for 9-residue and 13-residue loops. Fiser et al. used MD along with simulated annealing to predict loop conformations with an all-atom force field and a statistical treatment of solvation.[24] The percentage of predictions they report within 2 Å rmsd (described as good and medium predictions) is 55%.[24] Using a tighter rmsd cutoff of 1.5 Å, we obtain with PLOP and AGBNP+ an 86% success rate in our predictions for 9-residue loops. For a set of 13-residue loops, Fiser et al., using the same 2 Å rmsd cutoff, report a very low 15% success rate,[24] compared to the 77% success rate we obtained using the AGBNP+ scoring function. Xiang

et al. performed a search over a discrete rotamer library with scoring based on their colony energy. For 9-residue loops, they report an average rmsd of 2.68 Å.[25] In comparison the average rmsd we have obtained with PLOP and AGBNP+ is 1.00 Å. De Bakker et al.[62] generated loop conformations with their program RAPPER[63] and scored them with a knowledge-based potential and with a physics-based potential, AMBER/GBSA. For 9-residue loops from the Fiser set,[24] the average rmsd of the lowest-energy loops was over 2 Å when scored with the AMBER/GBSA potential which produced their best results.[62]

Jacobson et al.[26] performed loop prediction calculations on a large set of 9-residue loops using the SGB/NP model with the crystal symmetry included and using the standard conformational sampling algorithm used here.[26] Based on the Supporting Information they provided,[26] we were able to determine the number of energy and sampling errors using a 1.5 Å rmsd cutoff and a −5 kcal/mol energy cutoff. Based on our analysis of their data, they had obtained ten energy errors and eight sampling errors.[26] In comparison, we find 11 energy and seven sampling errors with SGB/NP without crystal symmetry, but we find only eight energy errors and five sampling errors with SGB/NP with crystal symmetry. This might indicate that crystal symmetry is important for prediction accuracy; however, we obtained two energy errors and five sampling errors using AGBNP+ without the presence of the crystal environment. A recent study based on the comparison of X-ray and NMR structures of identical proteins suggests that in most cases the impact of the crystal environment on protein structures is relatively small and not strongly correlated with crystal packing.[64] Recently, Zhu et al.[53,65] have reported loop prediction results for the same 35 13-residue loops investigated here using the SGB/NP potential with crystal symmetry supplemented by hydrophobic correction terms and a variable dielectric model. Zhu et al. show that these promising models lower the average backbone rmsds of the 13-residue predictions substantially, from 2.73 Å to 1.08 Å. In comparison, we obtain for the 13-residue loop set with AGBNP+ without crystal symmetry an average rmsd of 1.87 Å which is intermediate between the range of rmsd measures reported by Zhu et al.[53,65] The best performing model reported by Zhu et al. produces according to our definition five energy errors on the 13-residue loop set (see the Supporting Information of reference 65) compared with the two energy errors obtained here.

4.2. Accuracy of Scoring Functions. The ability of the effective potential model to consistently score native conformations more favorably than non-native conformations is essential for successful loop prediction. The results in Table 2 for the 9-residue loops indicate that significant differences, in terms of the number of energy errors, exist between the different solvation models we investigated. We observed that the occurrence of energy errors for each solvation model only depends weakly on the choice of conformational sampling as shown in Table 2. This is further confirmation that the energy errors are incorrect predictions mainly attributable to deficiencies of the energy functions, and as such they provide a means to analyze solvation models and suggest possible routes for improving them.

A more direct test of the potential energy functions used in loop prediction is to look at the relative percentage of energy errors rather than the relative percentage of correct predictions discussed previously which includes the effects of sampling errors. For the 9-residue loops in the absence of crystal symmetry, the largest percentage of energy errors (33.3%) was obtained for the distance-dependent dielectric. For the other implicit solvent models we tested in the absence of crystal symmetry, the percentage of energy errors decreases with, in order, SGB/NP (19.3%), AGB-γ (10.5%), AGBNP (7.0%), and AGBNP+ (3.5%).

The distance-dependent solvation model is clearly the worst in terms of accuracy, with nearly two-thirds of the incorrect predictions with extended sampling caused by energy errors (Table 2). Distance-dependent solvation models lack hydration free energy terms which provide the driving force toward solvent exposure of polar groups and vice versa the burial of hydrophobic groups. We have observed that a major structural problem with distance-dependent dielectric predictions is the occurrence of non-native salt bridges. Indeed after rescoring the distance-dependent dielectric predictions with AGBNP+, all are found to have energies greater than the native conformation due to the fact that Coulomb interaction energies of non-native ion pairs are countered by unfavorable electrostatic and nonpolar desolvation self-energy terms.

We observe about half as many energy errors with the SGB/NP solvation model as with the distance-dependent dielectric. However the occurrence of energy errors remains high; about half of the 21 incorrect predictions of 9-residue loops with SGB/NP in solution with extended sampling are attributed to the energy function. The reduction in the number of energy errors (11 to eight) with the inclusion of crystal symmetry can in principle be rationalized by the stabilization of the experimental structure due to crystal contacts not considered when evaluating the energy in solution, but we found very few examples (see below). In general the influence of the crystal environment appears to be secondary at this resolution in light of the fact that the occurrence of energy errors is significantly more pronounced with SGB/NP with crystal symmetry than with AGB-based solvation models without crystal symmetry (see Table 2). The reduction of SGB/NP energy errors with crystal symmetry is mainly due to crystal packing steric interactions preventing the formation of non-native low-energy conformations that occur in the absence of crystal contacts. Some examples illustrating the influence of the crystal environment on the loop conformation are discussed below.

Most SGB/NP predictions classified as energy errors were found to have electrostatic interaction energies significantly more negative than native conformations (results not shown), suggesting that SGB/NP overestimates the occurrence of salt bridges and intramolecular hydrogen bonds. When SGB/NP predictions are rescored with AGBNP+, all but two of the SGB/NP's energy errors are removed. Zhu et al.[53,65] recently obtained results indicating that the occurrence of energy errors with SGB/NP can be further reduced by including an empirical hydrophobic potential and a variable dielectric

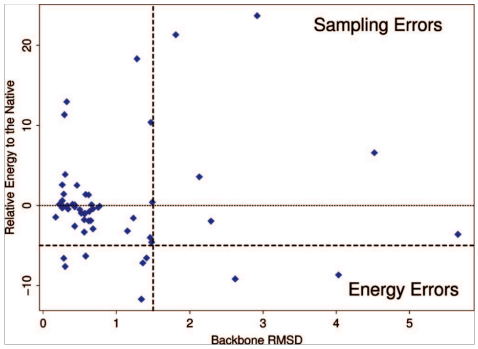

Figure 3. The results of the OPLS-AA/AGBNP+ loop predictions on the 57 9-residue loops in Table 1. The energies (in kcal/mol) relative to the native are plotted with respect to the backbone rmsd (in Å) to the native. The vertical dashed line is the rmsd cutoff, 1.5 Å. The bold, horizontal dotted-dashed line is the energy cutoff, −5 kcal/mol. Cases corresponding to the points to the left of the rmsd cutoff line are successful predictions, those in the top-right quadrant are sampling errors, and those in the bottom-right quadrant are energy errors.

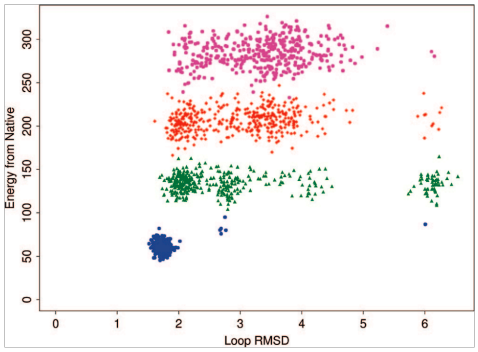

Figure 4. Energy gaps relative to the optimized native conformation (in kcal/mol) versus the rmsd (in Å) relative to the X-ray crystal conformation for the T-REMD prediction calculation of the 1fus (31−39) loop. The conformationas from the ensembles at 270 K, 400 K, 595 K, and 800 K are shown in blue, green, red, and magenta, respectively. Energies are in kcal/mol and rmsd is in Å.

model designed to favor conformations with packed hydrophobic cores and to disfavor the occurrence of salt bridges.

The AGBNP+ implicit solvent model with OPLS-AA yields only two energy errors for the 57 9-residue loops, the fewest among the solvation models tested (Table 2). The distribution of AGBNP+ results for 9-residue loops are plotted in Figure 3, where the energy errors are shown in the lower right of the plot. Only two of the 35 13-residue loop predictions with AGBNP+ are classified as energy errors. By analyzing the energy errors obtained with the various AGB-based models we are able to establish which features of the model aid in loop prediction. The number of energy errors for the 9-residue loops decreases consistently from six with the AGB-γ model, which is based on the simple surface area-only nonpolar model, to four with the AGBNP model,[44] which implements a nonpolar model that takes into account dispersive solute−solvent van der Waals interactions, to only two with the AGBNP+ model, which additionally adopts a parametrization designed to reduce the occurrence of salt bridges (see Methods).

These results indicate that the AGBNP model performs well for loop prediction applications regardless of the specific parametrization. Fine-tuning of the nonpolar model and salt bridge correction can yield, nevertheless, additional improvements. Two of the six energy errors with AGB-γ are removed when considering the AGBNP model, and, of the remaining four energy errors, two are removed when adopting ion pairing corrections in AGBNP+. One of these is the 1ivd(244−252) AGBNP prediction, which has an energy of −12.10 kcal/mol and an rmsd of 1.91 Å relative to the native. This incorrect prediction is stabilized by electrostatic interactions between Asp251 and Arg253. This interaction is absent in the 1.36 Å rmsd predicted conformation with AGBNP+, consistent with the fact that the energy of the incorrect prediction is raised above that of the correct prediction when rescored with AGBNP+. Similarly, the AGBNP incorrect prediction for 1sgp(109−117) is stabilized by a non-native ion-pair between residue Lys115 on the loop and the C-terminal carboxyl group of residue 242 which is avoided when using AGBNP+.

With AGBNP+ only two of the 13-residue loop predictions are classified as energy errors, moreover, as discussed below, the native conformations of these two loops are likely affected by intermolecular interactions present in the crystal that were not taken into account in the present calculations. In comparison, 13 of the 35 loops in this set were found to produce energy errors with the OPLS-AA/SGB/NP potential, and six of the loops are energy errors with the OPLS-AA/SGB/NP potential augmented by a hydrophobic contact correction term,[53] even though these calculations took into account crystallographic intermolecular interactions. The OPLS-AA/AGBNP+ potential function is in general able to identify the native conformation without the additional aid of knowledge-based empirical corrections, suggesting that the AGBNP solvation model captures the appropriate balance between polar and hydrophobic solvation and intramolecular interactions.

The small number of energy errors with the OPLS-AA/AGBNP+ force field are generally not very informative in terms of how to modify the potential in order to avoid them. The energy errors correspond to the 1xif(59−67) and 3pte(215−223) 9-residue loops and the 1hnj(A:191−203) and 1jp4(A:153−165) 13-residue loops. In all of these cases the native conformation is influenced by crystal contacts. Although we modeled 1xif as a monomer as did Fiser et al.[24] and Jacobson et al.,[26] the asymmetric unit of 1xif is a tetramer. However our attempt to model 1xif as a tetramer still resulted in an energy error possibly due to a native salt bridge not correctly modeled by AGBNP+. The native conformations of 3pte(215−223), 1hnj(A:191−203), and 1jp4(A:153−165) are clearly influenced by crystal packing

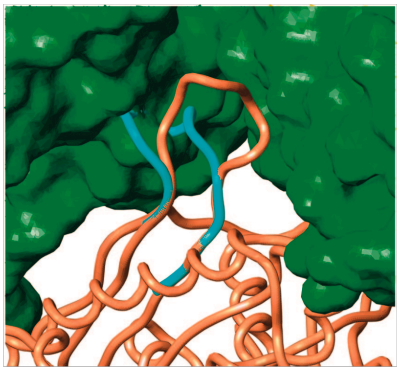

Figure 5. The X-ray (gold) and predicted (blue) conformations of the 13-residue loop 1jp4 (A:153−165). The surfaces of the crystallographically symmetric protein molecules are shown in green.

forces. As for example shown in Figure 5 for 1jp4, these loops extend away from the body of the protein, assuming a conformation unlikely to occur in solution. These loops make however extensive contacts with surrounding protein molecules in the crystal. The AGBNP+ predicted conformations without crystal symmetry instead pack closely against the protein body in a way which would not occur in the crystal due to steric repulsion. Moreover in the case of 1hnj and 1jp4, PLOP rejects backbone conformations that stray more than a certain distance from the protein body and prevents the evaluation of conformations near the native conformations. It should also be noted that, whereas we modeled only the monomer, the biological unit of 1hnj is a dimer and the loop in question (A191-A203) of one of the monomers makes contact with the same loop in the other monomer.

Apart from these cases, it appears that, within the resolution threshold we considered, the loop conformations predicted without using crystal symmetry are very close to the conformations seen in the crystal environment. This suggests that instead of crystal packing influencing loop conformations, in most cases it is the conformational propensity of the loop in solution which determines the packing arrangement in the crystal. This observation rationalizes the use of X-ray crystallographically determined structures as training sets in the development of homology modeling techniques for modeling protein loops in the solution environment.

4.3. Sampling Efficiency. Although they are indirectly influenced by properties of the energy function, such as its roughness and the level of degeneracy of native and non-native conformations, incorrect predictions classified as sampling errors primarily reflect limitations of the loop prediction algorithm. These are cases in which an incorrect prediction was made even though the energy of the native conformation is lower than that of the predicted conformation. It is important to reduce as much as possible the occurrence of sampling errors in order to decrease the overall number of mispredictions.

With the standard loop sampling procedure (Table 2) sampling errors generally represent a large fraction of incorrect predictions. This is in contrast to our results with the inclusion of crystal symmetry with the SGB/NP model in which only one-third of the incorrect predictions are classified as sampling errors. We conclude therefore that, although the parameters of the standard sampling algorithm (the value of the *ofac* parameter, the number of clusters, and the number of conformations that are passed from one stage of refinement to the next) work well when including crystallographically symmetry-related molecules,[26,53] the performance using standard sampling is significantly degraded when preforming loop prediction in the absence of the crystal environment. Evidently, the larger conformational space available to the loops in the absence of the crystal environment requires more extensive conformational search strategies. This has serious implications for loop prediction calculations as part of homology modeling projects which are typically carried out in the solution environment. Including the crystal environment is required to achieve high accuracy with the current sampling schemes. But in the majority of homology modeling applications, only the sequence and a related template protein is known. In most cases when the crystal parameters are known, so is the structure of the protein.

Sampling errors result from the sampling algorithm failing to produce near-native conformations of low enough energy or from failing to consider near-native conformations altogether. We refer to the first as a *local* sampling error and the latter as to a *global* sampling error. Global sampling errors typically occur when at the initial prediction stage the loop build-up procedure cannot find, within the resolution of the backbone and side chain rotamer library and the value of the *ofac* threshold parameter, any conformation free of clashes in the neighborhood of the native conformation. We also found that several of the global sampling errors with 9-residue loops are due to an insufficient preset number of clusters (36 for 9-residue loops), causing near-native conformations to sometimes be included in largely non-native conformational clusters. Local sampling errors are cases in which a near-native conformation produced by the initial prediction stage is abandoned prematurely and is not carried over to the subsequent refinement stages, which are responsible for adjusting the structure to lower the energy to a value closer to that of the native conformation. We found that the majority of 13-residue mispredictions are caused by local sampling errors.

Based on these observations we have modified the standard loop sampling procedure for 9-residue loops by decreasing one of the values of *ofac* tried at the initial prediction stage (from 0.75 to 0.5) and doubling the number of clusters (from 36 to 72) employed in the initial prediction stage. The standard loop sampling procedure for 13-residue loops was modified by increasing the number of candidate conformations carried over from one stage of refinement to the next (see Methods). These extended sampling schemes were then evaluated by applying them to the loops that resulted in sampling errors with the standard loop procedure. As Table 2 shows, the number of sampling errors was substantially

reduced for both the 13-residue and the 9-residue loops by using the extended sampling scheme. Interestingly, none of the sampling errors obtained with SGB/NP including crystal symmetry using the standard sampling scheme improved with the extended sampling scheme, confirming the results of earlier studies[26,53] that concluded that the standard sampling procedures were sufficient for loop predictions in the crystal environment.

4.4. Loop Prediction with Replica Exchange Molecular Dynamics. To better understand the origin of the observed sampling errors we investigated with T-REMD the six 9-residue loops that resulted in global sampling errors with the standard loop sampling procedure. As has been demonstrated,[26,53] the conformational search algorithms based on PLOP perform well for predicting the conformation of protein loops of up to 13 residues in length; however, because of the exponential explosion in the number of possible loop configurations that need to be examined, the application of this method to longer loops and situations which involve several interacting loops as well as simultaneous refinement of the protein region surrounding the loops is problematic. In contrast, importance sampling schemes concentrate sampling in the most thermodynamically relevant regions of the conformational space and scale linearly with the increase of the number of degrees of freedom.

The all-atom potential energy landscapes of proteins are rugged, containing many local minima separated from each other by high barriers. Because of this there are long dwell times in local minima which slows sampling rates making application of conventional room temperature MC or MD methods impractical for loop structure determination. New simulation strategies, called collectively generalized ensemble methods,[66] have been developed which overcome this sampling bottleneck. One of the most popular methods in this class is the temperature Replica Exchange Method (REM),[66,67] which can be paired with a constant temperature molecular dynamics engine (T-REMD).[55,56,68–70] The REM technique has been used to improve sampling of rough energy landscapes. The REM methodology has been used to predict the hypervariable regions of a llama VHH antibody domain[71] and has shown promise in other protein structure determination applications.[72–74]

Prior to applying the T-REMD procedure to the group of protein loops classified as sampling errors by the standard loop prediction routine, we tested the T-REMD protocol on a less challenging set of five 9-residue loops for which the PLOP conformational search scheme was able to locate near native conformations. The T-REMD approach produced matching results within reasonable simulation times, indicating that the T-REMD protocol can also easily provide good predictions in these cases. However, as the results summarized in Table 3 show, the more challenging cases of conformational sampling, although improved over the PLOP predictions, remain problematic. The T-REMD scheme was able to substantially improve within the allocated simulation time half of the PLOP sampling errors, resulting in much higher quality structures. The rmsds of the predictions for the 1onc, 1fus, and 1wer, improved from the range between 4 Å to 7.5 Å to ~2 Å or less. Only one case, however (1wer), resulted in a correct prediction based on the 1.5 Å rmsd threshold.

The T-REMD trajectory for the 1fus (31−39) loop is illustrated in Figure 4, where the energies of conformations sampled in the last 5 ns of simulation at various temperatures are plotted. The patchy pattern of the lowest temperature ensemble of loop configurations signifies the presence of high energy barriers which separate loop configurations into different conformational states. The absence of a direct path between these structurally distinct macrostates clearly shows that efficient sampling of the conformational space would not be possible with standard molecular dynamics conducted at room temperature. Transitions between the macrostates are accomplished by acquiring enough thermal energy (moving up the temperature ladder) to surmount the separating barrier. Afterward, there is a subsequent gradual annealing of the structure and temperature leading to the native conformation at low temperature. The numbers of transitions between macrostates during 5 ns is small.

5. Conclusion

We have conducted loop conformation prediction tests on challenging benchmark sets consisting of 9- and 13-residue loops using the conformational search schemes built into PLOP to investigate the accuracy of the AGBNP implicit solvation model in conjuction with the OPLS-AA intramolecular force field. For a set of 57 9-residue loops investigated previously[24–26] we accurately predicted 88% of the loops using the OPLS-AA/AGBNP+ potential. This is a substantial improvement over the use of a distance-dependent dielectric model (63%) or SGB/NP, with (77%) or without (67%) the inclusion of crystal symmetry, as the implicit solvent model. A more substantial difference between implicit solvent models is apparent when examining the relative percentage of energy errors. AGBNP+ has the lowest percentage of energy errors at 3.5%, which is less than one-fifth as many as for SGB/NP (19.3%) and one-ninth as many as for distance-dependent dielectric (33.3%).

The fact that we have obtained high accuracy without crystal symmetry when using AGBNP+ suggests that the presence of crystal symmetry in the model is not crucial for reproducing the loop structures which have been experimentally determined via X-ray crystallography. In general, although the side chain positions have been reported to be strongly influenced by the neighboring crystallographically symmetry-related molecules,[27] the backbone conformation does not appear to be as strongly influenced by crystal packing interactions at the resolution of the current study. A recent comparison between structures determined by X-ray crystallography and NMR of identical proteins showed little correlation between structural differences and crystal contacts.[64] We found, however, the conformation sampling schemes previously developed for loop predictions in the crystal environment needed to be extended in order to avoid sampling errors when crystal symmetry is not included in the model. We recommend the use of these updated extended sampling protocols for homology modeling applications in the solution environment.

We expect importance sampling conformational search methods such as T-REMD to become an important complement to traditional discrete conformational search methods in cases when the number of degrees of freedom is large such as interacting loops, imperfect frameworks for loop prediction, etc. We note that development of better implementations of REM ideas which will offer faster sampling in the context of structure prediction of protein loops is the subject of intensive ongoing research. This will go beyond simple temperature exchanges in REM and will involve modifying the system Hamiltonians and swapping replicas with different energy potentials, constructed to effectively increase the range of conformational motion.[71] Another avenue of improvement is to consider more rational ways of selecting pairs of replicas for exchanges of temperatures or Hamiltonian parameters,[75] with the goal being to examine how sampling can be enhanced through maximizing mixing among replicas. Such a multidimensional replica exchange procedure appears to be promising for exploring the conformational space of protein loops.

It should be noted that the success rates we obtained likely overestimate the success rate obtainable in actual homology modeling applications because these tests were performed in the idealized case in which the frame of the protein surrounding the loop is known. Successful prediction in this idealized situation is a necessary but not sufficient requirement for the ability to predict the correct nativelike loop conformation with partial knowledge of the protein framework. We have begun to investigate cases in which the conformations of the protein side chains surrounding the loop are predicted at the same time as the loop conformation. We find that the successful prediction rate for these cases is significantly reduced relative to the tests reported here with the conformations of the side chains of the protein frame fixed in their native conformations. Clearly more work is still needed to develop fast and accurate loop prediction protocols for "real life" homology modeling applications.

Acknowledgment. This project has been supported in part by the National Institutes of Health Grants (GM-30580 to R.M.L. and GM-52018 to R.A.F.). We thank Matt Jacobson for providing the PLOP program and Kai Liu for providing the script on which we based our 13-residue prediction protocol. We also thank Jennifer Knight for assistance with implementing PLOP and predicting loops with imperfect frames.

References

(1) Ginalski, K. *Curr. Opin. Struct. Biol.* **2006**, *16*, 172–177.

(2) Kryshtafovych, A.; Venclovas, C.; Fidelis, K.; Moult, J. *Proteins* **2005**, *61*, 225–236.

(3) Shiffer, C.; Hermans, J. *Methods Enzymol.* **2003**, *374*, 412–461.

(4) Xia, B.; Tsui, V.; Case, D.; Dyson, H.; Wright, P. *J. Biomol. NMR* **2002**, *22*, 317–331.

(5) Skolnick, J. *Curr. Opin. Struct. Biol.* **2006**, *16*, 166–171.

(6) Lazaridis, T.; Karplus, M. *Curr. Opin. Struct. Biol.* **2000**, *10*, 139–145.

(7) Rhee, Y. M.; Pande, V. S. *Biophys. J.* **2003**, *84*, 775–786.

(8) Huang, E. S.; Subbiah, S.; Tsai, J.; Levitt, M. *J. Mol. Biol.* **1996**, *257*, 716–725.

(9) Park, B.; Levitt, M. *J. Mol. Biol.* **1996**, *258*, 367–392.

(10) Park, B. H.; Huang, E. S.; Levitt, M. *J. Mol. Biol.* **1997**, *266*, 831–846.

(11) Samudrala, R.; Levitt, M. *Protein Sci.* **2000**, *9*, 1399–1401.

(12) Simons, K. T.; Bonneau, R.; Ruczinski, I.; Baker, D. *Proteins: Struct., Funct., Genet.* **1999**, *S3*, 171–176.

(13) Lazaridis, T.; Karplus, M. *J. Mol. Biol.* **1999**, *288*, 477–487.

(14) Petrey, D.; Honig, B. *Protein Sci.* **2000**, *9*, 2181–2191.

(15) Bursulaya, B. D.; Brooks III, C. L. *J. Phys. Chem. B* **2000**, *104*, 12378–12383.

(16) Dominy, B. N.; Brooks, C. L. *J. Comput. Chem.* **2002**, *23*, 147–160.

(17) Liu, Y.; Beveridge, D. L. *Proteins: Struct. Funct. Genet.* **2002**, *46*, 128–146.

(18) Feig, M.; Brooks, C. L., III *Proteins: Struct. Funct. Genet.* **2002**, *49*, 232–245.

(19) Felts, A. K.; Gallicchio, E.; Wallqvist, A.; Levy, R. M. *Proteins: Struct., Funct., Genet.* **2002**, *48*, 404–422.

(20) Zhang, Y.; Kolinski, A.; Skolnick, J. *Biophys. J.* **2003**, *85*, 1145–1164.

(21) Tsai, J.; Bonneau, R.; Morozov, A.; Kuhlman, B.; Rohl, C.; Baker, D. *Proteins* **2003**, *53*, 76–87.

(22) Wang, K.; Fain, B.; Levitt, M.; Samudrala, R. *BMC Struct. Biol.* **2004**, *4*, 8.

(23) Qiu, J.; Elber, E. *Proteins* **2005**, *61*, 44–55.

(24) Fiser, A.; Do, R. K. G.; Sali, A. *Protein Sci.* **2000**, *9*, 1753–1773.

(25) Xiang, Z. X.; Soto, C. S.; Honig, B. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7432–7437.

(26) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J. F.; Honig, B.; Shaw, D. E.; Friesner, R. A. *Proteins: Struct., Funct., Bioinform.* **2004**, *55*, 351–367.

(27) Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. *J. Mol. Biol.* **2002**, *320*, 597–608.

(28) Sherman, W.; Day, T.; Jacobson, M.; Friesner, R.; Farid, R. *J. Med. Chem.* **2006**, *49*, 534–553.

(29) Levy, R. M.; Gallicchio, E. *Annu. Rev. Phys. Chem.* **1998**, *49*, 531–67.

(30) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

(31) Dominy, B. N.; Brooks III, C. L. *J. Phys. Chem. B* **1999**, *103*, 3765–3773.

(32) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.

(33) Cortis, C. M.; Friesner, R. A. *J. Comput. Chem.* **1997**, *18*, 1591–1608.

(34) Rocchia, W.; Sridharan, S.; Nicholls, A.; Alexov, E.; Chiabrera, A.; Honig, B. *J. Comput. Chem.* **2002**, *23*, 128–137.

(35) Banks, J.; et al., *J. Comput. Chem.* **2005**, *26*, 1752–1780.

(36) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578–1599.

(37) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, C. W. *J. Phys. Chem. A* **1997**, *101*, 3005–3014.

(38) Tsui, V.; Case, D. A. *Biopolymers* **2000**, *56*, 275–291.

(39) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983–10990.

(40) Lee, M. S.; Feig, M., Jr.; Brooks, C. L. *J. Comput. Chem.* **2003**, *24*, 1348–1356.

(41) Feig, M.; Onufriev, A.; Lee, M.; Im, W.; Case, D.; Brooks, C., III *J. Comput. Chem.* **2004**, *25*, 265–284.

(42) Zhang, L.; Gallicchio, E.; Friesner, R.; Levy, R. M. *J. Comput. Chem.* **2001**, *22*, 591–607.

(43) Mongan, J.; Simmerling, C.; McCammon, J.; Case, D.; Onufriev, A. *J. Chem. Theory Comput.* **2007**, *3*, 156–169.

(44) Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479–499.

(45) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(46) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

(47) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824–19839.

(48) Gallicchio, E.; Kubo, M. M.; Levy, R. M. *J. Phys. Chem. B* **2000**, *104*, 6271–6285.

(49) Levy, R. M.; Zhang, L. Y.; Gallicchio, E. amd Felts, A. K. *J. Am. Chem. Soc.* **2003**, *25*, 9523–9530.

(50) Su, Y.; Gallicchio, E. *Biophys. Chem.* **2004**, *109*, 251–260.

(51) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983–10990.

(52) Gallicchio, E.; Zhang, L.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517–529.

(53) Zhu, K.; Pincus, D. L.; Zhao, S.; Friesner, R. A. *Proteins: Struct., Funct., Bioinform.* **2006**, *65*, 438–452.

(54) Hartigan, J. A.; Wong, M. A. *Appl. Stat.* **1979**, *28*, 100–108.

(55) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(56) Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. *Proteins: Struct., Funct., Bioinform.* **2004**, *56*, 310–321.

(57) Banks, J. L. *J. Comput. Chem.* **2005**, *26*, 1752–1780.

(58) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem. A* **1997**, *101*, 3005–3014.

(59) Wagoner, J. A.; Baker, N. A. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8331–8336.

(60) Geney, R.; Layten, M.; Gomperts, R.; Hornak, V.; Simmerling, C. *J. Chem. Theory. Comput.* **2006**, *2*, 115–127.

(61) Wang, G.; Dunbrack, R. *Bioinformatics* **2003**, *19*, 1589–1591.

(62) de Bakker, P. I. W.; DePristo, M. A.; Burke, D. F.; Blundell, T. L. *Proteins: Struct., Funct., Bioinform.* **2003**, *51*, 21–40.

(63) DePristo, M. A.; de Bakker, P. I. W.; Lovell, S. C.; Blundell, T. L. *Proteins: Struct., Funct., Bioinform.* **2003**, *51*, 44–55.

(64) Andrec, M; Snyder, D. A.; Zhou, Z.; Young, J. T. M. G.; Levy, R. M. *Proteins: Struct., Funct., Bioinform.* **2007**, *69*, 449–465.

(65) Zhu, K.; Shirts, M. R.; Friesner, R. A. *J. Chem. Theory Comput.* **2007**, *3*, 2108–2119.

(66) Sugita, Y.; Okamoto, Y. Free-energy calculations in protein folding by generalized-ensemble algorithms. In *Lecture Notes in Computational Science and Engineering*; Schlick, T.; Gan, H. H., Eds.; Springer-Verlag: Berlin, 2002.

(67) Nymeyer, H.; Gnanakaran, S.; García, A. E. *Methods Enzymol.* **2003**, *383*, 119–149.

(68) García, A. E.; Sanbonmatsu, K. Y. *Proteins: Struct., Funct., Genet.* **2001**, *42*, 345–354.

(69) Zhou, R.; Berne, B. J.; Germain, R. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 14931–14936.

(70) Cecchini, M.; Rao, F.; Seeber, M.; Caflisch, A. *J. Chem. Phys.* **2004**, *121*, 10748–10756.

(71) Fenwick, M. K.; Escobedo, F. A. *Biopolymers* **2003**, *68*, 160–177.

(72) Chen, J.; Im, W.; Brooks III, C. L. *J. Am. Chem. Soc.* **2004**, *126*, 16038–16047.

(73) Habeck, M.; Nilges, M.; Rieping, W. *Phys. Rev. Lett.* **2005**, *94*, 018105.

(74) Nanias, M.; Chinchio, M.; Oldziej, S.; Czaplewski, C.; Scheraga, H. A. *J. Comput. Chem.* **2005**, *26*, 1472–1486.

(75) Calvo, F. *J. Chem. Phys.* **2005**, *123*, 124106.

CT800051K

References

Fiser, A.; Kinh, R.; Do, G.; Šali, A. Modeling of loops in protein structures. *Protein Sci.* **2000**, *9*, 1753-1773.

Gallicchio, E.; Levy, R.M. AGBNP: an analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *J. Comput. Chem.* **2004**, *25*, 479-499.

Ghosh, A.; Rapp, C.S.; Friesner, R.A. Generalized Born model based on a surface integral formulation. *J. Phys. Chem. B* **1998**, *102*, 10983-10990.

Jacobson, M.P.; Friesner, R.A.; Xiang, Z.; Honig, B. On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol*. **2002**, *320*, 597-608.

Jacobson, M.P.; Pincus, D.L.; Rapp, C.S.; Day, T.J.; Honig, B.; Shaw, D.E.; Friesner, R.A. A hierarchical approach to all-atom protein loop prediction. *Proteins* **2004**, *55*, 351-67.

Jorgensen, W.L.; Maxwell, D.S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225-11236.

Kaminski, G.A.; Friesner, R.A.; Tirado-Rives, J.; Jorgensen, W.L. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **2001**, *105*, 6474-6487.

Krivov, G.G.; Shapovalov, M.V.; Dunbrack, R.L. Jr. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* **2009**, *77*, 778-95.

Lazaridis, T.; Karplus, M. Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* **2000**. *10*, 139-145.

Rhee, Y.M.; Pande, V.S. Multiplexed-Replica Exchange Molecular Dynamics Method for Protein Folding Simulation. *Biophys. J.* **2003**, *84*, 775-786.

Samudrala, R.; Levitt, M. Decoys 'R' Us: A database of incorrect conformations to improve protein structure prediction. *Protein Sci.* **2000**, *9*, 1399-1401.

Sherman, W.; Day, T.; Jacobson, M.P.; Friesner, R.A.; Farid R. Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem*. **2006**, *49*, 534-53.

Skolnick, J. In quest of an empirical potential for protein structure prediction. *Curr. Opin. Struct. Biol.* **2006**, *16*, 166-171.

Soto, C.S.; Fasnacht, M.; Zhu, J.; Forrest, L.; Honig, B. Loop modeling: Sampling, filtering, and scoring. *Proteins* **2008**, *70*, 834-43.

Zhang, L.Y.; Gallicchio, E.; Friesner, R.A.; Levy, R.M. Solvent models for protein-ligand binding: Comparison of implicit solvent Poisson and surface generalized Born models with explicit solvent simulations. *J. Comput. Chem.* **2001**, *22*, 591-607.

Chapter 3

Mini-proteins and AGBNP

3.1 Introduction

As described in Chapter 2, the accurate modeling of hydration and the ability of a scoring function to favorably score the native structure are important in computational modeling problems, including the study of protein folding, conformational equilibria, and binding. The results of earlier studies (Felts et al., 2004) and the study described in Chapter 2 (Felts et al., 2008) suggest that the OPLS-AA/AGBNP (Jorgenson et al., 1996; Kaminski et al., 2001; Gallicchio and Levy, 2004) model reproduces reasonably the backbone secondary structure features of proteins and peptides. This chapter, therefore, focuses primarily on the prediction of protein side chains.

Explicit solvent models are generally thought to provide the most complete and detailed description of hydration (Levy and Gallicchio, 1998) but are computationally demanding. Implicit solvent models, which are less computationally demanding, have been shown to be alternatives to explicit solvation. (Feig and Brooks, 2004; Roux and Simonson, 1999; Felts et al., 2004; Gallicchio et al., 2002) Here we test several incarnations of the Analytical Generalized Born plus Non-Polar (AGBNP) solvent model (Gallicchio and Levy, 2004) for its ability to predict the structure of several small proteins.

3.2 Hydrogen Bond Screening and AGBNP

Previous work with protein decoys (Felts et al., 2002), β-hairpin and α-helical peptides (Felts et al., 2004), and side chain and loop modeling (Jacobson et al., 2002; Jacobson et al., 2004; Yu et al., 2006) indicated that salt bridge formation is overestimated by generalized Born solvation models. A "quick fix" through use of dielectric screening was designed to reduce the occurrence of salt bridges between oppositely charged residues as well as to decrease repulsion between like-charged residues (Felts et al., 2004). The standard implementation of the GB model estimates the electrostatic component of the hydration free energy as a sum of the self energies and pair energies:

$$\Delta G_{elec} \cong \Delta G_{GB} = \Delta G_{self} + \Delta G_{pair} \tag{3.1}$$

where

$$\Delta G_{self} = -\frac{1}{2}\left(\frac{1}{\varepsilon_{in}} - \frac{1}{\varepsilon_{w}}\right)\sum_{i}\frac{q_i^2}{B_i} \tag{3.2}$$

and

$$\Delta G_{pair} = -\left(\frac{1}{\varepsilon_{in}} - \frac{1}{\varepsilon_{w}}\right)\sum_{i<j}\frac{q_i q_j}{f_{ij}} \tag{3.3}$$

In equations (3.2) and (3.3), ε_{in} is the dielectric constant of the interior of the solute (generally set to 1), ε_w is the dielectric constant of the solvent (generally set to 80 for water), q_i is the partial charge on atom i, B_i is the Born radius of atom i and f_{ij}, the generalized Born pair function, is defined as

$$f_{ij} = \sqrt{r_{ij}^2 + B_{ij}^2 \exp(-r_{ij}^2 / 4B_{ij}^2)} \tag{3.4}$$

where r_{ij} is the distance between atoms i and j, $B_{ij} = \sqrt{B_i B_j}$ is the geometric average of the Born radii of atoms i and j. The modified pair function used in this work

$$f_{ij} = \sqrt{r_{ij}^2 + S_{ij}^2 B_{ij}^2 \exp(-r_{ij}^2 / 4B_{ij}^2)} \tag{3.5}$$

differs from the original (eq. 3.4) by the introduction of the S_{ij} parameter which is defined as the geometric average of the screening constants S_i and S_j that are assigned to atoms i and j: $S_{ij} = \sqrt{S_i S_j}$. The screening constants are nonnegative, dimensionless parameters. When both S_i and S_j are set to 1, the original GB pair interaction energy is recovered (Felts et al., 2004).

Original work with the modified pair function aimed to reduce salt bridges and thus assigned screening values of 0.5 to the oxygen atoms of the carboxylate groups of the glutamate and aspartate residues and to the nitrogen atoms of the ammonium and guanadinium groups of the lysine and arginine residues while the screening constants for all other atom types were set to 1 (Felts et al., 2004). This preliminary study aims to lessen the number of any hydrogen bonds and thus reduces the values to 0.5, 0.7, 0.8 or 0.9 on both oxygen and hydrogen atoms on either the backbone or side chain.

We looked at a set of eight mini-proteins or peptides between 25 and 30 residues in length that have been shown to form stable secondary structures in solution. The set of eight structures included: neurotoxin III with sequence RSCCPCYWGGCPWGQNCYPEGCSGPKV (PDB id 1ANS; Manoleras and Norton, 1994), hpTX2 toxin with sequence DDCGKLFSGCDTNADCCEGYVCRLWCKLDW (PDB id 1EMX; Bernard et al., 2009), full sequence design 1 with sequence QQYTAKIKGRTFRNEKELRDFIEKFKGR (PDB id 1FSD; Dahiyat and Mayo, 1997), delta-conotoxin TxVIA with sequence WCKQSGEMCNLLDQNCCDGYCIVLVCT

(PDB id 1FU3; Kohno et al., 2002), computationally designed peptide 1 with sequence KPYTARIKGRTFSNEKELRDFLETFTGR (PDB id 1PSV; Dahiyat et al., 1997), an HPV E6-inhibiting Trp-cage with sequence XALQELLGQWLKDGGPSSGRPPPSX (PDB id 1RIJ; Liu et al., 2004), *Viola hederacea* root cyclotide-1 with sequence CAESCVWIPCTVTALLGCSCSNKVCYNGIP (PDB id 1VB8; Trabi and Craik, 2004), and the K channel blocker OmTx2 from the venom of the scorpion *Opisthacanthus madagascariensis* with sequence DPCYEVCLQQHGNVKECEEACKHPVEY (PDB id 1WQD; Chagot et al., 2005). These structures are not fragments of larger proteins and

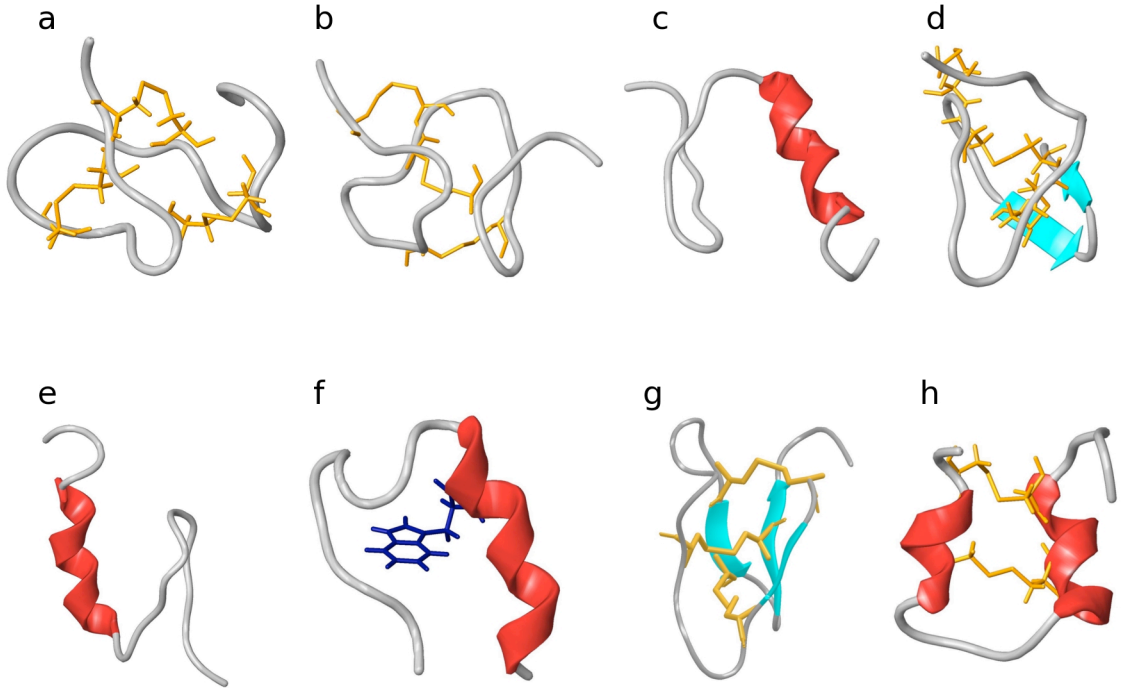

Figure 3.1. Graphical representations of eight mini-proteins. a: neurotoxin III (PDB id 1ANS); b: hpTX2 toxin (PDB id 1EMX); c: full sequence design 1 peptide (PDB id 1FSD); d: delta- conotoxin TxVIA (PDB id 1FU3); e: computationally designed peptide 1 (PDB id 1PSV); f: HPV E6-binding Trp-cage (PDB id 1RIJ); g: VHR1 cyclotide (PDB id 1VB8); h: potassium channel blocker OmTx2 (PDB id 1WQD). In each case the best representative structure as specified in the NMR model deposited in the PDB is shown. α-helices are shown in red, β-strands are shown as cyan arrows, loops and turns are shown in gray, Disulfide bonds formed in the unminimized native are shown in yellow. The tryptophan in the Trp-cage is also shown in dark blue for PDB id 1RIJ (f).

met the following requirements: no ligands, no non-standard residues, pH between 6.5 and 7.5, and inclusion of all heavy atom coordinates. They are a diverse set which include various secondary structures as shown in Figure 3.1.

Hybrid monte carlo (HMC) simulations were conducted for the complete set of eight mini-proteins up to 10ns (25ns for 1PSV) starting from the best representative structure from the NMR model deposited in the PDB. A time step of 2 fs was employed with 10 MD steps per HMC cycle. MD simulations were also conducted for a subset of four mini-proteins: 1EMX, 1FSD, 1PSV, 1RIJ for 10ns, set to 300 K with the Berendsen thermostat, and with a time step of 1fs. Simulations were performed with the OPLS-AA potential using the IMPACT program (Banks et al., 2005). Resulting structures were clustered using Cluster, a single-linkage hierarchical clustering algorithm (Shenkin and McDonald, 1994). Representatives from each cluster were then minimized using the OPLS-AA/AGBNP potential. 1ANS and 1PSV were further analyzed with minimizations using OPLS-AA/AGBNP and screening constants ranging from 0.5 to 1.0 for oxygen and hydrogen atoms on either the backbone or side chain. Explicit solvent (4ns with the TIP3P water model in a 45Å x 45Å x 45Å box) simulations were also employed for 1PSV using the IMPACT program. Each minimized representative structure was structurally analyzed for extent of molecular interactions including number of H-bonds, number of van der Waals contacts and solvent accessible surface area. H-bonds were detected using a minimum hydrogen-acceptor distance of 2.5 Å and a minimum donor angle of 120°. Van der Waals contacts were detected using a 1.3 cutoff for the ratio of the distance between atoms *i* and *j* to the sum of the vdW radii of atoms *i* and *j*. A comparison of the average values obtained using AGBNP-minimized

(predicted) structures versus the minimized native structures is shown in Table 3.1. For all eight mini-proteins, the predicted structures display much greater numbers of H-bonds and vdW contacts while the total SASA of the predicted structures is less than that of the native.

PDB id	*Av. H-bonds*	*Av. vdW Contacts*	*Av. SASA*
1ANS	21 / 7	2262 / 2023	2257 / 2344
1EMX	19 / 10	2580 / 2457	2541 / 2664
1FSD	22 / 18	3055 / 2904	2784 / 3012
1FU3	24 / 16	2455 / 2293	2149 / 2338
1PSV	22 / 9	2937 / 2668	2607 / 2865
1RIJ	21 / 13	2075 / 1957	1880 / 2093
1VB8	22 / 15	2488 / 2328	2256 / 2431
1WQD	19 / 13	2502 / 2250	2402 / 2604

Table 3.1. Comparison of OPLS-AA/AGBNP predicted structures and native structures. Average number of hydrogen bonds, average number of vdW contacts and average solvent accessible surface reported for predicted structures / native structure.

	Backbone					***Side chain***				
S_i	1.0	0.5	0.7	0.8	0.9	0.5	0.6	0.7	0.8	0.9
1ANS										
Pred. H-Bonds	20.5	8	9.5	14.5	15	13.5	16.5	15	17	19
Min. Native. H-bonds *Native H-bonds = 1*	6	3	3	3	3	6	6	3	3	3
Pred. BB H-bonds	10.5	2	3	6	2.5	7	6.25	5	8.5	8
Min. Native BB H-bonds *Native BB H-bonds = 1*	4	2	2	2	2	4	4	2	2	2
Pred. SB	0.5	2	0	0	0	0	0	0	0	0
Min. Native SB *Native SB = 0*	0	0	0	0	0	0	0	0	0	0
1PSV										
Pred. H-Bonds	26	16.7	16.7	19.3	21.5	18.6	23.3	20	21	21
Min. Native. H-bonds *Native H-bonds = 11*	11	8	7	7	9	12	12	10	10	12
Pred. BB H-bonds	14	4.7	7.7	9	10.5	9.6	8.3	9.7	10.5	11
Min. Native BB H-bonds *Native BB H-bonds = 11*	10	7	6	6	7	9	10	9	8	10
Pred. SB	2.5	3	1.7	0.5	2.5	1	1.6	0.7	1.75	1.3
Min. Native SB *Native SB = 0*	0	0	0	0	0	0	0	0	0	0

Table 3.2. Effect of differing screening constants S_i on the number of hydrogen bonds (H-bonds), backbone-to-backbone H-bonds (BB H-bonds) and salt bridges (SB). Average predicted values are compared with the minimized native. The native 1ANS before minimization displays 1 H-bond, of which 1

is a backbone to backbone H-bond and zero are salt bridges. The native 1PSV before minimization displays 11 H-bonds of which 11 are backbone-to-backbone H-bonds and zero are salt bridges.

Inclusion of the screening constants for 1ANS and 1PSV given in table 3.2 shows that addition of the screening constant does allow for a pronounced decrease in the number of H-bonds. However, comparison shows that whereas the native H-bonds are predominantly backbone-to-backbone, the predicted structures display a propensity for H-bond formation between side chain-side chain or side chain-backbone. In both cases, there is little effect on the number of salt bridges as salt bridges are difficult to form in the selected structures. Explicit solvent simulations on 1PSV are in agreement with the native conformations observed, with the majority of the H-bonds occurring in a backbone-to-backbone fashion.

3.3 Hydrogen Bonds in AGBNP2

The preliminary results from part 3.2 above pointed to the inability of the AGBNP solvent model to describe correct H-bond formation in the set of eight mini-proteins. The AGBNP2 model was thus presented as an evolution of the AGBNP model in which a new empirical component to model first solvation shell effects (such as H-bonding) is introduced. A subset of the eight mini-proteins above was utilized to test the new AGBNP2 model.

The procedures, results and discussion of this section are included in the following reprint of a paper published in *J. Chem. Theory Comput.* **2009**, *5*, 2544-2564.

2544 *J. Chem. Theory Comput.* **2009,** *5,* 2544–2564

JCTC Journal of Chemical Theory and Computation

The AGBNP2 Implicit Solvation Model

Emilio Gallicchio,* Kristina Paris, and Ronald M. Levy

Department of Chemistry and Chemical Biology and BioMaPS Institute for Quantitative Biology, Rutgers University, Piscataway, New Jersey 08854

Received May 11, 2009

Abstract: The AGBNP2 implicit solvent model, an evolution of the Analytical Generalized Born plus NonPolar (AGBNP) model we have previously reported, is presented with the aim of modeling hydration effects beyond those described by conventional continuum dielectric representations. A new empirical hydration free energy component based on a procedure to locate and score hydration sites on the solute surface is introduced to model first solvation shell effects, such as hydrogen bonding, which are poorly described by continuum dielectric models. This new component is added to the generalized Born and nonpolar AGBNP terms. Also newly introduced is an analytical Solvent Excluded Volume (SEV) model which improves the solute volume description by reducing the effect of spurious high dielectric interstitial spaces present in conventional van der Waals representations. The AGBNP2 model is parametrized and tested with respect to experimental hydration free energies of small molecules and the results of explicit solvent simulations. Modeling the granularity of water is one of the main design principles employed for the first shell solvation function and the SEV model, by requiring that water locations have a minimum available volume based on the size of a water molecule. It is shown that the new volumetric model produces Born radii and surface areas in good agreement with accurate numerical evaluations of these quantities. The results of molecular dynamics simulations of a series of miniproteins show that the new model produces conformational ensembles in substantially better agreement with reference explicit solvent ensembles than the original AGBNP model with respect to both structural and energetics measures.

1. Introduction

Water plays a fundamental role in virtually all biological processes. The accurate modeling of hydration thermodynamics is therefore essential for studying protein conformational equilibria, aggregation, and binding. Explicit solvent models arguably provide the most detailed and complete description of hydration phenomena.[1] They are, however, computationally demanding not only because of the large number of solvent atoms involved, but also because of the need to average over many solvent configurations to obtain meaningful thermodynamic data. Implicit solvent models,[2] which are based on the statistical mechanics concept of the solvent potential of mean force,[3] have been shown to be useful alternatives to explicit solvation for applications including protein folding and binding,[4] and small molecule hydration free energy prediction.[5]

Modern implicit solvent models[6,7] include distinct estimators for the nonpolar and electrostatic components of the hydration free energy. The nonpolar component corresponds to the free energy of hydration of the uncharged solute, while the electrostatic component corresponds to the free energy of turning on the solute partial charges. The latter is typically modeled treating the water solvent as a uniform high dielectric continuum.[8] Methods based on the numerical solution of the Poisson−Boltzmann (PB) equation[9] provide a virtually exact representation of the response of the solvent within the dielectric continuum approximation. Recent advances extending dielectric continuum approaches have focused on the development of Generalized Born (GB) models,[10] which have been shown to reproduce with good accuracy PB and explicit solvent[7,11] results at a fraction of

* Corresponding author e-mail emilio@biomaps.rutgers.edu.

10.1021/ct900234u CCC: $40.75
Published on Web 07/31/2009

the computational expense. The development of computationally efficient analytical and differentiable GB methods based on pairwise descreening schemes[6,12,13] has made possible the integration of GB models in molecular dynamics packages for biological simulations.[14−16]

The nonpolar hydration free energy component accounts for all nonelectrostatic solute−solvent interactions as well as hydrophobic interactions,[17] which are essential driving forces in biological processes such as protein folding[18−21] and binding.[22−25] Historically the nonpolar hydration free energy has been modeled by empirical surface area models[26] which are still widely employed.[10,27−35] Surface area models are useful as a first approximation; however, qualitative deficiencies have been observed.[29,36−41]

A few years ago we presented the Analytical Generalized Born plus NonPolar (AGBNP) implicit solvent model,[42] which introduced two key innovations with respect to both the electrostatic and nonpolar components. Unlike most implicit solvent models, the AGBNP nonpolar hydration free energy model includes distinct estimators for the solute−solvent van der Waals dispersion energy and cavity formation work components. The main advantages of a model based on the cavity/dispersion decomposition of the nonpolar solvation free energy stem from its ability to describe both medium-range solute−solvent dispersion interactions, which depend on solute composition, as well as effects dominated by short-range hydrophobic interactions, which can be modeled by an accessible surface area term.[40] A series of studies highlight the importance of the balance between hydrophobicity and dispersion interactions in regulating the structure of the hydration shell and the strength of interactions between macromolecules.[43−45] In AGBNP the work of cavity formation is described by a surface area dependent model,[37,46−48] while the dispersion estimator is based on the integral of van der Waals solute−solvent interactions over the solvent, modeled as a uniform continuum.[38] This form of the nonpolar estimator had been motivated by a series of earlier studies[5,37,49−52] and has since been shown by us[38,53−55] and others[39−41,56] to be qualitatively superior to models based only on the surface area in reproducing explicit solvent results as well as rationalizing structural and thermodynamic experimental observations.

