Title
New computational methods for ligand design

Author
Pitera, Jed W.

Publication Date
1999

Peer reviewed|Thesis/dissertation

New Computational Methods for Ligand Design

by

Jed W. Pitera

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA SAN FRANCISCO

Date University Librarian

Degree Conferred: .....

Acknowledgements

This thesis would not have been possible without the support of many people – my family, my friends, and most importantly, my research advisor, Dr. Peter Kollman.

Some Chapters of this thesis are a reprint of previously published material:

Chapter 2 was originally published in the Journal of Molecular Graphics, 1998.

Chapter 3 was originally published in the Journal of the American Chemical Society, 1998.

Chapter 5 was originally published in the Journal of Medicinal Chemistry, 1999.

Chapters 7 and 9 were submitted for publication in December 1998.

Abstract

New Computational Methods for Ligand Design

by Jed W. Pitera

The specific association of two molecules – a ligand and its receptor – is central to many problems in biochemistry and biology. The affinity of one molecule for another is specified by their free energy of association. This free energy is the result of competing contributions from a number of complex entropic and enthalpic effects. A class of computer calculations -- free energy calculations -- permit the calculation of relative free energies of association for different molecular species. However, the expense of these techniques has limited their practical application. This thesis describes the development and application of a novel free energy method, Chemical Monte Carlo/Molecular Dynamics (CMC/MD). CMC/MD allows one to compare many species in a single calculation, permitting the comparison of many ligands binding to a single receptor. We have applied this technique to compare families of guests binding to an organic host; families of ligands binding to a protein receptor; and families of amino acid side chains in the hydrophobic core of an enzyme. In each case, CMC/MD yields free energies in good agreement with prior results and rapidly ranks the species in question. By expanding free energy calculations to permit the simultaneous comparison of many species, their role has shifted from inquiry ("Why is A better than B?") to optimization ("Which of these A-Z is best, and why?").

Table of Contents

"As a young man. . . Callisto had learned a mnemonic device for remembering the laws of Thermodynamics: you can't win, things are going to get worse before they get better, who says they're going to get better."

Downstairs, Meatball Mulligan's lease-breaking party was moving into its 40th hour.

--Thomas Pynchon, "Entropy" (1960)

Chapter 1: Introduction

My graduate studies began in the fall of 1994, just as chemistry and biology were beginning to embrace the "combinatorial revolution" – synthesis and screening of many compounds or proteins at once. For years, biochemists and chemists had concentrated on the isolation and characterization of well-defined single molecular species. To the protein biochemist, this corresponded to the purification and characterization of a single wild-type or mutant protein in order to understand its amino acid sequence, chemical properties and biological function. For the pharmaceutical or medicinal chemist, the species of interest was a prospective ligand of defined chemical composition and stereochemistry, assayed to determine its affinity for a specific receptor. Traditionally, there was also a significant cost or effort associated with the production of each molecule of interest.

Both of these approaches are very difficult to apply to chemical optimization problems, where one wants to discover the best molecule for a given function. For the aforementioned biochemist, this might correspond to the most active enzyme for a particular reaction; for the medicinal chemist, it might be the highest-affinity ligand for a target receptor. Single-molecule approaches are defeated by both of these problems due to their scale. There are an enormous number of possible protein sequences (there are 20^{100} proteins one hundred amino acids in length). Likewise, there are an uncountably vast number of pharmaceutical-like organic compounds[1]. Finding the best sequence or compound thus requires comparison of an enormous number of molecules.

Combinatorial chemistry and biochemistry techniques allow the facile synthesis and

screening of thousands of compounds, and were developed in part to solve problems in chemical optimization – finding ligands for receptors[2], optimizing sequences for protein-protein interactions[3,4], or design of ideal catalysts and new materials for industrial processes[5].

The second chapter of this thesis details one of my initial projects in the Kollman Group. I became interested in the structure of water around solutes, both small organic molecules and proteins. To determine this structure, I developed tools to collect the probability distribution of water molecules around a solvent during a molecular dynamics trajectory. By counting the number of times a water molecule was in a particular location, one could build a picture of the "hydration shell." It was Professor Kuntz who first suggested that one might extract free energies from these probability distributions. This started me thinking about how to extract free energies from probabilities.

From basic statistical mechanics, the free energy difference between two states A and B is related to their probabilities P(A) and P(B) by

$$\Delta G(A->B) = -RT \ln[P(A)/P(B)]$$

So, if one observed a system one million times, and state B occurred once while state A occurred 999,999 times, A would be more favorable than B by ~13*RT. At 300K, this corresponds to 8.2 kcal/mol.

The application to conformational free energies was obvious – observe many molecules, count how many are in state A, how many are in state B, and calculate the free

energy difference. By the ergodic hypothesis, one can replace this sampling over many molecules with a time sampling over one molecule. Specifically,

$$\text{Lim}(N \to \infty)\ \Sigma(n: n=1 \text{ to } N)\ P(A,n,t_1) = \text{Lim}(T \to \infty)\ \Sigma(t: t=1 \text{ to } T)\ P(A,n_1,t)$$

Note that this is in the limit that time (t) goes to infinity. In other words, the single molecule must be observed for a sufficiently long time that the populations of all states of interest are converged. The long-time simulation of a single molecule or single system to produce ensemble behavior underlies virtually all molecular simulation – whether of neat liquids, proteins in solution, or crystalline phases.

Instead of comparing different conformations of the same compound (cis versus trans butane), a paper by C.H. Bennett[6] described a way to apply these methods to the comparison of two different chemical species (methane and ethane). This paper, suggested by Dr. Randall Radmer, then another student in the Kollman group, provided the major theoretical underpinning for the body of my thesis. Generalizing its conclusions, we developed a method called Chemical Monte Carlo/Molecular Dynamics (CMC/MD) that couples molecular dynamics for conformational sampling with Monte Carlo steps for "chemical sampling" – sampling between different molecules of interest. The third chapter of this thesis was originally published in the Journal of the American Chemical Society and describes the development and application of CMC/MD for solvation free energies and host-guest chemistry. One particular appeal of the CMC/MD method shown in this chapter is that it can be used to rapidly sort guests or ligands based

on their binding free energy, quickly identifying the best- and worst-binding compounds. The formal generalization of Bennett's derivation is also found in this paper.
Due in part to the constraints of journal publication, some of the specific details of the CMC/MD implementation within the AMBER molecular dynamics package were left out of the paper that composes Chapter 3. The specific implementation and accompanying thermodynamic issues are described in more detail in Chapter 4, as well as in Appendices 1 (pseudocode) and 2 (input files and specifications).

While the "tennis ball" host-guest system described in Chapter 3 was a good basic test for the CMC/MD method, a fruitful collaboration with Mats Eriksson, a postdoctoral researcher in the Kollman group showed that CMC/MD was a useful tool for the study of protein:ligand interactions. Chapter 5, which was published in the Journal of Medicinal Chemistry, describes the application of CMC/MD calculations to compare a number of related inhibitors of HIV Reverse Transcriptase. It also outlines a general strategy for computational structure-based lead optimization, and introduces PROFEC, a tool for lead optimization developed by Dr. Radmer. The particular challenges of the HIV-RT:inhibitor system led to the development of an adaptive form of the CMC/MD method, introduced in Chapter 5 and described in more detail in Chapter 6.

The HIV-RT project showed how useful CMC/MD could be in studies of protein:ligand interactions, but the family of ligands that we studied were all highly related, and only differed by small modifications. In addition, the host:guest chemistry and protein:ligand interactions studied in Chapters 3 and 5 are dominated by weak non-bonded interactions and solvation contributions. Some of the excitement surrounding the initial application of free energy calculations to biological molecules was due to the

parallels between the computational conversion of one sidechain to another and the biochemical technique of site-directed mutagenesis. Similarly, part of the "combinatorial revolution" mentioned above was the technique of exhaustive mutagenesis. Exhaustive mutagenesis is a biochemical method that allows the facile creation of all 20 natural amino acid mutants at a given position on the protein. In contrast to the weak noncovalent forces that define the interaction between proteins and their ligands, the differences between protein sidechains are influenced by internal bond, angle, and dihedral interactions and significant conformational entropy in addition to weak nonpolar interactions. All of these elements come into play in the solvation of amino acid side chains. More usefully, they define the contributions of various side chains in protein stability. Both peptide solvation and protein stability are studied in the calculations described in Chapter 7, which has also been submitted to the Journal of the American Chemical Society. The applications described in Chapter 7 also required the extension of CMC/MD to take advantage of the additional power provided by parallel computers. The pseudocode for a message passing (MPI) parallel version of CMC/MD is shown in Appendix 3. This major part of my thesis concludes with a general discussion of statistical issues associated with CMC/MD and the extraction of free energies from probabilities, presented as Chapter 8.

My interest in protein-ligand interactions and molecular recognition led to two other projects during my time at UCSF. Unlike the CMC/MD calculations described above, which attempted to produce quantitative or semiquantitative free energy comparisons of several ligands, these projects were directed towards a qualitative understanding of molecular recognition. In the first project, I studied complexes of the

thyroid hormone receptor ligand binding domain (TR-LBD) protein bound to its native ligand, thyroid hormone (T3). Using the PROFEC extrapolative free energy software written by Dr. Radmer (UCSF), I was able to reproduce the known structure-activity relationships for thyroid hormone analogs using only the structure of the TR-LBD/T3 complex. As part of this study, we also attempted to predict the structure and behavior of the TR-LBD in the absence of ligand. This work is still underway, and is not included in this thesis.

The second project, described in Chapter 9, was significantly more successful, and yielded a very rewarding synergy between theory and experiment. In a collaboration with Professor C.C. Wang and Dr. Narsimha Munagala, both here at UCSF, we compared the structure and dynamics of a human enzyme and its parasite analog bound to a number of different substrates. The enzymes are both phosphoribosyltransferases, and catalyze the conversion of a nucleobase and alpha-phosphoribosyl phosphate to the corresponding nucleotide monophosphate. While the human enzyme is relatively selective, accepting only hypoxanthine and guanine-based substrates, the parasite enzyme has the additional ability to process xanthine. The active sites of both enzymes are highly similar, so we attempted to explain the broader specificity of the parasite enzyme. Molecular dynamics simulations of several enzyme-substrate complexes showed a higher mobility and plasticity in the active site pocket of the parasite enzyme that contacts the crucial regions of the substrate. Our dynamics studies also helped to explain the function of several mutants of the parasite enzyme, and successfully suggested a secondary mutation that restored activity to a "dead" mutant enzyme. In addition to these calculations, we again

used the PROFEC software to aid in the design of prospective ligands, specifically those that show enhanced specificity for the parasite enzyme.

The final chapter of this thesis returns to the CMC/MD method and discusses some future directions and alternative approaches for free energy calculations in general.

Bibliography

1)Wolff, M. E. e. *Burger's Medicinal Chemistry and Drug Discovery*; 5 ed.; John Wiley & Sons: New York, 1995; Vol. 1, pp 1-1064.

2)Plunkett, M. J.; Ellman, J. A. *Sci Am* **1997**, *276*, 68-73.

3)Clackson, T.; Wells, J. A. *Trends Biotech* **1994**, *12*, 173-184.

4)Li, B.; Tom, J. Y. K.; Oare, D.; Yen, R.; Fairbrother, W. J.; Wells, J. A.; Cunningham, B. C. *Science* **1995**, *270*, 1657-1660.

5)Schultz, P. G.; Xiang, X. D. *Curr Opin Solid State Mat Sci* **1998**, *3*, 153-158.

6)Bennett, C. H. *Journal of Computational Physics* **1976**, *22*, 245-268.

Chapter 2:

Graphical visualization of mean hydration from molecular dynamics simulations

Jed Pitera[1] and Peter Kollman[2]*

University of California

[1]Graduate Group in Biophysics and

[2]Department of Pharmaceutical Chemistry

UCSF Box #0446

San Francisco, CA 94143 USA

*To whom correspondence should be addressed

Previously published in the Journal of Molecular Graphics, December 1998.

Abstract

How does one characterize water solvating a complex solute? Specific hydration of proteins and nucleic acids plays a key role in many biological processes. However, traditional pairwise descriptions of solvent structure (radial distribution functions, etc.) are incapable of adequately describing the hydration of these complex solutes. We have developed methods to visualize the average three-dimensional water structure surrounding a solute, as seen in a molecular dynamics (MD) simulation. Applications to simple solutes (sodium ion, n-methyl acetamide, 18-crown-6, (hydroxymethyl)phenols) will be presented, and the extension of the method to larger molecules of biochemical interest will be discussed.

Keywords

hydration visualization molecular dynamics AMBER MidasPlus

sodium ion n-methyl acetamide crown ether (hydroxymethyl)phenol

Introduction

Water is the solvent for virtually every biological process. Whether inside or outside a living cell, water surrounds the proteins, nucleic acids, and small molecules necessary for life. In many cases, water molecules make specific hydrogen bonds and electrostatic interactions with the surface atoms of these molecules. More generally, water is responsible for the hydrophobic effect that stabilizes the structure of proteins and drives many macromolecular interactions. Also, the water surrounding two molecules must be displaced if those molecules are to interact with one another. For example, both the enzyme active site and the substrate have to be desolvated before they can interact. More practically, water molecules must be displaced from a drug binding site on a protein before an inhibitor can bind. Clearly, it is important to describe and understand the hydration of biological molecules.

In this paper, we describe the development of some software tools that can be used in conjunction with molecular dynamics to describe the average structure of water molecules surrounding a solute. The present study is limited to small molecules, but these methods have been applied to biomolecular systems, including proteins and nucleic acids. The data we have collected show the position and orientation of water molecules in the tightly bound first and second shells of hydration surrounding each molecule. Molecular dynamics is used as a tool to generate a large number of realistic solvent conformations, and the average properties of these conformations are extracted and visualized.

Traditionally, water structure has been described by radial distribution functions (G(r)), which show the probability distribution of distances between an atom of the solute and a type of solvent atom (methane carbon and water oxygens, for example). For solutes

with a large number of atoms, like a typical protein (2000+ atoms), such an atom-based description is intractable. Instead, a description that is independent of the individual atoms of the solute is necessary. We have decided to describe the space surrounding the solute using a fixed Cartesian grid (F(x,y,z)). This permits our data collection and visualization software to be used on a broad range of systems -- from methane to DNA.

Two main types of data were collected in this study. The first is the Cartesian analog of the radial distribution function -- G(x,y,z) rather than G(r) -- the "water oxygen probability density". Properly normalized, this is the probability of finding a water oxygen in a particular volume of space during the molecular dynamics simulation. When converged, it gives a clear picture of the most favorable positions for water molecules near the solute, what is thought of as the "first shell of hydration.". This sort of technique has previously been applied to simulations of biological molecules by Lounnas, Pettit, et. al. (11). Beveridge and coworkers (12) have similarly used a superposition of snapshots along a molecular dynamics trajectory to suggest the hydration structure around nucleic acids. The superposition approach is hampered, however, by limited sampling: the twenty or so structures used can only suggest the highest maxima of the water probability density, and contain very little information about moderate- to low-probability regions. The continuous data collected in this study have been filtered and displayed based on simple statistical analysis, providing information about the structure of the water in both high- (traditional "hydration sites") and low-probability regions.

While these data are very useful, they do not completely describe the solvent structure. There is no information about the orientation of the water molecules in the "probability density." Consequently, we augmented the probability data with the "mean

dipolar vector field." This vector field shows the mean orientation of water molecules in a particular sector of space, indicating hydrogen bond and electrostatic interactions with the solute.

It must also be noted that prior studies of the hydration of both small solutes (13) and macromolecules (12) have examined the pairwise water-solute interactions in substantially more detail, including careful calculation of energetics and orientational parameters. The intent of this work, in contrast, is to describe a simple computational framework for the qualitative analysis of solvent structure that makes use of readily available software tools and interactive three-dimensional modeling to provide a vivid picture of solvation.

Methods

Simulation details

Molecular dynamics simulations were carried out for each solute solvated by a cube of water molecules. All simulations were run using the SANDER module of AMBER 4.1 (1) and the Cornell, et. al. force field (2) with RESP charges. The TIP3P water model was used. This model has been shown to accurately reproduce the free energies of solvation and hydrogen bond geometries for small solutes when used in conjunction with electrostatic potential derived charges (3). In addition, the TIP3P model adequately simulates the microscopic structure and bulk properties of water. Simulations were run with fully periodic boundary conditions. The systems were equilibrated at constant pressure, but data was collected from simulations at constant volume to facilitate the use of a fixed grid. Non-bonded interactions were truncated at 8 Angstroms. The

solute was held fixed in the center of the periodic box, but the solvent molecules were free to move and coupled to a temperature bath at 300K.

Data collection

Modifications were made to the SANDER molecular dynamics software to permit collection of information about the solvent structure on the fly. The simulation box was divided into cubic bins 0.5 Angstrom on a side. Data were collected every molecular dynamics timestep (2 femtoseconds), to give good statistics for the observed properties. For the probability densities, a three-dimensional array of integers is maintained, one per grid bin. At each timestep, the grid position of every water oxygen is determined. One "count" is then added to the integer array element at the corresponding grid position. At the end of the simulation, the array is output in a format suitable for the display software. The procedure for collecting the mean dipolar vectors is similar. Three variables are maintained per grid point in addition to the integer array: real accumulators for the x-, y- and z- components of the dipolar vector. At each timestep, the list of water molecules is again traversed. The position of each water oxygen is determined, and the integer counter incremented as above. In addition, cartesian components of the dipolar vector (vector sum of the O->H1 and O->H2 vectors) are calculated and added to the appropriate x-, y- and z- accumulators. When the simulation is complete, the mean x-, y-, and z- components of the dipolar vector are calculated for each grid bin. These are output as a list of grid coordinates and vector components.

Visualization

The MidasPlus software suite (4) was used to display and manipulate the data in this study, since it allows for real-time interactive manipulation of three-dimensional models as stereo images. All the data in this study are resolved in three spatial dimensions, but all the images included below are 2-D views for ease of viewing. The probability densities were projected onto a model of the solute using the MidasPlus program and its Density delegate, a facility for displaying electron densities (or other scalar fields) atop molecular models. The Density program allows interactive display, contouring and coloring of scalar fields. For each probability density, the mean density and standard deviation were calculated. All contouring was done at some number of standard deviations above the mean density, to differentiate the tightly associated hydration shell from the bulk solvent. The mean dipolar vectors were similarly displayed using MidasPlus. The Discern delegate was modified to permit interactive display, coloring and contouring of vector fields. Contours for the vector images were chosen by hand to show the significant features of the hydration shell, including hydrogen bond donors or acceptors, while minimizing clutter. Consequently, the vector images only show data points for high-probability regions of the solvent shell. The image of water residencies or lifetimes in the sodium ion hydration shell was also displayed using Discern. The wireframe images are direct screen captures from MidasPlus using the snapshot utility. All of the solid rendered images were generated using the MidasPlus rendering tools Conic (5) and Ribbonjr. All visualization and data display was carried out on a Silicon Graphics IRIS Indigo2 (150MHz R4400, Elan graphics) running IRIX 5.2.

Results

The methods described above were applied to four different solutes: a sodium ion, N-methyl acetamide, the crown ether 18-crown-6, and various isomers of (hydroxymethyl)phenols. The results for each system are presented below.

Sodium ion

The hydration of a sodium ion was selected as an initial test case for our methods. The simple structure of the Na+ hydration shell also permits a gradual introduction to our various graphical representations of hydration, from probability densities to mean dipolar vectors and lifetime data. Figure 1 shows these differing views of Na+ hydration. The first panel is a snapshot from a molecular dynamics trajectory, illustrating the difficulty in determining the structure of the solvent shell from instantaneous frames of molecular dynamics (Figure 1a). The second panel, displaying the water oxygen probability density around the ion from the same simulation, gives a much clearer picture of the first and second hydration shells (Figure 1b). Figure 1c adds orientational information to this view, showing the increased directional order of waters in the first hydration shell relative to the second. Our software also allows collection of dynamic information about the solvent structure, as seen in the final panel (Figure 1d). Water molecules remain in the first hydration shell much longer than the second; lifetimes in the second shell are only slightly longer than those in bulk water, indicating that these second shell waters are moving and exchanging rapidly with bulk solvent.

N-methyl acetamide (NMA)

NMA is a good example of a relatively simple molecule that is both a hydrogen bond donor and acceptor. The mean dipolar vector images in the first two panels of Figure 2 clearly show the hydrogen bonding patterns associated with both the hydrogen bond donor (N-H) and acceptor (C=O) groups of NMA (Figure 2a). The geometry of hydrogen bond donation by the N-H group is, as expected, more restricted than the geometry of hydrogen bonds accepted by the carbonyl (Figure 2b).

18-crown-6

The crown ether 18-crown-6 is a good example of where simple pairwise measurements fail to adequately describe hydration. The closely associated waters that donate hydrogen bonds to ether oxygens on either face of the ring typically interact with more than one ether group and often interfere or interact with one another, creating a complex, three-fold symmetric hydration shell. The network of water-solute and water-water hydrogen bonds in this shell was first observed in a Monte Carlo study (6), and our data replicate their observation of "bridging" water conformations. In these configurations, one water molecule sits just above the plane of the ring and donates two hydrogen bonds to ether oxygens. It accepts a hydrogen bond from a higher, "bridging" water that also hydrogen bonds to the third ether oxygen on the same face of the ring. The overall solvent structure around 18-crown-6 is shown in Figure 2c, and the high-probability regions, which correspond to waters making at least one hydrogen bond to the ring oxygens, are shown in more detail in the last panel (Figure 2d).

(Hydroxymethyl)phenols

The free energy of transfer from water to toluene is much more favorable (about 3 kcal/mol) for 1-3 (hydroxymethyl)phenol than either of the 1-2 or 1-4 isomers. Ben-Naim (7) has attributed this to the presence of a "bridging" water that forms strong hydrogen bonds to both hydroxyl groups of the 1-3 isomer (Figure 3a). The probability density from our MD simulations, however, shows no region of significant density that would correspond to this bridging water for the 1-3 (hydroxymethyl)phenol (Figure 3b, c), even when compared to the 1-4 isomer (Figure 3d). This confirms the free energy calculations of Sun, et. al. (8), who established that the free energies of transfer from the gas phase to water are very similar for all 3 phenols (within 1.0 kcal/mol), discounting any "anomalous" hydration of the 1-3 form.

Discussion

Our water probability density data clearly describe the average structure of a solute's hydration shell. In addition to these data, we have managed to collect information describing both the average orientation of solvent waters and approximate lifetime information, yielding a detailed picture of solvent structure and dynamics. The utility of this picture is shown by its application to a question of physical chemistry -- the anomalous transfer free energies of (hydroxymethyl)phenols.

These tools for graphical visualization of the mean solvent structure calculated from molecular dynamics simulation were developed for two reasons. First, it is difficult to extract information about the structure of the hydration shell from instantaneous coordinates of a molecular dynamics trajectory (for example, see Figure 1a). In addition, traditional radial distribution functions are inadequate for describing the hydration of complex solutes, especially large, moderately polar solutes like proteins and nucleic acids. Consequently, we have developed grid-based methods that capture the average structure of the hydration shell but are also extensible to permit study of large, biomolecular solutes. The on-the-fly software described above, where information is collected at every MD timestep, is too memory- and compute-intensive to use with large solutes. In collaboration with Thomas Cheatham, we have developed a post-processing utility that performs similar analysis and data collection on a previously calculated molecular dynamics trajectory. Application of these tools and methods to a 1 ns simulation of a DNA decamer clearly show the minor groove "spine of hydration" observed in high resolution X-ray crystallographic studies of DNA (9). Similar agreement

is seen between molecular dynamics simulations and high-resolution structures of RNA (10).

The programs developed in this article are available from the authors.

Acknowledgements

We are grateful for research support from the NSF (Grant CHE-94-17458) and the use of the the resources of the UCSF Computer Graphics Laboratory, Tom Ferrin, PI, supported by NIH RR-1081.

References

1. Pearlman, D., Case, D., Caldwell, J., Ross, W., Cheatham, T., Ferguson, D., Seibel, G., Singh, U., Weiner, P. and Kollman, P. AMBER 4.1, University of California, San Francisco. 1995.

2. Cornell, W., Cieplak, P., Bayly, C.I., Gould, I.R.; and others. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* 1995, **117**, 5179-5197.

3. Cornell, W., Cieplak, P., Bayly, C.I., and Kollman, P. Application of RESP Charges to Calculate Conformational Energies, Hydrogen Bond Energies, and Free Energies of Solvation. *J. Am. Chem. Soc.* 1993, **115**, 9620-9631.

4. MidasPlus, Computer Graphics Laboratory, University of California, San Francisco. 1994. Supported by NIH RR-01081

5. Huang, C.C., Pettersen, E.F., Klein, T.E., Ferrin, T.E., and Langridge, R. Conic: A Fast Renderer for Space-Filling Molecules With Shadows. *J. Mol. Graphics.* 1991, **9**, 230-236.

6. Ranghino, G., Romano, S., Lehn, J.M. and Wipff, G. Monte Carlo Study of the Conformation-Dependent Hydration of the 18-Crown-6 Macrocycle. *J. Am. Chem. Soc.* 1985, **107**, 7873-7877.

7. Ben-Naim, A. Solvent-induced interactions: Hydrophobic and hydrophilic phenomena. *J. Chem. Phys.* 1989, **90**, 7412-7425.

and Ben-Naim, A. Solvent Effects on Protein Association and Protein Folding. *Biopolymers.* 1990, **29**, 567-596.

8. Sun, Y., and Kollman, P. Are there water-bridge-induced hydrophilic interactions. *J. Phys. Chem.* 1996, **100**, 6760-6763.

9. Cheatham, T., Crowley, M., Fox, T., and Kollman, P. A molecular level picture of the stabilization of A-DNA in mixed ethanol-water solutions. *Proc. Natl. Acad. Sci.* 1997, **94**, 9626-9630.

10. Miller, J. Unpublished results.

11. Lounnas, V., Pettitt, B., Findsen, L. and Subramaniam, S.. A Microscopic View of Protein Solvation. *J. Phys. Chem.* 1992, **96**, 7157-7159.
and Rudnicki, W, and Pettitt, B. *Biopolymers.* 1997, **41**, 107-119.

12. for example, Subramanian, P.S. and Beveridge, D.L. A Theoretical Study of the Aqueous Hydration of Canonical B d(CGCGAATTCGCG): Monte Carlo Simulation and Comparison with Crystallographic Ordered Water Sites. *J. Biomol. Struct. Dyn.* 1989, **6**, 1093-1122.

13. Meng, E. and Kollman, P. Molecular Dynamics studies of the properties of water around simple organic solutes. *J. Phys. Chem.* 1996, **100**, 11460-11470.

Figure 1 :

A : Snapshot of sodium ion (Na+) in a box of water. Na+ is colored green.

B : Water oxygen probability density around Na+ from 2 nanosecond (ns) MD simulation. Na+ indicated by the green diamond. The probability density is contoured at 0.2 (white), 1.5 (blue), and 2.0 (navy) standard deviations above the mean density. Note the presence of both first and second hydration shells.

C : Mean water dipolar vectors from the same simulation. Again, Na+ is green. Vectors are only drawn for high-probability regions. Blue indicates the positive end of the dipole, red the negative end, and the body of the dipole is colored according to the probability density (blue -> red).

D : Water oxygen lifetimes in the Na+ hydration shell. A polygon is drawn at every grid point of significant density. The size of the polygon corresponds to the probability density, as seen in b. The polygon is colored to represent the average lifetime of a water molecule at that grid position, with the longest lifetimes (~0.3 picosec.) in red and the shortest (~0.05 picosec.) in blue and purple. Note that the longest lifetimes are associated with water in the first hydration shell. The second hydration shell shows lifetimes that are very similar to those of the bulk water.

Figure 2 :

A : Mean dipolar vectors around N-methyl acetamide, from 0.6 ns simulation. Vectors are displayed using the same scheme as figure 1c. Note the vectors corresponding to hydrogen bond acceptors near the N- H group, and the hydrogen bond donors surrounding the carbonyl.

B : Side view of the data in Figure 2a. The narrow directionality of the N-H group's hydrogen bond donation stands in contrast to the broader range of hydrogen bond orientations accepted by the carbonyl.

C : Water oxygen probability density surrounding 18-crown-6, from 0.24 ns trajectory. Density is contoured at 2 (yellow), 3 (blue), and 4 (navy) standard deviations above the mean. The highest probability is centrally located immediately above and below the plane of the ring, corresponding to a water that acts as hydrogen bond donor for two ether oxygens. The next highest probability consists of 3 lobes above the ether oxygens that project on either face of the ring. These are "bridging" waters, forming one hydrogen bond with an ether oxygen and another with the water immediately above the ring.

D : Water oxygen probability density as in Figure 2c, alternative view. Only the highest contour is displayed, showing the most preferred positions for water molecules on one face of the ring. 18-Crown-6 is drawn as a CPK model to make the water positions clearer.

Figure 3 :

A : Hand-built model of the putative "bridging water" interacting with 1,3-(hydroxymethyl)phenol. The bridging water oxygen is colored blue for contrast, and hydrogen bonding interactions shown with yellow dotted lines. Hydrogen bond donor-acceptor distances are reasonable, but we could not find a position for the bridging water that yielded both acceptable hydrogen bond distances and angles.

B : Water oxygen probability density surrounding 1,3- (hydroxymethyl)phenol. Density is contoured at 1.5 (blue) and 2.0 (navy) standard deviations above the mean. No peak in the density is visible between the hydroxymethyl and phenol functionalities, where a bridging water is proposed.

C : Side view of the data in Figure 3b, emphasizing the absence of a peak in the water oxygen probability density in the position expected for a bridging water.

D : Water oxygen probability density surrounding 1,4- (hydroxymethyl)phenol. Density is contoured at 1.5 (blue) and 2.0 (navy) standard deviations above the mean. This shows the expected hydration of the individual hydroxymethyl and phenol groups, for comparison with the 1,3 isomer.

Figure 1

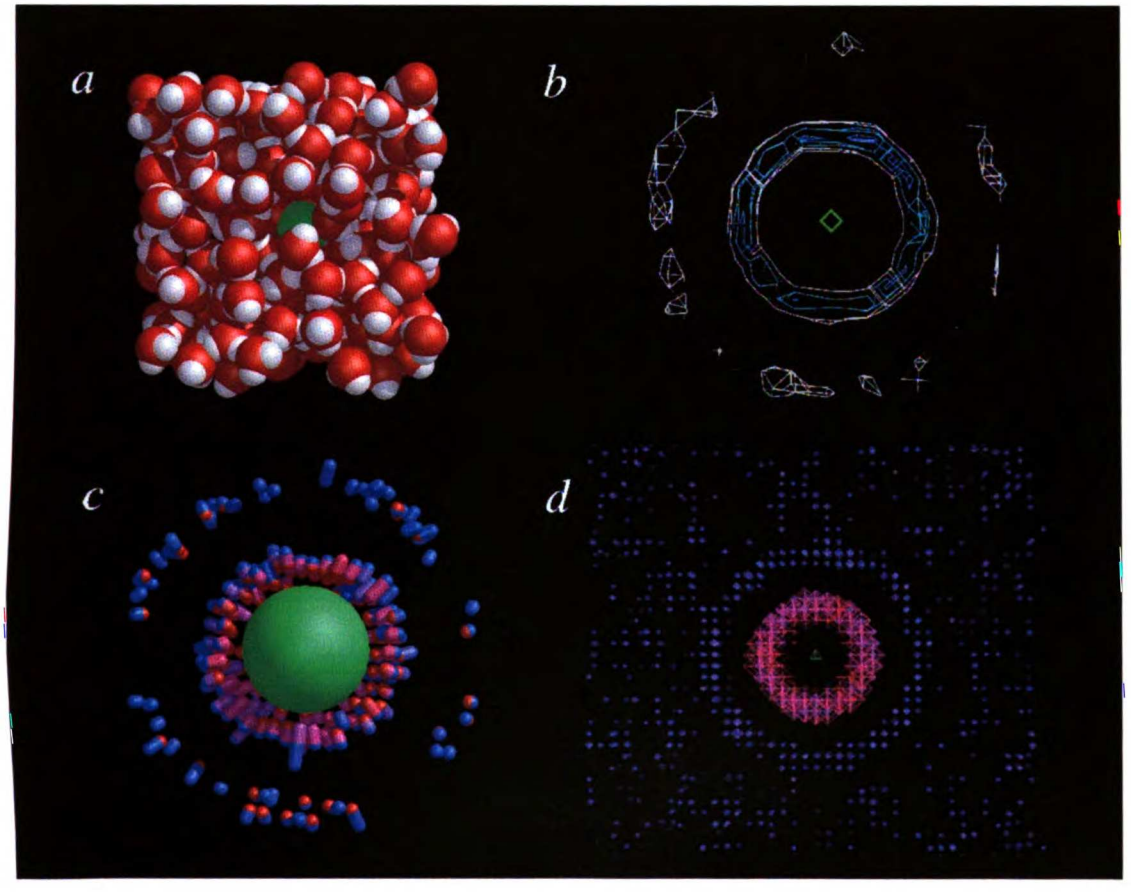


Figure 2

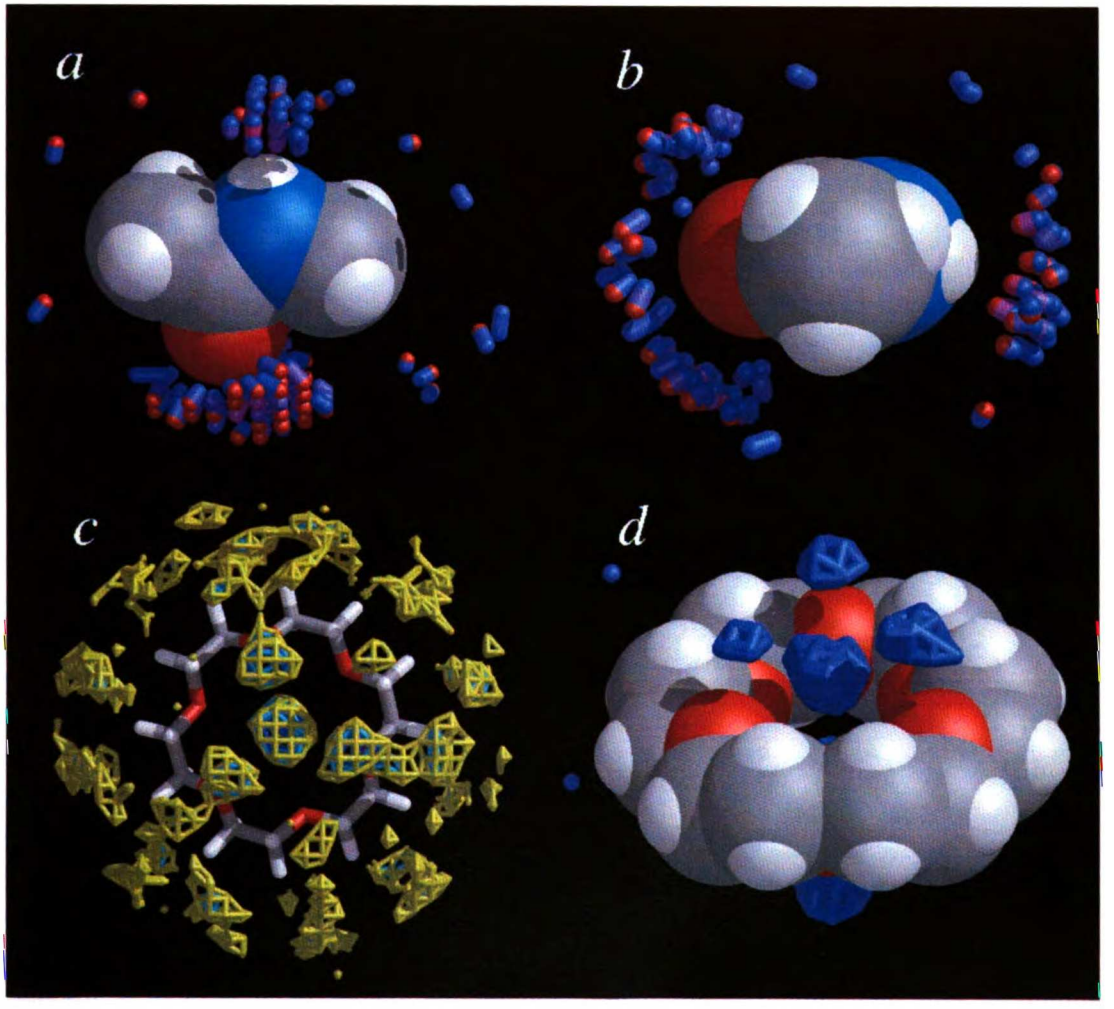
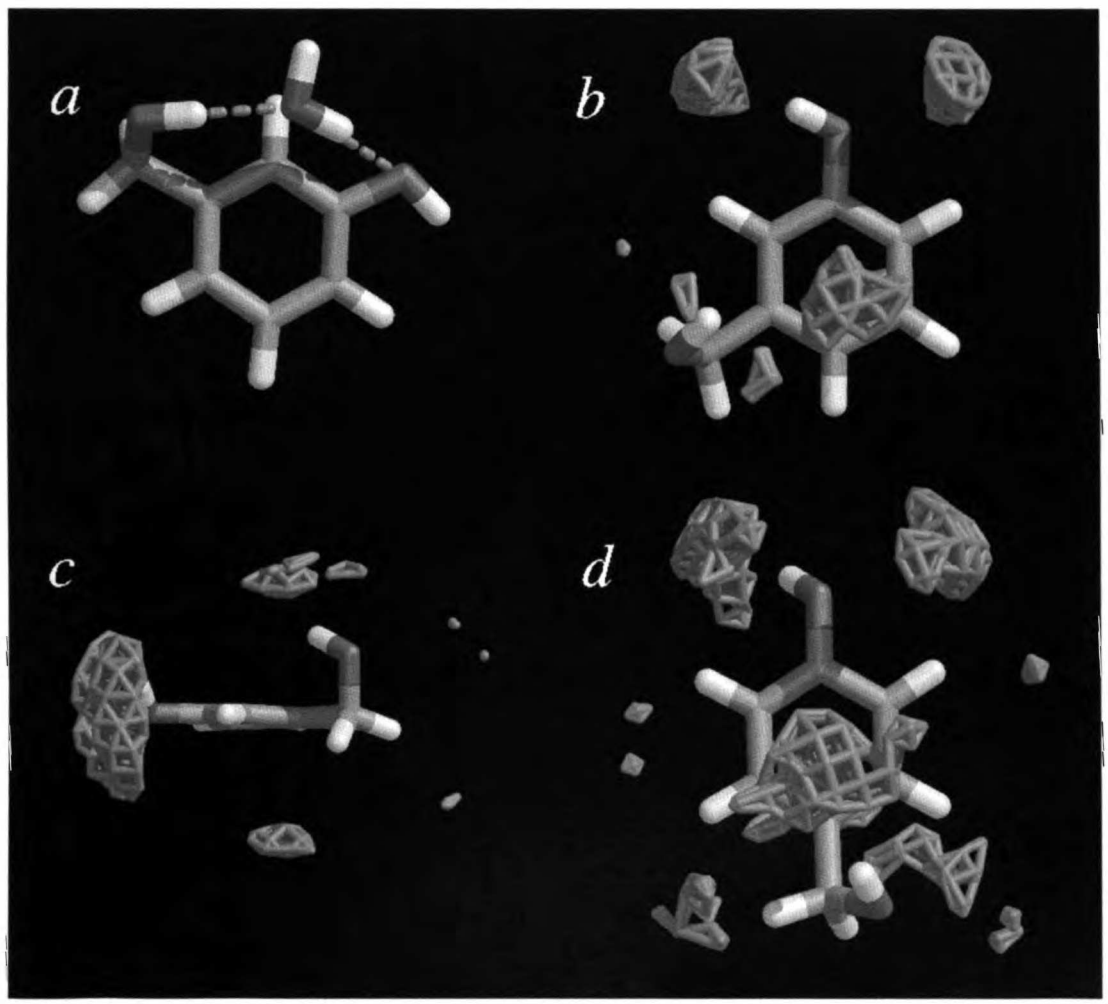


Figure 3



Chapter 3: Designing an optimum guest for a host using multi-molecule free energy calculations: predicting the best ligand for Rebek's "Tennis Ball"

Jed Pitera[1] and Peter Kollman[1,2,*]

[1]Graduate Group in Biophysics and

[2]Department of Pharmaceutical Chemistry

University of California at San Francisco

San Francisco, CA 94143-0446

August 22, 1997 revised 3 Feb 1998,

revised 19 Mar 1998

Published in the Journal of the American Chemical Society, Volume 120, Number 30,

Pages 7557-7567 (1998)

Abstract

We have predicted that difluoromethane (CH_2F_2) will be the highest-affinity guest for Rebek's "tennis ball" host[1] using a new approach to multi-molecule free energy calculations. The method, which we call chemical-Monte Carlo/Molecular Dynamics (CMC/MD), was first tested by calculating the relative free energies of solvation of a variety of molecules. Subsequently, we have used it to compare nine possible guests binding to the "tennis ball" host and predict that CH_2F_2 will bind more tightly to this host than CH_4, the strongest binding guest studied to date. This prediction has been supported by standard thermodynamic integration free energy calculations in which CH_4 was mutated into CH_2F_2 both in solution and in the host. Our results show the full power of such multi-molecule calculations -- namely, that they can be used to rapidly calculate and rank the relative binding free energies of many molecules from a single simulation, accelerating the discovery of novel ligands or guests.

Introduction

Molecular recognition is the selective, strong binding of a guest to a given host and is an essential element in biological systems, where receptor-ligand or receptor-inhibitor interactions are key to biological function. As a result, a detailed understanding of the process of molecular recognition and an ability to simulate it computationally could permit the efficient design of novel, viable drug candidates[2]. Thus, there are many computational approaches to ligand (or guest) design when the structure of the macromolecule (or host) is known. At one extreme of computational efficiency are approaches like DOCK[3], which can search databases of ~100,000 potential ligands using a very simple approach to "score" compounds and qualitatively suggest which will bind most tightly to the macromolecule.

At the other extreme are free energy calculation methods such as FEP[4] or thermodynamic integration (TI)[5], which have proven their utility in the detailed study of protein-ligand interactions. These use a thermodynamic cycle[6] (Figure 1) to analyze ligand binding. These methods calculate the relative free energies of the two ligands in the receptor (ΔGhost) and in solvent (ΔGsolv). The difference of these two values

$$\Delta G_{host} - \Delta G_{solv} = \Delta G_{bind}(2) - \Delta G_{bind}(1) = \Delta\Delta G_{bind} \qquad (1)$$

is the relative free energy of association, ΔΔGbind. Because the value of ΔΔGbind defines which ligand will bind to the receptor, it is crucial data for the design of novel inhibitors or ligands.

Free energy simulations have been successfully applied to calculate the relative binding free energy of protein-ligand complexes. Well known examples include the binding of trimethoprim and its congeners to dihydrofolate reductase[7], the comparison of various HIV protease inhibitors, and the relative binding of inhibitors to thermolysin and carbonic anhydrase [8]. However, these are expensive, pairwise comparisons between ligands. The detailed simulation of the protein-ligand complex required for just one such calculation currently requires anywhere from days to months of computer time. It is often cheaper and faster to simply carry out the relevant experiment. This has substantially limited the use of these methods in drug design or drug development applications. What, then, is the role of free energy methods in ligand design?

Free energy methods have the advantages of being thermodynamically rigorous and capable of fine distinctions between ligands ($\Delta G < 1$ kcal/mol) in favorable cases, so long as accurate potential functions are used[9]. However, this accuracy is not without a price -- these calculations are too slow for discovery of novel ligands. Lead discovery requires consideration and comparison of tens of thousands of compounds, a computationally prohibitive task using standard free energy methods.

In our opinion, the most efficient way to proceed in ligand design is to use a filtering strategy, where one uses rapid methods like DOCK to first suggest 10-100 possible leads from hundred thousand- or million-compound libraries. These compounds may then be examined by methods intermediate in accuracy and detail before resorting to traditional free energy calculations. One cannot realistically do full free energy calculations on many ligands because these methods are particularly inefficient when evaluating a family of related ligands. This is due to the pairwise comparison intrinsic to

standard free energy calculations -- to assess the relative free energies of ligands A, B, C, and D, at least three calculations must be carried out: one to compare A and B, one to compare B and C, and finally one to compare C and D. Since the lead refinement process often involves choosing between many possible modifications of a lead compound (each of which may involve a significant amount of synthetic chemistry), computational methods are needed that retain as much as possible of the accuracy of free energy calculations but have the ability to compare many ligands at a time.

Such "multi-molecule" free energy methods are actively being developed by many groups. Most notably, Kong and Brooks[10] have introduced "λ-dynamics": by expanding the extended Hamiltonian formalism[11,12] from one to several lambda variables, they have calculated relative solvent-state free energies for many species from a single simulation, and shown how expansion of the lambda-variable space can accelerate the convergence of traditional pairwise free energy calculations. The use of biasing potentials to improve the convergence of such simulations is also discussed in a general way. Other multi-molecule approaches include the calculation of relative free energies for many compounds by perturbation expansion from a single reference state, recently explored by Liu, et al.[13] as well as Radmer and Kollman[14].

In this paper we present a new multi-molecule free energy method and apply it to calculate the relative binding free energies for a series of small molecules binding to a rigid organic host. Specifically, we explored the binding of methane, ethane, and various halomethanes to the "tennis ball" dimer described by Branda, et. al.[1] Our chemical Monte Carlo - Molecular Dynamics method combines molecular dynamics to sample coordinate space with Metropolis Monte Carlo[15] to sample among various chemical

states of the system. The use of Monte Carlo sampling in "chemical space" was originally suggested by Bennett[16] and first used in a pairwise calculation of ion solvation by Tidor[17]. In the CMC/MD method, the solvation free energy of each ligand can also be included as a biasing potential in the Monte Carlo step to focus sampling towards the best binding ligands.

We have chosen the "tennis ball" host-guest system because it is experimentally well characterized and known to bind a range of ligands with varying affinity. It is also a case where theoretical calculations have complemented experiment. Specifically, Branda, et al.[1] were unable to detect the binding of tetrafluoromethane (CF_4) in their initial report. Free energy calculations carried out by Fox, et al.[18] suggested that CF_4 should have an affinity for this host intermediate between CH_4 and $CHCl_3$, the best and worst known guests. This prediction was subsequently confirmed by experiment. While the "tennis ball" had been shown to bind methane, fluoromethane, ethylene, dichloromethane and chloroform, we were interested in testing the entire range of fluoro- and chloro-substituted methanes binding to this host, an intractable series of calculations with the methods used previously. In light of the initial synergy between theory and experiment, we were excited to find our calculation predicts difluoromethane (CH_2F_2) to be an even better guest than methane.

CMC/MD is faster than the analogous thermodynamic integration calculations previously carried out by Fox et. al.[18], and it converges to the same relative free energies for each ligand. In addition, our method rapidly orders the ligands according to their binding free energies, well before the precise free energy values are completely converged. A similar effect is observed with both lambda-dynamics and Still's recent

work on enantioselectivity[19]. All of the above properties make these multi-molecule methods ideal for quickly comparing a family of related ligands and assessing their binding to a particular receptor. As such, we feel this chemical-MC/MD method will be useful in lead optimization and refinement, especially in comparison to traditional free energy methods.

Methods

The chemical Monte Carlo method is based on a derivation by Bennett[16]. This derivation shows how a Monte Carlo calculation can be used to determine the relative free energy of two chemical "states" (two solutes, two ligands, etc.) by a combination of Cartesian and chemical Monte Carlo steps. It is straightforward to generalize this formalism to the case of multiple chemical "states". The derivation and generalization are presented in Appendix I, along with a discussion of the similarities and differences between CMC/MD and other methods. It should be noted that Kong and Brooks' λ-dynamics derivation[10] is sufficiently general that it can also be extended to describe the CMC/MD approach, though both were developed independently.

Previously, combinations of Monte Carlo and molecular dynamics have primarily been used to improve the sampling of physical configurations. Notable examples are the hybrid Monte Carlo technique[20] and the MC(JBW)/SD method[21]. In the hybrid Monte Carlo method, molecular dynamics is used to generate "trial move" configurations which are then evaluated with Metropolis Monte Carlo criteria to generate a thermodynamic ensemble. The MC(JBW)/SD method uses Monte Carlo steps to "jump" between conformational minima that are separated by free energy barriers, thus allowing a single

simulation to explore a much broader set of configurations. These methods differ from the chemical-MC/MD approach described here in that they use a constant potential function. In contrast, the chemical-MC/MD method uses Monte Carlo steps to adjust the potential function, thereby representing the interaction of different ligands with the receptor. Instead of "jumping" between different Cartesian configurations and generating a Boltzmann ensemble of these configurations, we are essentially "jumping" between different ligands and generating a "Boltzmann ensemble" of ligands. In this respect, it is similar to Tidor[17]'s approach; however, we have extended this type of method to multiple, complex ligands in order to make it useful in the context of ligand design.

The use of Monte Carlo sampling between discrete chemical states allows us to further increase the utility of the chemical-MC/MD method. Specifically, there are two properties of interest when comparing ligands -- first, a rank order of the best binders, and second, the value of ΔΔGbind for each ligand. We want to find an optimal route to determine the relative free energy of binding, ΔΔGbind, for our ligands of interest.

Binding represents a balance between the free energies of the bound and free (solvated) states of the ligand. If we want to find the "best binders" our calculation must take into account the contributions of both these states. Drawing inspiration (and precedent) from the commonly used Monte Carlo technique of "umbrella sampling"[22], we can directly determine ΔΔGbind from our chemical-MC/MD simulation if we include the relative solvation free energies (ΔGsolv) as a "solvation offset" to the energy of each state. In the λ-dynamics derivation of Kong and Brooks[10], provisions are also made for

the inclusion of a biasing potential associated with each lambda-coordinate, though in the context of enhancing simulation convergence.

$$\Delta\Delta Gbind = \Delta Ghost - \Delta Gsolv \quad (7)$$

$$\Delta\Delta Gbind = -RT \ln < e^{-\Delta Ehost/RT} > -\Delta Gsolv \quad (8)$$

$$\Delta\Delta Gbind = -RT \ln < e^{-(\Delta Ehost - \Delta Gsolv)/RT} > \quad (9)$$

Equation 9 shows that if we know or can approximate ΔGsolv, we can include it as a biasing potential in our chemical-MC/MD simulation of the bound state. By its nature, the chemical-MC/MD method focuses sampling on the compounds with the most favorable free energy in a given environment. In solvent, these are the compounds with the most favorable solvation free energies. In the protein or host, these are the ligands with the most favorable ΔGhost. However, the quantity of interest is ΔΔGbind, not ΔGsolv or ΔGhost. Including the corresponding ΔGhost for each ligand as a biasing potential in a simulation of the bound state means that the calculated value is ΔΔGbind, and the simulation spends most of its time sampling the "best binders' rather than the ligands with the lowest free energy in the bound state (lowest ΔGhost). The net result is a rapid rank-order determination of the best binding ligands and a gradually converging determination of ΔΔGbind. A useful physical analogy suggested by Kong and Brooks[10] is that this process of finding the "best binder" truly corresponds to a competitive binding experiment in the laboratory, where many ligands present in solution are competing for a single binding site on a protein or host.

Computational Details

The chemical-MC/MD algorithm was implemented as part of the AMBER software package[23]. The SANDER molecular dynamics program was modified to carry out the Metropolis Monte Carlo sampling and collect, record, and report the necessary data.

During a simulation, all the solutes or ligands of interest are simultaneously included in the simulated system and their interactions calculated at every time step. However, the potential energy function is masked to reflect the chemical state of the system. At every time step, there is a single "real" ligand and the remainder are "ghosts". The "real" ligand interacts fully with the surroundings. The ghost ligands' interactions are calculated and recorded but do not affect the system energy or dynamics. In particular, the ghost ligands do not exert any forces on the surroundings. Also, no ligand ever interacts with another ligand. In effect, the "ghost" ligands are decoupled from the system surroundings. This is analogous to the "dual topology" approach to free energy calculations[24] except that we now have an "n-tuple topology" containing each of our n chemical species.

In the interests of simplicity and practicality, we have made a few approximations. First, the abrupt jumps between ligands mean that a newly "real" ligand does not have velocities appropriate for its surroundings. As a consequence, we randomly reassign the velocities of every particle in the simulation from a Maxwell-Boltzmann distribution whenever a Monte Carlo move occurs (Anderson temperature coupling)[25]. In addition, a single system temperature is calculated that includes the kinetic energy of every particle in the simulation, including the ghosts. This temperature

is maintained at 300K using a Berendsen temperature coupling scheme[26]. The error due to these approximations is small (there are <40 ghost particles in our 9-solute, 3377 atom simulation), and should be expected to cancel when considering the relative free energies of similar ligands from a single calculation.

The system is also maintained at constant pressure by a Berendsen algorithm[26]. In contrast to the temperature, the virial (and the pressure) only include interactions with the "real" ligand and the surroundings. We are presently evaluating alternative temperature- and pressure-coupling algorithms to improve the rigor of our calculations.

One issue in these calculations is ensuring that the ghosts sample configurations that are appropriate for the current configuration of the surrounding "context". If the ghosts are completely decoupled from the "context", sampling of ghost configurations is essentially random. This results in poor acceptance ratios for the Monte Carlo steps, since random ghost movements often generate unrealistic situations where ghost atoms overlap atoms of the "context". We have addressed this problem in two ways. First, all of the ligands are restrained to one another by harmonic potentials between their centers of mass. Second, the ghosts are allowed to feel the influence of the "context", but not vice versa. These "ghost forces" mean that atoms of the context exert forces on the ghosts but the ghosts remain invisible to the context. Ideally, we would correct the observed free energies for these biases, but we assume that they will cancel for comparisons of similar ligands from a single simulation. The net result of these approximations is a substantial improvement in the acceptance ratios for Monte Carlo steps, enhancing the efficiency of the calculation.

The chemical-MC/MD protocol is as follows. The system ("context" plus "real" and "ghost" ligands) is simulated for several steps (usually 1 picosecond) of molecular dynamics. This generates a novel configuration of the context, the real ligand, and the ghosts. Based on this configuration, the energies of each ligand are evaluated. A ligand is chosen at random (the "trial move"). The change in energy is evaluated and the trial move is accepted or rejected based on Metropolis Monte Carlo criteria[15].

$$\Delta E <= 0 \quad => P(\text{accept}) = 1 \tag{10}$$

$$\Delta E > 0 \quad => P(\text{accept}) = e^{-(\Delta E/RT)} \tag{11}$$

After the trial move is accepted or rejected, the outcome is recorded and molecular dynamics resumes, again simulating the interactions of the "context" and the currently "real" ligand. This cycle of coupled Monte Carlo and molecular dynamics steps is continued until the probability of observing each ligand converges.

While this approach is sufficient, it discards a great deal of information about each ligand. Specifically, we record the interaction energies of each ligand before selecting one for a Monte Carlo trial move. This history provides information about the "quality" of the Monte Carlo sampling and also allows us to estimate the free energy for poorly- or under-sampled states.

If an infinite number of Metropolis Monte Carlo steps were carried out on a given Cartesian configuration of the simulated system, the probabilities of each ligand would converge to the Boltzmann distribution for that configuration. That is,

$$\lim_{n\to\infty} P_j(r,\lambda_i) = \frac{e^{-\Delta E_j/RT}}{\sum_{i=1}^{n} e^{-\Delta E_i/RT}} \qquad (12)$$

Since we only carry out one Monte Carlo step for each Cartesian configuration considered, we record this "Boltzmann" probability data over the course of our simulation as a check on our Monte Carlo sampling. The Boltzmann-based P(ligand) values are averaged over every Monte Carlo step to yield an optimum probability P(ligand) for the simulation. In our converged simulations, these Boltzmann-based probabilities mirror the observed Monte Carlo history for each state.

Simulation specifics

1. Solvation

Relative free energies of solvation were calculated for solutes within a bath of TIP3P water molecules[27]. The parameters for each pair or family of compounds (including charges and geometries) were taken directly from the literature references to facilitate comparison between the chemical-MC/MD and FEP or TI calculations. Specifically the parameters for bromide and chloride were taken from Tidor's previously mentioned work[17]. The anisole and benzene data were from Kuyper, et. al.[28] , and Sun and Kollman's work on hydrophobic solvation provided the parameters for methane, ethane, and propane[29]. The charges, nonbonded parameters, and geometries for the substituted methanes were taken from Carlson, et. al.[30], supplemented by bond, angle, and torsional constants from the Cornell, et. al. AMBER force field[31]. In each case, the

simulation system consisted of all of the solutes of interest, plus anywhere from 500 to 800 TIP3P water molecules, simulated in a rectangular periodic box.

A modified version of the SANDER module of AMBER 4.1 was used for the molecular dynamics calculation [32]. A leapfrog integrator was used with a 2 femtosecond timestep. Metropolis Monte Carlo steps were evaluated every 1 picosecond (500 MD steps) for most systems. The system temperature was maintained at 300K by the previously described Andersen/Berendsen temperature coupling. The Andersen temperature coupling reassigned the velocities of every atom in the system in sync with the Monte Carlo steps (every 500 steps/ 1 ps). The pressure was kept at 1 atmosphere with the Berendsen coupling scheme, using the compressibility of bulk water (44.6×10^{-6} /bar) and a coupling constant of 0.2 ps^{-1}. An 8 Ångstrom cutoff was used for the nonbonded interactions, with updates to the pairlist made every 10 or 20 dynamics steps. All bonds were constrained to their equilibrium lengths using the SHAKE algorithm[33].

Since the ghosts are partially or completely decoupled from the rest of the system, something is necessary to keep them from drifting out of the vicinity of the binding cavity. For our initial test calculations, we simply constrained the analogous atoms (the carbon of methane and one carbon of ethane, or the phenyl rings of anisole and benzene, for example) of each solute to overlap through a nonphysical "bond" of length 0.0 Ångstroms.

Each set of solutes was solvated and then equilibrated at 300K for at least 100 ps of dynamics during which time no Monte Carlo moves were made. After the equilibration phase, Monte Carlo steps were initiated and the free energy calculation begun. Total simulation length for these calculations was anywhere from several

picoseconds to 2.4 nanoseconds. Standard deviations were calculated for converged calculations by dividing the statistics from the total simulation into 4 to 8 bins depending on the simulation length and calculating a mean and standard deviation over all the bins.

2. Binding

For our binding free energy calculations, we studied the "tennis ball" host-guest system synthesized and characterized by Branda, et. al.[1]. The host and solvent parameters were the same as described by Fox, et al[18]. This prior calculation also provided parameters for methane, chloroform, and tetrafluoromethane. Charges and parameters for fluoromethane were supplied by Reyes[34]. The values for chloromethane and dichloromethane were based on the chloroform parameters and tested as part of a new AMBER parameterization for organic solvents by Fox[35]. Ethylene parameters were developed by using default parameters for sp2 carbon and associated hydrogen from the Cornell force field. All charges were determined using the RESP procedure to fit charges to electrostatic potentials from *ab initio* Hartree-Fock calculations using a 6-31G* basis set[36].

In this "tennis ball" calculation, the simulation system consisted of 2 host molecules, 631 rigid chloroform solvent molecules, and either four (methane, fluoromethane, tetrafluoromethane, chloroform) or nine (methane, ethylene, fluoromethane, difluoromethane, trifluoromethane, tetrafluoromethane, chloromethane, dichloromethane, and chloroform) ligands. In contrast to the solvent, all ligands were treated as having flexible angles and torsions but rigid bonds. The total system size was either 3356 or 3377 atoms, and was simulated in a rectangular periodic box

approximately 46 Angstroms on a side. Figure 5 shows a stereo view of the "tennis ball" dimer with a representative configuration of difluoromethane in the binding cavity.

Again, the SANDER module of AMBER was used for the molecular dynamics calculation. A leapfrog integrator was used with a 2 femtosecond timestep. Metropolis Monte Carlo steps were evaluated every 1 picosecond (500 MD steps) for most systems. The system temperature was maintained at 300K by the previously described Andersen/Berendsen temperature coupling. The Andersen temperature coupling reassigned the velocities of every atom in the system in sync with the Monte Carlo steps (every 500 steps/ 1 ps). The pressure was kept at 1 atmosphere with the Berendsen coupling scheme, using the compressibility of bulk chloroform (108.60×10^{-6} /bar) and a coupling constant of 0.2 ps^{-1}. A 12 Ångstrom cutoff was used for the nonbonded interactions, with the pairlist update every 25 dynamics steps. A correction for the cutoff was included in the system energy and pressure[37]. All bonds were constrained to their equilibrium lengths using the SHAKE algorithm[33]. Aside from the chemical Monte Carlo steps and the Andersen temperature coupling, the dynamics simulation protocol is identical to that used by Fox, et al. for thermodynamic integration calculations.

Since the ghosts are partially or completely decoupled from the rest of the system, something is necessary to keep them from drifting out of the vicinity of the binding cavity. We chose to constrain the center of geometry of each ligand to that of one other ligand using a flat-well restraint. This restraint was used with a force constant of 500 kcal/mol and distances r1 = 1.0Å and r2 = 1.5Å. Since the purpose of the restraints was merely to keep the ligands in the vicinity of the binding cavity, the restraint energy was not included in the Monte Carlo calculation. Regardless, since this restraint is identical

for each ligand, its contribution to the relative free energy of any two ligands largely cancels. The sum of the average restraint energy for the eight ghosts in our 9 guest simulation was less than 1.0 kcal/mol, and we found that inclusion of the "ghost forces" substantially reduced the restraint energy while improving the sampling. Thus, it is highly unlikely that the restraint energy will differentially affect the calculated free energies for the different guests. This is further supported by the results of our 4 guest simulations, where the order of free energies is completely consistent with full TI calculations. In the future, it may be more appropriate to harmonically constrain each ligand to the center of the binding cavity, an approach that would permit analytic correction of the restraint contribution to the free energy, as outlined by Wang and Hermans[38], but this idea would also be limited to relatively simple ligands and binding geometries.

Total simulation length for our binding free energy calculations was either 400 or 800 picoseconds. This should be contrasted with the equivalent TI calculations, which required 200-800 picoseconds to calculate ΔGhost for a single pair of ligands[18].

The relative free energies of solvation in chloroform for each of our ligands were calculated using the GIBBS module of AMBER 4.1 and a simulation protocol similar to that described by Fox, et. al. We used the same general methodology, but instead of dividing our thermodynamic integration (TI) calculation into 101 windows of 3 ps each, we found better results from a simulation protocol of 26 larger windows each 12 ps in length. Improved convergence of the free energy value calculated for each window was seen, and the total free energy values were analogous to those determined by Fox and

Reyes. Our solvation free energy data are shown in Table 1 as the average free energy for forward and reverse calculations plus or minus the hysteresis between the two runs.

Results

1. Solvation

Before applying CMC/MD to a new problem, we first tested it by calculating the relative free energies of solvation in water for several families of compounds that had previously been studied by TI or FEP calculations. These results are presented in Table 2, along with the corresponding free energy data from the literature for comparison. While the data are not converged for every family of compounds studied, the CMC/MD method does a good job of determining the rank order and magnitude of the solvation free energies in each case. The relative solvation of bromide and chloride ion was studied by Tidor with the hybrid MC/MD method described previously, and our results are in reasonable agreement with his calculations. More difficult tests are the comparisons of methane versus ethane and anisole versus benzene. In particular, the comparison of anisole and benzene is significant because the steric difference between the two compounds is relatively large, yet our method gives a reasonable estimate of the free energy difference.

After these pairwise comparisons, we studied two families of compounds. Methane, ethane, and propane were studied in a single simulation that yielded quite accurate free energy estimates for all three compounds with a reasonable computational cost. Methanol and the substituted methanes formed the other family of compounds studied. They cover a broad range of polarity and free energy, yet our method rapidly

gets the correct rank order and order of magnitude of the relative free energies of solvation. This latter set of molecules had been studied by Kong and Brooks[10] using λ-dynamics, so it was appropriate to show that our procedure could also appropriately rank the free energies of solvation of these molecules.

Binding

Once we had achieved these promising results on relative free energies of solvation, we then applied the CMC/MD method to study the binding of four guests to the "tennis ball" host. The guests chosen were those previously studied by Fox, et. al (CH_4, CF_4, $CHCl_3$) and Reyes (CH_3F), so that thermodynamic integration data was readily available for comparison. As an initial test, we did not include the solvation offset in our calculation. The results of this determination of ΔGhost are shown in Figure 2. Figure 2a shows the relative populations of each ligand in the host accumulated over a 400 picosecond calculation. These data are converted into free energies relative to methane in Figure 2b. Clearly, our method rapidly indicates that CH_3F is the most favorably bound ligand. However, this calculation does not include the solvation free energies.

By including the solvation free energies as offsets to our Monte Carlo sampling, we can directly determine ΔΔGbind, as shown in Figure 3. This 1 nanosecond calculation is now dominated by CH_4 instead of CH_3F, in good agreement with the actual relative binding free energies (ΔΔGbind). This is one of the major strengths of our method -- most of the simulation time is spent sampling the ligands with favorable binding free energies. The calculation thus rapidly focuses on the real compounds of interest. In addition, the rank order of binding is rapidly determined (Figure 3a). Our

calculation shows that guests are preferred in the order $CH_4 > CH_3F > CF_4 > CHCl_3$, as observed experimentally. Extended calculations converge to well-defined values of the binding free energy (Figure 3b). Our calculated values (Table 3) are in good agreement with both experimental data and earlier free energy calculations.

We subsequently decided to apply our method in a predictive fashion to a simulation that included nine guests. We chose all of the guests that had been observed experimentally (CH_4, H_2CCH_2, CH_3F, CF_4, CH_2Cl_2, $CHCl_3$) as well as the remaining fluoromethanes (CH_2F_2, CHF_3) and chloromethane (CH_3Cl). We did not include carbon tetrachloride (CCl_4), since we expected it to be even less favorably bound than chloroform, the worst guest observed. Using the relative solvation free energy data from Table 1, we carried out a single 800 picosecond simulation on this family of compunds. The population and free energy data are shown in Figure 4. The data for the 9-guest case are substantially less well converged than the simpler 4-guest calculation, but several results are clear. Most importantly, the calculation quickly shows that difluoromethane is clearly the best binding compound and chloroform the worst. Our data also agree with Branda, et. al.[1] that methane and ethylene are approximately equally well bound by this host and that CH_2Cl_2 is preferred to $CHCl_3$. The predicted rank order from our calculation is $CH_2F_2 >> (H_2CCH_2, CF_4, CH_3F, CH_3Cl, CH_4, CHF_3) > CH_2Cl_2 >> CHCl_3$. However, we do not think these data are perfectly converged. Particularly, we are most confident about the prediction of the best- and worst-binding compounds and less certain of the ordering of "intermediate" binders. Still, the utility of this method in rapidly sorting the compounds by approximate binding free energy is clear.

Since the CMC/MD calculation strongly suggests that CH_2F_2 is the best guest for the "tennis ball", we decided to test this prediction with a thermodynamic integration calculation. Using a protocol identical to that used for the calculation of solvation free energies in chloroform and similar to that previously used to calculate ΔGhost for other guests binding to this host, we perturbed methane to difluoromethane in the cavity of the "tennis ball". The calculated ΔGhost was -1.88 +/- 0.03 kcal/mol; subtracting the previously calculated ΔGsolv of -1.44 kcal/mol yields the result that difluoromethane is preferred in this host by -0.44 kcal/mol. This is in good agreement with our chemical-MC/MD estimate of -0.76 kcal/mol, and provides a strong internal validation of our new method. We examined the complex of difluoromethane bound to the host dimer in detail in order to understand the structural basis for its affinity. Figure 5 shows a representative configuration of the complex. The guest is slightly off-center in the host cavity, and is oriented so that each fluorine projects towards one of the gaps between the two halves of the host. This arrangement appears to maximize the favorable van der Waals contacts between guest and host without straining the guest, either host monomer, or any inter-monomer hydrogen bonds. Energy minimization and analysis of the electrostatic and van der Waals interactions of the complex *in vacuo* support this conclusion.

Comparison of the minimized CH_4/host and CH_2F_2/host complexes leads to an interaction energy difference of ~4 kcal/mol favoring CH_2F_2. Of this difference, 3.5 kcal/mol is due to van der Waals energy and 0.5 kcal/mol from electrostatic interactions. We can include the solvation free energy of these two guests in a qualitative way using the data in Table 1: CF_4 is more favorably solvated than CH_4 by ~0.6 kcal/mol,

suggesting that each fluorine yields 0.6/4 = 0.15 kcal/mol solvation due to van der Waals interactions with the chloroform solvent. Thus, the ~1.4 kcal/mol improved solvation of CH_2F_2 relative to CH_4 has a ~1 kcal/mol contribution from electrostatic energies, which makes sense given the dipolar character of CH_2F_2 and the nonpolar nature of methane. Comparing these solvation free energies with the energy minimization results suggests that the "tennis ball"'s preference for CH_2F_2 is due to van der Waals interactions, since the favorable electrostatic contribution for CH_2F_2 versus CH_4 is even larger in solution than in the host cavity.

Of course, the above is only a qualitative analysis, but is unequivocal in the predominance of van der Waals forces. As noted, CH_2F_2 can gain van der Waals attractions for its fluorines by pointing them towards the inter-monomer gaps in the hosts. Directing the fluorines towards the aromatic ring leads to unfavorable repulsion and strain in the host. Similarly, the replacement of fluorines with chlorines is also disfavored, as CH_2Cl_2 is seen to be less favorably bound than CH_4 by both theory and experiment. One can now also rationalize the weaker binding of CHF_3 because the geometry of the host and guest preclude the formation of strong van der Waals interactions for the third fluorine group.

Discussion

We have developed and applied the chemical Monte Carlo/MD (CMC/MD) method to successfully determine the relative binding free energies of several nonpolar guests binding to an organic host. With sufficient sampling, our calculation yields free energies in close agreement with previous thermodynamic integration calculations, as

well as experiment (Table 2, figure 3, etc.). Our multi-molecule free energy method has a great deal in common with the previously published lambda-dynamics work of Brooks and Kong[10], though it was independently derived from theoretical work by Bennett[16] and the subsequent coupled MC/MD work of Tidor[17], as well as Radmer's work[14] on other multi-molecule free energy methods.

We have also shown that the solvation free energy may be included in the Monte Carlo stage of the calculation to focus sampling on the most favorably bound ligands. Application of this "solvation offset" to our 4-guest calculation shifts the predominant state from CH_3F (the guest with the lowest free energy in the bound state) to CH_4 (the most favorably bound guest). In addition, our calculations rapidly yield the observed preference of the host for various guests ($CH_4 > CH_3F > CF_4 > CHCl_3$). After this paper was submitted for review, works have appeared by both Guo, et al.[39] and Jarque and Tidor[40] which also demonstrate the feasibility and utility of sampling on the $\Delta\Delta$Gsolv (or $\Delta\Delta$Gbind) surface.

To demonstrate the real utility of multi-molecule free energy methods, we have carried out a predictive calculation -- the first using such techniques -- comparing 9 guests bound to the host. The rank order from our calculation correlates somewhat with that observed by Branda, et al.[1], for the five guests studied experimentally, and suggests that CH_2F_2 would be even more favorably bound to the host than methane. We have tested this prediction internally with a TI calculation that also finds CH_2F_2 a better guest than methane. This demonstrates the ideal application of multi-molecule methods -- they permit consideration of the relative binding free energy for many more compounds than

could be studied otherwise, and rapidly pick out promising binders for further computational or experimental study.

There are some limitations to our method. First, it is restricted to comparisons between relatively similar ligands, or at least compounds of similar volume. Ligands with substantial steric differences (methyl versus phenyl derivatives, for example) are difficult to compare with the CMC/MD method, since the abrupt jumps between states do not sample large changes in volume well. However, we have applied our method to accurately calculate the relative free energies of solvation of anisole and benzene (Table 2). This change from a hydrogen to a methoxy group gives us confidence that we can apply our method to pharmacologically relevant changes[7]. It should also be noted that free energy calculations which involve large steric changes are still a challenging prospect for more traditional FEP and TI calculations as well[41,42].

Furthermore, CMC/MD appears to be much more efficient for calculating ΔGhost or ΔΔGbind than it is for calculating ΔGsolv. The preorganized cavity of a host or protein binding site makes for more efficient sampling than the transient, rapidly fluctuating cavities that surround a solute in a solvent like water. Acceptance ratios are much higher for calculations of ΔGhost in our test system than they were for trial ΔGsolv calculations in water. In the present study, we avoided this issue by using ΔGsolv values from thermodynamic integration calculations. Since we expect to eventually use our chemical-MC/MD method to compare many ligands bound to a protein, we are exploring alternative, less expensive methods for calculating ΔGsolv, such as continuum solvent methods [43]. In addition, several projects are underway to use this method to compare the binding of multiple drug molecules to protein targets, with excellent initial results[44].

Both CMC/MD and lambda-dynamics[10] are "multi-molecule" free energy methods. They provide the framework for rapid comparison of the free energy of several molecules experiencing a common environment. Kong, et al.[10] and Guo, et al.[39] have both shown the power of lambda-dynamics in solvation free energy calculations and in accelerating the convergence of traditional free energy simulations. In contrast to lambda-dynamics, CMC/MD is a more approximate method -- the rapid jumps in chemical space permit us to save time by avoiding the simulation of intermediate states, but also appear to require longer simulation times to yield converged free energy statistics. Our "n-tuple topology" approach also means that CMC/MD is more readily extensible to comparisons between ligands of arbitrary topology, an essential issue in drug design calculations. This is illustrated here by our consideration not only of substituted methanes, but also ethylene as guests for the "tennis ball" host.

These multi-molecule methods occupy a middle ground of detail and accuracy in the range of computational methods that are applied to structure-based drug design. At one extreme there are docking and empirical scoring methods that can examine hundreds of thousands of compounds and possibilities. Traditional free energy perturbation methods occupy the other extreme, providing a detailed assessment of only two compounds. CMC/MD and λ-dynamics both give a relatively accurate free energy assessment for 5-10 compounds. A simpler dynamics-based free energy estimation method (the linear interaction approximation, or LIA) has been introduced by Aqvist[45]. Radmer and Kollman have introduced PROFEC, a tool for optimizing ligand affinity based on extrapolations from a single dynamics calculation[14]. A similar method from Liu, Mark and van Gunsteren uses extrapolations from a simulation of a single solute to

estimate free energies for a range of related compounds, with modest success[13]. Given the range of methods available, one can imagine a funneling process, where the best compounds found by a docking method are studied in more detail by LIA or chemical-MC/MD methods, possible modifications are suggested by PROFEC, and final lead optimization is guided by careful CMC/MD or FEP/TI calculations. At each stage of the process, the number of compounds studied is whittled down from thousands to hundreds or tens or even pairs of compounds, while the level of detail, accuracy, and computational expense per compound is simultaneously increased.

With the development and deployment of modern parallel supercomputers and workstation clusters, we also envision a "coarse-grained" parallel implementation of our method, where one chemical Monte Carlo - MD calculation is run on each of several processors. If we can compare 5-10 ligands per processor and one "reference" ligand is common to every processor, we can expect to compare and rank hundreds of ligands at once. In addition, the chemical Monte Carlo method is intrinsically suitable for more simple applications of coarse-grained parallelism. The results from two simulations of the same family of ligands can be added together directly to yield improved (or more rapid) free energy estimates. This is a sharp contrast to traditional FEP or TI calculations, where the need to smoothly integrate along the "reaction coordinate" means that one must either do additional preparatory simulations to divide the task among processors[46] or develop intrinsically fine-grained algorithms.

Finally, the use of computational methods to study the ideal guest for Rebek's "tennis ball" host has led to an exciting result -- the prediction that CH_2F_2 would be a better guest than CH_4. Analysis of the structure and energies yielded a rationalization of

this preference, based on several factors. First, fluorine groups are of the appropriate size to fit neatly in the inter-monomer interface. The geometry of host and guest permit only two positions on the guest to make such favorable interactions, which may also explain some of the host's preference for CH_2Cl_2 versus $CHCl_3$. Finally, the greater van der Waals well depth of fluorine relative to hydrogen makes this interaction stronger for CH_4 than for CH_2F_2. Thus, this study has met the fundamental requirements for any computational method -- it has qualitatively and semi-quantitatively reproduced known experimental data, made a prediction for a new guest, and provided mechanistic and structural insight into the origin of the increased affinity of this guest for the host. This offers encouragement for the continued utility of CMC/MD and other "multi-molecule" free energy calculations in the study of host-guest complexes, whether they be organic systems like the one described herein or biological problems like protein-ligand interactions.

Acknowledgements

PAK would like to acknowledge the support of NIH grants GM-29072 and GM-39552. JWP would like to acknowledge the support of the NSF, and thank R. Radmer and R. Stanton for many helpful discussions. T. Fox and C. Reyes were very helpful in supplying parameters and TI data for comparison.

Appendix I

Derivation

Consider n chemical states, numbered 0 through n, described by identical coordinates (**r**) but differing only in the potential functions (Un) describing them. The free energy difference between any two states is the ratio of their corresponding configurational integrals:

$$\Delta G(m \rightarrow n) = -kT \ln \frac{Qn}{Qm} \tag{13}$$

where such an integral has the form

$$Qn = \int e^{-Un(\mathbf{r})/kT} d\mathbf{r} \tag{14}$$

Bennett showed that this ratio of configurational integrals can be calculated by a simulation that samples various (**r**) and simultaneously carries out a special type of Metropolis Monte Carlo move. Specifically, the Monte Carlo move does not involve a change of coordinates (**r** -> **r**') but instead involves a change in potential function (Un -> Um). The Metropolis function

$$M(x) = \min\{1, e^{-x}\} \tag{15}$$

or more specifically

$$M(\Delta U/kT) = \min\{1, e^{-\Delta U/kT}\} \tag{16}$$

defines the acceptance probability for this potential-switching move just as in traditional cartesian applications of Metropolis Monte Carlo, where

$$\Delta U = U(\mathbf{r}') - U(\mathbf{r}) \tag{17}$$

In our case, however, ΔU is the change in energy involved in switching the system from potential function Um to Un:

$$\Delta U = Un(\mathbf{r}) - Um(\mathbf{r}) \tag{18}$$

For any physical configuration of the system ($\mathbf{r}$) the acceptance probabilities for any pair of potential-switching moves (m -> n and n -> m) are related by

$$M(Un - Um)/M(Um - Un) = e^{-(Un - Um)} \tag{19}$$

(where we have omitted the factor of kT from the exponential for clarity) which can be rearranged into the form

$$M(Un - Um)\, e^{-Um} = M(Um - Un)\, e^{-Un} \tag{20}$$

Since both potential functions apply to the same coordinate space ($\mathbf{r}$), one can integrate both sides of the above over all possible values of ($\mathbf{r}$), yielding equation (21)

$$\int M(Un-Um)e^{-Um}d\mathbf{r} = \int M(Um-Un)e^{-Un}d\mathbf{r} \tag{21}$$

Multiplying the left side of this equation by the identity Qm/Qm and the right by Qn/Qn gives equation (22):

$$\frac{Qm}{Qm}\int M(Un-Um)e^{-Um}d\mathbf{r} = \frac{Qn}{Qn}\int M(Um-Un)e^{-Un}d\mathbf{r} \tag{22}$$

The terms

$$\frac{Qm}{Qm}\int M(Un-Um)e^{-Um}d\mathbf{r}$$

and

$$\frac{Qn}{Qn}\int M(Um-Un)e^{-Un}d\mathbf{r}$$

are simply canonical averages in the Qm and Qn ensembles, respectively. A canonical average has the form

$$\langle F(U,\mathrm{r})\rangle = \frac{\int F(U,\mathrm{r}) e^{-U(\mathrm{r})} d\mathrm{r}}{Q} \tag{23}$$

so we can rearrange equation 22 to yield the ratio of interest:

$$\frac{Qm}{Qn} = \frac{\langle M(Um-Un)\rangle_n}{\langle M(Un-Um)\rangle_m} \tag{24}$$

The physical interpretation of the above is that a simulation which includes potential-switching Metropolis Monte Carlo moves in addition to some form of configurational sampling will sample the potential states Um and Un in proportions that reflect the free energy differences between states m and n. Bennett did go on to point out that it is often more efficient to evaluate the canonical integrals $\langle M(Um - Un)\rangle_n$ and $\langle M(Un - Um)\rangle_m$ directly. However, this is only true if one knows a priori which free energy differences (and states) are of interest.

For the multi-state case where we start with many states (0 . . . n), the full chemical Monte Carlo process has its own advantages. Specifically, we are interested in the relative free energies of each state, but our primary goal is finding the states of lowest free energy. Consequently, we do not want to waste computational time calculating detailed free energies for states that are not of interest. In the binding free energy

applications discussed in this paper, the states of lowest free energy correspond to the ligands that are the "best binders" for a given receptor.

In practice, the full chemical Monte Carlo method is implemented by adding a set of additional coordinates to the simulated system, one for each chemical state of interest. These coordinates (λ_i) are analogous to the "lambda" coordinates used in FEP and TI calculations. For a set of chemical states 0 to n, we have λ_0 to λ_n. The potential function used is of the form

$$U(\mathbf{r},\{\lambda_i\}) = \sum_{i=1}^{n} \lambda_i U_i(\mathbf{r}) \qquad (25)$$

where ($\mathbf{r}$) includes coordinates for each chemical state of interest plus the surrounding context (solvent, protein, or host molecule).

The lambda values are also subject to two constraints; first, each λ_i is either 0 or 1. Second, the sum of all λ_i is constrained to be 1. The result of these two constraints is that the calculation only simulates the end states of interest, and only simulates one at a time.

Comparison of methods

As noted in the introduction, Tidor has previously presented an implementation of Bennett's ideas that uses molecular dynamics to sample confguration space and Monte Carlo methods to take steps along a chemical "reaction coordinate" between two end states[17]. The "reaction coordinate", often called lambda (λ), typically couples the potential functions describing the two end states in a linear fashion:

$$V = \lambda * Va + (1 - \lambda) * Vb \quad (2)$$

where V is the simulated potential function and Va and Vb refer to the potentials appropriate for end states A and B, respectively. Both the aforementioned approach and traditional FEP or TI methods calculate the free energy difference between states A and B by integrating the free energy along λ. Tidor's method was successfully applied to calculate a free energy difference for two solvated ions via simulated annealing along the λ coordinate. The use of a continuous "reaction coordinate" means that this method has one of the limitations of traditional free energy calculations. Namely, much time is spent simulating nonphysical intermediate states rather than the end states of interest. This problem is compounded by allowing stochastic sampling along the reaction coordinate. Simulated annealing may be necessary since the simulation may get stuck in a free energy minimum that lies somewhere along the coordinate but is itself a poor representative of the end states.

While it would be possible to use Monte Carlo methods for both the chemical and Cartesian steps of the calculation, we were interested in eventually applying our method to studies of protein-ligand interactions. Consequently, we chose to use molecular dynamics methods instead of Monte Carlo methods for the sampling of configurational space. This rationale is partly historical, but also based on prior studies which showed MD was a better approach than MC for configurational sampling in proteins[47]. However, Jorgensen has recently made great strides in the application of MC techniques to proteins[48]. In contrast, Monte Carlo is a better configurational sampling tool in many

simpler systems, like solutions of small molecules[49]. For such systems, one could easily imagine using a chemical-MC/MC algorithm instead of our chemical-MC/MD approach.

To avoid the difficulties associated with "hybrid" or "in-between" states, we chose to restrict our chemical sampling to jumps between the end states of interest. In the formalism presented in the appendix,

$$\sum_{i=1}^{n} \lambda_i = 1 \text{ and } \lambda_i = \{0,1\} \tag{3}$$

This has the advantage that we are always simulating the end states of interest. However, the efficiency of the Monte Carlo sampling is now highly dependent on whether the simulation of state A samples configurations favorable for state B, or vice versa. The results of our simulations suggest that this is not an insurmountable problem, but its severity will be system dependent, an observation supported by Radmer and Kollman[14]'s work. In extreme cases, the barriers between states may be reduced by including a few carefully-chosen "hybrid" chemical states to bridge between the end points of interest, but we have not needed to take that approach for any of the calculations presented here.

A further advantage of our approach is that the extraction of relative free energies is very straightforward. The ratio of "populations" -- the number of times each chemical state is sampled in the calculations -- is directly related to the relative free energies of the chemical states by

$$\Delta G(A \rightarrow B) = -RT \ln (P(B)/P(A)) \tag{4}$$

This contrasts with the approaches that allow partial values of the reaction coordinate. In the two-state case, a simulation that has an average $\lambda = 0.5$ may not mean that the two end states are in equilibrium. Instead, it may mean the the free energy minimum of the potential is $\lambda = 0.5$. Similarly, the correct way to extract a free energy from these calculations is not

$$\Delta G(0 \rightarrow 1) = -RT \ln \lambda \qquad (5)$$

but rather, as Mezei, et al.[50] noted

$$\Delta G(0 \rightarrow 1) = -RT \ln (P(\lambda = 1)/P(\lambda = 0)) \qquad (6)$$

Since only those configurations where λ is fully representative of a single ligand contribute to the calculated free energy, it makes sense to avoid wasting time simulating intermediate states if that is feasible. If intermediate states are included in the calculation, however, one can calculate the potential of mean force or free energy integral along the reaction coordinate(s) as is done in a TI or FEP calculation.

Table 1: Relative free energies of solvation in chloroform calculated by thermodynamic integration

Ligand	ΔGsolv (CH_4 --> ligand) kcal/mol	note(s)
H_2CCH_2	-0.82 +/- 0.01	
CH_3F	-1.32 +/- 0.01	a
CH_2F_2	-1.44 +/- 0.1	
CHF_3	-1.31 +/- 0.01	
CF_4	-0.57 +/- 0.04	b
CH_3Cl	-2.42 +/- 0.01	
CH_2Cl_2	-3.31 +/- 0.25	
$CHCl_3$	-3.91 +/- 0.20	b

notes: a. value calculated by Reyes[34]; b. value calculated by Fox, et. al.[18]

Table 2: Calculated and reference small molecule ΔGsolv values. Reference values are from free energy perturbation calculations (no parentheses) or experiment (parentheses). Calculated values are from simulations using the chemical-MC/MD method +/- 1 standard deviation. Simulation times are the total amount used to calculate ΔGsolv. Times in brackets are the simulation times for the reference calculation.

System	reference ΔGsolv (kcal/mol)	calculated ΔGsolv (kcal/mol)	time (ps)
methane - methane	0	0.00	5
bromide - chloride	-3.22	-2.75	1000 [1000]
methane - ethane	0.15 +/- 0.07 (-0.17)	0.03	200 [1200]
anisole - benzene	0.90 (1.1 - 1.6)	0.99 +/- 0.48	1000 [~100]
methane, ethane and propane:			
methane - ethane	0.15 +/- 0.07	0.03 +/- 0.07	1600
ethane - propane	0.18 +/- 0.09 (0.12)	0.03 +/- 0.10	[2400]
methanol & substituted methanes:			
methanol - H_3CCN	-0.1 +/- 0.2 (1.2)	0.73	400 [~5000]
methanol - H_3CSH	4.6 +/- 0.1 (3.8)	2.95	
methanol - ethane	7.9 +/- 0.2 (6.9)	4.37	

Note for Table 2: All references are free energy perturbation calculations that include secondary references for the experimental values. Parameters for each system are taken from the references in question to facilitate direct comparison between the computational methods.

Table 3: Relative binding free energies vs. CH_4, 4-guest calculation

Guest	*ΔGhost (MC)*	*ΔGhost (TI)*	*ΔGsolv (TI)*	*ΔΔGbind (MC-TI)*	*ΔΔGbind (MC-offset)*	*ΔΔGbind (TI-TI)*	*ΔΔGbind (exp't)*
CH_3F	-0.77	-1.14	-1.32	+0.54	+0.17	+0.17	ND
CF_4	+0.20	+0.36	-0.57	+0.77	+0.41	+0.93	+2.8
$CHCl_3$	+2.50	+4.30	-3.91	+6.41	+3.57	+8.21	+5.2

All calculations were carried out using the parameters described in Fox, et al.[18]. Shaded columns show chemical Monte Carlo/MD calculations carried out in this work. Unshaded columns present experimental and thermodynamic integration data for comparison.

(MC): free energy from unbiased chemical-MC/MD calculation

(TI): Thermodynamic Integration data from Fox, et al.[18] and Reyes[34].

(MC-TI): ΔΔGbind calculated as ΔGhost from unbiased chemical-MC/MD calculation - ΔGsolv from TI

(MC-offset): ΔΔGbind calculated directly from a single chemical-MC/MD calculation using ΔGsolv from TI as a "solvation offset"

(TI-TI): ΔΔGbind calculated as ΔGhost from TI - ΔGsolv from TI

(exp't): Experimental binding data from Branda, et. al.[1]

References

1)Branda, N.; Wyler, R.; Rebek, J. *Science* **1994**, *263*, 1267-1268.

2)Kuntz, I. D. *Science* **1992**, *257*, 1078-1082.

3)Kuntz, I. D.; Meng, E. C.; Shoichet, B. K. *Account Chem Res* **1994**, *27*, 117-123.

4)Zwanzig, R. W. *The Journal of Chemical Physics* **1954**, *22*, 1420-1426.

5)Straatsma, T. P.; Berendsen, H. J. C. *J. Chem. Phys.* **1988**, *89*, 5876-5886.

6)Lybrand, T. P.; McCammon, J. A.; Wipff, G. *Proc Natl Acad Sci U S A* **1986**, *83*, 833-5.

7)Brooks, C. L.; Fleischman, S. H. *J Amer Chem Soc* **1990**, *112*, 3307-3312.

8)Kollman, P. *Chemical Reviews,* **1993**, *93*, 2395-2417.

9)Rao, B. G.; Kim, E. E.; Murcko, M. A. *J Comput Aid Molec Design* **1996**, *10*, 23-30.

10)Kong, X. J.; Brooks, C. L. *J Chem Phys* **1996**, *105*, 2414-2423.