The electrostatic solvation model in AGBNP is based on the pairwise descreening GB scheme[13] whereby the Born radius of each atom is obtained by summing an appropriate descreening function over its neighbors. The main distinction between the AGBNP GB model and conventional pairwise descreening implementations is that in AGBNP the volume scaling factors, which offset the overcounting of regions of space occupied by more than one atom, are computed from the geometry of the molecule rather than being introduced as geometry-independent parameters fit to either experiments or to numerical Poisson−Boltzmann results.[14,57−59] The reduction of the number of parameters achieved with this strategy improves the transferability of the model to unusual functional groups often found in ligand molecules, which would otherwise require extensive parametrization.[60]

Given its characteristics, the AGBNP model has been mainly targeted to applications involving molecular dynamics canonical conformational sampling, and to the study of protein−ligand complexes. Since its inception the model has been employed in the investigation of a wide variety of biomolecular problems ranging from peptide conformational propensity prediction and folding,[54,61−63] ensemble-based interpretation of NMR experiments,[64,65] protein loop homology modeling,[55] ligand-induced conformational changes in proteins,[66,67] conformational equilibria of protein−ligand complexes,[68,69] protein−ligand binding affinity prediction,[70] and structure-based vaccine design.[71] The AGBNP model has been reimplemented and adopted with minor modifications by other investigators.[72,73] The main elements of the AGBNP nonpolar and electrostatic models have been independently validated,[39,40,74,75] and have been incorporated in recently proposed hydration free energy models.[76,77]

In this work we present a new implicit solvent model named AGBNP2 which builds upon the original AGBNP implementation (hereafter referred to as AGBNP1) and improves it with respect to the description of the solute volume and the treatment of short-range solute−water electrostatic interactions.

Continuum dielectric models assume that the solvent can be described by a linear and uniform dielectric medium.[78] This assumption is generally valid at the macroscopic level; however, at the molecular level water exhibits significant deviations from this behavior.[1] Nonlinear dielectric response, the nonuniform distribution of water molecules, charge asymmetry, and electrostriction effects[79] are all phenomena originating from the finite size and internal structure of water molecules as well as their specific interactions which are not taken into account by continuum dielectric models. Some of these effects are qualitatively captured by standard classical fixed-charge explicit water models; however others, such as polarization and hydrogen bonding interactions, can be fully modeled only by adopting more complex physical models.[80] GB models make further simplifications in addition to the dielectric continuum approximation, thereby compounding the challenge of achieving with GB-based implicit solvent models the level of realism required to reliably model phenomena, such as protein folding and binding, characterized by relatively small free energy changes.

In the face of these challenges a reasonable approach is to adopt an empirical hydration free energy model motivated by physical arguments[81] parametrized with respect to experimental hydration free energy data.[20] Models of this kind typically score conformations on the basis of the degree of solvent exposure of solute atoms. Historically[82] the solvent accessible surface area of the solute has been used for this purpose, while modern implementations suitable for conformational sampling applications often employ computationally convenient volumetric measures.[83,84] In this work we take this general approach but we retain the GB model component which we believe is a useful baseline to describe the long-range influence of the water medium. The empirical parametrized component of the model then takes the form of an empirical first solvation shell correction function designed so as to absorb hydration effects not accurately described by the GB model. Specifically, as described below, we employ a short-range analytical hydrogen bonding correction

function based on the degree of water occupancy (taking into account the finite size of water molecules) of appropriately chosen hydration sites on the solute surface. The aim of this model is to effectively introduce some explicit solvation features without actually adding water molecules to the system as for example done in hybrid explicit/implicit models.[85,86]

In this work we also improve the description of the solute volume, which in AGBNP1 is modeled by means of atomic spheres of radius equal to the atomic van der Waals radius. The deficiencies of the van der Waals solute volume model have been recognized.[87,88] They stem from the presence of high dielectric interstitial spaces in the solute interior which are too small to contain discrete water molecules. These spurious high dielectric spaces contribute to the hydration of buried or partially buried atoms causing underestimation of desolvation effects. The volume enclosed by the molecular surface (MS), defined as the surface produced by a solvent spherical probe rolling on the van der Waals surface of the solute,[89] represents the region which is inaccessible to water molecules and is often referred as the Solvent Excluded Volume (SEV).[90] The SEV, lacking the spurious high dielectric interstitial spaces, provides a better representation of the low dielectric region associated with the solute. For this reason accurate Poisson−Boltzmann solvers[9,91,92] have employed the SEV description of the solute region.

Despite its clear advantages, the lack of analytical and computationally efficient representation of the SEV have hampered its deployment in conjunction with GB models for molecular dynamics applications. The Generalized Born Molecular Volume (GBMV) series of models[87,93,94] achieve high accuracy relative to numerical Poisson calculations in part by employing the SEV description of the solute volume. The analytical versions of GBMV[93,94] describe the SEV volume by means of a continuous and differentiable solute density function which is integrated on a grid to yield atomic Born radii. In this work we present a model for the SEV that preserves the analytical pairwise atomic descreening approach employed in the AGBNP1 model,[42] which avoids computations on a grid. We show that this approximate model reproduces some of the key features of the SEV while yielding the same favorable algorithmic scaling of pairwise descreening approaches.

This paper focuses primarily on the description and parametrization of the SEV model and the short-range hydrogen bonding function of AGBNP2. In section 2 we present a brief review of the AGBNP1 model, including the electrostatic and nonpolar models, followed by the derivation of the analytical SEV pairwise descreening model and the short-range hydrogen bonding function which are new for AGBNP2. In section 3 we validate the AGBNP2 analytical estimates for the Born radii and atomic surface areas using as a reference accurate numerical evaluations of these quantities. This is followed by the parametrization of the hydrogen bonding function against experimental hydration free energies of small molecules. This section concludes with a comparison between the structural and energetic properties of a series of structured peptides (miniproteins) predicted with the AGBNP2 model and those obtained with explicit solvation. The paper then concludes with a discussion and implications of the results, and with a perspective on future improvements and validation of the AGBNP2 model.

2. Methods

2.1. The Analytical Generalized Born plus Nonpolar Implicit Solvent Model (AGBNP). In this section we briefly review the formulation of the AGBNP1 implicit solvent model, which forms the basis for the new AGBNP2 model. A full account can be found in the original reference.[42] The AGBNP1 hydration free energy $\Delta G_h(1)$ is defined as

$$\begin{aligned}\Delta G_h(1) &= \Delta G_{elec} + \Delta G_{np} \\ &= \Delta G_{elec} + \Delta G_{cav} + \Delta G_{vdW}\end{aligned} \quad (1)$$

where ΔG_{elec} is the electrostatic contribution to the solvation free energy and ΔG_{np} includes nonelectrostatic contributions. ΔG_{np} is further decomposed into a cavity hydration free energy ΔG_{cav} and a solute−solvent van der Waals dispersion interaction component ΔG_{vdW}.

2.1.1. Geometrical Estimators. Each free energy component in eq 1 is ultimately based on an analytical geometrical description of the solute volume modeled as a set of overlapping atomic spheres of radii R_i centered on the atomic positions $\mathbf{r}_i$. Hydrogen atoms do not contribute to the solute volume. The solute volume is modeled using the Gaussian overlap approach first proposed by Grant and Pickup.[95] In this model the solute volume is computed using the Poincaré formula (also known as the inclusion−exclusion formula) for the volume of the union of a set of intersecting elements

$$V = \sum_i V_i - \sum_{i<j} V_{ij} + \sum_{i<j<k} V_{ijk} - \ldots \quad (2)$$

where $V_i = 4\pi R_i^3/3$ is the volume of atom i, V_{ij} is the volume of intersection of atoms i and j (second-order intersection), V_{ijk} is the volume of intersection of atoms i, j, and k (third-order intersection), and so on. The overlap volumes are approximated by the overlap integral, $V^g_{12\ldots n}$, available in analytical form (see for example eq 10 of ref 42), between n Gaussian density functions each corresponding to a solute atom:

$$V^g_{12\ldots n} \simeq \int d^3\mathbf{r}\, \rho_1(\mathbf{r})\, \rho_2(\mathbf{r}) \ldots \rho_n(\mathbf{r}) \quad (3)$$

where the Gaussian density function for atom i is

$$\rho_i(\mathbf{r}) = p \exp[-c_i(\mathbf{r} - \mathbf{r}_i)^2] \quad (4)$$

where

$$c_i = \frac{\kappa}{R_i^2} \quad (5)$$

and

$$p = \frac{4\pi}{3}\left(\frac{\kappa}{\pi}\right)^{3/2} \quad (6)$$

and κ is a dimensionless parameter that regulates the diffuseness of the atomic Gaussian function. In the AGBNP1 formulation $\kappa = 2.227$.

Gaussian integrals are in principle nonzero for any finite distances between the Gaussian densities. Although not mentioned in ref 42, to reduce computational cost AGBNP1 includes a switching function that reduces to zero the overlap volume between two or more Gaussians when the overlap volume is smaller than a certain value. If $V^g_{12\ldots}$ is the value of the Gaussian overlap volume between a set of atoms, the corresponding overlap volume $V_{12\ldots}$ used in eq 2 is set as

$$V_{12\ldots n} = \begin{cases} 0 & V^g_{12\ldots} \le \upsilon_1 \\ V^g_{12\ldots n} f_w(u) & \upsilon_1 < V^g_{12\ldots n} < \upsilon_2 \\ V^g_{12\ldots n} & V^g_{12\ldots} \ge \upsilon_2 \end{cases} \tag{7}$$

where

$$u = \frac{V^g_{12\ldots} - \upsilon_1}{\upsilon_2 - \upsilon_1} \tag{8}$$

$$f_w(x) = x^3(10 - 15x + 6x^2) \tag{9}$$

where, when using van der Waals atomic radii, $\upsilon_1 = 0.1$ and $\upsilon_2 = 1$ Å^3, and for the augmented radii used in the surface area model (see below), $\upsilon_1 = 0.2$ and $\upsilon_2 = 2$ Å^3. This scheme sets to zero Gaussian overlap volumes smaller than υ_1, leaves volumes above υ_2 unchanged, and smoothly reduces volumes between these two limits. It drastically reduces the number of overlap volumes that need to be calculated since the fact that an n-body overlap volume $V_{12\ldots n}$ between n atoms is zero guarantees that all of the $(n + 1)$-body overlap volumes corresponding to the same set of atoms plus one additional atom are also zero. (Note below that the formulation of AGBNP2 employs modified values of υ_1 and υ_2 to improve the accuracy of surface areas.)

The van der Waals surface area A_i of atom i, which is another geometrical descriptor of the model, is based on the derivative $\partial V/\partial R_i$ of the solute volume with respect to the radius R_i[96]

$$A_i = f_a\left(\frac{\partial V}{\partial R_i}\right) \tag{10}$$

where V is given by eq 2 and

$$f_a(x) = \begin{cases} \dfrac{x^3}{a^2 + x^2} & x > 0 \\ 0 & x \le 0 \end{cases} \tag{11}$$

with $a = 5$ Å^2, is a filter function which prevents negative values for the surface areas for buried atoms while inducing negligible changes to the surface areas of solvent-exposed atoms.

The model further defines the so-called self-volume V'_i of atom i as

$$V'_i = V_i - \frac{1}{2}\sum_j V_{ij} + \frac{1}{3}\sum_{j<k} V_{ijk} + \ldots \tag{12}$$

which is computed similarly to the solute volume and measures the solute volume that is considered to belong exclusively to this atom. Due to the overlaps with other atoms, the self-volume V'_i of an atom is smaller than the van der Waals volume V_i of the atom. The ratio

$$s_i = \frac{V'_i}{V_i} \le 1 \tag{13}$$

is a volume scaling factor used below in the evaluation of the Born radii.

2.1.2. Electrostatic Model. The electrostatic hydration free energy is modeled using a continuous dielectric representation of the water solvent using the Generalized Born (GB) approximation

$$\Delta G_{\text{elec}} = u_\epsilon \sum_i \frac{q_i^2}{B_i} + 2u_\epsilon \sum_{i<j} \frac{q_i q_j}{f_{ij}} \tag{14}$$

where

$$u_\epsilon = -\frac{1}{2}\left(\frac{1}{\epsilon_{\text{in}}} - \frac{1}{\epsilon_{\text{w}}}\right) \tag{15}$$

where ϵ_{in} is the dielectric constant of the interior of the solute and ϵ_{w} is the dielectric constant of the solvent; q_i and q_j are the charges of atom i and j, and

$$f_{ij} = \sqrt{r_{ij}^2 + B_i B_j \exp(-r_{ij}^2/4B_i B_j)} \tag{16}$$

In eqs 14−16 B_i denotes the Born radius of atom i which, under the Coulomb field approximation,[57] is given by the inverse of the integral over the solvent region of the negative fourth power of the distance function centered on atom i

$$\beta_i = \frac{1}{B_i} = \frac{1}{4\pi}\int_{\text{solvent}} \mathrm{d}^3\mathbf{r}\, \frac{1}{(\mathbf{r} - \mathbf{r}_i)^4} \tag{17}$$

In the AGBNP1 model this integral is approximated by a so-called pairwise descreening formula

$$\beta_i = \frac{1}{R_i} - \frac{1}{4\pi}\sum_{j\neq i} s_{ji} Q_{ji} \tag{18}$$

where R_i is the van der Waals radius of atom i, s_{ji} is the volume scaling factor for atom j (eq 13) when atom i is removed from the solute, and Q_{ji} is the integral (available in analytic form; see Appendix B of ref 42) of the function $(\mathbf{r} - \mathbf{r}_i)^{-4}$ over the volume of the sphere corresponding to solute atom j that lies outside the sphere corresponding to atom i. Equation 18 applies to all of the atoms i of the solute (hydrogen atoms and heavy atoms), whereas the sum over j includes only heavy atoms. The AGBNP1 estimates for the Born radii B_i are finally computed from the inverse Born radii β_i from eq 18 after processing them through the function

$$B_i^{-1} = f_b(\beta_i) = \begin{cases} \sqrt{b^2 + \beta_i^2} & \beta_i > 0 \\ b & \beta_i \le 0 \end{cases} \tag{19}$$

where $b^{-1} = 50$ Å. The filter function eq 19 is designed to prevent the occurrence of negative Born radii or Born radii larger than 50 Å. The goal of the filter function is simply to increase the robustness of the algorithm in limiting cases.

The filter function has negligible effect for the most commonly observed Born radii smaller than 20 Å.

In the AGBNP1 model the scaling factors s_{ji} are approximated by the expression

$$s_{ji} \simeq s_j + \frac{1}{2}\frac{V_{ij}}{V_j} \tag{20}$$

where s_j is given by eq 13 and V_{ij} is the two-body overlap volume between atoms i and j. Also, in the original AGBNP formulation the computation of the scaling factors and the descreening function in eq 18 employed the van der Waals radii for the atoms of the solute and the associated Gaussian densities. These two aspects have been modified in the new formulation (AGBNP2) as described below.

2.1.3. Nonpolar Model. The nonpolar hydration free energy is decomposed into the cavity hydration free energy ΔG_{cav} and the solute−solvent van der Waals dispersion interaction component ΔG_{vdW}:

$$\Delta G_{\mathrm{np}} = \Delta G_{\mathrm{cav}} + \Delta G_{\mathrm{vdW}} \tag{21}$$

The cavity component is described by a surface area model[37,46−48]

$$\Delta G_{\mathrm{cav}} = \sum_i \gamma_i A_i \tag{22}$$

where the summation runs over the solute heavy atoms, A_i is the van der Waals surface area of atom i from eq 10, and γ_i is the surface tension parameter assigned to atom i (see Table 1 of ref 42). Surface areas are computed using augmented radii R_i^c for the atoms of the solute and the associated Gaussian densities. Augmented radii are defined as the van der Waals radii (Table 1 of ref 42) plus a 0.5 Å offset. The computation of the atomic surface areas in AGBNP2 is mostly unchanged from the original implementation,[42] with the exception of the values of the switching function cutoff parameters υ_1 and υ_2 of eq 7, which in the new model are set as $\upsilon_1 = 0.01$ Å^3 and $\upsilon_2 = 0.1$ Å^3. This change was deemed necessary to improve the accuracy of the surface areas which in the new model also affect the Born radii estimates through eq 31 below.

The solute−solvent van der Waals free energy term is modeled by the expression

$$\Delta G_{\mathrm{vdW}} = \sum_i \alpha_i \frac{a_i}{(B_i + R_{\mathrm{w}})^3} \tag{23}$$

where α_i is an adjustable dimensionless parameter on the order of 1 (see Table 1 of ref 42) and

$$a_i = -\frac{16}{3}\pi\rho_{\mathrm{w}}\epsilon_{i\mathrm{w}}\sigma_{i\mathrm{w}}{}^6 \tag{24}$$

where $\rho_{\mathrm{w}} = 0.033\,28$ Å^{-3} is the number density of water at standard conditions, and $\sigma_{i\mathrm{w}}$ and $\epsilon_{i\mathrm{w}}$ are the OPLS force field[97] Lennard-Jones interaction parameters for the interaction of solute atom i with the oxygen atom of the TIP4P water model.[98] If σ_i and ϵ_i are the OPLS Lennard-Jones parameters for atom i

$$\sigma_{i\mathrm{w}} = \sqrt{\sigma_i \sigma_{\mathrm{w}}} \tag{25}$$

$$\epsilon_{i\mathrm{w}} = \sqrt{\epsilon_i \epsilon_{\mathrm{w}}} \tag{26}$$

where $\sigma_{\mathrm{w}} = 3.153\,65$ Å and $\epsilon_{\mathrm{w}} = 0.155$ kcal/mol are the Lennard-Jones parameters of the TIP4P water oxygen. In eq 23 B_i is the Born radius of atom i from eqs 18 and 19 and $R_{\mathrm{w}} = 1.4$ Å is a parameter corresponding to the radius of a water molecule.

2.2. The AGBNP2 Implicit Solvent Model. The AGBNP2 hydration free energy $\Delta G_{\mathrm{h}}(2)$ is defined as

$$\Delta G_{\mathrm{h}}(2) = \Delta G_{\mathrm{elec}} + \Delta G_{\mathrm{np}} + \Delta G_{\mathrm{hb}} \tag{27}$$

where ΔG_{elec} and ΔG_{np} have the same form as in the AGBNP1 model (eqs 14 and 21−23, respectively). The only major difference is the pairwise descreening model for the Born radii that in AGBNP2 is based on the solvent excluded volume described below rather than the van der Waals volume as in AGBNP1. ΔG_{hb}, described in section 2.2.2, is a novel term for AGBNP2 which represents a first solvation shell correction corresponding to the portion of the hydration free energy not completely accounted for by the uniform continuum model for the solvent. We think of this term as mainly incorporating the effect of solute−solvent hydrogen bonding. As described in detail below, the analytical model for ΔG_{hb} is based on measuring and scoring the volume of suitable hydration sites on the solute surface.

2.2.1. Pairwise Descreening Model Using the Solvent Excluded Volume. When using van der Waals radii to describe the solute volume, small crevices between atoms (Figure 1, panel A) are incorrectly considered as high dielectric solvent regions,[93,99,100] leading to underestimation of the Born radii, particularly for buried atoms. The van der Waals volume description implicitly ignores the fact that the finite size of water molecules prevents them from occupying sites that, even though they are not within solute atoms, are too small to be occupied by water molecules. Ideally a model for the Born radii would include in the descreening calculation all of the volume excluded from water either because it is occupied by a solute atom or because it is located in an interstitial region inaccessible to water molecules. We denote this volume as the solvent excluded volume (SEV). A realistic description of the SEV is the volume enclosed within the molecular surface[89] of the solute obtained by tracing the surface of contact of a sphere with a radius characteristic of a water molecule (typically 1.4 Å) rolling over the van der Waals surface of the solute. The main characteristic of this definition of the SEV (see Figure 2) is that, unlike the van der Waals volume, it lacks small interstitial spaces while it closely resembles the van der Waals volume near the solute−solvent interface. The molecular surface description of the SEV cannot be easily implemented into an analytical formulation. In this section we will present an analytical description of the SEV for the purpose of the pairwise descreening computation of the Born radii, as implemented in AGBNP2, that preserves the main characteristics of the molecular surface description of the SEV.

The main ideas underpinning the SEV model presented here are illustrated in Figure 1. We start with the van der Waals representation of the solute (model A) which presents an

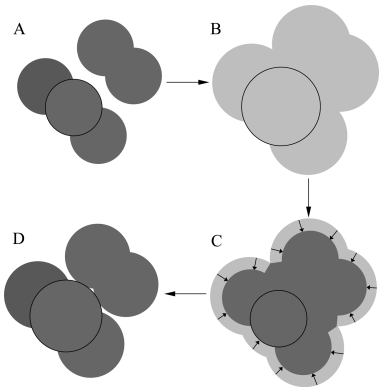

Figure 1. Schematic diagram illustrating the ideas underpinning the model for the solvent excluded volume descreening. Circles represent atoms of two idealized solutes placed in proximity of each other. The van der Waals description of the molecular volume (panel A) leaves high dielectric interstitial spaces that are too small to fit water molecules. The adoption of enlarged van der Waals radii (B) removes the interstitial spaces but incorrectly excludes solvent from the surface of solvent-exposed atoms. The solvent volume subtended by the solvent-exposed surface area is subtracted from the enlarged volume of each atom (C) such that larger atomic descreening volumes are assigned to buried atoms (circled) than exposed atoms (D), leading to the reduction of interstitial spaces while not overly excluding solvent from the surface of solvent-exposed atoms.

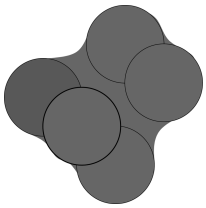

Figure 2. Illustration of the relationship between the van der Waals volume and the solvent excluded volume enclosed by the molecular surface.

undesirable high dielectric interstitial space between the two groups of atoms. Increasing the atomic radii leads to a representation (model B) in which the interstitial space is removed but that also incorrectly excludes solvent from the surface of solvent-exposed atoms. This representation is therefore replaced with one in which the effective volume of each atom in B is reduced by the volume subtended between the solvent-exposed surface of each atom and its van der Waals radius (Figure 1C). This process yields model D in which the effective volume of the most buried atom is larger than those of the solvent-exposed atoms. This SEV model covers the interstitial high dielectric spaces present in a van der Waals description of the solute volume, while approximately maintaining the correct van der Waals volume description of atoms at the solute surface as in the molecular surface description of the SEV (Figure 2).

These ideas have been implemented in the AGBNP2 model as follows. The main modification consists of adopting for the pairwise descreening generalized Born formulation the same augmented van der Waals radii as in the computation of the atomic surface areas. As in the previous model the augmented atomic radii, R_i^c, are set as

$$R_i^c = R_i + \Delta R \tag{28}$$

where R_i is the van der Waals radius of the atom and $\Delta R =$ 0.5 Å is the offset. The augmented radii are used in the same way as in the AGBNP1 formulation to define the atomic spheres and the associated Gaussian densities (eqs 3−6). Henceforth in this work all of the quantities (atomic volumes, self-volumes, etc.) are understood to be computed with the augmented atomic radii, unless otherwise specified. In AGBNP2 the form of the expression for the inverse Born radii (eq 18) is unchanged; however, the expressions for the volume scaling factors s_{ji} and the evaluation of the descreening function Q_{ji} are modified as follows to introduce the augmented atomic radii and the reduction of the atomic self-volumes in proportion to the solvent accessible surface areas as discussed above.

The pairwise volume scaling factors s_{ji}, that is the volume scaling factor for atom j when atom i is removed from the solute, are set as

$$s_{ji} = s_j + \frac{V'_{ji}}{V_j} \tag{29}$$

where s_j (defined below) is the volume scaling factor for atom j analogous to eq 13 computed with all the atoms present, and the quantity

$$V'_{ji} = V'_{ij} = \frac{1}{2}V_{ij} - \frac{1}{3}\sum_k V_{ijk} + \frac{1}{4}\sum_{k<l} V_{ijkl} - \ldots \tag{30}$$

subtracts from the expression for the self-volume of atom j all those overlap volumes involving both atoms i and j.

Two differences with respect to the original AGBNP1 formulation are introduced. The first is that s_j is computed from the self-volume after subtracting from it the volume of the region subtended by the solvent-exposed surface between the enlarged and van der Waals atomic spheres of atom j, according to the expression

$$s_j = \frac{V'_j - d_j A_j}{V_j} \tag{31}$$

where A_j is the surface area of atom j from eq 10. Referring to Figure 3, the volume of the subtended region is $d_j A_j$ as in eq 31 with

$$d_j = \frac{1}{3}R'_j\left[1 - \left(\frac{R_j}{R'_j}\right)^3\right] \tag{32}$$

The other difference concerns the V'_{ji} term which in the AGBNP1 formulation is approximated by the two-body overlap volume V_{ij} (see eq 13), the first term in the right-hand side of eq 30. This approximation is found to lack

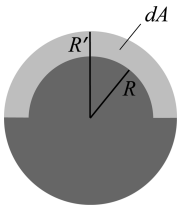

Figure 3. Graphical construction showing the volume subtracted from the atomic self-volume to obtain the surface area corrected atomic self-volume. R is the van der Waals radius of the atom; $R' = R + \Delta R$ is the enlarged atomic radius. dA is the volume of the region (light gray) subtended by the solvent-exposed surface area between the enlarged and van der Waals atomic spheres.

sufficient accuracy for the present formulation given the relative increase in size of all overlap volumes. Therefore in AGBNP2 V'_{ji} is computed including in eq 30 all nonzero overlap volumes after the application of the switching function from eq 7.

In the AGBNP2 formulation the functional form for the pair descreening function Q_{ji} is the same as in the original formulation (see Appendix of ref 42); however, in the new formulation this function is evaluated using the van der Waals radius R_i for atom i (the atom being "descreened") and the augmented radius R_j^c for atom j (the atom that provides the solvent descreening), rather than using the van der Waals radius for both atoms. Thus if the pair descreening function is denoted by $Q(r,R_1,R_2)$, where r is the interatomic distance, R_1 the radius of the atom being descreened, and R_2 the radius that provides descreening, we set in eq 18

$$Q_{ji} = Q(r_{ij}, R_i, R_j^c) \tag{33}$$

The alternative of using enlarged atoms for both atoms and the inclusion of a properly weighted self-descreening term (to take into account the SEV of the atom being descreened) was also tried and judged to be less accurate than eq 33 relative to numerical integration.

2.2.2. Short-Range Hydrogen Bonding Correction Function. In this section we present the analytical model that implements the short-range hydrogen bonding correction function for AGBNP2. The model is based on a geometrical procedure, described below, to measure the degree to which a solute atom can interact with hydration sites on the solute surface. The procedure is as follows. A sphere of radius R_s representing a water molecule is placed in a position that provides near-optimal interaction with a hydrogen bonding donor or acceptor atom of the solute. The position $\mathbf{r}_s$ of this water sphere s is function of the positions of two or more parent atoms that compose the functional group including the acceptor/donor atom:

$$\mathbf{r}_s = \mathbf{r}_s(\{\mathbf{r}_{ps}\}) \tag{34}$$

where $\{\mathbf{r}_{ps}\}$ represents the positions of the set of parent atoms of the water site s. For instance, the water site position in correspondence with a polar hydrogen is

$$\mathbf{r}_s = \mathbf{r}_D + \frac{\mathbf{r}_H - \mathbf{r}_D}{|\mathbf{r}_H - \mathbf{r}_D|} d_{HB}$$

where $\mathbf{r}_D$ is the position of the heavy atom donor, $\mathbf{r}_H$ is the position of the polar hydrogen, and d_{HB} is the distance between the heavy atom donor and the center of the water sphere (see Figure 4). Similar relationships (see the Appendix) are employed to place candidate water spheres in correspondence with hydrogen bonding acceptor atoms of the solute. These relationships are based on the local topology of the hydrogen bonding acceptor group (linear, trigonal, and tetrahedral). This scheme places one or two water spheres in correspondence with each hydrogen bonding acceptor atom (see Table 1).

The magnitude of the hydrogen bonding correction corresponding to each water sphere is a function of the predicted water occupancy of the location corresponding to the water sphere. In this work the water occupancy is measured by the fraction w_s of the volume of the water site sphere that is accessible to water molecules without causing steric clashes with solute atoms (see Figure 4)

$$w_s = \frac{V_s^{free}}{V_s} \tag{35}$$

where $V_s = (4/3)\pi R_s^3$ is the volume of the water sphere and

$$V_s^{free} = V_s - \sum_i V_{si} + \sum_{i<j} V_{sij} - \sum_{i<j<k} V_{sijk} \tag{36}$$

is the "free" volume of water site s, obtained by summing over the two-body, three-body, etc. overlap volumes of the water sphere with the solute atoms. Note that the expression of the free volume is the same as the expression for the self-volume (eq 12) except that it lacks the fractional coefficients 1/2, 1/3, etc. The overlap volumes in eq 36 are computed using radius R_s for the water site sphere (here set to 1.4 Å) and augmented radii R_i^c for the solute atoms. Equation 36 is derived similarly to the expression for the self-volumes by removing overlap volumes from the volume of the water sphere rather than evenly distributing them across the atoms participating in the overlap.

Given the water occupancy w_s of each water sphere, the expression for the hydrogen bonding correction for the solute is

$$\Delta G_{hb} = \sum_s h_s S(w_s; w_a, w_b) \tag{37}$$

where h_s is the maximum correction energy which depends on the type of solute−water hydrogen bond (see Table 1), and $S(w;w_a,w_b)$ is a polynomial switching function which is 0 for $w < w_a$, 1 for $w > w_b$, and smoothly (with continuous first derivatives) interpolates from 0 to 1 between w_a and w_b (see Figure 5). The expression of $S(w;w_a,w_b)$ is

$$S(w; w_a, w_b) = \begin{cases} 0 & w \le w_a \\ f_w\left(\dfrac{w - w_a}{w_b - w_a}\right) & w_a < w < w_b \\ 1 & w \ge w_b \end{cases} \tag{38}$$

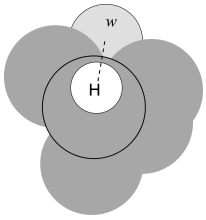

Figure 4. Schematic diagram for the placement of the water sphere (*w*, light gray) corresponding to the hydrogen bonding position relative to the a polar hydrogen (white sphere) of the solute (dark gray). The dashed line traces the direction of the hydrogen−parent heavy atom (circled) bond along which the water sphere is placed. The magnitude of hydrogen bonding correction grows as a function of the volume (light gray) of the water site sphere not occupied by solute atoms.

Table 1. Optimized Surface Tension Parameters and Hydrogen Bonding Correction Parameters for the Atom Types Present in Protein Molecules[a]

atom type	γ (cal/mol/Å^2)	geometry	N_w	h (kcal/mol)
C (aliphatic)	129			
C (aromatic)	120			
H on N		linear	1	−0.25
H on N (Arg)		linear	1	−2.50
H on O		linear	1	−0.40
H on S		linear	1	−0.50
O (alcohol)	117	tetrahedral	2	−0.40
O (carbonyl)	117	trigonal	2	−1.25
O (carboxylate)	40	trigonal	2	−1.80
N (amine)	117	tetrahedral	1	−2.00
N (aromatic)	117	trigonal	1	−2.00
S	117	tetrahedral	2	−0.50

[a] γ is the surface tension parameter, N_w is the number of water spheres, and h is the maximum correction corresponding to each atom type (eq 37). Atom types not listed do not have hydrogen bonding corrections and are assigned $\gamma = 117$ cal/mol/Å^2.

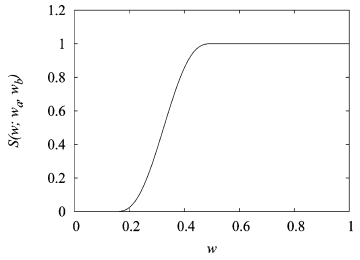

Figure 5. Switching function $S(w;w_a,w_b)$ from eqs 38 and 9 with $w_a = 0.15$ and $w_b = 0.5$.

where $f_w(x)$ is a switching function given by eq 9. In this work we set $w_a = 0.15$ and $w_b = 0.5$. This scheme establishes (see Figure 5) that no correction is applied if more than 85% of the water sphere volume is not water accessible, whereas maximum correction is applied if 50% or more of the water sphere volume is accessible.

2.3. Molecular Dynamics of Miniproteins. We conducted molecular dynamics simulations of what we will refer to as miniproteins (Figure 6), that is, peptides that have been shown to form stable secondary structures in solution: the 23-residue trp-cage peptide of sequence ALQELLGQWLKDGGPSSGRPPPS [Protein Data Bank (PDB) ID 1RIJ],[101] the 28-residue cdp-1 peptide of sequence KPYTARIKGRTFSNEKELRDFLETFTGR (PDB ID 1PSV),[102] and the 28-residue fsd-1 peptide of sequence QQYTAKIKGRTFRNEKELRDFIEKFKGR (PDB ID 1FSD).[103] The structure of trp-cage (see Figure 6) is characterized by a tryptophan side chain enclosed in a cage formed by an α-helix on one side and a proline-rich loop on the other. The cdp-1 and fsd-1 miniproteins (Figure 6) adopt a mixed $\alpha\beta$ conformation and are particularly rich in charged residues. The trp-cage miniprotein was chosen because it has been the target of several computational studies.[104−107] The cdp-1 and fsd-1 peptides were of interest because they showed in preliminary tests with AGBNP1 solvation a significant tendency to deviate from the experimental structures.

Molecular dynamics simulations were conducted for 12 ns starting with the first NMR model deposited in the PDB. The temperature was set to 300 K with the Nosé−Hoover thermostat,[108,109] a molecular dynamics (MD) time step of 2 fs was employed, and covalent bond lengths involving hydrogen atoms were fixed at their equilibrium positions. Backbone motion was restricted by imposing a positional harmonic restraint potential with a force constant of 0.3 kcal/mol/Å^2 on the positions of the Cα atoms, which allows for a range of motion of about 5 Å at the simulation temperature. These restraints are sufficiently weak to allow substantial backbone and side chain motion while preserving overall topology.

Molecular dynamics simulations were conducted with the OPLS-AA potential[97,110] with explicit solvation (SPC water model with 2450, 3110, and 3250 water molecules for trp-cage, cdp-1, and fsd-1, respectively) and with both AGBNP1 and AGBNP2 implicit solvation. The DESMOND program[111] was used for the explicit solvent simulations, and the IMPACT program[15] was used for those with implicit solvation. Identical force field settings were employed in these two programs. The explicit solvent simulations were conducted in the NPT ensemble using the Martyna−Tobias−Klein barostat[112] at 1 atm pressure and employed the smooth Particle Mesh Ewald (PME) method[113] for the treatment of long-range electrostatic interactions with a real-space cutoff of 9 Å. Equilibrium averages and energy distributions were obtained by analysis of the latter 10 ns of saved trajectories. Convergence was tested by comparing averages obtained using the first and second halves of simulation data. Hydrogen bonds were detected using a minimum hydrogen-acceptor distance of 2.5 Å and a minimum donor angle of 120°.

3. Results

3.1. Accuracy of Born Radii and Surface Areas. The quality of any implicit solvent model depends primarily on the reliability of the physical model on which it is based. The accuracy of the implementation, however, is also a critical aspect for the success of the model in practice. This is true in particular for models, such as AGBNP, based on

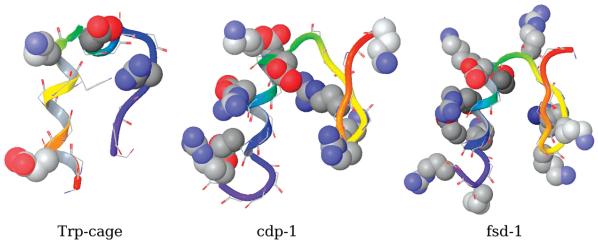

Figure 6. Graphical representations of the NMR structures of the three miniproteins investigated in this work: trp-cage (PDB ID 1RIJ), cdp-1 (PDB ID 1PSV), and fsd-1 (PDB ID 1FSD). In each case the first deposited NMR model is shown. Backbone ribbon is colored from the N-terminal (red) to the C-terminal (blue). Charged side chains are shown in space-filling representation.

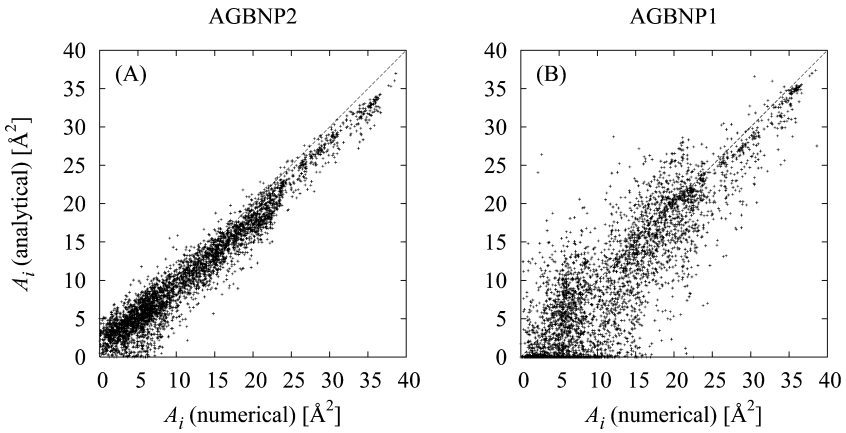

Figure 7. Comparisons between numerical and analytical molecular surface areas of the heavy atoms of the crystal structures (1ctf and 1lz1, respectively) of the C-terminal domain of the ribosomal protein L7/L12 (74 aa) and human lysozyme (130 aa), and of four conformations each of the trp-cage, cdp-1, and fsd-1 miniproteins extracted from the corresponding explicit solvent MD trajectories. (A) Analytical molecular surface areas computed using the present model and (B), for comparison, analytical surface areas computed using the original model from ref 42.

the generalized Born formula. It has been pointed out, for instance, that approximations in the integration procedure to obtain the Born radii may actually be of more significance than the physical approximations on which the GB model is based.[114] It is therefore important to test that the conformational-dependent quantities employed by AGBNP2 are a good approximation to the geometrical parameters that they are supposed to represent. The present AGBNP2 formulation relies mainly on three types of conformational-dependent quantities: Born radii (eq 18), solvent accessible surface areas (eq 10), and solvent accessibilities of hydration sites (eq 38). In this section we analyze the validity of the AGBNP2 analytical estimates for the Born radii and surface areas against accurate numerical results for the same quantities.

We employ the GEPOL program[90] to compute numerically atomic solvent accessible surface areas with a solvent probe diameter of 1 Å, the same probe diameter that defines the solute−solvent boundary in the AGBNP model. Figure 7A shows the comparison between the surface area estimates given by the present formulation of AGBNP and the numerical surface areas produced by GEPOL for a series of native and modeled protein conformations. In Figure 7B we show the same comparison for the surface areas of the original AGBNP1 model. These representative results show that the present analytical surface area implementation, which as described above employs a weaker switching function for the overlap volumes, produces significantly more accurate atomic surface areas than the original model. These improvements in the computation of the surface areas, introduced mainly to obtain more accurate Born radii through eq 31, are also expected to yield more reliable cavity hydration free energy differences.

Figure 8 illustrates on the same set of protein conformations the accuracy of the inverse Born radii, B_i^{-1}, obtained using the AGBNP2 pairwise descreening method using the SEV model for the solute volume described above (eq 18), by comparing them to accurate estimates obtained by evaluating the integral in eq 17 numerically on a grid. The comparison is performed for the inverse Born radii because these, being proportional to GB self-energies, are more reliable accuracy indicators than the Born radii themselves. The grid for the numerical integration was prepared as

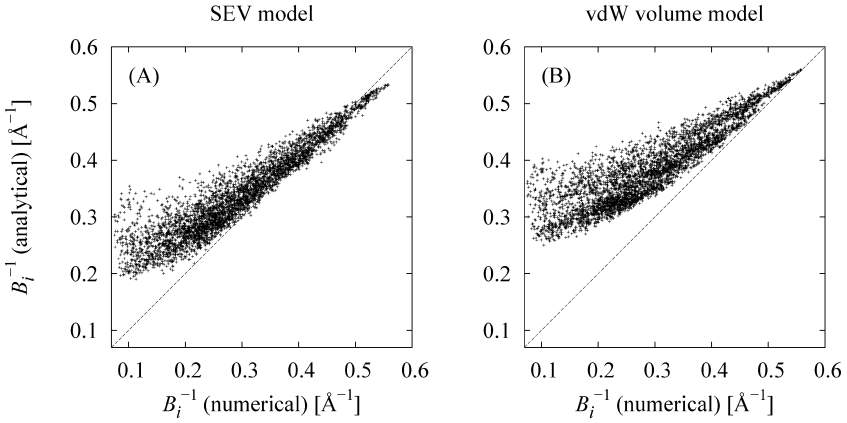

Figure 8. Comparisons between numerical and analytical inverse Born radii for the heavy atoms of the same protein conformations as in Figure 7. (A) Analytical Born radii computed using the present SEV model. (B) Analytical Born radii computed using the van der Waals volume model (ref 42).

previously reported,[42] except that the solvent excluded volume (SEV) of the solute was employed here rather than the van der Waals volume. The integration grid over the SEV was obtained by taking advantage of the particular way that the GEPOL algorithm describes the SEV of the solute; GEPOL iteratively places auxiliary spheres of various dimensions in the interstitial spaces between solute atoms in such a way that the van der Waals volume of the solute plus the auxiliary spheres accurately reproduces the SEV of the solute. Therefore in the present application a grid point was considered part of the SEV of the solute if it was contained within any solute atom or any one of the auxiliary spheres placed by GEPOL. The default 1.4 Å solvent probe radius was chosen for the numerical computation of the SEV with the GEPOL program to assess the accuracy of the model with respect to a full representation of the solute solvent excluded volume as in the GBMV series of models.[93,94] The results of this validation (Figure 8) show that the analytical SEV pairwise descreening model described above is able to yield Born radii which are not as affected by the spurious high dielectric interstitial spaces present in the van der Waals volume description of the solute. With the van der Waals volume model (Figure 8B) the Born radii of the majority of solute atoms start to significantly deviate from the reference values for Born radii larger than about 2.5 Å ($B^{-1} = 0.4$ Å^{-1}). Born radii computed with the analytical SEV model instead (Figure 8A) track the reference values reasonably well further up to about 4 Å ($B^{-1} = 0.25$ Å^{-1}). Despite this significant improvement most Born radii are still underestimated by the improved model (and, consequently, the inverse Born radii are overestimated—see Figure 8), particularly those of nonpolar atoms near the hydrophobic core of the larger proteins. These regions tend to be loosely packed and tend to contain interstitial spaces too large to be correctly handled by the present model. Because it mainly involves groups of low polarity, this limitation has a small effect on the GB solvation energies. It has however a more significant effect on the van der Waals solute−solvent interaction energy estimates through eq 23, which systematically overestimate the magnitude of the interaction for atoms buried in hydrophobic protein core. While the present model in general ameliorates in all respects the original AGBNP model, we are currently exploring ways to address this residual source of inaccuracy.

3.2. Small Molecule Hydration Free Energies. The validation and parametrization of the hydrogen bonding and cavity correction parameters have been performed based on the agreement between experimental and predicted AGBNP2 hydration free energies of a selected set of small molecules, listed in Table 2, containing the main functional groups present in proteins. This set of molecules includes only small and relatively rigid molecules whose hydration free energies can be reliably estimated using a single low energy representative conformation[115] as was done here. Table 2 lists for each molecule the experimental hydration free energy, the AGBNP2 hydration free energy computed without hydrogen bonding (HB) corrections and the default $\gamma = 117$ cal/mol/Å^2 surface tension parameter, denoted by AGBNP2/SEV, as well as the hydration free energy from the AGBNP2 model including the HB correction term and the parameters listed in Table 1. For comparison, the corresponding predictions with the original AGBNP1[42] model are reported in the Supporting Information.

Going down the results in Table 2, we notice a number of issues addressed by the new implementation. With the new surface area implementation and without corrections (third column in Table 2), the hydration free energies of the normal alkanes are too small compared to experiments; furthermore, in contrast with the experimental behavior, the predicted hydration free energies incorrectly become more favorable with increasing chain length. A similar behavior is observed for the aromatic hydrocarbons. Clearly this is due to the rate of increase of the positive cavity term with increasing alkane size which is insufficient to offset the solute−solvent van der Waals interaction energy term, which becomes more negative with increasing solute size. We have chosen to

Table 2. Experimental and Predicted Hydration Free Energies of a Set of Small Molecules

molecule	exptl[a,b]	AGBNP2/SEV[a,c]	AGBNP2[a,d]
n-ethane	1.83	0.98	1.80
n-propane	1.96	0.92	1.97
n-butane	2.08	0.88	2.14
n-pentane	2.33	0.78	2.26
n-hexane	2.50	0.70	2.40
cyclopentane	1.20	0.34	1.63
cyclohexane	1.23	0.05	1.50
benzene	−0.87	−1.50	−1.14
toluene	−0.89	−1.66	−0.94
acetone	−3.85	−1.09	−3.83
acetophenone	−4.58	−2.74	−5.07
ethanol	−5.01	−4.77	−5.30
phenol	−6.62	−4.51	−5.38
ethanediol	−9.60	−7.99	−9.87
acetic acid	−6.70	−2.73	−7.05
propionic acid	−6.48	−2.58	−6.38
methyl acetate	−3.32	−0.10	−3.92
ethyl acetate	−3.10	−0.02	−3.60
methyl amine	−4.56	−2.39	−4.37
ethyl amine	−4.50	−2.24	−3.95
dimethyl amine	−4.29	−1.95	−3.21
trimethyl amine	−3.24	−1.78	−2.39
acetamide	−9.71	−6.81	−10.45
N-methylacetamide	−10.08	−4.75	−7.51
pyridine	−4.70	−3.62	−5.30
2-methylpyridine	−4.63	−2.94	−4.22
3-methylpyridine	−4.77	−2.82	−4.13
methanethiol	−1.24	−0.61	−1.46
ethanethiol	−1.30	−0.57	−1.22
neutral AUE[a,e]		1.90	0.45
acetate ion	−79.90	−77.32	−87.70
propionate ion	−79.10	−76.29	−86.29
methylammonium ion	−71.30	−73.21	−73.54
ethylammonium ion	−68.40	−70.63	−70.75
methyl guanidinium	−62.02[f]	−57.30	−69.81
ions AUE[a,g]		2.85	5.47

[a] In kcal/mol. [b] Experimental hydration free energy from ref 116 except where indicated. [c] AGBNP predicted hydration free energies with the default γ parameter for all atoms types ($\gamma = 117$ cal/mol/Å^2) and without HB corrections. [d] AGBNP predicted hydration free energies with optimized parameters listed in Table 1. [e] Average unsigned error of the AGBNP predictions for the neutral compounds relative to the experiments. [f] From ref 117. [g] Average unsigned error of the AGBNP predictions for the ionic compounds relative to the experiments.

address this shortcoming by increasing by 10.2% and 2.5% respectively the effective surface tensions for aliphatic and aromatic carbon atoms rather than decreasing the corresponding α parameters of the van der Waals term since the latter had been previously validated against explicit solvent simulations. We have chosen to limit the increases of the surface tension parameters to aliphatic and aromatic carbon atoms since the results for polar functional groups did not indicate that this change was necessary for the remaining atom types. With this new parametrization we achieve (compare the second and fourth columns in Table 2) excellent agreement between the experimental and predicted hydration free energies of the alkanes and aromatic compounds. Note that the AGBNP2 model, regardless of the parametrization, correctly predicts the more favorable hydration free energies of the cyclic alkanes relative to their linear analogues. AGBNP2, thanks to its unique decomposition of the nonpolar solvation free energy into an unfavorable cavity term and an opposing favorable term, is, to our knowledge, the only analytic implicit solvent implementation capable of describing correctly this feature of the thermodynamics of hydration of hydrophobic solutes.

The AGBNP2 model without corrections markedly underpredicts the magnitudes of the experimental hydration free energies of the compounds containing carbonyl groups (ketones, organic acids, and esters). The hydration free energies of alcohols are also underpredicted but by smaller amounts. Much better agreement with the experimental hydration free energies of these oxygen-containing compounds (see Table 2) is achieved after applying hydrogen bonding corrections with $h = -1.25$ kcal/mol for the carbonyl oxygen and $h = -0.4$ kcal/mol for both the hydrogen and oxygen atoms of the hydroxy group (Table 1). Note that the same parameters employed individually for carbonyl and hydroxy groups in ketones and alcohols are applied to the more complex carboxylic groups of acids and esters as well as amides and carboxylate ions. The thiol groups of organic sulfides required similar corrections as the hydroxy groups (Tables 1 and 2). The AGBNP2 model without corrections also markedly underpredicted the magnitude of the experimental hydration free energies of amines and amides and, to a smaller extent, of compounds with nitrogen-containing heterocyclic aromatic rings. The addition of HB corrections of −0.25 kcal/mol for amine hydrogens and $h = -2.0$ kcal/mol for both amine and aromatic nitrogen atoms yields improved agreement (Table 2), although the effect of N-methylation is still overemphasized.

3.3. Miniprotein Results. As described in section 2.3, we have performed restricted MD simulations of a series of so-called miniproteins (trp-cage, cdp-1, and fsd-1) to study the extent of the agreement between the conformational ensembles generated with the original AGBNP implementation (AGBNP1) and the present implementation (AGBNP2) with respect to explicit solvent generated ensembles. The results of earlier studies[4,54,55] suggest that the AGBNP/OPLS-AA model correctly reproduces for the most part the backbone secondary structure features of protein and peptides. The tests in the present study are therefore focused on side chain conformations. The backbone atoms were harmonically restricted to remain within approximately 3 Å Cα root-mean-square deviation of the corresponding NMR experimental models. We structurally analyzed the ensembles in terms of the extent of occurrence of intramolecular interactions.

As shown in Table 3, we measured a significantly higher average number of intramolecular hydrogen bonds and ion pairing in the AGBNP1 ensembles relative to the explicit solvent ensembles for all miniproteins studied. The largest deviations are observed for cdp-1 and fsd-1, two miniproteins particularly rich in charged side chains, with on average nearly twice as many intramolecular hydrogen bonds compared to explicit solvent. Many of the excess intramolecular hydrogen bonds with AGBNP1 involve interactions between polar groups (polar side chains or the peptide backbone) and the side chains of charged residues. For example, for cdp-1 we observe (see Table 3) approximately eight hydrogen bonds between polar and charged groups on average compared to nearly none with explicit solvation.

Table 3. Average Number of Some Types of Intramolecular Electrostatic Interactions in the Explicit Solvent Conformational Ensembles, and the Ensembles Generated from Simulations Using the AGBNP1 and AGBNP2 Effective Potentials for the trp-cage, cdp-1, and fsd-1 Miniproteins

miniprotein	explicit	AGBNP1	AGBNP2
Intramolecular Hydrogen Bonds			
trp-cage	13.5	18.3	15.3
cdp-1	12.6	24.5	15.4
fsd-1	14.1	24.6	14.3
all	40.2	67.4	45.0
Polar−Polar Hydrogen Bonds			
trp-cage	12.9	17.1	13.9
cdp-1	12.5	16.4	14.1
fsd-1	12.0	15.0	12.9
all	37.4	48.5	40.9
Polar−Charged Hydrogen Bonds			
trp-cage	0.6	1.2	1.4
cdp-1	0.1	8.1	1.3
fsd-1	2.1	9.6	1.4
all	2.8	18.9	4.1
Ion Pairs			
trp-cage	0.3	1.0	1.0
cdp-1	2.5	2.9	2.7
fsd-1	1.4	4.6	4.0
all	4.2	8.5	7.7

Despite the introduction of empirical surface tension correction to penalize ion pairs,[55] AGBNP1 overpredicts ion pair formation. We found that ion pairing involving arginine was particularly overstabilized by AGBNP1 as we observed stable ion pairing between arginine and either glutamate or aspartate residues during almost the entire duration of the simulation in virtually all cases in which this was topologically feasible given the imposed backbone restrains. In contrast, with explicit solvation some of the same ion pairs were seen to form and break numerous times, indicating a balanced equilibrium between contact and solvent-separated conformations. This balance is not reproduced with implicit solvation, which instead strongly favors ion pairing. In any case, the relative stability of ion pairs appeared to depend in subtle ways on the protein environment as, for example, the two ion pairs between arginine and glutamate of cdp-1 were found to be stable with either explicit solvation or AGBNP1 implicit solvation whereas other Arg−Glu ion pairs in trp-cage and fsd-1 were found to be stable only with implicit solvation.

This analysis generally confirms quantitatively a series of past observations made in our laboratory indicating that the original AGBNP implementation tends to be biased toward conformations with excessive intramolecular electrostatic interactions, at the expense of more hydrated conformations in which polar groups are oriented so as to interact with the water solvent. During the process of development of the modifications to address these problems, we found it useful to rescore with varying AGBNP formulations and parametrizations the miniprotein conformational ensembles obtained with AGBNP1 and explicit solvation, rather than performing simulations with each new parametrizations. An example of this analysis is shown in the first row of plots of Figure 9, which compare the probability distributions of the

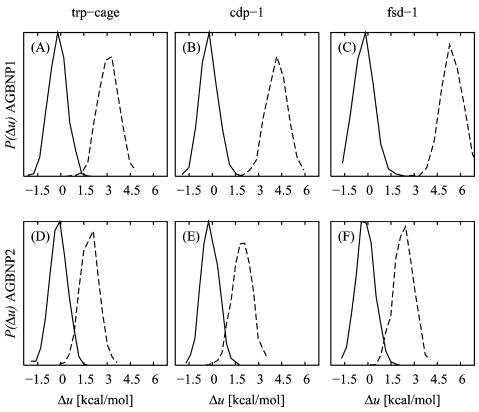

Figure 9. Potential energy distributions of the conformational ensembles for the trp-cage (first column, panels A, D), cdp-1 (second column, panels B, E), and fsd-1 (third column, panels C, F) miniproteins obtained using the AGBNP1/OPLS-AA (first row, panels A−C; full line) and AGBNP2/OPLS-AA (second row, panels D−F; full line) effective potentials and explicit solvation (dashed line). The distributions are shown as a function of the energy gap per residue (Δu) relative to the mean effective potential energy of the implicit solvent ensemble distribution.

AGBNP1 effective potential energies over the conformational ensembles generated with AGBNP1 implicit solvation and with explicit solvation. These results clearly show that the AGBNP1/OPLS-AA effective potential disfavors conformations from the explicit solvent ensemble relative to those generated with implicit solvation. The AGBNP1 energy scores of the explicit solvent ensembles of all miniproteins are shifted toward higher energies than those of the AGBNP1 ensemble, indicating that conformations present in the explicit solvent ensemble would be rarely visited when performing conformational sampling with the AGBNP1/OPLS-AA potential. AGBNP1/OPLS-AA assigns a substantial energetic penalty (see Figure 9A−C) to the explicit solvent ensemble relative to the AGBNP1 ensemble (on average 3.3, 4.4, and 5.7 kcal/mol per residue for, respectively, the trp-cage, cdp-1, and fsd-1 miniproteins). This energetic penalty, being significantly larger than thermal energy, rules out the possibility that conformational entropy effects could offset it to such an extent so as to equalize the relative free energies of the two ensembles. Detailed analysis of the energy scores shows that, as expected, the AGBNP1 implicit solvent ensemble is mainly favored by more favorable electrostatic Coulomb interaction energies due to its greater number of intramolecular electrostatic contacts relative to the explicit solvent ensemble (see above). Conversely, the AGBNP1 solvation model does not assign sufficiently favorable hydration free energy to the more solvent-exposed conformations obtained in explicit solvation so as to make them competitive with the more compact conformations of the AGBNP1 ensemble.

Similar energetic scoring analysis with the AGBNP2 model (see Figure 1 of the Supporting Information) with and

without hydrogen bonding to solvent corrections showed that the introduction of the SEV model for the solute volume significantly reduced the energetic gap between the explicit solvent and AGBNP1 conformational ensembles, and that the introduction of the hydrogen bonding corrections further favors the explicit solvent ensemble. We proceeded to vary the AGBNP2 parameters to achieve the best possible scoring of the explicit solvent ensembles relative to the AGBNP1 ensembles while maintaining an acceptable level of agreement with the small molecule experimental hydration free energies. This procedure eventually yielded the parameters listed in Table 1, which produce small molecule hydration free energies in good agreement with the experiments (Table 2), as well as energy distributions for the three miniproteins that, while still favoring the AGBNP1 ensembles, displayed energy gaps between the explicit and AGBNP1 implicit solvation ensembles comparable to thermal energy and smaller than the spread of the energy distributions.

The energy scoring experiments on the explicit solvent and AGBNP1 ensembles described above were very useful for tuning the formulation of the AGBNP2 model without requiring running lengthy MD simulations. They do not, however, guarantee that the conformational ensembles generated with the AGBNP2 solvation model will more closely match the explicit solvent ensembles than those generated with AGBNP1. This is because the new solvation model could introduce new energy minima not encountered with AGBNP1 or explicit solvation that would be visited only by performing conformational sampling with AGBNP2. To validate the model in this respect, we obtained MD trajectories with the AGBNP2 implicit solvent model and compared the corresponding probability distributions of the effective energy with those of the explicit solvent ensembles similarly as above. The results for the three miniproteins, shown in Figure 9D−F, indicate that the AGBNP2-generated ensembles display significantly smaller bias (mean energy gaps per residue of 2.0, 2.1, and 2.5 kcal/mol for, respectively, the trp-cage, cdp-1, and fsd-1 miniproteins) than AGBNP1 (Figure 9A−C), which yielded energy gaps of 3.3, 4.4, and 5.7 kcal/mol per residue, respectively. This observation shows that AGBNP2 produces conformational ensembles with energy distributions that more closely match on average that of the explicit solvent ensemble without producing unphysical minima that deviate significantly from it.

We have analyzed structural features of the conformational ensembles obtained with the AGBNP1 and AGBNP2 models to establish the degree of improvement achieved with the new model with respect to intramolecular interactions. The salient results of this analysis are shown in Table 3. This table reports for each miniprotein the average number of intramolecular hydrogen bonds and ion pairs. The number of hydrogen bonds is further decomposed into those involving only polar groups (including the backbone) and those involving a polar group and the side chain of a charged residue (arginine, lysine, aspartate, and glutamate). As noted above, it is apparent from these data that the AGBNP1 model produces conformations with too many hydrogen bonds and ion pairs. The majority of the excess hydrogen bonds with AGBNP1 involve residue side chains. Similarly, too many ion pairs are observed in the AGBNP1 ensemble particularly for the fsd-1 miniprotein (4.6 ion pairs on average with AGBNP1 compared to only 1.4 in explicit solvent). The AGBNP2 ensembles, in comparison, yield considerably fewer intramolecular hydrogen bonds. For instance, the average number of hydrogen bonds for cdp-1 is reduced from 24.5 with AGBNP1 to 15.4 with AGBNP2, which is to be compared with 12.6 in explicit solvent. With AGBNP2 the number of polar−polar hydrogen bonds is generally in good agreement with explicit solvation. However, the greatest improvement is observed with polar−charged interactions. For example, the number of polar−charged hydrogen bonds of fsd-1 is reduced by almost 10-fold in going from AGBNP1 to AGBNP2 to reach good agreement with explicit solvation. Importantly, a significant fraction of the excess polar−charged interactions observed with AGBNP1 corrected by AGBNP2 are interactions between the peptide backbone and charged side chains that would otherwise interfere with the formation of secondary structures.

With AGBNP2 we observe small but promising improvements in terms of ion pair formation. The average number of ion pairs of cdp-1 consistently agrees between all three solvation models, and the only possible ion pair in trp-cage is more stable in both implicit solvent formulations than in explicit solvent (it occurs in virtually all implicit solvent conformations compared to only 30% of the conformations in explicit solvent). However, the average number of ion pairs for fsd-1 is reduced from 4.6 with AGBNP1 to 4.0 with AGBNP2. We observe good agreement between the number of ion pairs involving lysine with either AGBNP1 or AGBNP2 and explicit solvation. However, ion pairs involving arginine are generally more stable with implicit solvation than with explicit solvation. The agreement in the number of ion pairs with cdp-1 is due to the fact that for this miniprotein the two possible ion pairs involving arginine result stable with explicit solvation as well as with implicit solvation. For the other two miniproteins, however, ion pairs involving arginine that are marginally stable with explicit solvation are found to be significantly more stable with implicit solvation, although less so with AGBNP2 solvation.

4. Discussion

Modern implicit solvent models for biomolecular simulations are generally based on the uniform dielectric continuum representation of the solvent which is accurately modeled by the Poisson−Boltzmann (PB) equation.[9] Generalized Born (GB) models,[10] which approximate the PB formalism, are applicable to molecular dynamics thanks to their low computational complexity. GB models have reached a high level of accuracy compared to PB following the introduction of more realistic solute volume descriptions[87,100] and of higher order corrections to the Coulomb field approximation.[118−120] However, at the molecular level water is sometimes poorly described by uniform continuum models. Even the best GB models have been found to deviate considerably from, for example, explicit solvent benchmarks.[121,127] The nonlinear and asymmetric dielectric response of water stems primarily from the finite extent and

internal structure of water molecules.[1] The modeling of effects due to water granularity is important for the proper description of molecular association equilibria. Integral equation methods[122] provide an accurate implicit solvation description from first principles; however, despite recent progress,[123] they are not yet applicable to molecular dynamics of biomolecules. The primary aim of the present study has been to formulate an analytical and computational efficient implicit solvent model incorporating solvation effects beyond those inherent in standard continuum dielectric models and, by so doing, achieving an improved description of solute conformational equilibria.

In this work we have developed the AGBNP2 implicit solvent model which is based on an empirical (but physically motivated) first solvation shell correction function parametrized against experimental hydration free energies of small molecules and the results of explicit solvent molecular dynamics simulations of a series of miniproteins. The correction function favors conformations of the solute in which polar groups are oriented so as to form hydrogen bonds with the surrounding water solvent, thereby achieving a more balanced equilibrium with respect to the competing intramolecular hydrogen bond interactions. A key ingredient of the model is an analytical prescription to identify and measure the volume of hydration sites on the solute surface. Hydration sites that are deemed too small to contain a water molecule do not contribute to the solute hydration free energy. Conversely, hydration sites of sufficient size form favorable interactions with nearby polar groups. This model thus incorporates the effects of both water granularity and nonlinear first shell hydration effects.

The GB and nonpolar models in the AGBNP2 implicit solvent model provide the linear continuum dielectric model basis of the model as well as a description of nonelectrostatic hydration effects.[42] In this work the GB and solute−solvent dispersion interaction energy models are further enhanced by replacing the original van der Waals solute volume model with a more realistic solvent excluded volume (SEV) model. The new volume description improves the quality of the Born radii of buried atoms and atoms participating in intramolecular interactions which would otherwise be underestimated due to high dielectric interstitial spaces present with the van der Waals volume description.[88] GB models with these characteristics have been previously proposed. The GBMV series of models[87,93,94] represent the SEV on a grid which, although accurate, is computationally costly and lacks frame of reference invariance. The pairwise descreening based GB^{OBC} model[120] introduced an empirical rectifying function to increase the Born radii of buried atoms in an averaged, geometry-independent manner. The GBn model[100] introduced the neck region between pairs of atoms as additional source of descreening, dampened by empirical parameters to account in an average way for overlaps between neck regions and between solute atoms and neck regions. The approach proposed here to represent the SEV consists of computing the atomic self-volumes, used in the pairwise descreening computation, using enlarged atomic radii so as to cover the interatomic interstitial spaces. The self-volume of each atom is then reduced proportionally to its solvent accessible surface area (see eq 31) to subtract the volume in van der Waals contact with the solvent. We show (Figure 8) that this model reproduces well Born radii computed from an accurate numerical representation of the SEV, noting that improvements for the Born radii of atoms in a loosely packed hydrophobic interior, while significant, are still not optimal. Although approximate, this representation of the SEV maintains the simplicity and computational efficiency of pairwise descreening schemes, while accounting for atomic overlaps in a consistent and parameter-free manner.

The new AGBNP2 model has been formulated to be employed in molecular dynamics conformational sampling applications, which require potential models of low computational complexity and favorable scaling characteristics, and with analytical gradients. These requirements have posed stringent constraints on the design of the model and the choice of the implementation algorithms. In the formulation of AGBNP2 we have reused as much as possible well-established and efficient algorithms developed earlier for the AGBNP1 model. For example, the key ingredient of the hydrogen bonding correction function is the free volume of a hydration site, which is computed using a methodology developed for AGBNP1 to compute atomic self-volumes. Similarly, the SEV-based pairwise descreening procedure employs atomic surface areas (see eq 31), computed as previously described.[42] AGBNP2 suffers additional computational cost associated with the SEV-based pairwise descreening procedure and the hydrogen bonding correction function. This handicap, however, is offset by having only one solute volume model in AGBNP2 rather than two in AGBNP1. AGBNP1 requires two separate invocations of the volume overlaps machinery (eq 2) for each of the two volume models it employs, corresponding to the van der Waals atomic radii for the pairwise descreening calculation and enlarged radii for the surface area calculation.[42] AGBNP2 instead employs a single volume model for both the pairwise descreening and surface area calculations. A direct CPU timing comparison between the two models cannot be reported at this time because the preliminary implementation of the AGBNP2 computer code used in this work lacks key data caching optimizations similar to those already employed in AGBNP1. Given the computational advantages of the new model discussed above, we expect to eventually obtain similar or better performance than with AGBNP1.

The AGBNP2 model has been parametrized against experimental hydration free energies of a series of small molecules and with respect to the ability of reproducing energetic and structural signatures of the conformational ensemble of three miniproteins generated with explicit solvation. These data sources are chosen so as to ensure that the resulting model would be applicable to both hydration free energy estimation and conformation equilibria, which are fundamental characteristics for models aimed at protein−ligand binding affinity estimation. On the other hand, experimental hydration free energies and explicit solvent conformational ensembles are to some extent incompatible with one another given the limitations of even the best fixed-charge force fields and explicit solvation models to reproduce experimental hydration free energies of small molecules with

high accuracy.[41,124,125] Mindful of this issue we did not seek a perfect correspondence with the experimental hydration free energy values. We first obtained parameters by fitting against the small molecule experimental hydration free energies and then adjusted the parameters to improve the agreement with the explicit solvent data, making sure that the predicted small molecule hydration free energies remained within a reasonable range relative to the experimental values. In practice this procedure yielded predicted hydration free energies in good agreement with the experimental values with the exception of the carboxylate and guanidinium ions (see Table 2), for which AGBNP2 predicts more favorable hydration free energies than the experiments, a consequence of the large hydrogen bonding corrections necessary to reduce the occurrence of intramolecular electrostatic interactions in the investigated proteins. As discussed further below, limitations in the description of hydration sites adopted for carboxylate and guanidinium ions may be partly the cause of the observed inconsistencies for these functional groups.

The parametrization and quantitative validation of the model, which is the primary focus of this work, has been based on comparing the effective potential energy distributions of implicit solvent conformational ensembles with those of explicit solvent ensembles. We observed that the AGBNP1 solvation model energetically ranked explicit solvent conformations significantly less favorably than implicit solvent conformations. The AGBNP2 model is characterized by smaller energetic bias relative to the explicit solvent ensembles, indicating that conformational sampling with the AGBNP2/OPLS-AA energy function produces conformations that more closely match those obtained with explicit solvation. This result is a direct consequence of employing the more realistic solvent excluded volume description of the solute, which yields larger Born radii for buried groups, as well as the hydrogen bonding to solvent corrections, which favor solvent exposed conformations of polar groups. Furthermore, comparison of the energy distributions of the AGBNP2 and explicit solvent ensembles for the three miniproteins (Figure 9D−F) shows, in contrast to the AGBNP1 results, that the AGBNP2 bias for the two more charge-rich miniproteins (cdp-1 and fsd-1) is similar to that of the least charged one (trp-cage). This suggests that the residual energetic bias of the AGBNP2 model is probably related to the nonpolar model rather than the electrostatic model. Future studies will address this particular aspect of the model.

The energy scoring studies conducted in this work indicate that AGBNP2 is a significant improvement over AGBNP1. They also show, however, that the new model falls short of consistently scoring explicit solvent conformations similarly to implicit solvent conformations. Although an optimal match between energy distributions is a necessary condition for complete agreement between implicit and explicit solvation results, it is unrealistic to expect to reach this ultimate goal at the present level of model simplification. Increasing the magnitude of the hydrogen bonding corrections can improve the agreement between the explicit and implicit solvation energy distributions, albeit at the expense of the quality of the predicted small molecule hydration free energies. It seems likely that the no parametrization of the current model would yield both good relative conformational free energies and hydration free energies. Future work will pursue the modeling of additional physical and geometrical features, such us the use of variable dielectric approaches to model polarization effects,[126] necessary to improve the agreement between implicit and explicit solvation energy distributions. The energy gap between the implicit solvent and explicit solvent energy distributions used here is, we believe, a meaningful measure of model quality and could serve as a useful general validation tool to compare the accuracy of implicit solvent models.

The excessive number of intramolecular electrostatic interactions involving charged groups has been one of the most noticeable shortcomings of GB-based implicit solvent models.[127] To correct this tendency, we have in the past adopted in the AGBNP1 model ad hoc strategies aimed at either destabilizing electrostatic intramolecular interactions[54] or, alternatively, stabilizing the competing solvent-separated conformations.[55] This work follows the latter approach using a more robust and physically motivated framework based on locating and scoring hydration sites on the solute surface as well as adopting a more realistic volume model. Structural characterization of the conformational ensembles has shown that AGBNP2 produces significantly fewer intramolecular interactions than AGBNP1, reaching good agreement with explicit solvent results. The reduction of intramolecular interactions has been the greatest for interactions between polar and charged groups. We believe the excessive tendency toward the formation of intramolecular interactions to be the root cause of the high melting temperatures of structured peptides[64] predicted with AGBNP1. Given the reduction of intramolecular interactions achieved with AGBNP2, we expect the new model to yield more reasonable peptide melting temperatures, a result which we hope to report in future publications.

Less-visible improvements have been obtained for ion pairs involving arginine side chains which remain more stable with implicit solvation than with explicit solvation. However, significantly, with AGBNP2 we observed a more dynamic equilibrium between ion-paired and solvent-separated conformations of arginine side chains that was not observed with AGBNP1. This result is promising because it indicates that the AGBNP2 solvation model, although still favoring ion-paired conformations, produces a more balanced equilibrium, which is instead almost completely shifted toward contact conformations with AGBNP1. Nevertheless it is apparent that the AGBNP treatment of the guanidinium group of arginine is not as good as for other groups. This limitation appears to be shared with other functional groups containing sp^2-hybridized nitrogen atoms as evidenced, for example, by the relatively lower quality of the hydration free energy predictions for amides and nitrogen-containing aromatic compounds (Table 2). Similar implicit solvent overstabilization solvation of arginine-containing ion pairs has been observed by Yu et al.[85] in their comparison of Surface Generalized Born (SGB) and SPC explicit solvation with the OPLS-AA force field. Despite quantitative differences, the explicit solvent studies (with the TIP3P water model) of Masunov and Lazaridis[128] and Hassan,[129] using the CHARMM force field, and that of Mandell at al.,[130] using the OPLS-AA force field, have confirmed that arginine forms the

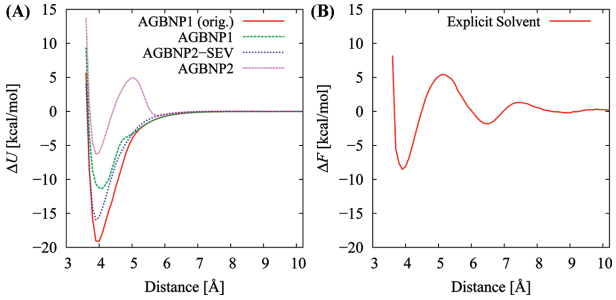

Figure 10. Potential of mean force of ion pair formation between propyl guanidinium and ethyl acetate in the coplanar orientation with AGBNP implicit solvation (A) and explicit solvation (B; ref 130). In (A) "AGBNP1 (orig.)" refers to the original AGBNP1 parametrization,[42] "AGBNP1" refers to the AGBNP1 model used in this work which includes a surface tension parameter correction for the carboxylate group aimed at reducing the occurrence of ion pairs,[55] "AGBNP2" refers to the current model, and "AGBNP2-SEV" refers to the current model without hydrogen bonding and surface tension corrections. The potentials of mean force are plotted with respect to the distance between the atoms of the protein side chain analogues equivalent to the Cζ of arginine and the Cγ of aspartate.

strongest ion pairing interactions, especially in the bidentate coplanar conformation. These observations are consistent with the present explicit solvent results using OPLS-AA and the SPC water model, where we find that most of the ion pairs of the miniproteins were found to involve arginine side chains. In contrast to our present implicit solvent results, however, the work of Masunov and Lazaridis[128] indicated that the GB-based implicit solvent model that they analyzed[14] produced potentials of mean force for arginine-containing ion pairs in general agreement with explicit solvation.

To rationalize the present implicit solvent results concerning ion pair formation, it has been instructive to analyze the potentials of mean force (PMFs) of ion pair association with the AGBNP model. As an example, Figure 10 shows the PMF for the association of propyl guanidinium (arginine side chain analogue) and ethyl acetate (aspartate and glutamate analogue) in a bidentate coplanar conformation (similar to the arrangement used previously)[85,128−130] for various AGBNP implementations. The corresponding explicit solvent PMF obtained by Mandell et al.[130] is also shown in Figure 10 for comparison. The original AGBNP1 parametrization[42] clearly leads to an overly stable salt bridge with the contact conformation scored at approximately −19 kcal/mol relative to the separated conformation, compared with −8.5 kcal/mol with explicit solvation. The AGBNP1 parametrization analyzed here, which includes an empirical surface area correction to reduce the occurrence of ion pairs,[55] yields a contact free energy (−11 kcal/mol) in much better agreement with explicit solvation, although the shape of the PMF is not properly reproduced. The present AGBNP2 model without hydrogen bonding corrections (labeled "AGBNP2-SEV" in Figure 10) yields a PMF intermediate between the original and corrected AGBNP1 parametrizations. The AGBNP2 model with hydrogen bonding corrections yields the PMF with the closest similarity to the one obtained in explicit solvent. Not only the contact free energy (−6.5 kcal/mol) is in good agreement with the explicit solvent result, but, importantly, it also reproduces the solvation barrier of the PMF at 5 Å separation, corresponding to the distance below which there is insufficient space for a water layer between the ions.

It is in this range of distances that the greatest discrepancies are observed between PMFs with explicit solvation and some GB-based implicit solvation models[85,128] that do not model effects due to the finite size of water molecules. Both the hydrogen bonding correction and the SEV volume description employed in AGBNP2, which are designed to take into account the granularity of the water solvent—the hydrogen bonding correction through the minimum free volume of water sites (eq 37) and the SEV model through the water radius offset (eq 28)—make it possible to reproduce the solvation barrier typical of molecular association processes in water.

It is notable in the PMF results shown in Figure 10 the lack of a free energy maximum with the AGBNP2/SEV model (AGBNP2 without HB corrections), which would be expected on the basis of results with the GBMV model, indicating that a SEV treatment of the GB model leads to a higher and much broader PMF maximum relative to explicit solvent.[88] There are two possible factors contributing to this discrepancy. The first is that the OPLS-AA force field used in this work seems to consistently produce stronger ionic interactions than the CHARMM force field (on which the GBMV model is based) as suggested by the relatively small free energies of salt bridge formation obtained with CHARMM-based implicit solvent models[88,128] relative to OPLS-AA-based ones (see for example ref 85 and the present results with AGBNP1). Because the shape of the PMF at intermediate separations is determined by a delicate balance between attractive electrostatic interactions and repulsive desolvation forces, stronger electrostatic interactions with OPLS-AA are potentially responsible in part for a missing or smaller PMF maximum. The lack of the PMF maximum with AGBNP2/SEV is most likely also due to the reduced radius offset used in AGBNP[42] used to construct the SEV. The small probe radius leads to a smaller reduction, compared to a full SEV treatment, of the high dielectric volume surrounding the ionic groups as they approach each other. The consequence is

a smaller rate of increase of the desolvation penalty which in turn leads to a smaller or absent PMF maximum.

Increasing the magnitude of the AGBNP radius offset is not feasible as we observed that the Gaussian overlap approximation for the overlap volumes (eq 10 of ref 42) breaks down for atomic radii much larger than the van der Waals radii. On the other hand, as the results in Figure 10 show, the added desolvation provided by the short-range HB function is able to properly correct this deficiency, yielding a PMF maximum in good correspondence with explicit solvation. This shows that the HB function as parametrized is likely taking into account not only short-range nonlinear hydration effects but also inaccuracies in the GB and nonpolar models, as well as approximations in the implementation such as the small probe radius discussed above.

The good correspondence between the AGBNP2 and explicit solvent PMFs for propyl guanidinium and ethyl acetate (Figure 10) stands in contrast with the residual AGBNP2 overprediction of arginine salt bridges compared to explicit solvation (Table 3). We observed that, in the majority of arginine salt bridges occurring with AGBNP2, the guanidinium and carboxylate groups interact at an angle rather than in the coplanar configuration discussed above. We have confirmed that the PMF of ion pair formation for an angled conformation (not shown) indeed shows a significantly more attractive contact free energy than the coplanar one. This result indicates that the in-plane placement of the hydration sites for the carboxylate groups (see the Appendix) does not sufficiently penalize angled ion pair arrangements. This observation is consistent with the need for introducing an isotropic surface area based hydration correction for carboxylate groups (the reduced γ parameter for the carboxylate oxygen atoms in Table 1), which showed some advantage in terms of reducing the occurrence of salt bridges. Future work will focus on developing a more general hydration shell description for carbonyl groups and related planar polar groups to address this issue.

5. Conclusions

We have presented the AGBNP2 implicit solvent model, an evolution of the AGBNP1 model we have previously reported, with the aim of incorporating hydration effects beyond the continuum dielectric representation. To this end a new hydration free energy component based on a procedure to locate and score hydration sites on the solute surface is used to model first solvation shell effects, such as hydrogen bonding, which are poorly described by continuum dielectric models. This new component is added to the generalized Born and nonpolar AGBNP models which have been improved with respect to the description of the solute volume description. We have introduced an analytical solvent excluded volume (SEV) model which reduces the effect of artifactual high dielectric interstitial spaces present in conventional van der Waals representations of the solute volume. The new model is parametrized and tested with respect to experimental hydration free energies and the results of explicit solvent simulations. The modeling of the granularity of water is one of the main principles employed in the design of the empirical first shell solvation function and the SEV model, by requiring that hydration sites have a minimum available volume based on the size of a water molecule. We show that the new volumetric model produces Born radii and surface areas in good agreement with accurate numerical evaluations. The results of molecular dynamics simulations of a series of miniproteins show that the new model produces conformational ensembles in much better agreement with reference explicit solvent ensembles than the AGBNP1 model with respect to both structural and energetics measures.

Future development work will focus on improving the modeling of some functional groups, particularly ionic groups involving sp^2 nitrogen, which we think are at the basis of the residual excess occurrence of salt bridges, and on the optimization of the AGBNP2 computer code implementation. Future work will also focus on further validation of the model on a wide variety of benchmarks including protein homology modeling and peptide folding.

Acknowledgment. This work was supported in part by National Institute of Health Grant GM30580. The calculations reported in this work have been performed at the BioMaPS High Performance Computing Center at Rutgers University funded in part by NIH shared instrumentation Grant 1 S10 RR022375.

Appendix A: Hydration Site Locations

Figure 11 shows the location of the hydration sites for the functional groups listed in Table 1. Each hydration site is represented by a sphere of 1.4 Å radius. The distance d_{HB} between the donor or acceptor heavy atom and the center of the hydration site sphere is set to 2.5 Å.

There is a single linear geometry for HB donor groups. The corresponding hydration site is placed at a distance d_{HB} from the heavy atom donor along the heavy atom−hydrogen bond.

Acceptor trigonal geometries have one or two hydration sites depending on whether the acceptor atom is bonded to, respectively, two or one other atom. In the former case the water site is placed along the direction given by the sum of the unit vectors corresponding to the sum of the NR_1 and NR_2 bonds (following the atom labels in Figure 11). In the latter case the W_1 site (see Figure 11) is placed in the R_1CO plane forming an angle of 120° with the CO bond. The W_2 site is placed similarly.

Acceptor tetrahedral geometries have one or two hydration sites depending on whether the acceptor atom is bonded, respectively, to three or two other atoms. In the former case the water site is placed along the direction given by the sum of the unit vectors corresponding to the sum of the NR_1, NR_2, and NR_3 bonds. In the latter case the positions of the W_1 and W_2 sites are given by

$$\mathbf{w}_1 = \mathbf{O} + d_{HB}(\cos\theta\mathbf{u}_1 + \sin\theta\mathbf{u}_2)$$

$$\mathbf{w}_2 = \mathbf{O} + d_{HB}(\cos\theta\mathbf{u}_1 - \sin\theta\mathbf{u}_2)$$

where $\mathbf{O}$ is the position of the acceptor atoms, $\theta = 104.4°$, and $\mathbf{u}_1$ and $\mathbf{u}_2$ are, respectively, the unit vectors corresponding to the OR_1 and OR_2 bonds.

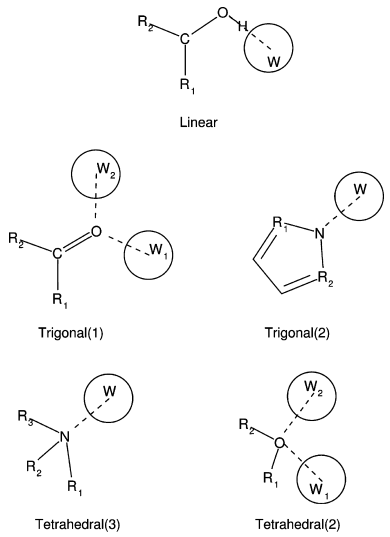

Figure 11. Diagram illustrating the hydration site locations for each of the functional group geometries used in this work. Linear, hydrogen bond donor; trigonal(1) and trigonal(2), trigonal planar geometries with, respectively, one and two covalent bonds on the acceptor atom; tetrahedral(2) and tetrahedral(3), tetrahedral geometries with, respectively, two and three covalent bonds on the acceptor atom. Representative molecular structures are shown for each geometry.

Appendix B: Gradients of GB and van der Waals Energies

The component of the gradient of the AGBNP2 van der Waals energy at constant self-volumes is the same as in the AGBNP1 model (see Appendix C of ref 42). In AGBNP2 the expression for the component of the gradient corresponding to variations in the atomic scaling factors, s_{ij}, includes pair corrections at all overlap levels because of the presence of multibody volumes in V''_{ij}. In addition, a new component corresponding to the change in surface areas appears:

$$\left(\frac{\partial \beta_j}{\partial \mathbf{r}_i}\right)_Q = -\frac{1}{4\pi}\sum_k \frac{\partial s_{kj}}{\partial \mathbf{r}_i} Q_{kj} = -\frac{1}{4\pi}\sum_k \frac{1}{V_k}\frac{\partial V'_k}{\partial \mathbf{r}_i} Q_{kj} \tag{39}$$

$$-\frac{1}{4\pi}\sum_k \frac{1}{V_k}\frac{\partial V'_{kj}}{\partial \mathbf{r}_i} Q_{kj} \tag{40}$$

$$+\frac{1}{4\pi}\sum_k \frac{1}{V_k} p_k \frac{\partial A_k}{\partial \mathbf{r}_i} Q_{kj} \tag{41}$$

Equation 39 leads to the same expression of the derivative component as in the AGBNP1 model (eq 72 in ref 42) (except for the extra elements in the two-body terms due to the inclusion of the $1/2V_{kj}$ correction term). Equation 40 corresponds to the component of the derivative due to variations in V'_{jk}, the volume to be added to the self-volumes of j and k to obtain the s_{jk} and s_{kj} scaling factors. In the AGBNP1 model this component included only two-body overlap volumes; in AGBNP2 this term instead includes all overlap volumes greater than zero. Finally, eq 41, where A_k is the surface area of atom k, leads to the component of the derivatives of the GB and vdW terms due to variations of the exposed surface area. The latter two terms are new for AGBNP2.

B.1. **Component of Derivative from eq 40.** From eq 63 in ref 42 and eq 40 we have

$$-4\pi\left(\frac{\partial \Delta G_{\mathrm{vdW}}}{\partial \mathbf{r}_i}\right)_{Q2} = \sum_{jk} W_{kj}\frac{\partial V'_{kj}}{\partial \mathbf{r}_i} \tag{42}$$

where W_{kj} has the same expression as in eq 69 in ref 42. In working with eq 42 it is important to note that, whereas V'_{kj} is symmetric with respect to swapping the j and k indices, W_{kj} and W_{jk} are different from each other. Substituting eq 30 into eq 42 and expanding over symmetric terms we obtain

$$-4\pi\left(\frac{\partial \Delta G_{\mathrm{vdW}}}{\partial \mathbf{r}_i}\right)_{Q2} = \frac{1}{2}\sum_{jk} W_{kj}\frac{\partial V_{kj}}{\partial \mathbf{r}_i} - \frac{1}{3}\sum_{jkl} W_{kj}\frac{\partial V_{jkl}}{\partial \mathbf{r}_i} + \frac{1}{2}\frac{1}{4}\sum_{jklp} W_{kj}\frac{\partial V_{jklp}}{\partial \mathbf{r}_i} - \ldots \tag{43}$$

Equation 43 is simplified by noting that

$$\frac{\partial V_{jk\ldots}}{\partial \mathbf{r}_i} = \delta_{ij}\frac{\partial V_{ik\ldots}}{\partial \mathbf{r}_i} + \delta_{ik}\frac{\partial V_{ji\ldots}}{\partial \mathbf{r}_i} + \ldots \tag{44}$$

Equation 44 is inserted in eq 43 and sums are reduced accordingly; then symmetric terms are collected into single sums by reindexing the summations, obtaining

$$-4\pi\left(\frac{\partial \Delta G_{\mathrm{vdW}}}{\partial \mathbf{r}_i}\right)_{Q2} = \frac{1}{2}\sum_j (W_{ij}+W_{ji})\frac{\partial V_{ij}}{\partial \mathbf{r}_i} - \frac{1}{3}\sum_{j<k}[(W_{ij}+W_{ji})+(W_{jk}+W_{kj})+(W_{ik}+W_{ki})]\frac{\partial V_{ijk}}{\partial \mathbf{r}_i} + \frac{1}{4}\sum_{j<k<l}[(W_{ij}+W_{ji})+(W_{ik}+W_{ki})+(W_{il}+W_{li})+(W_{jk}+W_{kj})+(W_{jl}+W_{lj})+(W_{kl}+W_{lk})]\frac{\partial V_{ijkl}}{\partial \mathbf{r}_i} - \ldots \tag{45}$$

The corresponding expression for the gradient of ΔG_{GB} is similar but employs the U_{ij} factors of eq 78 of ref 42 rather than W_{ij}.

B.2. **Component of Derivative from eq 41.** Inserting eq 41 in eq 63 of ref 42 gives

$$4\pi\left(\frac{\partial \Delta G_{\mathrm{vdW}}}{\partial \mathbf{r}_i}\right)_{Q3} = \sum_{jk} W_{kj} p_k \frac{\partial A_k}{\partial \mathbf{r}_i} = \sum_k W_k p_k \frac{\partial A_k}{\partial \mathbf{r}_i}$$

which is the same expression as that for the gradient of ΔG_{cav} (see Appendix A of ref 42) with the replacement

$$\gamma_k \rightarrow \frac{1}{4\pi} W_k p_k$$

The corresponding expression for the gradient of ΔG_{GB} follows from the substitution:

$$\gamma_k \rightarrow \frac{1}{4\pi} U_k p_k$$

B.3. Derivatives of HB Correction Energy. From eq 37 we have

$$\frac{\partial \Delta G_{\mathrm{hb}}}{\partial \mathbf{r}_i} = \sum_s h_s S'(w_s) \frac{\partial w_s}{\partial \mathbf{r}_i} \tag{46}$$

Inserting eqs 35 and 36 in eq 46 gives

$$\frac{\partial \Delta G_{\mathrm{hb}}}{\partial \mathbf{r}_i} = -\sum_{sj} \frac{h_s S'(w_s)}{V_s} \frac{\partial V_{sj}}{\partial \mathbf{r}_i} + \sum_{s,j<k} \frac{h_s S'(w_s)}{V_s} \frac{\partial V_{sjk}}{\partial \mathbf{r}_i} - \ldots \tag{47}$$

where

$$\frac{\partial V_{sjk\ldots}}{\partial \mathbf{r}_i} = \left(\frac{\partial V_{sjk\ldots}}{\partial \mathbf{r}_i}\right)_{\mathbf{r}_s} + \frac{\partial \mathbf{r}_s}{\partial \mathbf{r}_i} \frac{\partial V_{sjk\ldots}}{\partial \mathbf{r}_s} \tag{48}$$

where the first term on the right-hand side represents the derivative of the overlap volume with respect to the position of atom i keeping the position of the water site s fixed, and the second term reflects the change of overlap volume due to a variation of the position of the water site caused by a shift in position of atom i. The latter term is nonzero only if i is one of the parent atoms of the water site.

Supporting Information Available: Figure showing potential energy distributions of the AGBNP1 and explicit solvent conformational ensembles for the the trp-cage, cdp-1, and fsd-1 miniproteins scored with the AGBNP2-SEV/OPLS-AA and AGBNP2/OPLS-AA effective potentials; table listing experimental and AGBNP1 predicted hydration free energies of the set of small molecules in Table 2. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Levy, R. M.; Gallicchio, E. *Annu. Rev. Phys. Chem.* **1998**, *49*, 531–567.

(2) Feig, M.; Brooks, C. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217–224.

(3) Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1–20.

(4) Felts, A. K.; Andrec, M.; Gallicchio, E.; Levy, R. Protein Folding and Binding: Effective Potentials, Replica Exchange Simulations, and Network Models. In *Water and Biomolecules—Physical Chemistry of Life Phenomena*; Springer Science: New York, 2008.

(5) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517–529.

(6) Onufriev, A. *Annu. Rep. Comput. Chem.* **2008**, *4*, 125–137.

(7) Chen, J.; Brooks, C.; Khandogin, J. *Curr. Opin. Struct. Biol.* **2008**, *18*, 140–148.

(8) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027–2094.

(9) Baker, N. *Curr. Opin. Struct. Biol.* **2005**, *15*, 137–143.

(10) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrikson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–129.

(11) Zhang, L.; Gallicchio, E.; Friesner, R. A.; Levy, R. M. *J. Comput. Chem.* **2001**, *22*, 591–607.

(12) Schaefer, M.; Froemmel, C. *J. Mol. Biol.* **1990**, *216*, 1045–1066.

(13) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824–19839.

(14) Dominy, B. N.; Brooks, C. L. I. *J. Phys. Chem. B* **1999**, *103*, 3765–3773.

(15) Banks, J.; et al. *J. Comput. Chem.* **2005**, *26*, 1752–1780.

(16) Case, D. A.; Cheatham, T. E., III; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

(17) Ben-naim, A. *Hydrophobic Interactions*; Plenum Press: New York, 1980.

(18) Kauzmann, W. *Adv. Protein Chem.* **1959**, *14*, 1–63.

(19) Dill, K. A. *Biochemistry* **1990**, *29*, 7133–7155.

(20) Privalov, P. L.; Makhatadze, G. I. *J. Mol. Biol.* **1993**, *232*, 660–679.

(21) Honig, B.; Yang, A.-S. *Adv. Protein Chem.* **1995**, *46*, 27–58.

(22) Sturtevant, J. M. *Proc. Natl. Acad. Sci. U.S.A.* **1977**, *74*, 2236–2240.

(23) Williams, D. H.; Searle, M. S.; Mackay, J. P.; Gerhard, U.; Maplestone, R. A. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 1172–1178.

(24) Froloff, N.; Windemuth, A.; Honig, B. *Protein Sci.* **1997**, *6*, 1293–1301.

(25) Siebert, X.; Hummer, G. *Biochemistry* **2002**, *41*, 2965–2961.

(26) Ooi, T.; Oobatake, M.; Nemethy, G.; Sheraga, H. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 3086–3090.

(27) Lee, M. R.; Duan, Y.; Kollman, P. A. *Proteins* **2000**, *39*, 309–316.

(28) Hünenberger, P. H.; Helms, V.; Narayana, N.; Taylor, S. S.; McCammon, J. A. *Biochemistry* **1999**, *38*, 2358–2366.

(29) Simonson, T.; Brünger, A. T. *J. Phys. Chem.* **1994**, *98*, 4683–4694.

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

(31) Rapp, C. S.; Friesner, R. A. *Proteins: Struct., Funct., Genet.* **1999**, *35*, 173–183.

(32) Fogolari, F.; Esposito, G.; Viglino, P.; Molinari, H. *J. Comput. Chem.* **2001**, *22*, 1830–1842.

(33) Pellegrini, E.; Field, M. J. *J. Phys. Chem. A* **2002**, *106*, 1316–1326.

(34) Curutchet, C.; Cramer, C. J.; Truhlar, D. G.; Ruiz-Lòpez, M. F.; Rinaldi, D.; Orozco, M.; Luque, F. J. *J. Comput. Chem.* **2003**, *24*, 284–297.

(35) Jorgensen, W.; Ulmschneider, J.; Tirado-Rives, J. *J. Phys. Chem. B* **2004**, *108*, 16264–16270.

(36) Wallqvist, A.; Covell, D. G. *J. Phys. Chem.* **1995**, *99*, 13118–13125.

(37) Gallicchio, E.; Kubo, M. M.; Levy, R. M. *J. Phys. Chem. B* **2000**, *104*, 6271–6285.

(38) Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K. *J. Am. Chem. Soc.* **2003**, *25*, 9523–9530.

(39) Wagoner, J.; Baker, N. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8331–8336.

(40) Chen, J.; Brooks, C. *Phys. Chem. Chem. Phys.* **2008**, *10*, 471–481.

(41) Mobley, D.; Bayly, C.; Cooper, M.; Shirts, M.; Dill, K. *J. Chem. Theory Comput.* **2009**, *5*, 350–358.

(42) Gallicchio, E.; Levy, R. *J. Comput. Chem.* **2004**, *25*, 479–499.

(43) Wallqvist, A.; Gallicchio, E.; Levy, R. M. *J. Phys. Chem. B* **2001**, *105*, 6745–6753.

(44) Huang, D. M.; Chandler, D. *J. Phys. Chem. B* **2002**, *106*, 2047–2053.

(45) Zhou, R.; Huang, X.; Margulis, C.; Berne, B. *Science* **2004**, *305*, 1605–1609.

(46) Pierotti, R. A. *Chem. Rev.* **1976**, *76*, 717–726.

(47) Hummer, G.; Garde, S.; García, A. E.; Paulaitis, M. E.; Pratt, L. R. *J. Phys. Chem. B* **1998**, *102*, 10469–10482.

(48) Lum, K.; Chandler, D.; Weeks, J. D. *J. Phys. Chem. B* **1999**, *103*, 4570–4577.

(49) Pitarch, J.; Moliner, V.; Pascual-Ahuir, J.-L.; Silla, A.; Tuñón, I. *J. Phys. Chem.* **1996**, *100*, 9955–9959.

(50) Ashbaugh, H. S.; Kaler, E. W.; Paulaitis, M. E. *J. Am. Chem. Soc.* **1999**, *121*, 9243–9244.

(51) Pitera, J. W.; van Gunsteren, W. F. *J. Am. Chem. Soc.* **2001**, *123*, 3163–3164.

(52) Zacharias, M. *J. Phys. Chem. A* **2003**, *107*, 3000–3004.

(53) Su, Y.; Gallicchio, E. *Biophys. Chem.* **2004**, *109*, 251–260.

(54) Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 310–321.

(55) Felts, A.; Gallicchio, E.; Chekmarev, D.; Paris, K.; Friesner, R.; Levy, R. *J. Chem. Theory Comput.* **2008**, *4*, 855–858.

(56) Dong, F.; Wagoner, J.; Baker, N. *Phys. Chem. Chem. Phys.* **2008**, *10*, 4889–4902.

(57) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578–1599.

(58) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, C. W. *J. Phys. Chem. A* **1997**, *101*, 3005–3014.

(59) Tsui, V.; Case, D. A. *J. Am. Chem. Soc.* **2000**, *122*, 2489–2498.

(60) Schaefer, M.; Bartels, C.; Leclerc, F.; Karplus, M. *J. Comput. Chem.* **2001**, *22*, 1857–1879.

(61) Chekmarev, D.; Ishida, T.; Levy, R. *J. Phys. Chem. B* **2004**, *108*, 19487–19495.

(62) Andrec, M.; Felts, A. K.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6801–6806.

(63) Gallicchio, E.; Andrec, M.; Felts, A. K.; Levy, R. M. *J. Phys. Chem. B* **2005**, *109*, 6722–6731.

(64) Weinstock, D.; Narayanan, C.; Felts, A. K.; Andrec, M.; Levy, R.; Wu, K.; Baum, J. *J. Am. Chem. Soc.* **2007**, *129*, 4858–4859.

(65) Weinstock, D.; Narayanan, C.; Baum, J.; Levy, R. *Protein Sci.* **2008**, *17*, 950–954.

(66) Ravindranathan, K.; Gallicchio, E.; Levy, R. *J. Mol. Biol.* **2005**, *353*, 196–210.

(67) Messina, T.; Talaga, D. *Biophys. J.* **2007**, *93*, 579–585.

(68) Ravindranathan, K.; Gallicchio, E.; Friesner, R. A.; McDermott, A. E.; Levy, R. M. *J. Am. Chem. Soc.* **2006**, *128*, 5786–5791.

(69) Ravindranathan, P.; Gallicchio, E.; McDermott, A.; Levy, R. *J. Am. Chem. Soc.* **2007**, *129*, 474–475.

(70) Su, Y.; Gallicchio, E.; Das, K.; Arnold, E.; Levy, R. *J. Chem. Theory Comput.* **2007**, *3*, 256–277.

(71) Lapelosa, M.; Gallicchio, E.; Ferstandig Arnold, G.; Arnold, E.; Levy, R. M. *J. Mol. Biol.* **2009**, *385*, 675–691.

(72) Tjong, H.; Zhou, H. *J. Phys. Chem. B* **2007**, *111*, 3055–3061.

(73) Tjong, H.; Zhou, H. *J. Chem. Phys.* **2007**, *126*, 195102.

(74) Zhu, J.; Alexov, E.; Honig, B. *J. Phys. Chem. B* **2005**, *109*, 3008–3022.

(75) Fan, H.; Mark, A. E.; Zhu, J.; Honig, B. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6760–6764.

(76) Grant, J. A.; Pickup, B.; Sykes, M. J.; Kitchen, C.; Nicholls, A. *Phys. Chem. Chem. Phys.* **2007**, *9*, 4913–4922.

(77) Labute, P. *J. Comput. Chem.* **2008**, *29*, 1693–1698.

(78) Levy, R.; Belhadj, M.; Kitchen, D. *J. Chem. Phys.* **1991**, *95*, 3627–3633.

(79) Alper, H.; Levy, R. M. *J. Phys. Chem.* **1990**, *94*, 8401–8403.

(80) Morozov, A.; Kortemme, T. *Adv. Protein Chem.* **2005**, *72*, 1–38.

(81) Lazaridis, T.; Karplus, M. *Curr. Opin. Struct. Biol.* **2000**, *10*, 139–145.

(82) Eisenberg, D.; McLachlan, A. D. *Nature* **1986**, *319*, 199–203.

(83) Lazaridis, T.; Karplus, M. *Proteins* **1999**, *35*, 133–152.

(84) Vitalis, A.; Pappu, R. *J. Comput. Chem.* **2009**, *30*, 673–699.

(85) Yu, Z.; Jacobson, M.; Josovitz, J.; Rapp, C.; Friesner, R. *J. Phys. Chem. B* **2004**, *108*, 6643–6654.

(86) Okur, A.; Wickstrom, L.; Simmerling, C. *J. Chem. Theory Comput.* **2008**, *4*, 488–498.

(87) Lee, M. S.; Salsbury, F. R.; Brooks, C. L., III. *J. Chem. Phys.* **2002**, *116*, 10606.

(88) Swanson, J. M. J.; Mongan, J.; McCammon, J. A. *J. Phys. Chem. B* **2005**, *109*, 14769–14772.

(89) Lee, B.; Richards, F. *J. Mol. Biol.* **1971**, *55*, 379–400.

(90) Pascual-Ahuir, J. L.; Silla, E. *J. Comput. Chem.* **1990**, *11*, 1047–1060.

(91) Cortis, C. M.; Friesner, R. A. *J. Comput. Chem.* **1997**, *18*, 1591–1608.

(92) Rocchia, W.; Sridharan, S.; Nicholls, A.; Alexov, E.; Chiabrera, A.; Honig, B. *J. Comput. Chem.* **2002**, *23*, 128–137.

(93) Lee, M. S.; Feig, M.; Salsbury, F. R., Jr.; Brooks, C. L., III. *J. Comput. Chem.* **2003**, *24*, 1348–1356.

(94) Chocholousova, J.; Feig, M. *J. Comput. Chem.* **2006**, *27*, 719–729.

(95) Grant, J. A.; Pickup, B. T. *J. Phys. Chem.* **1995**, *99*, 3503–3510.

(96) Kratky, K. W. *J. Stat. Phys.* **1981**, *25*, 619–634.

(97) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(98) Jorgensen, W. L.; Madura, J. D. *Mol. Phys.* **1985**, *56*, 1381–1392.

(99) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.

(100) Mongan, J.; Simmerling, C.; McCammon, J.; Case, D.; Onufriev, A. *J. Chem. Theory Comput.* **2007**, *3*, 156–169.

(101) Liu, Y.; Liu, Z.; Androphy, E.; Chen, J.; Baleja, J. *Biochemistry* **2004**, *43*, 7421–7431.

(102) Dahiyat, B.; Sarisky, C.; Mayo, S. *J. Mol. Biol.* **1997**, *273*, 789–796.

(103) Dahiyat, B.; Mayo, S. *Science* **1997**, *278*, 82–87.

(104) Snow, C.; Zagrovic, B.; Pande, V. *J. Am. Chem. Soc.* **2002**, *124*, 14548–14549.

(105) Pitera, J.; Swope, W. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 7587–7592.

(106) Zhou, R. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13280–13285.

(107) Paschek, D.; Hempel, S.; García, A. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 17754–17759.

(108) Nosè, S. *J. Chem. Phys.* **1984**, *81*, 511–519.

(109) Hoover, W. *Phys. Rev. A* **1985**, *31*, 1695–1697.

(110) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

(111) Bowers, K.; Chow, E.; Xu, H.; Dror, R.; Eastwood, M.; Gregersen, B.; Klepeis, J.; Kolossváry, I.; Moraes, M.; Sacerdoti, F.; Salmon, J.; Shan, Y.; Shaw, D. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In *Proceedings of the ACM/IEEE Conference on Supercomputing (SC06)*; IEEE: Tampa, FL, 2006.

(112) Martyna, G.; Tobias, D.; Klein, M. *J. Comput. Phys.* **1994**, *101*, 4177–4189.

(113) Essman, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(114) Onufriev, A.; Case, D. A.; Bashford, D. *J. Comput. Chem.* **2002**, *23*, 1297–1304.

(115) Mobley, D.; Chodera, J.; Dill, K. *J. Phys. Chem. B* **2008**, *112*, 938–946.

(116) Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. *J. Solution Chem.* **1981**, *10*, 563–595.

(117) Vorobyov, I.; Li, L.; Allen, T. *J. Phys. Chem. B* **2008**, *112*, 9588–9602.

(118) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983–10990.

(119) Im, W.; Lee, M.; Brooks, C. *J. Comput. Chem.* **2003**, *24*, 1691–1702.

(120) Onufriev, A.; Bashford, D.; Case, D. *Proteins* **2004**, *55*, 383–394.

(121) Roe, D. R.; Okur, A.; Wickstrom, L.; Hornak, V.; Simmerling, C. *J. Phys. Chem. B* **2007**, *111*, 1846–1857.

(122) Yoshida, N.; Imai, T.; Phongphanphanee, S.; Kovalenko, A.; Hirata, F. *J. Phys. Chem. B* **2009**, *113*, 873–886.

(123) Miyata, T.; Hirata, F. *J. Comput. Chem.* **2008**, *29*, 871–882.

(124) Deng, Y.; Roux, B. *J. Phys. Chem. B* **2004**, *108*, 16567–16576.

(125) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 134508.

(126) Zhu, K.; Shirts, M.; Friesner, R. *J. Chem. Theory Comput.* **2007**, *3*, 2108–2119.

(127) Zhou, R.; Berne, B. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12777–12782.

(128) Masunov, A.; Lazaridis, T. *J. Am. Chem. Soc.* **2003**, *125*, 1722–1730.

(129) Hassan, S. *J. Phys. Chem. B* **2004**, *108*, 19501–19509.

(130) Mandell, D. J.; Chorny, I.; Groban, E.; Wong, S.; Levine, E.; Rapp, C.; Jacobson, M. *J. Am. Chem. Soc.* **2007**, *129*, 820–827.

CT900234U

References

Banks, J.L.; Beard, H.S.; Cao, Y.; Cho, A.E.; Damm, W.; Farid, R.; Felts, A.K.; Halgren, T.A.; Mainz, D.T.; Maple, J.R.; Murphy, R.; Philipp, D.M.; Repasky, M.P.; Zhang, L.Y.; Berne, B.J.; Friesner, R.A.; Gallicchio, E.; Levy, R.M. Integrated Modeling Program, Applied Chemical Theory (IMPACT). *J. Comp. Chem.* **2005**, *26*, 1752-1780.

Bernard, C.; Legros, C.; Ferrat, G.; Bischoff, U.; Marquardt, A.; Pongs, O.; Darbon, H. Solution structure of hpTX2, a toxin from Heteropoda venatoria spider that blocks Kv4.2 potassium channel. *Protein Sci*. **2000**, *9*, 2059-67.

Chagot, B.; Pimentel, C.; Dai, L.; Pil, J.; Tytgat, J.; Nakajima, T.; Corzo, G.; Darbon, H. Ferrat, G. An unusual fold for potassium channel blockers: NMR structure of three toxins from the scorpion Opisthacanthus madagascariensis. *Biochem. J*. **2005**, *388*, 263-71.

Dahiyat, B.I.; Mayo, S.L. De novo protein design: fully automated sequence selection. *Science* **1997**, *278*, 82-7.

Dahiyat, B.I.; Sarisky, C.A.; Mayo, S.L. De novo protein design: towards fully automated sequence selection. *J. Mol. Biol*. **1997**, *273*, 789-96.

Feig, M.; Brooks, C. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Op. Struct. Biol.* **2004**, *14*, 217-224.

Felts, A.K.; Gallicchio, E.; Wallqvist, A.; Levy, R.M. Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the surface generalized Born solvent model. *Proteins* **2002**, *48*, 404-422.

Felts, A.K.; Harano, Y.; Gallicchio, E.; Levy, R.M. Free energy surfaces of β-hairpin and α-helical peptides generated by replica exchange molecular dynamics with the AGBNP implicit solvent model. *Proteins* **2004**, *56*, 310-321.