11)Nose, S. *J. Chem. Phys.* **1984**, *81*, 511-519.

12)Ji, J., Cagin, T., and Pettit, B.M. *J. Chem. Phys.* **1992**, *96*, 1333-1342.

13)Liu, H. Y.; Mark, A. E.; Vangunsteren, W. F. *J Phys Chem* **1996**, *100*, 9485-9494.

14)Radmer, R. J.; Kollman, P. A. *J Comput Chem* **1997**, *18*, 902-919.

15)Metropolis, N. R., A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087-1092.

16)Bennett, C. H. *Journal of Computational Physics* **1976**, *22*, 245-268.

17)Tidor, B. *J Phys Chem* **1993**, *97*, 1069-1073.

18)Fox, T.; Thomas, B. E.; McCarrick, M.; Kollman, P. A. *J Phys Chem* **1996**, *100*, 10779-10783.

19)Senderowitz, H., and Still, W.C. *Journal of Physical Chemistry B* **1997**, *101*, 1409-1412.

20)Duane, S. K., A. D.; and Pendleton, B. J. *Physics Letters B* **1987**, *195*, 216-222.

21)Senderowitz, H.; Guarnieri, F.; Still, W. C. *J. Amer. Chem. Soc.* **1995**, *117*, 8211-8291.

22)Torrie, G. M. and Valleau, J.P. *Journal of Computational Physics* **1977**, *23*, 187-199.

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

24)Pearlman, D. *Journal of Physical Chemistry* **1994**, *98*, 1487-1493.

25)Anderson, H. C. *J. Chem. Phys.* **1980**, *52*, 2384-2393.

26)Berendsen, H. J. C.; Potsma, J. P. M.; van Gunsteren, W. F.; DiNola, A. D.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684-3690.

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

28)Kuyper, L. F.; Hunter, R. N.; Ashton, D.; Merz, K. M.; Kollman, P. A. *J Phys Chem* **1991**, *95*, 6661-6666.

29)Sun, Y. and Kollman, P. *J. Comp. Chem.* **1995**, *16*, 1164-1169.

30)Carlson, H. A.; Nguyen, T. B.; Orozco, M.; Jorgensen, W. L. *J Comput Chem* **1993**, *14*, 1240-1249.

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

32)Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, T. E., DeBolt, S., Ferguson, D., Seibel, G., Kollman, P. A. *Comp. Phys. Com.* **1995**, *91*, 1-41.

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

34)Reyes, C. **1997**.

35)Fox, T. and Kollman, P. *J. Phys. Chem., submitted* **1997**.

36)Bayly, C. I. *J. Phys. Chem.* **1993**, *97*, 10260-10280.

37)Allen, M. P., Tildesley, D. J. *Computer Simulations of Liquids.*; Oxford University Press: Oxford, 1987.

38)Wang, L. and Hermans, J. *Journal of the American Chemical Society* **1997**, *119*, 2707-2714.

39)Guo, Z., Kong, X., and Brooks, C.L. *Journal of Physical Chemistry* **submitted**.

40)Jarque, C. and Tidor, B. *Journal of Physical Chemistry B* **1997**, *101*, 9362-9374.

41)Merz, K. M., Jr.; Murcko, M. A.; Kollman, P. A. *J. Am. Chem. Soc.* **1991**, *113*, 4484-4490.

42)Daura, X., Hunenberger, P.H., Mark, A.E., Querol, E., Aviles, F.X. and Vangunsteren, W.F. *Journal of the American Chemical Society* **1996**, *118*, 6285-6294.

43)Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Amer. Chem. Soc.* **1990**, *112*, 6127-6128.

44)Wang, L., Eriksson, M., Pitera, J., and Kollman, P. *New Free Energy Calculation Methods for Structure-based Drug Design and Prediction of Protein Stability*; Wang, L., Eriksson, M., Pitera, J., and Kollman, P., Ed.; in press, 1998.

45)Aqvist, J., Warshel, A. *J. Am. Chem. Soc.* **1990**, 2860-2868.

46)DeBolt, S. E.; Pearlman, D. A.; Kollman, P. A. *J. Comp. Chem.* **1994**, *15*, 351-373.

47)McCammon, J. A.; Harvey, S. C. *Dynamics of proteins and nucleic acids*; Cambridge University Press: Cambridge, 1987.

48)Jones-Hertzog, D. K. and Jorgensen, William L. *Journal of Medicinal Chemistry* **1997**, *40*, 1539-1549.

49)Jorgensen, W. L., Tirado-Rives, J. *J. Phys. Chem.* **1996**, *100*, 14508-14513.

50)Mezei, M., Mehrotra, P.M. and Beveridge, D.L. *Journal of the American Chemical Society* **1985**, *107*, 2239-2245.

Figure Captions

Figure 1: Thermodynamic cycle for calculating the relative free energies (ΔΔGbind) for two ligands binding to a common receptor.

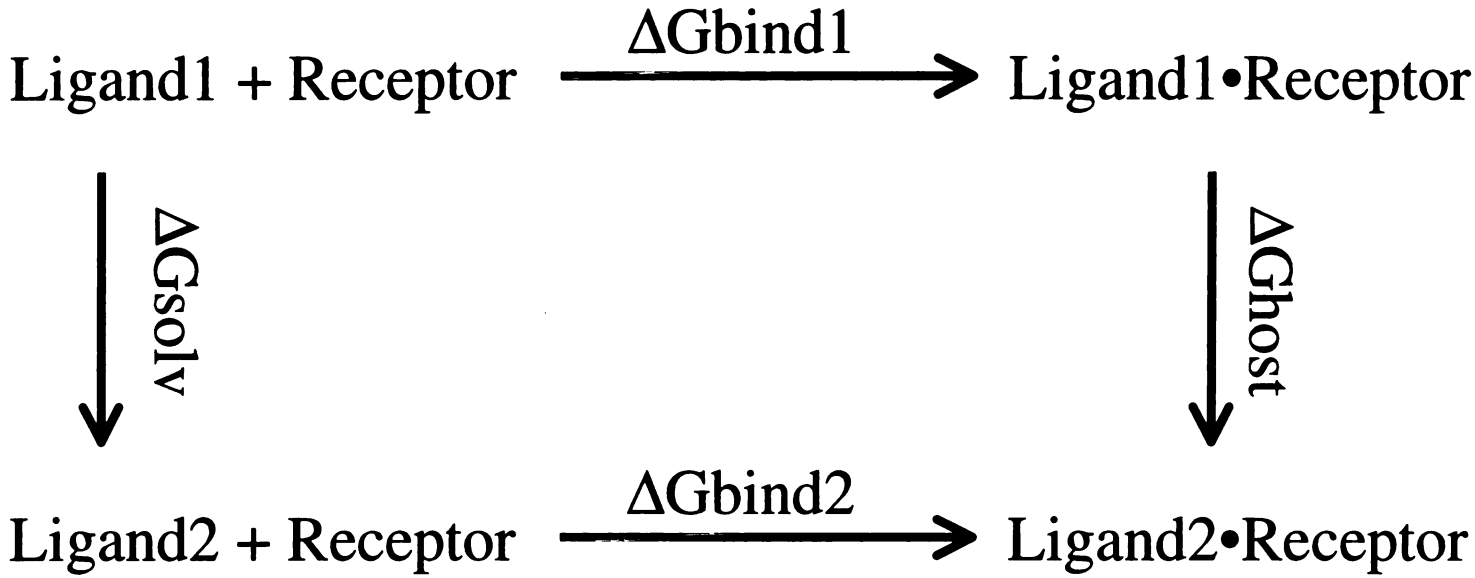
Figure 2: Population (**a**) and ΔGhost (**b**) data for the unbiased 4-guest calculation. (**a**) shows the relative populations of each ligand in the simulation. (**b**) shows these population data converted to ΔGhost free energies relative to CH_4. Solid circles show data for CH_3F; Solid squares are data for CF_4; solid diamonds are the data for $CHCl_3$; and CH_4 is shown in part (**a**) as the heavy line. The calculation is dominated by CH_3F, the guest with the most favorable ΔGhost.

Figure 3: Population (**a**) and ΔΔGbind (**b**) data for the "solvation offset" 4-guest calculation. (**a**) shows the relative populations of each ligand in the simulation. (**b**) shows these population data converted to ΔΔGbind free energies relative to CH_4. Solid circles show data for CH_3F; Solid squares are data for CF_4; solid diamonds are the data for $CHCl_3$; and CH_4 is shown in part (**a**) as the heavy line. By including ΔGsolv as a "solvation offset" to the Monte Carlo sampling, the calculation is now dominated by CH_4, the guest with the most favorable ΔΔGbind.

Figure 4: Population (**a**) and ΔΔGbind (**b**) data for the "solvation offset" 9-guest calculation. (**a**) shows the relative populations of each ligand in the simulation. (**b**) shows these population data converted to ΔΔGbind free energies relative to CH_4. Again, solid circles show data for CH_3F; Solid squares are data for CF_4; solid diamonds are the

data for $CHCl_3$; and CH_4 is shown in part (**a**) as the heavy line. In addition, data for H_2CCH_2 are indicated with open triangles pointing up, data for CH_2F_2 with stars, CHF_3 with solid triangles pointing down, CH_3Cl with plus signs and CH_2Cl_2 with X marks. These ΔΔGbind data are clearly not converged, but the calculation readily and rapidly determines the best (CH_2F_2) and worst ($CHCl_3$) guests for this host.

Figure 5: Single molecular dynamics snapshot of the "tennis ball" binding CH_2F_2 from the 9-guest calculation. The two halves of the tennis ball are drawn in black, and CH_2F_2 is shown in grey with both fluorines colored black. Each fluorine fits neatly into one of the major gaps between host monomers, with little strain of the host or guest molecules. Chloroform solvent molecules have been omitted for clarity.



$$\Delta\Delta Gbind = \Delta Gbind2 - \Delta Gbind1 = \Delta Ghost - \Delta Gsolv$$

Figure 1: Thermodynamic Cycle

Figure 2a,b

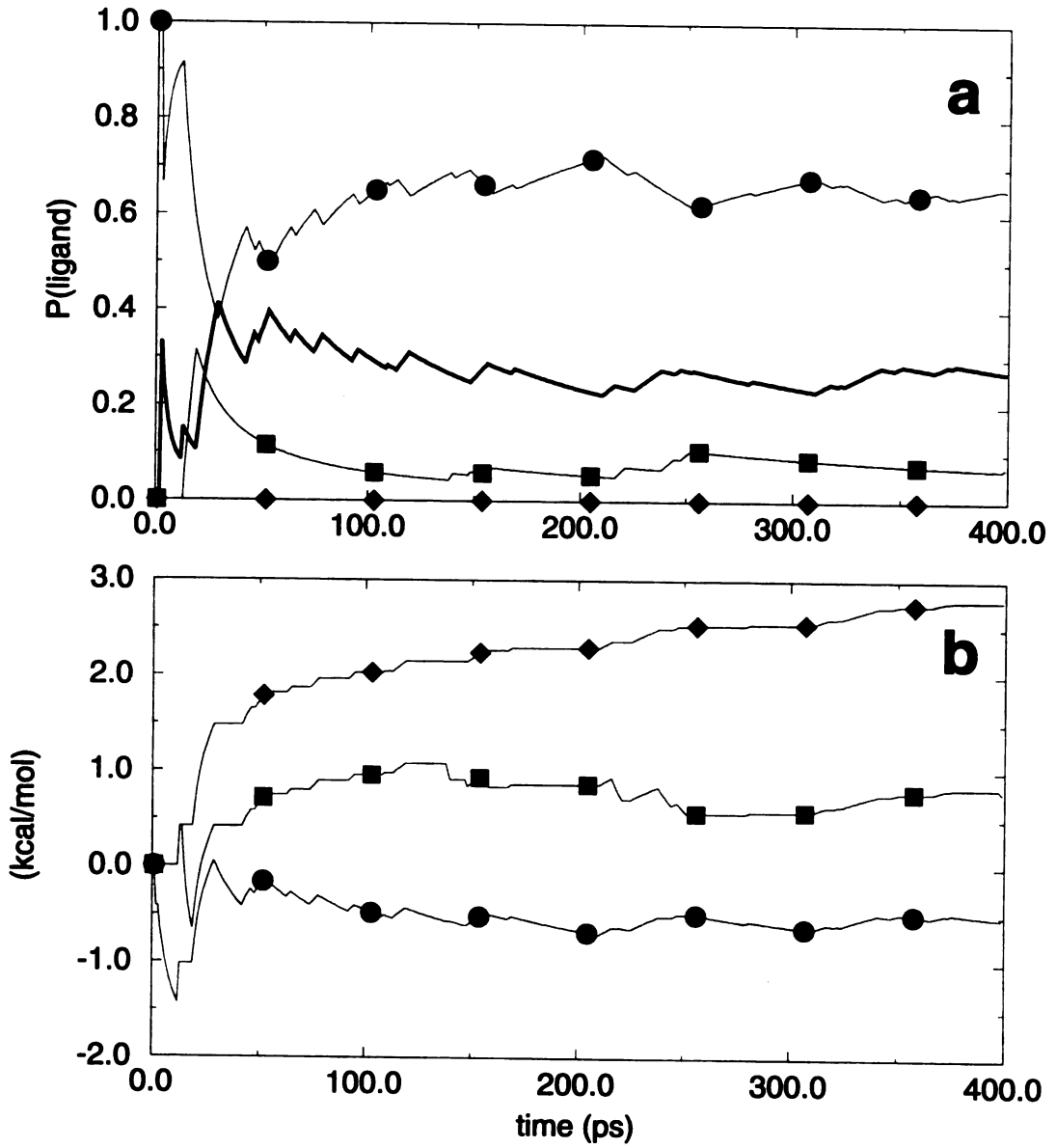

a
b
P(ligand)
(kcal/mol)
time (ps)
1.0
0.8
0.6
0.4
0.2
0.0
3.0
2.0
1.0
0.0
-1.0
-2.0
0.0
100.0
200.0
300.0
400.0

Figure 3a,b

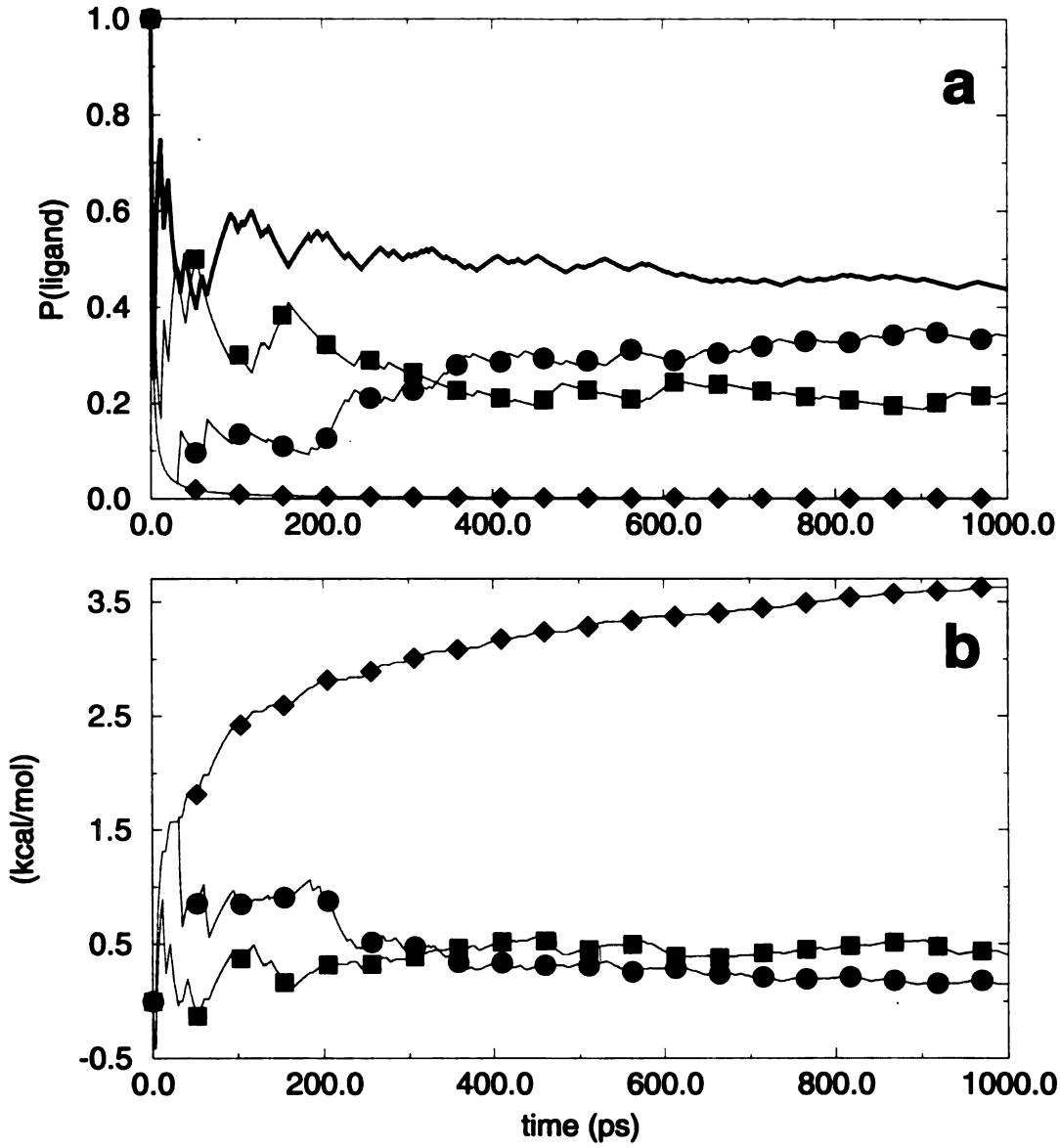
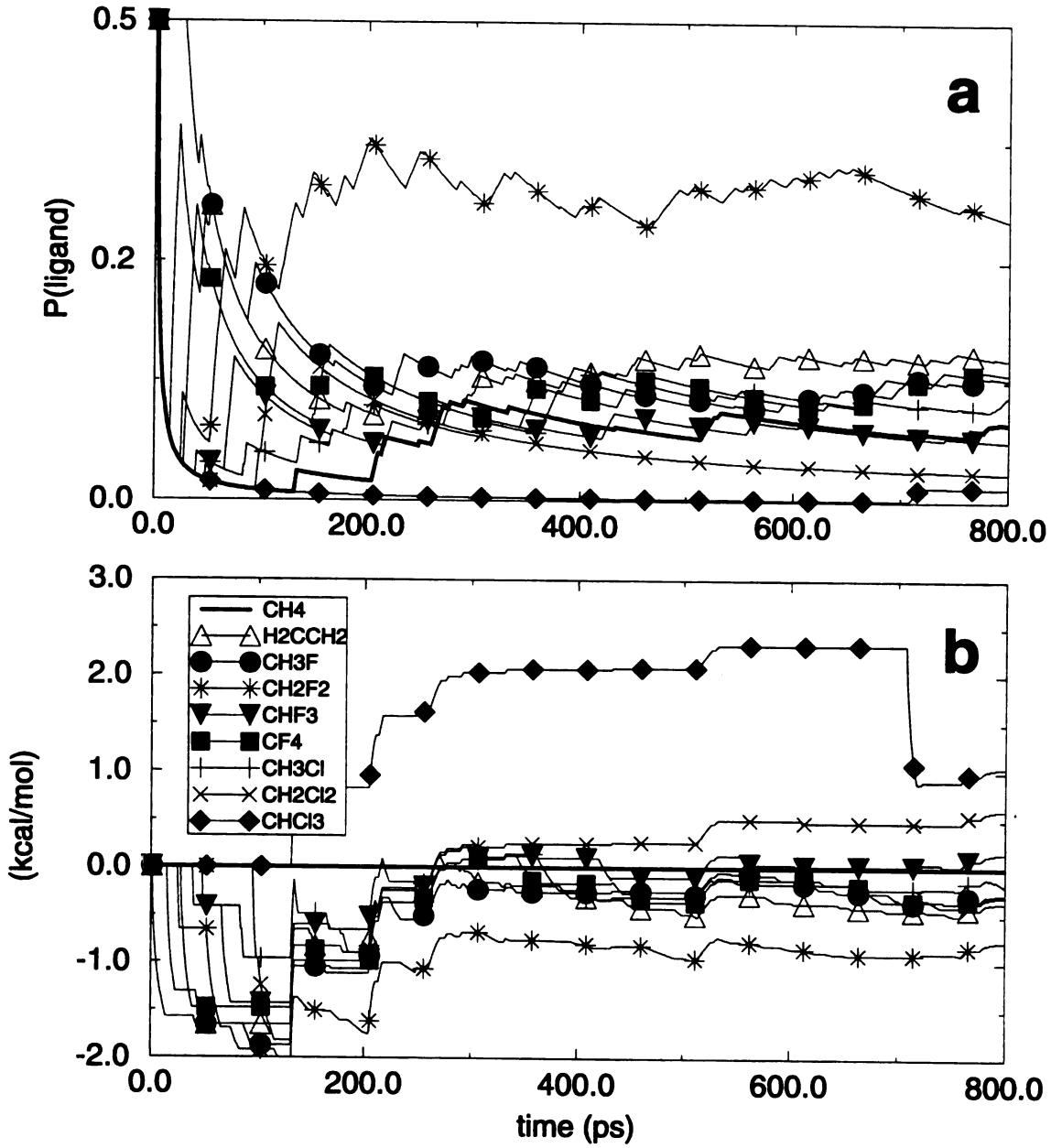
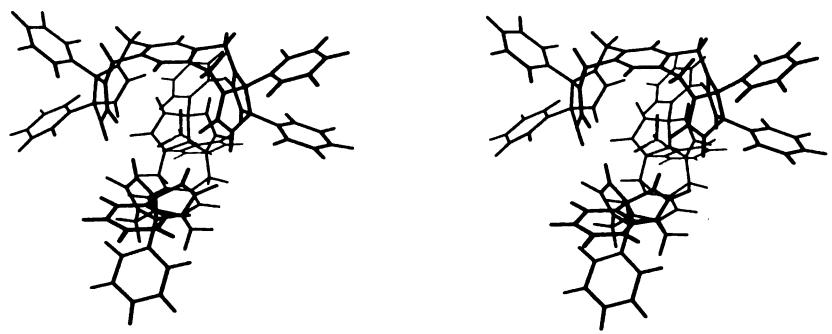

a
b
P(ligand)
(kcal/mol)
time (ps)
1.0
0.8
0.6
0.4
0.2
0.0
3.5
2.5
1.5
0.5
-0.5
0.0
200.0
400.0
600.0
800.0
1000.0

Figure 4a,b


a
P(ligand)
0.5
0.2
0.0
0.0
200.0
400.0
600.0
800.0
b
(kcal/mol)
3.0
2.0
1.0
0.0
-1.0
-2.0
CH4
H2CCH2
CH3F
CH2F2
CHF3
CF4
CH3Cl
CH2Cl2
CHCl3
time (ps)



UCSF MidasPlus

Figure 5

Chapter 4: Theoretical and practical considerations in Chemical Monte Carlo/Molecular Dynamics

While the prior Chapter outlined the general ideas and underlying formalism of CMC/MD, some details of the basic CMC/MD method deserve further explanation. First, it is important to clearly describe the system simulated in CMC/MD, including the masking of the potential function. This includes the precise nature of each chemical state or "endpoint." Second, the energy terms included in the Monte Carlo comparison are detailed, for cases when intra-MC residue energies are either ignored or included. Finally, some formal issues regarding the correctness of the hybrid Monte Carlo/molecular dynamics procedure are discussed.

First, consider the CMC/MD system used for calculation of a solvation free energy between two monoatomic solutes A and B. The simulated CMC/MD system is logically divided into two groups. One is the set of MC residues – in this example, solutes A and B. The remainder of the system (in this case the solvent) we call the "surroundings" or "surrounding residues". This division is shown in Figure 1a. At any one time, one of the MC residues is treated as "real". This residue, plus the "surroundings", forms the complete "real" system. Only the "real" elements of the system are used to determine the total system energy, and the molecular virial (for pressure coupling). The other Monte Carlo residue is a noninteracting "ghost". The two "real" systems possible in this case are shown in Figure 1b and 1c. Chemical Monte Carlo moves switch between these two systems. A microscopic "state" of this system is specified by the positions of all the particles (r) their velocities (v) and a variable (λ)

that specifies which MC residue is "real". The molecular dynamics steps serve to sample different coordinates (r') and velocities (v'), while the Monte Carlo steps, as previously noted, only alter the lambda variable.

The Monte Carlo steps in CMC/MD are sampled using a Metropolis Monte Carlo algorithm [1]. This algorithm generates a Markov chain of states that are populated in proportion to their relative free energies. If state I is the current state and J is the trial move, the trial move is accepted with a probability

$$P(I->J) = \min(1,\exp[-(H(J) - H(I))/kT])$$

Where H(I) is the total (potential plus kinetic) energy of state I:

$$H(I) = E(I) + K(I)$$

Since our chemical Monte Carlo moves only alter the lambda variables, and thus the potential energy of the system (r(I) = r(J), v(I) = v(J), lambda(I) ≠ lambda(J)), the acceptance criterion is simplified:

$$H(J) - H(I) = E(J) + K(J) - (E(I) + K(I)) = E(J) - E(I) = \Delta E$$

$$P(I->J) = \min(1,\exp[-\Delta E/kT])$$

A further simplification derives from our use of the AMBER molecular mechanics potential function[2]. AMBER, like many other molecular mechanics potentials, is built on an additive approximation. That is, the energy for a system of three atoms A, X, and Y is the sum of simpler interactions:

$$E(A,X,Y) = E(A) + E(X) + E(Y) + E(A,X) + E(A,Y) + E(X,Y)$$

It is useful to construct another idealized three-atom system B, X, Y. This system has energy

$$E(B,X,Y) = E(B) + E(X) + E(Y) + E(B,X) + E(B,Y) + E(X,Y)$$

And the energy difference between these two systems is

$$\Delta E = E(B,X,Y) - E(A,X,Y) = E(B) - E(A) + [E(A,X) + E(A,Y)] - [E(B,X) + E(B,Y)]$$

Our two constructed systems correspond directly to the two chemical states in Figure 1a, or the two real systems (Figure 1b, 1c). The energy difference derived above shows that it is not necessary to evaluate the total potential energy of both systems when considering Monte Carlo moves between them. Instead, it is only necessary to recalculate the contributions associated with the difference between solutes A and B. Our CMC/MD implementation makes extensive use of this fact. It is particularly useful since most of the systems of interest consist of a small MC region (less than 50 residues, typically with less than 25 atoms each) embedded in a large surrounding region (at least 200 water molecules (600 particles) and typically closer to 2000-10000 particles). The additional

overhead associated with calculating the interaction energy for an additional MC residue is minimal compared to the total cost of evaluating the potential energy for the entire system. Consequently, we calculate interactions for each MC residue every time we evaluate the potential energy for the system. These interactions are calculated and added or masked from the total potential energy as appropriate, and terms corresponding to interactions of the surrounding atoms with a MC residue are collected and used for the Metropolis Monte Carlo routine. This allows us to apply the "ghost forces" mentioned in the previous Chapter, as well as readily calculate the "Boltzmann probabilities" for every MC state as a check on our sampling. Use of nonadditive molecular mechanics models or more advanced (semiempirical or ab initio) quantum mechanical models (also nonadditive) would require much more expensive full-system energy calculations for each end state.

There are two options for which energy terms are included in the MC evaluation. Both are correct, but correspond to the calculation of free energies between subtly different systems. In Chapters 3 and 5, we neglected the intra-MC energy terms in our energy evaluation (E(A) and E(B), above) basing our MC sampling on:

$$\Delta E = [E(A,X) + E(A,Y)] - [E(B,X) + E(B,Y)]$$

This is somewhat analogous to the neglect of intra-perturbed group contributions in traditional free energy calculations. It is actually formally correct to do this, but one must take care to describe the precise nature of the systems being compared. When intra-perturbed group terms are not included in the MC calculation, the free energy difference we calculate corresponds to the difference between two non-physical systems: one where

solute A and the surroundings are real, while solute B is a noninteracting ghost; and another where solute B and the surroundings are real, while A is a ghost. The additional noninteracting particles mean that our simulated systems are not identical to the experimental systems we attempt to compare to. In the case we are discussing, the noninteracting solute is effectively in the gas phase.

As with traditional free energy calculations, the neglect of intra-perturbed group contributions may be appropriate when studying simple partitioning processes, for which molecular strain is not a factor. This includes the partitioning of a family of related, rigid ligands from solvent into a protein binding site, as we show in the next Chapter. In these situations, the intra-perturbed contributions are expected to cancel when a full thermodynamic cycle is used.

The other possibility with the MC evaluation is to include the intra-MC residue energy terms.

This is necessary when intramolecular interactions and strain energies make significant contributions to the free energy of interest, as in the peptide solvation we study in Chapter 7. When this is the case, both intra- and inter-MC terms are included in the energy evaluation:

$$\Delta E = E(B) - E(A) + [E(A,X) + E(A,Y)] - [E(B,X) + E(B,Y)]$$

When we include all of the energy terms in the MC energy evaluation, the free energy we calculate corresponds to the free energy difference between the two "real" systems: one where solute A and the surroundings are real, and one where solute B and the

surroundings are real. The noninteracting residues are part of the chemical Monte Carlo book-keeping, but are not included in the end states we are comparing.

There are also some important issues we must consider with regard to the mix of molecular dynamics and Monte Carlo sampling used in CMC/MD. Classical Hybrid Monte Carlo schemes have been shown to sample from the correct ensemble so long as the molecular dynamics algorithm used is symplectic and reversible[3]. However, the dynamics trajectories in these techniques are simply used to generate trial moves for a Metropolis Monte Carlo algorithm – trial moves in coordinate space. As such, it is essential that the dynamics is perfectly reversible in order to preserve detailed balance. That is, state **r'** must be accessible from state **r**, and state **r** must be identically accessible from state **r'**. However, CMC/MD differs from these other hybrid techniques in that the Molecular Dynamics trajectory is just used to generate coordinates sampled from a Boltzmann distribution. The trial move consists of a change in λ, not in **r**, and all λ-states are equivalently accessible at a given **r**.

It may be helpful to think of CMC/MD as related to Andersen temperature coupling[4]. Andersen, or stochastic temperature coupling, generates a series of coordinates sampled from the canonical (N,V,T) ensemble. An Andersen trajectory consists of a number of sub-trajectories. Each sub-trajectory is sampled from the microcanonical (N,V,E) ensemble. The energies for these sub-trajectories are themselves sampled from a Maxwell-Boltzmann distribution at the specified temperature, T. This is achieved by occasionally randomly reassigning the kinetic energy of a few particles of the system, selecting it from that same distribution. The integrator used for the trajectories is not required to be reversible, though it must be symplectic and conserve

energy. In a similar way, CMC/MD generates something akin to a grand-canonical (μ, V, T) ensemble trajectory by generating a number of microcanonical sub-trajectories from the appropriate distribution. The simple Maxwell-Boltzmann distribution used for the kinetic energies in Andersen temperature coupling is replaced by a Boltzmann-weighted distribution over the chemical (λ) states. This correspondence is illustrated in Figure 2. The upper pair of panels (Fig. 2a) show an Andersen "trajectory" and the corresponding distribution of particle velocities for a 1-dimensional system, while the lower pair of panels (Fig. 2b) show a CMC/MD trajectory and the corresponding (free energy) distribution of λ-values.

Since CMC/MD was built by modifying existing software, we made use of the leap-frog molecular dynamics integrator[5,6] of SANDER, which is symplectic but not reversible. In the leap-frog integration scheme, velocities and coordinates are not available at precisely the same time. Instead, velocities lag a half-timestep behind the coordinates. This means that the leap-frog integrator is not strictly time-reversible, unless additional half-step velocity integrations are used to synchronize velocities and coordinates. Fortunately, the discontinuities introduced by the MC steps correspond to discontinuities in the forces, rather than the energy or velocities. Frequent use of stochastic (Andersen) temperature coupling in sync with the Monte Carlo steps ensures that this small discontinuity does not produce any systematic error. The CMC/MD process does not absolutely conserve energy, but it is not expected to – Metropolis Monte Carlo trajectories are not isoenergetic. Instead, they correspond to samples from an isothermal ensemble, as we noted above.

A very helpful general reference for both free energy calculations and the theory and practice of hybrid MC/MD simulation is "Understanding Molecular Simulation: From Algorithms to Applications", by Frenkel and Smit (Academic Press 1996).

Bibliography

1)Metropolis, N. R., A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087-1092.

2)Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J Amer Chem Soc* **1996**, *118*, 2309-2309.

3)Duane, S. K., A. D.; and Pendleton, B. J. *Physics Letters B* **1987**, *195*, 216-222.

4)Andersen, H. C. *J. Chem. Phys.* **1980**, *52*, 2384-2393.

5)Hockney, R. W. *Methods Comput. Phys.* **1970**, *9*, 136-211.

6)Hockney, R. W. E., J.W. *Computer Simulation Using Particles*; McGraw-Hill: New York, 1981.

Figure 1

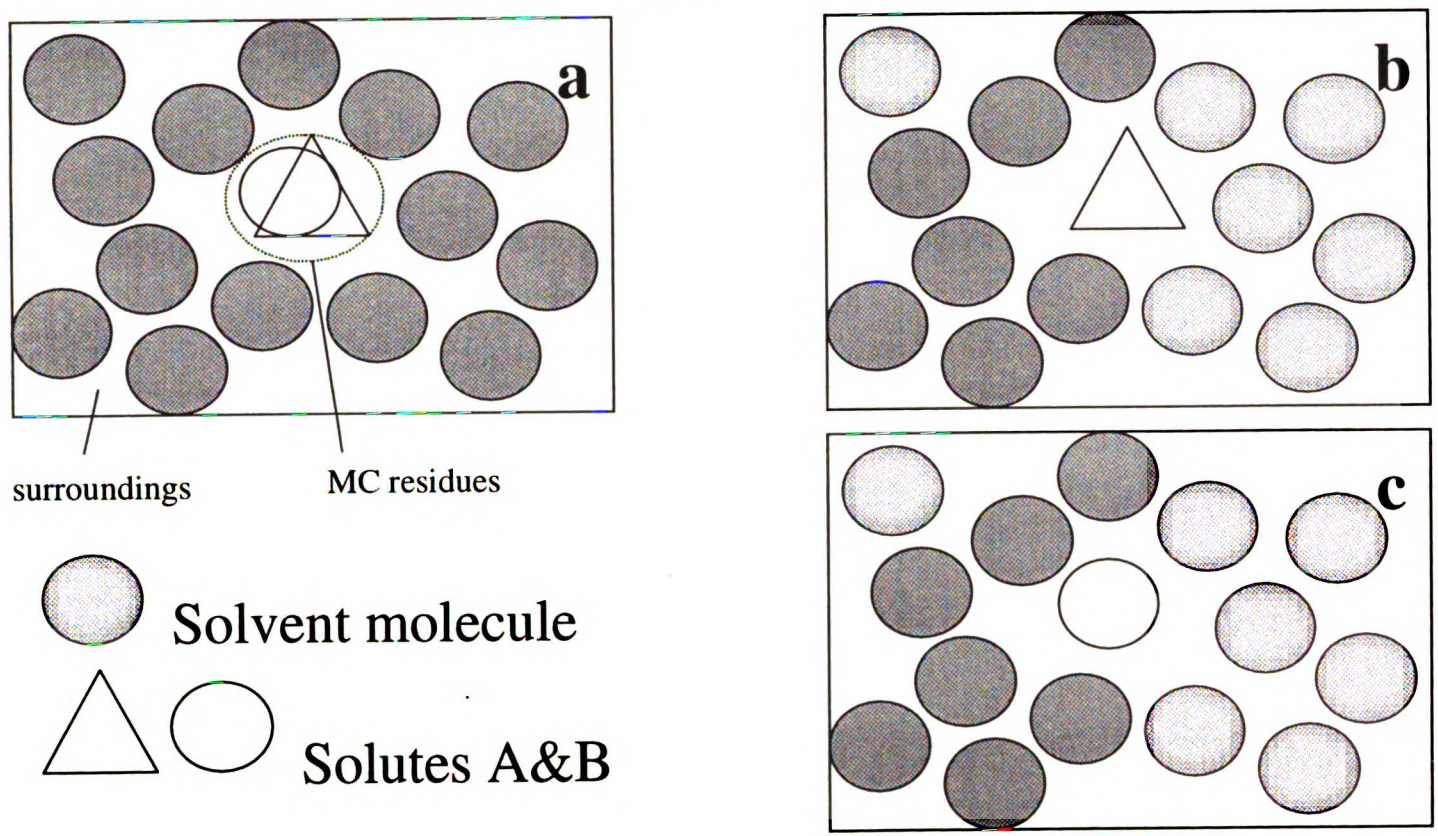
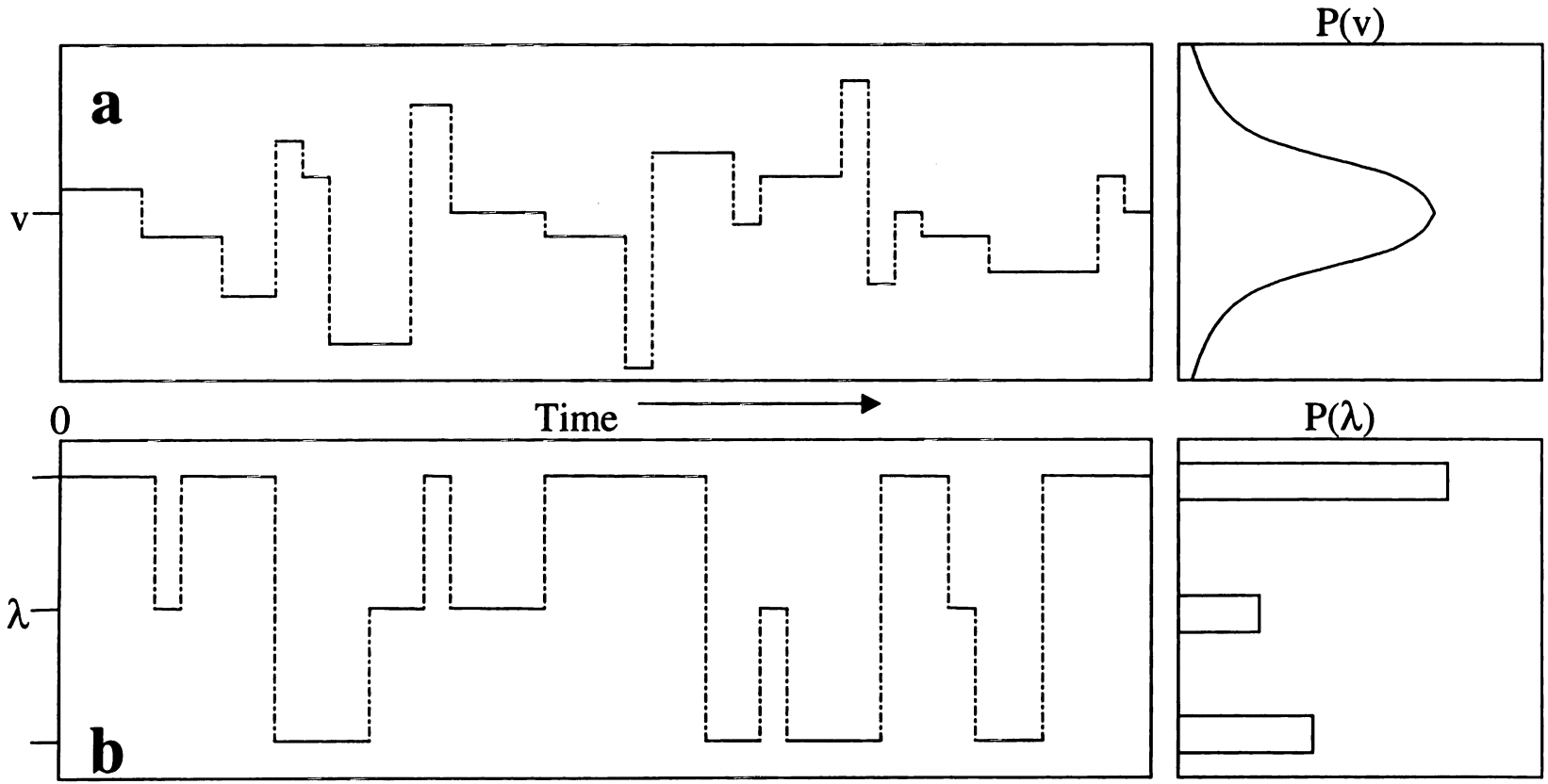


Figure 2



Chapter 5: Prediction of the binding free energies of new TIBO-like HIV-1 reverse transcriptase inhibitors using a combination of PROFEC, PB/SA, CMC/MD and free energy calculations.

Mats A. L. Eriksson[1,2], Jed Pitera[3] and Peter A. Kollman[1,4]

[1]Department of Pharmaceutical Chemistry

[3]Graduate Group in Biophysics

University of California at San Francisco

San Francisco, California 94143-0446

[2]Present address: Department of Biochemistry

University of Stockholm, SE-106 91 Stockholm, Sweden

[4]author to whom correspondence should be addressed

tel. (415) 476-4637; fax (415) 476-0688; e-mail: pak@cgl.ucsf.edu

Previously published in the Journal of Medicinal Chemistry, Volume 42, Number 5, Pages 868-881 (1999)

ABSTRACT

We have ranked 13 different TIBO derivatives with respect to their relative free energies of binding using two approximate computational methods - adaptive chemical Monte Carlo/molecular dynamics (CMC/MD) and Poisson-Boltzmann/Solvent Accessibility (PB/SA) calculations. Eight of these derivatives have experimentally determined binding affinities. The remaining new derivatives were constructed based on contour maps around R86183 (8Cl-TIBO), generated with a program, PROFEC (pictorial representation of free energy changes). The rank order among the derivatives with known binding affinity was in good agreement with experimental results for both methods, with average errors in the binding free energies of 1.0 kcal/mole for CMC/MD and 1.3 kcal/mole for the PB/SA method. With both methods, we found that one of the new derivatives was predicted to bind 1-2 kcal/mole better than R86183, which is the hitherto most tightly binding derivative. This result was subsequently supported by the most rigorous free energy computational methods - free energy perturbation (FEP) and thermodynamic integration (TI). The strategy we have used here should be generally useful in structure-based drug optimization. An initial ligand is derivatized based on PROFEC suggestions, and the derivatives are ranked with CMC/MD and PB/SA to identify promising compounds. Since these two methods rely on different sets of approximations, they serve as a good complement to each other. Predictions of the improved affinity can be reinforced with FEP or TI, and the best compounds synthesized and tested. Such a computational strategy would allow many different derivatives to be tested in a reasonable time, focusing synthetic efforts on the most promising modifications.

Introduction

Inhibitors to HIV-1 RT are one of the cornerstones in the treatment of AIDS patients, preventing the progression of HIV infection. The enzyme is an attractive target for drug therapy not only because it is essential for HIV replication, but also since it is not required for normal host cell replication. HIV-1 RT is a multifunctional enzyme that copies the RNA genome of HIV-1 into DNA which is subsequently integrated in the host cell. The enzyme is a heterodimer composed of the two subunits p66 and p51 and its unliganded structure has been determined at 2.35 Å resolution.[1] The active site (or the dNTP site), which contains the catalytically essential amino acids (primarily a triad of aspartic acids) is located in the p66 palm subdomain with the 3'-OH of the primer terminus near the active site. The types of inhibitors currently discovered can be divided into two classes; nucleoside inhibitors [NIs, for example, AZT, ddI and ddC (for reviews, see refs. [2-5])] and non-nucleoside inhibitors (NNIs, for example TIBO, HEPT, α-APA and nevirapine, reviewed in refs. [4, 6-8]). The NIs cause termination of the growing DNA chain because elongation is blocked due to the lacking 3'-OH functional group, which is essential for incorporation of additional nucleosides (reviewed in ref. [3]). However, the NIs can also be incorporated into cellular DNA by the host DNA polymerases and therefore cause serious side effects. Unlike the NIs, NNIs are HIV-1 RT specific and do not inhibit host cell polymerases. The binding site of the NNIs is located in the p66 palm subdomain near, but distinct from, the dNTP-binding site. Recently, the two similar crystal structures of HIV-1 RT in complex with the TIBO NNIs R86183 (8-Cl TIBO)[9] and R82913 (9-Cl TIBO)[10] have been solved at 3.0 Å resolution. The NNI binding

pocket constitutes mainly of hydrophobic and aromatic residues (green residues in fig. 1). Comparisons of structures of HIV-1 RT, complexed with different NNIs[1] reveal that there is a significant rearrangement of a three-stranded β-sheet in the p66 subunit (containing the catalytic triad of aspartic acids), with respect to the rest of the polymerase site. This suggests that NNIs inhibit HIV-1 RT by locking the polymerase active site in an inactive conformation, similar to the conformation observed in the inactive p51 subunit[1]. In addition, the NNIs have low cytotoxcity and produce a few side effects.[11] A serious problem with the NNI HIV-1-RT inhibitors is the emergence of viral strains that have point mutations in the region encoding HIV-1 RT which prevent these drugs from inhibiting RT.