Felts, A.K.; Andrec, M.; Gallicchio, E.; Levy, R. Protein folding and binding: Effective potentials, replica exchange simulations, and network models. In *Water and Biomolecules – Physical Chemistry of Life Phenomena*; Springer Science: 2008.

Felts, A.K.; Gallicchio, E.; Chekmarev, D.; Paris, K.A.; Friesner, R.; Levy, R. Prediction of Protein Loop Conformations using the AGBNP Implicit Solvent Model and Torsion Angle Sampling. *J. Chem. Theory Comput.* **2008**, *4*, 855-868.

Gallicchio, E.; Zhang, L.Y.; Levy, R.M. The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. *J. Comp. Chem.* **2002**, *23*, 517-529.

Gallicchio, E.; Levy, R.M. AGBNP: an analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *J. Comp. Chem.* **2004**, *25*, 479-499.

Jacobson, M.P.; Friesner, R.A.; Xiang, Z.; Honig, B. On the role of the crystal environment in determining side-chain conformations. *J. Mol. Biol.* **2002**, *320*, 597-608.

Jacobson, M.P.; Pincus, D.L.; Rapp, C.S.; Day, T.J.F.; Honig, B.; Shaw, D.E.; Friesner, R.A. A hierarchical approach to all-atom protein loop prediction. *Proteins* **2004**, *55*, 351-367.

Jorgensen, W.L.; Maxwell, D.S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225-11236.

Kaminski, G.A.; Friesner, R.A.; Tirado-Rives, J.; Jorgensen, W.L. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **2001**, *105*, 6474-6487.

Kohno, T.; Sasaki, T.; Kobayashi, K.; Fainzilber, M; Sato, K. Three-dimensional solution structure of the sodium channel agonist/antagonist delta-conotoxin TxVIA. *J. Biol. Chem.* **2002**, *277*, 36387-91.

Levy, R.M.; Gallicchio, E. Computer simulations with explicit solvent: recent progress in the thermodynamic decomposition of free energies and in modeling electrostatic effects. *Annu. Rev. Phys. Chem.* **1998**, *49*, 531-567.

Liu, Y.; Liu, Z.; Androphy, E.; Chen, J.; Baleja, J.D. Design and characterization of helical peptides that inhibit the E6 protein of papillomavirus. *Biochemistry* **2004**, *43*, 7421-31.

Manoleras, N.; Norton, R.S. Three-dimensional structure in solution of neurotoxin III from the sea anemone Anemonia sulcata. *Biochemistry* **1994**, *33*, 11051-61.

Roux, B.; Simonson, T. Implicit solvent models. *Biophys. Chem.* **1999**, *78*, 1-20.

Shenkin, P.S.; McDonald, D.Q. Cluster analysis of molecular conformations. *J. Comput. Chem.* **1994**, *15*, 899-916.

Trabi, M.; Craik, D.J. Tissue-specific expression of head-to-tail cyclized miniproteins in Violaceae and structure determination of the root cyclotide Viola hederacea root cyclotide1. *Plant Cell*. **2004**, *16*, 2204-16.

Yu, Z.; Jacobson, M.P.; Friesner, R.A. What role do surfaces play in GB models? A new-generation of surface-generalized born model based on a novel Gaussian surface for biomolecules. *J. Comput. Chem*. **2006**, *27*, 72-89.

PART II.

RECEPTOR REORGANIZATION AND LIGAND BINDING

Chapter 4

Introduction to Receptor Reorganization in Protein-Ligand Binding

4.1 Folding Funnels and Ligand Binding

Prediction of receptor-ligand affinities is one of the key tasks for computer-aided drug design. The overall affinity of a ligand for a receptor can be expressed as a balance between the strength of the interactions of the ligand for a particular binding-competent conformation of the receptor and the probability of occurrence of that conformation in the absence of a ligand. This concept can be seen in the proposed thermodynamic cycle for binding in Figure 4.1 where one portion of the cycle focuses on the interactions between

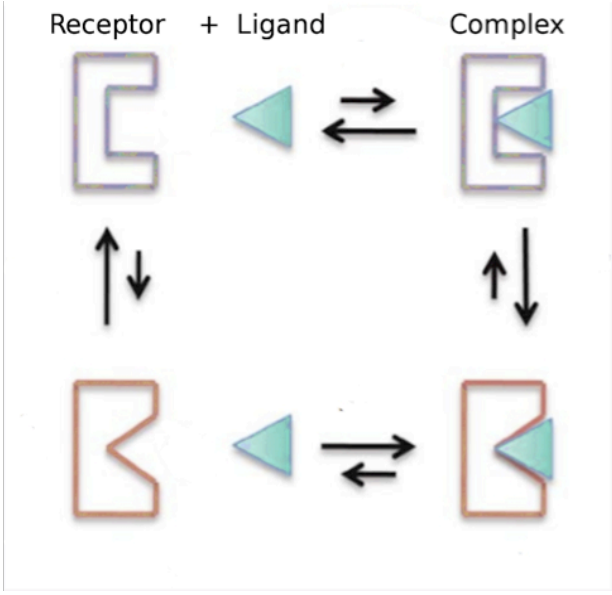

Figure 4.1. Thermodynamic cycle for binding. "Induced fit" starts in the top left and proceeds to the right where the ligand binds to the receptor and then incurs a conformational change in the receptor. Conformational selection starts in the top left and proceeds down where the receptor adopts a conformation to which the ligand binds.

the protein and ligand and another focuses on the conformational change of the receptor. Much work has been done on the former part of the problem of determining the strength of interactions between a ligand and receptor (Friesner et al., 2004; Halgren et al., 2004; Friesner et al., 2006; Ewing and Kuntz, 1997; Lang et al., 2009; Jones et al., 1995; Verdonk et al., 2008; Kramer et al., 1999; Jain, 2007; Venkatachalam et al., 2003; Zhou et al., 2007; Ferrara et al. 2004). The latter part of the problem has recently come back into focus with the idea of conformational selection (Boehr et al., 2009; Bakan and Bahar, 2009; Ma et al., 2002; Ma et al., 1999; Frauenfelder et al., 1991; Miller and Dill, 1997). Previously, ligand binding was often approached via either Fischer's "lock-and-key" model (Fischer, 1894) or Koshland's "induced fit" hypothesis (Koshland, 1958). In the "lock-and-key" model, the free and ligand-bound proteins have the same rigid conformation whereas in the "induced fit" model, the ligand induces a complementary conformational change in the protein. The conformational selection hypothesis approaches binding from a "folding funnel" point of view where protein folding is viewed as a parallel process where an ensemble of molecules goes downhill through an energy funnel (Dill and Chan, 1997; Lazaridis and Karplus, 1997; Becker and Karplus, 1997; Martinez et al., 1998; Onuchic et al., 1997; Ravindranathan et al., 2005). Folding funnels are rugged in the vicinity of the native fold of the protein, suggesting energetically competitive and similar conformations that provide an enhanced means of interactions between the protein and either ligands or other proteins. The model of conformational selection takes into account this rugged terrain and argues that ligand binding can shift the populations towards the weakly populated, higher energy conformations that are more suitable for binding. (Ma et al., 1999) Figure 4.2 gives a

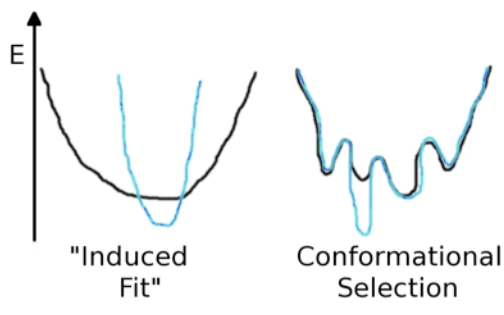

Figure 4.2. Cartoon receptor conformational landscapes for changes due to "induced fit" and conformational selection theories. Black lines represent the landscape prior to binding; blue lines represent the landscape after binding. It the "induced fit" model, binding causes a conformational change in the receptor. In the conformational selection model, binding causes a population shift, deepening a well that was not previously as populated. Both models are thought to play roles in binding (Boehr et al., 2009).

cartoon before and after "landscape" comparison of induced fit and conformational selection. Both conformational selection and induced fit appear to play roles in ligand binding (Boehr et al., 2009; Bakan and Bahar, 2009).

4.2 Receptor Reorganization in Ligand Binding

As shown above, receptor reorganization can potentially be an important part of the measure of the affinity of a ligand for that particular receptor. Receptors that undergo little to no conformational change upon binding can be handled in a "lock-and-key" fashion where the receptor is held rigid as a ligand is docked. Receptors that do undergo conformational change upon binding may require inclusion of receptor reorganization or strain energy to properly model the binding of ligands to that protein. Many medically relevant receptors undergo conformational changes upon binding, including several of the human immunodeficiency viral enzymes as well as a variety of kinases that have been implicated in certain cancers and other diseases. There is no cookbook recipe for modeling receptor reorganization in ligand binding and several methods have been attempted. These include MD and MC methods (Armen et al., 2009; Cheng et al., 2008;

Bowman et al., 2007; Carlson, 2002; Hart and Read, 1992; Dixon and Oshiro, 1995), use of rotamer libraries (Schaffer and Verkhivker, 1998; Desmet et al., 1992; Leach, 1994; Trosset and Sheraga, 1999), protein ensemble docking (Armen et al., 2009; Totrov and Abagyan, 2008; Knegtel et al., 1997; Ferrari et al., 2004; Claussen et al., 2001), and soft-receptor modeling (Ferrari et al., 2004; Osterberg et al., 2002; Knegtel et al., 1997). MD and MC methods can be computationally expensive and have the drawback of potentially introducing significant error and "noise" that could decrease docking accuracy. Methods based on rotamer libraries represent the receptor as a set of experimentally observed and preferred rotameric states for side chains that surround the binding pocket. However, this technique does not include backbone flexibility. Ensemble docking methods, where the ligand is docked to an ensemble of receptors with varying structures, have been explored but some studies have shown that docking to an ensemble may give worse results than rigid docking (Polgár and Keserü, 2006; Barril and Morley, 2005). Soft-receptor modeling combines information from several protein conformations to generate a single weighted average grid to which the ligand is docked. Another version of "soft" docking employs reduced van der Waals radii or deleting side chains of residues predicted to be flexible, thus potentially eliminating close contacts (Carlson and McCammon, 2000). A study in 2006 rather successfully combined the "soft" docking technique with iterations of rigid receptor docking using reduced vdW radii and protein structure prediction techniques (Sherman et al., 2006). However "soft" techniques are not able to handle large changes in conformation.

The subsequent chapters delve into several plausible methods for modeling receptor reorganization through clustering, simulation and QSAR-like techniques.

References

Armen, R.S.; Chen, J.; Brooks, C.L. III. An evaluation of explicit receptor flexibility in molecular docking using molecular dynamics and torsion angle molecular dynamics. *J. Chem. Theory Comput*. **2009**, *5*, 2909-2923.

Bakan, A.; Bahar, I. The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proc. Natl. Acad. Sci.* **2009**, *106*, 14349-14354.

Barril, X.; Morley, S.D. Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. *J. Med. Chem.* **2005**, *48*, 4432– 4443.

Becker, O.M.; Karplus, M. The topology of multidimensional potential energy surfaces: theory and application to peptide structure and kinetics. *J. Chem. Phys.* **1997**, *106*, 1495-1517.

Boehr, D.D.; Nussinov, R.; Wright, P.E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol*. **2009**, *5*, 789-96.

Bowman, A.L.; Nikolovska-Coleska, Z.; Zhong, H.; Wang, S.; Carlson, H.A. Small molecule inhibitors of the MDM2-p53 interaction discovered by ensemble-based receptor models. *J. Am. Chem. Soc.* **2007**, *129*, 12809–12814.

Carlson, H.A.; McCammon, J.A. Accommodating protein flexibility in computational drug design. *Mol. Pharmacol.* **2000**, *57*, 213-218.

Carlson, H.A. Protein flexibility and drug design: how to hit a moving target. *Curr. Opin. Chem. Biol.* **2002**, *6*, 447-452.

Cheng, L.S.; Amaro, R.E.; Xu, D.; Li, W.W.; Arzberger, P.W. McCammon, J.A., Ensemble-based virtual screening reveals potential novel antiviral compounds for avian influenza neuraminidase. *J. Med. Chem.* **2008**, *51*, 3878–3894.

Claussen, H; Buning, C.; Rarey, M; Lengauer, T. FlexE: Efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **2001**, *308*, 377-395.

Desmet, J.; Maeyer, M.D.; Hazes, B.; Lasters, I. The dead end elimination theorem and its use in protein side-chain positioning. *Nature* **1992**, *356*, 539 –542.

Dill, K.A.; Chan, H.S. From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **1997**, *4*, 10-19.

Ewing, T.J.A.; Kuntz, I.D. Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem.* **1997**, *18*, 1175-1189.

Ferrara, P.; Gohlke, H.; Price, D.J.; Klebe, G.; Brooks, C.L. III. Assessing scoring functions for protein-ligand interactions. *J. Med. Chem*. **2004**, *47*, 3032-47.

Ferrari, A.M.; Wei, B.Q.; Costantino, L.; Shoichet, B.K. Soft docking and multiple receptor conformations in virtual screening. *J. Med. Chem.* **2004**, *47*, 5076-5084.

Fischer, E. Einfluss der configuration auf die wirkung der enzyme. *Ber. Dtsch. Chem.* **1894**, *27*, 2984-2993.

Frauenfelder, H.; Sligar, S.G.; Wolynes, P.G. The energy landscapes and motions of proteins. *Science* **1991**, *254*, 1598-1603.

Friesner, R.A.; Banks, J.L.; Murphy, R.B.; Halgren, T.A.; Klicic, J.J.; Mainz, D.T.; Repasky, M.P.; Knoll, E.H.; Shelley, M.; Perry, J.K.; Shaw, D.E.; Francis, P.; Shenkin, P.S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem*. **2004**, *47*, 1739–1749.

Friesner, R.A.; Murphy, R.B.; Repasky, M.P.; Frye, L.L.; Greenwood, J.R.; Halgren, T.A.; Sanschagrin, P.C.; Mainz, D.T. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem*. **2006**, *49*, 6177-96.

Halgren, T.A.; Murphy, R.B.; Friesner, R.A.; Beard, H.S.; Frye, L.L.; Pollard, W.T.; Banks, J.L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem*. **2004**, *47*, 1750–1759.

Hart, T.N.; Read, R.J. A multiple-start Monte Carlo docking method. *Proteins* **1992**,*13*, 206 –222.

Jain, A.N. Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J. Comput. Aided Mol. Des*. **2007**, *21*, 281-306.

Jones, G.; Willett, P.; Glen, R.C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43-53.

Knegtel, R.M.A.; Kuntz, I.D.; Oshiro, C.M. Molecular docking to ensembles of protein structures. *J. Mol. Biol.* **1997**, *266*, 424 – 440.

Koshland, D.E. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. USA* **1958**, *44*, 98-104.

Kramer, B.; Rarey, M.; Langauer, T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins* **1999**, *37*, 228-241.

Lang, P.T., Brozell, S.R., Mukherjee, S., Pettersen, E.F., Meng, E.C., Thomas, V., Rizzo, R.C., Case, D.A., James, T.L., and Kuntz, I.D. DOCK 6: Combining techniques to model RNA-small molecule complexes. *RNA* **2009**, *15*, 1219-1230.

Lazaridis, T.; Karplus, M. "New view" of protein folding reconciled with the old through multiple unfolding simulations. *Science* **1997**, *278*, 1928-1931.

Leach, A.R. Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.* **1994**, *235*, 345–356.

Ma, B.; Kumar, S.; Tsai, C.J.; Nussinov, R. Folding funnels and binding mechanisms. *Protein Eng.* **1999**, *12*, 713-720.

Ma, B.; Shatsky, M.; Wolfson, H.J.; Nussinov, R. Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. *Protein Sci.* **2002**, *11*, 184-197.

Martinez, J.C.; Pisabarro, M.T.; Serrano, L. Obligatory steps in protein folding and the conformational diversity of the transition state. *Nat. Struct. Biol.* **1998**, *5*, 721-729.

Miller, D.W.; Dill, K.A. Ligand binding to proteins: the binding landscape model. *Protein Sci.* **1997**, *6*, 2166-2179.

Onuchic, J.N.; Luthey-Schulten, Z.; Wolynes, P.G. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545-600.

Oshiro, C.M.; Kuntz, I.D.; Dixon, J.S. Flexible ligand docking using a genetic algorithm. *J. Comput. Aided Mol. Des.* **1995**, *9*, 113–130.

Osterberg, F.; Morris, G.M.; Sanner, M.F.; Olson, A.J.; Goodsell, D.S. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins* **2002**, *46*, 34 – 40.

Polgár, T.; Keserü, G.M. Ensemble docking into flexible active sites. Critical evaluation of FlexE against JNK-3 and β-Secretase. *J. Chem. Inf. Model.* **2006**, *46*, 1795-1805.

Ravindranathan, K.P.; Gallicchio, E.; Levy, R.M. Conformational equilibria and free energy profiles for the allosteric transition of the ribose-binding protein. *J Mol Biol* **2005**, *353*, 196-210.

Schaffer, L.; Verkhivker, G.M. Predicting structural effects in HIV-1 protease mutant complexes with flexible ligand docking and protein side-chain optimization. *Proteins* **1998**, *33*, 295–310.

Sherman, W.; Day, T.; Jacobson, M.P.; Friesner, R.A.; Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* **2006**, *49*, 534-553.

Totrov, M.; Abagyan, R. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr. Opin. Struct. Biol.* **2008**, *18*, 178-184.

Trosset, J.Y.; Scheraga, H.A. Prodock: software package for protein modeling and docking. *J. Comput. Chem.* **1999**, *20*, 412– 427.

Venkatachalam, C.M.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graph. Model.* **2003**, *21*, 289-307.

Verdonk, M.L.; Mortenson, P.N.; Hall, R.J.; Hartshorn, M.J.; Murray, C.W. Protein-ligand docking against non-native protein conformers. *J. Chem. Inf. Model*. **2008**, *48*, 2214-25.

Zhou, Z.; Felts, A.K.; Friesner, R.A.; Levy, R.M. Comparative performance of several flexible docking programs and scoring functions: enrichment studies for a diverse set of pharmaceutically relevant targets. *J. Chem. Inf. Model*. **2007**, *47*, 1599-608.

Chapter 5

Introduction to the Human Immunodeficiency Virus and Reverse Transcriptase

5.1 The Human Immunodeficiency Virus

Human immunodeficiency virus (HIV) infection has reached near-global pandemic proportions in the past few years. At the end of 2008, the Joint United Nations Programme in HIV/AIDS (UNAIDS) and the World Health Organization (WHO) reported 31.1-35.8 million infections worldwide with 1.7-2.4 million deaths (UNAIDS and WHO, 2009). There is no cure or vaccine for HIV, only medication that can slow the process of acquired immune deficiency syndrome (AIDS) formation and, eventually, death. AZT, which was introduced in 1987, was the first drug to show some success in combating HIV infection. Significant chemotherapeutic progress was not achieved until 1996 with the advent of combination therapy, Highly Active Antiretroviral Therapy (HAART) (Kaufman and Cooper, 2000). Though further improvements have made the dosing more tolerable, HAART is only palliative, it is not globally available, and serious side-effects are common (Kallings, 2008). Coupled with the lack of success in vaccine development (Walker and Burton, 2010), it is essential to seek new anti-HIV agents that feature efficacy against a broad spectrum of HIV variants, low cost, easy storage and administration, and reduced side effects. Therefore, the ability to discover drugs for HIV treatment is at the forefront of scientific research and combines a number of fields including chemistry, biology and computer science.

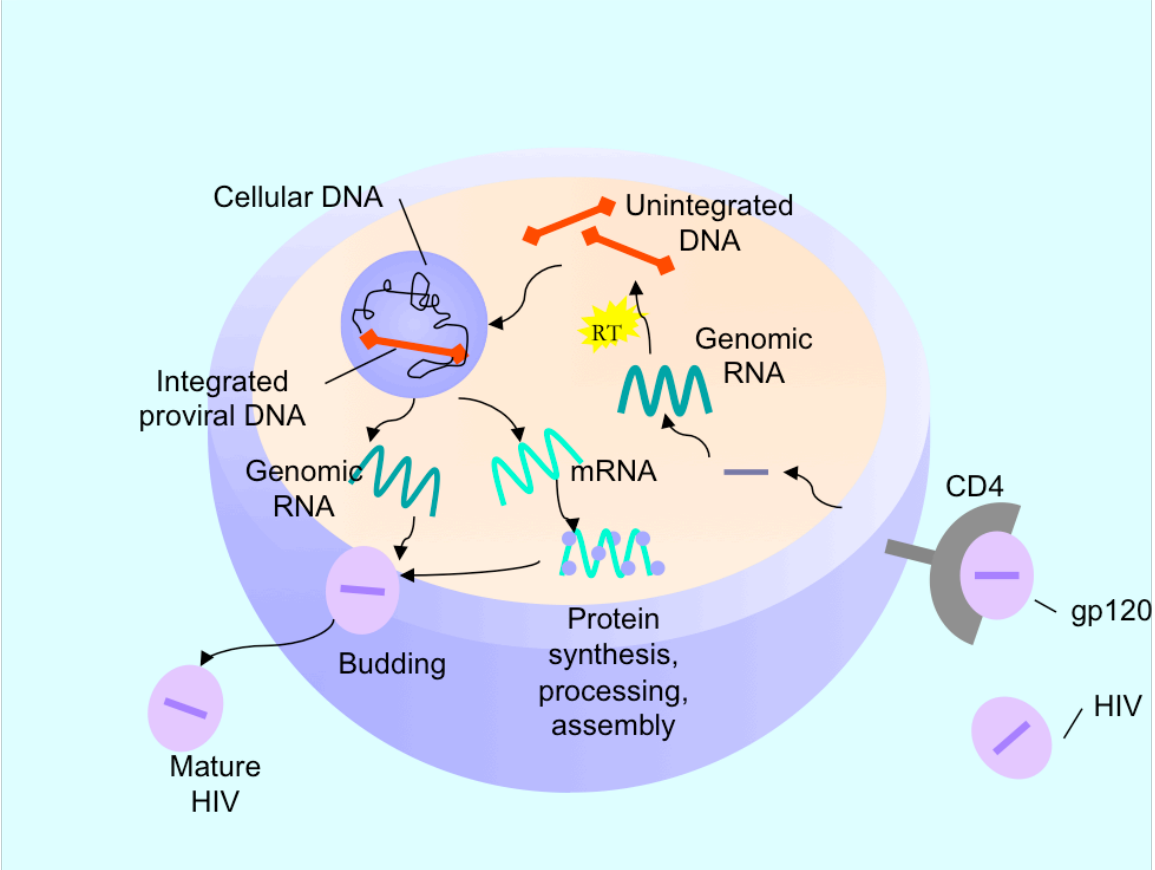

Figure 5.1 Life cycle of HIV-1.

HIV is a retrovirus that attacks the human immune system by infecting $CD^{+}4$ T cells (white blood cells), microphages, and dendrite cells. The decrease in $CD^{+}4$ T cells marks progression to AIDS, a state in which the human body can no longer defend itself from disease. Study of the life cycle of the virus has led to several pathways to control the replication that leads to infection of additional cells. Figure 5.1 outlines the viral replication cycle. The virus enters the cell via a pathway thought to be initiated by the interaction of the HIV glycoprotein gp120 and CD4 on the target cell and the injection of the viral genomic material into the target cell. Once inside the cell, an enzyme called reverse transcriptase (RT) liberates the single strand RNA from the viral proteins and copies it, creating a vDNA that is transported into the cell nucleus. Integration of the

proviral DNA into the host genome is carried out by the viral enzyme integrase, leading to what is called the latent stage of HIV infection. HIV provirus may lie dormant within a cell for a long time. When the cell becomes activated, it treats HIV genes in much the same way as human genes: it converts them into copies of the HIV genome and messenger RNA (mRNA) using the host cell's RNA polymerase. The mRNA is transported outside the nucleus, where it is used as a blueprint to make HIV proteins, including Tat, Rev, Gag and Env. The HIV enzyme protease cuts the HIV proteins into smaller individual proteins that come together with copies of the viral RNA to assemble a new virion particle. The newly assembled virion pushes out ("buds") from the host cell, "stealing" part of the cell's outer envelope which is studded with the HIV glycoproteins gp41 and gp120 that are necessary for the virus to bind CD4 on a new host cell. As would be expected, the cleavage of the bud from the $CD^{+}4$ T cell causes cell death. (Zheng et al., 2005)

Hindrance of the replication process can be achieved at several steps in the viral life cycle. Entry and fusion inhibitors (EIs and FIs) affect the entry of the virus into the cell. Nucleoside reverse transcriptase inhibitors (NRTIs) and non-nucleoside reverse transcriptase inhibitors (NNRTIs) slow the replication of the viral RNA. Integrase inhibitors (InIs) target the splicing of the viral genomic material into the host cell DNA. Finally, the formation and maturation of the new virion can be inhibited by protease inhibitors (PIs). Typical therapies require a combination of different drug inhibitors in what is called HAART (Kaufman and Cooper, 2000). Currently there are 26 approved antiretroviral drugs of which nine are NRTIs, four are NNRTIs, 10 are PIs, one is an EI, one is an InI and one is an FI. 12 inhibitors are also currently under investigation.

5.2 Inhibition of HIV-1 Reverse Transcriptase

The HIV-1 viral enzyme reverse transcriptase is a heterodimer composed of subunits p66 and p51. They share a common sequence although p51 is truncated. The p66 subunit holds two domains: an N-terminal polymerase domain of ~440 residues and a C-terminal RNase H domain of ~120 residues. The N-terminal polymerase domain is often described as a "right hand" with fingers, thumb, and palm subdomains. A connection subdomain connects the "hand" with the RNase H domain (Kohlstaedt et al., 1992; Jacobo-Molina et al., 1993; Das et al., 1996). Nucleic acid binds in a cleft that measures 17-18 base pairs in length and is situated in the palm subdomain that extends from the polymerase active site (defined by the active site catalytic residues D185, D186, D110) to the RNase H active site. A conformational change, depicted in Figure 5.2, occurs upon binding nucleic acid: the thumb and fingers move to "clasp" the nucleic acid (Jacobo-Molina et al., 1993; Ding et al., 1998). The enzyme builds a DNA strand based on the viral RNA at the polymerase active site. The original RNA strand is then cleaved into pieces at the RNase H active site at the opposite end of the enzyme. Finally, a second DNA strand matched to the one that was just created is built at the polymerase active site to form the final DNA double helix. Both the viral RNA and the single-strand viral DNA are believed to be guided by the polymerase primer grip region of the enzyme. The primer grip, situated in the palm subdomain, has been proposed to be essential for positioning the primer 3' terminus at the polymerase active site (Jacobo-Molina et al., 1993) and movement of the primer grip and associated thumb subdomain are thought to be critical for the translocation of nucleic acid following incorporation of nucleotides during polymerization (Tantillo et al., 1994). HIV replicates approximately every two

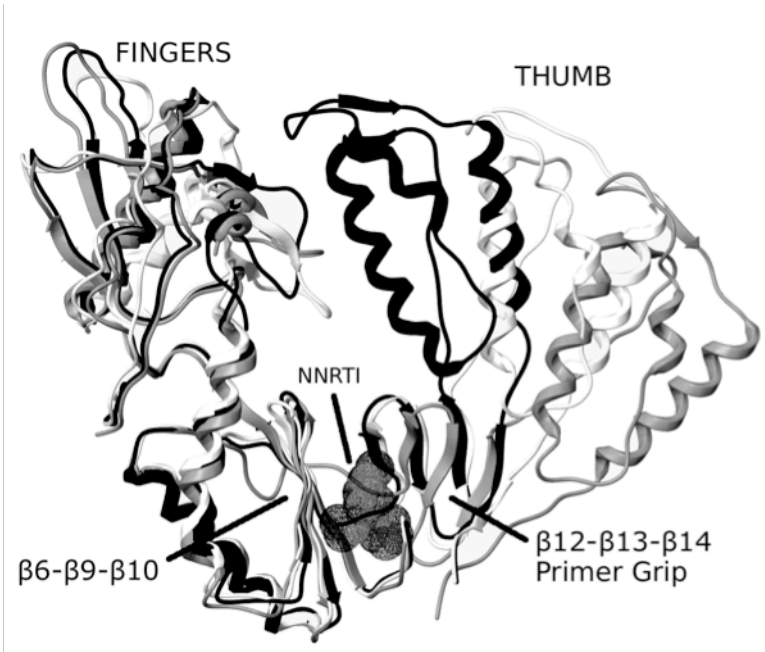

Figure 5.2. Rearrangement of HIV-1 RT upon binding. Upon binding substrates such as DNA or NNRTIs such as nevirapine, RT undergoes a global conformational change in the thumb of the enzyme. Upon binding an NNRTI, there is also a shift in the primer grip region shown in the movement from black/light gray to dark gray. Black: Unbound RT (PDB ID 1DLO); Light gray: RT bound to DNA (PDB ID 1RTD); Dark gray: RT bound to the NNRTI nevirapine (PDB ID 1VRT).

days and this, combined with the high mutation rate (the error rate per nucleotide is between 1 in 10^4 and 1 in 10^5 errors per base pair per cycle), results in production of a large number of viruses that vary by one or more nucleotides (Coffin, 1995).

Currently, there are two types of inhibitors that target RT: nucleoside inhibitors (NRTIs) and non-nucleoside inhibitors (NNRTIs). NRTIs are competitive inhibitors that mimic normal nucleotides but lack the 3'-OH required for elongation and thus terminate chain elongation by preventing addition of more nucleotides. NNRTIs are non-competitive, specific inhibitors that bind to a pocket called the non-nucleoside reverse transcriptase binding pocket (NNIBP), which lays approximately 10 Å from the enzyme's polymerase active site (Kohlstaedt et al., 1992). The NNIBP undergoes large structural rearrangements upon binding of an NNRTI where the aromatic side chains Y181 and

Y188 swivel and the primer grip region moves to create space for the ligand (Hsiou et al., 1996; Das et al., 2007). It can be defined as being bounded by the "YMDD loop"-containing β6-β9-β10 sheet (which also contains the polymerase active site) and the primer grip-containing β12-β13-β14 sheet.

Analysis of crystal structures has suggested three possible mechanisms of inhibition (which may or may not work in conjunction) of HIV-1 RT by NNRTIs: 1) restriction of the p66 thumb ("molecular arthritis"); 2) distortion of catalytically essential residues at the polymerase active site; and 3) displacement of the primer grip. In each of these mechanisms, the binding of NNRTIs is proposed to lead to conformational perturbations that, in turn, limit conformational flexibility required for efficient DNA synthesis by RT. In the "molecular arthritis" mechanism, conformational restriction of the p66 thumb subdomain has been suggested to limit flexibility of the enzyme required for catalysis (Kohlstaedt et al., 1992). NNRTI binding may restrict the mobility of the thumb subdomain (Kohlstaedt et al., 1992; Tantillo et al., 1994; Shen et al., 2003) or may change the direction of movement of the thumb subdomain (Temiz and Bahar, 2002), thus slowing down or preventing template-primer translocation and inhibiting facile elongation of nascent viral DNA. Binding of an NNRTI causes perturbation of the configuration of the RT polymerase active site region, including the catalytically essential D110, D185, and D186 residues (Esnouf et al., 1995), and limits conformational changes of the "YMDD loop" containing M184 and D185 (Ding et al.,1998), both of which are believed to be important in the activity of the enzyme. The conserved primer grip, one of the boundaries defining the NNIBP, is a structural element in HIV-1 RT that has been proposed to be essential for positioning the primer 3' terminus at the active site (Jacobo-

Molina et al., 1993), and movements of the primer grip and the associated thumb subdomain are thought to be critical for the translocation of nucleic acid following incorporation of nucleotides during polymerization (Tantillo et al., 1994). NNRTI binding causes a significant displacement (~4 Å) of the primer grip, leading to possible inappropriate positioning of the primer terminus at the active site; this conformational alteration and possible restriction of primer grip mobility may be a major contributor to inhibition by NNRTIs (Das et al., 1996). Movement in the primer grip is also thought to affect allosteric hinge-bending movements in the position of the thumb subdomain (the tip of which lies ~30 Å from the NNIBP) (Kohlstaedt et al., 1992; Ding et al., 1995; Hsiou et al., 1996). Figure 5.2 shows structures superimposed based on the β6-β9-β10 strands where the thumb subdomain of NNRTI-bound RT is rotated by ~ 40° relative to that in the unbound apo enzyme (Hsiou et al., 1996; Rodgers and Harrison, 1995).

The NNIBP is very flexible, changing conformation when different NNRTIs are bound in an effort to optimize stabilizing interactions with the ligand (Das et al., 2004; Das et al., 2008). Although available inhibitors have different shapes, sizes, functional groups, and binding modes, they display a number of common features in their interactions with the NNIBP residues: aromatic ring(s) capable of forming π-π stacking interactions with aromatic residues as well as making hydrophobic contacts with other nonpolar pocket residues, and (usually) hydrogen bond (H-bond) donors capable of forming an H-bond with the backbone carbonyl oxygen of K101. This is exemplified in Figure 5.3 that depicts several NNRTIs that are currently used or in drug trials.

As replication of the HIV genome is imperfect (Coffin, 1995), mutation within the NNIBP is very common and is a major concern in drug development. Commonly

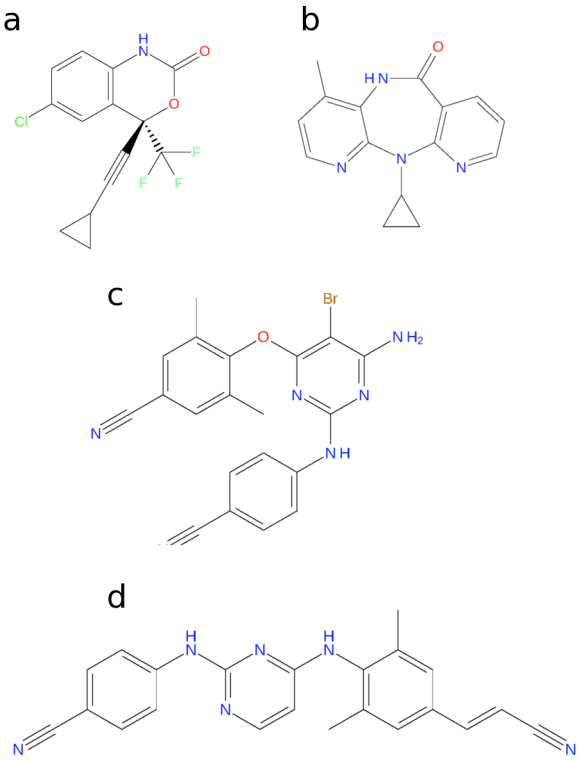

Figure 5.3. Several NNRTIs used for HIV treatment or in clinical trials. (a) efavirenz; (b) nevirapine; (c) etravirine (TMC-125) (d) rilpivirine (TMC-278).

observed mutations include those that reduce the favorable π-π stacking interactions (e.g. Y188C/L and Y181C), cause steric interference within the NNIBP with bulkier side chains (e.g. L100I, V108I, G190A/S), cause indirect allosteric conformational changes within the binding pocket (V106A/I) and stabilize the “closed” apo form of the enzyme (K103N). (Das et al., 2004; Das et al., 2005; Lewi et al., 2003) Current design of inhibitors focuses on developing the "required" interactions in such a way to ensure high activity, reasonable solubility, and broad potency against drug-resistant variants (Jorgensen, 2009; Jorgensen et al., 2006; Barreca et al., 2007; Wolber and Langer, 2005).

References

AIDS epidemic update; UNAIDS and WHO, WHO Library Cataloguing-in-Publication Data: Geneva, Switzerland, 2009.

Barreca, M.L.; De Luca, L.; Iraci, N.; Rao, A.; Ferro, S.; Maga, G.; Chimirri, A. Structure-based pharmacophore identification of new chemical scaffolds as non-nucleoside reverse transcriptase inhibitors. *Bioorg. Med. Chem. Lett*. **2007**, *17*, 1956-60

Coffin, J.M. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science* **1995**, *267*, 483-489.

Das, K.; Ding, J.; Hsiou, Y.; Clark, A.D., Jr.; Moereels, H.; Koymans, L,; Andries, K.; Pauwels, R.; Janssen, P.A.J.; Boyer, P.L.; Clark, P.; Smith, R.H., Jr.; Kroeger Smith, M.B.; Michejda, C.J.; Hughes, S.H.; Arnold, E. Crystal structures of 8-Cl and 9-Cl TIBO complexed with wild-type HIV-1 RT and 8-Cl TIBO complexed with the Tyr181Cys HIV-1 RT drug-resistant mutant. *J. Mol. Biol.* **1996**, *264*, 1085-1100.

Das, K.; Clark, A.D., Jr.; Lewi, P.J.; Heeres, J.; De Jonge, M.R.; Koymans, L.M.; Vinkers, H.M.; Daeyaert, F.; Ludovici, D.W.; Kukla, M.J.; De Corte, B.; Kavash, R.W.; Ho, C.Y.; Ye, H.; Lichtenstein, M.A.; Andries, K.; Pauwels, R.; De Béthune, M.P.; Boyer, P.L.; Clark, P.; Hughes, S.H.; Janssen, P.A.; Arnold, E. Roles of conformational and positional adaptability in structure-based design of TMC125-R165335 (etravirine) and related non-nucleoside reverse transcriptase inhibitors that are highly potent and effective against wild-type and drug-resistant HIV-1 variants. *J. Med. Chem.* **2004**, *47*, 2550-2560.

Das, K,; Lewi, P.J.; Hughes, S.H.; Arnold, E. Crystallography and the design of anti-AIDS drugs: conformational flexibility and positional adaptability are important in the design of non-nucleoside HIV-1 reverse transcriptase inhibitors. *Prog. Biophys. Mol. Biol.* **2005**, *88*, 209-231.

Das, K.; Sarafianos, S.G.; Clark, A.D., Jr.; Boyer, P.L.; Hughes, S.H.; Arnold, E. Crystal structures of clinically relevant Lys103Asn/Tyr181Cys double mutant HIV-1 reverse transcriptase in complexes with ATP and non-nucleoside inhibitor HBY 097. *J. Mol. Biol.* **2007**, *365*, 77-89.

Das, K.; Bauman, J.D.; Clark, A.D., Jr.; Frenkel, Y.V.; Lewi, P.J.; Shatkin, A.J.; Hughes, S.H.; Arnold, E. High-resolution structures of HIV-1 reverse transcriptase/TMC278 complexes: strategic flexibility explains potency against resistance mutations. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 1466-1471.

Ding, J.; Das, K.; Tantillo, C.; Zhang, W.; Clark, A.D., Jr.; Jessen, S.; Lu, X.; Hsiou, Y.; Jacobo-Molina, A.; Andries, K.; Pauwels, R.; Moereels, H.; Koymans, L.; Janssen, P.A.J.; Smith, R.H., Jr.; Koepke, M.K.; Michejda, C.J.; Hughes, S.H.; Arnold, E. Structure of HIV-1 reverse transcriptase in a complex with the non-nucleoside inhibitor alpha-APA R 95845 at 2.8 A resolution. *Structure* **1995**, *3*, 365-379.

Ding, J.; Das, K.; Hsiou, Y.; Sarafianos, S.G.; Clark, A.D., Jr.; Jacobo-Molina, A.; Tantillo, C.; Hughes, S.H.; Arnold, E. Structure and functional implications of the polymerase active site region in a complex of HIV-1 RT with a double-stranded DNA template-primer and an antibody Fab fragment at 2.8 Å. *J. Mol. Biol.* **1998**, *284*, 1095-1111.

Esnouf, R.; Ren, J.; Ross, C.; Jones, Y.; Stammers, D.; Stuart, D. Mechanism of inhibition of HIV-1 reverse transcriptase by non-nucleoside inhibitors. *Nat. Struct. Biol.* **1995**, *2*, 303-308.

Hsiou, Y.; Ding, J.; Das, K.; Clark, A.D., Jr.; Hughes, S.H.; Arnold, E. Structure of unliganded HIV-1 reverse transcriptase at 2.7 Å resolution: implications of conformational changes for polymerization and inhibition mechanisms. *Structure* **1996**, *4*, 853-860.

Jacobo-Molina, A.; Ding, J.; Nanni, R.G.; Clark, A.D., Jr.; Lu, X.; Tantillo, C.; Williams, R.L.; Kamer, G.; Ferris, A.L.; Clark, P.; Hizi, A.; Hughes, S.H.; Arnold, E. Crystal structure of Human Immunodeficiency Virus Type 1 reverse transcriptase complexed with double-stranded DNA at 3.0 Å resolution shows bent DNA. *Proc. Natl. Acad. Sci.* **1993**, *90*, 6320-6324.

Jorgensen, W.L.; Ruiz-Caro, J.; Tirado-Rives, J.; Basavapathruni, A.; Anderson, K.S.; Hamilton, A.D. Computer-aided design of non-nucleoside inhibitors of HIV-1 reverse transcriptase. *Bioorg. Med. Chem. Lett*. **2006**, *16*, 663-7.

Jorgensen, W.L. Efficient drug lead discovery and optimization. *Acc. Chem. Res*. **2009**, *42*, 724-33.

Kallings, L. O.The first postmodern pandemic: 25 years of HIV/AIDS. *J. Intern. Med.* **2008**, *263*, 218– 243.

Kaufman, G.; Cooper, D. Antiretroviral therapy of HIV-1 infection: established treatment strategies and new therapeutic options. *Curr. Opin. Microbio.* **2000**, *3*, 508-514.

Kohlstaedt, L.A.; Wang, J.; Friedman, J.M.; Rice, P.A.; Steitz, T.A. Crystal structure at 3.5 Å resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science* **1992**, *256*, 1783-1790.

Lewi, P.J.; de Jonge, M.; Daeyaert, F.; Koymans, L.; Vinkers, M.; Heeres, J.; Janssen, P.A.; Arnold, E.; Das, K.; Clark, A.D., Jr.; Hughes, S.H.; Boyer, P.L.; de Béthune, M.P.; Pauwels, R.; Andries, K.; Kukla, M.; Ludovici, D.; De Corte, B.; Kavash, R.; Ho, C. On the detection of multiple-binding modes of ligands to proteins, from biological, structural, and modeling data. *J. Comput. Aided Mol. Des.* **2003**, *17*, 129-134.

Rodgers, D.W.; Harrison, S.C. The structure of unliganded reverse transcriptase from the human immunodeficiency virus type 1. *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 1222–1226.

Shen, L.; Shen, J.; Luo, X.; Cheng, F.; Xu, Y.; Chen, K.; Arnold, E.; Ding, J.; Jiang, H. Steered molecular dynamics simulation on the binding of NNRTI to HIV-1 RT. *Biophys J*, **2003**, *84*, 3547-63.

Tantillo, C.; Ding, J.; Jacobo-Molina, A.; Nanni, R.G.; Boyer, P.L.; Hughes, S.H.; Pauwels, R.; Andries, K.; Janssen, P.A.; Arnold, E. Locations of anti-AIDS drug binding sites and resistance mutations in the three-dimensional structure of HIV-1 reverse transcriptase. Implications for mechanisms of drug inhibition and resistance. *J. Mol. Biol.* **1994**, *24*, 369-387.

Temiz, N.A.; Bahar, I. Inhibitor binding alters the directions of domain motions in HIV-1 reverse transcriptase. *Proteins* **2002**, *49*, 61-70.

Walker, L.M.; Burton, D.R. Rational antibody-based HIV-1 vaccine design: current approaches and future directions. *Curr. Opin. Immunol*. **2010**, *22*, 1-9.

Wolber, G.; Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J Chem Inf Model*. **2005**, *45*, 160-9.

Zheng, Y.; Lovsin, N.; Peterlin, B.M. Newly identified host factors modulate HIV infection. *Immunol. Lett.* **2005**, *97*, 225-234.

Chapter 6

Conformational Landscape of HIV-1 Reverse Transcriptase Binding to Non-Nucleoside Inhibitors From a Large Data Set of Many Crystal Structures

6.1 Introduction

As discussed in the preceding chapters, the HIV-1 reverse transcriptase (RT) non-nucleoside inhibitor binding pocket (NNIBP) is very flexible, conforming to the shape and needs of each bound ligand (Das et al., 2004; Das et al., 2008). Present day crystallization techniques have allowed determination of a large number of X-ray structures, which are deposited in the publically available Protein Data Bank (PDB; Berman et al., 2000). As there are many different NNRTIs, it is reasonable to assume that the many X-ray structures associated with these NNRTIs offer many possible conformations of the NNIBP. Each X-ray structure can be thought of as a point on the conformational landscape for binding NNRTIs to HIV-1 RT. Previous studies have compared and contrasted limited numbers of receptor conformations (Das et al., 2008; Das et al., 2005; Das et al., 2007; Ren et al., 2000; Ren et al., 1995; Spallarossa et al., 2008; Pata et al., 2004) while others have focused on the composition and conformations of the ligands without regard to the conformation of the receptor (Xu et al., 2006; O'Brien et al., 2005). The present study infers information about the conformational landscape of the NNIBP from a large set of 99 RT crystal structures.

6.2 Procedures and Results

The procedures and results of this part of the thesis are presented below as a reprint of a paper published in the *Journal of Medicinal Chemistry* **2009**, *52*, 6413-6420.

Journal of Medicinal Chemistry Article

J. Med. Chem. **2009,** *52,* 6413–6420 **6413**
DOI: 10.1021/jm900854h

Conformational Landscape of the Human Immunodeficiency Virus Type 1 Reverse Transcriptase Non-Nucleoside Inhibitor Binding Pocket: Lessons for Inhibitor Design from a Cluster Analysis of Many Crystal Structures

Kristina A. Paris,[†,‡] Omar Haq,[†,‡] Anthony K. Felts,[†,‡] Kalyan Das,[†,§] Eddy Arnold,*[,†,§] and Ronald M. Levy*[,†,‡]

[†]*Department of Chemistry and Chemical Biology, Rutgers University, 610 Taylor Road, Piscataway, New Jersey 08854,* [‡]*BioMaPS Institute for Quantitative Biology, Rutgers University, Piscataway, New Jersey 08854, and* [§]*Center for Advanced Biotechnology and Medicine, Rutgers University, 679 Hoes Lane, Piscataway, New Jersey 08854*

Received June 11, 2009

Clustering of 99 available X-ray crystal structures of HIV-1 reverse transcriptase (RT) at the flexible non-nucleoside inhibitor binding pocket (NNIBP) provides information about features of the conformational landscape for binding non-nucleoside inhibitors (NNRTIs), including effects of mutation and crystal forms. The ensemble of NNIBP conformations is separated into eight discrete clusters based primarily on the position of the functionally important primer grip, the displacement of which is believed to be one of the mechanisms of inhibition of RT. Two of these clusters are populated by structures in which the primer grip exhibits novel conformations that differ from the predominant cluster by over 4 Å and are induced by the unique inhibitors capravirine and rilpivirine/TMC278. This work identifies a new conformation of the NNIBP that may be used to design NNRTIs. It can also be used to guide more complete exploration of the NNIBP free energy landscape using advanced sampling techniques.

Introduction

Current strategies for treatment of HIV involve hindering different steps in the retrovirus' life cycle. This study focuses on the inhibition of the viral enzyme reverse transcriptase (RT[a]) by non-nucleoside reverse transcriptase inhibitors (NNRTIs). NNRTIs are noncompetitive inhibitors that bind to a pocket called the non-nucleoside reverse transcriptase binding pocket (NNIBP) which lies ∼10 Å from the enzyme's polymerase active site.[1]

Analysis of crystal structures has suggested three possible mechanisms of inhibition (which are not mutually exclusive) of HIV-1 RT by NNRTIs: (1) restriction of the p66 thumb flexibility; (2) distortion of catalytically essential residues at the polymerase active site; (3) displacement of the primer grip. In each of these mechanisms, the binding of NNRTIs is proposed to lead to conformational perturbations and to limit conformational flexibility required for efficient DNA synthesis by RT. In the "molecular arthritis" mechanism, conformational restriction of the p66 thumb subdomain was suggested to limit flexibility of the enzyme required for catalysis.[1] NNRTI binding may restrict the mobility of the thumb subdomain[1−3] or may change the direction of movement of the thumb subdomain,[4] thus slowing down or preventing template-primer translocation and inhibiting facile elongation of nascent viral DNA. NNRTI binding perturbs the configuration of the RT polymerase active site region, including the catalytically essential D110, D185, and D186 residues,[5] and limits conformational changes of the "YMDD loop" containing M184 and D185.[6] The primer grip is a structural element in HIV-1 RT that has been proposed to be essential for positioning the primer 3′ terminus at the active site,[7] and movements of the primer grip and the associated thumb subdomain are thought to be critical for the translocation of nucleic acid following incorporation of nucleotides during polymerization.[2] NNRTI binding causes a significant displacement (∼4 Å) of the primer grip, leading to possible inappropriate positioning of the primer terminus at the active site; this conformational alteration and possible restriction of primer grip mobility may be a major contributor to inhibition by NNRTIs.[8] Movement in the primer grip is also thought to affect allosteric hinge-bending movements in the position of the thumb subdomain (the tip of which lies ∼30 Å from the NNIBP) .[1,9,10] If structures are superimposed on the basis of the $\beta6$-$\beta9$-$\beta10$ strands, the thumb subdomain of NNRTI-bound RT is rotated by ∼40° relative to that in the unbound apo enzyme[10,11] (see Figure 1).

The NNIBP is very flexible, changing conformation when different NNRTIs are bound.[12] This has been described as a "shrink-wrap" effect where the binding pocket residues change conformation to form stabilizing interactions with a ligand.[13] Although available inhibitors have different shapes, sizes, functional groups, and binding modes, they display a number of common features in their interactions with the NNIBP residues: aromatic ring(s) capable of forming $\pi-\pi$ stacking interactions with aromatic residues, as well as making hydrophobic contacts with other nonpolar pocket residues, and (usually) hydrogen bond (H-bond) donors capable

*To whom correspondence should be addressed. For E.A.: (address) Center for Advanced Biotechnology and Medicine, Rutgers University, 679 Hoes Lane, Piscataway, NJ 08854; (phone) 732-235-5323; (fax) 732-235-5788; (e-mail) arnold@cabm.rutgers.edu. For R.M.L.: (address) Department of Chemistry and Chemical Biology, Rutgers University, 610 Taylor Road, Piscataway, NJ 08854; (phone) 732-445-3947; (fax) 732-445-5312; (e-mail) ronlevy@lutece.rutgers.edu.

[a] Abbreviations: HIV-1, human immunodeficiency virus type 1; RT, reverse transcriptase; NNIBP, non-nucleoside reverse transcriptase inhibitor binding pocket; NNRTI, non-nucleoside reverse transcriptase inhibitor; PDB, Protein Data Bank; REMD, replica exchange molecular dynamics.

Published on Web 09/08/2009

of forming an H-bond with the backbone carbonyl oxygen of K101. Current design of inhibitors focuses on developing the "required" interactions in such a way to ensure high activity, reasonable solubility, and broad potency against drug-resistant variants.

As there are many different NNRTIs, it is reasonable to assume that there are many different conformations of the NNIBP. Here, 99 X-ray structures of HIV-1 RT from the Protein Data Bank (PDB)[14] are examined. Of these, 52 are wild-type (WT) RT bound to NNRTIs, 30 are mutant forms of RT bound to NNRTIs, 3 are unliganded WT RT, 3 are unliganded mutant RT, 10 contain DNA or RNA substrates or ATP, and 1 is bound to the RNase H inhibitor DHBNH. (Note: In this instance, WT refers to an enzyme with no mutations within 15 Å of the NNIBP.) These 99 structures represent an ensemble of observed conformations of the NNIBP with perturbations created by mutation, binding of different ligands (induced fit effects), and different crystal forms and are listed in Table 1.

The conformational elasticity of the binding pocket plays an important role in drug design. Here we report the analysis of conformational states of the NNIBP in 99 available crystal structures of HIV-1 RT. Each X-ray structure represents a point on the conformational landscape for binding NNRTIs.

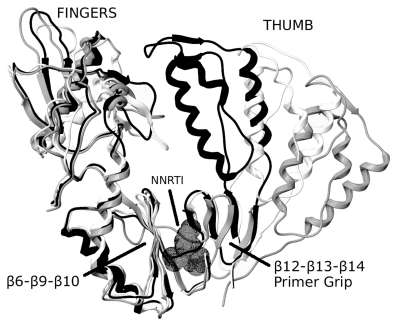

Figure 1. Rearrangement of HIV-1 RT upon binding. Upon binding substrates such as DNA or NNRTIs such as nevirapine, RT undergoes a global conformational change in the thumb of the enzyme. Upon binding an NNRTI, there is also a shift in the primer grip region shown in the movement from black/light-gray to dark-gray: black, unbound RT (PDB ID 1DLO); light-gray, RT bound to DNA (PDB ID 1RTD); dark-gray, RT bound to the NNRTI nevirapine (PDB ID 1VRT).

The goal of characterizing a conformational landscape and its corresponding energy landscape has come to occupy a central role in biophysics.[15−19] This study infers information about the landscape for ligand binding to HIV-1 RT by performing a cluster analysis of this large data set of X-ray crystal structures. Our cluster analysis focuses on the conformational states of the binding pocket, whereas previously published studies have used clustering primarily to characterize the flexibility, chemical class, and binding mode of the ligand.[20,21]

The availability of a large data set of HIV-1 RT crystal structures in the PDB and their clustering provides information about the locations of free energy basins and their shapes. Ideally, the populations of the different X-ray resolved conformations of the NNIBP of HIV-1 RT could be transformed through Boltzmann statistics into a free energy landscape of the receptor in the spirit of free energy folding funnels proposed for proteins in general. Folding funnels are rugged in the vicinity of the native fold of the protein, suggesting energetically competitive and similar conformations that provide an enhanced means of interaction between the protein and either ligands or other proteins.[15−19] The landscape provides useful information about both the different means for inhibitors to bind to HIV-1 RT and the strain free energy required to adopt a particular conformation for binding.

Highly populated clusters may suggest that the deformations within the NNIBP are locally elastic with small free energy penalties. In contrast, sparsely populated clusters are suggestive of more steeply sloped free energy basins. However, as the 99 X-ray structure data set does not represent a systematic sampling of the landscape, the populations may reflect the bias found in the drug design process where inhibitors are often designed on the basis of earlier inhibitors or are designed for previously determined structures of the NNIBP. Even so, the NNIBP conformations representative of the sparsely populated basins provide opportunities for exploiting new interactions and ligand conformational freedom in developing new more potent NNRTIs.

Results

The average root-mean-square deviation (rmsd) of all Cα atoms within 15 Å of any NNRTI across the set of 99 X-ray structures of RT is only 1.23 ± 0.48 Å when the superposition is performed on the same set of Cα atoms. This increases to 1.58 ± 0.59 Å for all Cα atoms within 10 Å of any NNRTI. An analysis of the radii of gyration for each of the Cα atoms within 10 Å of any NNRTI across the ensemble of 82 RT/NNRTI complex conformations shows large variation in

Table 1. The 99 Crystal Structures of HIV-1 RT Used in This Analysis[a]

	PDB code
WT/NNRTI	1BQM, 1C0T, 1C0U, 1C1B, 1C1C, 1DTQ, 1DTT, 1EET, 1EP4, 1FK9, 1HNI, 1HNV, 1IKW, 1JLQ, 1KLM, 1LW0, 1LW2, 1LWE, 1REV, 1RT1, 1RT2, 1RT3, 1RT4, 1RT5, 1RT6, 1RT7, 1RTH, 1RTI, 1S6P, 1S6Q, 1S9E, 1S9G, 1SUQ, 1TKT, 1TKX, 1TKZ, 1TL1, 1TL3, 1TV6, 1TVR, 1 VRT, 1 VRU, 2B5J, 2B6A, 2BAN, 2BE2, 2OPP, 2 VG5, 2 VG6, 2 VG7, 2ZD1, unpublished
Mut/NNRTI	1FKO, 1FKP, 1HPZ, 1HQU, 1IKV, 1IKX, 1IKY, 1JKH, 1JLA, 1JLB, 1JLC, 1JLF, 1JLG, 1LWC, 1LWF, 1S1T, 1S1U, 1S1 V,1S1W, 1S1X, 1SV5, 2HND, 2HNY, 2HNZ, 2IC3, 2OPQ, 2OPR, 2OPS, 2ZE2, 3BGR
WT unbound	1DLO, 1HMV, 1RTJ
Mut unbound	1HQE, 1JLE, 1QE1
DNA/RNA/ATP bound	1HYS, 1J5O, 1N5Y, 1N6Q, 1R0A, 1RTD, 1T03, 1T05, 2HMI, 2IAJ
RT/RNase H I	2I5J

[a] WT/NNRTI: WT RT complexed with an NNRTI. Mut/NNRTI: mutant RT complexed with an NNRTI. WT unbound: Apo WT RT. Mut unbound: Apo mutant RT. DNA/RNA/ATP bound: RT bound to substrates DNA, RNA or ATP. RT/RNase H I: RT complexed with a RNase H inhibitor.

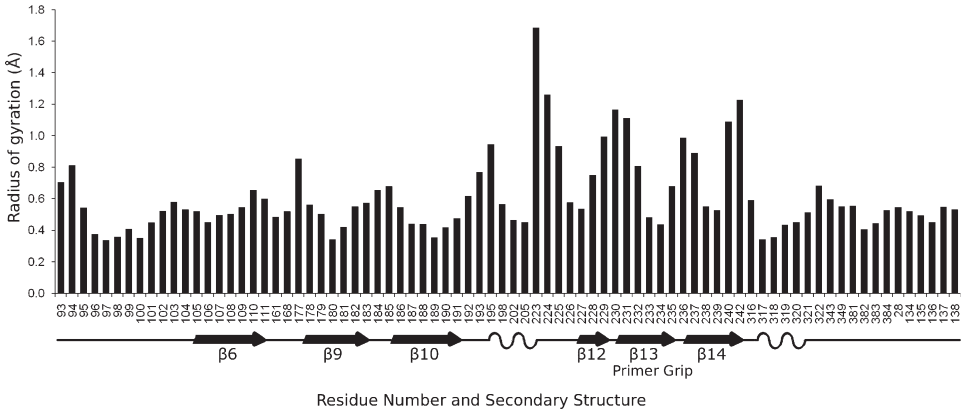

Figure 2. Radius of gyration (R_g) for each Cα atom of the residues within 10 Å of any NNRTI over the set of RT/NNRTI complexes (superposition was based on all Cα atoms in the set). The radius of gyration is an estimate for the spread of an atom's position across the ensemble of structures. A higher R_g signifies a large spread, whereas a smaller R_g signifies little fluctuation of position of the atom across different structures. Secondary structure is labeled with arrows for β strands and with curves for α helices. A straight line designates regions with no evident secondary structure using DSSP.[39]

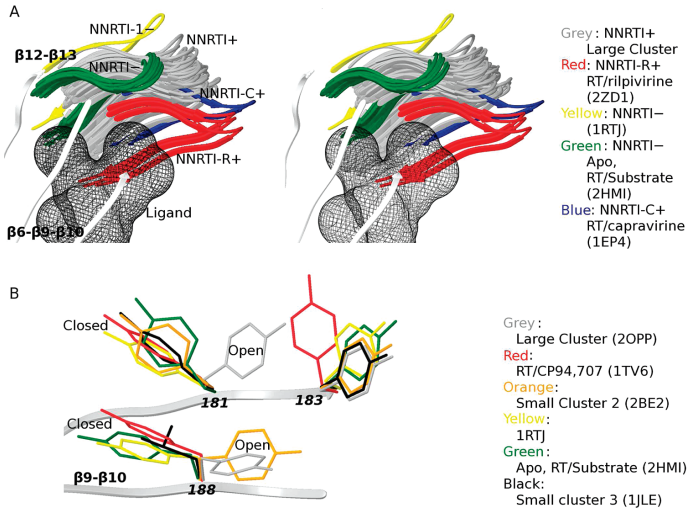

Figure 3. Separation of 99 X-ray structures. Structure representatives from each cluster are listed in parentheses. A further description can be found in Table 2. (A) Stereoview of the primer grip regions of all 99 X-ray structures (only β12-β13 shown) with alignment on β6-β9-β10. (B) Cluster representatives showing separation in residues 181, 183, and 188.

some regions in the vicinity of the NNIBP such as the primer grip β12-β13-β14 and the loop around P95 through L100 while other regions such as the β6-β9-β10 sheet remain more static (see Figure 2). This analysis serves to indicate regions on which to align the ensemble of 99 X-ray structures (the least variable) and on which to focus clustering experiments (the most variable).

Analysis of the side chains that point into the NNIBP of the 52 WT NNRTI-bound conformations of RT displays only three residues that give discrete clusters when clustered individually: Y181, Y183, and Y188. The remaining side chains (L100, K103, D186, F227, W229, and L234 from p66; N136 and E138 from p51) fluctuate across the ensemble of structures, but the distribution is quasi-continuous and therefore does not allow separation into meaningful clusters.

Further investigation of the NNIBP backbone (primarily in the primer grip region) and side chains (residues 181, 183, 188) utilizing hierarchical clustering techniques elucidates eight

Table 2. Clustering Results: Eight Basins with Different Features[a]

cluster	cluster members, PDB IDs	residue 181, 188	$\beta12$-$\beta13$	rmsd of $\beta12$-$\beta13$-$\beta14$ (Å)		NNRTI bound
				to large Clust Rep (2OPP)	to Apo (1DLO)	
large	**2OPP**, etc. (73 structures)	open, open	NNRTI+	0.0	4.0	y
small 1	2IAJ,1HMV,1RTD, 1HQE,1T05, 1N6Q, **2HMI**,1R0A,1T03, 1N5Y, 1J5O,1QE1, 1HYS,1DLO	closed, closed	NNRTI−	3.4	1.8	n
small 2	1FKO,1RTI,**2BE2**, 2B5J,1RT3	closed, open	NNRTI+	1.2	3.3	y
small 3	**1JLE**,2I5J	closed, closed	NNRTI+	2.2	3.8	n
small 4	**2ZD1**,2ZE2,3BGR	open, open	NNRTI-R+	3.9	6.6	y
1EP4	1EP4	open, open	NNRTI-C+	4.5	6.2	y
1TV6[b]	1TV6	closed, closed	NNRTI+	2.6	2.8	y
1RTJ	1RTJ	closed, closed	NNRTI-1−	4.1	3.6	n

[a] Structure representatives from each cluster are in bold. A cluster representative is a structure in the cluster that has the smallest rmsd when compared to the collection of centroids of each of the comparison atoms.[38] [b] 1TV6 represents the only case in which Y183 is in a largely differing position in the ensemble of structures.

basins that depict varying conformations of the flexible binding pocket. The eight basins include one large cluster of 73 structures, four small clusters composed of 2−14 structures, and three singletons (Figure 3 and Table 2). The large cluster is composed solely of NNRTI-bound structures with Y181 and Y188 side chains both occupying bound "open" conformations where the two side chains have moved to open a pocket that accommodates the ligand. The small cluster of 14 (small cluster 1) includes unliganded structures and those containing dsDNA, RNA/DNA, and/or dNTP. Since no NNRTI is bound in these 14 conformations, W229, F227, Y181, and Y188 fill the space where the ligand would be found. This difference in positioning of W229 and its connected primer grip as well as the positioning of the side chains of residues 181 and 188 in the unbound "closed" position allows for separation from structures that are bound to NNRTIs. The primer grip conformation seen in this small cluster will be referred to as NNRTI−, while the primer grip conformation seen in the large cluster, where there is a shift in the primer grip of ~3.4 Å (see Table 2), will be referred to as NNRTI+.

Most interesting is the identification of a small cluster of three RT/rilpivirine complexes (small cluster 4; PDB IDs 2ZD1, 2ZE2, 3BGR) and two singletons (1EP4, bound to the NNRTI capravirine, and 1RTJ, where an NNRTI was washed out prior to structural determination), which have different primer grip conformations from those seen in the large cluster. The small cluster of three RT/rilpivirine complexes will be called NNRTI-R+, the singleton 1EP4 will be called NNRTI-C+, and the singleton 1RTJ will be called NNRTI-1−. The unexpected effect of rilpivirine and capravirine on the distortion of the primer grip was discovered through the cluster analysis performed in this work. These conformations will be discussed at length in the Discussion.

Inclusion of the side chains of residues 181, 183, and 188 in the clustering analysis provides separation of several structures that have the large cluster primer grip NNRTI+ conformation but have differing conformations of these side chains. The small cluster of 5 (small cluster 2 in Table 2) is composed of RT/NNRTI complexes where Y181 is in the "closed" position, while Y188 is in the "open" position. The small cluster of 2 (small cluster 3 in Table 2) displays both Y181 and Y188 in the "closed" position. Both structures in this cluster represent interesting cases in which an NNRTI is not bound but the primer grip is in a bound conformation: an NNRTI was removed prior to structure determination of 1JLE, and 2I5J is bound to the RNase H inhibitor DHBNH. DHBNH does not bind in, but instead binds adjacent to, the NNIBP[22] and therefore would not be expected to have a primer grip conformation similar to the NNRTI-bound large cluster. This small cluster therefore offers insight into possible unique interactions near the NNIBP that may be exploited for design of new NNRTIs that can stabilize the primer grip in a perturbed conformation that disrupts polymerase activity. The final singleton, 1TV6, bound to the large ligand CP-94,707, is the only structure in which a different discrete conformation of Y183 is seen. 1TV6 also is a case in which both Y181 and Y188 are in the "closed" position.

Additional Cluster Analysis of the RT "Thumb" Region. A large conformational change occurs upon binding of nucleic acid where the thumb and fingers of RT move to "clasp" the nucleic acid; a similar change in the thumb conformation is also apparent upon binding of an NNRTI[7] (see Figure 1). As movement in the primer grip is thought to affect allosteric hinge-bending movements in the position of the thumb subdomain,[1,9,10,23] an additional cluster analysis on three residues at the tip of the p66 thumb subdomain of the structures was performed in an attempt to give more information about the large cluster. The clustering level with the largest separation ratio yields nearly identical results to clustering on the primer grip. Further analysis using a smaller separation ratio for selection of the cluster level results in the separation of the 80 RT/NNRTI complexes that occupy the large cluster into one singleton (1JLE) and two subclusters of 28 and 51 structures between which a small shear or twist of the primer grip is seen. The separation of the two subclusters is due to a shift in position of the tip of the thumb corresponding to an average rmsd between clusters of 5.7 ± 1.7 Å. The shift in thumb position is most likely due to the different crystal forms used in structure determination, as the structures in the cluster of 51 utilize one crystal form while the cluster of 28 utilize one of two differing crystal forms. Influence of the crystal packing propagates to the NNIBP, causing a slight shear or twist seen in the primer grip. However, these subclusters overlap in the primer grip region and so are not discernible by clustering on the primer grip alone; the effect of crystal form on the primer grip conformation is minimal.

Discussion

The existence of several clusters indicates that structural variability is present, but since most of the structures are in one cluster, that variability is not evenly distributed across the NNIBP landscape. Most obvious, the NNRTI-bound structures are separated from those of RTs not bound to NNRTIs.

This was anticipated, as the NNIBP undergoes large structural rearrangements upon binding of an NNRTI where the aromatic side chains of Y181 and Y188 swivel out of the binding pocket and the primer grip region moves to create space for the NNRTI; the non-nucleoside inhibitor binding pocket only exists in structures with an NNRTI present.[10,24]

A cluster distribution like that observed for the 82 HIV-1 RT/NNRTI complexes, where the great majority of the structures are found in one large cluster, suggests the possibility that the receptor pocket has not been interrogated by as extensive a variety of ligands as may have been previously thought. However, a few other basins do emerge from this analysis of the large data set of (99) RT crystal structures. We can speculate that the sparsely populated basins are separated by relatively high free energy barriers from the largest basin representing 73 structures; otherwise we would expect to have observed structures populating these "barrier" regions of the landscape among the large number of RT complexes in the PDB. The large cluster, on the other hand, appears to represent a sampling from what is effectively a continuum of accessible states.

The large cluster can be pictured as a basin within which there are low energy barriers separating many minima. Several of the residues within 10 Å of the ligands sample a more or less continuous distribution of conformations; the flexible loop consisting of residues 95−100 is an example. The primer grip β-sheet displays a "shrink-wrap" effect involving small adjustments of position within the large basin to optimize interactions with different functional groups and chemically diverse NNRTIs. The average backbone rmsd of the β12-β13-β14 strands within the large cluster is 1.4 ± 0.5 Å; all NNIBP residues sample somewhat continuous distributions that span 1−3 Å of conformational space. The ligands bound to the complexes in the large cluster are diverse in their shapes, sizes, functional groups, and binding modes, creating a large basin within which many conformations of the binding pocket are explored.

The most significant new observation to arise from this cluster analysis of the RT data set is the description of four basins that are sampled by the functionally important primer grip β12-β13-β14 sheet: the large NNRTI+ cluster, the small NNRTI-R+ cluster of three structures bound to the NNRTI rilpivirine/TMC278,[13] the NNRTI-C+ singleton bound to the NNRTI capravirine/S-1153,[25] and the NNRTI-1- singleton 1RTJ in which a HEPT ligand was washed out prior to structure determination.[5] Whereas the majority of the β12-β13 strands of the primer grip are repositioned upon binding a non-nucleoside inhibitor by 3.4 ± 0.5 Å, the NNRTI-R+ and NNRTI-C+ forms differ from the NNRTI− structure 1DLO by > 6.2 Å, setting them ∼4 Å from the large NNRTI+ cluster representative (Table 2). All three clusters, NNRTI-R+, NNRTI-C+ and NNRTI-1−, are also separated via clustering on the thumb. The NNRTI-R+ and NNRTI-C+ forms, separated from each other by ∼2.2 Å in the primer grip and ∼6.6 Å in the thumb, can be rationalized as the interrogation of the NNIBP by larger ligands that interact with residues in the NNIBP in distinct ways.

The ligand bound to the NNRTI-R+ form, rilpivirine/TMC278 (2ZD1, 2ZE2, 3BRG), is a diaryl pyrimidine (DAPY) analogue. DAPY compounds have been found to be effective against many mutant forms of RT by utilizing multiple binding modes.[12,26] The binding of the DAPY rilpivirine differs from that seen in other DAPYs; its cyanovinyl group extends into a hydrophobic tunnel formed by the side chains of Y188, F227, W229, and L234. The extensive interaction of this cyanovinyl group with the hydrophobic tunnel is thought to explain why rilpivirine is the most potent of the DAPY analogues.[13] The formation of the tunnel is apparently also responsible for the shift of the β12-β13-β14 strands over the binding pocket as the positions of F227, W229, and L234 are reconfigured to make room for the cyanovinyl group. One other crystallized NNRTI, seen in 2B5J, acts similarly because its cyanovinyl group also extends into the hydrophobic tunnel.[27] However, instead of causing a displacement of the β12-β13-β14 sheet to form the tunnel, binding of this NNRTI is accompanied by a displacement of Y188. The three RT/rilpivirine complexes also correspond to a new crystal form of HIV-1 RT.[13] To examine whether the crystal contacts in the NNRTI-R+ structures (2ZD1, 2ZE2, 3BGR) induce changes in the primer grip, a complex of RT with a non-DAPY ligand in the new crystal form was included in the clustering study.[28] This structure clusters in the large cluster, implying that the NNRTI-R+ conformation is not due to the crystal contacts in the new crystal form but rather to the novel interactions of the inhibitor's cyanovinyl group with the hydrophobic tunnel of the enzyme.

The ligand found in the NNRTI-C+ form, the imidazole capravirine (S-1153), is larger and more branched than others. Novel in this RT/capravirine complex (1EP4) is the formation of a main-chain hydrogen bond with P236.[25] This H-bond causes the 3,5-dichlorophenyl ring to be in proximity with W229, which is shifted by ∼4 Å over the NNIBP relative to the large cluster NNRTI+ representative.

The NNRTI-1− form shows a subtler shift in the binding pocket. Our clustering revealed that the different crystal forms of RT do not induce significant perturbations of the binding pocket structure except in the case of the NNRTI-1− form (PDB ID 1RTJ), where a weakly bound NNRTI was washed out to obtain an unliganded RT structure.[5] Crystal contacts appear to stabilize the NNRTI-1− structure in the inhibited "open" form of the primer grip; this suggests that fluctuations of the binding pocket to the open form may occur even when no ligand is present.

The conformations of the primer grip β12-β13-β14 strands that are identified in the cluster analysis as NNRTI-R+ and NNRTI-C+ suggest routes for further exploration of new ligands that interrogate the NNIBP in ways that sample new and sparsely populated regions of the conformational landscape. Such conformations highlight receptor−ligand interactions such as additional H-bonds and formation of a stabilizing hydrophobic tunnel that appear resistant to several common mutations and may not be attainable in other conformations of the binding pocket. Design strategies based on the NNRTI-R+ and NNRTI-C+ basins can be utilized. These include further optimization of analogues of the highly active DAPY and imidazole compounds, focusing on interactions with the hydrophobic tunnel similar to rilpivirine and focusing on forming main-chain hydrogen bonds with P236 similar to capravirine.

NNIBP Mutations: Conformational Effects. Across the 82 RT/NNRTI complexes, mutations appear to have little effect on conformational change in the NNIBP but instead mainly affect the chemical signatures of the binding pocket, causing energetic penalties in binding of inhibitors. Minor changes in the NNIBP do occur in response to repositioning of the ligands, but these changes are minimal, causing mutants to be found in the conformational basins associated with their WT counterparts.

Interestingly, the ligands associated with both the NNRTI-R+ and NNRTI-C+ conformational basins, capravirine and rilpivirine, are highly active not just against the WT form of RT but also against many mutant forms. Both capravirine and rilpivirine have lower EC_{50} values than many of the other NNRTIs, showing greater activity toward WT and commonly mutated forms of HIV-1 RT.[29,30] The higher activities have been attributed to the interactions with the enzyme for these ligands as discussed above.

Energy Landscape View of NNRTI Binding to HIV-1 RT. The overall affinity of a ligand for a receptor can be expressed as a balance between the strength of the interactions of a ligand for any particular binding-competent conformation of the receptor and the probability of occurrence of that conformation in the absence of the ligand. Another name for these receptor conformation probability distributions is the free energy landscape of the receptor from which the strain free energy required to move from one conformation to another in the absence of a ligand may be estimated.

Clustering of the 99 available X-ray RT structures based on the functionally important primer grip residues has identified five clusters or strain free energy basins. These basins are depicted in Figure 4 where the rmsd of the primer grip β12-β13 strands, relative to the apo structure 1DLO and relative to the large cluster representative 2OPP, were chosen as order parameters. These coordinates best describe the degree to which the primer grip has moved due to binding of an inhibitor. The clusters described in the previous section are further illustrated by Figure 4: the main large cluster (80 structures), a cluster of 10 substrate-bound and 4 apo RTs, a cluster of 3 RT/rilpivirine complexes with 1 RT/capravirine complex (separated in the above section into a cluster of 3 and a singleton, respectively), and a singleton represented by 1RTJ (NNRTI-1−). The large cluster (NNRTI+) is described by a broad and rugged region of the landscape corresponding to fine-tuning of the NNIBP to fit various inhibitors. The region corresponding to the RT/capravirine (NNRTI-C+) and RT/rilpivirine (NNRTI-R+) complexes reflects inhibitors that have stretched the primer grip region, creating novel conformations of the NNIBP.

The populations of the different NNIBP conformational basins shown in Figure 4 cannot be directly inverted to estimate receptor strain free energies. The observed locations and populations of the basins depend not only on the receptor strain free energies but also on the averaged interaction energies of the ligands with the receptor. Additionally, the crystal structure database represents a nonsystematic sampling of the landscape, as many of the inhibitors have been designed on the basis of an earlier inhibitor through QSAR techniques[31] or designed for previously determined structures of the receptor. Both design approaches limit the potential to discover novel conformations of the receptor and partially explain why the cluster analysis produces a large cluster of NNIBP structures with similar inhibitors and similar receptor conformations.

One possible route to construct the receptor strain free energy landscape for the binding of NNRTIs to the NNIBP is to integrate information from the cluster analysis with molecular simulations. We can use structures representative of the different basins as "landmarks" to guide and test physics-based simulations using modern effective potentials and advanced sampling techniques like replica exchange molecular dynamics (REMD).[19,32,33] As the enzyme is very large, performing simulations using the whole protein may not be the most effective way to carry out free energy simulations of the binding pocket. Information about flexibility acquired from the cluster analysis of the X-ray structures described here can be used to both create a suitable fragment of the enzyme and develop constraints on the system to limit the computational time needed while optimizing the sampling of the conformational landscape of the NNIBP. For example, the regions of the NNIBP that pertain to areas of little flexibility, e.g., the β6-β9-β10 strands, can be held fixed while the highly variable regions such as the β12-β13-β14 strands of the primer grip and neighboring residues can be allowed to move. The clustering results presented here also provide a benchmark for the performance of the conformational sampling of the landscape. Initial simulations appear promising, as all of the basins illustrated in Figure 4 are found to have substantial statistical weight in the physics-based exploration of the receptor free energy landscape using temperature replica exchange molecular dynamics. However, several of the basins are not fully explored. Incorporation of umbrella sampling and/or utilization of distance restraints will allow for a more complete picture.

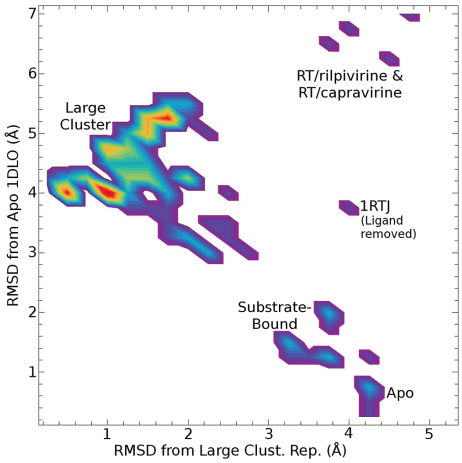

Figure 4. Populations based on rmsd (to apo 1DLO and large cluster representative 2OPP) of the primer grip with alignment on β6-β9-β10. Red designates a higher population, with tones progressing to dark blue being regions with lesser occupancy. The 99 experimental X-ray structures are as follows: locations of apo RT, substrate bound RT, the large NNRTI+ cluster, the NNRTI-1− singleton (1RTJ) and the RT/rilpivirine and RT/capravirine complexes (NNRTI-R+ and NNRTI-C+, respectively). The large cluster (NNRTI+) is descibed as broad and rugged.

Conclusion

Previous structural studies have compared and contrasted a limited number of HIV-1 RT NNIBP receptor conformations,[13,23−25,34−36] while other studies have focused on the composition and conformations of the ligands (NNRTIs) without regard to the conformation of the NNIBP. This study, the first to take a comprehensive look at the conformational fluctuations of the NNRTI receptor pocket, fills in missing pieces by utilizing a clustering algorithm to compare

and contrast 99 available conformations of the non-nucleoside inhibitor binding pocket. The cluster analysis reported here has identified the locations of several conformational basins of the receptor pocket. The separation found is very similar across multiple clustering algorithms and, as such, suggests that the results reported here are intrinsic to the data and represent a "natural" clustering of the experimentally determined NNIBP conformations. The different basins reflect the variation in the NNIBP; however, the basins are not evenly populated. The sparsely populated basins provide opportunities for the design and/or optimization of potent ligands that inhibit RT in conformations of the NNIBP that exhibit varied positions of the functionally important primer grip. Two of the sparsely populated basins highlight receptor−ligand interactions that may not be attainable in other conformations of the binding pocket and that may be exploited in drug design strategies. These include main-chain hydrogen bonds with P236 and interactions with the hydrophobic tunnel surrounded by Y188, F227, W229, and L234 and formed by repositioning the primer grip.

Information from this study also serves as an essential guide for theoretical studies to map the free energy landscape of the NNIBP using modern all atom effective potentials and advanced sampling techniques like replica exchange molecular dynamics (REMD). A free energy landscape for the NNIBP would allow calculation of the strain free energy of the receptor required to adopt the ligand-bound conformation. Simulations may also highlight previously unexplored conformations of the NNIBP that may be suitable for ligand design and lead to novel potent NNRTIs. The construction of a model for the free energy landscape of the NNIBP using REMD guided by the cluster landmarks described in this paper will be the subject of a future communication.

Experimental Section

Selection and Preparation of X-ray Structures. Careful preparation of the structural data set was essential, as simply clustering on the unedited set revealed mostly noise. X-ray crystal structures from the Protein Data Bank[14] were first analyzed to determine an atom sequence common to all structures. Entries that were found to be missing a large amount of structural information were removed to leave the data set of 99 structures used in this study. The 99 structures were then renumbered and reordered to follow the common atom sequence discovered from the analysis of each entry. This allowed for a normalization of the entries. Residues that experienced mutations were stripped of their side chain atoms that were not shared by each residue type. For example, as residue 103 is found as either a lysine or an asparagine, only the backbone atoms and Cβ and Cγ of the side chain were included. Regions where many structures were found to be missing atoms were also removed.

Analysis of Backbone and Side Chain Fluctuation. A fluctuation analysis of the backbones of residues within 10 Å of any NNRTI in RT/NNRTI complexes (residues 91, 93−111, 161, 168, 177−193, 195, 198, 202, 205, 223−240, 242, 316−322, 343, 381−384 from p66 and 28, 134−138 from p51) was performed by aligning the ensemble of structures based on the Cα of each of the 81 residues and calculating the radius of gyration (R_g) of the point cloud of all the positions for each Cα atom in the entire ensemble of structures. A low R_g coincides with little movement of the position of the atom across conformers, whereas a high R_g coincides with an atom that takes on many different positions in the ensemble.

Fluctuation of the side chains within 10 Å of any NNRTI that point into the NNIBP was analyzed by clustering on each side chain individually using single-linkage hierarchical clustering.[37,38] The best clustering was chosen as that which gave the highest minimum separation ratio (MSR), an empirical measure of the degree of separation.

Alignment of RT Structures. Clustering results are partially dependent on the alignment of conformers. In this study, the structures were aligned on the Cα atoms of residues 105−111, 178−183, and 186−191 that correspond to β6, β9, and β10, respectively. This alignment was chosen on the basis of the backbone analysis above, since their Cα atoms have low R_g values across the 82 RT/NNRTI complexes (see Figure 2). Superposition on the β6-β9-β10 is also often used in the literature to show movement within the NNIBP as well as global changes in conformation due to binding different ligands and substrates. Results using alignment on β6-β9-β10 were compared to alternative alignments, including alignment on the backbone atoms of all residues within 15 Å of the NNIBP, and were shown to give similar results. However, alignment on β6-β9-β10 was found to give the best separation of primer grip conformations with respect to the minimum separation ratio and minimum distances between clusters.

Clustering of NNIBP Conformations. Clustering was performed using two different techniques: single-linkage hierarchical clustering and complete linkage hierarchical clustering.[37] Single linkage forms clusters that are more connected, while complete linkage forms clusters that are optimally compact. However, in this case, both algorithms gave very similar results, which points to a clustering that is intrinsic to the data and not an artifact of the chosen method.

Several initial clustering experiments using different alignments, different atoms on which to perform the clustering analysis, and clustering of only torsion angles (which does not require alignment) were attempted. However, the results of these experiments were clouded by a large amount of noise.

Therefore, a more systematic approach to determine the alignment and clustering parameters was employed. The choice of atoms on which clustering was performed was based on the analysis of the backbone and side chain fluctuations above. The Cα atoms of residues associated with the primer grip region gave the highest radii of gyration across the 82 RT/NNRTI complexes (see Figure 2) and were therefore chosen for clustering. This corresponds to the Cα atoms of residues 224−242. Side chains were also chosen by reviewing their fluctuation analysis above. Side chains that gave clustering levels with both high minimum separation ratios and minimum distances between clusters were picked. This corresponds to the side chains of residues 181, 183, and 188. As all of these residues are tyrosines in the WT form of RT and either cysteines or leucines when mutated, the χ_1 dihedral angle was chosen for clustering these side chains. Clustering on a dihedral angle also alleviates the need for alignment of the structures. The best clustering was chosen as that which gave the highest MSR and a minimum rmsd between clusters of greater than 1 Å.

Acknowledgment. This work was supported by NIH Grants AI27690 (MERIT award to E.A.) and GM30580 (to R.M.L.).

References

(1) Kohlstaedt, L. A.; Wang, J.; Friedman, J. M.; Rice, P. A.; Steitz, T. A. Crystal structure at 3.5 Å resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science* **1992**, *256*, 1783–1790.
(2) Tantillo, C.; Ding, J.; Jacobo-Molina, A.; Nanni, R. G.; Boyer, P. L.; Hughes, S. H.; Pauwels, R.; Andries, K.; Janssen, P. A.; Arnold, E. Locations of anti-AIDS drug binding sites and resistance mutations in the three-dimensional structure of HIV-1 reverse transcriptase. Implications for mechanisms of drug inhibition and resistance. *J. Mol. Biol.* **1994**, *243*, 369–387.
(3) Shen, L.; Shen, J.; Luo, X.; Cheng, F.; Xu, Y.; Chen, K.; Arnold, E.; Ding, J.; Jiang, H. Steered molecular dynamics simulation on the binding of NNRTI to HIV-1 RT. *Biophys. J.* **2003**, *84*, 3547–3563.

(4) Temiz, N. A.; Bahar, I. Inhibitor binding alters the directions of domain motions in HIV-1 reverse transcriptase. *Proteins* **2002**, *49*, 61–70.

(5) Esnouf, R.; Ren, J.; Ross, C.; Jones, Y.; Stammers, D.; Stuart, D. Mechanism of inhibition of HIV-1 reverse transcriptase by non-nucleoside inhibitors. *Nat. Struct. Biol.* **1995**, *2*, 303–308.

(6) Ding, J.; Das, K.; Hsiou, Y.; Sarafianos, S. G.; Clark, A. D., Jr.; Jacobo-Molina, A.; Tantillo, C.; Hughes, S. H.; Arnold, E. Structure and functional implications of the polymerase active site region in a complex of HIV-1 RT with a double-stranded DNA template-primer and an antibody Fab fragment at 2.8 Å resolution. *J. Mol. Biol.* **1998**, *284*, 1095–1111.

(7) Jacobo-Molina, A.; Ding, J.; Nanni, R. G.; Clark, A. D., Jr.; Lu, X.; Tantillo, C.; Williams, R. L.; Kamer, G.; Ferris, A. L.; Clark, P.; Hizi, A.; Hughes, S. H.; Arnold, E. Crystal structure of human immunodeficiency virus type 1 reyerse transcriptase complexed with double-stranded DNA at 3.0 Å resolution shows bent DNA. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 6320–6324.

(8) Das, K.; Ding, J.; Hsiou, Y.; Clark, A. D., Jr.; Moereels, H.; Koymans, L.; Andries, K.; Pauwels, R.; Janssen, P. A. J.; Boyer, P. L.; Clark, P.; Smith, R. H., Jr.; Kroeger Smith, M. B.; Michejda, C. J.; Hughes, S. H.; Arnold, E. Crystal structures of 8-Cl and 9-Cl TIBO complexed with wild-type HIV-1 RT and 8-Cl TIBO complexed with the Tyr181Cys HIV-1 RT drug-resistant mutant. *J. Mol. Biol.* **1996**, *264*, 1085–1100.

(9) Ding, J.; Das, K.; Tantillo, C.; Zhang, W.; Clark, A. D., Jr.; Jessen, S.; Lu, X.; Hsiou, Y.; Jacobo-Molina, A.; Andries, K.; Pauwels, R.; Moereels, H.; Koymans, L.; Janssen, P. A. J.; Smith, R. H., Jr.; Koepke, M. K.; Michejda, C. J.; Hughes, S. H.; Arnold, E. Structure of HIV-1 reverse transcriptase in a complex with the non-nucleoside inhibitor alpha-APA R 95845 at 2.8 A resolution. *Structure* **1995**, *3*, 365–379.

(10) Hsiou, Y.; Ding, J.; Das, K.; Clark, A. D., Jr.; Hughes, S. H.; Arnold, E. Structure of unliganded HIV-1 reverse transcriptase at 2.7 Å resolution: implications of conformational changes for polymerization and inhibition mechanisms. *Structure* **1996**, *4*, 853–860.

(11) Rodgers, D. W.; Harrison, S. C. The structure of unliganded reverse transcriptase from the human immunodeficiency virus type 1. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 1222–1226.

(12) Das, K.; Clark, A. D., Jr.; Lewi, P. J.; Heeres, J.; De Jonge, M. R.; Koymans, L. M.; Vinkers, H. M.; Daeyaert, F.; Ludovici, D. W.; Kukla, M. J.; De Corte, B.; Kavash, R. W.; Ho, C. Y.; Ye, H.; Lichtenstein, M. A.; Andries, K.; Pauwels, R.; De Béthune, M. P.; Boyer, P. L.; Clark, P.; Hughes, S. H.; Janssen, P. A.; Arnold, E. Roles of conformational and positional adaptability in structure-based design of TMC125-R165335 (etravirine) and related non-nucleoside reverse transcriptase inhibitors that are highly potent and effective against wild-type and drug-resistant HIV-1 variants. *J. Med. Chem.* **2004**, *47*, 2550–2560.

(13) Das, K.; Bauman, J. D.; Clark, A. D., Jr.; Frenkel, Y. V.; Lewi, P. J.; Shatkin, A. J.; Hughes, S. H.; Arnold, E. High-resolution structures of HIV-1 reverse transcriptase/TMC278 complexes: strategic flexibility explains potency against resistance mutations. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 1466–1471.

(14) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(15) Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545–600.

(16) Becker, O. M.; Karplus, M. The topology of multidimensional potential energy surfaces: theory and application to peptide structure and kinetics. *J. Chem. Phys.* **1997**, *106*, 1495–1517.

(17) Dill, K. A.; Chan, H. S. From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **1997**, *4*, 10–19.

(18) Tsai, C. J.; Kumar, S.; Ma, B.; Nussinov, R. Folding funnels, binding funnels, and protein function. *Protein Sci.* **1999**, *8*, 1181–1190.

(19) Ravindranathan, K. P.; Gallicchio, E.; Levy, R. M. Conformational equilibria and free energy profiles for the allosteric transition of the ribose-binding protein. *J. Mol. Biol.* **2005**, *353*, 196–210.

(20) Xu, Q. S.; Daeyaert, F.; Lewi, P. J.; Massart, D. L. Studies of relationship between biological activities and HIV reverse transcriptase inhibitors by multivariate adaptive regression splines with curds and whey. *Chemom. Intell. Lab. Syst.* **2006**, *82*, 24–30.

(21) O'Brien, S. E.; Brown, D. G.; Mills, J. E.; Phillips, C.; Morris, G. Computational tools for the analysis and visualization of multiple protein−ligand complexes. *J. Mol. Graphics Modell.* **2005**, *24*, 186–194.

(22) Himmel, D. M.; Sarafianos, S. G.; Dharmasena, S.; Hossain, M. M.; McCoy-Simandle, K.; Ilina, T.; Clark, A. D., Jr.; Knight, J. L.; Julias, J. G.; Clark, P. K.; Krogh-Jespersen, K.; Levy, R. M.; Hughes, S. H.; Parniak, M. A.; Arnold, E. HIV-1 reverse transcriptase structure with RNase H inhibitor dihydroxy benzoyl naphthyl hydrazone bound at a novel site. *ACS Chem. Biol.* **2006**, *1*, 702–712.

(23) Das, K.; Lewi, P. J.; Hughes, S. H.; Arnold, E. Crystallography and the design of anti-AIDS drugs: conformational flexibility and positional adaptability are important in the design of non-nucleoside HIV-1 reverse transcriptase inhibitors. *Prog. Biophys. Mol. Biol.* **2005**, *88*, 209–231.

(24) Das, K.; Sarafianos, S. G.; Clark, A. D., Jr.; Boyer, P. L.; Hughes, S. H.; Arnold, E. Crystal structures of clinically relevant Lys103Asn/Tyr181Cys double mutant HIV-1 reverse transcriptase in complexes with ATP and non-nucleoside inhibitor HBY 097. *J. Mol. Biol.* **2007**, *365*, 77–89.

(25) Ren, J.; Nichols, C.; Bird, L.; Fujiwara, T.; Sugimoto, H.; Stuart, D. I.; Stammers, D. K. Binding of the second generation non-nucleoside inhibitor S-1153 to HIV-1 reverse transcriptase involves extensive main chain hydrogen bonding. *J. Biol. Chem.* **2000**, *275*, 14316–14320.

(26) Lewi, P. J.; de Jonge, M.; Daeyaert, F.; Koymans, L.; Vinkers, M.; Heeres, J.; Janssen, P. A.; Arnold, E.; Das, K.; Clark, A. D., Jr.; Hughes, S. H.; Boyer, P. L.; de Béthune, M. P.; Pauwels, R.; Andries, K.; Kukla, M.; Ludovici, D.; De Corte, B.; Kavash, R.; Ho, C. On the detection of multiple-binding modes of ligands to proteins, from biological, structural, and modeling data. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 129–134.

(27) Himmel, D. M.; Das, K.; Clark, A. D., Jr.; Hughes, S. H.; Benjahad, A.; Oumouch, S.; Guillemont, J.; Coupa, S.; Poncelet, A.; Csoka, I.; Meyer, C.; Andries, K.; Nguyen, C. H.; Grierson, D. S.; Arnold, E. Crystal structures for HIV-1 reverse transcriptase in complexes with three pyridinone derivatives: a new class of non-nucleoside inhibitors effective against a broad range of drug-resistant strains. *J. Med. Chem.* **2005**, *48*, 7582–7591.

(28) Das, K.; Bauman, J. D.; Arnold, E. Unpublished results.

(29) Fujiwara, T.; Sato, A.; el-Farrash, M.; Miki, S.; Abe, K.; Isaka, Y.; Kodama, M.; Wu, Y.; Chen, L. B.; Harada, H.; Sugimoto, H.; Hatanaka, M.; Hinuma, Y. S-1153 inhibits replication of known drug-resistant strains of human immunodeficiency virus type 1. *Antimicrob. Agents Chemother.* **1998**, *42*, 1340–1345.

(30) Janssen, P. A.; et al. In search of a novel anti-HIV drug: multidisciplinary coordination in the discovery of 4-[[4-[[4-[(1*E*)-2-cyanoethenyl]-2,6-dimethylphenyl]amino]-2-pyrimidinyl]amino]benzonitrile (R278474, rilpivirine). *J. Med. Chem.* **2005**, *48*, 1901–1909.

(31) Debnath, A. K. Application of 3D-QSAR techniques in anti-HIV-1 drug design—an overview. *Curr. Pharm. Des.* **2005**, *11*, 3091–3110.

(32) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(33) Ravindranathan, K. P.; Gallicchio, E.; Friesner, R. A.; McDermott, A. E.; Levy, R. M. Conformational equilibrium of cytochrome P450 BM-3 complexed with *N*-palmitoylglycine: a replica exchange molecular dynamics study. *J. Am. Chem. Soc.* **2006**, *128*, 5786–5791.

(34) Ren, J.; Esnouf, R.; Garman, E.; Somers, D.; Ross, C.; Kirby, I.; Keeling, J.; Darby, G.; Jones, Y.; Stuart, D.; Stammers, D. High resolution structures of HIV-1 RT from four RT-inhibitor complexes. *Nat. Struct. Biol.* **1995**, *2*, 293–302.

(35) Pata, J. D.; Stirtan, W. G.; Goldstein, S. W.; Steitz, T. A. Structure of HIV-1 reverse transcriptase bound to an inhibitor active against mutant reverse transcriptases resistant to other nonnucleoside inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 10548–10553.

(36) Spallarossa, A.; Cesarini, S.; Ranise, A.; Ponassi, M.; Unge, T.; Bolognesi, M. Crystal structures of HIV-1 reverse transcriptase complexes with thiocarbamate non-nucleoside inhibitors. *Biochem. Biophys. Res. Commun.* **2008**, *365*, 764–770.

(37) Johnson, S. C. Hierarchical clustering schemes. *Psychometrika* **1967**, *32*, 241–254.

(38) Shenkin, P. S.; McDonald, D. Q. Cluster analysis of molecular conformations. *J. Comput. Chem.* **1994**, *15*, 899–916.

(39) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–637.

References

Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235-242.

Das, K.; Clark, A.D., Jr.; Lewi, P.J.; Heeres, J.; De Jonge, M.R.; Koymans, L.M.; Vinkers, H.M.; Daeyaert, F.; Ludovici, D.W.; Kukla, M.J.; De Corte, B.; Kavash, R.W.; Ho, C.Y.; Ye, H.; Lichtenstein, M.A.; Andries, K.; Pauwels, R.; De Béthune, M.P.; Boyer, P.L.; Clark, P.; Hughes, S.H.; Janssen, P.A.; Arnold, E. Roles of conformational and positional adaptability in structure-based design of TMC125-R165335 (etravirine) and related non-nucleoside reverse transcriptase inhibitors that are highly potent and effective against wild-type and drug-resistant HIV-1 variants. *J. Med. Chem.* **2004**, *47*, 2550-2560.

Das, K,; Lewi, P.J.; Hughes, S.H.; Arnold, E. Crystallography and the design of anti-AIDS drugs: conformational flexibility and positional adaptability are important in the design of non-nucleoside HIV-1 reverse transcriptase inhibitors. *Prog. Biophys. Mol. Biol.* **2005**, *88*, 209-231.

Das, K.; Sarafianos, S.G.; Clark, A.D., Jr.; Boyer, P.L.; Hughes, S.H.; Arnold, E. Crystal structures of clinically relevant Lys103Asn/Tyr181Cys double mutant HIV-1 reverse transcriptase in complexes with ATP and non-nucleoside inhibitor HBY 097. *J. Mol. Biol.* **2007**, *365*, 77-89.

Das, K.; Bauman, J.D.; Clark, A.D., Jr.; Frenkel, Y.V.; Lewi, P.J.; Shatkin, A.J.; Hughes, S.H.; Arnold, E. High-resolution structures of HIV-1 reverse transcriptase/TMC278 complexes: strategic flexibility explains potency against resistance mutations. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 1466-1471.

O'Brien, S.E.; Brown, D.G.; Mills, J.E.; Phillips, C.; Morris, G. Computational tools for the analysis and visualization of multiple protein-ligand complexes. *J. Mol. Graph. Model.* **2005**, *24*, 186-194.

Pata, J.D.; Stirtan, W.G.; Goldstein, S.W.; Steitz, T.A. Structure of HIV-1 reverse transcriptase bound to an inhibitor active against mutant reverse transcriptases resistant to other nonnucleoside inhibitors. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 10548-10553.

Ren, J.; Esnouf, R.; Garman, E.; Somers, D.; Ross, C.; Kirby, I.; Keeling, J.; Darby, G.; Jones, Y.; Stuart, D.; Stammers, D. High resolution structures of HIV-1 RT from four RT-inhibitor complexes. *Nat. Struct. Biol*. **1995**, *2*, 293-302.

Ren, J.; Nichols, C.; Bird, L.; Fujiwara, T.; Sugimoto, H.; Stuart, D.I.; Stammers, D.K. Binding of the second generation non-nucleoside inhibitor S-1153 to HIV-1 reverse transcriptase involves extensive main chain hydrogen bonding. *J. Biol. Chem*. **2000**, *275*, 14316-14320.

Spallarossa, A.; Cesarini, S.; Ranise, A.; Ponassi, M.; Unge, T.; Bolognesi, M. Crystal structures of HIV-1 reverse transcriptase complexes with thiocarbamate non-nucleoside inhibitors. *Biochem. Biophys. Res. Commun*. **2008**, *365*, 764-770.

Xu, Q.S.; Daeyaert, F.; Lewi, P.J.; Massart, D.L. Studies of relationship between biological activities and HIV Reverse Transcriptase Inhibitors by Multivariate Adaptive Regression Splines with Curds and Whey. *Chemo. Intell. Lab. Sys*. **2006**, *82*, 24-30.

Chapter 7

Exploration of the HIV-1 Reverse Transcriptase Non-Nucleoside Inhibitor Binding Pocket With the Advanced Sampling Method Replica Exchange Molecular Dynamics

7.1 Replica Exchange Molecular Dynamics as a Sampling Method

Replica exchange molecular dynamics (REMD; Sugita and Okamoto, 1999; Felts et al. 2004) is an advanced sampling algorithm that may allow efficient sampling of the conformational and associated free energy landscape of biomolecular systems. REMD simulations have been used to study peptide folding (Paschek et al. 2007; Rhee and Pande, 2003; Nymeyer et al. 2004), NMR structure refinement (Chen et al. 2005), loop modeling (Felts et al. 2008; Velez-Vega et al. 2009) and ligand binding (Okumura et al. 2010; Ravindranathan et al. 2006). In this method, a number of simulations (replicas) are run in parallel over different specified temperatures. An exchange of adjacent replicas (T_i and T_j) is attempted periodically and is accepted based on the following Metropolis transition probability:

$$W\left\{T_i,T_j\right\}\rightarrow\left\{T_j,T_i\right\}=\min\left(1,\exp\left[-\left(\beta_j-\beta_i\right)\left(E_j-E_i\right)\right]\right) \tag{7.1}$$

where $\beta_{i(j)}=1/KT_{i(j)}$ and $E_{i(j)}$ is the potential energy of the *i*th (*j*th) replica. The exchanges allow rapid interconversion between stable conformations through high energy

intermediates sufficiently populated only at higher temperatures. Thus, REMD allows efficient exploration of the free energy landscape while giving a canonical (NVE) distribution of conformations at each temperature. REMD has been implemented in the IMPACT simulation package (Banks et al. 2005). Here, we perform simulations using the AGBNP implicit solvent model (Gallicchio and Levy, 2004) and the OPLS-AA force field (Jorgenson et al. 1996; Kaminski et al. 2001).

Based on the clustering study of the human immunodeficiency virus type 1 (HIV-1) enzyme reverse transcriptase (RT) non-nucleoside inhibitor binding pocket (NNIBP) in the previous chapter, two conformations were chosen as starting structures: the substrate-bound 2HMI and the capravirine-bound structure with the largest deviation in the primer grip region: 1EP4. Two temperature ranges were also utilized. Starting from 1EP4, 20 temperatures (and therefore 20 replicas) were used: 298, 313, 328, 344, 360, 376, 392, 408, 424, 440, 456, 472, 488, 504, 520, 536, 552, 568, 584, and 600 K. Starting from 2HMI, 30 temperature (and therefore 30 replicas) were used: 298, 307, 316, 326, 335, 345, 356, 366, 377, 388, 400, 412, 424, 437, 450, 464, 477, 492, 506, 521, 537, 553, 570, 587, 604, 622, 641, 660, 680, and 700K. The higher temperatures were utilized for 2HMI since it exhibits a closed NNIBP and may require higher temperatures to breach the barriers necessary to open the pocket.

As the HIV-1 RT enzyme is very large, it was truncated for simulation using a simple distance restriction: only residues within 20 Å of any ligand studied in Chapter 6 were included. This was further simplified for simulation by designation of free, buffered (harmonically restrained), and fixed atoms. The 2HMI simulations used a distance cutoff from Trp229: 81 residues within 15 Å of W229, including all of the

primer grip region, were free; 21 residues between 15 - 17 Å of W229 were placed in harmonically constrained buffer; and 97 residues that were greater than 15 Å from W229 were held fixed. The 1EP4 simulation was designed based on the NNIBP X-ray structure clustering in Chapter 6. Regions shown to be very variable were labeled as free and those that were shown to be rather rigid were held fixed. This resulted in 55 free residues which include the primer grip, residues in the β15 strand that backs the primer grip, the P95 loop, which includes important residues 100-103 and side chains that stick into the

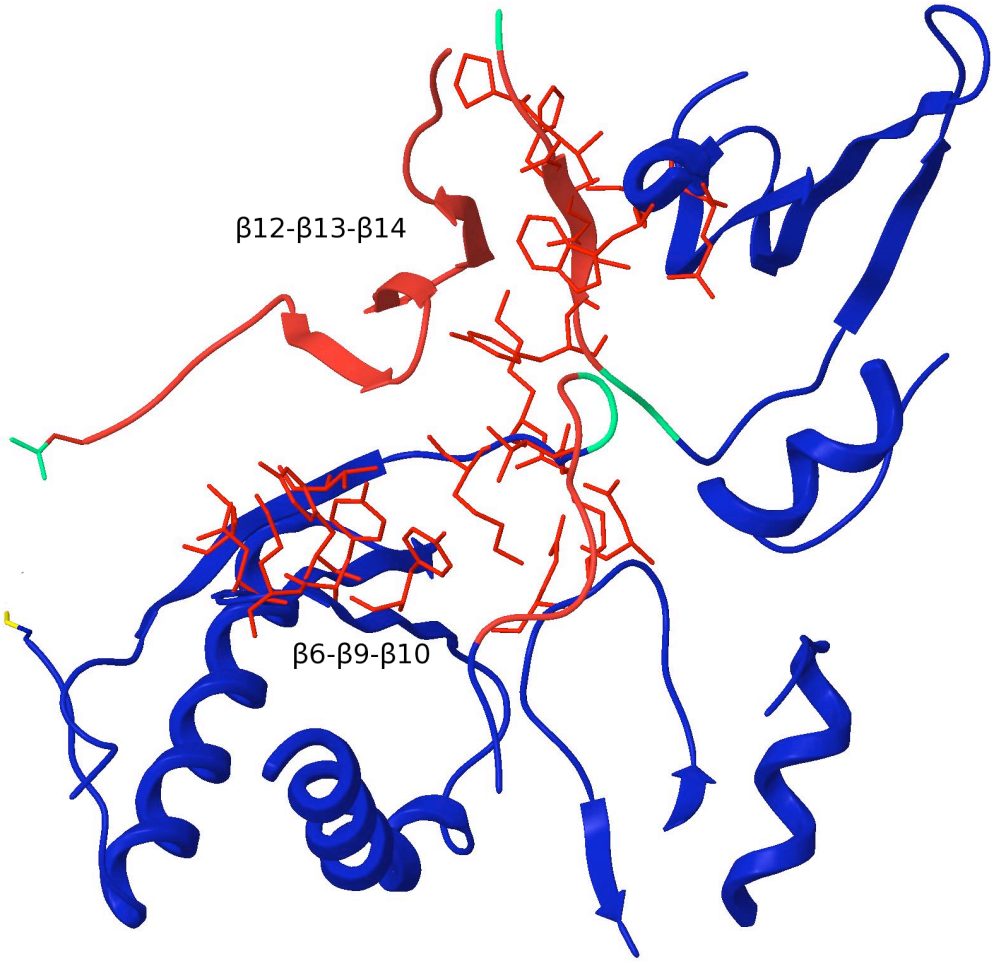

Figure 7.1. Designation of free, buffered and fixed regions for REMD starting from PDB id 1EP4. Regions were designed based on information obtained from clustering of 99 X-ray structures. Dark blue: fixed; Cyan: buffer; Red: free.

binding pocket: V106, V108, Y184, Y183, Y181, Y188, W266, Q269, and I270 in p66 and D136 and E138 in p51. 162 residues were held fixed including the β6-β9-β10 sheet that was used for superpositioning of X-ray structures in Chapter 6. Lastly, seven residues in regions between free and fixed were harmonically restrained in buffer. Figure 7.1 diagrams the constraints for 1EP4. The 1EP4 simulation was run for 5ns while the 2HMI simulation was run for only 1ns as it includes higher temperatures, which were expected to speed up the sampling of the conformational landscape.