There is a considerable interest in developing computational methods that are sufficiently efficient to allow ranking of several (10 -100) inhibitors with respect to their binding free energy to a common receptor. This stems from the fact that the most rigorous computational methods — free energy perturbation (FEP) and thermodynamic integration (TI) calculations (for reviews, see for example, refs. [12, 13]) — are both too slow for practical use in drug optimization. These two methods typically give good (< 1 kcal/mole[14]) estimates of the relative binding energies. However, only one pair of inhibitors can be compared in a single FEP/TI run, which may require from days to weeks to complete due to their computationally intensive nature. Since lead optimization requires the comparison of *many* possible modifications to the lead compound, there is a need for more rapid methods. One such method, denoted chemical Monte Carlo/molecular dynamics (CMC/MD), has recently been developed by Pitera and Kollman.[15] CMC/MD combines the Monte Carlo method for sampling the chemical space of a system and the MD method for generating a set of coordinates for a distinct chemical system. The chemical space can be typically 5-10 different derivatives of an

inhibitor. During the course of a CMC/MD run the probabilities of each derivative are generated, which can be related to their relative binding free energies. The CMC/MD method has successfully been applied to estimate relative solvation free energies for small organic molecules and to study the strength of small ligands binding to an organic host.[15] In CMC/MD the solvent is typically described with explicit water molecules, which is a computational bottleneck since a great portion of computer time is spent calculating forces on the solvent molecules. The problem is circumvented in the Poisson-Bolzmann/Solvent Accessibility (PB/SA) method,[16, 17] where the solvent is treated implicitly as a dielectric continuum. The protein and inhibitor are modeled as low-dielectric cavities containing fixed partial charges. By solving the Poisson-Boltzmann equation the electrostatic contribution to the binding free energy is calculated. The non-polar contribution to the binding free energy is estimated, assuming an empirical linear dependence on the solvent accessibility areas.[18, 19] This method has been applied to a number of protein-ligand complexes,[20-25] for estimation of absolute and relative ligand binding free energies. A third method is the linear interaction energy (LIE) method, developed by Åqvist.[26] In this method, the binding free energy is approximated to be linearly dependent on the ligands interaction energies in the protein or in solution. The LIE method (sometimes with slight modifications) has also been applied to a variety of ligand-protein complexes,[26-30] where absolute as well as relative binding free energies have been estimated.

In order to generate suggestions for modifications on a lead inhibitor that would improve its binding, Radmer & Kollman have developed PROFEC[31] (pictorial representation of free energy changes). This approach uses MD trajectories to estimate the average cost of adding a particle around the inhibitor in the protein and in the solution, respectively. The difference cost can then be visualized as contour maps around the inhibitor. Favorable (negative) regions of the contour maps indicate positions where modifications to the inhibitor should improve its binding free energy.

In this study, we have used the PROFEC method to suggest modifications on R86183, which is the so far the tightest binding inhibitor (see Table I). Five new derivatives was constructed with the help of the contour maps from PROFEC. These new inhibitors were ranked together with 8 derivatives with experimentally[32] known binding affinities, using both the CMC/MD and the PB/SA methods. The purposes of the study are 1/ to test these relatively new methods (especially CMC/MD) against experimental results and against each other; 2/ to develop a feasible computational strategy for structure-based lead optimization; and 3/ To generate a better binding TIBO derivative than the previously known. We show that both methods work surprisingly well, given the approximations involved, and rank the inhibitors in good agreement with the experiment. Since the two methods are based on different sets of approximations they serve as good complements to each other and a consistent result between them increases its validity. Both methods predict that one of the new PROFEC derivatives should bind HIV-1 RT about 1-2 kcal/mole stronger than R86183. Subsequently, we performed "full" free energy calculations (FEP and TI) on this best binding inhibitor. The full free energy calculations also suggest that this new inhibitor binds better than R86183. The strategy we have employed herein (outlined in fig. 2) could generally be used as one of the final stages in a structure-based lead-optimization using computational methods.

Computational methods

Force field parameters for the TIBO derivatives. The van der Waals (vdW), bond, angles, dihedrals and improper dihedrals parameters for the sulfur atom were adopted from a parameterization of thiobiotin[33] and vdW parameters of the chlorine atom were taken from parameters used for chloroform.[34] To estimate the partial atomic charges of 8Cl-TIBO (R86183, see Table I), 9Cl-TIBO (R82913) and unchlorinated TIBO (R82150), we used both the conformation of 8Cl-TIBO in complex with HIV-1 RT[9] as well as the A-form of the crystal structure of 9Cl-TIBO (R82913).[35] The two respective conformers were geometry optimized using Gaussian94[36] at the STO-3G level. The electrostatic potential around the TIBO derivatives was then calculated with the 6-31G* basis set and atomic partial charges were fitted to the electrostatic potentials around the two structures using the RESP method.[37] Since the partial charges evaluated from the two conformers individually were very similar, we evaluated the partial charges of the remaining TIBO derivatives (Table I) using only the conformation of 8Cl-TIBO in HIV-1 RT.

Setup and equilibration of HIV-1 RT in complex with 8Cl-TIBO. The simulations were carried out with the AMBER 4.1[38] program "Sander" using the Cornell *et al.* force field.[39] Starting with the 3.0 Å resolution crystal structure of 8Cl-TIBO in HIV-1 RT[9] we added unresolved residues, modeled as alanines in the crystal structure, as well as hydrogen atoms. The hydrogen atoms were then minimized for 200 steps (steepest descent) *in vacuo*. To let the protein relax in an aqueous environment the complex was

immersed in a 55 Å radius sphere of TIP3P-water.[40] The solvent sphere together with the protein-inhibitor complex were minimized with a gradual decrease in the position restraints of the protein atoms. Thereafter, all water molecules beyond the first hydration shell (i.e. at a distance > 3.5 Å from any protein atom) were removed and to achieve electroneutrality 11 chloride ions were added, using the program module "CION" within Amber 4.1. Protein residues with any atom closer than 12 Å from 8Cl-TIBO were chosen to be flexible in the simulations. All protein residues, water molecules and counterions further away than 15 Å from any flexible residue were then deleted, due to the size of HIV-1 RT. We then centered a 20 Å radius spherical cap of TIP3P-water around TIBO, including the hydrating water molecules within the sphere from the previous step. The water cap was equilibrated for 50 ps at 300 K, keeping the protein, 8Cl-TIBO and the hydrating water molecules outside the water cap rigid. Thereafter, the flexible residues (as defined above) and 8Cl-TIBO together with the cap of water molecules were then heated (50 ps) and equilibrated for 300 ps at 300 K. A time step of 2 fs was used, with the non-bonded list updated every 20 time step and all bonds were constrained with the SHAKE algorithm[41]. We applied a dual cutoff of 9 and 13 Å, respectively, where energies and forces due to interactions between 9 and 13 Å were updated every 20 time step. The temperature was maintained using the Berendsen method,[42] with separate couplings of the solute and solvent to the heat. This system was then run for 500 ps for a subsequent analysis with the PROFEC program (see below).

Setup and equilibration of 8Cl-TIBO in solution. For 8Cl-TIBO in solution we started with the A-form of the crystal structure of 9Cl-TIBO,[35] with a substitution of the

atoms at positions 8 and 9. 8Cl-TIBO was then immersed in a box of TIP3P water with dimensions 34 x 33 x 29 $Å^3$. The water molecules were equilibrated at constant pressure for 100 ps, keeping the inhibitor rigid. We then released the TIBO atoms and the system was equilibrated for 200 ps, using the same dual cutoff and time step as for 8Cl-TIBO in HIV-1 RT. Also here, we performed an additional 500 ps MD simulation for the PROFEC analysis.

Pictorial representation of free energy changes (PROFEC). The contour maps that are generated from the program PROFEC[31] can be used as guides to where atoms/groups should be added or deleted on the inhibitor in order to improve its binding free energy to a protein. The maps are generated from trajectories of two MD simulations - one of the protein-inhibitor complex and the other of the inhibitor in solution. The insertion free energy of a test particle (ΔG_{ins}) at various grid points close to the inhibitor is calculated according to:

$$\Delta G_{ins}(i,j,k) = -RT\ln\langle\exp(-\Delta V(i,j,k)/RT\rangle_0 \qquad (1)$$

where i, j and k are the coordinates of a grid point, $\Delta V(i,j,k)$ is the interaction energy between the test particle and the surrounding atoms, and $\langle ...\rangle_0$ is an average over the trajectories. To generate the coordinate system of the grid points, three atoms in the inhibitor determine a coordinate plane and the third axis is formed as a vector product of two axes in that plane. Since the coordinate system is molecule-fixed, corresponding grid points are comparable for the inhibitor in solution and in the protein, respectively. The

difference, $\Delta\Delta G_{ins}$, of particles in the inhibitor-protein complex and in the inhibitor in solution, respectively, is formed for each grid point and contour maps of $\Delta\Delta G_{ins}$ can be constructed and displayed. The electrostatic properties of the added test particles can also be estimated by calculating the derivative of $\Delta\Delta G_{ins}(i,j,k)$ with respect to charge at each grid point. This derivative is then displayed by coloring the contour map (at, for example, $\Delta\Delta G_{ins}$ =0) and might thus suggest how the charge distribution should be changed for an improved binding.

The PROFEC contour maps were calculated from the two 500 ps MD trajectories of 8Cl-TIBO in HIV-1 RT and in solution, respectively. In each PROFEC calculation we chose the grid size to be 7.5 Å with a grid spacing of 0.5 Å. We selected different atoms of 8Cl-TIBO in each calculation to obtain detailed contour maps centered on various regions of the ligand. Through a special delegate program written by R.J. Radmer (UCSF), the contour maps were visualized with UCSF MidasPlus.[43]

Chemical Monte Carlo (CMC)/molecular dynamics (MD). The CMC/MD method (described in detail in ref. [15]) has recently been developed for determination of relative free energies of a series of ligands binding to a common receptor. The method employs MD to generate a set of coordinates for one distinct chemical system and MC to sample the *chemical* space of the system, which can be 5-10 different derivatives of an inhibitor in a protein-inhibitor system. The derivatives are all present in the protein binding pocket during the simulation but they do not exert forces on each other. In addition, the protein only feels the presence of one ("real") inhibitor at a given time. An MC-step consists of

choosing an inhibitor 'i' at random and this inhibitor will be accepted as the new "real" ligand, according to the Metropolis[44] criteria:

$$\text{if } \Delta E_i \leq 0 \Rightarrow P_i=1, \qquad \text{if } \Delta E_i > 0 \Rightarrow P_i=\exp(-\Delta E_i/RT) \tag{2}$$

where ΔE_i is the difference in protein-inhibitor interaction energy between a derivative 'i' and the old derivative and P_i is the acceptance probability. We use the protein-ligand interaction energy instead of the total system energy in our Monte Carlo step since the only thing that changes in the MC move is which ligand is "real", i.e. interacting with the protein. During the course of the MC/MD run, the probability of each derivative being the "real" ligand 'P_i' is accumulated, resulting in a probability distribution that mirrors the relative free energies of the bound state of the derivatives. To better determine the relative free energies of unfavorable states, the "Boltzmann" probabilities of each inhibitor 'i' can also be calculated prior to each MC-step according to:

$$P_i= \exp(-\Delta E_i/RT)/\Sigma\exp(-\Delta E_i/RT) \tag{3}$$

If an infinite number of MC-steps were performed on a *single* Cartesian conformation, the resulting probability distribution $\{P_i\}$ would be exactly the same as that calculated with equation 3. We used the averaged P_I's from eq. 3 herein, since they also allow for estimations of the relative free energies of poorly sampled derivatives. The

resulting probability distribution is then related to the relative free energy of the bound state (ΔG_{bound}) for derivatives 'j' and 'i' is according to:

$$\Delta G_{bound,j}-\Delta G_{bound,i}=-RT\ln (P_j/P_i) \quad (4)$$

We found that CMC/MD, as outlined above, converged very slowly when applied to the TIBO derivatives in HIV-1 RT. In order of increase the convergence rate, a variant of this method - herein called the "adaptive CMC/MD" method (J. Pitera, unpublished) - was developed. The goal with adaptive CMC/MD is to sample the chemical space *evenly* instead of sampling this space according to the relative free energies of the derivatives. This can be achieved by introducing biasing offsets, $\Delta G_{offs,i}$, that for each inhibitor 'i' reflects its relative free energy in the bound state. These biasing offsets are introduced by umbrella sampling, as previously described.[15, 45] MC-sampling by testing the acceptance against (ΔE_i-$\Delta G_{offs,i}$), rather than ΔE_i as in eq. 2, would then result in an even sampling of all inhibitors, since all (ΔE_i-$\Delta G_{offs,i}$) would equal zero, on average. Starting with all $\Delta G_{offs,i}$=0, the offsets are solved for iteratively and the probabilities of each inhibitor are calculated according to eq. 3, averaged over a certain number of MC/MD-cycles (a CMC/MD-run). A first set of $\Delta G_{offs,I}$'s, relative to some arbitrarily chosen ligand, is estimated from eq. 4, and these offsets are used in the next CMC/MD-run. The offsets are then adjusted iteratively after each CMC/MD-run by averaging the P_I's from eq. 3 and using the $\Delta G_{bound,i}$ obtained from eq. 4 to adjust $\Delta G_{offs,i}$. Upon convergence, all P_I's are roughly equal and the relative free energies of the bound state ($\Delta G_{bound,i}$) are equal to -$\Delta G_{offs,i}$. Finally, the relative free energies of binding ($\Delta\Delta G_{bind}$) are calculated by

subtracting the solvation free energies (ΔG_{solv}) from ΔG_{bound}. This adaptive procedure is effectively the same as the WHAM procedure[46] for calculating conformational free energy differences. However, chemical-MC/MD allows us to use it for the calculation of chemical free energy differences.

The adaptive CMC/MD method was applied to 8 different TIBO derivatives with experimentally known[32] relative binding affinities (Table I) together with 5 new derivatives that were suggested by visualization of the PROFEC contour maps. The derivatives were created by substituting and/or deleting atoms of the HIV-1 RT conformer of 8Cl-TIBO and positioned in the equilibrated HIV-1 RT - 8Cl-TIBO complex (see above). The inhibitors were then allowed to relax in the binding pocket by individually minimizing them, keeping everything but the inhibitors rigid. In the subsequent adaptive CMC/MD calculations the MD time step was reduced to 1.5 fs due to problems with the SHAKE algorithm and one MC step was performed every 20 MD time steps. We applied the adaptive CMC/MD method for two sets of inhibitors, each containing 10 derivatives. The $\Delta G_{offs,i}$'s were iteratively adjusted every 500 MC steps for set 1. For set 2 we shortened that interval to every 125 MC steps. The values of $\Delta G_{offs,i}$ were graphed and monitored until they appeared converged by visual inspection. This required 450 ps (30 iterations) for set 1 and 560 ps (150 iterations) for set 2.

The solvation free energies of the TIBO derivatives were estimated from Generalized Born/Solvent Accessibility (GB/SA)[47] calculations, using the program MacroModel/BatchMin, version 4.5,[48] with our RESP derived charges on the derivatives. The derivatives were substitutions from the A-form of the 9Cl-TIBO crystal structure[35] that were minimized *in vacuo* prior to the calculations. While this approach does not

include the relative internal entropies of each compound in solution, we expect those contributions to be small among our family of highly similar and relatively rigid compounds.

Poisson-Boltzmann/Solvent Accessibility (PB/SA) calculations. In the PB/SA calculations, which were carried out with the latest Delphi package,[49, 50] the solvent is represented by a continuum with a dielectric constant ε=80, with or without implicit ions. In this work we added implicit ions to an ionic strength of 0.13 M. The protein and the TIBO derivatives are represented by a cavity, both with a dielectric constant ε=2, containing fixed partial charges on their atoms. The relative free energy of solvation ($\Delta\Delta G_{solv}$) for two TIBO derivatives L_1 and L_2 was estimated according to:

$$\Delta\Delta G_{solv} = \Delta G_{react}^{g\rightarrow aq}(L_2) - \Delta G_{react}^{g\rightarrow aq}(L_1) + \Delta\Delta G_{nonpol} \qquad (5)$$

where $\Delta G_{react}^{g\rightarrow aq}$ is the reaction field energy when transferring the derivative from vacuum (ε=1) to aqueous solution (ε=80). The non-polar contribution ($\Delta\Delta G_{nonpol}$) can be estimated according to the following empirical relation, which correlates the solvation free energies of non-polar solutes to their solvent accessible surface area:[18, 19]

$$\Delta\Delta G_{nonpol} = \sigma\Delta A \qquad (6)$$

where ΔA is the difference in solvent accessible area between L_1 and L_2. ΔA was calculated with Connolly's MS program[51] using van der Waals radii from the Cornell *et al.* force field.[39] σ is the empirical solvation parameter and we used a value of 5 cal mol^{-1} A^{-2} in this work. The same structures as for the GB/SA calculations (above) was used for estimations of ΔG_{solv}.

A corresponding calculation of the relative free energies in the bound state ($\Delta\Delta G_{bound}$) involves an estimation of the difference in solvation free energies between L_1P and L_2P. This is very difficult in practice since these energies are large numbers and subtracting them might result in large errors. Therefore, as a first approximation, we estimated the polar contribution to $\Delta\Delta G_{bound}$ simply as the difference in reaction field energy on the two ligands (L_1 and L_2) in the protein [$\Delta G^{g\to aq}_{react,L_2}(L_2P) - \Delta G^{g\to aq}_{react,L_1}(L_1P)$] plus the difference in ligand-protein electrostatic energy ($\Delta\Delta G_{lig\text{-}prot,elec}$). A more rigorous PB calculation would also include the intramolecular energies of the protein-ligand complexes and of the free ligands. However, the intramolecular energies of the proteins in the protein-ligand complexes are large numbers and might add large errors when subtracting them. We assume therefore to a first approximation that these energies cancel then forming the difference between the complexes. Moreover, the intramolecular energies of the free ligands are sensitive to their conformation and a correct inclusion of these energies would thus require multiple conformations of the ligands, generated, for example, by using MD simulations. This would make the PB method relatively slow and tedious and since we are interested in rapid, approximate methods that are able to rank the derivatives, we only use one conformation in solution and assume to a first approximation that the relative intramolecular energies of the ligands cancel in the protein and in solution, respectively. Finally, since the TIBO derivatives are completely buried in the hydrophobic binding pocket (i.e. their accessible surface areas are zero), the non-polar contribution can simply be estimated as the difference in ligand-protein van der Waals energy ($\Delta\Delta G_{lig\text{-}prot,\ vdw}$). The resulting simplified expression for $\Delta\Delta G_{bound}$ is thus:

$$\Delta\Delta G_{bound} = \Delta G^{g\to aq}_{react.L_2}(L_2P) - \Delta G^{g\to aq}_{react.L_1}(L_1P) + \Delta\Delta G_{lig-prot.elec} + \Delta\Delta G_{lig-prot.vdw} \quad (7)$$

The resulting relative free energies of binding ($\Delta\Delta G_{bind}$) are then estimated from the difference $\Delta\Delta G_{bound}$-$\Delta\Delta G_{solv}$. The 13 different TIBO - HIV-1 RT complexes were further minimized, now with flexible residues, water molecules and counterions as in the MD simulations (see above). Prior to the PB/SA calculations, all water molecules and counterions were removed.

Free energy calculations (TI and FEP). To estimate the relative binding free energies ($\Delta\Delta G_{bind}$) of two TIBO derivatives (L_1 and L_2) to HIV-1 RT (P) we made use of the following thermodynamic cycle:

$$\begin{array}{ccc} L_1 + P & \xrightarrow{\Delta G_1} & L_1P \\ \downarrow \Delta G_s & & \downarrow \Delta G_b \\ L_2 + P & \xrightarrow{\Delta G_2} & L_2P \end{array} \quad (8)$$

where ΔG_1 and ΔG_2 are the experimentally determined[32] binding affinities. Since G is a state function the following applies:

$$\Delta\Delta G_{bind} = \Delta G_2 - \Delta G_1 = \Delta G_b - \Delta G_s \quad (9)$$

where ΔG_b and ΔG_s are calculated with the TI or FEP methods (see below).

Except for one set of calculations (described below) we used the TI algorithm to estimate ΔG_b and ΔG_s (eq. 8). In this method a coupling parameter λ is introduced, which varies from λ=0 [for $L_1(P)$] to λ=1 [for $L_2(P)$]. The free energy change is then evaluated according to:

$$\Delta G = \int_0^1 \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda \qquad (10)$$

where $H(\lambda)$ is the potential energy of the system as a function of the coupling parameter λ and $< >_\lambda$ is an ensemble average at λ. The integral is evaluated numerically from a number of evenly spaced windows (spacing = $\Delta\lambda$) with λ values ranging from 0 to 1. $<\partial H(\lambda)/\partial\lambda>_\lambda$ is calculated by averaging over molecular dynamics trajectories run at a certain number of steps in each window. The calculations were run with the AMBER 4.1 program "GIBBS" and we applied the same parameters and protocol as for the MD simulations (above). Starting with the equilibrated systems of 8Cl-TIBO in HIV-1 RT and in solution, respectively, the 8-chloro atom was perturbed into a hydrogen (R82150). We continued with a perturbation where the position of the chlorine was changed from 8 to 9 (i.e. R86183 to R82913, see Table I). A window size ($\Delta\lambda$) of 0.02 was used, i.e. 51 windows in the λ-interval [0,1] and for TIBO in solution each window was equilibrated for 2 ps prior to a data collection time of 5 ps per window. The corresponding equilibration/data collection times for TIBO in HIV-1 RT were 3 ps and 8 ps, respectively. In the final, full free energy calculation of derivatives with experimentally known binding affinities, we perturbed the sulfur of R82150 into an oxygen (R80902,

Table I). In the first set (1b and 1s, respectively, see Table II), we used the same parameters and protocol as for the perturbations described above. This yields a $\Delta\Delta G_{bind}$ (R82150 - R80902) = +1.7 kcal/mole which is far from the experimental value of -2.69 kcal/mole. This perturbation involves much larger changes in the electrostatics than the two previous and the free energy change might thus be more sensitive to the treatment of long-ranged electrostatics as well as the local counterion configuration. We therefore changed the protocol as follows (set 2b and 2s, respectively): for TIBO in HIV-1 RT we added counterions (Cl^-) at salt-bridges that were truncated during the setup of the TIBO-HIV-1 RT system (see above) to obtain a net electroneutral system. One of the counterions was constrained to be 3.8 Å from the sulfur/oxygen, since three lysine residues are very close to this part of TIBO (two of them are shown in fig. 1). The non-bonded cutoff was increased to 10 Å and all interactions with TIBO closer than 100 Å were included (i.e. all atoms in the system). Since the perturbation of set 1 was well converged we decreased the equilibration and data collection times to 2 and 5 ps per window, respectively, and we ran the remaining perturbations only in one direction. For TIBO in solution (set 2b), we eliminated possible discrepancies between free energy estimates from a periodic box of water and the "cap" protein simulation, by instead simulating TIBO in a 20 Å radius sphere of TIP3P water. Also here, we applied the 10/100 Å cutoff as described above. Set 2 gives a slightly lower free energy difference (0.7 kcal/mole) than set 1, but is still relatively far from the experimental value. Next, we suspected that the non-bonded parameters of the sulfur atom, which have been developed for sp^3-sulfur[33] might contribute of the erroneous value that we obtain. Therefore, R^* was changed from 2.0 to 1.9 and ε from 0.2 to 0.381, which results in a slightly smaller sulfur

atom but with an unchanged repulsive contribution to the vdW energy (set 3b and 3s, respectively). With these parameters (and with the same protocol as in set 2) we obtain a relative binding free energy of -0.4 kcal/mole, which is considerably closer, but still relatively far away, from the experimental value. The vdW contribution to the free energy (ΔG_{vdw}) is very sensitive to the choice of non-bonded parameters, as is seen from Table II. A further reduction of R^* to 1.844, which with ε=0.55 gives an unchanged repulsive contribution to the vdW energy (set 4b and 4s, respectively), results in a relative binding free energy of -1.0 kcal/mole, which is closer to the experimental value. It does not make physical sense to reduce R^* further.

Similarly, differences in free energy of solvation ($\Delta\Delta G_{solv}$) for two TIBO derivatives can be estimated from the following thermodynamic cycle of the derivatives in gas phase (g) and in solution (aq), respectively:

$$\begin{array}{ccc} L_1(g) & \xrightarrow{\Delta G_1^{'}} & L_1(aq) \\ \downarrow \Delta G_g & & \downarrow \Delta G_s \\ L_2(g) & \xrightarrow{\Delta G_2^{'}} & L_2(aq) \end{array} \qquad (11)$$

and $\Delta\Delta G_{solv}$ is obtained from the relation

$$\Delta\Delta G_{solv} = \Delta G_2^{'} - \Delta G_1^{'} = \Delta G_s - \Delta G_g \qquad (12)$$

ΔG_g was calculated by perturbing the TIBO derivatives *in vacuo* with the same $\Delta\lambda$, equilibration and data collection times as when calculating ΔG_s above.

Finally, we performed perturbations on the next best binding TIBO derivative - "HET" - according to both the adaptive CMC/MD and the PB calculations. Starting from a 500 ps MD equilibration of the HIV-1 RT - HET complex (same parameters and protocol as for 8Cl-TIBO in HIV-1 RT), HET was perturbed into 8Cl-TIBO in two steps (see fig. 3). In the first step the cyclohexyl ring was perturbed into dummy atoms and to avoid the singularity when these atoms disappears at λ=1, we used a soft-core non-bonded potential energy function.[52, 53] For atoms that disappears at λ=1, this function has the form:

$$V_{nb} = (1-\lambda)\sum_{i<j}\left\{\left[\frac{A_{ij}}{\left(\alpha_{LJ}\sigma_{ij}\lambda^2 + r_{ij}^6\right)^2} - \frac{B_{ij}}{\left(\alpha_{LJ}\sigma_{ij}\lambda^2 + r_{ij}^6\right)}\right] + \frac{q_i q_j}{4\pi\varepsilon\left(\alpha_c\lambda^2 + r_{ij}^2\right)^{1/2}}\right\} \qquad (13)$$

where α_{LJ} and α_C are the soft-core parameters for the Lennard-Jones and the electrostatic terms, respectively. We used values of α_{LJ}=0.5 Å^2 and α_{LJ}=15.0 Å^2, that previously[54] has been found suitable. This function is identical to the regular AMBER non-bonded potential function at the perturbation endpoints (λ=0 and λ=1), and they should thus give identical results. The function has the advantage of smoothing the interactions at short interatomic distances, which results in a well-behaved $\Delta G(\lambda)$ function. We used the slightly different free energy perturbation (FEP) scheme for this

step, since the soft-core potential energy function currently is implemented only for that method. The FEP method relies on the following master equation instead of eq. 10 for TI:

$$\Delta G = \sum_{\lambda=0}^{1} - RT \ln \left\langle e^{-[H(\lambda+\delta\lambda)-H(\lambda)]/RT} \right\rangle_{\lambda} \qquad (14)$$

For this perturbation we used the same $\Delta\lambda$ and equilibration/data collection times in solution and protein, respectively, as above. The partial charges of HET were also perturbed into those of 8Cl-TIBO in this step. Bond potential of mean force (pmf) calculations cannot be performed when changing bond lengths of systems belonging to closed rings (as here), when using the FEP method. We therefore kept all bond lengths at their initial values by keeping the atoms bound to positions 8 and 9 in HET (see fig. 3) as carbons (atom type CT). In the second step, the two carbon atoms at positions 8 and 9 were perturbed into 8Cl and 9H, respectively, using the TI method. $\Delta\lambda$ in step 2 was increased to 0.05 and the equilibration/data collection times in solution was chosen to 3 and 8 ps, respectively. The corresponding times in the protein were prolonged to 4 and 10 ps, respectively. Instead of running the perturbations forward and reverse, as in the previous perturbations, we performed two forward (i.e. HET $\rightarrow$ R86183) perturbations (run 1 and 2), which differ by an equilibration of 100 ps of HET in HIV-1 RT and in solution, respectively.

Test of chlorine parameters. To test whether the van der Waals parameters for chlorine, that were adopted from chloroform,[34] also could be used for the TIBO

derivatives, we estimated the relative free energy of solvation ($\Delta\Delta G_{solv}$) between benzene and chlorobenzene. Partial charges of these two compounds (Table III) were obtained from RESP fits[37] of the 6-31G* electrostatic potentials, calculated with the program Gaussian94.[36] Benzene was perturbed into chlorobenzene using the TI method. We used a window size ($\Delta\lambda$)of 0.01 with equilibration and data collection times of 1 ps in each window for both the perturbation in solution and *in vacuo*. Non bonded interactions were cut off at 9 A and a time step of 2 fs was used. For the perturbation in solution benzene was immersed in a box of TIP3P water[40] of dimensions 25 x 27 x 21 A^3 and equilibrated for 50 ps

RESULTS

PROFEC contour maps. We were able to extract meaningful information from the PROFEC contour maps centered around C4 and Cl8 (fig. 4). Fig. 4 (top) shows the zero level (i.e. $\Delta\Delta G_{ins}$=0) PROFEC contour map centered around atom C4. The cavity around C4 suggests that addition of an atom/group would improve the binding of the inhibitor. Therefore, we added a methyl group in a *cis* position relative to the methyl group at position 5 and we denoted this derivative 45MeT (see Table I). The contour map also partly overlaps the methyl group at position 5 and since it is unfavorable to have groups outside the 'cage' formed by the contour map (see fig. 4, top), a removal of this methyl group was predicted to improve binding. Thus, for the following two compounds we removed the C5 methyl group and in one of these, we kept the methyl at the C4 position (4MeT). In the other, we added a chlorine in the C4 position (4ClT), since $\partial\Delta\Delta G_{ins}/\partial q$ (see "Methods") is positive in the map around the C4 cavity (blue in fig. 4), suggesting

that the added group/atom should be electronegative. There is also a cavity around the atoms at positions 8 and 9, as seen from the contour map centered around Cl8 (fig. 4, bottom) and the multicolored plot suggests that the added group should be electroneutral. We added two new substituents at this position - a cyclohexyl group and a benzene ring condensed at positions 8 and 9 - and these compounds are denoted "HET" and "BET", respectively (see Table I).

Adaptive CMC/MD. Adaptive CMC/MD were run for two sets of derivatives and the first set consisted of R86183, R82913, R84963, R80150, R84194, R80902, HET, BET, 4MeT and 4ClT (see Table IV). The free energies of solvation (ΔG_{solv}) were subtracted from the energy "offsets" (see "Methods"), that were obtained from the 450 ps simulation, and the relative free energies of binding are shown in Table IV. The values of ΔG_{solv}, that were estimated from GB/SA calculations, are shown in Table V. From the first adaptive CMC/MD run we note that R80902 is clearly ranked as being the poorest inhibitor, in agreement with experiments, and we therefore discarded this compound in the next set. Moreover, HET and the similar BET were shown to be tight binding inhibitors (HET is about 2 kcal/mole better than R86183), so we discarded also these two derivatives in the next run. In set 2, the three discarded compounds were replaced with two derivatives with experimentally known binding affinities, R87027 and R84674, and we also included one new compound (45MeT, see Table I). The second set was run for 560 ps and for derivatives present in both sets we averaged the two estimates of the relative binding free energy. The rank order of the 8 different TIBO derivatives with known experimental binding affinity according to adaptive CMC/MD is in good

agreement with experiments (Table IV and fig. 5, top) with an unsigned average error of the relative binding free energy of 1.0 kcal/mole. In this context, we should however point out that discrepancies between computed and experimental values also can be due to an imperfect agreement between HIV-1 RT activity and binding affinity, caused by differences in cell penetrating ability and metabolic stability between the TIBO derivatives. The three best binding derivatives, according to experimental results, were also ranked as the three best among the inhibitors with known binding affinity (bold numbers in Table IV). The binding free energy of R87027 and R84674 have both been underestimated with the adaptive CMC/MD method, which erroneously ranks then as better binders than R86183. The next three derivatives have almost the same experimental binding affinity, and they are also ranked between 4 and 6 among the derivatives with known experimental binding affinities. Finally the two derivatives with poorest experimental binding free energy, R84194 and R80902, have also been ranked as the worst binders among the derivatives with known binding affinity according to the adaptive CMC/MD method. Among the new derivatives, HET is ranked as being the best with a binding free energy of 2 kcal/mole better than R86183. BET is also one of the better inhibitors, whereas the other three PROFEC compounds were found to be poor binders.

PB/SA calculations. ΔG_{solv}, as estimated with the PB/SA calculations (see "Methods") are in reasonable agreement with those obtained from the GB/SA method (Table V). The relative solvation free energies ($\Delta\Delta G_{solv}$), which is the property of interest in this comparative study, have an average (unsigned) error of 0.4 kcal/mol between the

two methods. Also from the PB/SA calculations, we obtain a relatively good agreement with the experimental rank order of binding to HIV-1 RT (Table IV and fig. 5, bottom) and the unsigned average error is 1.3 kcal/mole. The binding free energy of the derivative R84914 has been underestimated and it thus has a too favorable ranking. When omitting this derivative, the rank order among the derivatives with known binding affinity coincides with that of the experiment except for R87027 and R84674, where the rank order is reversed. The derivative 45MeT is ranked as being the best binding derivative, closely followed by HET, which according to this method is ranked as number 2. BET is also a good binder with this method whereas 4MeT and 4ClT are both poor binding inhibitors as also was found in the adaptive CMC/MD method. We will discuss possible reasons for the large discrepancy between the two methods for 45MeT below. Dissecting the terms of ΔG_{bound} (see "Computational Methods") we find that the vdW energy between the derivative and HIV-1 RT (Table VI) is the most favorable of all derivatives for HET closely followed by 45MeT and BET. From this table it is also apparent that the vdW energy is the most important term, determining almost solely the strength of the binding the derivatives to HIV-1 RT. This is expected since most of the variation between the derivatives consist of modifications to hydrophobic groups.

Free energy calculations (TI and FEP). The relative free energies of solvation ($\Delta\Delta G_{solv}$) for R86183, R82150 and R80902 as estimated with the TI method (Table VII) are all well converged and in qualitative agreement with those estimated from the GB/SA and PB/SA calculations (Table V). A much poorer convergence is found for the perturbations in HIV-1 RT in spite of the prolonged equilibration and data collection

times (Table VIII). In both the R82150→R86183 and R82913→R86183 perturbations the vdW interaction is almost solely responsible for the hysteresis between the forward and reverse runs. We also note that differences in vdW interactions are responsible for the difference in binding free energy between R86183 and R82150. This is consistent with that these interactions are dominating the differences in binding free energies as found from the PB calculations (Table VI). Considering that TIBO is bound in a pocket, with predominantly hydrophobic and aromatic residues (see fig. 1) it is not surprising that differences in binding strength are governed by vdW interactions. The relative binding free energy of -1.9±0.5 kcal/mole that we obtain is in close agreement with experimental results[32] (-1.34 kcal/mole). In the R82913→R86183 perturbations, the contributions to the differences in binding affinities are shared between vdW and pmf contributions, whereas the electrostatic contribution to the difference in binding affinity is negligible. Here, we get a $\Delta\Delta G_{bind}$ of -3.2±0.5 kcal/mole, in qualitative agreement with experiments (-1.17 kcal/mole).

The relative free energy of perturbing HET to R86183, via the intermediate (see fig. 3) is estimated to -0.8±0.7 kcal/mole (Table IX). This agrees with the other two methods, supporting the prediction that HET should have improved affinity for HIV-1 RT compared to the parent compound R86183. A comparison of the individual contributions to the stability, summing over both steps of the perturbation, shows that HET is a tighter binder mainly because of a stronger vdW interaction (Table IX). This is also consistent with the PB/SA calculations, where the inhibitor-protein vdW interaction energy was strongest for HET.

Test of chlorine parameters. The results from the TI calculations are well converged, with very close values for the forward and reverse runs (Table X). We obtain a $\Delta\Delta G_{solv}$ (benzene - chlorobenzene) of -0.19 ± 0.06 kcal/mole which is in reasonable agreement with the experimental result[55] of 0.12 kcal/mole.

DISCUSSION

The "full" free energy calculations on TIBO derivatives with known experimental binding affinity were performed as an initial check whether it was possible to reproduce the relative binding affinities with the most rigorous method prior to applying the more approximate approaches to this system. In spite of the relatively poor convergence for the R82150 → R86183 and R89193 → R86183 perturbations, they both give reasonable estimates for the relative free energy. However, we were only able to qualitatively reproduce the relative binding free energy of R80902 versus R82150, with significant changes in the non-bonded parameters of sulfur (described in the "Computational Methods" section). These calculations find that the non-bonded interactions of the sulfur atom are strongly contributing to the erroneous result we obtained for our initial set of parameters (Table II). Since our initial non-bonded parameters for sulfur was the same as used in cysteine and methionine, it is not unreasonable that the non-bonded parameters for sp^2-sulfur (as in TIBO) have a smaller R^* and a larger ε than found for an sp^3 single bonded sulfur. However, even with the modified sulfur parameters there is a significant quantitative difference (1.7 kcal/mole) between calculated and experimental $\Delta\Delta G_{bind}$ for R80902 → R82150 (C=O → C=S). Interestingly, the PB/SA and CMC/MD methods both estimate this $\Delta\Delta G_{bind}$ in close agreement with the experimental value (Table IV) even with the initial set of sulfur non-bonded parameters. We do not understand why the "most rigorous" approach is less accurate in this regard, but this is further support for the use of multiple methods to make binding free energy predictions.

The derivatives 4MeT and 4CIT are both estimated as being poor binders in spite of the fact that we would expect them to bind better than R86183 from the PROFEC contour

maps. However, it is not obvious that the cavity that was found around atom C4 is present when the C5 methyl group is also removed. Since the contour maps are based on simulations of a single parent compound (R86183), they give no information about what would happen with the cavity if other changes are made to the inhibitor. Comparing the minimized structures of 45MeT, 4MeT and 4ClT from the PB calculations, we also observe small tendencies for the two latter derivatives to be pushed away from the original cavity at position C4. This is probably also reflected when comparing the inhibitor-protein vdW energies (Table VI). Both 4MeT and 4ClT have unfavorable vdW energies compared to 45MeT which is the major reason for their poor binding according to the PB calculations. The high solubility of 4ClT (Table V) makes this inhibitor an even weaker binder. The very tight binding of 45MeT that was predicted by the PB method is not consistent with the results from the adaptive MC/MD runs, where this derivative instead is estimated as being a poor binder (Table VI). This discrepancy illustrates one of the most severe problems with the adaptive CMC/MD as currently implemented. Within the limited time of sampling, certain derivatives might be over- or undersampled, due to the fact that the binding mode of one derivative might not be favorable for another derivative. Therefore, during the course of the CMC/MD run, a certain derivative may never (or rarely) interact optimally with the protein, especially if such an interaction would require rather large structural changes of the protein. This might be the case for 45MeT, in that rearranging nearby protein residues for an optimal interaction is a slower process than can be caught within our sampled time of 560 ps. The oversampling of R87027 (leading to a too favorable relative binding free energy) is most probably an example of the same problem, but reversed. The surrounding side chains of the protein

have to rearrange in order to accommodate the relatively long "tail" of R87027 that results from the two added methyl groups (see Table I). When performing an MC step with R87027 as the sampled inhibitor, it might be difficult for the other derivatives to find an optimal conformation of the protein, resulting in rejections of these trial steps (see "Computational Methods"). The algorithm might therefore get temporarily "stuck" sampling R87027 due to incompatible binding modes, leading to an overestimation of its binding free energy. The "adaptive" CMC/MD version was partly constructed in order to overcome this problem and it is an improvement over the non adaptive protocol, where the convergence was extremely slow for this system. However, convergence of the calculated free energies is still hampered by the problem described above. This difficulty might further be reduced by adding multiple conformers (rotamers) of some critical side-chains of the protein and also permitting them to participate in the CMC sampling. Work is currently in progress (J. Pitera) to implement multiple copies of protein side chains, in order to improve the convergence rate. Another observation, comparing set 1 and 2, is that the relative binding free energies are consequently higher in set 2 (except for 4ClT). The stems from the fact that the "reference derivative", R86183, is being estimated as a relatively better binder in set 2 than in set 1, which thus shifts the relative free energies of all the other compounds.

Most encouraging is that we find HET to be a much better binder than R86183 with both the adaptive CMC/MD and the PB/SA method. "Full" free energy calculations also support this prediction (albeit less clearly due to the large error estimate), lending support to the conclusions of the other two methods. The physical picture of HET binding to HIV-1 RT also suggests that this affinity is plausible. The hydrophobic cyclohexyl

moiety of TIBO fits very well into a pocket, composed of many hydrophobic side chains; Val106, Tyr188, Phe227 and Leu234 (fig. 6). This is also consistent with the calculation that HET has the highest vdW interaction with the protein of all compared derivatives. Coupled with a more unfavorable solvation free energy, this yields a much greater affinity for HIV-1 RT than the parent compound, R86183.

From this study is it not possible to judge which of the two approximate methods (adaptive CMC/MD and PB/SA) performs best in ranking the derivatives with respect to the binding free energies since both methods have their strengths and shortcomings. In CMC/MD, on one hand, the solvent with counterions is modeled explicitly and a large number of conformations are sampled, but as now implemented suffers from the slow convergence when the sampled inhibitors have differing binding modes. The PB/SA method, on the other hand, does not suffer from that problem, but the solvent is instead approximately described as a continuum and we only consider one conformation of the ligand and the complex, respectively. Since the two method have different sets of approximations we feel that they complement each other and that a consistent result between them increases its validity.

CONCLUSIONS

In this study we have used two methods in order to rank 13 different TIBO derivatives - the adaptive CMC/MD method and PB/SA calculations. Five of these derivatives were new modifications, that were made from suggestions generated by PROFEC contour maps. The rest of the TIBO derivatives have experimentally determined[32] binding affinities. Both methods work surprisingly well, yielding rank orders in good agreement with experimental results. Since the two methods are quite different in nature, each with their own sets of approximations, they serve as a good complement to each other. That is, if both methods predict the same rank order, the reliability of this prediction will significantly increase. The methods are also relatively rapid - a total of 0.8 ns simulation time was needed in order to rank the 13 TIBO derivatives with the adaptive CMC/MD method. To put this in perspective, considerable simulation times (1.1 ns or more) were required in order of obtain an estimate of the relative binding free energy between only one pair of derivatives with a FEP/TI calculation. We found that one of the new modifications (HET), as suggested from PROFEC, was binding about 1-2 kcal/mole better than R86183 according to both methods. This result was confirmed with subsequent FEP/TI calculations. The hydrophobic cyclohexyl moiety that was added to TIBO fits well into a cavity in HIV-1 RT that consists of many non-polar residues. This is also consistent with the observation that HET has the most favorable vdW interaction with HIV-1 RT among the TIBO derivatives studied.

The protocol we have used in this paper (outlined in fig. 2) is a general strategy for computational structure-based lead optimization. While we have not considered crucial pharmacological issues like bioavailability and toxicity, our approach appears to be useful for optimization of affinity. Starting from a parent lead compound (or a family of compounds), PROFEC can be used to suggest where modifications of the lead should be made to improve the binding affinity of the lead compound. PB/SA and adaptive CMC/MC can then be applied for ranking of the PROFEC derivatives, preferably together with derivatives of known binding affinity if such are available. Thereafter FEP/TI can be used to study particularly interesting derivatives and to confirm results from the more approximate methods, followed by synthesis and *in vitro* testing of the best binding derivative(s). However, as one reviewer noted, the expense of FEP/TI calculations suggests a revised strategy, where the compounds selected by CMC/MD and PB/SA are synthesized and tested, without carrying out a FEP/TI calculation. We agree that one can simply use any predictions from PROFEC, CMC/MD, or PB/SA directly, but if significant synthetic efforts are required, it is certainly worth a confirmatory calculation with FEP/TI to see if such efforts are justified. Regardless of whether FEP/TI calculations are used, we feel that the strategy used in this paper provides an excellent blueprint for lead optimization in structure-based drug design.

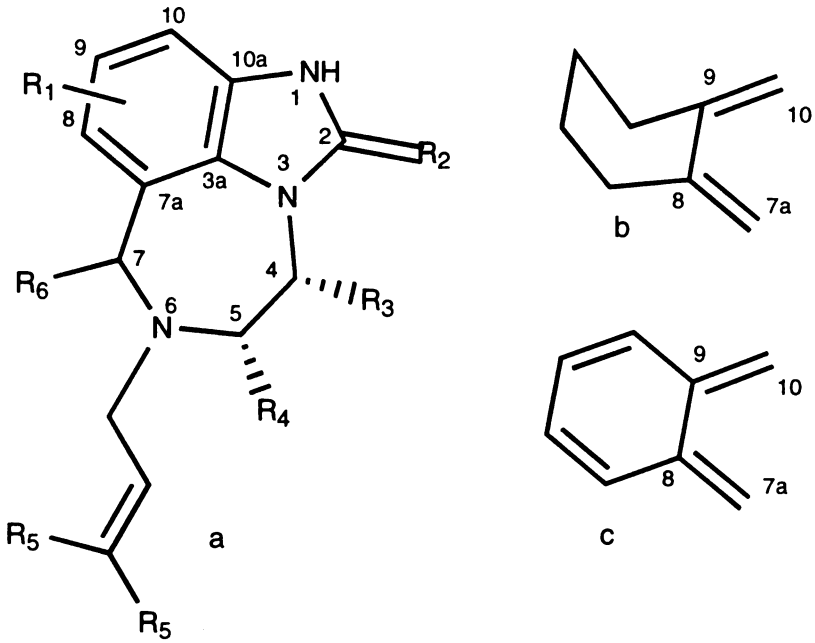
SUPPORTING INFORMATION

Partial charges of the 13 TIBO derivatives (2 pages).