7.2 Comparison With a Benchmark

	1EP4	*2HMI*
Average RMSD from start	2.03	2.31
Min RMSD from start	0.62	1.11
Max RMSD from start	5.02	3.13
RMSD to 1DLO	3.30	1.77
RMSD to 2OPP	1.48	2.09
RMSD to 2ZD1	1.50	4.83
RMSD to 1TV6	1.40	1.40
RMSD to 1S9E	0.77	3.14
RMSD to 2I5J	1.28	1.46
RMSD to 1RT1	1.99	3.54
RMSD to 1EP4	0.62	4.29

Table 7.1. Comparison of simulation results with experimental representative structures. Representative structures include 1DLO, an apo structure; 2OPP, the representative structure from the large cluster; 2ZD1, RT bound to the novel TMC278 inhibitor with an extended primer grip orientation; 1TV6, 1S9E, 2I5J, 1RT1, all members of the large cluster that sit on one of the outer skirts of the smear of conformations exhibited in that cluster; and 1EP4, the capravirine-bound structure with the largest observed deviation in primer grip conformation. RMSD fluctuation from the starting structures is also presented.

Results form both simulations were compared with the benchmark X-ray structure clustering from Chapter 6 to test whether REMD is capable, under the restrictions placed on the simulations listed above, to sample important conformations of the HIV-1 NNIBP. Root mean square deviation was calculated between each resulting conformation at any temperature from each simulation to selected representative structures chosen from the experimental landscape: 1DLO, an apo structure; 2OPP, the representative structure from

the large cluster; 2ZD1, RT bound to the novel TMC278 inhibitor with an extended primer grip orientation; 1TV6, 1S9E, 2I5J, 1RT1, all members of the large cluster that sit on one of the outer skirts of the smear of conformations exhibited in that cluster; and 1EP4, the capravirine-bound structure with the largest observed deviation in primer grip conformation.

Landscape cartoons like that created for the X-ray clustering are shown in Figure 7.2 and show results from the lowest temperature and highest temperature used. Both simulations appear to sample large areas of the NNIBP. Starting from 2HMI, regions close to the substrate-bound and ligand bound are highly populated at both high and low temperatures. The 1EP4 simulations offer a broader sampling for the NNIBP than that from 2HMI. The low temperature results show two highly populated basins in the area of the starting structure and close to (and partially overlapping) the large cluster seen in experiment. High temperature displays a highly populated valley in between the two low temperature basins. Using a cutoff of 1.5 Å rmsd, the 1ns 2HMI REMD was able to sample two experimentally-determined conformations found in the large cluster while the 5ns 1EP4 REMD "found" four out of the five representative X-ray structures in the large cluster as well as one of the TMC278-bound conformations (2ZD1) and the capravirine-bound 1EP4 conformation.

These "test" simulations show the need for the careful choice of starting structure as well as constraints and temperatures. Starting from 2HMI results in a large population of structures in a region not sampled by experiment. It is possible that with a longer simulation time, this population will shift to either the ligand-bound large cluster or back to substrate or apo form. Starting from 1EP4 shows a shift from the very open form of

the NNIBP seen in 1EP4 and the TMC278-bound conformations to a sampling of the outer regions of the large cluster region. With longer simulations, other conformations of the NNIBP may be found, such as the substrate and apo forms and the large cluster may be sampled more efficiently. It is interesting that these two experiments that start from essentially opposite sides of the NNIBP spectrum - the Closed and the most Open conformations - both manage to sample regions of space close to or overlapping the large cluster found experimentally. With 2HMI we see an opening of the pocket while with 1EP4 we see a partial closing of the pocket. It should be noted that previous restraints tested on 1EP4 where the β15 strand that backs the primer grip region was held fixed caused the structure to be "stuck" in the well surrounding the starting conformation. Therefore, it was with careful selection of freed regions surrounding and within the NNIBP that we were able to see larger conformational changes and the ability to climb the potential energetic barrier between the very open 1EP4 conformation and the large cluster.

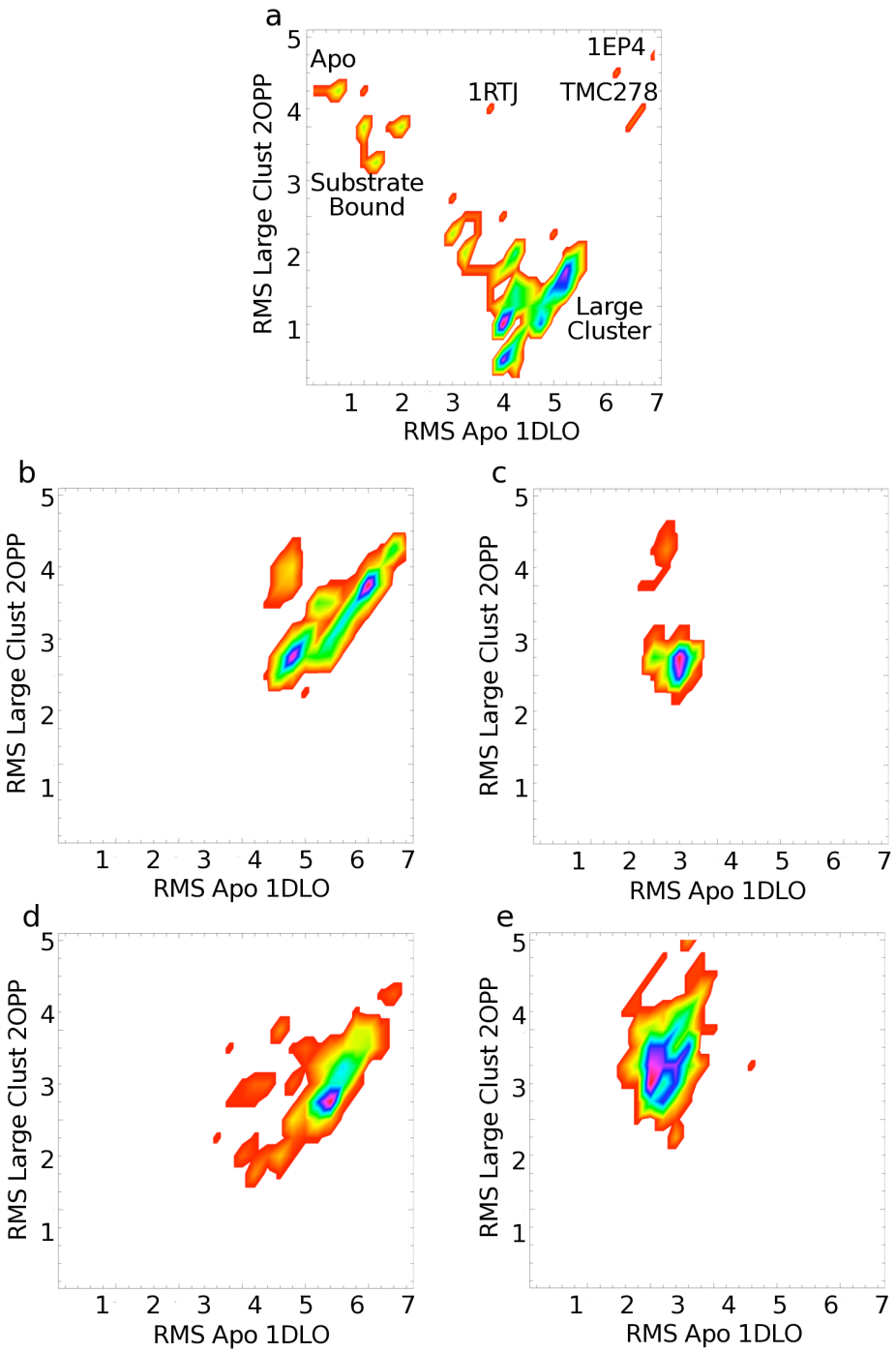

Figure 7.2. Conformation landscape cartoons for primer grip fluctuation from experiment and from simulations with coordinates selected to separate primer grip conformations as discussed in Chapter 6. RMSD in Å calculated from the representative X-ray structures 2OPP and 1DLO. (a) Experimental landscape created from analysis of 99 X-ray structures. The majority of the structures are found in the large, bound cluster. (b) Results from 5ns simulation starting from restricted 1EP4 at 298 K. (c) Results from 1ns simulation starting from substrate-bound 2HMI at 298 K. (d) 600 K structures from 1EP4 simulation. (e) 700 K structures from 2HMI simulation. Relative populations are colored from white (no population) to red (sparsely populate) to fuscia (high population).

References

Banks, J.L.; Beard, H.S.; Cao, Y.; Cho, A.E.; Damm, W.; Farid, R.; Felts, A.K.; Halgren, T.A.; Mainz, D.T.; Maple, J.R.; Murphy, R.; Philipp, D.M.; Repasky, M.P.; Zhang, L.Y.; Berne, B.J.; Friesner, R.A.; Gallicchio, E.; Levy, R.M. Integrated Modeling Program, Applied Chemical Theory (IMPACT). *J. Comput. Chem*. **2005**, *26*, 1752-80.

Chen, J.; Won, H.S.; Im, W.; Dyson, H.J.; Brooks, C.L., 3rd. Generation of native-like protein structures from limited NMR data, modern force fields and advanced conformational sampling. *J. Biomol. NMR*. **2005**, *31*, 59-64.

Felts, A.K.; Harano, Y.; Gallicchio, E.; Levy, R.M. Free energy surfaces of beta-hairpin and alpha-helical peptides generated by replica exchange molecular dynamics with the AGBNP implicit solvent model. *Proteins*. **2004**, *56*, 310-21.

Felts, A.K.; Gallicchio, E.; Chekmarev, D.; Paris, K.A.; Friesner, R.; Levy, R. Prediction of Protein Loop Conformations using the AGBNP Implicit Solvent Model and Torsion Angle Sampling. *J. Chem. Theory Comput.* **2008**, *4*, 855-868.

Gallicchio, E.; Levy, R.M. AGBNP: an analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *J. Comp. Chem.* **2004**, *25*, 479-499.

Jorgensen, W.L.; Maxwell, D.S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225-11236.

Kaminski, G.A.; Friesner, R.A.; Tirado-Rives, J.; Jorgensen, W.L. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **2001**, *105*, 6474-6487.

Nymeyer, H.; Gnanakaran, S.; García, A.E. Atomic simulations of protein folding, using the replica exchange algorithm. *Methods Enzymol*. **2004**, *383*, 119-49.

Okumura, H.; Gallicchio, E.; Levy, R.M. Conformational populations of ligand-sized molecules by replica exchange molecular dynamics and temperature reweighting. *J. Comput. Chem*. **2010**, *31*, 1357-67.

Paschek, D.; Nymeyer, H.; García, A.E. Replica exchange simulation of reversible folding/unfolding of the Trp-cage miniprotein in explicit solvent: on the structure and possible role of internal water. *J. Struct. Biol*. **2007**, *157*, 524-3.

Ravindranathan, K.P.; Gallicchio, E.; Friesner, R.A.; McDermott, A.E.; Levy, R.M. Conformational equilibrium of cytochrome P450 BM-3 complexed with N-palmitoylglycine: a replica exchange molecular dynamics study. *J. Am. Chem. Soc*. **2006**, *128*, 5786-91.

Rhee, Y.M.; Pande, V.S. Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophys. J*. **2003**, *84*, 775-86.

Sugita, Y.; Okamoto, Y. Replica exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141-151.

Velez-Vega, C.; Fenwick, M.K.; Escobedo, F.A. Simulated mutagenesis of the hypervariable loops of a llama VHH domain for the recovery of canonical conformations. *J. Phys. Chem. B*. **2009**, *113*, 1785-95.

Rhee, Y.M.; Pande, V.S. Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophys. J*. **2003**, *84*, 775-86.

Sugita, Y.; Okamoto, Y. Replica exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141-151.

Velez-Vega, C.; Fenwick, M.K.; Escobedo, F.A. Simulated mutagenesis of the hypervariable loops of a llama VHH domain for the recovery of canonical conformations. *J. Phys. Chem. B*. **2009**, *113*, 1785-95.

Chapter 8

Introduction to HIV-1 Protease Structure and Inhibition

8.1 HIV-1 Protease Structure

As discussed in Chapter 5, the viral enzyme protease (PR) plays an essential role in the life cycle of the human immunodeficiency virus (HIV). It generates mature virion particles through cleavage of the viral Gag and GagPol precursor proteins (Kohl et al., 1988). HIV-1 PR is composed of 99 amino acids and is a member of the aspartic acid PR family (Oroszlan and Luftig, 1990). However, unlike cellular PRs, this viral PR requires symmetric dimer formation for catalytic activity (Wlodawer and Erickson, 1993). PR is generally described as having the shape of a bull dog's head, with eyes, ears, nose and cheek (see Figure 8.1). The active site is formed along the dimer interface and the two active site residues (D25 and D25') are contributed by each monomer (Oroszlan and Luftig, 1990). Water acts as a nucleophile, in conjunction with the well-placed aspartic acids, to hydrolyze the scissile peptide bond to cleave viral peptides (Jaskólski et al., 1991). Binding of substrate causes the flaps of the dimer to move by as much as 7 Å in an opening and closing motion (Miller et al., 1989). To date, there are no X-ray structures demonstrating the completely open conformation of the flap but NMR data suggests that there are rapid fluctuations between open and closed forms of the flap in solution (Freedburg et al., 2002; Hornak and Simmerling, 2007).

As PR recognizes the asymmetric shape of peptide substrates rather than amino acid sequence, it does not require a particular amino acid sequence for cleavage. Instead, it cleaves peptides that have similar secondary structures that fit in a defined "substrate envelope" (Prabu-Jeyebalan et al., 2002). When PR binds a substrate, the structural symmetry of the homodimer is broken as the monomers adjust to accommodate the substrate. The substrate binding pocket or substrate envelope conformation can be divided into four main pockets: P1/P3, P2, P1'/P3' and P2' shown in Figure 8.1.

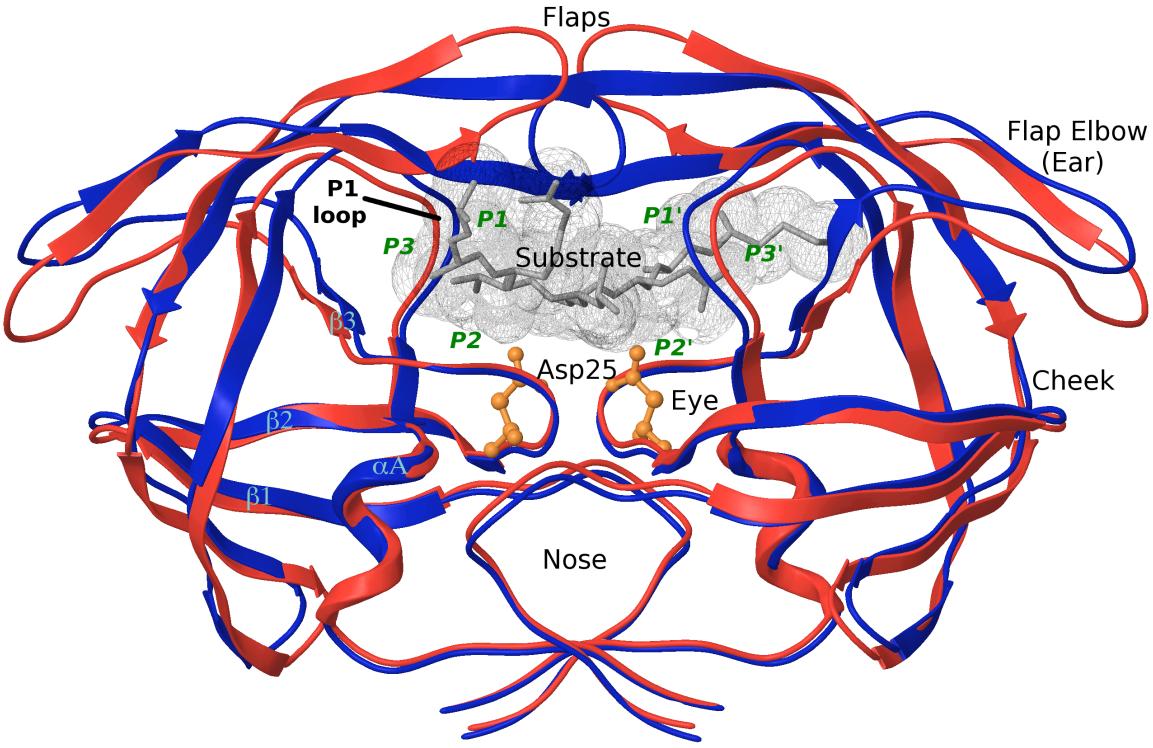

Figure 8.1. Cartoon of "semi-open" HIV-1 PR and substrate bound "closed" PR . Blue: PR bound to substrate with PDB id 1F7A. Substrate is shown in gray and gray mesh. Red: "Semi-open" unbound PR with PDB id 1PC0. PR is often seen as a bull-dog's face, with eyes, ears, cheek and nose as labeled. The active site residues D25 and D25' are shown in orange. Important regions P1, P1', P2, P2', P3, P3' and the P1 loop are labeled as well.

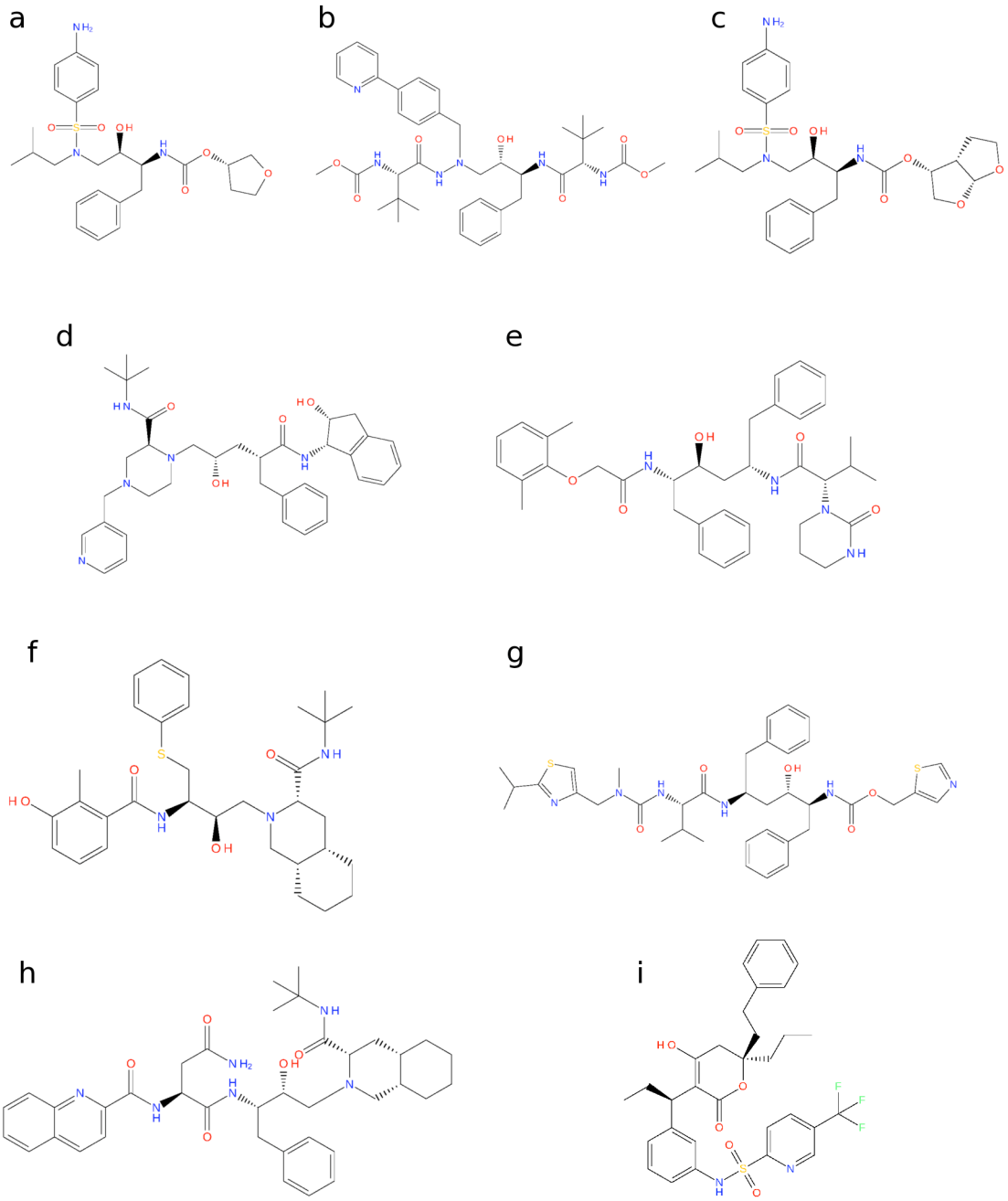

Figure 8.2. Inhibitors of HIV-1 PR. (a) Amprenavir (b) Atazanavir (c) Darunavir (d) Indinavir (e) Lopinavir (f) Nelfinavir (g) Ritonavir (h) Saquinavir (i) Tipranavir.

8.2 Mutation

Despite its critical function in the viral life cycle, HIV has shown great plasticity and mutations have been observed at one third of the 99 amino acid sites (Rhee et al., 2003). Some polymorphisms have been seen to occur naturally while keeping some regions invariant: the dimer interface, the active site floor, the P3-P3' substrate binding region and the β-hairpin loops of the flaps. However, upon exposure to various inhibitors, the enzyme develops a number of mutations to combat the ability of inhibitors to bind. The drug-resistant mutations are especially observed in the flap region and parts of the P1 loop and P3-P3' binding cleft (Galiano et al., 2009).

Only a subset of mutations affect inhibitor binding via alteration of a direct point of contact: D30N, G48V, V82A, I84V, I50V, and I50L. Some of these mutations are primarily associated with particular inhibitors. D30N is generally associated with nelfinavir, G48V with saquinavir, I50V with amprenavir and Darunavir, and I50L with atazanavir. V82A and I84V impact almost all inhibitors. Patients treated with a variety of PR inhibitors often experience between 5 and 15 mutations in the PR gene (Wu et al., 2003; Rhee et al., 2005). These mutations are often in specific combinations of mutations in the active site and compensatory mutations outside the active site (Shafer et al., 1999; Rhee et al., 2007; Hoffman et al., 2003). Some common sites outside the active site are L10I, I54V/T, A71V/T, V77I, and L90M. These mutations may not only impact inhibitor binding through allosteric conformational changes but may also compensate for the viability and fitness of the enzyme.

8.3 Inhibition

Detailed knowledge of the enzyme's structure and its interactions with substrate has led to the development of a number of PR inhibitors. Currently, there are nine PR inhibitors: saquinavir, ritonavir, indinavir, nelfinavir, amprenavir, lopinavir, atazanavir, tipranavir, and Darunavir. These ligands are depicted in Figure 8.2. They bind competitively, mimicking natural substrates. Generally, they contain a central hydroxyl group that increases affinity by interacting with the catalytic residues D25 and D25'. In addition, large hydrophobic groups on either side of the hydroxyl group bind in hydrophobic subsites and polar groups form hydrogen bonds with the enzyme. First generation inhibitors such as saquinavir and indinavir maximize hydrophobic interactions and their binding is entropically driven while most second

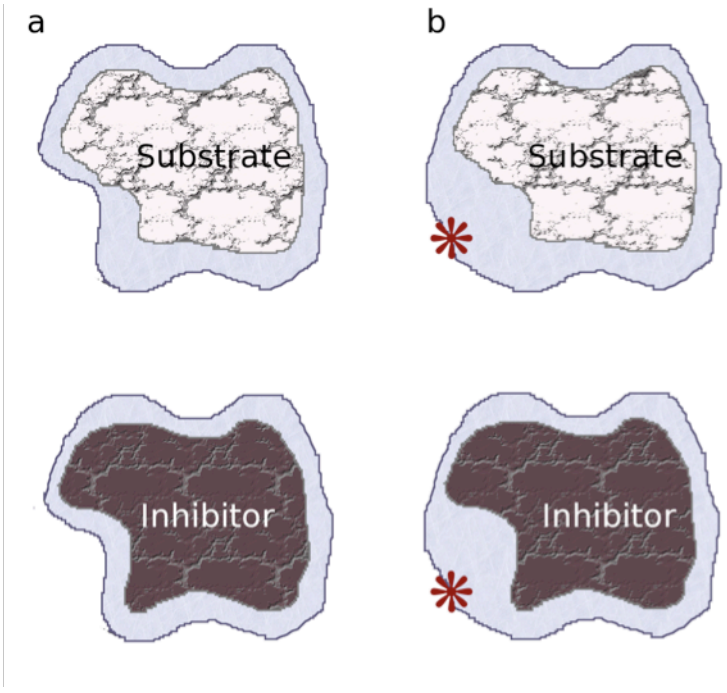

Figure 8.3. Illustration of the "substrate envelope" hypothesis. (a) In the wild-type receptor, the inhibitor (bottom) makes more contacts and occupies more space in the binding pocket than the substrate (top). (b) A mutation occurs in the receptor binding pocket that enlarges the pocket. The inhibitor suffers from lack of the contacts it once made while the substrate loses negligible affinity since it did not interact with that portion of the binding pocket in the wild-type form (Prabu-Jeyabalan et al., 2002; Altman et al., 2008).

generation inhibitors such as Darunavir have been designed to maximize polar interactions with main chain PR atoms and binding of these inhibitors is generally enthalpically driven (Velazquez-Campoy et al., 2000a,b; Freire, 2008).

The large degree of potential drug-induced mutation in PR makes drug design of PR inhibitors very difficult. One strategy is to design drugs that mimic the structural features of substrates (Prabu-Jeyabalan et al., 2002; King et al., 2004; Tuske et al., 2004; Atlman et al., 2008). Ideally, mutations would then render the enzyme inactive. The "substrate envelope hypothesis" illustrated in Figure 8.3 has taken the front row in this argument. Studies of substrate-PR complexes have shown a consensus substrate envelope (Prabu-Jeyabalan et al., 2002). Following the idea of mimicking substrate binding, inhibitors that are designed into this envelope-defined boundary should be less likely to induce resistant mutations than those that jut out from the envelope and provide regions where mutations could occur (Altman et al., 2008; Prabu-Jeyabalan et al., 2002). Design strategies using the substrate envelope hypothesis have been promising (Nalam et al., 2010; Altman et al., 2008). A recent study compared over 130 HIV-1 PR inhibitors designed with and without substrate-envelope constraints and found that those nanomolar to picomolar inhibitors that fit within the substrate envelope have flatter resistance profiles than those that do not fit within the substrate envelope (Nalam et al., 2010).

Knowledge of structural fluctuations of a target enzyme has been shown to be important for the understanding of the enzyme's function and for drug design (Nalam et al., 2010; Altman et al., 2008; Hornak and Simmerling, 2007; Prabu-Jeyabalan et al., 2002; Yang et al., 2008; Kurt et al., 2003; Zoete et al., 2002). The substrate envelope hypothesis discussed above was introduced based on a study of six complexes bound to

six decameric peptides that correspond to the cleavage sites within the Gag and Pol polyproteins (Prabu-Jeyabalan et al., 2002). In the study, the side chains of the P1/P3 pocket (R8, I47, F53, V82', and I84') were found to adopt a variety of conformations, but the P1'/P3' pocket side chains remain rather static with the exception of V82. There is also a larger backbone rearrangement in residues 45-50 (the flap region) and 78'-82' (the P1' loop). The P2 and P2' pockets (defined by N/D25, G27, A28, D29, and D30) also remain rather rigid (Prabu-Jeyebalan et al., 2002).

Other studies have also utilized multiple crystal structures in an effort to analyze PR flexibility and motions (Zoete et al., 2002; Kurt et al., 2003; Yang et al., 2008). A combination of comparison of 73 X-ray structures, molecular dynamics simulations, normal mode analyses, and X-ray *B*-factor analyses pointed toward a potential energy surface for HIV-1 PR that is characterized by many local minima with small energy differences (Zoete et al., 2002). The backbone root mean square deviation (rmsd) of the 73 inhibitor-bound structures was found to be unevenly distributed, with rigid regions around the active site triplet (residues 25-27) having an rmsd of 0.25 Å while most variable regions located in loops around residues 18, 40, 52-53 (the flap region), 68 and 82 have rmsds closer to 1 Å. Mutation was shown to have a minimal effect on structural fluctuation; location of the average structure only varied slightly with mutation (Zoete et al., 2002). A similar study in 2008 used principle component analysis (PCA) and an elastic network model (ENM) with 156 X-ray structures to obtain information about motions within the PR enzyme (Yang et al., 2008). Results suggested that the similar variations among the observed structures from PCA and the corresponding conformational changes from the normal modes from the ENM are facilitated by low-

frequency, global motions that are intrinsic to the enzyme. It also showed that a large number of experimental structures could directly provide important information about protein dynamics (Yang et al., 2008)

The global motions of PR have been studied in detail using computational methods such as MD and with analytical tools such as PCA, ENM, and Gaussian network models (GNM) (Kurt et al., 2003; Hornak and Simmerling, 2007; Piana et al., 2002a; Piana et al., 2002b; Perryman et al., 2004; Damm et al., 2008; Hornak et al., 2006a; Hornak et al., 2006b). Classical and *ab initio* MD simulations of various mutants reveal that PR flexibility modulates the activation free energy barrier of the enzymatic cleavage reaction. Active-site mutations are often associated with compensatory mutations that enhance the catalytic rate of the mutant by affecting the flexibility of the protein (Piana et al., 2002a; Piana et al., 2002b). MD simulations of ligand-bound and unliganded apo structures have shown that the overall modes of motion of different conformations are generally conserved but the most mobile and least flexible regions differ between bound and unbound structures. The flaps and loop containing residue 40 of the unliganded structure are the most mobile regions. In the ligand-bound structure these regions lose mobility while the terminal regions become more flexible (Kurt et al., 2003). Later all-atom simulations of PR showed that introduction of a ligand to an open apo structure caused the enzyme to spontaneously close to the bound conformation (Hornak et al., 2006b) and removal of the ligand favored the semi-open conformation of the enzyme (Hornak et al., 2006a). Discussion of the changes in flexibility upon binding has led to the possibility of utilizing allosteric inhibitors to control flexibility within PR (Perryman et al., 2003). Various promising regions are currently being studied including the Ear-

Cheek (or Elbow) region (Perryman et al., 2003; Perryman et al., 2006), the dimer interface (Hwang et al., 2005; Shultz et al., 2004), and the Eye region (Damm et al., 2008).

References

Altman, M.D.; Ali, A.; Reddy, G.S.; Nalam, M.N.; Anjum, S.G.; Cao, H.; Chellappan, S.; Kairys, V.; Fernandes, M.X.; Gilson, M.K.; Schiffer, C.A.; Rana, T.M.; Tidor, B. HIV-1 protease inhibitors from inverse design in the substrate envelope exhibit subnanomolar binding to drug-resistant variants. *J. Am. Chem. Soc*. **2008**, *130*, 6099-113.

Damm, K.L.; Ung, P.M.U.; Quintero, J.J.; Gestwicki, J.E.; Carlson, H.A. A poke in the eye: Inhibiting HIV-1 protease through its flap-recognition pocket. *Biopolymers* **2008**, *89*, 643-652.

Freedberg, D.I.; Ishima, R.; Jacob, J.; Wang, Y.X.; Kustanovich, I.; Louis, J.M.; Torchia, D.A. Rapid structural fluctuations of the free HIV protease flaps in solution: relationship to crystal structures and comparison with predictions of dynamics calculations. *Protein Sci*. **2002**, *11*, 221-32.

Freire, E. Do enthalpy and entropy distinguish first in class from best in class? *Drug Disc. Today* **2008**, *13*,869-874.

Galiano, L.; Ding, F.; Veloro, A.M.; Blackburn, M.E.; Simmerling, C.; Fanucci, G.E. Drug pressure selected mutations in HIV-1 protease alter flap conformations. *J. Am. Chem. Soc*. **2009**, *21*, 430-1.

Hoffman, N.G.; Schiffer, C.A.; Swanstrom, R. Covariation of amino acid positions in HIV-1 protease. *Virology* **2003**, *314*, 536-548.

Hornak, V.; Okur, A.; Rizzo, R.C.; Simmerling, C. HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proc. Natl. Acad. Sci.* **2006**, *103*, 915-920.

Hornak, V.; Okur, A.; Rizzo, R.C.; Simmerling, C. HIV-1 protease flaps spontaneously close to the correct structure in simulations following manual placement of an inhibitor into the open state. *J. Am. Chem. Soc.* **2006**, *128*, 2812-2813.

Hornak, V.; Simmerling, C. Targeting structural flexibility in HIV-1 protease inhibitor binding. *Drug Disc. Today* **2007**, *12*, 132-138.

Hwang, Y.S.; Chmielewski, J. Development of low molecular weight HIV-1 protease dimerization inhibitors. *J. Med. Chem.* **2005**, *48*, 2239-2242.

King, N.M.; Prabu-Jeyabalan, M.; Nalivaika, E.A.; Wigerinck, P.; de Béthune, M.P.; Schiffer, C.A. Structural and thermodynamic basis for the binding of TMC114, a next-generation human immunodeficiency virus type 1 protease inhibitor. *J. Virol*. **2004**, *78*, 12012-21.

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

Kurt, N.; Scott, W.R.; Schiffer, C.A.; Haliloglu, T. Cooperative fluctuations of unliganded and substrate-bound HIV-1 protease: a structure-based analysis on a variety of conformations from crystallography and molecular dynamics simulations. *Proteins* **2003**, *51*, 409-22.

Jaskólski, M.; Tomasselli, A.G.; Sawyer, T.K.; Staples, D.G.; Heinrikson, R.L.; Schneider, J.; Kent, S.B.; Wlodawer, A. Structure at 2.5-A resolution of chemically synthesized human immunodeficiency virus type 1 protease complexed with a hydroxyethylene-based inhibitor. *Biochemistry* **1991**, *30*,1600–9.

Miller, M.; Schneider, J.; Sathyanarayana, B.K.; Toth, M.V.; Marshall, G.R.; Clawson, L.; Selk, L.; Kent, S.B.; Wlodawer, A. Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 A resolution. *Science* **1989**, *246*, 1149–52.

Nalam, M.N.; Ali, A.; Altman, M.D.; Reddy, G.S.; Chellappan, S.; Kairys, V.; Ozen, A.; Cao, H.; Gilson, M.K.; Tidor, B.; Rana, T.M.; Schiffer, C.A. Evaluating the substrate-envelope hypothesis: structural analysis of novel HIV-1 protease inhibitors designed to be robust against drug resistance. *J. Virol***. 2010**, *84*, 5368-5378.

Oroszlan, S.; Luftig, R.B. Retroviral proteinases. *Curr. Top. Microbiol. Immunol*. **1990**, *157*, 153-85.

Perryman, A.L.; Lin, J.; McCammon, J.A. HIV-1 protease molecular dynamics of a wild-type and of the V82F/I84V mutant: Possible contributions to drug resistance and a potential new target site for drugs. *Prot. Science* **2004**, *13*, 1108-1123.

Perryman, A.L.; Lin, J.; McCammon, J.A. Optimization and computational evaluation of a series of potential active site inhibitors of the V82F/I84V drug-resistant mutant of HIV-1 protease: an application of the relaxed complex method of structure-based drug design. *Chem. Biol. Drug Des.* **2006**, *67*, 336-345.

Piana, S.; Carloni, P.; Parrinello, M. Role of conformational fluctuations in the enzymatic reaction of HIV-1 protease. *J. Mol. Biol.* **2002**, *319*, 567-583.

Piana, S.; Carloni, P.; Rothlisberger, U. Drug resistance in HIV-1 protease: Flexibility-assisted mechanism of compensatory mutations. *Prot. Science* **2002**, *11*, 2393-2402.

Prabu-Jeyabalan, M.; Nalivaika, E.; Schiffer, C.A. Substrate shape determines specificity of recognition for HIV-1 protease: analysis of crystal structures of six substrate complexes. *Structure* **2002**, *10*, 369-381.

Rhee, S.Y.; Gonzales, M.J.; Kantor, R.; Betts, B.J.; Ravela, J.; Shafer, R.W. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucl. Acids Res*. **2003**, *31*, 298–303.

Rhee, S.Y.; Fessel, W.J.; Zolopa, A.R.; Hurley, L.; Liu, T.; Taylor, J.; Nguyen, D.P.; Slome, S.; Klein, D.; Horberg, M.; et al. HIV-1 protease and reverse-transcriptase mutations: correlations with antiretroviral therapy in subtype B isolates and implications for drug-resistance surveillance. *J. Infect. Dis.* **2005**, *192*, 456-465.

Rhee, S.Y.; Liu, T.F.; Holmes, S.P.; Shafer, R.W. HIV-1 subtype B protease and reverse transcriptase amino acid covariation. *PLoS Comput. Biol.* **2007**, *3*, e87.

Shafer, R.W.; Hsu, P.; Patrick, A.K.; Craig, C.; Brendel V. Ientification of biased amino acid substitution patterns in human immunodefiency virus type 1 isolates from patients treated with protease inhibitors. *J. Virol.* **1999**, *73*, 6197-6202.

Shultz, M.D.; Ham, Y.; Lee, S.; Davis, D.A.; Brown, C.; Chmielewski, J. Small-molecule dimerization inhibitors of wild-type and mutant HIV protease: a focused library approach. *J. Am. Chem. Soc.* **2004**, *126*, 9886-9887.

Tuske, S.; Sarafianos, S.G.; Clark, A.D., Jr.; Ding, J.; Naeger, L.K.; White, K.L.; Miller, M.D.; Gibbs, C.S.; Boyer, P.L.; Clark, P.; Wang, G.; Gaffney, B.L.; Jones, R.A.; Jerina, D.M.; Hughes, S.H.; Arnold, E. Structures of HIV-1 RT-DNA complexes before and after incorporation of the anti-AIDS drug tenofovir. *Nat. Struct. Mol. Biol*. **2004**, *11*, 469-74

Velazquez-Campoy, A.; Luque, I.; Todd, M.J.; Milutinovich, M.; Kiso, Y.; Freire, E. Thermodynamic dissection of the binding energetics of KNI-272, a potent HIV-1 protease inhibitor. *Protein Sci*. **2000**, *9*, 1801-9.

Velazquez-Campoy, A.; Todd, M.J.; Freire, E. HIV-1 protease inhibitors: enthalpic versus entropic optimization of the binding affinity. *Biochemistry*, **2000**, *39*, 2201-7.

Wlodawer, A.; Erickson, J.W. Structure-based inhibitors of HIV-1 protease. *Annu. Rev. Biochem*. **1993**, *62*, 543-85.

Wu, T.D.; Schiffer, C.A.; Gonzales, M.J.; Taylor, J.; Kantor, R.; Chou, S.; Israelski, D.; Fessel, W.J.; Shafer, R.W. Mutation patterns and structural correlates in human immunodefiency virus type 1 protease following different protease inhibitor treatments. *J. Virol.* **2003**, *77*, 4836-4847.

Yang, L.; Song, G.; Carriquiry, A.; Jernigan, R.L. Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. *Structure* **2008**, *16*, 321-30.

Zoete, V.; Michielin, O.; Karplus, M. Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. *J. Mo.l Biol*. **2002**, *315*, 21-52.

Chapter 9

Estimation of the Conformational Landscape for Binding of HIV-1 Protease

9.1 Introduction

Present day crystallization techniques have allowed determination of a large number of X-ray structures, which are deposited in the publically available Protein Data Bank (PDB; Berman et al., 2000). As there are many different HIV-1 protease (PR) inhibitors, as well as many mutational variations of the enzyme, it is reasonable to assume that the many associated X-ray structures may be used to construct a rough conformational landscape for binding to HIV-1 PR as was previously done for HIV-1 reverse transcriptase (RT) in Chapter 6. Unlike RT, much work has been done in attempts to study ligand binding to PR, especially on the front of receptor reorganization (Prabu-Jeyebalan et al., 2002; Yang et al., 2008; Zoete et al., 2002, Kurt et al., 2003; Piana et al., 2002; Damm et al., 2008; Hornak et al., 2006a; Hornak et al., 2006b; Hornak and Simmerling, 2007). This is most likely a result of the high frequency of mutation of the enzyme that has the potential to drastically change the atmosphere of the binding pocket.

An important study by Prabu-Jeyabalan et al. compared and contrasted six substrate-bound PR and found that main chain hydrogen bonds and much of the binding pocket conformation is conserved (2002). However, no substrate side chain hydrogen bond is conserved. Several regions of the receptor pocket are also distorted, including the

flap regions, P1/P3 pocket and P1 loop. This led to the development of the "substrate envelope" hypothesis where it is believed that ligands that form less interactions with the receptor and occupy less space within the binding "envelope" are preferred as they will be less effected by potential mutational variation (Altman et al., 2008; Prabu-Jeyabalan et al., 2002).

Analysis of multiple conformations has also assisted in understanding the dynamics and function of HIV PR. It has been proposed by use of a combination of molecular dynamics (MD) simulations, X-ray structures and a Gaussian network model (GNM) that binding of ligands reduces the mobility in the flap and 40's loop regions (Kurt et al., 2003). Several studies have focused on identifying essential motions and regions of flexibility of PR. Zoete et al. examined 73 X-ray structures and compared the flexible regions reported by experimental B factors to MD predicted regions using root mean square deviations (rmsds) and normal mode analysis (NMA) as tools (2002). In 2008, a further study looked at 150 available X-ray structures, an NMR ensemble of 28 models and structures from a 10ns MD simulation and identified key motions of the enzyme using principle component analysis (PCA) and an elastic network model (ENM) (Yang et al., 2008). Both studies found that the observed motions are intrinsic to the nature of the enzyme.

In this current work, we look at 327 available experimental (X-ray and NMR) structures in an attempt to produce a rough conformational landscape for binding to HIV-1 PR using hierarchical clustering techniques. All forms of PR were included: apo, substrate-bound and ligand-bound. The PDB ids and class (bound, apo) are listed in Table 9.1. As the many complexes of PR bound to asymmetric inhibitors do not uniquely

orient the crystal cell, we separated dimers into monomers so that 327 structures became 629 monomers. (There is not an exact doubling in number as several of the original 327 structures only included chain A and several others included more than two differing chains in the PDB entry). Careful preparation of the structural data set was essential as simply clustering on the unedited set produced mostly noise. X-ray crystal structures from the PDB were first analyzed to determine an atom sequence common to all structures. Entries that were found to be missing a large amount of structural information were removed to leave the data set of 629 monomeric structures used in this study. The 629 structures were then renumbered and reordered to follow the common atom sequence discovered from the analysis of each entry. This allowed for a normalization of the entries. Residues that experienced mutations were stripped of their side-chain atoms that were not shared by each residue type. For example, if a residue is found as either a lysine or an asparagine, only the backbone atoms and Cβ and Cγ of the side chain were included.

A fluctuation analysis of the backbones of all 99 residues was performed by aligning the ensemble of structures based on the Cα of each of the 99 residues and calculating the radius of gyration (R_g) of the point cloud of all the positions for each Cα atom in the entire ensemble of structures. A low R_g coincides with little movement of the position of the atom across conformers whereas a high R_g coincides with an atom that takes on many different positions in the ensemble.

Fluctuation of the side chains that point into the binding pocket (residues 8, 23, 25, 29, 32, 45, 47, 50, 54, 56, 58, 76, 81, 84, and 87) was analyzed by clustering on each side chain individually using single-linkage hierarchical clustering (Shenkin and

McDonald, 1994; Johnson, 1967). The best clustering was chosen as that which gave the highest minimum separation ratio (MSR), an empirical measure of the degree of separation.

Clustering results are partially dependent on the alignment of conformers. In this study, the structures were aligned on the Cα atoms of residues 10-14, 20-24, 31-34, and 87-93 that correspond to β1, β2, β3, and αA respectively. This alignment was chosen based on the backbone analysis above since their Cα atoms have low R_gs across the PR structures. (See Figure 9.1). Clustering was performed using two different techniques: single-linkage hierarchical clustering and complete linkage hierarchical clustering (Johnson, 1967). Single linkage forms clusters that are more connected while complete linkage forms clusters that are optimally compact. However, in this case, both algorithms gave very similar results, which points to a clustering that is intrinsic to the data and not an artifact of the chosen method.

The choice of atoms on which clustering was performed was based on the analysis of the backbone and side chain fluctuations of residues surrounding the binding pocket above. The Cα atoms of residues associated with the flap and P1 loop region gave the highest radii of gyration across the 629 structures (see Figure 9.1) and were therefore chosen for clustering. This corresponds to the Cα atoms of residues 49-51 and 80-82, respectively. Side chains were also chosen by reviewing their fluctuation analysis. Only one side chain demonstrated a clustering level with both high minimum separation ratio and minimum distance between clusters: that of R8, a highly conserved residue that is important for dimer formation. A manufactured dihedral angle was chosen for clustering R8 using the C, Cα, Cβ and Cζ. Clustering on a dihedral angle also alleviates

	PDB ids
Apo	1hhp_a; 1rpi_a,b; 2g69_a; 2hb4_a; 2pc0_a; 3hvp_a; 3phv_a
Bound	1a30_a,b; 1a8g_a,b; 1a8k_a,b; 1a94_a,b; 1a9m_a,b; 1aaq_a,b; 1aid_a,b; 1ajv_a,b; 1ajx_a,b; 1axa_a,b; 1b6j_a,b; 1b6k_a,b; 1b6l_a,b; 1b6m_a,b; 1b6p_a,b; 1bdl_a,b; 1bdq_a,b; 1bdr_a,b; 1bv7_a,b; 1bv9_a,b; 1bvg_a,b; 1bwa_a,b; 1bwb_a,b; 1c6x_a,b; 1c6y_a,b; 1c6z_a,b; 1c70_a,b; 1cpi_a,b; 1d4h_a,b; 1d4i_a,b; 1d4j_a,b; 1d4k_a,b; 1d4l_a,b; 1d4s_a,b; 1d4y_a,b; 1dif_a,b; 1dmp_a,b; 1ebw_a,b; 1eby_a,b; 1ebz_a,b; 1ec0_a,b; 1ec1_a,b; 1ec2_a,b; 1ec3_a,b; 1f7a_a,b; 1fb7_a, 1fqx_a,b; 1g2k_a,b; 1g35_a,b; 1gnm_a,b; 1gnn_a,b; 1gno_a,b; 1hbv_a,b; 1hih_a,b; 1hiv_a,b; 1hos_a,b; 1hpo_a,b; 1hps_a,b; 1hpv_a,b; 1hpx_a,b; 1hsg_a,b; 1hte_a,b; 1htf_a,b; 1htg_a,b; 1hvh_a,b; 1hvi_a,,b; 1hvj_a,b; 1hvk_a,b; 1hvl_a,b; 1hvr_a,b; 1hvs_a,b; 1hwr_a,b; 1hxb_a,b; 1hxw_a,b; 1iiq_a,b; 1izh_a,b; 1izi_a,b; 1k1t_a,b; 1k1u_a,b; 1k2b_a,b; 1k2c_a,b; 1k6c_a,b; 1k6p_a,b; 1k6t_a,b; 1k6v_a,b; 1kj4_a,b; 1kj7_a,b; 1kjf_a,b; 1kjg_a,b; 1kjh_a,b; 1kzk_a,b; 1lzq_a,b; 1m0b_a,b; 1mer_a,b; 1mes_a,b; 1met_a,b; 1meu_a,b; 1mrw_a,b; 1mrx_a,b; 1msm_a,b; 1msn_a,b; 1mt7_a,b; 1mt8_a,b; 1mt9_a,b; 1mtb_a,b; 1mtr_a,b; 1mui_a,b; 1n49_a,b; 1nh0_a,b; 1npa_a,b; 1npv_a,b; 1npw_a,b; 1odw_a,b; 1odx_a,b; 1ody_a,b; 1ohr_a,b; 1pro_a,b; 1qbr_a,b; 1qbs_a,b; 1qbt_a,b; 1qbu_a,b; 1rl8_a,b; 1rq9_a,b; 1rv7_a,b; 1sbg_a,b; 1sdt_a,b; 1sdu_a,b; 1sdv_a,b; 1sgu_a,b; 1sh9_a,b; 1sp5_a,b; 1t3r_a,b; 1t7i_a,b; 1t7j_a,b; 1t7k_a,b; 1tcx_a,b; 1tsq_a,b; 1tsu_a,b; 1tw7_a,b; 1u8g_a,b; 1upj_a; 1vij_a,b; 1vik_a,b; 1w5v_a,b; 1w5w_a,b; 1w5x_a,b; 1w5y_a,b; 1wbk_a,b; 1wbm_a,b; 1xl2_a,b; 1xl5_a,b; 1yt9_a,b; 1z1h_a,b; 1z1r_a,b; 1z8c_a,b; 1zbg_a,b; 1zj7_a; 1zlf_a,b; 1zp8_a; 1zpa_a; 1zpk_a,b; 1zsf_a,b; 1zsr_a,b; 1ztz_a,b; 2a1e_a,b; 2a4f_a,b; 2aid_a,b; 2aoc_a,b; 2aod_a,b; 2aoe_a,b; 2aof_a,b; 2aog_a,b; 2aoh_a,b; 2aoi_a,b; 2aoj_a,b; 2aqu_a,b; 2avm_a,b; 2avo_a,b; 2avq_a,b; 2avs_a,b; 2avv_a,b; 2az8_a; 2az9_a; 2azc_a,b; 2b60_a,b; 2b7z_a,b; 2bb9_a,b; 2bbb_a,b; 2bpv_a,b; 2bpw_a,b; 2bpx_a,b; 2bpy_a,b; 2bpz_a,b; 2bqv_a,b; 2cej_a,b; 2cem_a,b; 2cen_a,b; 2f3k_a,b; 2f80_a,b; 2f81_a,b; 2f8g_a,b; 2fdd_a,b; 2fde_a,b; 2fgu_a,b; 2fgv_a,b; 2fle_a,b; 2fns_a,b; 2fnt_a,b; 2fxd_a,b; 2fxe_a,b; 2hb3_a,b; 2hc0_a,b; 2hs1_a,b; 2hs2_a,b; 2i0a_a,b; 2i0d_a,b; 2i4d_a,b; 2i4u_a,b; 2i4v_a,b; 2i4w_a,b; 2i4x_a,b; 2idw_a,b; 2ien_a,b; 2ieo_a,b; 2j9j_a,b; 2j9k_a,b; 2je4_a,b; 2nmw_a,b; 2nmy_a,b; 2nmz_a,b; 2nnk_a,b; 2nnp_a,b; 2nph_a,b; 2nxd_a,b; 2nxl_a,b; 2nxm_a,b; 2o4k_a,b; 2o4l_a,b; 2o4n_a,b; 2o4p_a,b; 2o4s_a,b; 2p3a_a,b; 2p3b_a,b; 2p3c_a,b; 2p3d_a,b; 2pk5_a,b; 2pk6_a,b; 2pqz_a,b; 2psu_a,b; 2psv_a,b; 2pwc_a,b; 2pwr_a,b; 2pym_a,b; 2pyn_a,b; 2q3k_a,b; 2q54_a,b; 2q55_a,b; 2q5k_a,b; 2q63_a,b; 2q64_a,b; 2qak_a,b; 2qci_a,b; 2qd6_a,b; 2qd7_a,b; 2qd8_a,b; 2qhc_a,b; 2qhy_a,b; 2qhz_a,b; 2qi0_a,b; 2qi1_a,b; 2qi3_a,b; 2qi4_a,b; 2qi5_a,b; 2qi6_a,b; 2qi7_a,b; 2qmp_a,b; 2qnn_a,b; 2qnp_a,b; 2qnq_a,b; 2r5p_a,b; 2r5q_a,b; 2rkf_a,b; 2upj_a,b; 2uxz_a,b; 2uy0_a,b; 2z4o_a,b; 2z54_a,b; 3aid_a,b; 3b7v_a,b; 3b80_a,b; 3bva_a,b; 3bvb_a,b; 3bxr_a,b; 3bxs_a,b; 3cyw_a,b; 3cyx_a,b; 3d1x_a,b; 3d1y_a,b; 3d1z_a,b; 3d20_a,b; 3dck_a,b; 3dcr_a,b; 3hvp_a, 3phv_a, 3tlh_a, 4hvp_a,b; 4phv_a,b; 7hvp_a,b; 7upj_a,b; 8hvp_a,b; 9hvp_a,b; 1daz_d; 1dw6_d; 1fej_d; 1ff0_d; 1fff_d; 1ffi_d; 1fg6_d; 1fg8_d; 1fgc_d; 1hef_e; 1heg_e

Table 9.1. HIV-1 PR PDB ids analyzed with clustering. If more than one chain was available in the PDB, they are designated after the underscores.