ACKNOWLEDGEMENTS

Mats Eriksson gratefully acknowledges a postdoctoral grant from the Swedish Natural Science Research Council (NFR). Jed Pitera is grateful to the NSF and the UCSF Chancellor's Office for predoctoral support. Peter Kollman thanks the NIH for research support through grant GM56609 (Prof. E. Arnold, P.I.). We are grateful to Professor Edward Arnold for providing us with the coordinates of the R86183 - HIV-1 RT complex and for the computer time provided by him on the NCI Cray C90.

Table I. The selected set of TIBO derivatives.



Compound	R_1	R_2	R_3	R_4	R_5	R_6	EC_{50} (nM)*
R86183	8-Cl	S	H	CH_3	H	H	4.6
R82913	9-Cl	S	H	CH_3	H	H	33
R82150	H	S	H	CH_3	H	H	44
R80902	H	O	H	CH_3	H	H	4200
R84674	8-CH_3	S	H	CH_3	H	H	14
R84963	H	S	H	CH_3	H	-CH_3 (*trans*)**	39
R84914	H	S	H	CH_3	H	-CH_3 (*cis*)**	790
R87027	8-Cl	S	H	CH_3	CH_3	H	5.1
45MeT	8-Cl	S	CH_3	CH_3	H	H	
4MeT	8-Cl	S	CH_3	H	H	H	
4ClT	8-Cl	S	Cl	H	H	H	
HET	b	S	H	CH_3	H	H	
BET	c	S	H	CH_3	H	H	

* ref. 32

**relative stereochemistry of the methyl groups at positions 5 and 7.

Table II. Free energy perturbation (TI), R80902 → R82150 using various protocols and parameters. The free energies are in kcal/mole.

	in HIV-1 RT				in solution			
	set 1b[†]	set 2b	set 3b	set 4b	set 1s[†]	set 2s	set 3s	set 4s
ΔG_{tot}	27.92 ±0.01	27.00	24.54	23.03	26.20 ±0.05	26.29	24.91	24.03
$\Delta G_{el.stat}$	23.82 ±0.03	22.61	23.07	22.92	23.14 ±0.03	23.06	22.95	22.86
ΔG_{vdw}	3.03 ±0.01	3.58	1.13	0.24	3.01 ±0.02	3.08	1.89	1.08
ΔG_{pmf}	1.07 ±0.04	0.81	0.35	-0.13	0.05 ±0.04	0.015	0.076	0.089

[†] Set 1b: dual cutoff (9/13 Å) $\delta\lambda$=0.02, t_{eq}=3 ps, $t_{sampl.}$= 8 ps; Set 2b: cutoff 10 Å, cutoff for TIBO 100 Å, electroneutral system, one Cl^- constrained to be 3.8 Å from sulfur/oxygen atom in TIBO, $\delta\lambda$=0.02, t_{eq}=2 ps, $t_{sampl.}$= 5 ps; Set 3b: as set 2, but with R^* changed from 2.0 to 1.9 and ε from 0.2 to 0.381; Set 4b: as set 2, but with R^*=1.844 and ε=0.55; Set 1s: as set 1b, but in a water box of water and with t_{eq}=2 ps, $t_{sampl.}$= 5 ps; Set 2s: as set 1s, but in a water sphere with radius 20 Å and cutoff 10 Å, cutoff for TIBO 100 Å; Set 3s: as set 2s, but with R^* changed from 2.0 to 1.9 and ε from 0.2 to 0.381; Set 4s: as set 2s, but with R^*=1.844 and ε=0.55.

Table III. RESP derived[37] partial charges for benzene and chlorobenzene.

atom	q (benzene)	q (chlorobenzene)
1C	-0.146	-0.033
1H/Cl	0.146	-0.115
2C	-0.146	-0.053
2H	0.146	0.133
3C	-0.146	-0.156
3H	0.146	0.147
4C	-0.146	-0.145
4H	0.146	0.150

Table IV. Binding free energies (relative to R86183, kcal/mole) and rank order of binding to HIV-1 RT according to adaptive CMC/MD, PB calculations and experimental[32] values.

	adaptive CMC/MD		PB		experimental	
derivative	$\Delta\Delta G_{bind}$	rank	$\Delta\Delta G_{bind}$	rank	$\Delta\Delta G_{bind}$	rank
R86183	$(0^a+0^b)/2=0$	4 (**3**)	0	3 (**1**)	0	1
R87027	-2.56^b	1 (**1**)	1.87	7 (**4**)	0.06	2
R84674	-0.74^b	3 (**2**)	0.21	5 (**2**)	0.66	3
R82913	$(0.69^a+1.70^b)/2=1.19$	8 (**6**)	2.24	8 (**5**)	1.17	4
R84963	$(-0.12^a+1.28^b)/2=0.58$	6 (**4**)	2.31	9 (**6**)	1.27	5
R82150	$(0.67^a+1.15^b)/2=0.91$	7 (**5**)	2.67	11 (**7**)	1.34	6
R84914	$(0.78^a+2.01^b)/2=1.39$	9 (**7**)	0.86	6 (**3**)	3.05	7
R80902	3.71^a	12 (**8**)	5.19	13 (**8**)	4.04	8
HET	-1.94^a	2	-1.28	2		
45MeT	1.80^b	11	-1.47	1		
BET	0.50^a	5	0.09	4		
4MeT	$(1.11^a+1.74^b)/2=1.42$	10	2.36	10		
4ClT	$(2.02^a+1.58^b)/2=1.80$	11	4.19	12		

[a] set 1, 450 ps adaptive CMC/MD
[b] set 2, 560 ps adaptive CMC/MD

Table V. Estimated ΔG_{solv} (kcal/mole) of the TIBO derivatives, using GB/SA-, PB/SA- and TI- calculations,

TIBO derivative	ΔG_{solv} GB/SA	ΔG_{solv} - ΔG_{solv}(R86183) GB/SA	ΔG_{solv} PB	ΔG_{solv} - ΔG_{solv}(R86183) PB	ΔG_{solv}- ΔG_{solv}(R86183) TI
R86183	-5.14	0	-3.92	0	0
R82150	-5.50	-0.36	-4.18	-0.26	-0.73
R82913	-5.21	-0.07	-4.18	-0.26	-0.60
R80902	-7.12	-1.98	-5.00	-1.08	-3.04
R84674	-4.48	0.66	-3.99	-0.07	
R84963	-4.38	0.76	-4.30	-0.38	
R84914	-5.20	-0.06	-3.89	0.03	
R87027	-4.99	0.15	-3.69	0.23	
4ClT	-6.41	-1.27	-4.86	-0.94	
4MeT	-4.96	0.18	-3.87	0.05	
45MeT	-4.45	0.69	-3.59	0.33	
HET	-4.19	0.95	-3.43	0.49	
BET	-5.28	-0.14	-4.27	-0.35	

Table VI. Energy quantities (kcal/mole, eq. 5 and 7) for calculation of binding free energies, according to the PB/SA-method.

derivative	$\Delta G^{g\to aq}_{react,L}(LP)$	$\Delta G_{lig-prot,vdw}$	$\Delta G_{lig-prot,elec}$	ΔG_{solv}	ΔG_{bind}
R86183	0.175	-52.71	-6.45	-3.92	-55.06
R87027	0.209	-50.64	-6.47	-3.69	-53.19
R84674	0.505	-52.05	-7.30	-3.99	-54.86
R82913	0.555	-51.00	-6.56	-4.18	-52.82
R84963	0.460	-50.12	-7.40	-4.30	-52.76
R82150	0.520	-49.84	-7.26	-4.18	-52.40
R84914	0.447	-51.50	-7.04	-3.89	-54.20
R80902	0.378	-48.95	-6.30	-5.00	-49.87
HET	0.520	-53.00	-7.29	-3.43	-56.34
45MeT	0.307	-52.87	-7.56	-3.59	-56.53
BET	0.505	-52.76	-6.99	-4.27	-54.97
4MeT	0.294	-49.52	-7.35	-3.87	-52.71
4ClT	0.196	-48.52	-7.41	-4.86	-50.87

Table VII. Relative solvation free energies (kcal/mole) of a selected set of TIBO derivatives estimated with the TI method.

a/ R82150 → R86183:

	in solution			in vacuo		
	fwd	rev	avg	fwd	rev	avg
ΔG_{tot}	1.531	1.691	1.611 ±0.080	0.858	0.905	0.882 ±0.023
$\Delta G_{el.stat}$	2.522	2.697	2.609 ±0.072	1.843	1.848	1.846 ±0.002
ΔG_{vdw}	0.699	0.703	0.701 ±0.002	1.319	1.202	1.261 ±0.059
ΔG_{pmf}	-1.690	-1.709	-1.700 ±0.010	-2.303	-2.145	-2.224 ±0.080

$\Delta\Delta G_{solv}$ (R86183- R82150) = 0.73±0.08 kcal/mole

b/ R82913 → R86183:

	in solution			in vacuo		
	fwd	rev	avg	fwd	rev	avg
ΔG_{tot}	1.658	1.787	1.722 ±0.064	1.113	1.081	1.118 ±0.006
$\Delta G_{el.stat}$	1.341	1.377	1.35 ±0.018	1.134	1.140	1.137 ±0.003
ΔG_{vdw}	1.677	1.750	1.714 ±0.036	1.595	1.605	1.600 ±0.005
ΔG_{pmf}	-1.358	-1.341	-1.350 ±0.009	-1.616	-1.621	-1.618 ±0.003

$\Delta\Delta G_{solv}$ (R86183- R82913) = 0.60±0.06 kcal/mole

c/ R80902 → R82150:

	in solution			in vacuo		
	fwd	rev	avg	fwd	rev	avg
ΔG_{tot}	26.148	26.250	26.199 ±0.051	23.87	23.903	23.886 ±0.016
$\Delta G_{el.stat}$	23.111	23.174	23.142 ±0.032	22.052	21.964	22.008 ±0.044
ΔG_{vdw}	3.024	2.986	3.005 ±0.019	0.549	0.456	0.502 ±0.046
ΔG_{pmf}	0.013	0.090	0.052 ±0.038	1.269	1.480	1.374 ±0.106

$\Delta\Delta G_{solv}$ (R82150 - R80902) = 2.31±0.05 kcal/mole

Table VIII. Relative free energies (kcal/mole) of binding to HIV-1 RT for a selected set of TIBO derivatives estimated with the TI method.

a/ R82150 → R86183

	in HIV-1 RT			in solution		
	fwd	rev	avg	fwd	rev	avg
ΔG_{tot}	-0.777	0.127	-0.325 ±0.452	1.531	1.691	1.611 ±0.080
$\Delta G_{el.stat}$	2.707	2.202	2.454 ±0.253	2.522	2.697	2.609 ±0.072
ΔG_{vdw}	-1.653	-0.271	-0.962 ±0.691	0.699	0.703	0.701 ±0.002
ΔG_{pmf}	-1.832	-1.727	-1.780 ±0.105	-1.690	-1.709	-1.700 ±0.010

$\Delta\Delta G_{bind}$ (R86183 - R82150) = -1.9±0.5 kcal/mole, experimental value:[32] -1.34

b/ R82913 → R86183

	in HIV-1 RT			in solution		
	fwd	rev	avg	fwd	rev	avg
ΔG_{tot}	-2.109	-0.784	-1.447 ±0.662	1.658	1.787	1.722 ±0.064
$\Delta G_{el.stat}$	1.131	1.372	1.25 ±0.12	1.341	1.377	1.35 ±0.018
ΔG_{vdw}	-0.658	0.091	-0.28 ±0.37	1.677	1.750	1.714 ±0.036
ΔG_{pmf}	-2.582	-2.247	-2.41 ±0.17	-1.358	-1.341	-1.350 ±0.009

$\Delta\Delta G_{bind}$ (R86183 - R82913) = -3.2 ±0.7 kcal/mole, experimental value:[32] -1.17

Table IX. a/ A two-step (see, fig. 3) free energy calculation of R86183 → HET. The energies are given in kcal/mole.

	step 1 (FEP)					
	in HIV-1 RT			in solution		
	Run 1*	run 2	Avg	run 1	run 2	avg
ΔG_{tot}	-4.043	-5.333	-4.668 ±0.645	-2.830	-2.232	-2.531 ±0.299
ΔG_{elstat}	-3.475	-4.308	-3.892 ±0.417	-3.094	-3.178	-3.136 ±0.042
ΔG_{vdw}	-1.723	-2.403	-2.063 ±0.340	-0.657	0.268	-0.194 ±0.462
ΔG_{14vdw}	2.854	2.818	2.836 ±0.018	3.020	3.021	3.020 ±0.001
$\Delta G_{14elstat}$	-1.655	-1.410	-1.532 ±0.122	-2.094	-2.088	-2.091 ±0.003
	step 2 (TI)					
	Run 1	run 2	Avg	run 1	run 2	avg
ΔG_{tot}	1.963	2.324	2.144 ±0.180	0.838	0.806	0.822 ±0.016
ΔG_{vdw}	0.513	0.719	0.616 ±0.103	-0.088	-0.133	-0.110 ±0.022
ΔG_{14vdw}	1.048	1.036	1.042 ±0.006	0.860	0.897	0.878 ±0.018
ΔG_{badh}**	0.027	0.021	0.024 ±0.003	0.020	0.035	0.028 ±0.008
ΔG_{pmf}	0.389	0.556	0.472 ±0.083	0.050	0.010	0.075 ±0.020

* see "Computational Methods".

** free energy contribution from bonds, angles and dihedrals.

b/ summary table from the two perturbation steps.

	ΔG_{tot} (in HIV-1 RT)	ΔG_{tot} (in solution)	ΔG_{tot} (in HIV-1 RT) - ΔG_{tot} (in solution)
step1	-4.668±0.645	-2.531±0.299	-2.137±0.711
step2	2.144±0.180	0.822±0.016	1.322±0.181
step1 + step 2	-2.524±0.670	-1.709±0.299	-0.815±0.733

$\Delta\Delta G_{bind}$ (HET - R86183) = -0.82±0.73 kcal/mole.

Table X. Thermodynamic integration, chlorobenzene → benzene. The energies are given in kcal/mole.

	in solution			in vacuo		
	Fwd	rev	avg	fwd	rev	avg
ΔG_{tot}	1.980	1.887	1.93 ±0.05	2.072	2.161	2.12 ±0.04
$\Delta G_{el.stat}$	0.764	0.600	0.68 ±0.08	1.485	1.486	1.485 ±0.001
ΔG_{vdw}	0.905	0.931	0.92 ±0.01	0.150	0.151	0.15 ±0.001
ΔG_{pmf}	0.311	0.356	0.33 ±0.02	0.437	0.525	0.48 ±0.04

$\Delta\Delta G_{solv}$ (benzene - chlorobenzene) = -0.19 ±0.06 kcal/mole, experimental value [55]: 0.12 kcal/mole.

Figure captions.

Figure 1. 8Cl-TIBO (R86183, red) in HIV-1 RT. The non-polar (green) and polar (white) residues shown have any atom closer than 3.5 Å from any TIBO atom. This is a snapshot from the 500 ps MD simulation and some close water molecules are also shown as blue spheres. This picture and the other molecular graphics images in this paper have been created with the program MidasPlus.[43]

Figure 2. General outline of a procedure that can be applied as one of the final steps in computational structure-based lead optimization. We have used this strategy for the TIBO derivatives in this work.

Figure 3. The two-step perturbation of HET → R86183 (only the aromatic part of TIBO is shown). DH and DC are dummy hydrogen and carbon atoms, respectively. These atoms have the same masses are the original atoms, but no interactions with the surrounding.

Figure 4. (stereo views) PROFEC[31] contour maps (contour level, $\Delta\Delta G_{ins}=0$) around R86183, centered around atom C4 (top) and 8Cl (bottom). Outside the cages that are formed from the contour maps, is it unfavorable to add a particle ($\Delta\Delta G_{ins}>0$), whereas inside, an addition of a particle improves the binding free energy ($\Delta\Delta G_{ins}<0$). The color of the map ranges from blue when $\partial\Delta\Delta G_{ins}/\partial q >0$ to red when $\partial\Delta\Delta G_{ins}/\partial q < 0$. Green and yellow thus correspond to areas with $\partial\Delta\Delta G_{ins}/\partial q$ closer to 0. (see "Computational methods").

Figure 5. Estimated binding free energies (kcal/mole) relative to R86183 according to CMC/MD (top) and PB/SA (bottom) calculations, plotted against experimental values.[32] The least-square linear fits are also shown in the plots.

Figure 6. A snapshot from the MD simulation of HET (red) in HIV-1 RT (white). Hydrophobic residues with any atom < 3.5 A from the cyclohexyl moiety atoms are shown in green and water molecules are shown as yellow spheres.

References

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

(2) Larder, B. A. Inhibitors of HIV reverse transcriptase as antiviral agents and drug resistance. *Cold Spring Harbor Lab. Press, Cold Spring Harbor, New York* **1993,** 205-222.

(3) Schinazi, R. F. Competitive inhibitors of human immunodeficiency virus reverse transcriptase. *Perspect. Drug Discovery Design.* **1993,** *1*, 151-180.

(4) De Clercq, E. Antiviral therapy of human immunodeficiency virus infections. *Clin. Microbiol. Rev.* **1995,** *8*, 200-239.

(5) DeClercq, E. HIV resistance to reverse transcriptase inhibitors. *Biochem. Pharmacol.* **1994,** *47*, 155-169.

(6) Young, S. D. Non-nucleoside inhibitors of HIV-1 reverse transcriptase. *Perspect. Drug Discov. Design* **1993,** *1*, 181-192.

(7) De Clercq, E. Toward improved anti-HIV chemotherapy: therapeutic strategies for intervention with HIV infections. *J. Med. Chem.* **1995,** *38*, 2491-2517.

(8) Arnold, E.; Das, K.; Ding, J.; Yadav, P. N. S.; Hsiou, Y.; Boyer, P. L. ; Hughes, S. H. Targeting HIV reverse transcriptase for anti-AIDS drug design. *Drug Design Discov.* **1996,** *13*, 29-47.

(9) Ding, J.; Das, K.; Moereels, H.; Koymans, L.; Andries, K.; Janssen, P. A. J.; Hughes, S. H. ; Arnold, E. Structure of HIV-1 RT/TIBO R 86183 complex reveals

similarity in the binding of diverse nonnucleoside inhibitors. *Structural Biology* **1995,** *2*, 407-415.

(10) Das, K.; Ding, J.; Hsiou, Y.; Clark, A. D. J.; Moereels, H.; Koymans, L.; Andries, K.; Pauwels, R.; Janssen, P. A. J.; Boyer, P. L.; Clark, P.; Smith, R. H. J.; Kroeger-Smith, M. B.; Michedja, 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.

(11) Tantillo, C.; Ding, J.; Jacobo-Molina, A.; Nanni, R. G.; Boyer, P. L.; Hughes, S. H.; Pauwels, R.; Andries, K.; Janssen, P. A. J. ; Arnold, E. Locations of anti-AIDS drug binding sites and resistance mutations in the three-dimensional structure of HIV-1 reverse transcriptase. *J. Mol. Biol.* **1994,** *243*, 369-387.

(12) Beveridge, D. L. ; DiCapua, F. M. Free energy via molecular simulation: Application of chemical and biochemical systems. *Annu. Rev. Biophys. Biophys. Chem.* **1983,** *18*, 431-492.

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

(14) Rao, B. G.; Kim, E. E. ; Murcko, M. A. Calculation of solvation and binding free energy differences between VX-478 and its analogs by free energy perturbation and AMSOL methods. *J. Comput. Aid. Molec. Des.* **1996,** *10*, 23-30.

(15) Pitera, J. ; Kollman, P. Designing an optimum guest for a host using multimolecule free energy calculations: Predicting the best ligand for Rebek's "Tennis Ball" . *J. Am. Chem. Soc.* **1998,** *120*, 7557-7567.

(16) Gilson, M. K. ; Honig, B. Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies and conformational analysis. *Proteins: Struct. Funct. & Genet.* **1988,** *4*, 7-18.

(17) Sharp, K. A. ; Honig, B. Electrostatic interactions in macromolecules: theory and applications. *Annu. Rev. Biophys. Biophys. Chem.* **1990,** *19*, 301-332.

(18) Nozaki, Y. ; Tanford, C. The solubility of amino acids and two glycine peptides in aqueous solution and dixonane solutions. Establishment of a hydrophobicity scale. *J. Biol. Chem.* **1971,** *246*, 2211-2217.

(19) Hermann, R. B. Theory of hydrophobic binding. II. The correlation of hydrocarbon solubility in water with solvent cavity surface area. *J. Phys. Chem.* **1971,** *76*, 2754 -2759.

(20) Wendoloski, J. J.; Shen, J.; Oliva, M. T. ; Weber, P. C. Biophysical tools for structure-based drug design. *Pharmac. Ther.* **1993,** *66*, 169-183.

(21) Jackson, R. M. ; Sternberg, M. J. E. A continuum model for protein-protein interactions: Application to the docking problem. *J. Mol. Biol.* **1995,** *250*, 258-275.

(22) Shen, J. ; Quiocho, F. A. Calculation of binding energy differences receptor-ligand systems using the Poisson-Boltzmann method. *J. Comp. Chem.* **1995,** *16*, 445-448.

(23) Shen, J. ; Wendoloski, J. Binding of phosphorus-containing inhibitors to thermolysin studied by the Poisson-Boltzmann method. *J. Comp. Chem.* **1996,** *17*, 350-357.

(24) Zhang, T. ; Koshland, J. D. E. Computational method for relative binding energies of enzyme-substrate complexes. *Prot. Sci.* **1996,** *5*, 348-356.

(25) Froloff, N.; Windemuth, A. ; Honig, B. On the calculation of binding free energies using continuum methods: Application to MHC class I protein-peptide interactions. *Prot. Sci.* **1997,** *6*, 1293-1301.

(26) Åqvist, J.; Medina, C. ; Samuelsson, J.-E. A new method for predicting binding affinity in computer-aided drug design. *Prot. Eng.* **1994,** *7*, 385-391.

(27) Åqvist, J. ; Mowbray, S. L. Sugar recognition by a glucose/galactose receptor. *J. Biol. Chem.* **1995,** *270*, 9978-9981.

(28) Hansson, T. ; Åqvist, J. Estimation of binding free energies for HIV-1 protease inhibitors by molecular dynamics simulations. *Prot. Eng.* **1995,** *8*, 1137-1144.

(29) Paulsen, M. D. ; Ornstein, R. L. Binding free energy calculations for P450cam-substrate complexes. *Prot. Eng.* **1996,** *9*, 567-571.

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

(31) Radmer, R. J. ; Kollman, P. A. The application of three approximative free energy calculation methods to structure based ligand design: Trypsin and its complex inhibitors. *J. Comput.-Aided Mol. Design* **1998,** *12*, 215-227.

(32) Pauwels, R.; Andries, K.; Debyser, Z.; Kukla, M. J.; Schols, D.; Breslin, H. J.; Woestenborghs, R.; Desmyter, J.; Janssen, M. A. C.; de Clercq, E. ; Janssen, P. A. J. New tetrahydroimidazo[4,5,1-*jk*][1,4]-benzodiazepin-2(1*H*)-one and -thione derivatives are potent inhibitors of human immunodeficiency virus type 1

replication and are synergistic with 2',3'-dideoxynucleoside analogs. *Antimicrob. Agents. Chemother.* **1994,** *38*, 2863-2870.

(33) Miyamoto, S. ; Kollman, P. A. Absolute and relative binding free energy calculations of the interaction of biotin and its analogs with streptavidin using molecular dynamics/free energy perturbation approaches. *Proteins: Struct., Funct. & Genet.* **1993,** *16*, 226-245.

(34) Fox, T.; Thomas, B. E.; McCarrick, M. ; Kollman, P. A. Application of free energy perturbation calculations to the "tennis ball" dimer: Why is CF4 not encapsulated by this host? *J. Phys. Chem.* **1996,** *100*, 10779-10783.

(35) Liaw, Y. C.; Gao, Y. G.; Robinson, H. ; Wang, A. H. J. Molecular structure of a potent HIV-1 inhibitor belonging to TIBO family. *J. Am. Chem. Soc.* **1991,** *113*, 1857-1859.

(36) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Gill, P. M. W.; Jonhson, B. G.; Robb, M. A.; Cheeseman, J. R.; Keith, T.; Petersson, G. A.; Montegomery, J. A.; Rahavachari, K.; Al-Laham, M. A.; Zakrzewski, V. G.; Ortiz, J. V.; Foresman, J. B.; Peng, C. Y.; Ayala, P. Y.; Chen, W.; Wong, M. W.; Andres, J. L.; Replogle, E. S.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Binkley, J. S.; Defrees, D. J.; Baker, J.; Stewart, J. P.; Head-Gordon, M.; Gonzalez, C. ; Pople, J. A. *Gaussian 94* **1995,** *Revision B.3 Gaussian Inc. Pittsburg, PA.*.

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

(38) Pearlman, D. A.; Case, D. A.; Caldwell, J. C.; Ross, W. S.; Cheatham III, T. E.; Ferguson, D. M.; Seibel, G. L.; Singh, U. C.; Weiner, P. ; Kollman, P. A. AMBER 4.1 (UCSF), University of California, San Francisco. **1994.**

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

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

(41) Ryckaert, J. P.; Ciccotti, G. ; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comp. Phys.* **1977,** *23*, 327-341.

(42) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Di Nola, A. ; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984,** *81*, 3684-3690.

(43) Ferrin, T. E.; Huang, C. C.; Jarvis, L. E. ; Langridge, R. The Midas display system. *J. Mol. Graph.* **1988,** *6*, 13-27.

(44) Metropolis, N. R.; Rosenbluth, M. N.; Teller, A. H. ; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953,** *21*, 1087-1092.

(45) Torrie, G. M. ; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free energy estimation: Umbrella sampling. *J. Comp. Phys.* **1977,** *23*, 187-199.

(46) Kumar, S.; Swendsen, R. H.; Kollman, P. A. ; Rosenberg, J. M. The weighted histogram analysis for free energy calculations on biomolecules. 1. The method. *J. Comp. Chem.* **1992,** *13*, 1011-1021.

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

(48) Mohamadi, F.; Richards, N. G. J.; Guida, W. C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T. ; Still, W. C. MacroModel - An integrated software system for modeling organic and bioorganic molecules using molecular mechanics. *J. Comp. Chem.* **1990,** *11*, 440-467.

(49) Gilson, M. K.; Sharp, K. A. ; Honig, B. Calculating the electrostatic potential of molecules in solution: method and error assessment. *J. Comp. Chem.* **1988,** *9*, 327-335.

(50) Honig, B. ; Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **1995,** *268*, 1144-1149.

(51) Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **1983,** *221*, 709-713.

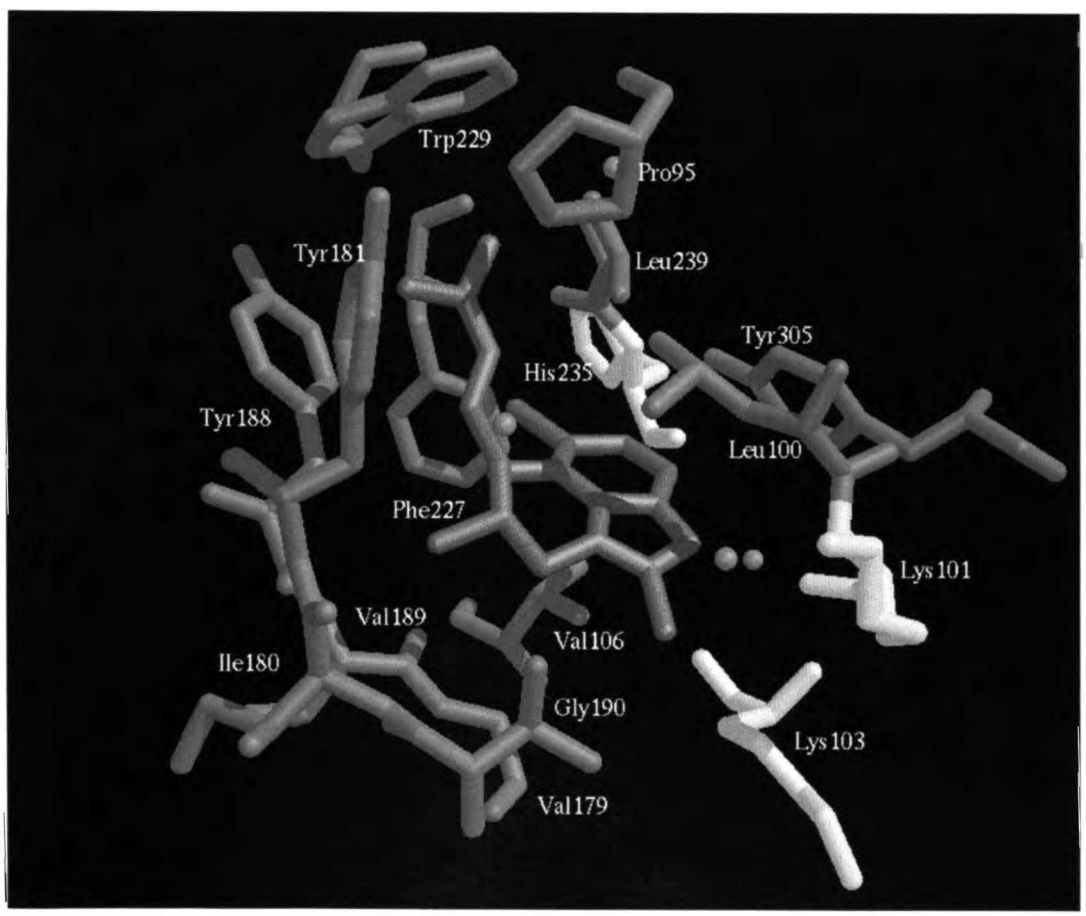
(52) Zacharias, M.; Straatsma, T. P. ; McCammon, J. A. Separation-shifted scaling, a new method for Lennard-Jones interactions in thermodynamic integration. *J. Chem. Phys.* **1994,** *100*, 9025-9031.

(53) Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R. ; van Gunsteren, W. F. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chem. Phys. Lett.* **1994,** *222*, 529-539.

(54) Simmerling, C.; Fox, T. ; Kollman, P. Use of locally enhanced sampling in free energy calculations: Testing and application to the α→β anomerization of glucose. *J. Am. Chem. Soc.* **1998,** *120*, 5771-5782.

(55) Hine, J. ; Mookerjee, P. K. The intrinsic character of organic compounds. Correlations in terms of structural contributions. *J. Org. Chem.* **1975,** *40*, 292-298.

Figure 1


Trp229
Pro95
Tyr181
Leu239
Tyr305
His235
Tyr188
Leu100
Phe227
Lys101
Val189
Val106
Ile180
Gly190
Lys103
Val179

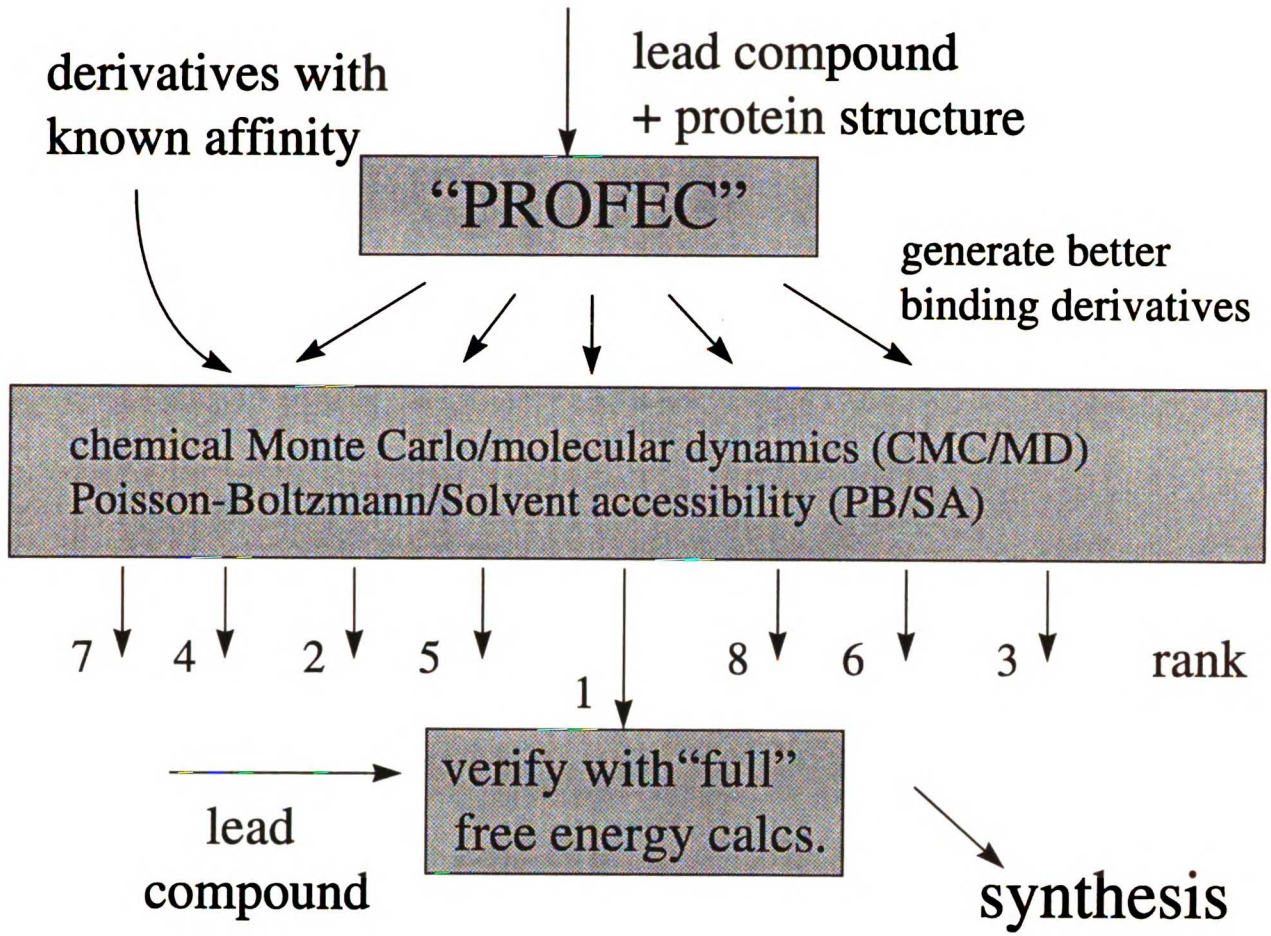

derivatives with
known affinity
lead compound
+ protein structure
"PROFEC"
generate better
binding derivatives
chemical Monte Carlo/molecular dynamics (CMC/MD)
Poisson-Boltzmann/Solvent accessibility (PB/SA)
7
4
2
5
1
8
6
3
rank
verify with"full"
free energy calcs.
lead
compound
synthesis

Figure 3

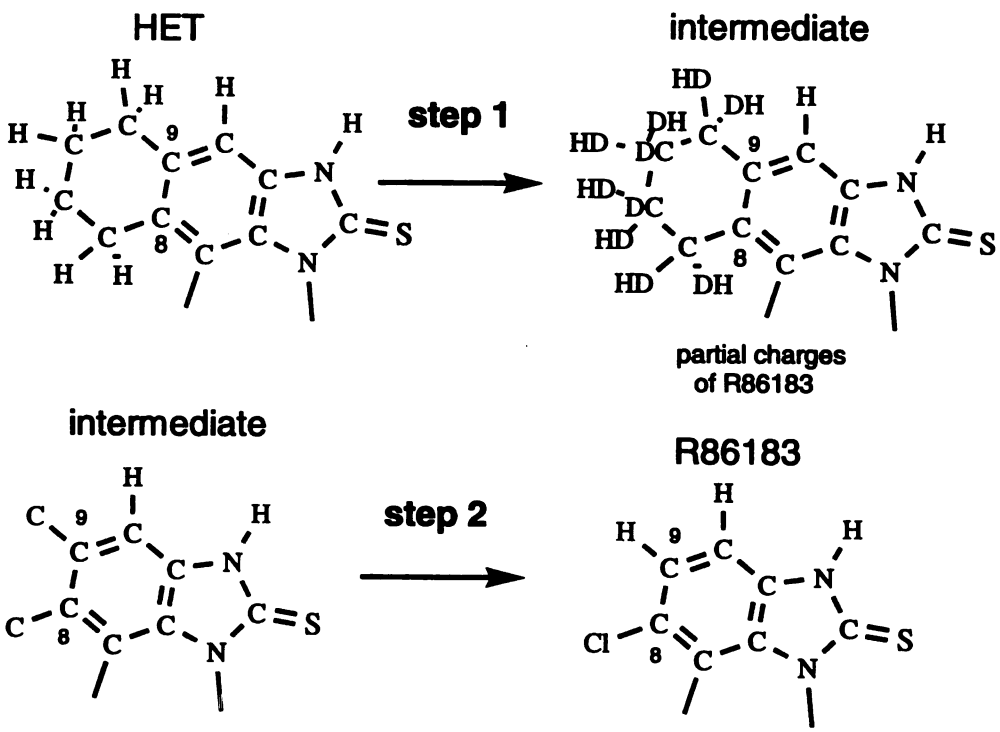

HET
step 1
intermediate
partial charges
of R86183
intermediate
step 2
R86183

Figure 4

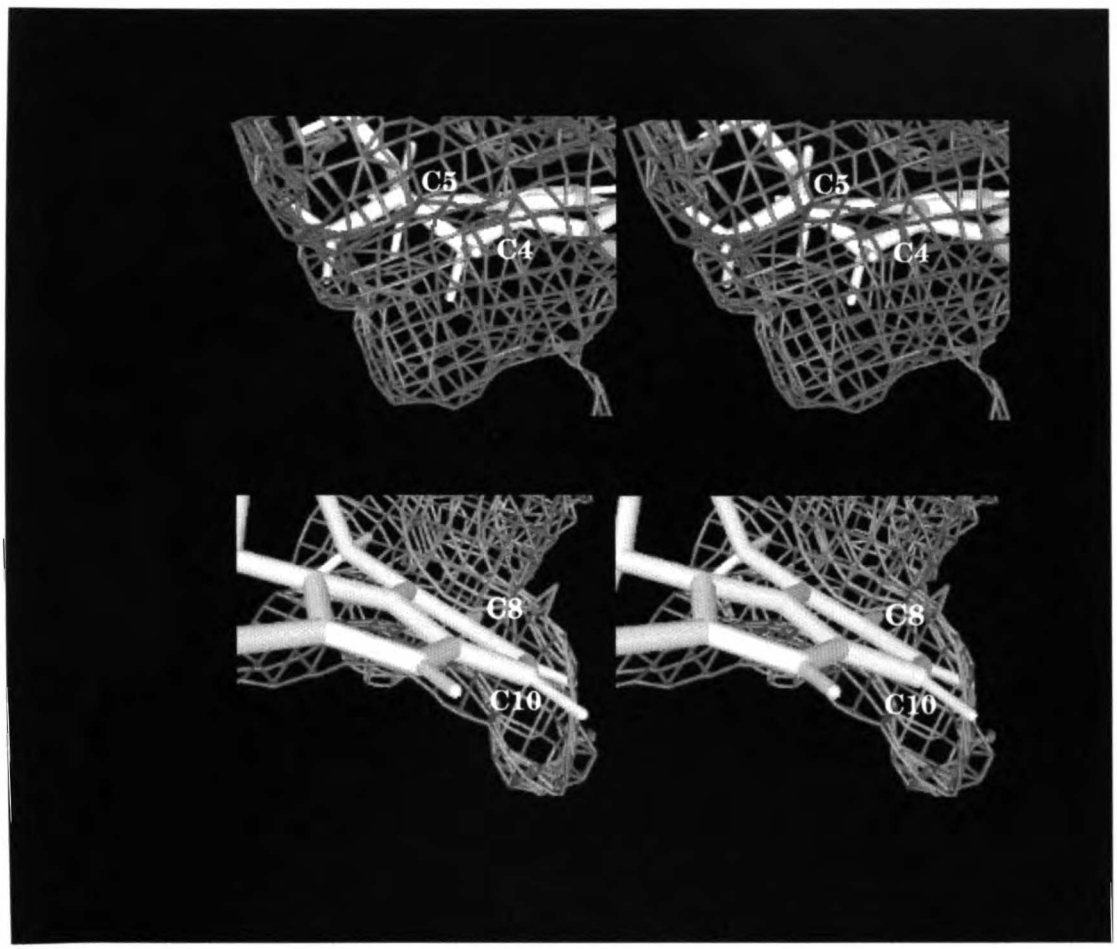
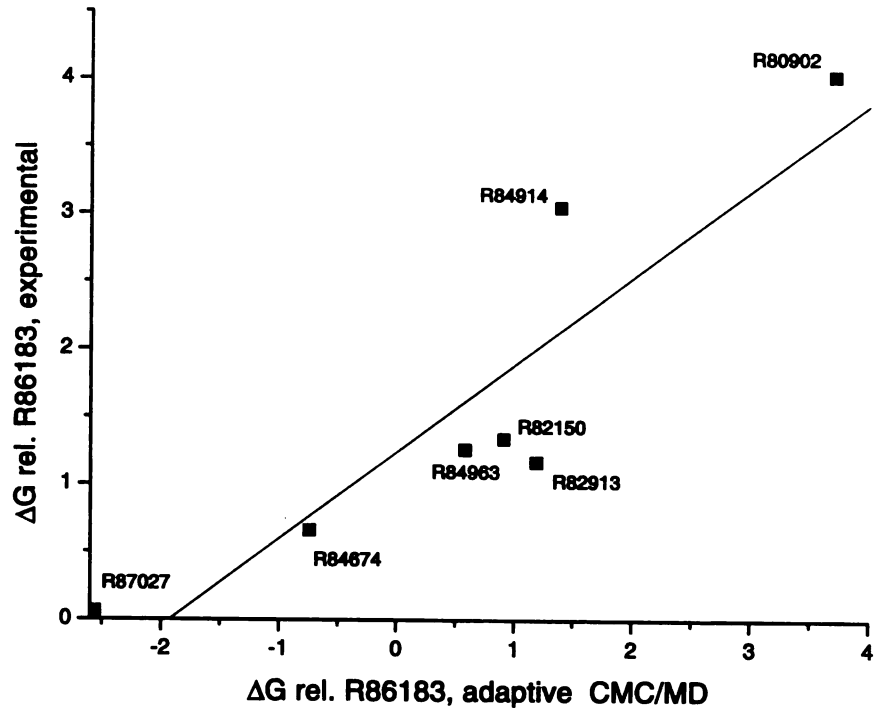


Figure 5


R80902
R84914
R82150
R84963
R82913
R84674
R87027
ΔG rel. R86183, experimental
ΔG rel. R86183, adaptive CMC/MD

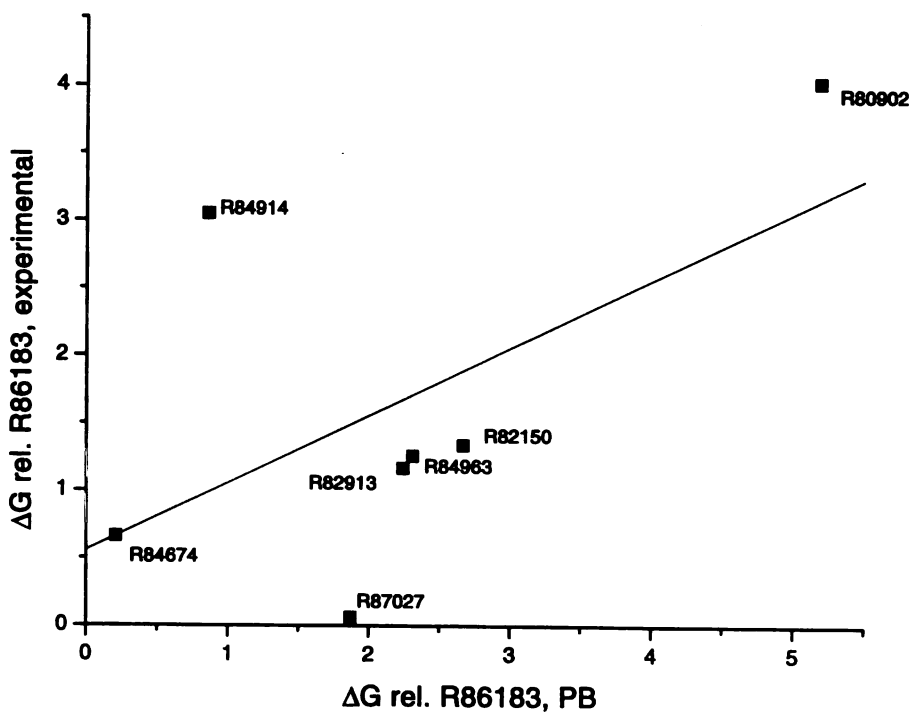
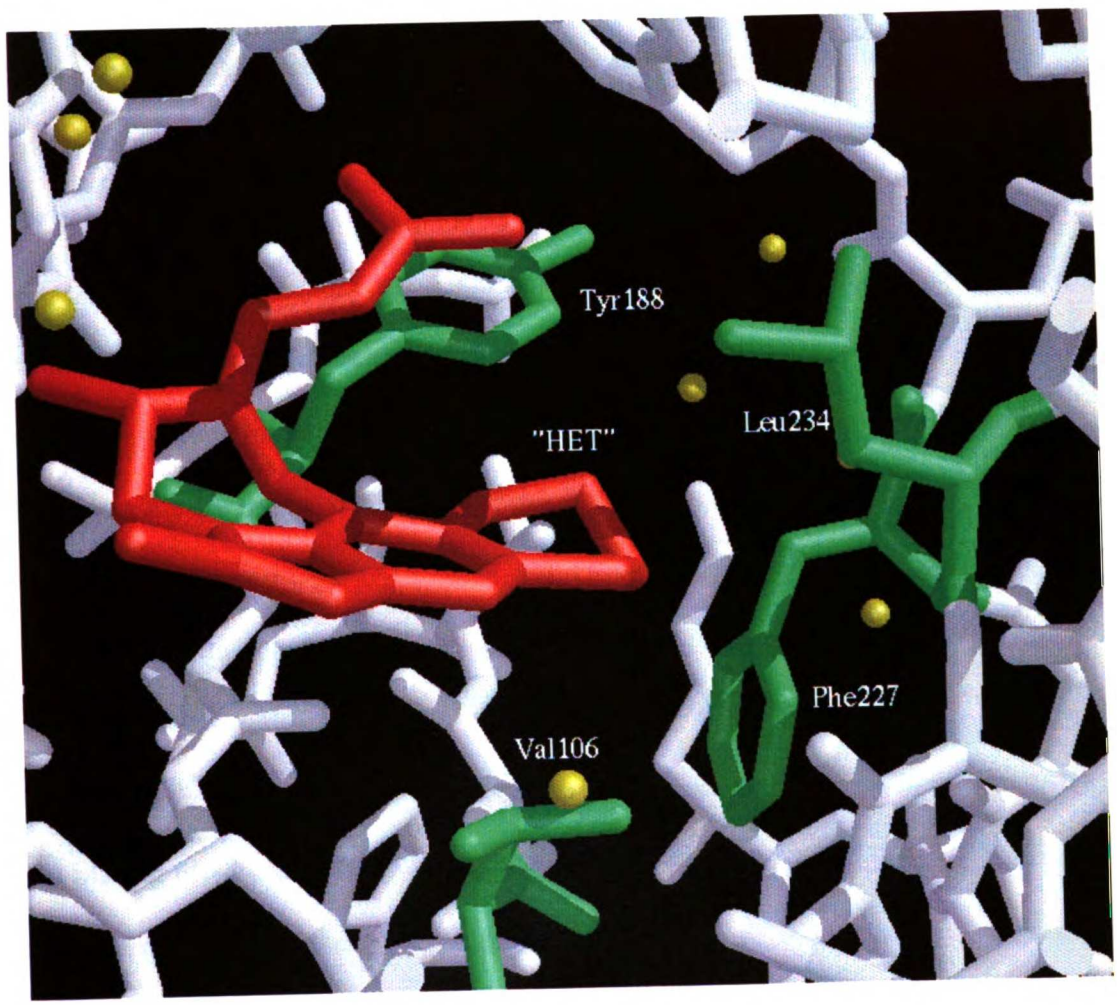

R80902
R84914
R82150
R84963
R82913
R84674
R87027
ΔG rel. R86183, experimental
ΔG rel. R86183, PB

Figure 6



Chapter 6: Adaptive Chemical Monte Carlo/Molecular Dynamics

One weakness of our initial implementation of the CMC/MD method arises directly from the use of observed populations to determine free energy differences. Highly unfavorable species are virtually never sampled, while favorable species dominate the sampling. It is difficult to determine free energy differences greater than 4-5 kcal/mol from practical length simulations. This slow convergence of large free energy differences led us to implement an iterative procedure – adaptive CMC/MD – to allow our calculations to efficiently span a larger range of free energies.