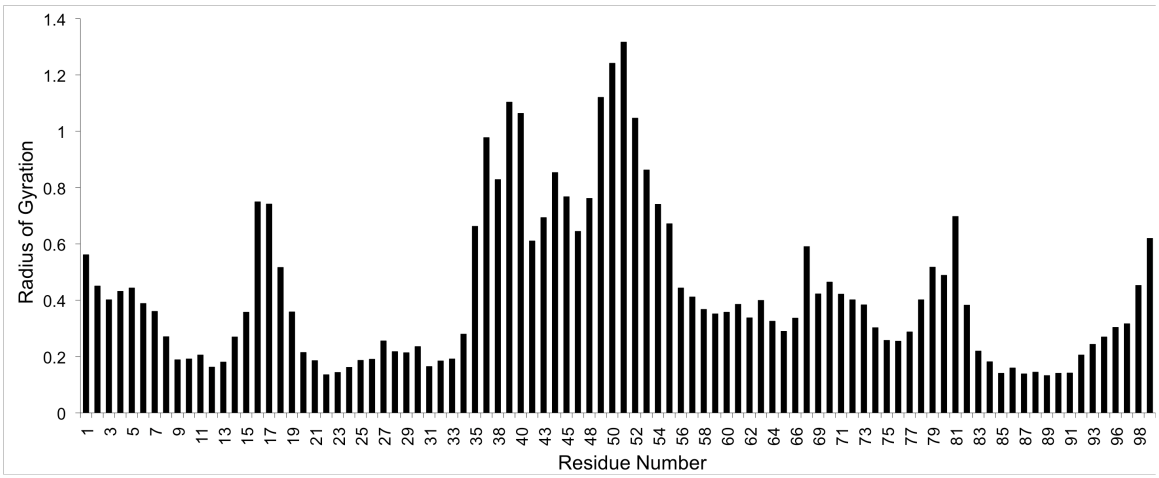

Figure 9.1. Radius of gyration for each HIV-1 PR residue (1-99) after superimposition on β1 (residues 10-14), β2 (20-24), β3 (31-34) and αA (87-93).

the need for alignment of the structures. The best clustering was chosen as that which gave the highest MSR and a minimum RMSD between clusters of greater than 1Å.

9.2 Having Many Structures Does not Denote an Abundance of Discrete Conformational Variability

Analysis of backbone and side chain configurations across the 629 monomeric PR structures results in 10 clusters as shown in Figures 9.2a,b and Table 9.2 using both single linkage (SLC) clustering and complete linkage clustering (CLC). However, the majority of the structures (586 out of 629) are found in one cluster. Other clusters include: three small clusters of seven, four, and 26 structures and 6 singletons. Separately, the flap region is separated into four clusters which correspond to a "closed" conformation, and three expanded semi-open conformations; the P1 loop is separated into 2 clusters where only one structure is found to have an expanded conformation; and the

side chain conformation of R8 is separated into 2 clusters, one of which points into the binding pocket ("in") while the other points away from the binding pocket ("out").

The majority of the structures (613) sample the "closed" flap conformation. As expected, all of the "closed" structures are either substrate-bound or inhibitor-bound. The expanded semi-open conformations of the flap are split into three classifications: expanded-1 (Exp-1), expanded-2 (Exp-2) and expanded-M (Exp-M) and are shown in

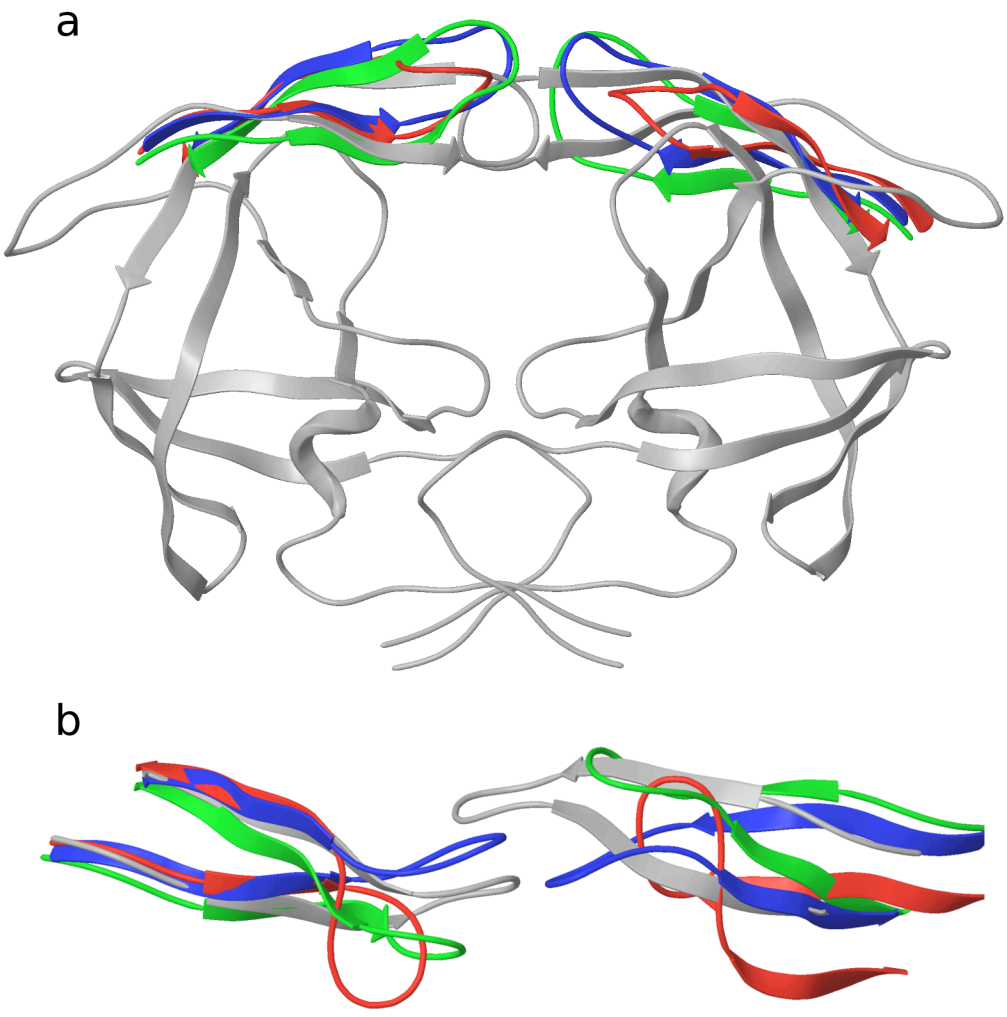

Figure 9.2a. Four conformations of the HIV-1 PR flexible flap. Gray: Large cluster representative. Blue: Apo semi-open Exp-1. Green: Apo and mutated ligand-bound semi-open Exp-2. Red: Metallocarborane-bound Exp-M. (a) Side view of HIV-1 PR. (b) Top view of PR shows a change in the "handedness" between apo and bound structures.

Figure 9.2a. Exp-1 includes only apo structures. Exp-2 includes both apo and ligand bound structures that have been described as having expanded binding sites. 1TW7 has been described as a "wide-open" structure that results from a large degree of mutation in residues 10, 36, 46, 54, 62, 63, 71, 82, 84, and 90 (Martin et al., 2005). In 2009 it was demonstrated via both experimental techniques and simulation that the "wide open" structure seen in 1TW7 is in fact a result of extensive crystal contacts. Simulations have shown that, if allowed to relax, 1TW7 takes on a semi-open form (instead of the

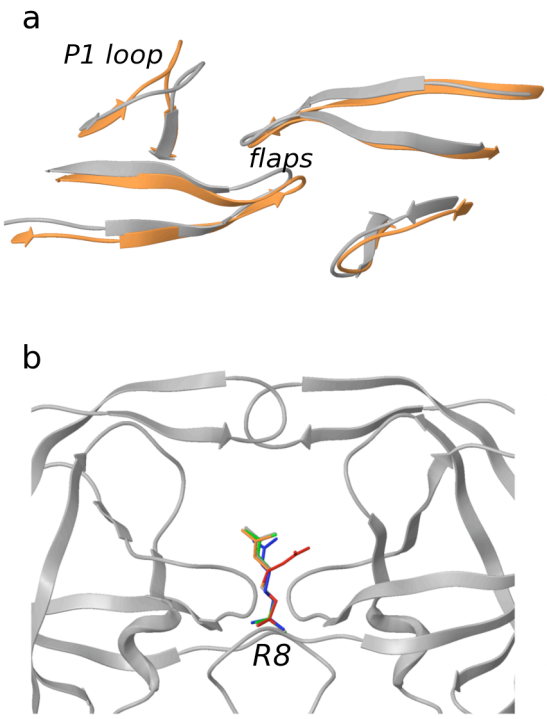

Figure 9.2b. Variations in the P1 loop and conserved side chain Arg8. (a) Gray: Large cluster representative or the "common" conformation. Orange: singleton 2FXD_a. (b) Fluctuations in the conformation of the Arg8 side chain. Red: "out" conformation. All other colors: "in" conformations.

described "wide open" form) (Lexa et al., 2009). Therefore, the mutations may still be somewhat responsible for this semi-open, expanded form of the flap. 1RV7 and 1RQ9 are also highly mutated with mutations at residues 10, 36, 46, 54, 63, 71, 82 or 84, and

90. The active sites are deemed "expanded" sites as a result of the reduced size of side chains due to the V82A and I84V mutations. They are also reported as having semi-open flap configurations possibly as a result of the novel binding modes of the inhibitor MDR-769 (Logsdon et al., 2004). Finally, the Exp-M is a novel conformation of the flaps that is a result of binding to a set of metallocarboranes. It experiences an odd crystal formation

Cluster	*Cluster Member PDB ids*	*Flap*	*P1 Loop*	*Arg8*
Large	[586]	Closed	Common	IN
Small Cluster1	1rpi_b 1tw7_a 1tw7_b 1rv7_a 1rq9_a 1rq9_b 1rv7_b	Exp-2	Common	IN
1rpi_a	1rpi_a	Exp-2	Common	OUT
Small Cluster 2	1hhp_a 3hvp_a 3phv_a 2g69_a	Exp-1	Common	IN
2hb4_a	2hb4_a	Exp-1	Common	OUT
1ztz_a	1ztz_a	Exp-M	Common	IN
1ztz_b	1ztz_b	Exp-M	Common	OUT
2pc0_a	2pc0_a	Exp-2	Common	OUT
2fxd_a	2fxd_a	Closed	Expanded	IN
Small Cluster 3	1bvg_a 1bvg_b 3dcr_b 2rkf_a 2rkf_b 2avm_b 2avq_b 2aod_b 2aog_b 1htf_a 2nnp_a 3dck_b 2idw_a 1mt7_a 3d1x_a 3dcr_a 3cyx_a 1sh9_b 1tsu_b 3bva_b 1sgu_a 2avv_b 2b60_a 2p3c_a 2p3c_b 1c70_b	Closed	Common	OUT

Table 9.2. Ten conformational basins from clustering of 629 HIV-1 PR monomers. Combinations of structural features are given for each cluster or singleton. There are four conformations of the flap: one Closed and three semi-open or expanded (Exp-1, Exp-2, Exp-M). There are two conformations for the P1 loop: the Common conformation that presides in all but one structure and the Expanded conformation which is found only in chain A of PDB id 2FXD. Finally, there are two conformations for the conserved Arg8 sidechain: "In," where the side chain faces into the binding pocket and "out," where it faces solvent. PDB ids are given for singletons and small clusters. There are 586 structures populating the large cluster.

where four metallocarboranes are situated in a new pocket defined as a tetramer instead of the typical dimer (Cígler et al., 2005). It should also be pointed out that the

conformations of the semi-open expanded flaps for the apo structures in Exp-1 display the typical opposite handedness of the bound structures, including those bound structures in Exp-2 and Exp-M, as shown in Figure 9.2a.

Clustering of R8 results in 2 "smears" of conformations that are separated by about 20°. 599 PR monomers display the "in" conformation of R8; 30 PR monomers display the "out" conformation. Apo, substrate bound and ligand bound structures may exhibit either conformation. Finally, only one structure, a M46I, V82F, I84V, and L90M mutant, is separated based on the conformation of its P1 loop: 2FXD_a. The other 2FXD monomer has a "common" configuration of the P1 loop. This is a result of the asymmetric inhibitor that is bound, which forms π-π stacking interactions with the mutated F82 (Klei et al., 2007).

Information from the clustering was used to create a conformational landscape that maps the fluctuation of the flap region. Coordinates were chosen as the large cluster representative and an apo Exp-1 conformation. Structures were binned based on these quantities and the resulting populations are shown in Figure 9.3. All four conformations of the flap are separated using these coordinates. Even though there are many structures, the majority are found in the large cluster, suggesting a rugged landscape in that region where conformations are separated by small energy barriers. This is consistent with the trends seen in previous studies (Zoete et al., 2002; Yang et al., 2008). As discussed in Chapter 6, these populations may not be directly converted to free energies as biases exist in the dataset from different ligands and substrates, crystal contacts, and mutations. However, this conformational landscape does offer a comprehensive rough estimate for possible locations of energy basins and can serve as a benchmark for computational

modeling of the HIV-1 PR "substrate envelope" or a variety of possible "substrate envelopes" for use in inhibitor design.

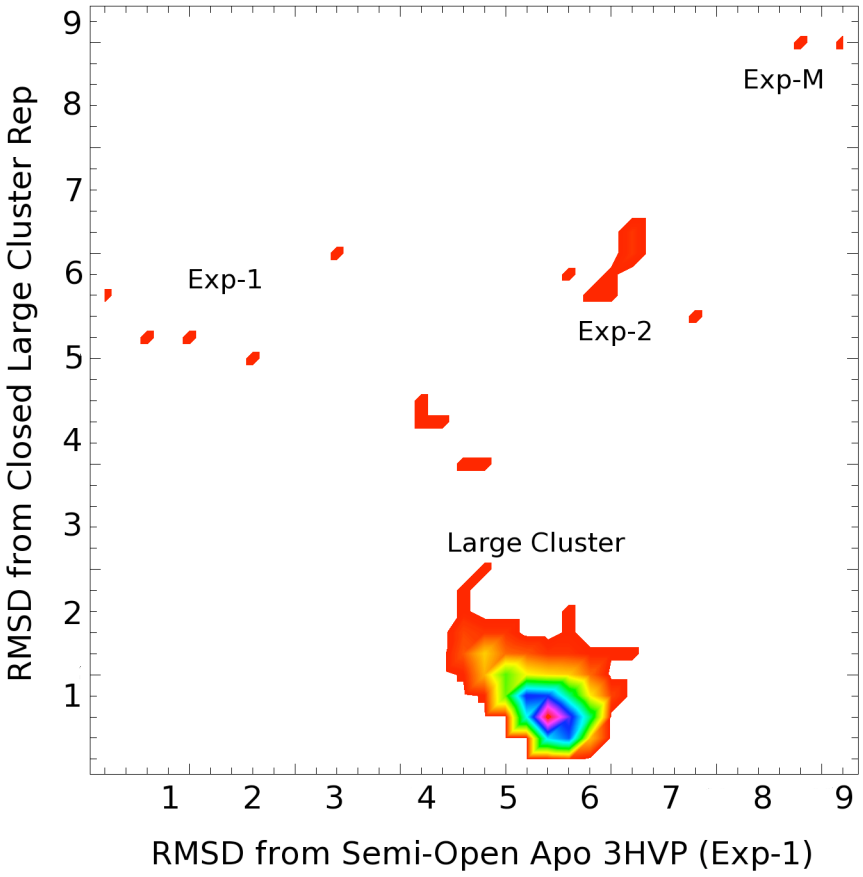

Figure 9.3. Sample conformational landscape for HIV-1 PR showing the four clusters of the flexible flap region. The closed Large Cluster, and the semi-open Exp-1, Exp-2, and Exp-M. are labeled. Colors scale with the population in that region with white being zero structures, red being a small number and fuchsia being the highest population observed. Most of the structures fall into the Large Cluster region around 4-6.5 Å rmsd from the semi-open apo structure and around 0-2 Å rmsd from the large cluster representative.

Reference

Altman, M.D.; Ali, A.; Reddy, G.S.; Nalam, M.N.; Anjum, S.G.; Cao, H.; Chellappan, S.; Kairys, V.; Fernandes, M.X.; Gilson, M.K.; Schiffer, C.A.; Rana, T.M.; Tidor, B. HIV-1 protease inhibitors from inverse design in the substrate envelope exhibit subnanomolar binding to drug-resistant variants. *J. Am. Chem. Soc*. **2008**, *130*, 6099-113.

Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235-242.

Cígler, P.; Kozísek, M.; Rezácová, P.; Brynda, J.; Otwinowski, Z.; Pokorná, J.; Plesek, J.; Grüner, B.; Dolecková-Maresová, L.; Másа, M.; Sedlácek, J.; Bodem, J.; Kräusslich, H.G.; Král, V.; Konvalinka, J. From nonpeptide toward noncarbon protease inhibitors: metallacarboranes as specific and potent inhibitors of HIV protease. *Natl. Acad. Sci. U S A*. **2005**, *102*, 15394-9.

Damm, K.L.; Ung, P.M.U.; Quintero, J.J.; Gestwicki, J.E.; Carlson, H.A. A poke in the eye: Inhibiting HIV-1 protease through its flap-recognition pocket. *Biopolymers* **2008**, *89*, 643-652.

Hornak, V.; Okur, A.; Rizzo, R.C.; Simmerling, C. HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proc. Natl. Acad. Sci.* **2006**, *103*, 915-920.

Hornak, V.; Okur, A.; Rizzo, R.C.; Simmerling, C. HIV-1 protease flaps spontaneously close to the correct structure in simulations following manual placement of an inhibitor into the open state. *J. Am. Chem. Soc.* **2006**, *128*, 2812-2813.

Hornak, V.; Simmerling, C. Targeting structural flexibility in HIV-1 protease inhibitor binding. *Drug Disc. Today* **2007**, *12*, 132-138.

Johnson, S.C. Hierarchical clustering schemes. *Psychometrika* **1967**, *32*, 241-254.

Klei, H.E.; Kish, K.; Lin, P.F.; Guo, Q.; Friborg, J.; Rose, R.E.; Zhang, Y.; Goldfarb, V.; Langley, D.R.; Wittekind, M.; Sheriff, S. X-ray crystal structures of human immunodeficiency virus type 1 protease mutants complexed with atazanavir. *J. Virol*. **2007**, *81*, 9525-35.

Kurt, N.; Scott, W.R.; Schiffer, C.A.; Haliloglu, T. Cooperative fluctuations of unliganded and substrate-bound HIV-1 protease: a structure-based analysis on a variety of conformations from crystallography and molecular dynamics simulations. *Proteins* **2003**, *51*, 409-22.

Lexa, K.W.; Damm, K.L.; Quintero, J.J.; Gestwicki, J.E.; Carlson, H.A. Clarifying allosteric control of flap conformations in the 1TW7 crystal structure of HIV-1 protease. *Proteins* **2009**, *74*, 872-80.

Logsdon, B.C.; Vickrey, J.F.; Martin, P.; Proteasa, G.; Koepke, J.I.; Terlecky, S.R.; Wawrzak, Z.; Winters, M.A.; Merigan, T.C.; Kovari, L.C. Crystal structures of a multidrug-resistant human immunodeficiency virus type 1 protease reveal an expanded active-site cavity. *J. Virol*. **2004**, *78*, 3123-32.

Martin, P.; Vickrey, J.F.; Proteasa, G.; Jimenez, Y.L.; Wawrzak, Z.; Winters, M.A.; Merigan, T.C.; Kovari, L.C. "Wide-open" 1.3 A structure of a multidrug-resistant HIV-1 protease as a drug target. *Structure*. **2005**, *13*, 1887-95.

Piana, S.; Carloni, P.; Rothlisberger, U. Drug resistance in HIV-1 protease: Flexibility-assisted mechanism of compensatory mutations. *Prot. Science* **2002**, *11*, 2393-2402.

Prabu-Jeyabalan, M.; Nalivaika, E.; Schiffer, C.A. Substrate shape determines specificity of recognition for HIV-1 protease: analysis of crystal structures of six substrate complexes. *Structure* **2002**, *10*, 369-381.

Shenkin, P.S.; McDonald, D.Q. Cluster analysis of molecular conformations. *J. Comput. Chem*. **1994**, *15*, 899-916.

Yang, L.; Song, G.; Carriquiry, A.; Jernigan, R.L. Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. *Structure* **2008**, *16*, 321-30.

Zoete, V.; Michielin, O.; Karplus, M. Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. *J. Mo.l Biol*. **2002**, *315*, 21-52.

Lexa, K.W.; Damm, K.L.; Quintero, J.J.; Gestwicki, J.E.; Carlson, H.A. Clarifying allosteric control of flap conformations in the 1TW7 crystal structure of HIV-1 protease. *Proteins* **2009**, *74*, 872-80.

Logsdon, B.C.; Vickrey, J.F.; Martin, P.; Proteasa, G.; Koepke, J.I.; Terlecky, S.R.; Wawrzak, Z.; Winters, M.A.; Merigan, T.C.; Kovari, L.C. Crystal structures of a multidrug-resistant human immunodeficiency virus type 1 protease reveal an expanded active-site cavity. *J. Virol*. **2004**, *78*, 3123-32.

Martin, P.; Vickrey, J.F.; Proteasa, G.; Jimenez, Y.L.; Wawrzak, Z.; Winters, M.A.; Merigan, T.C.; Kovari, L.C. "Wide-open" 1.3 A structure of a multidrug-resistant HIV-1 protease as a drug target. *Structure*. **2005**, *13*, 1887-95.

Piana, S.; Carloni, P.; Rothlisberger, U. Drug resistance in HIV-1 protease: Flexibility-assisted mechanism of compensatory mutations. *Prot. Science* **2002**, *11*, 2393-2402.

Prabu-Jeyabalan, M.; Nalivaika, E.; Schiffer, C.A. Substrate shape determines specificity of recognition for HIV-1 protease: analysis of crystal structures of six substrate complexes. *Structure* **2002**, *10*, 369-381.

Shenkin, P.S.; McDonald, D.Q. Cluster analysis of molecular conformations. *J. Comput. Chem*. **1994**, *15*, 899-916.

Yang, L.; Song, G.; Carriquiry, A.; Jernigan, R.L. Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. *Structure* **2008**, *16*, 321-30.

Zoete, V.; Michielin, O.; Karplus, M. Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. *J. Mo.l Biol*. **2002**, *315*, 21-52.

Chapter 10

Modeling Receptor Strain Energy in Protein-Ligand Binding

10.1 Introduction to Protein-Ligand Binding and GLIDE

An understanding of molecular recognition and its principles would result in more efficient applications in medicinal chemistry. Drug discovery, in principle, should capture the physical properties that are responsible for recognition of the drug by its target protein. As illustrated in chapters 6 and 9, the amount of available structural data is ever-growing as experimental techniques improve. This large amount of structural data along with a large amount of available inhibition data (as shown in Appendix A.1) allows computer-aided structure-based ligand design to serve as an alternative strategy to experimental high-throughput screening to find novel leads in drug development.

Computer-aided drug design methods take aim at two tasks: predicting the binding mode of the ligand in the binding pocket and estimating the binding affinity. The estimation of the binding affinity is important to correctly rank possible lead molecules. For example, in virtual screening, weak binders should be distinguishable from strong binders and non-binders. Predicting binding modes and estimating affinities are generally accomplished in two steps: docking and scoring. In the docking step, multiple protein-ligand configurations, called poses, are generated. Then, the poses are scored using a scoring function to calculate the binding affinity of the ligand in that pose for the receptor. Conformations of the ligand close to the "native" conformation should be

ranked above (have more favorable binding energies than) those farther from the "native."

Scoring functions can be grouped into three classes: force-field based, knowledge based and empirical scoring functions. Force-field based scoring functions, such as those found in CHARMm (Momany and Rone, 1992) and Dock-chemical (Ewing and Juntz, 1997), apply classical molecular mechanics energy functions where they approximate the free energy of binding as a sum of van der Waals and electrostatic interactions. Knowledge based scoring functions like DrugScore (Gohlke et al. 2000) and PMF (Muegge and Martin, 1999) represent the binding affinity as a sum of protein-ligand atom pair interactions. These potentials utilize distance-dependent interaction free energies of protein-ligand atom pairs derived from probability distributions of interatomic distances from protein-ligand complexes with known structures. Empirical scoring functions like ChemScore (Eldridge et al. 1997), Gold (Jones et al. 1995; Jones et al. 1997), AutoDock (Morris et al. 1998; Goodsell and Olson, 1990; Morris et al. 1996) and Glide (Friesner et al. 2006; Friesner et al. 2004; Halgren et al. 2004) estimate the binding free energy by summing interaction terms derived from fitting the scoring function to experimental binding constants of a training set of protein-ligand complexes. The archetypical empirical scoring function consists of five main terms that represent hydrogen bonds, ionic and lipophilic interactions, and the loss of external and configurational entropy upon binding (Böhm, 1994; Böhm, 1998).

Glide from Schrödinger, Inc. (Grid-based Ligand Docking with Energetics) has recently been shown to outperform other powerful empirical scoring functions in both correct pose identification and in virtual screening (Zhou et al. 2007; Cross et al. 2009; Li

et al. 2010; Friesner et al. 2006). The Glide XP algorithm uses a series of hierarchical filters to search positions, orientation and conformations of the ligand in the receptor's binding site (Friesner et al. 2006; Friesner et al. 2004; Halgren et al. 2004). The shape and properties of the receptor are represented on a grid by varying sets of fields that are computed prior to docking. The ligand's translation ability is limited by the box that defines the binding site. Initially, a set of ligand conformations is generated through an exhaustive torsional search which is clustered in a combinatorial fashion. In the first stage, the clusters which are characterized by a common "core" conformation and a set of rotamer group conformations are docked as single objects (Friesner et al. 2004). Rough positioning and scoring allows a reduction of the possible poses that will be considered in the next step. Step two minimizes selected poses using precomputed van der Waals and electrostatic grids for the receptor. The precomputed values were acquired with the OPLS-AA force field (Jorgenson et al. 1996; Kaminski et al. 2001). Finally, the five to ten lowest energy poses are subjected to a Monte Carlo procedure that examines nearby torsional minima to refine the peripheral groups of the ligand. The minimized poses are then rescored using the XP GlideScore function shown below in equations 10.1-3 (Friesner et al. 2006):

$$XP\ GlideScore = E_{Coul} + E_{vdW} + E_{bind} + E_{penalty} \tag{10.1}$$

where $E_{bind} = E_{hyd\_enclosure} + E_{hb\_nn\_motif} + E_{PI} + E_{hb\_pair} + E_{phobic\_pair}$ (10.2)

and $E_{penalty} = E_{desolv} + E_{ligand\_strain}$ (10.3)

The XP GlideScore is an expanded ChemScore (Eldridge et al. 1997) function with force-field components and additional terms accounting for solvation and repulsive interactions. The Glide XP scoring function applies desolvation penalties by docking

explicit waters into the highest scored docked complexes and evaluating the solvation of polar and charged ligand and protein groups by counting the number of neighboring waters and comparing these values to statistics extracted from a database of correctly docked ligands. Incremental increases in binding affinity are added to the ligand score when appropriate motifs are recognized. Additional terms that take into account hydrogen bonding, treatment of salt bridges π-cation interactions and other specialized medicinal chemistry motifs are described by Friesner et al. (2006). The XP scoring function was parametrized using a training set of 15 receptor structures and affiliated "fitting" ligands (Friesner et al. 2006).

10.2 Advantages and limitations in utilizing structural descriptors for characterization of receptor reorganization free energy in protein ligand binding

An understanding of molecular recognition and its principles leads towards more efficient applications in medicinal chemistry. Drug discovery should capture the physical properties that are responsible for recognition of the drug by its target protein. The amount of available structural data is ever-growing as experimental techniques improve. This, along with a large amount of available inhibition data, allows computer-aided structure-based ligand design to serve as an alternative strategy to experimental high-throughput screening to find novel leads in drug development. Computer-aided drug design methods take aim at two tasks: predicting the binding mode of the ligand in the binding pocket and estimating the binding affinity. The estimation of the binding affinity is important to correctly rank possible lead molecules. For example, in virtual screening, weak binders should be distinguishable from strong binders and non-binders.

The overall affinity of a ligand for a receptor can be expressed as a balance between the strength of the interactions of the ligand for a particular binding-competent conformation of the receptor and the probability of occurrence of that conformation in the absence of a ligand. Much work has been done on the former part of the problem of determining the strength of interactions between a ligand and receptor. (Friesner, et al. 2004; Halgren, et al. 2004; Friesner, et al. 2006; Ewing and Kuntz 1997; Moustakas, et al. 2006; Jones, et al. 1995; Verdonk, et al. 2008; Kramer, et al. 1999; Jain 2007; Venkatachalam, et al. 2003; Zhou, et al. 2007; Ferrara, et al. 2004) The latter part of the problem has recently come back into focus with the idea of conformational selection or binding funnels. (Boehr, et al. 2009; Bakan and Bahar 2009; Ma, et al. 2002; Ma, et al. 1999; Frauenfelder, et al. 1991; Miller and Dill1997) Previously, ligand binding was often approached via either Fischer's "lock-and-key" model (Fischer 1894) or Koshland's "induced fit" hypothesis. (Koshland 1958) In the "lock-and-key" model, the free and ligand-bound proteins have the same rigid conformation whereas in the "induced fit" model, the ligand induces a complementary conformational change in the protein. The conformational selection hypothesis approaches binding from a "folding funnel" point of view where protein folding or binding is viewed as a parallel process in which an ensemble of molecules goes downhill through an energy funnel. (Dill and Chan 1997; Lazaridis and Karplus 1997; Becker and Karplus 1997; Martinez, et al. 1998; Onuchic, et al. 1997; Ravindranathan, et al. 2005) Folding funnels are rugged in the vicinity of the native fold of the protein, suggesting energetically competitive and similar conformations that provide an enhanced means of interactions between the protein and either ligands or other proteins. The binding funnel model takes into account this rugged terrain and

argues that ligand binding can shift the populations towards the weakly populated, higher energy conformations that are more suitable for binding. (Ma, et al. 1999) Both conformational selection and induced fit appear to play roles in ligand binding. (Boehr, et al. 2009; Bakan and Bahar 2009)

Receptor reorganization can potentially be an important part of the measure of the affinity of a ligand for that particular receptor. Receptors that undergo little to no conformational change upon binding can be handled in a "lock-and-key" fashion where the receptor is held rigid as a ligand is docked. However, receptors that do undergo conformational change upon binding may require inclusion of receptor reorganization or strain free energy to properly model the binding of ligands to that protein. Many medically relevant receptors undergo conformational changes upon binding, including several of the human immunodeficiency viral enzymes as well as a variety of kinases that have been implicated in certain cancers and other diseases. The problem of receptor reorganization in protein-ligand binding represents the most difficult challenge; there is no cookbook recipe for modeling receptor reorganization in ligand binding and several methods have been attempted. These include MD and MC methods, (Armen, et al. 2009; Cheng, et al. 2008; Bowman, et al. 2007; Carlson 2002; Hart and Read 1992; Oshiro, et al. 1995) use of rotamer libraries, (Schaffer and Verkhivker 1998; Desmet, et al. 1992; Leach 1994; Trosset and Scheraga 1999) protein ensemble docking, (Armen, et al. 2009; Totrov and Abagyan 2008; Knegtel, et al. 1997; Ferrari, et al. 2004; Claussen, et al. 2001) and soft-receptor modeling. (Knegtel, et al. 1997; Ferrari, et al. 2004; Osterberg, et al. 2002) MD and MC methods can be computationally expensive and have the drawback of potentially introducing significant error and "noise" that could decrease

docking accuracy. Methods based on rotamer libraries represent the receptor as a set of experimentally observed and preferred rotameric states for side chains that surround the binding pocket. However, this technique does not include backbone flexibility. Ensemble docking methods, where the ligand is docked to an ensemble of receptors with varying structures, have been explored but some studies have shown that docking to an ensemble may give worse results than rigid docking. (Polgar and Keseru 2006; Barril and Morley 2005) Soft-receptor modeling combines information from several protein conformations to generate a single weighted average grid to which the ligand is docked. Another version of "soft" docking employs reduced van der Waals radii or deletion of side chains of residues predicted to be flexible, thus potentially eliminating close contacts. (Carlson and McCammon 2000) A study in 2006 rather successfully combined the "soft" docking technique with iterations of rigid receptor docking using reduced vdW radii and protein structure prediction techniques. (Sherman, et al. 2006) However, "soft" techniques are not able to handle large changes in conformation.

While progress is being made in generating new receptor conformations for binding in connection with docking, we will use ensembles of experimentally determined receptor conformations to test our model for receptor reorganization free energy. A linear response model for incorporation of receptor strain in combination with a modern protein-ligand binding affinity estimator is proposed. For the majority of targets, most scoring functions perform rather well. However, receptors that undergo large conformational changes present problems in both the estimation of binding affinities and in ranking of ligands. Here, we study a set of five targets in an effort to develop a protocol for adding receptor strain estimators to the scoring function included in Glide

(Grid-based Ligand Docking with Energetics) by Schrödinger, Inc. (Friesner, et al. 2004; Halgren, et al. 2004; Friesner, et al. 2006): HIV-1 reverse transcriptase (HIV RT), HIV-1 protease (HIV PR), p38 mitogen-activated protein kinase kinase (p38), Abl kinase (Abl), and phosphodiesterase 4 (PDE4). The PDB ids associated with each target are given in Table 1 and structures are shown in Figure 1. Each target undergoes a significant conformational change upon binding, has significant errors (greater than 1.5 kcal/mol) between their associated estimated scores and experimental binding energies, has low rank-order correlations (lower than 0.5), and offers a large number of crystallographically determined structures with different ligands bound. They also allow for easy definition of a receptor binding pocket and show flexibility within said binding pocket.

HIV RT, the HIV-1 viral enzyme responsible for the replication of the viral genomic material, experiences a large conformational change, depicted in Figure 1, upon binding nucleic acid as the enzyme moves to "clasp" the nucleic acid. (Jacobo-Molina, et al. 1993; Ding, et al. 1998) One type of HIV RT inhibitor, the non-nucleoside inhibitor (NNRTI), is a non-competitive, specific inhibitor that binds to a pocket called the non-nucleoside reverse transcriptase binding pocket (NNIBP), which lays approximately 10 Å from the enzyme's polymerase active site. (Kohlstaedt, et al. 1992) The NNIBP undergoes large structural rearrangements upon binding of an NNRTI where the aromatic side chains Y181 and Y188 swivel and the primer grip region (contained in the β12-β13-β14 sheet) moves to create space for the ligand (as shown in Figure 1). (Hsiou, et al. 1996; Das, et al. 2007) As NNRTIs come in many shapes and sizes, the NNIBP has been found to be quite flexible, and has been likened to "shrink wrap" that changes form to optimize interactions with different ligands (Das, et al. 2008).

The viral enzyme HIV PR, an aspartic acid protease shown in Figure 1, plays an essential role in the life cycle of HIV in generation of mature virion particles through cleavage of the viral Gag and GagPol precursor proteins. (Kohl, et al. 1988) It is composed of 99 amino acids and requires symmetric dimer formation for catalytic activity. (Wlodawer and Erickson 1993) The active site is formed along the dimer interface and the two active site residues (D25 and D25') are contributed by each monomer. (Oroszlan and Luftig 1990) Binding of substrate causes the flaps of the dimer to move by as much as 7 Å in an opening and closing motion. (Miller, et al. 1989) To date, there are no X-ray structures demonstrating the completely open conformation of the flap but NMR data suggests that there are rapid fluctuations between open and closed forms of the flap in solution. (Freedberg, et al. 2002; Hornak and Simmerling 2007) As protease recognizes the asymmetric shape of peptide substrates rather than amino acid sequence, it does not require a particular amino acid sequence for cleavage. Instead, it cleaves peptides that have similar secondary structures that fit in a defined "substrate envelope." (Prabu-Jeyabalan, et al. 2002) When protease binds a substrate, the structural symmetry of the homodimer is broken as the monomers adjust to accommodate the substrate.

P38 is a mitogen-activated protein kinase (MAPK) kinase and is a target for anti-inflammatory therapy for treatment of rheumatoid arthritis, psoriasis, multiple sclerosis and inflammatory bowel disease. (Han, et al. 1994; Saklatvala 2004) Abl kinase is a tyrosine kinase that has been linked to chronic myelogenous leukemia (CML). (Lombardo, et al. 2004; Tokarski, et al. 2006; Wong and Witte 2004) Inhibition of p38 and Abl can be accomplished with two variations of inhibitors that target two very

different conformations of the enzyme as shown in Figure 1. Type I kinase inhibitors bind at the ATP binding site in the active conformation while type II kinase inhibitors target the inactive (non phosphorylated) conformation and bind to the ATP binding cleft and an adjacent hydrophobic pocket created by the aspartate-phenylalanine-glycine (DFG) loop being in an "out" conformation. The DFG loop is the kinase's activation loop where the conserved DFG motif is found at the start of the loop. The active conformations of both p38 and Abl are often referred to as DFG-"in" whereas the inactive conformations are often referred to as DFG-"out." (Liu and Gray 2006; Munoz, et al. 2010)

Phosphodiesterase 4 (PDE4) is the major enzyme that degrades cAMP in cells and is a therapeutic target of high interest for central nervous system, inflammatory and respiratory diseases. (Houslay, et al. 2005; Spina 2008; Burgin, et al. 2010) Currently explored inhibitors bind the active site competitively with cAMP; the upstream conserved region 2 (UCR2), a signature regulatory domain, is necessary for the binding of at least one of these inhibitors, rolipram. (Burgin, et al. 2010; Bolger, et al. 1993; Jacobitz, et al. 1996) UCR2 has also been found to be partially responsible for the regulation of cAMP hydrolysis as it can adopt a "closed" conformation that blocks the active site; it is possible that the binding of inhibitors serve to stabilize this "closed" conformation. The PDE4 inhibitor binding site is the least flexible of the five targets explored in this study and is similar to that of the apo structure (as shown in Figure 1); the primary source of conformational variation is in the side chains of UCR2 as well as other side chains surrounding the binding pocket.

In an effort to refine estimated binding free energy scores with an addition of receptor strain free energy using a linear response model, we postulate that most of the errors between the scores and experimental binding energies are due primarily to the neglect of receptor strain free energy and that there exists a response between the strain energies and the geometry of the receptor pockets. A brute-force, general semi-automated procedure is introduced that uses pair-wise residue-residue contacts in a defined binding pocket as descriptors for changes in receptor conformation and, therefore, a possible measure of receptor strain. Residue-residue contact information can be easily assembled without detailed knowledge of the receptor and has the ability to account for all receptor conformational changes. Increases and decreases in residue-residue contacts may be both favorable or unfavorable as they can account for ligand-induced repacking of the receptor binding pocket through breaks in contacts between residues on opposite sides of the pocket and creation of contacts between neighboring residues. As such, they can be useful for evaluating nonbonded interaction energies. Residue-residue contact counts are used as a basis for several highly simplified models for protein motion and dynamics, such as the anisotropic network model (Atilgan, et al. 2001; Eyal, et al. 2006) and the Gaussian network model. (Haliloglu, et al. 1997; Bahar and Jernigan 1998; Yang, et al. 2009; Lin, et al. 2008) These motions are generally characterized by a high degree of collectivity and are defined by the overall architecture or topology of interresidue contacts in the native structure. (Tama and Brooks 2006; Nicolay and Sanejouand 2006) Applications of such simplified models have become increasingly popular in modeling protein-ligand intereactions, especially in cases where the receptor is flexible. (May and Zacharias 2008; Floquet, et al. 2006; Cavasotto, et al.

2005) However, moving from the study of protein motion to a quantification of a resulting energy term has proven difficult and far from straightforward. The linear model here is an attempt at pairing a simplified model of receptor motion and the concept of receptor reorganization free energy. It is constructed to fit the strain free energy, defined by the difference between the experimental binding free energy and the estimated score from Glide (GlideScore), to a combination of pair-wise receptor-receptor contacts. Analysis of the selected combinations of residue-residue contact counts may offer valuable information as to the type of conformational changes, if any, that may contribute to the receptor reorganization free energy. Ligand reorganizational energies and entropic loss are ignored here as they are partially accounted for in the estimated GlideScores.

As we are using such a large number of potential descriptors, it is possible that the fits produced are simply by chance and have no real structural meaning or implications. Several statistical tests including jack-knife "leave one out" and the Bonferroni ad hoc test are utilized to test the fits. A more sophisticated null hypothesis test is also introduced and constructed where random data was generated and fit via linear regression to the difference between the experimental binding free energy and the estimated GlideScore, here defined as the GlideScore "error," in the same manner as the real data. Since the use of fewer variables will more likely avoid over-fitting and fitting by chance alone, an alternative to using receptor residue-residue contact counts as potential descriptors is offered for sample difficult targets.

Results and Discussion

Correlation of the GlideScores and the experimental binding free energies are depicted in Figure 2 and in the Spearman rank order correlations (ρ's) listed in Table 2.

The rank order correlations of the GlideScores to experiment are very low (no correlation) for Abl, HIV RT and p38 and marginal for HIV PR and PDE4. In four out of the five cases, the majority of the complexes have GlideScores lower than the experimental binding free energy. This may suggest that a positive energy penalty due to receptor strain can be added to improve the scores. For PDE4, which has a smaller unsigned error of 1.53 kcal/mol, a marginal rank order correlation, and the majority of GlideScores greater than corresponding experimental binding free energies, it is still interesting to test the concept of receptor reorganization where there is less ligand-induced variation in the binding pocket and where the bound conformations sample a more rugged landscape about a conformation close to the inactive apo structure.

Figure 2 and Table 2 give the results of the optimal linear model. Comparison can be made to the errors and correlations of the initial GlideScores numerically and by eye from Figure 2. Comparison can also be made to a very simple regression of the GlideScores to the experimental energies where errors in the binding free energies are reduced but the rank order correlation does not improve. It should be reinforced that the linear model proposed here does not fit the GlideScore directly to the experimental energies as is done in the simple linear regression; only contact descriptors have coefficients in the linear model. For all five targets, the linear response model provided apparent improvement in reducing the error *and* increasing the rank order correlations. The largest improvement in predicting binding free energies is found in Abl where the error decreases from 3.78 to 1.10 kcal/mol while the rank order correlation increases slightly. HIV RT, p38, HIV PR and PDE4 experience reductions in binding free energy

error of close to 1 kcal/mol while also increasing the Spearman rank order correlation to significantly higher values.

Conformational contributions to receptor reorganization free energy. The contact pairs that were selected for the linear model may offer useful information about each targets' free energy landscape for binding in the spirit of free energy folding funnels proposed for proteins in general. Folding funnels are rugged in the vicinity of the native fold of the protein suggesting energetically competitive and similar conformations which provide an enhanced means of interaction between the protein and either ligands or other proteins. (Dill and Chan 1997; Becker and Karplus 1997; Onuchic, et al. 1997; Ravindranathan, et al. 2005; Tsai, et al. 1999) The landscape provides useful information about both the different means for inhibitors to bind a receptor and the strain free energy required to adopt a particular conformation for binding. Limited fluctuation of the receptor reorganization free energies may suggest that the deformations within the binding pocket are locally elastic with small free energy penalties. In contrast, large changes in reorganization free energies are suggestive of more steeply sloped free energy basins. One of the challenges of exploring a free energy landscape is the determination of useful order parameters. This study may offer some insight into which parameters may be useful to describe the receptor reorganization free enery landscape for binding to each target through examination of the receptor contact pairs selected in the linear model.

The selected Abl kinase descriptors describe receptor changes that are evident to the naked eye. One pair, E286 and F382 tracks the "in" and "out" conformations of the enzyme. All structures with zero contacts made between this pair are in the "out" conformation while all structures with at least one contact are "in" conformations. F382

is also an important conserved residue and makes up the "F" of the DFG motif. The other pair, I293 and F359 track a rotation in the χ_1 dihedral angle of F359. Estimated reorganization free energies for Abl are positive and have a large range from 0.66 kcal/mol to 6.55 kcal/mol with ~70% of the structures having reorganization free energies within the first standard deviation (stdev; 1.88 kcal/mol) from the mean (3.68 kcal/mol); all reorganization free energies fall within two standard deviations from the mean. The largest reorganization free energies are associated with the DFG "out" conformations while those with smaller energies are associated with DFG "in" conformations, suggesting a possible free energy barrier between the two conformations.

The two pairs from the linear model for HIV RT also describe large changes in the binding pocket of the enzyme. The Y188 and W229 pair scales with the distance between the conserved primer grip region and the β6-β9-β10 sheet, which accounts for an expansion of the NNRTI binding pocket via movement of the biologically important primer grip (Jacobo-Molina, et al. 1993) in response to different ligands. The Y181 and E138 (on p51) pair track conformational changes seen in the χ_1 angle of Y181 as the residue swivels between "Closed" and "Open" conformations. These descriptors mirror those from a previous study that utilized clustering to probe the conformational landscape for binding NNRTIs. (Paris, et al. 2009) The estimated reorganization free energies experience a large unevenly distributed range from -0.29 kcal/mol to 5.36 kcal/mol; 80% are within the first stdev (1.02 kcal/mol) from the mean (2.00 kcal/mol) but one structure is found over three stdevs from the mean: 1TV6. 1TV6 was found to be a singleton in the clustering study (Paris, et al. 2009) and experiences "Closed" conformations of the Y181 and Y188 side chains with a primer grip region in a more open position than most

of the other structures. As 1EP4, which has the most expanded positioning of the primer grip (Paris, et al. 2009), demonstrates a reorganization free energy close to the mean, it can be suggested that the largest influence on HIV RT's reorganization free energy may not be the change in conformation of the primer grip alone but instead the swiveling of the Y181 side chain in combination with primer grip movement.

The three pairs found for p38 take aim at side chain fluctuation. M109-D112 and I141-I147 scale with the χ_1 angles of M109 and I141, which sample conformations that are separated by 100° to180°, respectively. The third pair, R67 and R70 changes with fluctuations in R67, which is quite variable throughout the set of X-ray structures. The range of reorganization free energies is quite large, from -3.67 kcal/mol to 7.58 kcal/mol but 78% are within the first stdev (2.59 kcal/mol) from the mean that is closer to 0 (0.77 kcal/mol). For most of the complexes, the addition of reorganization free energy is very small and is found to be less than 1 kcal/mol. However, there are two structures where the contributed energy is at the upper limit (close to 7.5 kcal/mol); one of these structures is a DFG "out conformation with a rather large ligand and the other represents a unique conformation that has its DFG loop in an "in" position but has a large DFG "out" type ligand bound. It is interesting to note that, although p38 and Abl are from the same family and display similar conformational changes of their DFG loops upon binding, the large backbone rearrangements are only found to be contributors to the reorganization free energy of Abl and not p38.

Finally, both PDE4 and HIV PR's contact pairs are associated with more subtle reorganizations in small fluctuations of side chains throughout the structure. Linear regression for PDE4 selects residue pairs: M347/273-L393/319, M347/273-I450/376,

I410/336-M411/337 (where the first residue number is for PDE4B and the second is for PDE4D). HIV PR fitting results in T12_a-I66_a, D25_a-D25_b, L5_a-P9_a, R8_a-P9_a pairs (where the a's and b's designate the chain/monomer). PDE4 samples energies from -3.54 kcal/mol to 1.22 kcal/mol with an average of -1.34 kcal/mol while HIV PR samples energies from -1.92 kcal/mol to 3.56 kcal/mol with an average of 1.22 kcal/mol. Both targets sample small near-evenly distributed ranges in reorganization free energy that point to rugged energetic basins around their averages.

The analysis of the selected residue pairs used for fitting offers valuable information as to the type of conformational changes, if any, may contribute to the receptor reorganization free energy. Both HIV RT and Abl kinase pairs track large changes in back bone and side chains. P38 pairs select for large side chain fluctuations, and HIV PR and PDE4 pairs select for smaller, subtle side chain variations. For PDE4, this was expected as the backbone of the binding pocket has been described as rather rigid. The HIV PR structures included in this study also do not include large backbone changes; all flaps are in the closed, bound conformation.

Statistical examination of the linear response model. As we are using a large number of potential descriptors and selecting them based on experimental data, it is possible that the fits produced are simply by chance and have no real structural or energetic meaning or implication. This problem of statistical significance is oftentimes probed by jack-knife or "leave one out" tests and calculation of p-values for the linear fits. Jack-knife tests successively leave one data point out during fitting and determination of the linear model. Errors of each successive model are then averaged and compared with errors in the initial all-inclusive model. The p-value for the linear fit

is defined as the probability, under the assumption of no effect (the null hypothesis), of obtaining a result equal to or more extreme than what was actually observed. Low p-values (less than 0.05 or 0.01) are argued to provide evidence against the null hypothesis (Fisher 1950) but this practice has found itself highly contested. (Neyman and Pearson 1933; Sterne and Davey Smith 2001) The Bonferroni ad hoc test calls for a reduction in the acceptable p-value by $1/n$ where n is the initial number of variables. In this case, the number of initial values for each receptor is represented by the number of contact pairs within 5 Å of any associated ligand and ranges from 27 to 153 depending on the target and size of the ligands. Both jack-knife and Bonferroni tests are shown in Table 2 and, for the most part, point toward good fits that are likely not produced by chance alone. Jack-knife tests result in expected small increases of the binding free energy error except in the case of p38. In three of the five cases (Abl, HIV PR and PDE4), the p-values for the fits are more than three orders of magnitude less than the Bonferroni threshold. The remaining two (HIV RT and p38) are slightly less than the threshold.

A further, more sophisticated null hypothesis test was constructed where data was randomly generated and fit via linear regression to the difference between the GlideScores and experimental binding free energies in the same manner as the real data (see the Experimental Section). This null model works as a test to see if structural knowledge is necessary to model the reorganization free energy of binding or if random numbers could be selected to produce similar results. Outcomes and significance tests from the null model and comparison with the real model are shown in Table 2. The null model was able to give comparable fits to those generated with real data. Jack-knife tests on the null model produce similar results to those for the real model; however, the real

model performs better (has a lower jack-knife error) than the null model in three out of five cases (Abl, PDE4, and HIV PR). Again, the jack-knife test for p38 produces unfavorable results. Only two of the five targets pass the Bonferroni ad hoc test with the null model in comparison with the real contact descriptors where all five targets produce p-values lower than the Bonferroni threshold.

The apparent success of the null model is most likely a consequence of the method chosen where the first pass pares down the large number of contact counts to a smaller set based on Spearman rank order correlations. Selection of the contact counts for the linear regression was also done using different ρ cutoff values (shown in Table 2; ρ cutoffs are described in the Experimental Section below) for random versus real and, in all five cases, the cutoffs were lower for the random data than for the real. This is especially prevalent in Abl where the cutoff used for real data is 0.7 while that used for the random is 0.3. The different cutoff values were thought to be necessary to provide similar numbers of contact counts for each target as input to the linear model. In other words, the "random" data used for the null model regression was inadvertently chosen to mimic the real data and thus was able to produce comparable fits. Use of the Bonferroni test and comparison of the jack-knife results between the null model and the real model allow for a clearer picture of which targets benefit from the reorganization free energy model posited here (PDE4, Abl, HIV PR), which results are not truly structurally significant (p38), and those that are borderline (HIV RT).

An alternative route: structurally significant intrareceptor distance descriptors. In the two cases where the model utilizing residue-residue contact counts produce structurally insignificant (p38) or borderline results (HIV RT), an alternative

route using intrareceptor distances was attempted. This method also has the ability to overcome the possibility of over-fitting or fitting by chance alone as a smaller set of variables, in the form of hand-selected distances that describe some of the structural variation of the receptors, is utilized. HIV RT is described by Y181cz-E138cd, Y188ca-W229ca, V108ca-186ca, and L228cg-F227cg (residue type, residue number, PDB atom type; E138 is located on the p51 monomer while the others are located on p66). These distances were chosen based on a previous study. (Paris, et al. 2009) The distance descriptors for p38 were selected by observations of several superimposed structures and focus on variations of the DFG loop along with possible fluctuations of the DFG "in" and "out" binding pockets. As they must include fluctuations in both binding regions, there are more potential descriptors for p38 than for HIV RT, which only has one binding pocket; they include: F169_M109ca, F169-I84ca, F169-A51ca, M109ca-V30ca, D112ca-V30ca, T106ca-I84ca.

Unlike the previous fits, no descriptors were thrown out prior to determination of the linear model; all descriptors were used as input for the linear model. The "best" fit combination of distance descriptors was determined as was done previously in the contact count model discussed in the Experimental Section below. One of the targets – p38 – showed no improvement over the initial GlideScores while HIV RT showed considerable improvement with a binding free energy error reduction from 2.13 kcal/mol to 1.21 kcal/mol and an increase in the Spearman rank order correlation from 0.25 to 0.66. A null model was again constructed where data was randomly generated and then fit to the difference in GlideScore and experimental binding free energy in the same fashion as the real data. In this case, since there are fewer descriptors and no need for a ρ cutoff filter

that may inadvertently choose "random" numbers that mimic the real data, the null model produced results that showed *no* improvement over the initial GlideScores shown in Table 2.

As the model assumes that the difference between experimentally observed and computationally predicted binding affinities is due primarily to the reorganization free energy and that there is a linear response between the reorganization free energy and the geometry of the pockets, it may not be beneficial in cases in which these assumptions are not valid. One such case may be p38 where the large backbone fluctuations found between the DFG "in" and DFG "out" conformations are not correlated with the difference between observed and predicted energies. Other targets such as HIV RT, where the energy errors are correlated with large structural fluctuations, are largely improved with incorporation of reorganization free energy.

The use of selected structurally significant intrareceptor distance descriptors does show some promise in the use of the proposed fitting protocol for modeling receptor strain in ligand binding for more difficult targets such as HIV RT. However, whereas use of contact counts to estimate receptor reorganization is statistically limited by the large volume of data used as input to the protocol, selected distance descriptors are limited by the human intuition needed to make the descriptor choices and the method is not easily transferrable to other receptor pockets.