Adaptive CMC/MD makes use of umbrella sampling offsets to decrease the effective free energy difference between states. The offset necessary to eliminate the free energy difference between two states is simply the opposite of that free energy difference. However, we are trying to determine these free energy differences – we do not know them before beginning our calculation. As a consequence, we implemented a simple iterative procedure to determine these offsets over the course of a number of CMC/MD runs.

The first residue is arbitrarily chosen as the reference state, and all free energies and biasing offsets are determined relative to this residue. An initial unbiased CMC/MD run is carried out, and the "Boltzmann" probability of each state is recorded. At the end of the simulation these statistics are used to calculate the relative free energies of each state:

$$\Delta G\ (1 \rightarrow n) = -RT \ln[P(n)/P(1)]$$

These free energies are then used as biasing offsets for the next CMC/MD run:

$$\text{Offset}(n) = -\Delta G\ (1 \rightarrow n)$$

Like the "solvation offsets" used in our host:guest calculation, these biasing offsets are directly applied to the Monte Carlo sampling. In considering a trial move between states I and J, the relevant energy difference is not

$$\Delta E = E(J) - E(I)$$

But

$$\Delta E = (E(J) + \text{Offset}(J)) - (E(I) + \text{Offset}(I))$$

Again, a CMC/MD run is carried out, but this time with biased sampling. The Boltzmann probabilities are again accumulated, but corrected for the inclusion of the biasing offset. This allows statistics from successive runs to be accumulated and added together, rather than discarded. The correct Boltzmann probabilities are calculated without the offset term, unlike the Monte Carlo sampling:

$$P'(I) = \exp[-E(I)/kT]/\{\Sigma(J, J=1 \text{ to } n)\ \exp[-E(J)/kT]\}$$

The probabilities from this run are added to those previously accumulated, and used to calculate new free energies and offsets:

$$P_{total}(I) = P(I) + P'(I)$$

$$\Delta G'(1 \rightarrow n) = -RT \ln[P_{total}(n)/P_{total}(1)]$$

One continues with cycles of biased CMC/MD runs and adjustment of the biasing offsets until the offsets appear to have converged – that is, they do not change significantly between successive runs. This was generally evaluated by running a fixed number of adaptive cycles, and graphically inspecting the resultant free energy offsets.

Ideally, one would continue each individual CMC/MD run for a time significantly longer than the relaxation times of significant processes in the simulated system. In water, this might be 20-40 picoseconds, while 50-100 picoseconds per run might be more appropriate for an enzyme active site. Note that there is a good test for whether the adaptive CMC/MD calculation has converged. Once the adaptive calculation appears to be complete, the final offsets are taken and used in a single long non-adaptive calculation (typically 10-100 times longer than the individual adaptive runs). If the offsets are perfectly converged, the expected uniform distribution of populations will result. Otherwise, the free energy difference between states can be calculated as usual. This value, corrected for the offsets, yields the best determination of the free energy difference.

A graphical depiction of this iterative procedure is shown in Figure 1. The populations or free energy offsets are shown as histograms – observed populations are white bars, free energies are grey bars, and black bars depict the biasing offsets. The cycle starts at point A. The CMC/MD run (biased or unbiased) is point B. C depicts the populations resulting from this run, and the corresponding free energies. Point D shows the conversion of these free energies into biasing offsets to be fed into the next CMC/MD run. Point E shows the ideal final result, a CMC/MD calculation with uniform population distributions and converged offsets.

This iterative and adaptive procedure is related to several other types of free energy calculations. Bennett's original acceptance ratio method determines the free energy difference between two states I and J by determining the offset (C(I<->J)) necessary to yield equal populations of those two states in a CMC/MD-like simulation. In practice, however, acceptance ratio calculations are performed by performing two separate simulations – one of each chemical state. During the simulation of state I, the energy of state J is calculated and recorded. Similarly, during the simulation of state J, the energy of state I is registered. These energies are then used to iteratively solve for the free energy offset. Instead of using a simple exponential weighting to determine C(I<->J),

$$\langle \exp[-\Delta V(I\to J) + C(I<\to J)] \rangle_I = \langle \exp[-\Delta V(J\to I) + C(J<\to I)] \rangle_J$$

a Fermi-Dirac weighting

$$f[x] = 1/(1 + \exp[\beta x])$$

is used:

$$\langle f[\Delta V(I\to J) + C(I<\to J)] \rangle_I = \langle f[\Delta V(J\to I) + C(J<\to I)] \rangle_J$$

This modification means that small values of ΔV contribute more to the estimate of C than large ones, which serves to minimize the estimated error in the calculation C(I<->J) for finite sample sizes. In the adaptive CMC/MD calculation, one determines the free energy offsets directly by observing their effects on the MC sampling. While our calculation is less efficient than the acceptance ratio method for comparing the free energies of two species, it is more readily generalized to the multi-state case we are interested in.

Similar statistical issues drove the development of another free energy method related to our adaptive calculation – the Weighted Histogram Averaging Method, or WHAM[1,2]. WHAM was originally developed in the context of conformational free energy differences or potentials of mean force, and provides a mechanism whereby several separate simulations can be combined to give an accurate estimate of the free energy change along a coordinate. WHAM is necessary since a single simulation can usually only sample over a small range of the coordinate of interest. Thus, a number of simulations dispersed evenly along the coordinate are often necessary to provide complete sampling. Each sub-simulation will give a good estimate of the free energy in the region it samples and a poor estimate for free energy values outside that region. Reconstructing the overall free energy difference by aligning adjacent sub-simulations has the unfortunate side effect of maximizing the error – since errors in each sub-simulation add in the final result.

While WHAM-like techniques have been used to calculate the free energy difference between chemical species [3,4], they are not ideally applicable to our CMC/MD calculation. The species at either end of a conformational coordinate are often

enormously different (helix vs. coil, for instance) – they are separated by a large distance in coordinate space. In contrast, the species at either end of a chemical coordinate (our CMC/MD end states) can be close in coordinate space but generally differ in energy. An accurate comparison of the free energies for a helix and a coil requires consideration of many intervening conformational states – thus the need for a technique like WHAM that samples carefully along a coordinate. Chemical comparisons, at least between related species, are better suited to energy-oriented techniques like acceptance ratio calculations or CMC/MD, neither of which require the simulation of intervening states.

References:

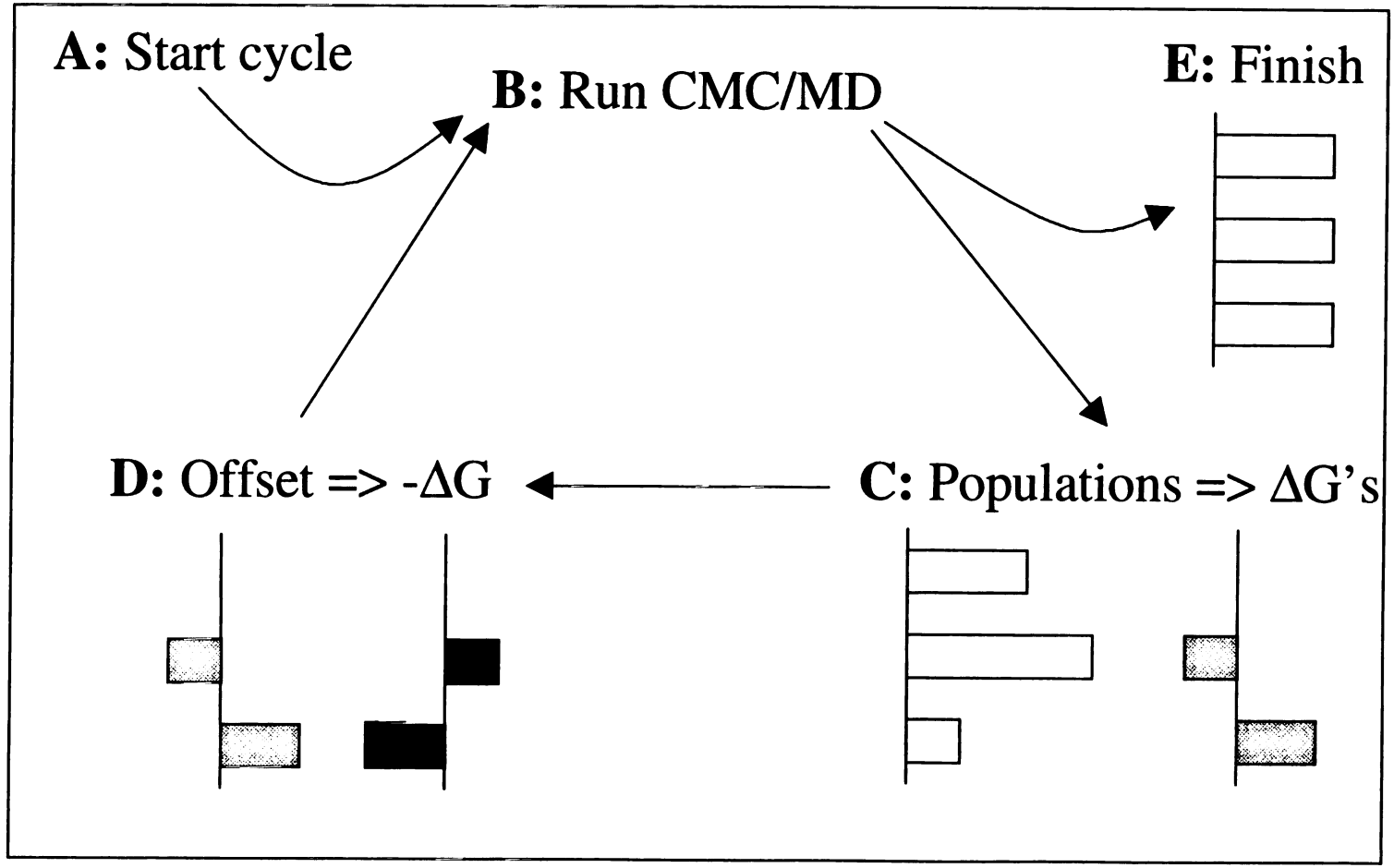
1)Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comp. Chem.* **1992**, *13*, 1011-1021.

2)Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comp. Chem.* **1995**, *16*, 1339-1350.

3)Wilding, N. B.; Muller, M. *J Chem Phys* **1994**, *101*, 4324-4330.

4)Guo, Z.; Brooks, C. L.; Kong, X. *J Phys Chem B* **1998**, *102*, 2032-2036.

Figure 1


A: Start cycle
B: Run CMC/MD
E: Finish
D: Offset => -ΔG
C: Populations => ΔG's

Chapter 7: Exhaustive Mutagenesis *in silico*: multicoordinate free energy calculations on proteins and peptides

Jed W. Pitera[1] and Peter A. Kollman[1,2,*]

[1]Graduate Group in Biophysics and

[2]Department of Pharmaceutical Chemistry,

University of California – San Francisco

San Francisco, CA 94143-0446, USA

e-mail: pak@cgl.ucsf.edu

*: to whom correspondence should be addressed

Submitted for publication, December 1998.

Abstract

We have extended and applied a multicoordinate free energy method, chemical Monte Carlo/Molecular Dynamics (CMC/MD) to calculate the relative free energies of different amino acid side chains. CMC/MD allows the calculation of the relative free energies for many chemical species from a single free energy calculation. We have previously shown its utility in host:guest chemistry[1] and ligand design[2], and here demonstrate its utility in calculations of amino acid properties and protein stability.

We first study the relative solvation free energies of N-methylated and acetylated alanine, valine, and serine amino acids. With careful inclusion of rotameric states, internal energies, and both the solution and vacuum states of the calculation, we calculate relative solvation free energies in good agreement with thermodynamic integration (TI) calculations. Interestingly, we find that a significant amount of the unfavorable solvation of valine seen in prior work[3] is caused by restraining the peptide in an unfavorable extended conformation. In contrast, the solvation free energy of serine is calculated to be less favorable than expected from experiment, due to the formation of a favorable intramolecular hydrogen bond in the vacuum state. Our peptide calculations emphasize the need to accurately consider all significant conformations of flexible molecules in free energy calculations.

This development of the CMC/MD method paves the way for computations of protein stability analogous to the biochemical technique of "exhaustive mutagenesis". We have carried out just such a calculation at position 133 of T4 lysozyme, where we use CMC/MD to calculate the relative stability of eight different side chain mutants in a

single free energy calculation. Our T4 calculations show good agreement with the prior free energy calculations of Veenstra, et. al.[4] and excellent agreement with the experiments of Mendel, et. al.[5]

Introduction

Following Wolfenden's[6] pioneering experimental studies on the solubility of protein side chain analogs, and the recognition of the importance of hydrophobic contributions to protein folding and function[7], there has been substantial interest in the solvation free energies of amino acids and the importance of these solvation properties in determining overall protein stabilities. Computational techniques, especially free energy calculations, have been used to provide insight into the microscopic picture of amino acid solvation. In particular, Sun, et. al.[3] studied the relative solvation of blocked alanine and valine molecules and pointed out the crucial role of the backbone amide groups in the aqueous solvation of these compounds. As others have noted[8], accurate prediction of amino acid solvation free energies is one of the first steps toward prediction of protein stability from free energy calculations. We have developed and implemented the necessary techniques to apply a multi-coordinate free energy method, chemical-Monte Carlo/Molecular Dynamics (CMC/MD) to free energy calculations on protein and peptide systems. In this paper we describe the application of CMC/MD to both types of systems; the acetylated and N-methylated alanine, valine and serine peptides, and eight mutants of the T4 lysozyme protein. When contributions from rotamers, internal energies, and both reference states are adequately represented, we calculate relative free energies in good agreement with traditional free energy calculations or experiment.

Figure 1 shows the thermodynamic cycle used for the peptide calculations. The relative solvation free energy of two compounds ($\Delta\Delta G_{sol}$) can be directly determined from experiments that measure the free energy cost of transferring each compound from

the gas phase to water (ΔG1, ΔG2). With free energy calculations, however, ΔΔGsol is calculated by determining the work necessary to convert one molecule to another in vacuum (ΔGvac) and solvent (ΔGsol). Since the free energy is a state function, one can therefore determine the relative free energies of solvation (ΔΔGsol):

$$\Delta\,\Delta Gsol = \Delta G2 - \Delta G1 = \Delta Gsol - \Delta Gvac$$

The same thermodynamic cycle can be used for the calculation of relative solvation free energies by both traditional free energy calculation methods (FEP, TI) and CMC/MD. However, one of the problems common to all microscopic free energy methods is the need for adequate conformational sampling. In both traditional free energy methods and multi-coordinate methods, the free energy is calculated as the ensemble average of a function. For free energy perturbation, this is

$$\langle e^{-\Delta V/RT} \rangle$$

whereas for thermodynamic integration it is the ensemble average of the derivative of the potential:

$$\langle \Delta V/\Delta\lambda \rangle$$

For CMC/MD, it is the ensemble average of the Metropolis[9] transition probability between each chemical state:

<P(A => B)>, <P(B => A)>

Both molecular dynamics and Monte Carlo techniques are commonly used to generate the ensemble averages listed above. The degree of sampling required for accurate free energy estimates is often extensive[10]. Molecular dynamics, in particular, has difficulty in sampling between conformations that are separated by appreciable free energy barriers (ca. 3-4 kcal/mol), often causing substantial difficulties in systems with multiple rotatable bonds. The locally enhanced sampling (LES) method of Elber[11] has been applied to surmount this difficulty in calculating the anomeric effect in glucose[12]. A flattening of the free energy surface by a perturbation to the simulated potential and subsequent correction has been explored by Ota and Brunger[13]. Straatsma and McCammon have suggested a multistep method that first calculates the free energy cost for constraining the geometry of the system, then the free energy of the chemical perturbation, followed by the free energy cost of releasing the constraint[14].

However, the ability of Monte Carlo techniques to sample between states separated by significant barriers provides an alternative approach to solve the problem. To jump between two separated states via Monte Carlo, that transition has to be included in the "move set" used in the calculation. Still, et. al.[15] have used this to dramatically improve sampling in the MC(JBW) methods, where molecular or stochastic dynamics are used to sample within wells, while Monte Carlo steps are occasionally used to shift the simulated molecule between wells. MC(JBW) methods require an initial scan of the potential energy surface to enumerate the "wells" to be sampled between. In our

calculations on amino acids, we have decided to include a limited form of this approach in order to properly sample the rotamers of each amino acid. The use of rotamers provides us with a defined move set for each amino acid. In this respect, our work is similar to the "Boltzmann ensemble" simulated by Huber, et al[16] where multiple conformations of a molecule are enumerated and simulated in a "mixed state" where each conformation is populated according to its Boltzmann weight. Our peptide CMC/MD differs from these prior two approaches since it allows significant sampling in both conformational and chemical space, permitting accurate free energy calculations on peptides and proteins.

The need to accurately include each rotamer conformation in free energy calculations on amino acids has been emphasized by Wilson, et al[17] as well as Hermans[18, 19] in the case of exhaustive simulations. As a simple example, if valine is constrained to one of its three rotamers when it is transferred from vacuum to solvent, there is a corresponding free energy penalty of

$$-RT \ln 3 = 0.6 \text{ kcal/mol}$$

For side chains with several rotamers (leu, met, ile, etc.), this penalty can be as large as 2-3 kcal/mol. Clearly, in calculations of amino acid solvation or protein stability, this is a crucial contribution. It must be recognized that this is but a small part of the overwhelming conformational sampling problems associated with protein stability calculations[8] though qualitative[20] and quantitative[4] free energy calculations have been reported. It is also interesting to note that calculations with exhaustive rotamer and

side-chain sampling (but a single backbone conformation) have successfully been used to design protein sequences that stably adopt a given fold[21]. The method we describe herein is robust and applicable to any chemical species for which conformational heterogeneity is an issue.

Our intent in this paper is to describe the assembly and application of a multi-coordinate free energy method to calculations on amino acids. With CMC/MD, we have enhanced both the conformational and chemical sampling relative to traditional free energy methods, allowing us to carry out single calculations that compare many amino acid side chains at a given position. Like the analogy between multicoordinate binding free energy calculations and "competitive binding" experiments[22], our CMC/MD protein stability calculation can be compared to the biochemical technique of "exhaustive mutagenesis". In these experiments, all twenty amino acids are inserted at a particular position in the peptide chain, and their effects on protein properties observed.

We have successfully carried out just such a calculation, comparing eight different residues (alanine, ethylglycine, valine, norvaline, O-methyl serine, leucine, isoleucine, and phenylalanine) at position 133 of T4 lysozyme. This system was chosen due to the wealth of thermodynamic, mutational, and structural data available[5, 23, 24], making it an ideal test case. Several crystal structures of position 133 mutants have been solved, and they show only small readjustments in the hydrophobic residues near the mutated side chain. The successful free energy calculations of Veenstra, et. al.[4] and Wang, et. al.[25] also suggested that T4 lysozyme was a reasonable system for our calculations. We chose a panel of mutations that vary substantially in size (alanine ⇔ phenylalanine), chemistry (O-methyl serine ⇔ norvaline) and branching (valine ⇔

norvaline, isoleucine ⇔ leucine). By allowing us to calculate relative free energies for many more side chains at once, CMC/MD also permits extensive and specific insight into the balance of chemical forces that influence side chain contributions to protein stability. Our protein stability calculations used the thermodynamic cycle shown in Figure 2, calculating the relative stability (ΔΔGfold) of two mutant proteins by comparing the free energy difference of the mutants in the folded protein (ΔGprot) and in a model of the unfolded state (ΔGwat).

Theory

The chemical-Monte Carlo/Molecular Dynamics method has been described previously[1]. It is based on a derivation by Bennett[26], where he observed that Metropolis Monte Carlo steps between different chemical states of a system (alanine and valine, for example), coupled with some coordinate sampling, would populate those states in proportion to their relative free energies (alanine being preferred in water). In this paper, we have modified the CMC/MD calculation somewhat to include both internal energies and rotameric states, crucial issues for amino acid properties.

In our initial host:guest and protein:ligand calculations, one copy of each ligand of interest was simulated in the binding site. At any point in time, the potential function is masked so that one copy is "real" and interacts with the surroundings. The other copies are "ghosts", and they do not interact with the surrounding binding cavity or solvent. Chemical Monte Carlo moves consist of switching which ligands are "ghosts" and which ligands are "real". The fraction of time each ligand is real is related to its relative free energy, by

$$\Delta G(a->b) = -RT \ln [(P(b)/P(a)]$$

For such Monte Carlo methods to work, trial moves must be selected from a uniform distribution over the states of interest. For a system of N chemical states, the probability of any single state being selected for a trial move is

$$P(\text{trial move} => j) = 1/N$$

However, when we add rotamers for some of the chemical states the assignment of trial moves is more complicated. If each species has f(n) rotamers (one for alanine, three for valine, etc.), the probability of a trial move to a particular rotamer becomes

$$P(\text{trial move} => j, \text{rotamer } k) = 1/N * 1/f(n)$$

This renormalization allows for the proper uniform distribution of trial moves, while including conformational heterogeneity for those states where it is necessary.

Another issue in the CMC/MD calculation that is raised by amino acids is the need to include internal energies in the calculated free energy. For our ligand calculations, there was no covalent connection between the Monte Carlo residues and the protein surroundings. For our peptide work, and subsequent calculations on protein side chains, there are covalent connections between the Monte Carlo residues and the prior and subsequent residues of the polymer. In the case of the present systems, each residue

copy is bound to both the N- and C-terminal residues (Figure 3a). While this topology allows for the possibility of protein stability calculations, there is now a need to include internal terms in the free energy calculation – specifically, all bonds, angles, dihedrals, and 1-4 nonbonded terms can now be included in the calculated free energy. In our initial host:guest studies, restraints were necessary to hold the "ghost" ligands in the binding cavity of the host. In a polymer, however, the covalent bonds between Monte Carlo and flanking regions serve to restrain the Monte Carlo residues in appropriate positions and conformations. As these constraints are consistent between both vacuum and solvent states of the calculation, they do not appear in the solvation free energy we determine.

We have modified the CMC/MD software so that internal energy contributions can be included in the calculation in two ways. First, we added the option of including internal energies in the calculated free energy. This is analogous to the inclusion of all "intra-perturbed group" contributions in traditional free energy calculations[27]. Second, there is an option to adjust the forces felt by the "ghost" Monte Carlo residues. While the ghost residues never exert forces on the neighboring residues or on the solvent, the current software has the option to separately allow the ghosts to feel either the internal forces and/or the nonbonded forces resulting from these surroundings. These forces are masked on a per-atom basis, ensuring that they only affect atoms in the Monte Carlo residues.

Computational Details and Methods

Peptide Calculation

Our peptide CMC/MD calculation was carried out with a modified version of the SANDER program from AMBER 5.0[28]. A 1 femtosecond timestep was used, along with the SHAKE algorithm[29] to constrain bonds to hydrogen atoms. Standard Cornell, et. al.[30] parameters were used for the amino acids and blocking groups, and the TIP3P[31] water model was employed. Several different calculations were carried out, using different combinations of free energy calculation and conformational restraints. For all calculations, the side chain $\chi 1$ torsions for each rotamer of valine and serine were restrained using flat-well torsional restraints to an angle of 60, 180 or 300 degrees with a well width of +/- 40.0 degrees and a force constant of 50 kcal/mol. For the first two calculations ("CMC/MD-1","CMC/MD-2a"), the peptide was maintained in an extended conformation by 50 kcal/mol harmonic torsional restrains that kept phi and psi at –160 and 160 degrees, respectively. In our final calculation, this backbone restraint was relaxed to a very wide flat-well: using the same 50 kcal/mol force constant, phi was allowed to range from -175 to -5 degrees without restraint, and psi was permitted to sample freely between 5 and 175 degrees ("CMC/MD-2b"). In all these calculations, Monte Carlo steps were carried out every 100 dynamics steps, and Andersen (stochastic) temperature coupling[32] was used with a periodicity of 5 picoseconds.

In order to develop a methodology that was extensible to proteins and other polymers, a complex topology was constructed for our amino acid "solute" (Figure 3a). The Monte Carlo region consisted of one alanine residue, three valine residues (each restrained to a different rotamer) and three serine residues (likewise restrained). This region was flanked by the methyl and acetyl blocking groups. Each of our Monte Carlo residues was bonded to both the N-terminal and C-terminal blocking groups. These

bonds, as well as all angles, dihedrals and 1-4 van der Waals interactions were masked from the simulated potential as described above. As can be seen from Figure 3a, this topology is readily extensible to study a residue within a peptide chain, where the Monte Carlo region may be flanked by N- and C- terminal amino acids instead of blocking groups. Since there were three copies each of valine and serine, and one alanine, each valine and serine copy was chosen for a trial move 1/9 of the time, versus 1/3 of the time for the sole alanine (Figure 3b).

As mentioned above, three sets of CMC/MD calculations were carried out. All included a gas phase and a solvent leg, corresponding to the scheme(s) in Figure 1. However, the first set of calculations did not include contributions from intra-perturbed group energies (bonds, angles, dihedral and nonbonded interactions within each amino acid residue) much like the prior free energy calculations of Sun, et. al[3]. This set is denoted in our Tables and Figures as "CMC/MD-1." The second set included all intra-perturbed group energies in the calculated free energies, resulting in a substantially greater magnitude in the free energies calculated for each leg and concomitant convergence problems. With this protocol, we also evaluated two different types of backbone restraints. The "CMC/MD-2a" calculation restrained the peptide backbone to (-160,160) with harmonic restraints, while the "CMC/MD-2b" calculation allowed the backbone some degree of conformational freedom using the flat-well restraints described above.

To improve the statistics of the calculated free energies and minimize the total real time necessary for our calculations, we used the computational strategy schematized in Figure 5. An initial iterative calculation is run in the "reference state." Specifically,

we took the "reference state" to be whichever leg of the thermodynamic cycle (Figure 1, Figure 2) contained fewer atoms and was less computationally expensive to simulate. This was the vacuum state for our peptide calculations, and the unfolded state for our T4 lysozyme simulations. In this first adaptive phase, several (10-100) short (10-25ps) simulations were run to get an initial estimate of the relative free energy of each species. After each short simulation, the current estimate of the relative free energy of each species was calculated. These values were then used as an umbrella sampling[33] biasing potential for the next simulation cycle. In this fashion, the sampling is gradually forced toward a uniform distribution over the chemical species – once this is achieved, the biasing potentials reflect the relative free energies of each state. The "adaptive CMC/MD" procedure is equivalent to the iterative WHAM[34] algorithm for the calculation of conformational free energy differences, and was successfully used in our CMC/MD calculations on HIV Reverse Transcriptase inhibitors[2].

These adaptive estimates were used to bias the sampling of two separate simulations in the second phase of the calculation. Here, we carried out long, non-iterative CMC/MD runs (~0.5-10 ns) for each leg of the thermodynamic cycle. The biasing potential was then subtracted from the results of these second calculations to yield the final calculated values for each leg of the thermodynamic cycle. The difference of these final values yields the relative free energy of interest.

There are two major rationales for the strategy described above. First, the inclusion of intramolecular terms in some of our calculations meant that the relative free energies along each leg of the thermodynamic cycle were often large and of the same sign (Table 2) in both phases, representing significant intramolecular contributions. For

example, valine has a number of favorable 1-4 nonbonded interactions that are not present in alanine, regardless of whether it is in the gas phase or vacuum. Calculating an estimate of the free energy for one leg, then, provided a good "first guess" for the free energy difference in the second leg, and allowed the noniterative CMC/MD calculations to more accurately reflect the environmental effects (vacuum, water, protein) of each state. Secondly, this allows the two long, non-iterative calculations to be run in parallel, minimizing the total real-world time necessary to calculate an accurate relative free energy. The biasing potentials used are of the exact same form as our prior calculations of relative binding free energies ($\Delta\Delta$Gbind)[1].

Calculations were carried out on either a 275 MHz DEC Alpha workstation, a four-processor SGI Origin 200, or 4-8 processors of a Convex Exemplar X-class. The parallel calculations used a message passing interface (MPI) implementation of the code. Long simulation times (~5-10 nanoseconds) were required to adequately converge the calculated free energies in solvent, while the vacuum calculations were appreciably converged in 200ps when intra-perturbed group contributions were not included. As noted above, the inclusion of intra-perturbed group contributions substantially increases the free energy differences in both legs (in vacuum, ala => val goes from +1.5 kcal/mol to –14.1 kcal/mol) of the calculation, requiring extensive (>1 nanosecond) calculations to yield converged results.

For comparison, vacuum-state thermodynamic integration (TI) calculations were carried out using the GIBBS module of AMBER 4.1, using identical simulation parameters to the corresponding CMC/MD calculation. The perturbations were not exactly analogous to the CMC/MD calculation since they used a single rather than

multiple topology protocol[27]. To calculate the contributions from each rotamer, and the inter-rotamer free energies, alanine was separately perturbed to each valine or serine rotamer while the $\chi 1$ torsion was restrained with a flat-well restraint identical to that used in the CMC/MD calculation. The TI calculation was carried out with 25 windows each consisting of 3 picoseconds each of equilibration and data collection. The calculated free energies for each rotamer were Boltzmann weighted[4, 18] to yield the overall free energy difference between alanine and valine or alanine and serine. For example,

$$\Delta Gvac\ (ala->ser) = -RT\ \ln\ [\Sigma_{j=1,2,3}\ (1/3 * e^{-\Delta G(ala->ser,j)/RT})]$$

T4 calculation

T4 lysozyme was simulated with eight different amino acids at position 133; alanine (ALA), ethylglycine (ETH), valine (VAL), norvaline (NVL), O-methylserine (MSE), leucine (LEU), isoleucine (ILE) and phenylalanine (PHE). For all residues except alanine, each rotamer was included in the calculation. A total of forty-nine different side chains, varying in both chemistry and conformation, were used in this CMC/MD calculation. The parameters used were either from the Cornell, et. al. force field[30] or those used in prior free energy calculations[4] (NVL, MSE) and molecular dynamics modeling[35] (ETH) on T4 lysozyme.

Starting from the x-ray crystal structure of wild-type (L133) T4 lysozyme[24], each CMC/MD residue was modeled in by hand using the MidasPlus molecular modeling software. Like the blocked dipeptide, the CMC/MD residues were superimposed and each covalently bonded to the N- (D132) and C-terminal (A134) residues of the amino

acid chain. Each side chain was fixed in its rotamer via flat-well torsional restraints, as before. For the T4 lysozyme calculation, we used wells of +/- 30 degrees that were harmonic between 30 and 50 degrees and linear for torsions beyond 50 degrees, with a force constant of 50.0 kcal/mol*A^2. No backbone torsional restraints were used in the protein.

Once the model was built, it was solvated with the crystallographically observed water molecules plus a 17.0 angstrom spherical cap centered on alanine 133, using the TIP3P water model. As with prior calculations, neutralizing counterions were placed next to solvent-exposed charged residues. Single chloride ions were placed near R8, R14, K16, K19, K35, K43, K65, K85, R96, R119, K124, K135, and K147. Sodium ions were placed near D20, D47, D61, and the terminal carboxyl group of L164. The total system was 4770 atoms in size, including 433 water molecules and the 49 Monte Carlo residues. Following the protocol established by Veenstra and Kollman, we allowed the entire simulated system to move during molecular dynamics, but restrained the backbone (CA, N, C) atoms of residues distant from the cavity to their crystallographic positions with harmonic restraints of 10.0 kcal/mol*A^2. This corresponds to the "cavity-constraints" protocol of the prior work. Figure 6 shows position 133 in the protein, and depicts the ALA and LEU residues as they are modeled in at this position.

Also following Veenstra and Kollman, the unfolded state of T4 lysozyme was modeled as a blocked, solvated pentapeptide. This model was shown to represent the unfolded state of the molecule sufficiently well enough to allow thermodynamic integration (TI) calculations in good agreement with experiment. The "true" unfolded state environment of residue 133 is expected to be a mix of solvent water and transient,

weakly-structured contacts with other protein residues, so we might expect our extended peptide model to exaggerate the solvent exposure in the unfolded state. This is somewhat balanced, however, by the limited sampling possible in our folded-state model, which does not necessarily permit sampling of all low-energy protein conformations for each mutant. The CMC/MD residue coordinates from our protein model, plus one flanking residue on either side, were used to build a blocked peptide model (ACE-ASN-XXX-ALA-NME, where XXX denotes the CMC/MD residues). Again, the side chain coordinates were restrained with flat well torsional restraints of +/- 30 degrees, 50.0 kcal/mol*A^2. In addition, the ϕ and ψ angles of the CMC/MD residues were restrained with flat well torsional restraints. ϕ was unrestrained between 185 and 355 degrees, with harmonic restraints from 185 to 170 and from 355 to 370 degrees. The ψ torsion angle was unrestrained from 5 to 175 degrees, with harmonic restraints from –10 to 5 degrees and from 175 to 190 degrees. Outside of these regions, the restraints were linear. Like the side chain restraints, a force constant of 50.0 kcal/mol*A^2 was used. This peptide model was solvated with a periodic box of 1165 TIP3P water molecules, forming a total system of 4360 atoms.

For both T4 simulations, a 1 femtosecond (1 fs) timestep was used in conjunction with SHAKE to constrain the length of all bonds containing hydrogen. MC steps occurred every 100 MD steps. The temperature of the system was kept at 300K with Andersen temperature coupling at a frequency of 2500 MD steps. For the iterative calculation, the biasing offsets were adjusted every 25 picoseconds, and their convergence was monitored graphically. The unfolded state was simulated at a constant pressure of 1 atmosphere using Berendsen pressure coupling[36] with a time constant of

0.2 ps. Since the folded state was simulated as a finite system, no pressure coupling was used. Nonbonded interactions were truncated with an 8 Angstrom cutoff.

The overall strategy described in Figure 5 was used for these calculations as well. An initial 2.0 ns of adaptive CMC/MD (80 iterations of 25 ps each) was carried out with the unfolded state model. The biasing offsets from this calculation were then used for extensive non-adaptive calculations in both the folded and unfolded states. This required 1.0 ns of simulation for the folded state and 4.0 ns of simulation for the unfolded state. Two separate trajectories using different initial velocities and random number seeds were run for each state and the free energies reported are the average of those two trajectories. The results of these non-adaptive calculations were then used to calculate the relative free energies of folding for each residue.

Results

Peptide

Figures 6 and 7 show the results of our vacuum calculations for the peptide model, where Figure 6 depicts the convergence of the adaptive phase. The non-adaptive vacuum calculations are shown in Figure 7. The adaptive calculation takes a very long time (14 ns) to achieve stable values when the intra-perturbed groups are included (CMC/MD-2b, Figure 6b). Graph 7a shows the relative free energy of valine and serine versus alanine when the intraperturbed group terms are not included in the calculation. Graph 7b is identical except that here we show the results of calculations where the intraperturbed group contributions are included. Here the agreement between CMC/MD

and TI calculations (Table 1) is not as good, but that is expected given the large magnitude of the free energy changes involved and the topological difference in the calculations. Population-based free energy calculations like CMC/MD necessarily have problems with large free energy differences, since a free energy difference of 20 kcal/mol corresponds to a population ratio of less than 1 in 10^14. The judicious use of umbrella sampling biasing potentials (Figure 5) permitted the convergence of the calculation shown in Graph 6b, though it is still slow.

Figure 8 shows the convergence of our solvent calculations, again showing the simulations without (CMC/MD-1, Figure 8a) and with (CMC/MD-2b, Figure 8b) intra-perturbed group contributions and flat-well backbone restraints. As expected, these calculations require much more simulation time to converge, with free energies only reaching stable values after nanoseconds of simulation. The final free energy values (ΔGvac, ΔGsol) and corresponding relative free energies of solvation ($\Delta\Delta$Gsolv) are reported in Table 1, which includes reference values from both thermodynamic integration free energy calculations and from the experiments of Wolfenden, et. al[6]. The relative free energies of the solvated state were also extracted and reported separately in Table 1. Inclusion of the intramolecular energy terms in the calculated free energy makes valine and serine both significantly more favorable than alanine in either phase, as noted above.

While the inclusion of rotamers for valine and serine was necessary to calculate accurate relative free energies, our approach also allows us to evaluate the relative free energies of each rotamer for a particular side chain. Table 2 shows the relative free energy of each valine and serine rotamer in vacuum, as calculated by both CMC/MD and

TI. The values are in good agreement, while the inter-residue free energies differ more substantially between our calculations. These inter-rotamer free energy differences can be calculated directly from the CMC/MD calculation, by comparing the relative probability of observing different rotamers of the same side chain. For the TI calculation, however, 2 separate simulations were required to determine each value in Table 2. First, the free energy difference between alanine and the side chain (valine or serine) restrained to $\chi 1 = 60$ degrees was determined. The second calculation determined the free energy difference between alanine and an alternative rotamer ($\chi 1 = 180$ or 300) of the side chain. The free energy difference for the two rotamers can then be calculated as

$$\Delta G\ (V60 \Rightarrow V180) = \Delta G(V60 \Rightarrow A) - \Delta G(V180 \Rightarrow A)$$

Both the CMC/MD and TI calculations detailed in Table 2 show that the inter-rotamer free energy differences in vacuum are significant (greater than 0.5 kcal/mol) but small enough that several rotamers are significantly populated for each side chain.

As noted, one interesting advantage of the CMC/MD calculation is that it is very straightforward to observe which side chain rotamers are populated in a given environment, providing detailed insight into the complex mechanism of amino acid solvation. Table 3 shows the normalized populations of valine and serine rotamers in vacuum and in solution, from each CMC/MD calculation. For CMC/MD-1 in vacuum, valine populates mostly the $\chi 1 = 60$ and $\chi 1 = 300$ rotamers, which make favorable van der Waals contacts with the peptide backbone. The "trans" ($\chi 1 = 180$) rotamer is less favorable in vacuum, since the methyl groups are maximally distant from the peptide

backbone. In water, these populations shift substantially, and the $\chi 1 = 300$ rotamer is the only populated species. Interestingly, Sun, et. al.[3]'s TI calculation on ALA ⇔ VAL in water sampled only the $\chi 1 = 180$ "trans " rotamer and found valine disfavored by 1.14 +/- 0.05 kcal/mol. However, both our vacuum TI and CMC/MD-1 calculations confirm that this is the least favorable valine rotamer. By indirectly excluding the more favorable $\chi 1 = 60$ and $\chi 1 = 300$ rotamers from their calculation, Sun, et. al. appear to have overestimated the free energy for valine. Even when intramolecular terms are ignored, the vacuum state also makes a significant contribution to the relative free energy of solvation, another contribution neglected by the prior TI calculation.

When intramolecular terms are included (CMC/MD-2a), we calculate that the solvation free energies of alanine and valine differ by 1.2 kcal/mol in favor of alanine. This is analogous to the preference of 1.14 kcal/mol observed above. However, the "trans" rotamer is still the least populated species. More importantly, this preference for alanine is substantially reduced if the peptide backbone is not held in an extended conformation (CMC/MD-2b). Allowing the backbone to relax moves the free energy difference closer to its expected value (+0.5 kcal/mol vs. the calculated value of +0.3 kcal/mol for Me => Prp). This relaxation also shifts the rotamer preferences to favor $\chi 1 = 300$ (and, to a lesser extent $\chi 1 = 180$) both in the gas phase and in solution. The data from CMC/MD-2a and CMC/MD-2b suggest that the strong preference for alanine seen by Sun, et. al., is primarily the result of the relatively unfavorable extended backbone conformation used in their calculations.

The rotamer preferences for serine are smaller, but still interesting. Table 1 shows that serine is always preferred to alanine in the gas phase. This is due to the ability

of the serine hydroxyl group to make a strong intramolecular hydrogen bond to either the carbonyl of the N-acetyl blocking group or the carbonyl group of the serine itself. In our first calculation, CMC/MD-1, we see a strong hydrogen bond between the $\chi 1 = 300$ rotamer and the N-acetyl group's carbonyl. The dominance of this rotamer decreases somewhat when intramolecular interactions are included (CMC/MD-2a), since some strain is induced by the formation of the hydrogen bond. In solution, the contribution of an intramolecular hydrogen bond is insignificant given the preponderance of intermolecular hydrogen bonds between the solute and the water solvent. When the peptide backbone is in an extended conformation, an almost uniform distribution of serine rotamers is observed in solvent (Table 3). However, when the backbone is allowed to relax (CMC/MD-2b), the $\chi 1 = 180$ rotamer is significantly disfavored in both phases, and the contribution of the $\chi 1 = 300$ intramolecular hydrogen bond becomes less significant. Instead, there is a significant contribution from a hydrogen bond between the $\chi 1 = 60$ rotamer and the amino acid's own carbonyl group. In vacuum, the backbone adopts a C7 equatorial (C7eq) conformation, while it samples more broadly in solution. The shifting contributions of different rotamers depending on the precise conformation of the peptide and the simulation protocol used emphasize the need for careful inclusion of rotameric states in free energy calculations.