Conclusion

Accounting for receptor reorganization free energy is one of the most difficult problems in modeling protein-ligand binding. Development and application of a semi-automated protocol that makes use of a set of descriptors based on a large set of available

structural data to model receptor reorganization in combination with commercially available docking programs shows promise in reduction of binding free energy errors and in increasing rank-order correlation. A major potential pitfall of such a model is the possibility of producing fits purely by chance that may not have any structural significance. The sophisticated null hypothesis test presented here in combination with the Bonferroni ad hoc test offers a possible solution for examination of the significance of models based on culling from large data sets. As use of a linear response model for receptor reorganization is founded on the assumption that the difference between the experimentally observed and computationally predicted binding affinities is due primarily to receptor reorganization, such a model may not be valid in some cases where conformational variability does not correlate well with errors in predicted binding affinities. However, this model does have the potential to allow for coarse-grained investigation of the conformational landscapes for binding inhibitors and may offer insight into which structural features may influence receptor reorganization free energy as well as information about the characteristics of the receptor strain free energy landscape.

Experimental Section

Data preparation. Each structure listed in Table 1 was prepared using Schrödinger, Inc.'s Protein Preparation Wizard which enumerates bond orders, determines optimal protonation states for both protein and ligand, and adds missing hydrogen atoms. It also allows for optimization of the protein's hydrogen bond network by means of a systematic, cluster-based approach. After preparation, each structure underwent a restrained minimization that allows hydrogen atoms to be freely minimized

while allowing some heavy atom movement to relax possible strained bonds and angles as well as possible clashes.

Docking and calculation of binding affinities. Glide 5.0 XP, the current version of Glide (Friesner, et al. 2004; Halgren, et al. 2004; Friesner, et al. 2006) available in Schrödinger's 2009 suite, was used to dock the ligands into their respective receptors to provide an optimally docked complex for analysis. Glide has recently been compared to other powerful empirical scoring functions and has been shown to perform well in both correct pose identification and in virtual screening. (Friesner, et al. 2006; Zhou, et al. 2007; Li, et al. 2010; Cross, et al. 2009) The Glide XP algorithm uses a series of hierarchical filters to search positions, orientation and conformations of the ligand in the receptor's binding site. (Friesner, et al. 2004; Halgren, et al. 2004; Friesner, et al. 2006) The XP GlideScore function is an expanded ChemScore (Friesner, et al. 2006; Eldridge, et al. 1997) function with force-field components and additional terms accounting for solvation and repulsive interactions. The Glide XP scoring function applies desolvation penalties by docking explicit waters into the highest scored docked complexes and evaluating the solvation of polar and charged ligand and protein groups by counting the number of neighboring waters and comparing these vales to statistics extracted from a database of correctly docked ligands. Incremental increases in binding affinity are added to the ligand score when appropriate motifs are recognized. Additional terms that take into account hydrogen bonding, treatment of salt bridges π-cation interactions and other specialized medicinal chemistry motifs are described by Friesner et al. (2006). The XP scoring function was parametrized using a training set of 15 receptor structures and affiliated "fitting" ligands. (Friesner, et al. 2006)

Estimation of binding modes and energies was performed using three different protocols: score in place (SIP), where the prepared complex is scored without docking; refined docking, where the ligand is refined based on its initial condition (using the XP version of Glide, this means that the ligand is regrown in place); and flexible docking, where Glide generates different conformations of the ligand by varying acyclic torsional angles and sampling low-energy ring conformations. The complex with root mean square deviation (rmsd) less than 2 Å from the starting structure and with the lowest energy was chosen as the representative structure whose energy served as input for the linear regression model. (In the majority of cases, the refined docking gave the lowest energy structure.) Correlation of the lowest GlideScores and the experimental binding energies are depicted in Figure 2 and in the Spearman rank order correlations (ρ's) listed in Table 2.

Contact count determination. The binding pockets for each target were defined as all residues within 5 Å from any included ligand bound to that target. Receptor-receptor contact maps were generated for each receptor utilizing the defined binding pockets for each target. The contact count between a pair of residues was calculated as the sum of contacts between each atom *i* in residue 1 and each atom *j* in residue 2. A contact was defined as any two atoms *i* and *j* that had a ratio

$$C = \frac{D_{ij}}{R_i + R_j} < 1.3 \tag{10.4}$$

where D_{ij} is the distance between the centers of atoms *i* and *j* and R_i and R_j are the Lennard Jones radii of atoms *i* and *j*, respectively. In this way, we hoped to be able to catch any possible conformational variability between structures without having to decide on a set of coordinates for superimposition of the receptors. It should be noted that the

issue of HIV PR's possible asymmetry due to the binding of asymmetric inhibitors was treated prior to determining contact counts. Monomers were relabeled based on the number of interactions they create with their bound ligands. Monomers within the dimer that have the greater number of receptor-ligand contacts are labeled as chain A whereas those with less are labeled B. Symmetric dimers that displayed no differences in monomeric interaction with bound symmetric ligands retained their original chain name designation. This method was utilized in an attempt to include interactions between monomers.

Estimation of receptor reorganization free energy. Receptor strain was estimated using linear combinations of residue-residue contact counts as conformational descriptors. Each target started with between 27 (HIV RT) and 153 (HIV PR) residue pair contact counts. This initial selection was first pared down to a more manageable number (4 to 12) by calculating the Spearman rank order correlation ρ between each pairwise contact count and the "error" of the GlideScore (ddG = experimental dG – predicted GlideScore dG), ordering the pairwise contact counts based on ρ and selecting a threshold to be used for the first round of contact count selection. The chosen threshold for each target is listed in Table 2 as ρ CO and was chosen as the optimal threshold that provided as close to 10 descriptors as possible. The ρ CO's and number of filtered descriptors differ between targets primarily as a result of the differing sizes of the receptor pockets (larger pockets have more descriptors). A linear model is then constructed as in equation 2 with the smaller filtered set of 4 to 12 contact counts C_i, and optimal weights w_i and intercept b. The best fit to the difference of the experimental binding free energy and the GlideScore is then determined.

$$ddG = b + \sum_i w_i C_i \tag{10.5}$$

The model is sequentially tested to acquire the optimal contact count combination using the Mallow's C_p value as a test. The Mallow's C_p is often used to remove collinearity or variable redundancy and to reduce the chance of overfitting in a regression model. (Hocking 1976) All combinations of the contact count regressors are tried and the combination with the lowest C_p value is chosen as the "best" model. The C_p is defined in equation 3 as:

$$C_p = \frac{SSE_p}{S^2} - N + 2P \tag{10.6}$$

where SSE_p is the error sum of squares for the model with P regressors, N is the sample size (number of receptors for the said target), and S^2 is the residual mean square after regression on the complete set of regressors. (Mallows 1973) Table 2 shows the final number of regressors chosen for each target.

After determining the "best fit" linear model, new predicted binding energies are calculated and the new mean absolute error is calculated along with the new Spearman rank order correlation ρ between the new predicted scores and the experimental binding energies. Table 2 gives the results of the optimal linear regressions with associated errors and ρ's. Comparison can be made to the errors and correlations of the initial GlideScores.

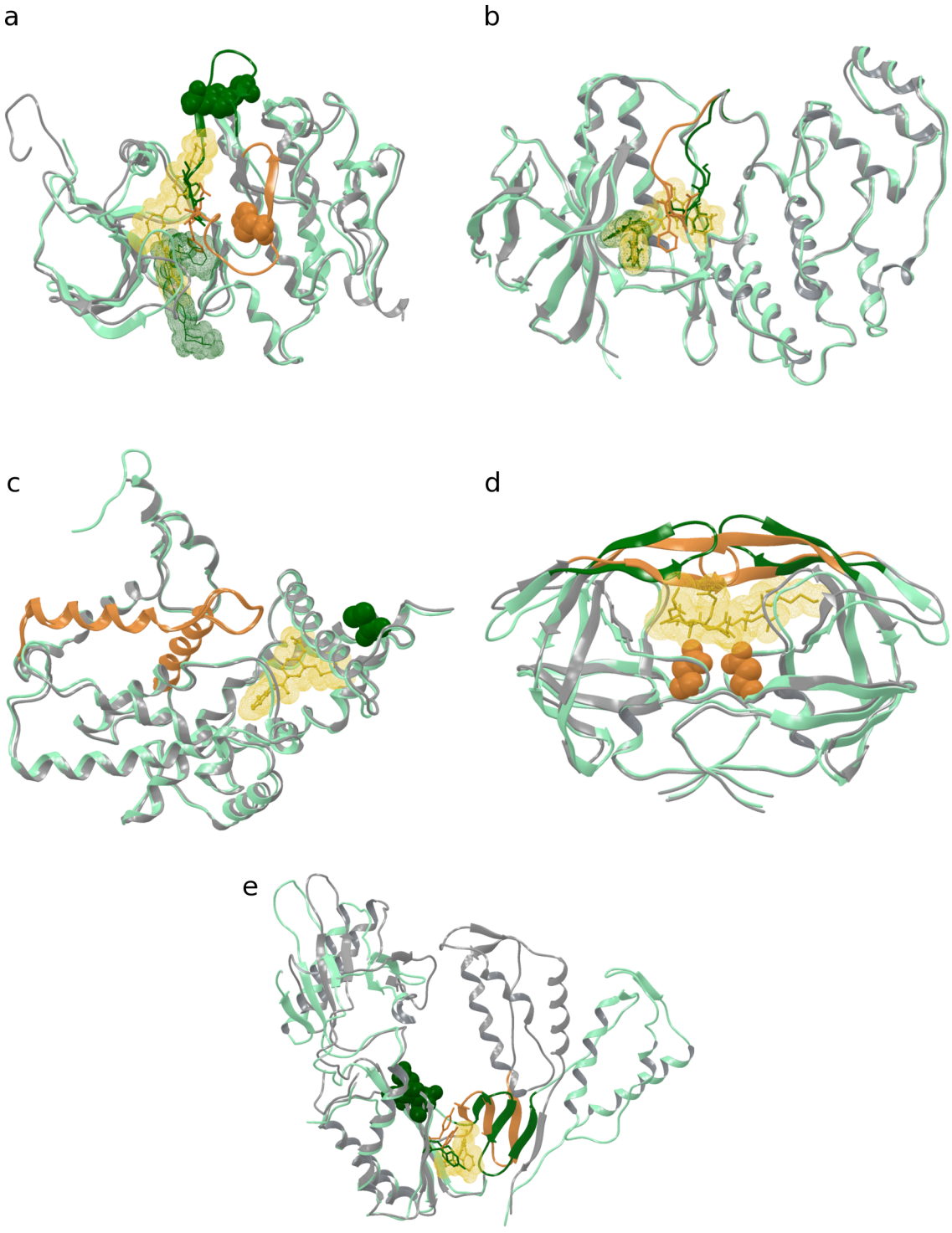

Figure 10.1. Targets for reorganizational free energy analysis. (a) Abl. Green: DFG "in," PDB id 2V7A; Orange/Gray: DFG ""out," PDB id 1OPJ. Ligands are shown in ball and stick with mesh surfaces. Light green ligand: DFG "in" ligand; Yellow ligand: DFG

"out." Active site residue Y393 is shown in CPK space fill. DFG motif is shown as tubes. The activation loop is colored dark green (DFG "in") or orange (DFG "out"). (b) p38. Green: DFG "in," PDB id 1W84; Gray/Orange: DFG "out," PDB id 1W83. Ligands are shown as ball-and-stick with mesh surfaces; yellow: DFG "out" ligand, green: DFG "in" ligand which sits in the ATP binding site. DFG motif is shown as tubes. The activation loop is colored dark green (DFG "in") or orange (DFG "out"). (c) PDE4. Green: Apo, PDB id 1F0J; Gray: Inhibitor-bound, PDB id 1W84. Orange: portion of UCR2 that acts as a gate to the active site (also inhibitor binding site). One active site residue is shown in dark green CPK space fill: M439. A sample ligand is shown in ball-and-stick with yellow mesh. (d) HIV PR. Green: Apo, PDB id 2PC0; Gray/Orange: Substrate-bound, PDB id 1F7A. Substrate is shown with yellow mesh surface and active site residues D25, D25' are shown in orange space-fill. (e) HIV RT. Gray/Orange: Apo, PDB id 1DLO; Green: NNRTI-bound, PDB id 1VRT. NNRTI nevirapine is shown with yellow mesh surface and active site residues D185, D186 and D110 are shown in dark green space-fill. Primer grip region and Y181 and Y188 are colored dark green (bound) and orange (apo) to show changes in the NNRTI binding pocket due to ligand binding. The thumb region connected to the primer grip also exhibits a large conformational change from closed (apo; gray) to open (NNRTI bound; green).

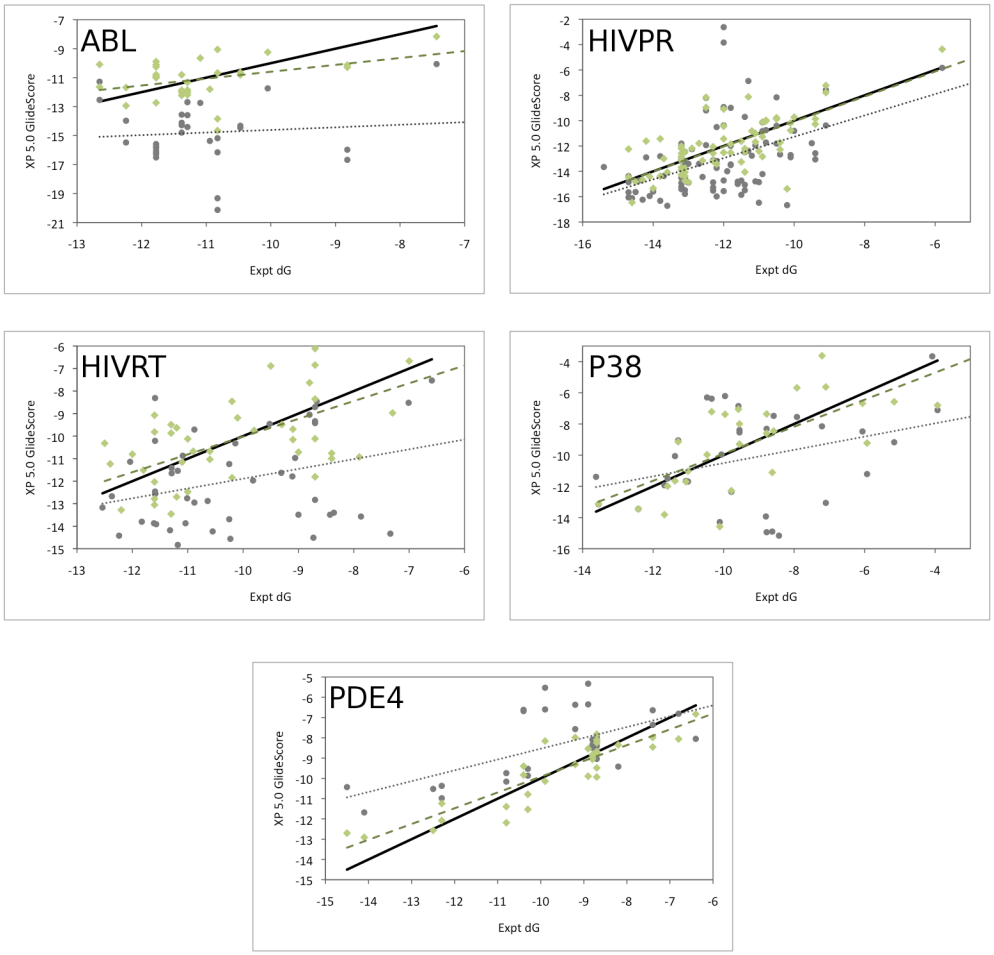

Figure 10.2. Comparison and correlation of experimental binding free energies and GlideScores or binding free energies estimated from the linear model. The black line is the 1-to-1 line and points represent each target complex. Gray points represent initial GlideScores before fitting with the linear model and the gray dotted line represents the linear fit of the GlideScores to the experimental binding free energies. Percent of complexes with GlideScore < experimental binding free energy for each target are: 94% (Abl), 72% (HIV PR), 84% (HIV RT), 50% (p38; this does not coincide with DFG "in" vs. DFG "out" conformations although DFG "out" conformations have a slightly higher percentage of 62%), and 14% (PDE4). Green points represent scores after incorporation of receptor reorganization free energies from the linear model and the green dashed line depicts the linear fit to the experimental values. Spearman rank order correlation coefficients for each target before and after fitting with the linear model are listed in Table 2.

Table 1. Targets and PDB ids included in receptor reorganization study

Target	*PDB ids*
Abl	1IEP_A, 1IEP_B, 1M52_A, 1M52_B, 1OPJ_A, 1OPJ_B, 1OPK_A, 1OPL_A, 1OPL_B, 2E2B_A, 2E2B_B, 2F4J, 2FO0, 2GQG_A, 2GQG_B, 2HIW_A, 2HIW_B, 2HYY_A, 2HYY_B, 2HYY_C, 2HYY_D, 2HZ0_A, 2HZ0_B, 2HZ4, 2HZI_A, 2HZI_B, 2HZN, 2QOH_A, 2QOH_B, 2V7A_A, 2V7A_B, 2Z60_A
HIV PR	1A30, 1A9M, 1AAQ, 1AJV, 1AJX, 1B6J, 1B6L, 1B6M, 1BDQ, 1BV7, 1C6Y, 1C70, 1CPI, 1D4I, 1D4J, 1DIF, 1DMP, 1EBW, 1EBY, 1EBZ, 1EC0, 1EC2, 1EC3, 1G2K, 1G35, 1HIH, 1HPO, 1HPV, 1HSG, 1HVC, 1HVI, 1HVJ, 1HVK, 1HVL, 1HVR, 1HWR, 1HXW, 1IIQ, 1IZH_A, 1IZH_B, 1IZI, 1K6C, 1K6T, 1KZK, 1LZQ_A, 1LZQ_B, 1M0B_A, 1M0B_B, 1MER, 1MES, 1MET, 1MEU, 1MRW, 1MRX, 1MSM, 1MSN, 1N49, 1NPA, 1NPV, 1NPW, 1ODW, 1ODY, 1OHR, 1PRO, 1QBS, 1QBT, 1RL8_A, 1RL8_B, 1S65, 1S6S, 1SBG, 1SDU, 1SDV, 1SGU, 1T3R, 1T7J, 1T7K, 1TCX, 1W5V, 1W5W, 1W5X, 1W5Y, 1WBK, 1WBM, 1XL5, 1Z1H, 1Z1R, 2AVO, 2AVS_A, 2AVS_B, 2BB9, 2BPY, 2BPZ, 2BQV, 3AID
HIV RT	1C0T, 1C0U, 1C1B, 1C1C, 1DTQ, 1EET, 1EP4, 1FK9, 1FKO, 1FKP, 1HNI, 1HQU, 1IKX, 1JKH, 1JLF, 1JLG, 1KLM, 1LWE, 1REV, 1RT1, 1RT2, 1RT4, 1RT5, 1RT6, 1RTH, 1S1T, 1S1V, 1S1W, 1S1X, 1S6P, 1S9E, 1S9G, 1SV5, 1TKT, 1TKX, 1TKZ, 1TL1, 1TL3, 1TV6, 1VRT, 1VRU, 2B5J, 2B6A, 2BAN, 2BE2, 3HVT
p38	1A9U, 1BL6, 1BL7, 1BMK, 1DI9, 1IAN, 1KV1, 1KV2, 1M7Q, 1OUK, 1OUY, 1OVE, 1W7H, 1W82, 1W83, 1W84, 1WBN, 1WBO, 1WBS, 1WBT, 1WBV, 1WBW, 1YQJ, 1YW2, 1ZYJ, 1ZZ2, 1ZZL, 2BAJ, 2BAK, 2BAL, 2BAQ, 2GFS
PDE4	1Q9M_B, 1Q9M_C, 1Q9M_D, 1RO6_AA, 1RO6_AB, 1RO6_BA, 1RO6_BB, 1XLX_A, 1XLX_B, 1XLZ, 1XM4, 1XMU_A, 1XMU_B, 1XMY_A, 1XMY_B, 1XN0_A, 1XN0_B, 1XOM_A, 1XOM_B, 1XON, 1XOQ, 1XOR_A, 1XOR_B, 1XOS, 1XOT_A, 1XOT_B, 1Y2C_A, 1Y2C_B, 1Y2E_A, 1Y2E_B, 1Y2H_A, 1Y2H_B, 1Y2K_A, 1Y2K_B, 1ZKN

_A , _B, _C, _E refer to different binding modes for the same ligand/receptor.

Table 2. Analysis of GlideScores before and after incorporation of receptor reorganization free energy from the linear model using pairwise contact counts and comparison to the null model.

	Orig GS		*Orig GS LR*		*Real Contact Descriptor Model*						*Null Model*						*p-value Bonf*
Target	*Error*	ρ	*Error*	ρ	*Error*	ρ	*JK*	ρ *CO*	*P*	*p-value*	*Error*	ρ	*JK*	ρ *CO*	*P*	*p-value*	
Abl	3.78	<0.00	0.78	<0.00	1.10	0.22	1.43	0.70	2	5.2e-7	1.33	0.24	2.00	0.30	3	1.1e-3	5.2e-4
HIV PR	1.97	0.49	1.13	0.49	1.22	0.65	1.61	0.40	4	1.3e-8	1.40	0.61	1.96	0.25	4	3.0e-4	3.3e-4
HIV RT	2.13	0.25	1.21	0.25	1.39	0.51	1.96	0.25	2	1.8e-3	1.17	0.55	1.71	0.20	3	2.8e-4	1.9e-3
p38	2.41	0.30	1.74	0.30	1.64	0.75	3.63	0.45	3	3.0e-4	1.79	0.42	3.44	0.40	3	2.0e-3	4.3e-4
PDE4	1.53	0.43	1.22	0.43	0.69	0.82	0.89	0.50	3	5.3e-7	0.93	0.60	1.89	0.30	3	3.7e-3	9.1e-4

For each target, the original GlideScore unsigned errors and Spearman rank order correlations ρ (Orig GS Error, ρ) are higher than both after simple linear regression of the original GlideScore fit to the experimental binding energy (Orig GS LR Error; the ρ does not change as a result of this linear regression) and after utilizing the optimal contact count linear model (Real Contact Descriptor Model). Results with the Null Model, where random data was generated and fit as in the real model, are reported for comparison with the Real Contact Model. Jack knife (JK, leave one out) results, ρ cutoffs (ρ CO) for selection of initial pairwise contact counts (that allowed a paring down of the large contact count data to 4 to 12 descriptors) and final number of descriptors P used in the model is also shown. Bonferroni p-value thresholds (p-value Bonf) are given for comparison.

References

Armen, R. S.; Chen, J.; Brooks, C. L. An Evaluation of Explicit Receptor Flexibility in Molecular Docking Using Molecular Dynamics and Torsion Angle Molecular Dynamics. *J. Chem. Theory Comput.* **2009**, *5,* 2909-2923.

Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* **2001**, *80,* 505-515.

Bahar, I.; Jernigan, R. L. Vibrational dynamics of transfer RNAs: comparison of the free and synthetase-bound forms. *J. Mol. Biol.* **1998**, *281,* 871-884.

Bakan, A.; Bahar, I. The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106,* 14349-14354.

Barril, X.; Morley, S. D. Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. *J. Med. Chem.* **2005**, *48,* 4432-4443.

Becker, O. M.; Karplus, M. The topology of multidimensional potential energy surfaces: theory and application to peptide structure and kinetics. *J. Chem. Phys.* **1997**, *106,* 1495-1517.

Boehr, D. D.; Nussinov, R.; Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **2009**, *5,* 789-796.

Böhm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput.- Aided Mol. Des*. **1994**, *8*, 243-256.

Böhm, H. J. Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3d database search programs. *J. Comput.-Aided Mol. Des*. **1998**, *12*, 309-323.

Bolger, G.; Michaeli, T.; Martins, T.; St John, T.; Steiner, B.; Rodgers, L.; Riggs, M.; Wigler, M.; Ferguson, K. A family of human phosphodiesterases homologous to the dunce learning and memory gene product of Drosophila melanogaster are potential targets for antidepressant drugs. *Mol. Cell. Biol.* **1993**, *13,* 6558-6571.

Bowman, A. L.; Nikolovska-Coleska, Z.; Zhong, H.; Wang, S.; Carlson, H. A. Small molecule inhibitors of the MDM2-p53 interaction discovered by ensemble-based receptor models. *J. Am. Chem. Soc.* **2007**, *129,* 12809-12814.

Burgin, A. B.; Magnusson, O. T.; Singh, J.; Witte, P.; Staker, B. L.; Bjornsson, J. M.; Thorsteinsdottir, M.; Hrafnsdottir, S.; Hagen, T.; Kiselyov, A. S.; Stewart, L. J.; Gurney, M. E. Design of phosphodiesterase 4D (PDE4D) allosteric modulators for enhancing cognition with improved safety. *Nat. Biotechnol.* **2010**, *28,* 63-70.

Carlson, H. A. Protein flexibility and drug design: how to hit a moving target. *Curr. Opin. Chem. Biol.* **2002**, *6,* 447-452.

Carlson, H. A.; McCammon, J. A. Accommodating protein flexibility in computational drug design. *Mol. Pharmacol.* **2000**, *57,* 213-218.

Cavasotto, C. N.; Kovacs, J. A.; Abagyan, R. A. Representing receptor flexibility in ligand docking through relevant normal modes. *J. Am. Chem. Soc.* **2005**, *127,* 9632-9640.

Cheng, L. S.; Amaro, R. E.; Xu, D.; Li, W. W.; Arzberger, P. W.; McCammon, J. A. Ensemble-based virtual screening reveals potential novel antiviral compounds for avian influenza neuraminidase. *J. Med. Chem.* **2008**, *51,* 3878-3894.

Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **2001**, *308,* 377-395.

Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **2009**, *49,* 1455-1474.

Das, K.; Bauman, J. D.; Clark, A. D.,Jr; Frenkel, Y. V.; Lewi, P. J.; Shatkin, A. J.; Hughes, S. H.; Arnold, E. High-resolution structures of HIV-1 reverse transcriptase/TMC278 complexes: strategic flexibility explains potency against resistance mutations. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105,* 1466-1471.

Das, K.; Sarafianos, S. G.; Clark, A. D.,Jr; Boyer, P. L.; Hughes, S. H.; Arnold, E. Crystal structures of clinically relevant Lys103Asn/Tyr181Cys double mutant HIV-1 reverse transcriptase in complexes with ATP and non-nucleoside inhibitor HBY 097. *J. Mol. Biol.* **2007**, *365,* 77-89.

Desmet, J.; Maeyer, M. D.; Hazes, B.; Lasters, I. The dead end elimination theorem and its use in protein side-chain positioning. *Nature* **1992**, *356,* 539-542.
Dill, K. A.; Chan, H. S. From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **1997**, *4,* 10-19.

Ding, J.; Das, K.; Hsiou, Y.; Sarafianos, S. G.; Clark, A. D.,Jr; Jacobo-Molina, A.; Tantillo, C.; Hughes, S. H.; Arnold, E. Structure and functional implications of the polymerase active site region in a complex of HIV-1 RT with a double-stranded DNA template-primer and an antibody Fab fragment at 2.8 A resolution. *J. Mol. Biol.* **1998**, *284,* 1095-1111.

Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des.* **1997**, *11,* 425-445.

Ewing, T. J. A.; Kuntz, I. D. Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem* **1997**, *18,* 1175-1189.
Eyal, E.; Yang, L. W.; Bahar, I. Anisotropic network model: systematic evaluation and a new web interface. *Bioinformatics* **2006**, *22,* 2619-2627.

Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L.,3rd Assessing scoring functions for protein-ligand interactions. *J. Med. Chem.* **2004**, *47,* 3032-3047.

Ferrari, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K. Soft docking and multiple receptor conformations in virtual screening. *J. Med. Chem.* **2004**, *47,* 5076-5084.
Fischer, E. Einfluss der configuration auf die wirkung der enzyme. *Ber. Dtsch. Chem.* **1894**, *27,* 2984-2993.

Fisher, R. A. *Statistical methods for research workers;* Oliver and Boyd: London, 1950; .
Floquet, N.; Marechal, J. D.; Badet-Denisot, M. A.; Robert, C. H.; Dauchez, M.; Perahia, D. Normal mode analysis as a prerequisite for drug design: application to matrix metalloproteinases inhibitors. *FEBS Lett.* **2006**, *580,* 5130-5136.

Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. The energy landscapes and motions of proteins. *Science* **1991**, *254,* 1598-1603.

Freedberg, D. I.; Ishima, R.; Jacob, J.; Wang, Y. X.; Kustanovich, I.; Louis, J. M.; Torchia, D. A. Rapid structural fluctuations of the free HIV protease flaps in solution: relationship to crystal structures and comparison with predictions of dynamics calculations. *Protein Sci.* **2002**, *11,* 221-232.

Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47,* 1739-1749.

Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **2006**, *49,* 6177-6196.

Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol*. **2000**, *295*, 337-356.

Goodsell, D. S.; Olson, A. J. Automated docking of substrates to proteins by simulated annealing. *Proteins* **1990**, *8*, 195-202.

Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47,* 1750-1759.

Haliloglu, T.; Bahar, I.; Erman, B. Gaussian dynamics of folded proteins. *Phys. Rev. Lett.* **1997**, *79,* 3090-3093.

Han, J.; Lee, J. D.; Bibbs, L.; Ulevitch, R. J. A MAP kinase targeted by endotoxin and hyperosmolarity in mammalian cells. *Science* **1994**, *265,* 808-811.
Hart, T. N.; Read, R. J. A multiple-start Monte Carlo docking method. *Proteins* **1992**, *13,* 206-222.

Hocking, R. R. The analysis and selection of variables in linear regression. *Biometrics* **1976**, *32,* 1-50.

Hornak, V.; Simmerling, C. Targeting structural flexibility in HIV-1 protease inhibitor binding. *Drug Discov. Today* **2007**, *12,* 132-138.

Houslay, M. D.; Schafer, P.; Zhang, K. Y. Keynote review: phosphodiesterase-4 as a therapeutic target. *Drug Discov. Today* **2005**, *10,* 1503-1519.

Hsiou, Y.; Ding, J.; Das, K.; Clark, A. D.,Jr; Hughes, S. H.; Arnold, E. Structure of unliganded HIV-1 reverse transcriptase at 2.7 A resolution: implications of conformational changes for polymerization and inhibition mechanisms. *Structure* **1996**, *4,* 853-860.

Jacobitz, S.; McLaughlin, M. M.; Livi, G. P.; Burman, M.; Torphy, T. J. Mapping the functional domains of human recombinant phosphodiesterase 4A: structural requirements for catalytic activity and rolipram binding. *Mol. Pharmacol.* **1996**, *50,* 891-899.

Jacobo-Molina, A.; Ding, J.; Nanni, R. G.; Clark, A. D.,Jr; Lu, X.; Tantillo, C.; Williams, R. L.; Kamer, G.; Ferris, A. L.; Clark, P. Crystal structure of human immunodeficiency virus type 1 reverse transcriptase complexed with double-stranded DNA at 3.0 A resolution shows bent DNA. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90,* 6320-6324.

Jain, A. N. Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J. Comput. Aided Mol. Des.* **2007**, *21,* 281-306.

Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245,* 43-53.

Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol*. **1997**, *267*, 727-748.

Jorgensen, W.L.; Maxwell, D.S.; Tirado-Rives, J. Development and testing of the OPLS

all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225-11236.

Kaminski, G.A.; Friesner, R.A.; Tirado-Rives, J.; Jorgensen, W.L. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **2001**, *105*, 6474-6487.

Knegtel, R. M.; Kuntz, I. D.; Oshiro, C. M. Molecular docking to ensembles of protein structures. *J. Mol. Biol.* **1997**, *266,* 424-440.

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

Kohlstaedt, L. A.; Wang, J.; Friedman, J. M.; Rice, P. A.; Steitz, T. A. Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science* **1992**, *256,* 1783-1790.

Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **1958**, *44,* 98-104.

Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins* **1999**, *37,* 228-241.

Lazaridis, T.; Karplus, M. "New view" of protein folding reconciled with the old through multiple unfolding simulations. *Science* **1997**, *278,* 1928-1931.

Leach, A. R. Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.* **1994**, *235,* 345-356.

Li, X.; Li, Y.; Cheng, T.; Liu, Z.; Wang, R. Evaluation of the performance of four molecular docking programs on a diverse set of protein-ligand complexes. *J. Comput. Chem.* **2010**, *31,* 2109-2125.

Lin, C.; Huang, S.; Lai, Y.; Yen, S.; Shih, C.; Lu, C.; Huang, C.; Hwang, J. Deriving protein dynamical properties from weighted protein contact number. *Proteins* **2008**, *72,* 929-935.

Liu, Y.; Gray, N. S. Rational design of inhibitors that bind to inactive kinase conformations. *Nat. Chem. Biol.* **2006**, *2,* 358-364.

Lombardo, L. J.; Lee, F. Y.; Chen, P.; Norris, D.; Barrish, J. C.; Behnia, K.; Castaneda, S.; Cornelius, L. A.; Das, J.; Doweyko, A. M.; Fairchild, C.; Hunt, J. T.; Inigo, I.; Johnston, K.; Kamath, A.; Kan, D.; Klei, H.; Marathe, P.; Pang, S.; Peterson, R.; Pitt, S.; Schieven, G. L.; Schmidt, R. J.; Tokarski, J.; Wen, M. L.; Wityak, J.; Borzilleri, R. M. Discovery of N-(2-chloro-6-methyl- phenyl)-2-(6-(4-(2-hydroxyethyl)- piperazin-1-yl)-2-

methylpyrimidin-4- ylamino)thiazole-5-carboxamide (BMS-354825), a dual Src/Abl kinase inhibitor with potent antitumor activity in preclinical assays. *J. Med. Chem.* **2004**, *47,* 6658-6661.

Ma, B.; Kumar, S.; Tsai, C. J.; Nussinov, R. Folding funnels and binding mechanisms. *Protein Eng.* **1999**, *12,* 713-720.

Ma, B.; Shatsky, M.; Wolfson, H. J.; Nussinov, R. Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. *Protein Sci.* **2002**, *11,* 184-197.

Mallows, C. L. Some comments on Cp. *Technometrics* **1973**, *15,* 661-675.

Martinez, J. C.; Pisabarro, M. T.; Serrano, L. Obligatory steps in protein folding and the conformational diversity of the transition state. *Nat. Struct. Biol.* **1998**, *5,* 721-729.

May, A.; Zacharias, M. Protein-ligand docking accounting for receptor side chain and global flexibility in normal modes: evaluation on kinase inhibitor cross docking. *J. Med. Chem.* **2008**, *51,* 3499-3506.

Miller, D. W.; Dill, K. A. Ligand binding to proteins: the binding landscape model. *Protein Sci.* **1997**, *6,* 2166-2179.

Miller, M.; Schneider, J.; Sathyanarayana, B. K.; Toth, M. V.; Marshall, G. R.; Clawson, L.; Selk, L.; Kent, S. B.; Wlodawer, A. Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 A resolution. *Science* **1989**, *246,* 1149-1152.

Momany, F. A.; Rone, R. Validation of the general-purpose QUANTA. 3.2/CHARMm force-field. *J. Comput. Chem*. **1992**, *13*, 888-900.

Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. Distributed automated docking of exible ligands to proteins: Parallel applications of AutoDock 2.4. *J. Comput.-Aided Mol. Des*. **1996**, *10*, 293-304.

Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem*. **1998**, *19*, 1639-1662.

Moustakas, D. T.; Lang, P. T.; Pegg, S.; Pettersen, E.; Kuntz, I. D.; Brooijmans, N.; Rizzo, R. C. Development and validation of a modular, extensible docking program: DOCK 5. *J. Comput. Aided Mol. Des.* **2006**, *20,* 601-619.

Muegge, I.; Martin, Y. C. A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *J. Med. Chem*. **1999**, *42*, 791-804.

Munoz, L.; Selig, R.; Yeung, Y. T.; Peifer, C.; Hauser, D.; Laufer, S. Fluorescence polarization binding assay to develop inhibitors of inactive p38alpha mitogen-activated protein kinase. *Anal. Biochem.* **2010**, *401,* 125-133.

Neyman, J.; Pearson, E. On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. A* **1933**, *231,* 289-337.

Nicolay, S.; Sanejouand, Y. H. Functional modes of proteins are among the most robust. *Phys. Rev. Lett.* **2006**, *96,* 078104.

Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* **1997**, *48,* 545-600.

Oroszlan, S.; Luftig, R. B. Retroviral proteinases. *Curr. Top. Microbiol. Immunol.* **1990**, *157,* 153-185.

Oshiro, C. M.; Kuntz, I. D.; Dixon, J. S. Flexible ligand docking using a genetic algorithm. *J. Comput. Aided Mol. Des.* **1995**, *9,* 113-130.

Osterberg, F.; Morris, G. M.; Sanner, M. F.; Olson, A. J.; Goodsell, D. S. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins* **2002**, *46,* 34-40.

Paris, K. A.; Haq, O.; Felts, A. K.; Das, K.; Arnold, E.; Levy, R. M. Conformational landscape of the human immunodeficiency virus type 1 reverse transcriptase non-nucleoside inhibitor binding pocket: lessons for inhibitor design from a cluster analysis of many crystal structures. *J. Med. Chem.* **2009**, *52,* 6413-6420.

Polgar, T.; Keseru, G. M. Ensemble docking into flexible active sites. Critical evaluation of FlexE against JNK-3 and beta-secretase. *J. Chem. Inf. Model.* **2006**, *46,* 1795-1805.

Prabu-Jeyabalan, M.; Nalivaika, E.; Schiffer, C. A. Substrate shape determines specificity of recognition for HIV-1 protease: analysis of crystal structures of six substrate complexes. *Structure* **2002**, *10,* 369-381.

Ravindranathan, K. P.; Gallicchio, E.; Levy, R. M. Conformational equilibria and free energy profiles for the allosteric transition of the ribose-binding protein. *J. Mol. Biol.* **2005**, *353,* 196-210.

Saklatvala, J. The p38 MAP kinase pathway as a therapeutic target in inflammatory disease. *Curr. Opin. Pharmacol.* **2004**, *4,* 372-377.

Schaffer, L.; Verkhivker, G. M. Predicting structural effects in HIV-1 protease mutant complexes with flexible ligand docking and protein side-chain optimization. *Proteins* **1998**, *33,* 295-310.

Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* **2006**, *49,* 534-553.

Spina, D. PDE4 inhibitors: current status. *Br. J. Pharmacol.* **2008**, *155,* 308-315.

Sterne, J. A.; Davey Smith, G. Sifting the evidence-what's wrong with significance tests? *BMJ* **2001**, *322,* 226-231.

Tama, F.; Brooks, C. L. Symmetry, form, and shape: guiding principles for robustness in macromolecular machines. *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35,* 115-133.

Tokarski, J. S.; Newitt, J. A.; Chang, C. Y.; Cheng, J. D.; Wittekind, M.; Kiefer, S. E.; Kish, K.; Lee, F. Y.; Borzillerri, R.; Lombardo, L. J.; Xie, D.; Zhang, Y.; Klei, H. E. The structure of Dasatinib (BMS-354825) bound to activated ABL kinase domain elucidates its inhibitory activity against imatinib-resistant ABL mutants. *Cancer Res.* **2006**, *66,* 5790-5797.

Totrov, M.; Abagyan, R. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr. Opin. Struct. Biol.* **2008**, *18,* 178-184.

Trosset, J. Y.; Scheraga, H. A. Prodock: software package for protein modeling and docking *J. Comput. Chem.* **1999**, *20,* 412-427.

Tsai, C. J.; Kumar, S.; Ma, B.; Nussinov, R. Folding funnels, binding funnels, and protein function. *Protein Sci.* **1999**, *8,* 1181-1190.

Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graph. Model.* **2003**, *21,* 289-307.

Verdonk, M. L.; Mortenson, P. N.; Hall, R. J.; Hartshorn, M. J.; Murray, C. W. Protein-ligand docking against non-native protein conformers. *J. Chem. Inf. Model.* **2008**, *48,* 2214-2225.

Wlodawer, A.; Erickson, J. W. Structure-based inhibitors of HIV-1 protease. *Annu. Rev. Biochem.* **1993**, *62,* 543-585.

Wong, S.; Witte, O. N. The BCR-ABL story: bench to bedside and back. *Annu. Rev. Immunol.* **2004**, *22,* 247-306.

Yang, L.; Song, G.; Jernigan, R. L. Protein elastic network models and the ranges of cooperativity. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106,* 12347-12352.

Zhou, Z.; Felts, A. K.; Friesner, R. A.; Levy, R. M. Comparative performance of several flexible docking programs and scoring functions: enrichment studies for a diverse set of pharmaceutically relevant targets. *J. Chem. Inf. Model.* **2007**, *47,* 1599-1608.

Chapter 11

Conclusions, Implications and Future Directions

We have applied various methods to describe the conformational fluctuation of flexible receptors in ligand binding. The ideas behind our studies are centered in the conformational selection or landscape model where the ligand binds to a conformation of the receptor that, in absence of the ligand, is not highly populated. This notion leads to the idea of receptor reorganization or strain energy which can be defined as the free energy required to access the conformational states to which ligands bind. Several "hot" targets are studied, including HIV-1 reverse transcriptase (RT), HIV-1 protease (PR), Abelson kinase (Abl), Phosphodiesterase 4 (PDE4), and p38 MAPK kinase using the large amount of structural data available today in the Protein Data Bank (PDB).

Detailed descriptions of conformational variability through clustering of both HIV-1 RT and PR offer rough looks at the conformation landscapes for binding and illuminate regions of the landscape that are not highly sampled by inhibitors. These regions offer conformations of the binding pockets that can be further explored by drug discovery processes, including optimization of the few ligands that are currently found to bind to these sparsely populated configurations. The clustering studies also can double as benchmarks for computational simulations to explore flexibility, motions and functions of these enzymes.

Replica exchange simulations on HIV-1 RT show the possibility of developing complete conformational landscapes that could be used to calculate receptor strain. However, a shift from temperature REMD to Hamiltonian REMD or a combination of other sampling methods may greatly decrease the amount of time required to acquire said landscape.

Finally, use of a set of descriptors, such as receptor-receptor contact counts or "hand-picked" internal distances that are based on the large set of available structural data, to model receptor reorganization in combination with commercially available docking programs show some promise. This experiment also highlights potential pitfalls of such an exercise where large sets of parameters are used to fit observables. It is also possible that the method may only work with some targets as it assumes that the error between the experimental binding affinity and the predicted binding affinity determined by a docking program is mostly due to receptor reorganization. This may be the case for some targets such as HIV-1 RT, which showed the best results with structurally significant intrareceptor distance descriptors for reorganization, but may be incorrect for cases such as p38 MAPK kinase where conformational variability does not appear to be well correlated with the errors in the binding affinities.

Appendix

A.1. Analysis of the Binding DB for Ligand-Binding Data

The BindingDB (www.bindingdb.org; Liu,T., et al., 2007) is a free web-accessible database of experimentally determined protein-ligand binding affinities, with focus on proteins that are drug targets or candidate drug targets and for which structural data are present in the Protein Data Bank (PDB; Berman et al., 2000). As of October 2009, the BindingDB contains ~58,800 experimentally determined binding affinities for protein-ligand complexes, ~31,300 small molecule ligands and ~619 protein targets, 168 of which have at least one structural match in the PDB (where both the protein sequence and the ligand associated with the binding affinity matches that of the PDB structure). The data are extracted from the scientific literature and can be queried by chemical structure, chemical similarity, protein sequence, ligand name, protein name, affinity range, molecular weight and PDBID (which searches by protein sequence and ligand similarity). Links are provided from the data in the BindingDB to structural data in the PDB, to the literature in PubMed, and to UniProt.

2115 PDBs were found to match BindingDB data based on 85% sequence similarity and 0.9 ligand similarity; 1184 were found to match based on 100% sequence similarity and exact ligand match. As many targets may contain mutations that do no effect binding, the larger set of 2115 PDB IDs was searched for structures which had an exact ligand match and at least 99% sequence similarity with any mutations present being those which do not effect binding. This led to 1658 instances that span 168 targets in which there is structural data paired with affinity data in the Binding DB.

	Schrödinger	*Binding DB*	*Common*	*Schrödinger Only*	*Binding DB Only*
ABL	36	13	13	19	0
ALR2	36	49	34	2	15
CDK2	80	90	58	22	32
ER-A	14	22	9	5	13
ER-B	13	7	2	11	5
fVIIa	18	8	6	12	2
fXa	31	37	20	11	17
HIV PR	91	107	56	35	51
HIV RT	44	36	22	22	14
HSP90	19	16	8	11	8
JNK1	8	8	8	0	0
JNK3	9	5	4	5	1
OppA	27	0	0	27	0
p38	31	30	20	11	10
PDE4B	16	18	16	0	2
PDE4D	16	22	15	1	7
PKA	38	38	25	13	13
PPAR-g	19	12	6	13	6
PTP1B	23	38	12	11	26
Thrombin	49	30	20	29	10
Total	618	586	354	260	232

Table A.1. Binding affinity data points in the Binding DB versus the Schrödinger dataset.

The Schrödinger data set includes a total of 618 data points that span 17 targets (see Table A.1) where there is both structural and affinity data. Of these 17 targets, 16 are found in the BindingDB (OppA is not included in the BindindDB). For the 16 targets, the Binding DB offers 586 data points. Table A.1 gives the counts for the number of data points available for each target within the Schrödinger set, the Binding DB set and combinations of the two sets. The information available from each of the two sets is comparable.

The BindingDB offers 232 additional structures which may be added to the Schrödinger set or may be utilized as a test set for docking calculations using algorithms trained on the Schrödinger set. It also offers additional targets that have both structural and binding affinity data. Of the 168 available targets, for which there is both binding

and structural data, in the BindingDB, five stand out with a large number of data points across different ligands and mutant forms: androgen receptor (39 data points), β-secretase beta-site APP-cleaving enzyme-1 (BACE-1: 39 data points), dihydrofolate reductase (DHFR: 32 data points), thymidylate synthase (33 data points), and trypsin (88 data points).

The BindingDB binding affinities are given in the form of IC_{50} values, inhibitor constant K_i values, and isothermal titration calorimetry (ITC) measurements. The Schrödinger set includes solely IC_{50} value data. The binding free energies or dGs are then calculated from the IC_{50} values using $dG= RT \ln(IC_{50})$. As this transformation is not perfect, the dGs from the Schrödinger IC50s has been compared to the dGs from the BindingDB's K_i values ($dG= RT\ln(K_i)$) and ITC measurement. The comparison is done for the 158 data points which lie in the "Common" region of the Schrödinger + BindingDB set that have K_i or ITC values available in the BindindDB. The average of the unsigned difference between the two is 0.40 kcal/mol, with the max difference of 5.47 kcal/mol and min difference of 0 kcal/mol. The distribution over the 158 data points is shown in Figure A.1. The distributions by target are shown in Figure A.2. The use of IC_{50} values to estimate the binding free energy appears sound as the differences between these values and values calculated using other methods are, on average, very small. Looking closer at each target, no target stands out as an instance in which the estimation of dG from IC_{50} values "fails" or strays greatly from the other experimentally determined dGs.

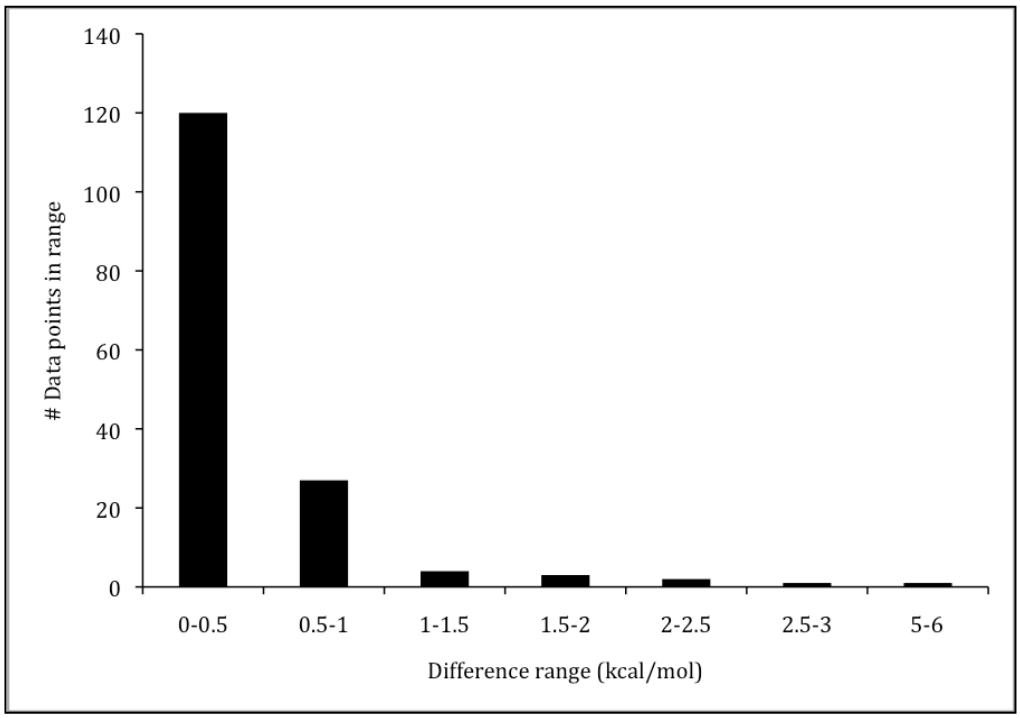

Figure A.1. Distribution of $|dG_{(Ki \text{ or } ITC)} - dG_{(Schrödinger)}|$

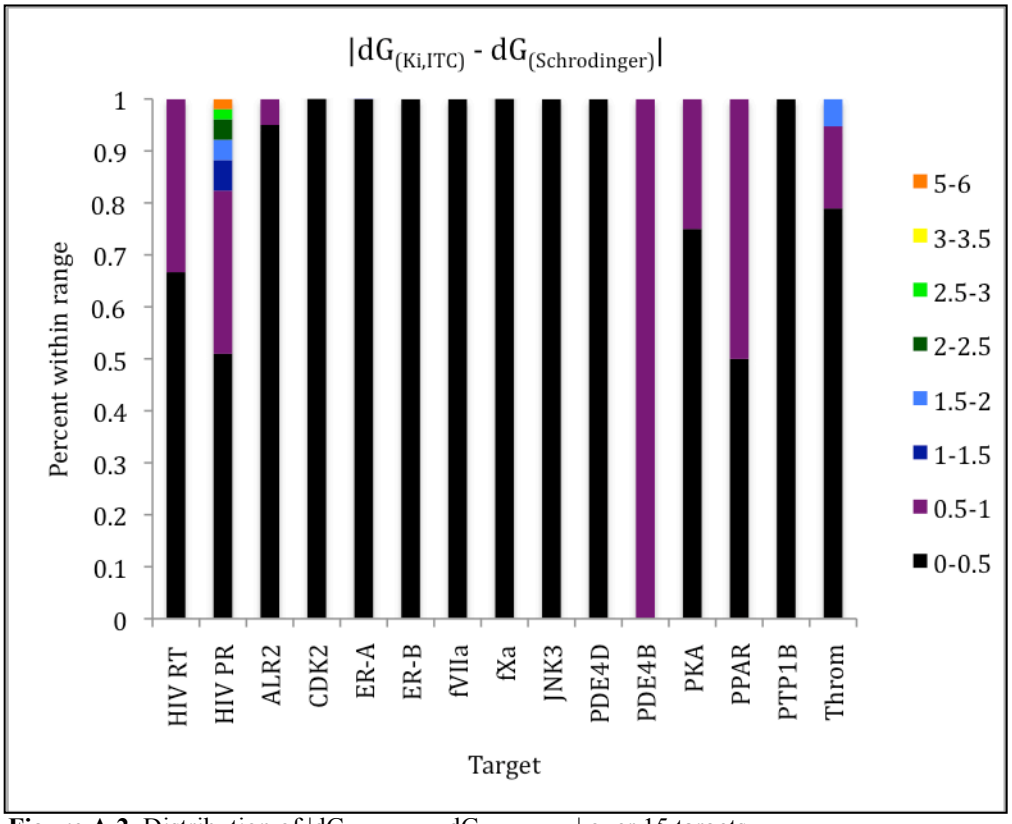

Figure A.2. Distribution of $|dG_{(Ki \text{ or } ITC)} - dG_{(Schrödinger)}|$ over 15 targets

References

Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235-242.

Liu, T.; Lin, Y.; Wen, X.; Jorissen, R.N.; Gilson, M.K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res*. **2007**, *35*, D198-201.

Curriculum Vita

Kristina A. Paris

2002 B.S. in Chemistry, New Jersey Institute of Technology, Newark, New Jersey
2011 Ph.D. in Chemistry, Rutgers University, Piscataway, New Jersey

Publications

Gilbert, K.M.; Skawinski, W.J.; Misra, M.; Paris, K.A.; Naik, N.H.; Buono, R.A.; Deutsch, H.M.; Venanzi, C.A. Conformational analysis of methylphenidate: comparison of molecular orbital and molecular mechanics methods. *J. Comput. Aided. Mol. Des*. **2004**, *18*, 719-738.

Felts, A.K.; Gallicchio, E.; Chekmarev, D.; Paris, K.A.; Friesner, R.A.; Levy, R.M. Prediction of protein loop conformations using the agbnp implicit solvent model and torsion angle sampling. *J. Chem. Theory Comput.* **2008**, *4*, 855-868.

Gallicchio, E.; Paris, K.; Levy, R.M. The AGBNP2 Implicit Solvation Model. *J. Chem. Theory Comput.* **2009**, *5*, 2544–2564.

Paris, K.A.; Haq, O.; Felts, A.K.; Das, K.; Arnold, E.; Levy, R.M. Conformational landscape of the human immunodeficiency virus type 1 reverse transcriptase non-nucleoside inhibitor binding pocket: lessons for inhibitor design from a cluster analysis of many crystal structures. *J. Med. Chem*. **2009**, *52*, 6413-6420.

Paris, K.A.; Felts, A.K. Advantages and limitations in utilizing structural descriptors for characterization of receptor reorganization free energy in protein-ligand binding. *Submitted.*