T4 results

The results of our T4 lysozyme free energy calculations are shown in Table 4, along with values from experiment and traditional thermodynamic integration (TI) free energy calculations, where available. There is reasonable agreement with the

experimental data, with an average absolute error of 0.8 kcal/mol. The data presented are the average results of two separate CMC/MD calculations for each state. As others have noted, the contribution of the residue 133 sidechain to the stability of T4 lysozyme is primarily driven by the burial of hydrophobic groups. The cavity itself is highly structured, showing only small differences between wild-type and L133A crystal structures[37]. Thus, the hydrophobic side chain that fills the cavity best makes for the most stable protein.

As expected, the β-branched side chains valine and isoleucine are relatively unfavorable in the helical environment of residue 133. Isoleucine is able to compensate somewhat for its β-branching by burial of substantial hydrophobic surface area, but is still much less favorable than norvaline, leucine, or phenylalanine. The values for the extended norvaline and O-methylserine side chains are calculated in excellent agreement with experiment.

While the CMC/MD calculation successfully identifies LEU and PHE as the most favorable mutants, PHE is calculated to be somewhat less favored than expected and the preference for leucine is somewhat overestimated. This may be in part due to the use of the L133 crystal structure as the starting point for the simulation, but it is also due to the uncertainty and difficulty in calculating free energy values for phenylalanine. Due to the very large steric volume of the PHE aromatic ring, Monte Carlo moves that select PHE as a trial move are often rejected due to unfavorable van der Waals overlaps between the aromatic ring and the surroundings. This makes the relative solvation free energy of phenylalanine difficult to determine (especially versus alanine). It must be pointed out, however, that ALA ⇔ PHE perturbations are also very challenging for traditional free

energy calculations. Bias due to the L133 crystal structure might be ameliorated by carrying out separate calculations starting from both the wild type and L133A structures and comparing their predictions.

In addition to allowing the calculation of free energies for many side chains at one position in T4 lysozyme, the CMC/MD calculation also provides some important general insights. Like our dipeptide calculations, the explicit inclusion of each rotamer in the T4 calculation allows us to monitor the relative population of each rotamer for a given residue type. Fundamentally, rotamers are just conformational minima separated by significant (~3 kcal/mol) free energy barriers. The height of these barriers means that it is difficult to sample over them with short (sub-nanosecond) molecular dynamics or free energy calculations. This is a particularly crucial issue for protein free energy calculations. Prior work has suggested that single preferred rotamers dominate free energy differences between side chains. Figure 10 shows the relative population of each rotamer for the folded and unfolded states of our CMC/MD calculation. Clearly, several side chains populate multiple rotamers in each state, each of which contributes to the calculated free energy. In solution, most side chains populate several rotamers, while only a few are populated in the protein. $\chi 1 = 60$ rotamers are excluded for most side chains in either environment, probably due to the local structure of the peptide backbone. The highly flexible side chains (NVL, MSE) populate several species in both environments, while the larger species (LEU, ILE, PHE – data not shown) populate only a single rotamer in the restricted environment of the protein. As expected, LEU populates the wild-type $\chi 1 = 300$, $\chi 2 = 180$ rotamer, while PHE is restricted to the $\chi 1 = 300$, $\chi 2 = 0$ flat-well, compatible with its experimental $\chi 1$ value of 273 degrees.

Discussion and Conclusions

We have successfully extended CMC/MD calculations to the realm of protein side chains. Our present calculations have yielded quantitative determination of the relative solvation free energies of alanine, valine, and serine in good agreement with prior free energy calculations. As expected, the order of solvation free energies is VAL > ALA >> SER, with alanine preferred to valine by 0.3 to 0.5 kcal/mol and serine preferred to alanine by 0.9 to 2.3 kcal/mol. However, these are substantial deviations from the expected solvation free energies based on the gas phase to water transfer free energies of the model compounds studied by Wolfenden, et. al[6]. Since the solvation free energies of methane and propane are very similar, it was a surprise that valine appeared so unfavorable relative to alanine in water during prior free energy calculations[3]. However, the previous calculations did not take into account the entropy loss associated with the rotamers of valine's $\chi 1$ torsion or the contribution of the gas phase free energy difference to the calculated solvation free energy. More importantly, a non-optimum extended backbone model was used. These effects all contribute significantly to the calculated free energy difference between alanine and valine. When the vacuum state and rotamer contributions are properly accounted for, valine is calculated to have an unfavorable solvation free energy of +0.3 kcal/mol relative to alanine. If the intramolecular contributions are taken into account, this rises to +1.2 kcal/mol if the peptide backbone is still restrained to an extended conformation. When the backbone is allowed to relax to an optimum conformation in each phase, the relative solvation free

energy of valine versus alanine returns to +0.5 kcal/mol. In the present work, a different effect is observed with serine. Ala => ser is calculated to be –0.9 to –2.0 kcal/mol, yet their analogs (methane and methanol) differ in solvation free energy by –6.8 kcal/mol[38]. An intramolecular hydrogen bond formed by the serine side chain means that serine is favored in vacuum by –1.9 kcal/mol, when intramolecular terms are not included. This is in contrast with the vacuum free energy difference of methane and methanol, which is expected to be small.

In a more ambitious calculation, we determined the relative stability of eight different mutants of T4 lysozyme. Our calculation clearly identifies leucine and phenylalanine as the most stabilizing residues at position 133, with alanine being the least-favorable residue. This is compatible with the observations of Karpusas, et. al.[23] that residue 133 occupies a well-structured hydrophobic cavity. The relative free energies of folding (ΔΔGfold) versus alanine are summarized in Table 4, and show good agreement with both the experimental data of Mendel, et. al. and the prior free energy calculations of Veenstra, et. al. The sole exception is ethylglycine (ETH), where we calculate this residue to be preferred by –2.4 kcal/mol over alanine in comparison to the experimental –0.2 kcal/mol. It is important, to note, however, that the ethylglycine mutant shows a significantly lowered Tm from the wild-type enzyme (~31 vs. 43.5 degrees C), requiring an extrapolation to determine the value of ΔΔGfold. Even with the ethylglycine outlier, the average absolute error of the 5 free energies calculated from CMC/MD is 0.8 kcal/mol. We somewhat overestimate the stability of the two large hydrophobic mutants (LEU and PHE), especially phenylalanine. For PHE, this error appears to be due to significant uncertainties in its solvation free energy (ΔΔGwat).

These uncertainties are expected, given the large steric differences between ALA and PHE and the difficulty of sampling between them during a Monte Carlo step. Reassuringly, however, we find that neither β-branched amino acid (VAL, ILE) is highly favored at this position. As expected, β-branching is incompatible with the helical location of residue 133, though ILE is able to compensate for this somewhat by the burial of significant hydrophobic surface area.

In carrying out these calculations, we needed to extend the CMC/MD method to include both conformational and chemical sampling, allowing Monte Carlo steps to occur in a hybrid space of rotamers and chemical species. The appreciable free energy penalty associated with the "freezing" of a protein side chain in a single rotamer (ca. 0.6-2.6 kcal/mol) is a substantial contribution that must be considered in quantitative free energy calculations on proteins and peptides. In our case, we see significant shifts in the relative populations of rotamers depending on the environment, underscoring the complexity of the issue (Table 3, Figure 10). In particular, side chains need not be "frozen" in a single rotamer – some rotamers are simply more or less populated depending on the context. As previously observed, the contribution of each rotamer to the calculated free energy must be determined and included in order to yield an accurate free energy. Highly flexible side chains (NVL and MSE in the current work; MET, LYS) can often occupy several different rotamers that each satisfy the steric constraints of a tightly-packed cavity.

While substantial progress has been made in computational approaches to large-scale protein engineering[21], quantitative calculations of the effect of single amino acid changes on protein properties are still difficult computational tasks[4]. In contrast, our CMC/MD approach brings multi-coordinate free energy methods to the arena of protein

free energy calculations, allowing the quantitative comparison of many different side chains in a single computer "experiment". As we have noted, this is analogous to "exhaustive mutagenesis" of a single residue. Unlike the biochemical experiment, however, CMC/MD has some difficulty in sampling between chemical states separated by large changes in volume (arginine and glycine, for instance) or charge (glutamic acid versus lysine). Instead, we expect that CMC/MD will best be used to compare a family of similarly-sized amino acids (ala-ser-val-ile-leu-thr, for instance). This is particularly useful in the context of non-natural side chains[39], where the significant difficulty of synthesis and *in vitro* transcription for a single substitiution prompt a need for accurate predictions.

With our calculations on the blocked dipeptide and T4 lysozyme, we have now entered the era of "exhaustive mutagenesis" *in silico* – quantitative free energy calculations on many different side chains at one position of a protein. While there are still substantial technical and practical limitations to the determination of protein stability with free energy calculations, substantial progress has been made. Multicoordinate free energy methods like CMC/MD and the lambda-dynamics of Brooks, et. al.[22], serve to expand the predictive power of theoretical methods by allowing the rapid comparison of many chemical states in a single calculation. Species of interest are quickly picked out for further study by either computational or experimental means. This "winnowing" and discarding of uninteresting states allows more time to be spent synthesizing and studying the compounds of interest – whether they are novel guests[1], new ligands[2] or new proteins.

The developments outlined in this paper transform CMC/MD into a uniquely powerful tool for free energy calculations. By allowing a combination of chemical and conformational sampling in the same calculation, we can accurately include both in studies of free energy differences at modest cost. It is difficult to do this efficiently with traditional free energy calculations, as demonstrated by our vacuum state TI calculations on the dipeptide. Our peptide calculation underscores the need for extensive and detailed conformational sampling in any free energy calculation on flexible molecules. The inclusion of intramolecular energies allows CMC/MD calculations to be carried out on "mutations" in any polymeric species, including but not limited to proteins and nucleic acids. With the present additions, CMC/MD will also become an even better tool for ligand design, able to rapidly rank the binding free energies of a series of ligands that vary both in composition and flexibility.

Acknowledgements

JWP would like to acknowledge the support of the NSF, IBM and the Minnesota Supercomputer Center, and the University of California Office of the Chancellor. PAK would like to thank the NIH for support through GM-29072. Dr. David Veenstra and Dr. Lu Wang were invaluable in providing parameters and protocols for the T4 lysozyme calculations. We would also like to thank the NCSA supercomputing center at University of Illinois Urbana-Champaign for computational resources.

Table 1: ΔΔGsolv for alanine, valine, and serine

Method	alanine => valine			Alanine => serine			Internal Energies	Sidechain Restraint	Backbone restraint
	ΔGvac	ΔGwat	ΔΔGsolv	ΔGvac	ΔGwat	ΔΔGsolv			
TI	1.5	1.14	(-0.4)	-3.7			N	$\chi 1$ +/- 40	(-160,160)
CMC/MD-1	0.5	0.8	+0.3	-1.9	-4.2	-2.3	N	$\chi 1$ +/- 40	(-160,160)
TI	-18.9			-9.4			Y	$\chi 1$ +/- 40	(-160,160)
CMC/MD-2a	-15.3	-14.1	+1.24	-7.9	-8.5	-0.6	Y	$\chi 1$ +/- 40	(-160,160)
CMC/MD-2b	-15.8	-15.3	+0.5	-8.1	-9.0	-0.9	Y	$\chi 1$ +/- 40	(-175<ϕ<-5, 5<ψ<175)
Experiment	Me => Prp		+0.0	Me => MeOH		-6.8			

Table 2: Inter-rotamer relative free energies, vacuum state, harmonic backbone restraints

Residue	$\chi 1$ (degrees)	ΔG, No internal contributions (kcal/mol)		ΔG, internal contributions (kcal/mol)	
		CMC/MD	TI	CMC/MD	TI
Serine	60 => 180	+0.1	+1.0	-0.9	-0.5
	60 => 300	-0.8	-1.4	-1.2	-0.9
Valine	60 => 180	+2.4	+4.0	+0.4	+1.3
	60 => 300	+0.7	+0.5	+1.2	+1.7

Table 3: Rotamer populations for the peptide calculations

CMC/MD Method	Phase	Valine $\chi 1$ =			Serine $\chi 1$ =		
		60	180	300	60	180	300
1	Vacuum	0.77	0.02	0.21	0.17	0.13	0.68
1	Water	0.01	0.00	0.99	0.67	0.22	0.11
2a	Vacuum	0.54	0.06	0.40	0.04	0.71	0.24
2a	Water	0.48	0.02	0.50	0.39	0.25	0.36
2b	Vacuum	0.14	0.33	0.52	0.81	0.02	0.16
2b	Water	0.01	0.18	0.81	0.71	0.01	0.28

Table 4: ΔΔGfold for T4 lysozyme side chains

Alanine =>	ΔGprot	ΔGwat	ΔΔGfold	ΔΔGfold, exp't
ETH	6.3	8.8	-2.5 +/- 0.7	-0.2 (*)
VAL	3.2	2.6	+0.5 +/- 0.1	ND
NVL	2.6	4.9	-2.3 +/- 0.4	-2.4 (-3.4, TI)
MSE	0.8	2.0	-1.2 +/- 0.6	-0.8 (-1.6, TI)
LEU	2.9	7.7	-4.8 +/- 0.2	-3.5
ILE	5.0	6.0	-1.0 +/- 0.7	ND
PHE	7.7	10.4	-2.7 +/- 1.4	-3.3

*: All experimental ΔΔGfold values reported by Mendel, et. al. are at the wild-type T_m of 316.6 K except for the ETH mutant. In this case, the reported ΔΔGfold is at the mutant T_m of 304 K, and the wild-type value has been extrapolated down from the higher temperature based on a ΔH_{fold} of 96 kcal/mol and ΔC_p of 1.80 kcal/mol.

Bibliography

1. Pitera, J., Kollman, P., *Designing an Optimum Guest for a Host using Multi-Molecule Free Energy Calculations: Predicting the Best Ligand for Rebek's "Tennis Ball".* Journal of the American Chemical Society, 1998. **120**(30): p. 7557-7567.
2. Eriksson, M.A.L., Pitera, J., Kollman, P.A., *Construction and ranking of new HIV-1 reverse transcriptase TIBO inhibitors combining computational methods at different levels of accuracy.* J. Med. Chem., 1998. **submitted**.
3. Sun, Y., Spellmeyer, D., Pearlman, D., and Kollman, P., *Simulation of the Solvation Free Energies of Methane, Ethane, and Propane and Corresponding Amino Acid Dipeptides: A Critical Test of the "Bond-PMF" correction, a New Set of Hydrocarbon Parameters, and the Gas Phase-Water Hydrophobicity Scale.* Journal of the American Chemical Society, 1992. **114**(17): p. 6798-6801.
4. Veenstra, D.L., Kollman, P.A., *Modeling protein stability: a theoretical analysis of the stability of T4 lysozyme mutants.* Protein Engineering, 1997. **10**(7): p. 789-807.
5. Mendel, D., *et al.*, *Probing protein stability with unnatural amino acids.* Science, 1992. **256**(5065): p. 1798-802.
6. Wolfenden, R., Andersson, L., Cullis, P.M., Southgate, C.C.B., *Affinities of Amino Acid Side Chains for Solvent Water.* Biochemistry, 1981. **20**: p. 849-855.
7. Dill, K.A., *Dominant forces in protein folding.* Biochem., 1990. **29**(31): p. 7133-7155.

8. Yunyu, S., Mark, A.E., Wang, C.X., Huang, F.H., Berendsen, H.J.C, Vangunsteren, W.F., *Can the stability of protein mutants be predicted by free energy calculations.* Protein Engineering, 1993. **6**(3): p. 289-295.

9. Metropolis, N.R., A. W., Rosenbluth, M. N., Teller, A. H., Teller, E., *Equation of State Calculations by Fast Computing Machines.* J. Chem. Phys., 1953. **21**: p. 1087-1092.

10. Pearlman, D., Kollman, P., *The lag between the Hamiltonian and the system configuration in free energy perturbation calculations.* J Chem Phys, 1989. **91**(12): p. 7831-7839.

11. Elber, R., Karplus, M., *Enhanced Sampling in Molecular Dynamics - Use of the Time-Dependent Hartree Approximation for a Simulation of Carbon Monoxide Diffusion Through Myoglobin.* Journal of the American Chemical Society, 1990. **112**(25): p. 9161-9175.

12. Simmerling, C., Fox, T., Kollman, P.A., *Use of locally enhanced sampling in free energy calculations: Testing and application to the alpha->beta anomerization of glucose.* Journal of the American Chemical Society, 1998. **120**(23): p. 5771-5782.

13. Ota, N., Brunger, A., *Overcoming barriers in macromolecular simulations: non-Boltzmann thermodynamic integration.* Theor Chem Acc, 1997. **98**: p. 171-181.

14. Mark, A.E., van Gunsteren, W.F., Berendsen, H.J.C., J Chem Phys, 1991. **94**(3808).

15. Guarnieri, F. and W.C. Still, *A Rapidly Convergent Simulation Method - Mixed Monte Carlo Stochastic Dynamics.* J Comput Chem, 1994. **15**(11): p. 1302-1310.

16. Huber, T., Torda, A., Vangunsteren, W.F., *Optimization methods for conformational sampling using a Boltzmann-weighted mean field approach.* Biopolymers, 1996. **39**(1): p. 103-114.

17. Wilson, C., Mace, J.E., Agard, D.A., *Computational Method for the Design of Enzymes with Altered Substrate Specificity.* J Mol Biol, 1991. **220**(2).

18. Yun, R.H. and J. Hermans, *Conformational equilibria of valine studied by dynamics simulation.* Protein Eng, 1991. **4**(7): p. 761-6.

19. Hermans, J., A.G. Anderson, and R.H. Yun, *Differential helix propensity of small apolar side chains studied by molecular dynamics simulations.* Biochemistry, 1992. **31**(24): p. 5646-53.

20. Sun, Y.C., Veenstra, D. L., Kollman, P. A., *Free Energy Calculations of the Mutation of ILE96-ALA in Barnase - Contributions to the Difference in Stability.* Prot. Eng., 1996. **9**: p. 273-281.

21. Dahiyat, B.I., Mayo, S.L., *De novo protein design: Fully automated sequence selection.* Science, 1997. **278**(5335): p. 82-87.

22. Kong, X.J. and C.L. Brooks, *Lambda-Dynamics - a New Approach to Free Energy Calculations.* J Chem Phys, 1996. **105**(6): p. 2414-2423.

23. Karpusas, M., *et al.*, *Hydrophobic packing in T4 lysozyme probed by cavity-filling mutants.* Proc Natl Acad Sci U S A, 1989. **86**(21): p. 8237-41.

24. Weaver, L.H. and B.W. Matthews, *Structure of bacteriophage T4 lysozyme refined at 1.7 A resolution.* J Mol Biol, 1987. **193**(1): p. 189-99.

25. Wang, L., *et al.*, *Can one predict protein stability? An attempt to do so for residue 133 of T4 lysozyme using a combination of free energy derivatives, PROFEC, and*

free energy perturbation methods. Protein-Struct Funct Genet, 1998. **32**(4): p. 438-458.

26. Bennett, C.H., *Efficient Estimation of Free Energy Differences from Monte Carlo Data.* Journal of Computational Physics, 1976. **22**: p. 245-268.

27. Pearlman, D., *A comparison of alternative approaches to free energy calculations.* J Phys Chem, 1994. **98**(5): p. 1487-1493.

28. Pearlman, D.A., *et al.*, *AMBER, a Package Of Computer Programs For Applying Molecular Mechanics, Normal Mode Analysis, Molecular Dynamics and Free Energy Calculations to Simulate the Structure and Energetic Properties of Molecules.* Comp. Phys. Comm., 1995. **91**(1-3): p. 1-41.

29. Ryckaert, J.P., G. Ciccotti, and H.J.C. Berendsen, *Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-Alkanes.* J. Comput. Phys., 1977. **23**: p. 327-341.

30. Cornell, W.D., *et al.*, *A Second Generation Force Field For the Simulation Of Proteins, Nucleic Acids, and Organic Molecules (Vol 117, Pg 5179, 1995).* J Amer Chem Soc, 1996. **118**(9): p. 2309-2309.

31. Jorgensen, W.L., *et al.*, *Comparison of Simple Potential Functions for the Simulation of Liquid Water.* J. Chem. Phys., 1983. **79**: p. 926.

32. Andersen, H.C., *Molecular Dynamics Simulations at Constant Pressure and/or Temperature.* J. Chem. Phys., 1980. **52**: p. 2384-2393.

33. Torrie, G.M.a.V., J.P., *Nonphysical sampling distributions in Monte Carlo free energy estimation: umbrella sampling.* Journal of Computational Physics, 1977. **23**: p. 187-199.

34. Kumar, S., *et al.*, *The weighted histogram analysis method for free-energy calculations on biomolecules. 1. The method.* J . Comp. Chem., 1992. **13**(8): p. 1011-1021.

35. Cornish, V.W., *et al.*, *Stabilizing and destabilizing effects of placing beta-branched amino acids in protein alpha-helices.* Biochemistry, 1994. **33**(40): p. 12022-31.

36. Berendsen, H.J.C., *et al.*, *Molecular Dynamics with Coupling to an External Bath.* J. Chem. Phys., 1984. **81**: p. 3684-3690.

37. Zhang, X.J., W.A. Baase, and B.W. Matthews, *Multiple alanine replacements within alpha-helix 126-134 of T4 lysozyme have independent, additive effects on both structure and stability.* Protein Sci, 1992. **1**(6): p. 761-76.

38. Ben-Naim, A., Marcus, Y., J Chem Phys, 1984. **81**: p. 2016-.

39. Ellman, J., *et al.*, *Biosynthetic method for introducing unnatural amino acids site-specifically into proteins.* Methods Enzymol, 1991. **202**: p. 301-36.

Figure Captions

Figure 1: Thermodynamic cycle used to calculate relative solvation free energies ($\Delta\Delta$Gsolv) the peptide calculation.

Figure 2: Thermodynamic cycle used to calculate relative stabilities ($\Delta\Delta$Gfold) of T4 lysozyme mutants.

Figure 3: Schematic representation of the topology (3a) and Monte Carlo trial move space (3b) for the peptide calculation.

Figure 4: Stick representation of the ACE-(ALA,VALx3,SERx3)-NMA molecule, in a representative conformation from the vacuum CMC/MD-2b calculation. Alanine is colored light grey and all other residues are colored by atom type.

Figure 5: Computational strategy used for the CMC/MD calculations described in this paper.

Figure 6: Ribbon model of T4 lysozyme from our CMC/MD trajectory, showing some of the residues studied at position 133. The dominant leucine rotamer is depicted in black, and the other 8 leucine rotamers are shown in light gray.

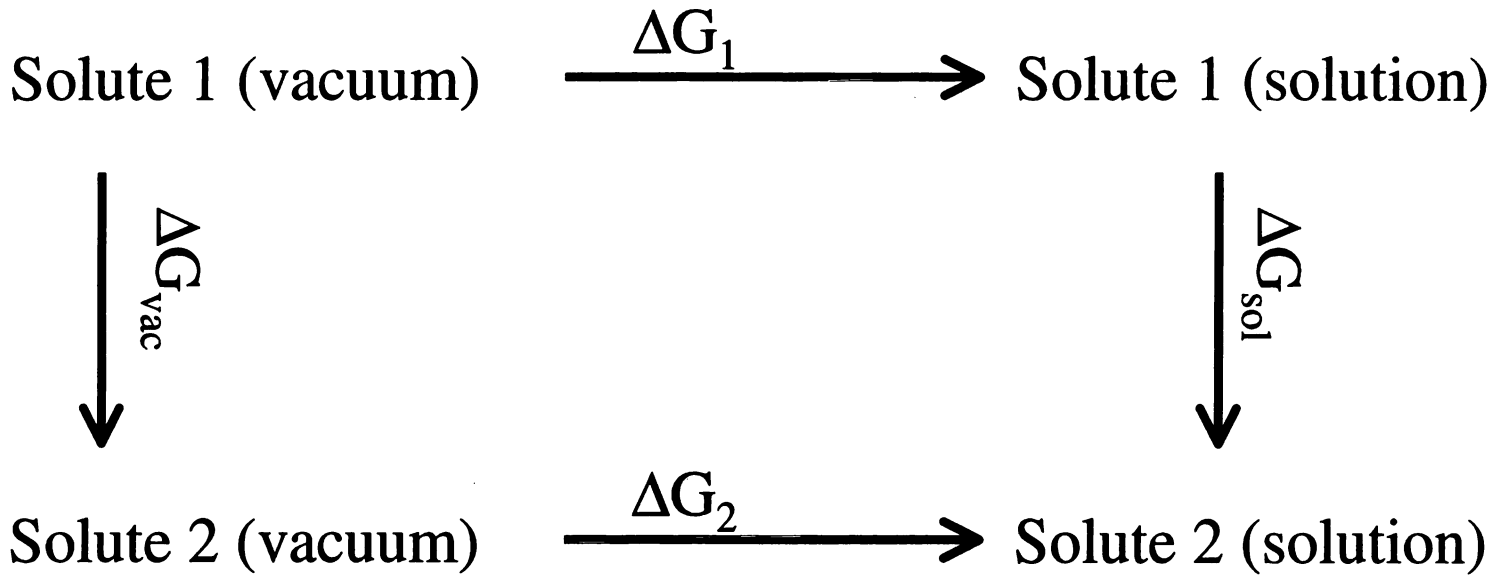
Figure 7: Convergence of the adaptive vacuum phase of the peptide calculation. Figure 7a shows the results when intraperturbed contributions are not included (CMC/MD-1) while Figure 7b depicts the results when intraperturbed groups are included as well as flat-well backbone restraints (CMC/MD-2b). In all cases, the solid line shows the free energy value for ALA => VAL, while the dashed line depicts ALA => SER.

Figure 8: Convergence of the non-adaptive vacuum phase of the peptide calculation. Again, 8a shows the CMC/MD-1 result, and the CMC/MD-2b result is shown in 8b. The solid line is the data for ALA => VAL and the dashed line is the value for ALA => SER.

Figure 9: Convergence of the non-adaptive solvent phase of the peptide calculation. Figure 9a depicts the convergence of the CMC/MD-1 calculation. The results of including intraperturbed group contributions and flat-well backbone restraints (CMC/MD-2b) are shown in Figure 9b. Again, ALA => VAL is the solid line and ALA => SER is the dotted line.

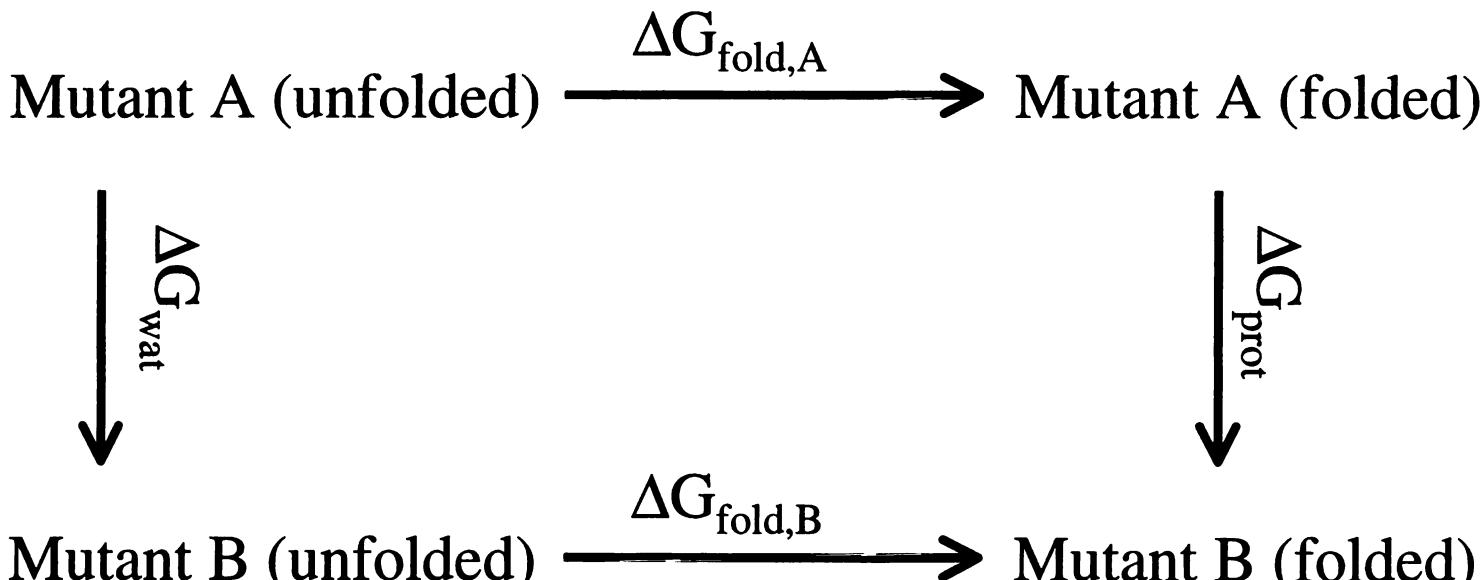
Figure 10: Relative rotamer populations from T4 lysozyme calculations in the unfolded (white) and folded (grey) states. Populations are depicted for ETH, VAL, NVL, MSE, LEU and ILE. Statistics for PHE were too poor to determine meaningful populations.

Figure 1



$$\Delta\Delta G_{sol} = \Delta G_2 - \Delta G_1 = \Delta G_{sol} - \Delta G_{vac}$$

Figure 2



$$\Delta\Delta G_{fold} = \Delta G_{fold,B} - \Delta G_{fold,A} = \Delta G_{prot} - \Delta G_{wat}$$

Figure 3a

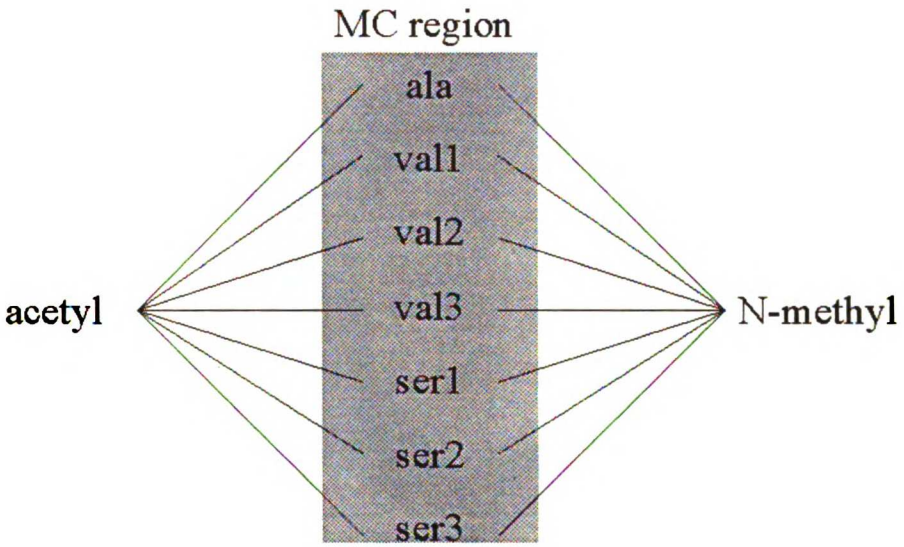


Figure 3b

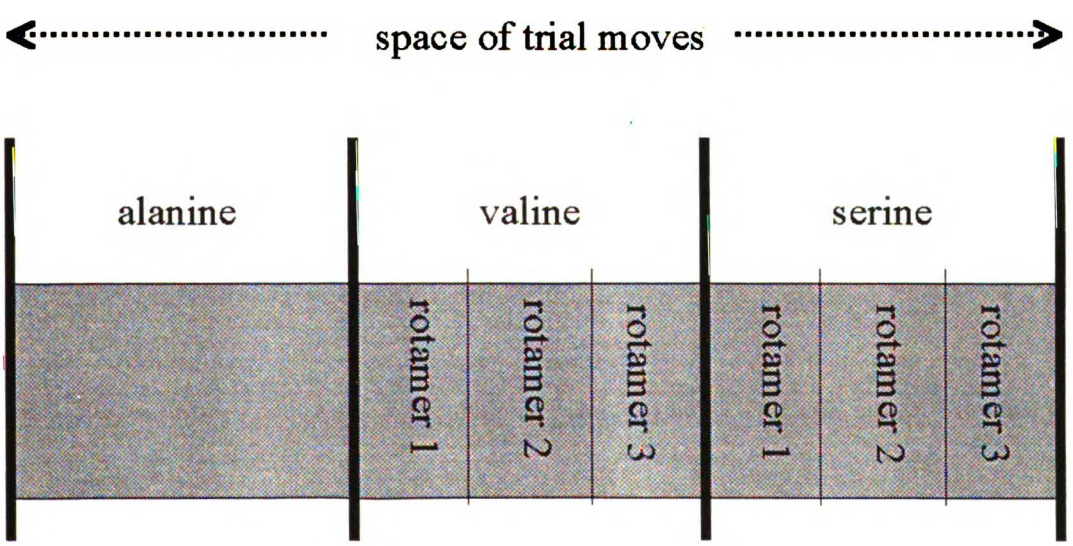
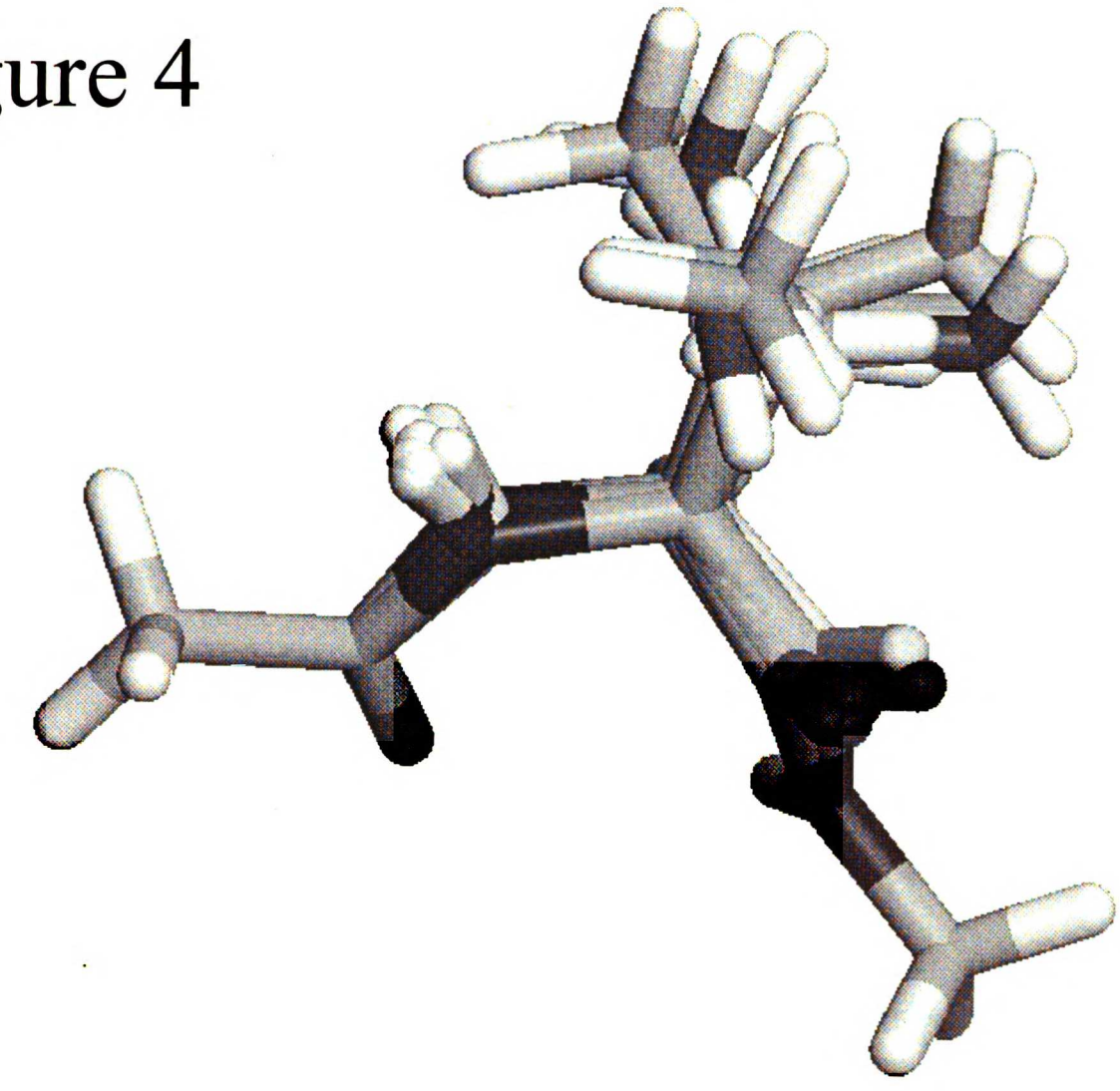


Figure 4



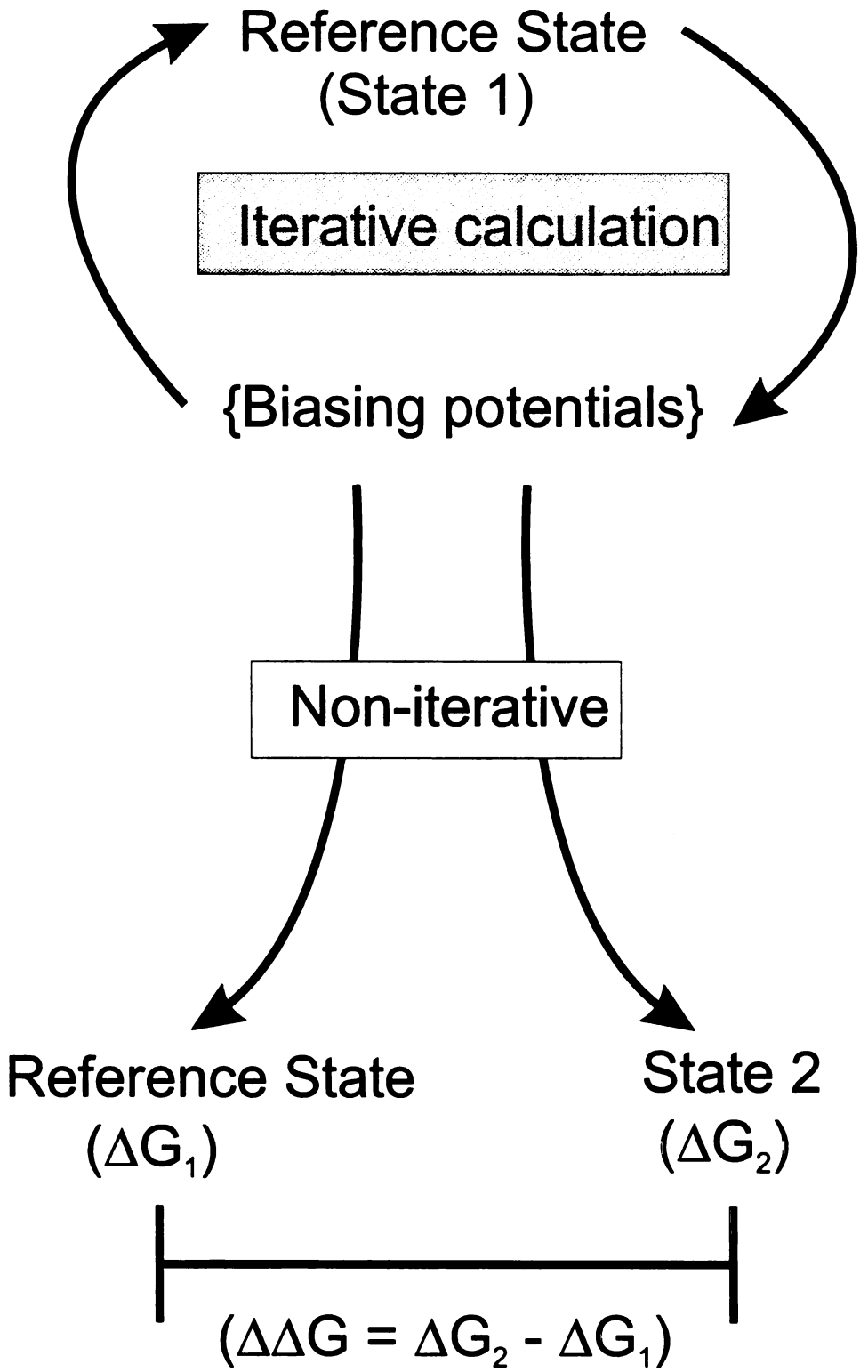

Reference State
(State 1)
Iterative calculation
{Biasing potentials}
Non-iterative
Reference State
(ΔG1)
State 2
(ΔG2)
(ΔΔG = ΔG2 - ΔG1)

Figure 6

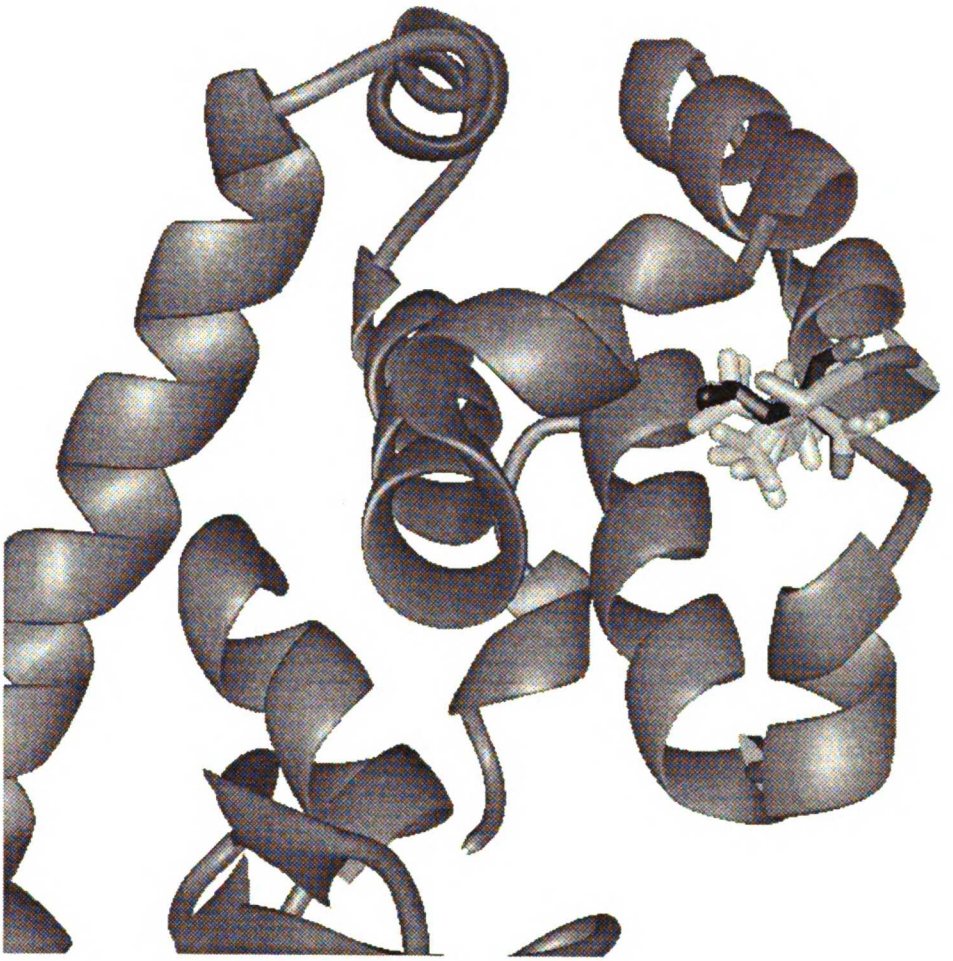


Figure 7

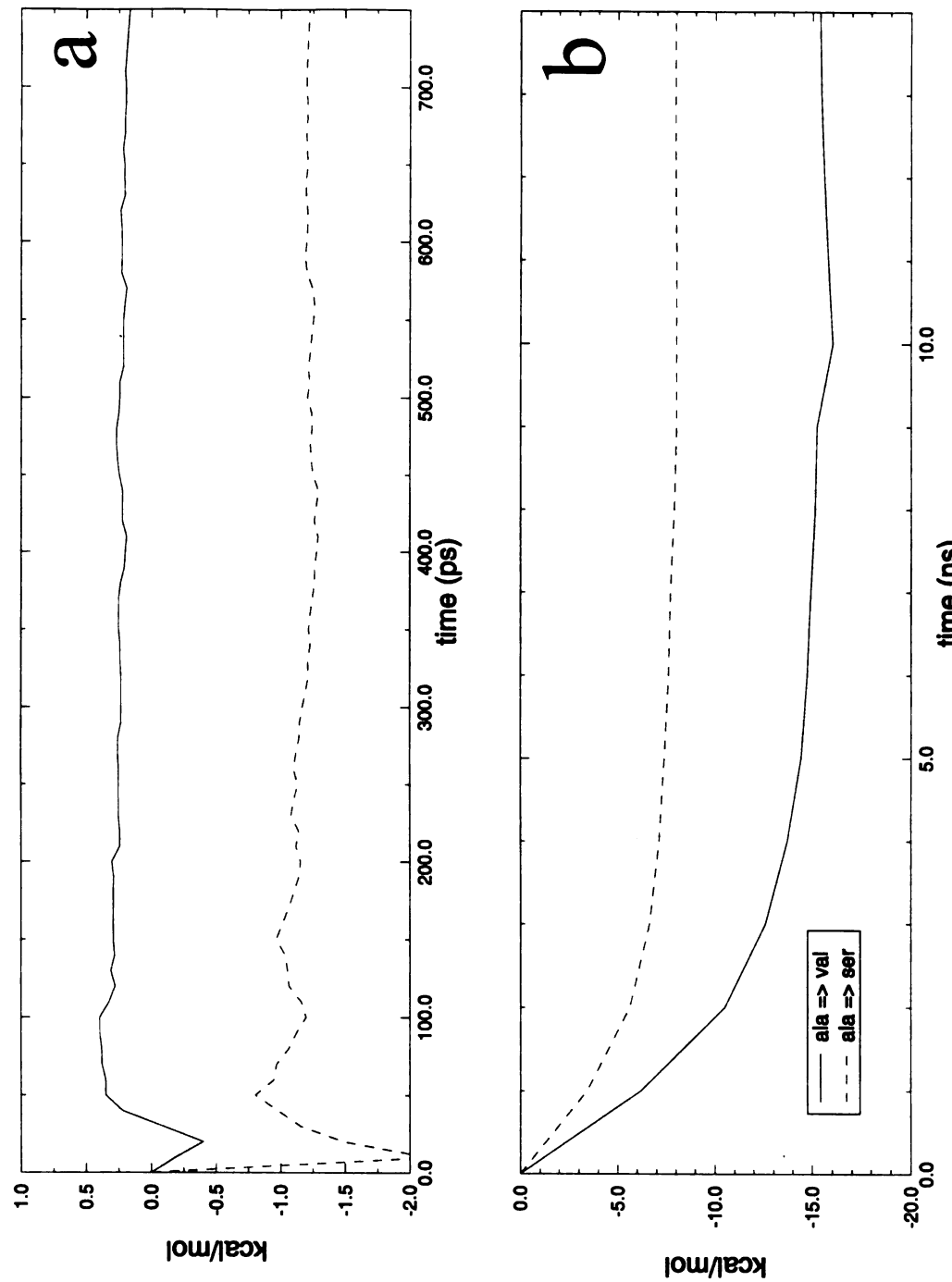

a
kcal/mol
time (ps)
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
0.0
100.0
200.0
300.0
400.0
500.0
600.0
700.0
b
kcal/mol
time (ns)
0.0
-5.0
-10.0
-15.0
-20.0
0.0
5.0
10.0
ala => val
ala => ser

Figure 8

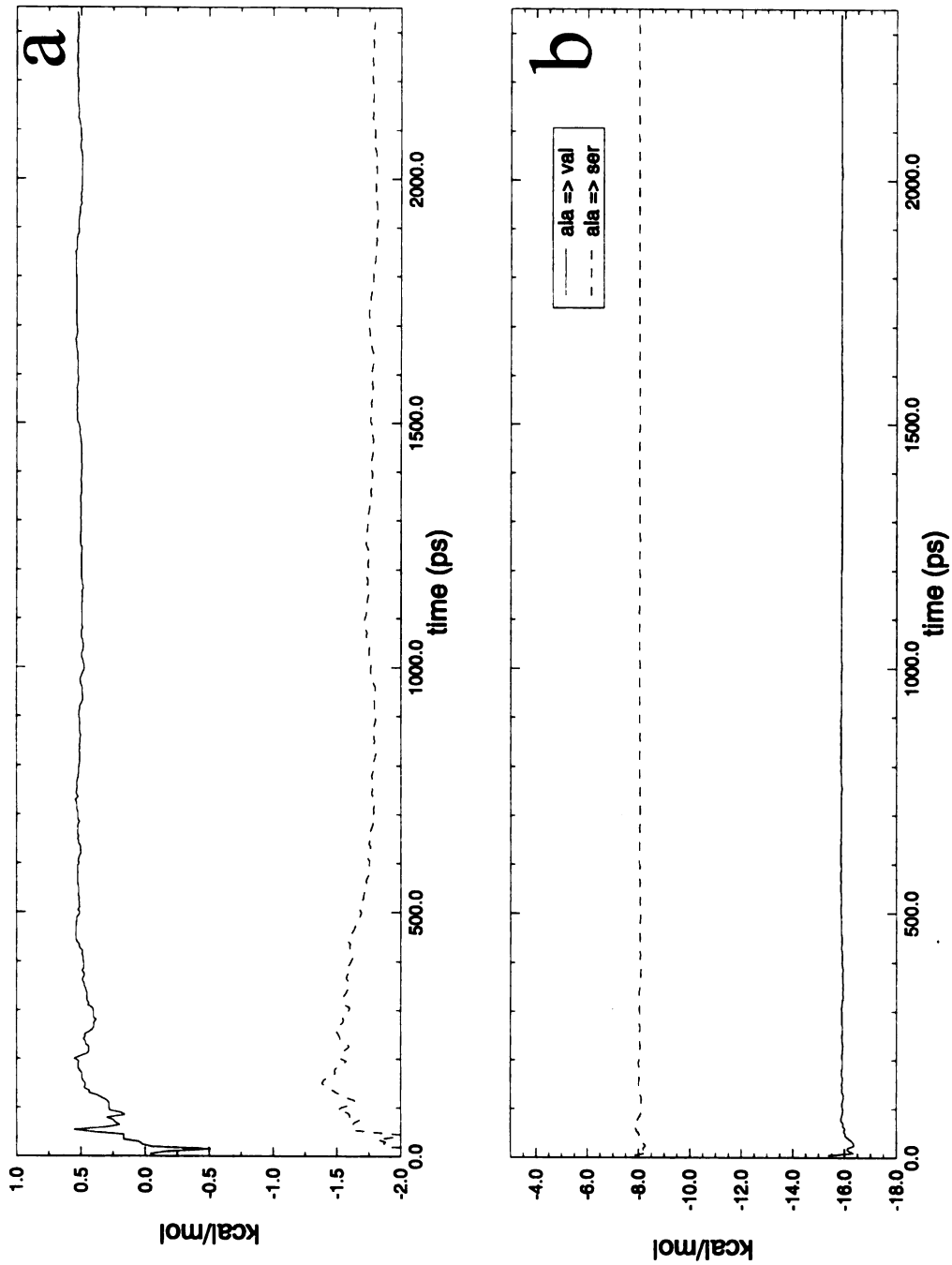

a
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
kcal/mol
0.0
500.0
1000.0
1500.0
2000.0
time (ps)
b
ala => val
ala => ser
-4.0
-6.0
-8.0
-10.0
-12.0
-14.0
-16.0
-18.0
kcal/mol
0.0
500.0
1000.0
1500.0
2000.0
time (ps)

Figure 9

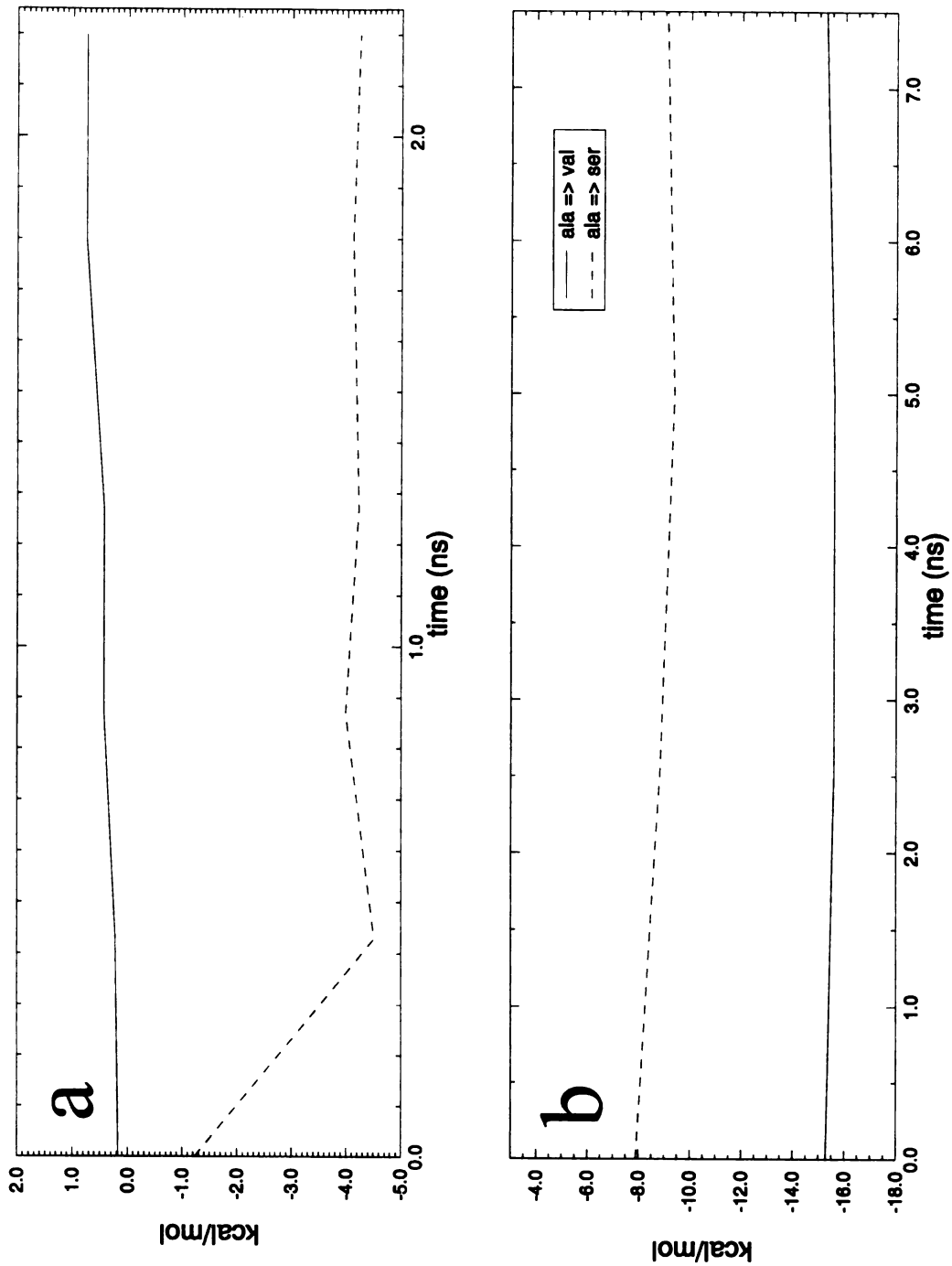
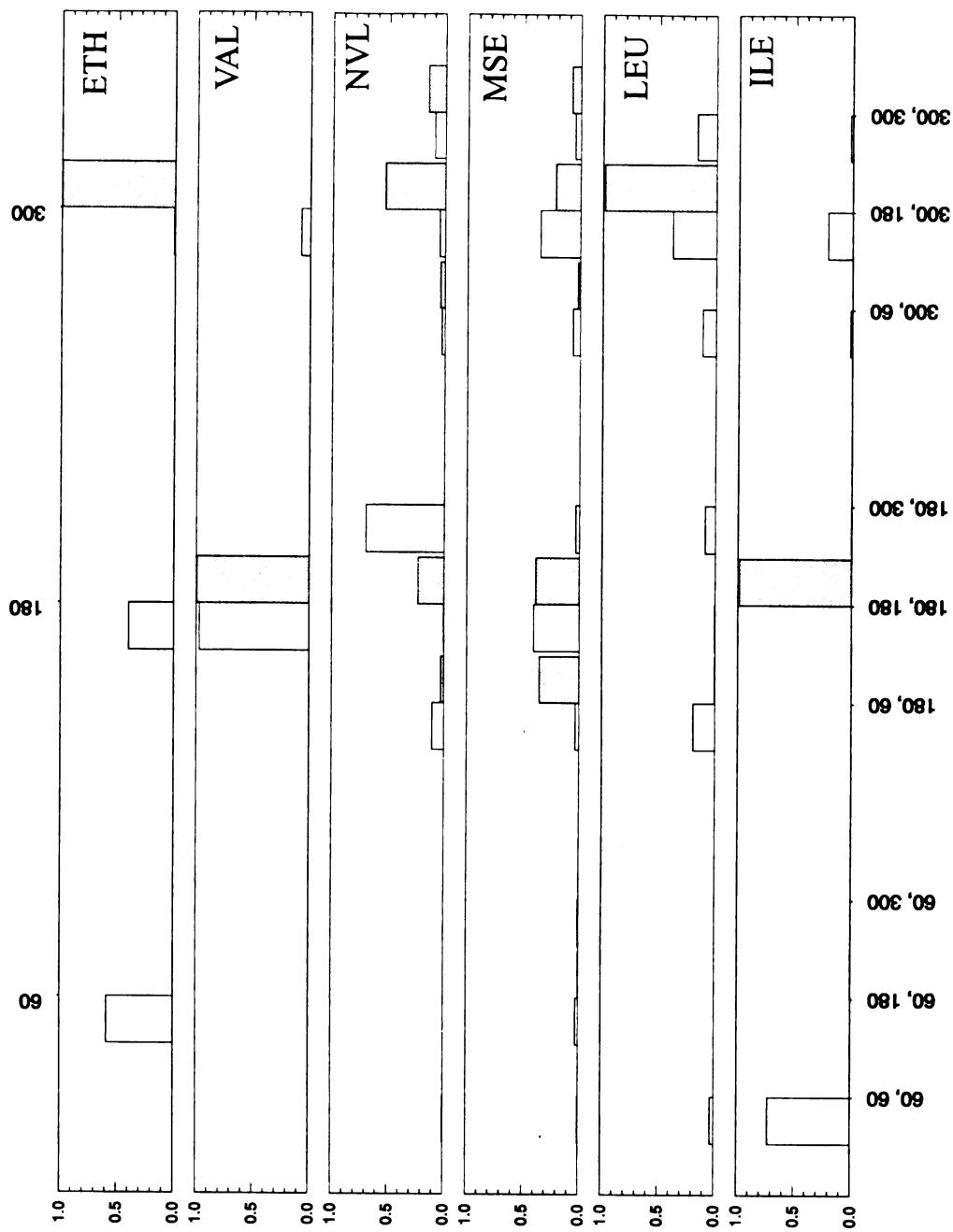

a
kcal/mol
time (ns)
2.0
1.0
0.0
-1.0
-2.0
-3.0
-4.0
-5.0
0.0
1.0
2.0
b
kcal/mol
time (ns)
ala => val
ala => ser
-4.0
-6.0
-8.0
-10.0
-12.0
-14.0
-16.0
-18.0
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0

Figure 10


ETH
VAL
NVL
MSE
LEU
ILE
300
180
60
1.0
0.5
0.0
60, 60
60, 180
60, 300
180, 60
180, 180
180, 300
300, 60
300, 180
300, 300

Chapter 8: Statistical issues in free energy determination with CMC/MD.

The population-based free energy determination used in CMC/MD has some statistical limitations. Specifically, when one determines free energy differences by taking the ratio of two observed populations

$$\Delta G(i->j) = -RT \ln (P(j)/P(i))$$

there are two major difficulties that arise. The first is simply the logarithmic relationship of the population ratio to the free energy. This is illustrated in Table 1, which shows the population ratios that correspond to a range of free energies.

Table 1

Free energy difference at 300K (kcal/mol)	Population ratio
0	1:1
0.6	1:2.7
1.0	1:5.3
2.0	1:28.6
3.0	1:153.3
4.0	1:820.0
5.0	1:4393
10.0	$1:1.9 \times 10^{7}$
20.0	$1:3.7 \times 10^{14}$

Now, if we make one observation every picosecond, it will take over 10^7 picoseconds (10 microseconds) to sample a population ratio that corresponds to a 10 kcal/mol free energy difference. At present, available computer power restricts practical calculations to less than 10 nanoseconds, several orders of magnitude below the time scale necessary. This

first issue is largely addressed by the use of an adaptive CMC/MD calculation, as described in Chapter 6. In the adaptive CMC/MD calculation, a biasing potential is added to the energy of each state, and these values are iteratively adjusted until they yield a uniform population distribution in a CMC/MD calculation. In this fashion, chemical species that have large free energy differences (10-20 kcal/mol) can be efficiently compared.

Free energy differences that are very large relative to kT (greater than 60 kcal/mol) are still not accessible with CMC/MD, since the huge biasing potential required allows the acceptance of Monte Carlo steps that would normally be impossible at 300K. Fortunately, this is not an issue in the comparison of families of highly related ligands. Free energy differences of this magnitude can arise in biomolecular systems, however, especially in the comparison of ionic and neutral systems. The free energy of solvation of a monoatomic ion in water is favorable by hundreds of kcal/mol, compared to less than ten kcal/mol for a similarly-sized noble gas atom. It should be noted that these sorts of comparisons are difficult for traditional free energy calculations as well, requiring addition of a Born or reaction-field correction[1].

The second statistical issue arises from the ability of CMC/MD to compare many different species in a single calculation. The number of species that can be compared is limited by the need to determine accurate populations for each of them. Due to the logarithmic relationship between population ratios and free energies (as shown above), this process is largely limited by the population of the least favorable species. For example, the minimum number of samples to determine population ratios for a system of 7 solutes with relative free energies of 0,1,2,3,4,5 and 10 kcal/mol is approximately

($1.9x10^7 + 7x10^6 + 3.5x10^6 + 6.6x10^5 + 1.2x10^5 + 2.3x10^4 + 4393 + 1 = 3.0x10^7$). The adaptive CMC/MD procedure described in Chapter 6 was specifically developed to sidestep this limitation and allow the determination of free energies for unfavorable states as well as favorable ones. This is achieved by iteratively determining biasing potentials that convert the calculation from a comparison of species with differing free energies into a sampling between isoenergetic species.

One can also learn something by considering these systems with small, or zero, free energy differences. Consider two calculations: the first, comparing two solutes with zero free energy difference; and the second, comparing ten solutes all with equivalent free energies. A calculation of many (N) steps will yield populations of N/2 for each solute in the first calculation. In the second calculation, the same computation will yield populations of N/10. To determine free energy differences with equivalent accuracy, the populations observed from the two calculations should be equal in magnitude. Thus, our 10-solute calculation will require 5 times longer (5N) to yield equivalently accurate free energy values, since (5N)/10 = N/2. More generally, a CMC/MD calculation comparing M isoenergetic states will require M/2 times more calculations to determine free energies of the same accuracy as a two-state calculation. More complicated (non-isoenergetic) situations are very case-dependent. As a practical rule, I have found that 5-10 species can be accurately ranked with CMC/MD in a calculation of 1-2 nanoseconds, but convergence to quantitative free energy values requires longer (2-4 nanoseconds).

In considering quantitative free energy calculations, one question is the relationship of error in the population ratio to error in the determined free energy. The propagation of error is relatively straightforward. If we compare two states I and J with

populations P(I), P(J), and a corresponding free energy difference $\Delta G(I->J)$, how does an error in one population (dP) translate to an error in the determined free energy (dG)?

$$\Delta G = -RT \ln [P(J)/P(I)]$$

$$\Delta G + dG = -RT \ln [(P(J)+dP)/P(I)]$$

$$dG = -RT \{ \ln [(P(J)+dP)/P(I)] - \ln [P(J)/P(I)] \}$$

$$dG = -RT \ln [(P(J)+dP)/P(J)]$$

$$dG = -RT \ln [1 + dP/P(J)]$$

What does this mean? For a fixed magnitude of error, one can calculate how doubling the (long) simulation will affect the expected error in the free energies:

$$dG2/dG1 = \{ \ln [1 + dP/2P(J)] \}/\{ \ln [1 + dP/P(J)] \}$$

In a very unfavorable case, dP = P(J), yielding

$$dG2/dG1 = \{ \ln [1 + ½] \}/\{ \ln [1 + 1] \}$$

$$dG2/dG1 = \{ \ln [3/2] \}/\{ \ln [2] \} = 0.58$$

As P(J) increases relative to dP, the expected error should be halved by doubling the simulation time:

$$\text{Lim } (P(J) -> \infty) \; dG2/dG1 = 0.5$$

So, for simple propagation of a fixed error in the probabilities, extending the simulation time is expected to decrease the error in the calculated free energy in a straightforward way – doubling the simulation time will roughly halve the expected error in the calculated free energy. Now, our assumption of a fixed error is not entirely realistic. Instead, one might expect that the magnitude of the error increases in parallel with the magnitude of the observed probability (i.e. $dP \propto P(I)$). In this case, the error in the calculated free energy is constant regardless of the simulation length. Practically speaking, the behavior of the error is expected to lie somewhere between these two extremes (fixed vs. proportional to P(I)). Fortunately for our calculations, in this regime increasing the length of the simulation can only have the effect of decreasing the expected error or at worst keeping it constant.

The first two issues we have discussed are general properties of probability-based free energy calculations. There is one final statistical issue that should also be considered with regard to the specific use of CMC/MD in molecular systems. Our previous discussions of minimum sampling and error are based on the assumption that our Monte Carlo steps are <u>independent</u> samples. That is, between any two Monte Carlo steps there are enough molecular dynamics steps that the system has a chance to relax and explore phase space. A historically useful rule for molecular systems[2] is to base the sampling frequency on the relaxation time associated with significant processes in the system. For the calculation of an ensemble average property, it is not necessary to discretize the sampling with a frequency longer than the relaxation time. Instead, the sampling just needs to be carried out over a significantly longer length of time than the relaxation time – if the significant relaxation time in the system is 1 picosecond, averaging needs to be

carried out over 100+ picoseconds. This is actually a significant and little-discussed limitation of traditional free energy calculations. The numerical integration used in FEP or TI calculations usually requires breaking the calculation into several discrete windows and calculating a free energy difference for each window. These windows are effectively separate simulations, and each needs to be carried out for at least the sampling time described above. However, typical free energy calculations use less than 25 picoseconds of sampling time for each window. This is insufficient to sample contributions from minima separated by appreciable free energy barriers, like protein side chain rotamers[3].

Table 2 shows some of these time constants for processes in biomolecular systems, as listed in McCammon and Harvey[4]. For the specific case of CMC/MD in solution, a good lower bound on the sampling frequency is the pairlist update frequency. In a correctly run molecular simulation, the pairlist update is set to occur more frequently than significant changes in the conformation of the system. For the typical AMBER molecular dynamics simulation in solution, the pairlist is updated every 10 to 20 femtoseconds. Repeating Monte Carlo steps more frequently than this value will just produce the "Boltzmann probability" distribution described in Chapter 5 at additional computational expense. The studies reported in this thesis have typically used a 10 fs spacing for CMC/MD steps in vacuo, and a spacing of 100-500 fs (0.1-0.5 ps) for CMC/MD calculations in proteins or in solution. The total simulation lengths were, as noted previously, typically greater than or equal to 1 nanosecond (10^3 ps). The stochastic nature of the Metropolis Monte Carlo process helps to prevent the CMC/MD simulation from getting trapped in non-ergodic regions of phase space. This is augmented by the use

of stochastic temperature coupling methods to maintain the system at the desired temperature.

Motion	Amplitude (Angstrom)	Characteristic time
Vibration of bonded atoms	0.01 to 0.1	1-10 femtoseconds
Elastic vibration of globular protein	0.05 to 0.5	1 to 10 picoseconds
Surface sidechain rotation	5 to 10	10 to 100 picoseconds
Libration of buried sidechains	0.5	100 picoseconds to 1 nanosecond
Hinge bending, domain motion	1 to 5	100 picoseconds to 0.1 microsecond
Buried sidechain rotation	5	Microseconds to seconds

References:

1)Straatsma, T. P.; Berendsen, H. J. C. *J. Chem. Phys.* **1988**, *89*, 5876-5886.

2)Allen, M. P., Tildesley, D. J. *Computer Simulations of Liquids.*; Oxford University Press: Oxford, 1987.

3)Caldwell, J.; Agard, D.; Kollman, P. *Proteins-Structure Function and Genetics* **1991**, *10*, 140-148.

4)McCammon, J. A.; Harvey, S. C. *Dynamics of proteins and nucleic acids*; Cambridge University Press: Cambridge, 1987.

Chapter 9: Understanding substrate specificity in human and parasite phosphoribosyltransferases through calculation and experiment

Jed W. Pitera, Narsimha R. Munagala, Ching C. Wang and Peter A. Kollman*†

Graduate Group in Biophysics and

Department of Pharmaceutical Chemistry, School of Pharmacy

University of California, San Francisco

San Francisco, California 94143-0446

†This work was supported by the NIH through grants GM-29072 (PAK, JWP) and AI-19391 (NRM, CCW).

*to whom correspondence should be addressed c/o

Department of Pharmaceutical Chemistry

University of California, San Francisco

San Francisco, CA 94143-04446

(415)476-4637

fax: (415)476-0688, e-mail: pak@cgl.ucsf.edu

RUNNING TITLE: Understanding PRTase substrate specificity

This work was submitted for publication in December 1998.

Abstract

We present molecular dynamics (MD) simulations on two enzymes: a human hypoxanthine-guanine-phosphoribosyltransferase (HGPRTase) and its analog in the protozoan parasite *Tritichomonas foetus*. The parasite enzyme has an additional ability to process xanthine as a substrate, making it a hypoxanthine-guanine-xanthine phosphoribosyltransferase (HGXPRTase).(1) X-ray crystal structures of both enzymes complexed to guanine monoribosylphosphate (GMP) have been solved, and show only subtle differences in the two active sites.(2,3) Most of the direct contacts with the base region of the substrate are made by the protein backbone, complicating the identification of residues significantly associated with xanthine recognition. Our calculations suggest that the broader specificity of the parasite enzyme is due to a significantly more flexible base-binding region, and rationalize the effect of two mutations, R155E and D163N, that alter substrate specificity.(4) In addition, our simulations suggested a double mutant (D106E/D163N) that might rescue the D163N mutant. This double mutant was expressed and assayed, and its catalytic activity confirmed.

Our molecular dynamics trajectories were also used with a structure-based design program, Pictorial Representation Of Free Energy Changes (PROFEC), to suggest parasite-selective derivatives of GMP. Our calculations here successfully rationalize the parasite-selectivity of two novel inhibitors derived from the computer-aided design of Somoza, et. al.(5) and demonstrate the utililty of PROFEC in the design of species-selective inhibitors.

Phosphoribosyltransferases (PRTases) are enzymes that catalyze the addition of a nucleobase to the alpha-carbon of alpha-phosphoribosylpyrophosphate to form a nucleoside monophosphate and pyrophosphate (Figure 1). They also typically catalyze the reverse reaction at high efficiency. For many parasitic organisms, purine PRTases are essential enzymes of the purine salvage pathway. In the protozoan *Tritichomonas foetus* (*T. foetus*) the essential purine PRTase operates with relatively high efficiency on hypoxanthine (H;R2 = H), xanthine (X;R2 = O) and guanine (G;R2 = NH2). In contrast to the protozoan HGXPRTase, the corresponding human enzyme shows a substantial preference for hypoxanthine and guanine and minimal xanthine activity (human HGPRTase). Competitive inhibition data (Ki) values show that the human HGPRTase has a 100-fold reduced affinity for xanthine (250 μM) versus hypoxanthine or guanine (1.8 and 2.4 μM)(6). The kinetics of both enzymes have also been extensively studied(7-9), and kcat's, Km's, and catalytic efficiencies for the forward reactions of both human and parasite enzymes are summarized in Table 1. Detailed data for the reaction of xanthine with the human enzyme have not been reported in the literature, presumably due to the low affinity of the enzyme for this substrate. Also listed in Table 1 are the properties of two mutant parasite enzymes (R155E, D163N). These two mutants were designed by Munagala and Wang(4) to help understand the broad substrate specificity of the parasite enzyme. R155E is particularly interesting since this mutation distant from the active site serves to substantially reduce the ability of the enzyme to process xanthine, XMP and GMP (data not shown). Closer to the substrate, the backbone of residue 163 forms direct hydrogen bonds to the C2 substituent of the base. The D163N mutant is

interesting since the side chain of this residue does not form any direct contacts with the substrate. However, the substitution of asparagine for aspartic acid substantially decreases the xanthine activity of the HGXPRTase while only slightly affecting the other two substrates.

X-ray crystal structures of the human HGPRTase complexed with GMP and the parasite HGXPRTase-GMP complex were solved by Eads, et. al.(2) and Somoza, et. al.(3). The two enzymes show about 30% sequence identity. Both active sites are very similar in structure, and use almost identical residues to recognize the substrate. The two complexes are superimposed and shown in Figure 2, with several key residues labeled. The similarity of the two active sites does not answer why the parasite enzyme recognizes xanthine or XMP while the human enzyme does not. To attempt to answer this question, we carried out molecular dynamics (MD) simulations of the human-GMP, human-XMP, parasite-GMP and parasite-XMP complexes. In addition, we simulated the two mutant parasite complexes (R155E-XMP and D163N-XMP) in order to attempt to understand the effects of these alterations on enzyme specificity. Our molecular dynamics calculations provide a detailed, microscopic picture of the structure and dynamics of each complex, and are discussed in detail below. After this work was initiated, several structures of complete purine phosphoribosyltransferase catalytic complexes were solved (10,11). These structures show the positions of the substrate base, phosphoribosyl phosphate, and catalytic metal ions just prior to the formation of the nucleotide. As expected, the crucial protein contacts that recognize the base are similar in these new structures and the previous nucleotide complexes which we have simulated.

Methods

Setup and model building

The models of each complex were prepared using the LEAP module of AMBER 5.0(12). The PDB coordinates of the human-GMP (1hmp) and parasite-GMP (1hgx) complex were used as starting points for all calculations. They were modified as necessary to represent alternative substrates (XMP) or mutant structures (D163N, R155E). XMP was built onto the GMP coordinates by hand, while the mutant coordinates were built using the "swapaa" function of MidasPlus(13). A sulfate ion ($SO4^{-2}$) distant from the ligand binding site was deleted from the parasite structures. All ionizable amino acids were set to their most probable protonation state at pH 7.0, with no exceptions.

It was necessary to build molecular mechanics models of both XMP and GMP for our calculations, since these molecules are not part of the usual AMBER parameter set. While all necessary bond, angle, dihedral and van der Waals terms were already present in the Cornell, et. al.(14) force field, we had to derive charges for both substrates. The restrained electrostatic potential (RESP) method(15) was used to determine a set of point charges for each molecule that best fit their electrostatic potential. These electrostatic potentials were derived from single-point restricted Hartree-Fock quantum mechanical calculations carried out with the Gaussian 94 computer program(16). It was necessary to split the nucleotides into two parts: the base and sugar, which were treated as a single neutral species; and the phosphate group, which was treated as a singly charged anion (17). In both calculations, a 6-31G* basis set was used. The two calculations were combined to yield ribonucleoside phosphates each with a net charge of –1.0. The charges

and parameters used are available from the authors. In preparing our model of XMP, we considered only the 2-oxo tautomer, as enolic tautomers of the nucleobases are rarely populated in solution.(18)

Equilibration

Using these molecular mechanics models and the protein structures described above, including any counterions and crystallographic waters, we solvated each complex with a 25.0 Angstrom cap of TIP3P(19) water molecules centered on the substrate. Though both the human and parasite structures are dimeric enzymes, we concentrated our attentions on a single active site in each case. In fact, only atoms closer than 25 Angstroms from the substrate were permitted to move in each simulation.

Each complex was equilibrated and simulated using an identical protocol and the SANDER molecular dynamics package of AMBER 5.0(12). The molecular mechanics model described above was first subjected to 100 steps of steepest descent minimization followed by 1000 steps of conjugate gradient minimization to fix any errors introduced in the model-building process. After this minimization, a short (2 picoseconds (2 ps), heating from 0 to 300K) molecular dynamics run was started where only the water molecules were allowed to move while the protein, ligand, and counterions remained fixed. This served to equilibrate the water structure around the complex prior to the production runs. The equilibrated complex was then gradually heated from 0K to 300K over 8 ps. During this heating period, the protein and ligand heavy atoms were restrained to their original positions with weak (1.0 kcal/mol/A^2) positional restraints. Once the heating was over, the positional restraints were released, and each complex was simulated

for a total of 300 ps. For all runs, Berendsen temperature coupling(20) was used to maintain the system at its assigned temperature, while the SHAKE algorithm constrained the length of all hydrogen containing bonds to their equilibrium values. The latter serves to permit the use of a 2 femtosecond timestep in the dynamics calculations. In order to minimize computational expense, long-ranged energies and forces (electrostatic and van der Waals) were only calculated out to a fixed (residue-based) cutoff distance. In the past, typical molecular dynamics or Monte Carlo protocols have used 8 or 9 Angstrom cutoffs(21). The recent development of Particle Mesh Ewald (PME) electrostatics(22) and related algorithms(23) allow all long-range contributions to be included at a reasonable cost in periodic systems. Since our complexes were aperiodic, it was necessary to use a cutoff-based approach for our calculations. We initially simulated each complex using a 9 Angstrom cutoff. However, this resulted in rapid distortions of the binding complex including movement of XMP or GMP phosphate groups out of the phosphate-binding loop and into solution. These distortions are commonly seen when using cutoffs in simulations of highly charged systems(24). Due to the distortions, these calculations were discarded and a 14.0 Angstrom cutoff was used for all subsequent calculations. The longer cutoff resulted in much lower structural distortions and relatively stable structures for each complex. A similar effect has also been observed in complexes of isocitrate dehydrogenase(25). All calculations reported in this paper used the longer 14 Angstrom cutoff. Attempts to simulate the complex in vacuum with a 14 Angstrom cutoff and a distance-dependent dielectric of $4R_{ij}$ were unsuccessful due to unrealistic fluctuations of the ligand and Y156 or F186, the protein side chain that stacks atop the base portion of the ligand.

Analysis

The CARNAL trajectory analysis software of AMBER 5.0 was used to calculate the root-mean-square (RMS) deviations and interatomic distances for each trajectory. The MDANAL part of the AMBER 5.0 package was used to calculate the atomic fluctuations of each complex. Structures were displayed and compared using the MidasPlus graphical visualization software(13).

PROFEC

The PROFEC free energy extrapolation software(26) was used to suggest modifications to GMP or guanine that would increase the parasite selectivity of the resulting compound. Both the parasite-GMP and human-GMP trajectories were analyzed to find locations where the ligand could be favorably derivatized. This is done by calculating the free energy cost of inserting a test particle at various locations near the C2 of GMP. The difference of these two analyses yields a map of positions where GMP could be changed to yield a parasite-selective compound. For these calculations, an uncharged carbonyl oxygen-sized particle was used as the test particle. Once the map was calculated, it was projected on the parasite-XMP complex and displayed using MidasPlus, which also served to superimpose the DOCKed coordinates of the compounds described by Somoza, et. al.(5) The free energy map was contoured at levels of –2.0, 0.0 and +1.0 kcal/mol in order to show regions where modifications to the ligand would be favorable, neutral, or unfavorable, respectively.

Simulations

A total of six simulations were carried out to permit a detailed comparison of the structure and dynamics of human and parasite enzymes bound to each substrate (GMP, XMP) as well as model the behavior of two parasite mutants (D163N, R155E) bound to XMP. In all cases, the binding complex is well maintained, and the overall geometry of the initial model-built complex was preserved.

Site-directed Mutagenesis

Site-directed mutagenesis and expression and purification of the D106E/D163N double mutant was carried out as previously described(1)[(4). Oligonucleotide primers were designed, synthesized, and used with a Stratagene kit for site-directed mutagenesis. The plasmid (pBTfprt) containing the full-length gene encoding *T. foetus* was transformed into an *E. coli* mutant strain Sϕ606. Expression of the mutant *T. foetus* HGXPRTase gene in the plasmid was induced in the low-phosphate culture medium. The recombinant mutant protein was purified to homogeneity from the transformed cells, and steady state kinetic analysis was performed on the purified enzyme. The kinetic constants were obtained by monitoring the catalysis spectrophotometrically, as previously described(9).

Results

Simulations of wild-type complexes

The enzymological data (Table 1) show that the human enzyme substantially prefers guanine over xanthine, whereas the parasite HGXPRTase shows only a slight

preference. Our simulations of each wild-type enzyme bound to both substrates were intended to address this issue. The structural changes and fluctuations seen in each simulation help explain the observed substrate specificity.

A rough measure of structural change during each simulation is the root-mean-square (RMS) deviation of the protein backbone from its initial position. Compared to the corresponding parasite complexes, the protein backbones of the human complexes diverge further from the x-ray structure, as measured by RMS deviation (Figure 3a,3b). The average structures of the human-GMP and human-XMP simulations are shown superimposed on the x-ray structure in Figure 4, highlighting a few differences. The salt bridge between D193 and K68 is broken in the human-XMP complex, while it is preserved in the human-GMP complex. In the human-XMP complex, K68 moves to form a salt bridge with D134, which normally interacts with the sugar hydroxyls. In this complex, D193 is also displaced slightly away from the base, minimizing unfavorable interactions between its backbone carbonyl and the C2 carbonyl of XMP. In contrast, this same residue superimposes closely with the crystal structure in our human-GMP simulation, maintaining a good hydrogen bond between its main chain carbonyl and the GMP amino group. Despite these structural differences, both simulations show uniformly low fluctuations of the protein backbone in this ligand-binding region. Also, the phosphate- and sugar-binding regions of both complexes are well preserved, as expected for substrates that differ only at the C2 position.

In contrast to the human complexes, the parasite-XMP and parasite-GMP backbones show smaller RMS deviations from the initial x-ray structure (Figure 3b). However, the two parasite structures are more different from each other than the two

human structures are from each other, especially in the C2-pocket region. This is best shown by Figure 5, which superimposes the C2 pockets of each parasite complex. From this figure, one can see that both structures have diverged somewhat from the x-ray crystal structure, primarily due to the loss of the hydrogen bond between the D163 backbone carbonyl and the Y156 hydroxyl group. In the parasite-XMP complex, D163 has moved away from the C2 carbonyl of xanthine, allowing water molecules more access to donate hydrogen bonds to the substrate. With the Y156 hydrogen bond broken, D163 shifts to make a much better hydrogen bond to the GMP NH_2 in the parasite-GMP complex. Again, the phosphate- and sugar-binding regions of both the parasite-GMP and parasite-XMP complexes remain well-structured throughout the calculation.

Simulations of mutant complexes

In addition to the wild-type simulations, we carried out simulations on two mutants of the parasite enzyme bound to XMP. During the extensive enzymological and mutagenic studies of *T. foetus* HGXPRTase by Munagala et. al.(4,9), two anomalous mutations were observed. First, the mutation of arginine 155 to a glutamic acid (R155E) disrupts a salt bridge relatively distant (~15 Angstrom) from the C2 pocket, yet it substantially reduces the affinity of the enzyme for some C2-substituted substrates (xanthine, XMP and GMP) while influencing hypoxanthine/IMP binding much less. Our simulation of the parasite R155E-XMP complex shows no large structural differences from the wild type enzyme, and has a low overall RMS deviation (Figure 3c). However, the backbone fluctuations of residue 163 (and the C2 pocket) are substantially decreased

in the mutant complex. This suggests that R155E may exert its effect by reducing the ability of the parasite C2 pocket to reorganize and recognize C2-substituted substrates.

In contrast, the D163N mutant changes the amino acid that forms most of the C2 pocket in the parasite enzyme. Since the C2 pocket is largely formed by the backbone of residue 163, it is interesting that this mutation substantially affects the affinity of the parasite for XMP and xanthine (Table 1). In our simulations, exchange of the negatively charged carboxylic acid for the neutral amide group of asparagine causes substantial structural distortions of the ligand-binding complex. In the wild-type structure, D163 acts as a salt bridge partner for R169. In the D163N mutant, R169 lacks a nearby negatively charged salt bridge partner. Over the course of our simulations, this arginine moves towards the carboxylic acids of D103 and E102 which normally recognize the hydroxyl groups of the ligand. While we do not expect that these specific distortions occur in the actual D163N mutant, it does suggest that the deleterious effect of the D163N mutant is due to a substantial reorganization of the enzyme, rather than a specific contact with the substrate. In support of this observation, we must note that Munagala, et. al. found the D163E mutant to have almost wild-type activity(4). The D163E mutant preserves the negative charge at this position and presumably maintains the D163-R169 salt bridge as well.

Analysis of the simulations

The backbone (N,CA,C,O) atomic fluctuations (B-factors) of each complex are presented in Figure 6. They are shown for residues 153-173 (in the parasite enzyme; 183-203 in the human), which includes the amino acids that recognize the C6, N1, and

C2 positions of the base. Figure 6A shows the B-factors in this region for the two starting crystal structures. Both the human-GMP and parasite-GMP complexes show low fluctuations in this region according to the crystal structures. For comparison, panels B and C show the fluctuations calculated from our wild-type simulations. Here, both human complexes show the expected low fluctuations in this region. The parasite-GMP and parasite-XMP simulations show much larger fluctuations in this region. This suggests that the parasite enzyme has a far less rigid C2 pocket which permits it to recognize both the amino group of GMP and the carbonyl oxygen of XMP. While the main-chain carbonyl of residue D163/D193 provides an appropriate hydrogen-bond acceptor for GMP's amino group, the C2 pocket needs to reorganize to properly recognize XMP's carbonyl group. While buried hydrogen bonds are not necessarily expected to add to binding affinity, the failure to form an expected enzyme-substrate hydrogen bond can cost substantial affinity or stability(27). The necessary hydrogen bonds to recognize XMP appear to be provided largely by nearby water molecules in our simulations. In fact, our parasite-XMP simulations shows a very intermittent hydrogen bond between the D163 carbonyl and the hydroxyl of tyrosine Y156. Breaking this hydrogen bond allows the tyrosine and aspartic acid to separate enough that a water molecule can approach the oxygen of XMP from above the plane of the base.

Interestingly, the C2 pocket fluctuations of both mutant parasite complexes are decreased relative to their wild-type counterparts (Figure 6, panel D). In the case of R155E-XMP, this may be the mechanism by which the R155E mutation exerts its effects on some C2-substituted substrates (xanthine, XMP, GMP). In contrast, the D163N mutant shows moderate fluctuations in the backbone of residue 163, but slightly

increased fluctuations in residue 153-154 and the region around R165. Since this latter residue shifts substantially during our simulation, its increased fluctuations are expected.

To get a picture of the specific contacts that form the C2 pocket, we monitored the distance between nearby hydrogen bond donors and acceptors during our simulations. There are three major hydrogen bonds that can be formed with the base C2 substituent in both enzymes, as seen in Figure 2. First, the main-chain carbonyl of I157/I187 is seen to accept a hydrogen bond from GMP in both human and parasite crystal structures. Second, the backbone carbonyl of D163/D193 can also accept a hydrogen bond from the amino group of guanine. Third, the amide nitrogen of that same residue can donate a hydrogen bond to the C2 carbonyl of xanthine or XMP if they are present, though this would require some reorganization to achieve the optimal geometry. We monitored the time course of these three distances over each of our simulations, and found that only a few varied significantly over our trajectories. As expected, the D163 backbone – C2 distances are significantly longer in the two wild type parasite complexes. The averages and standard deviations of each distance are shown in Table 2, along with the values from the two crystal structures for comparison.

Both mutant complexes (R155E, D163N) also show very short distances between the amide nitrogen of residue 163 and the C2 oxygen of XMP, indicating that the base has "tipped down" in the C2 pocket to form this good hydrogen bond.

In addition to contacts along the edges of the base, we also monitored the distance between the N7 atom of each base and the oxygen atoms of aspartic acid D106/D137 (both shown in Figure 2). Proton transfer between this carboxylic acid and the N7

nitrogen has been implicated as the critical catalytic step in the human enzyme(28). Distances from the nearest carboxyl oxygen to N7 of each substrate are shown in Table 3.

These distances were calculated over the last 200 picoseconds of each trajectory, to allow for the initial relaxation of each structure. Although we are not necessarily simulating a catalytically competent complex (D106/D137 would need to be protonated to initiate a proton transfer to N7 of the base), the human-XMP and D163N-XMP complexes stand out in the above table as having larger average and maximum distances than the other complexes. Interestingly, these are two of the three complexes that show reduced affinity and catalytic activity (Table 1). The last, R155E, appears to be impaired due to its decreased fluctuations in the C2 pocket, as mentioned above.

Our observation of an increased D106-XMP N7 distance in the D163N-XMP simulation prompted the design of a double mutant D106E/D163N. The substitution of glutamic for aspartic acid at position 106 allows the double mutant to partially compensate for the deleterious effects of the D163N mutation. This double mutant was made, expressed and assayed by N. Mungala using the previously-described protocols for biochemical studies of the parasite HGXPRTase(1,4). The double mutant restores catalytic activity to the D163N enzyme, as shown in table 4. As well as supporting our model, the activity of the double mutant is additional evidence that the carboxylic acid at position 106 (137 in the human enzyme) is essential to the catalytic mechanism of these enzymes(28).

A model of the double mutant was hand-built from the structure of the simulated D163N-XMP complex. The side chain dihedrals of glutamic acid were kept to canonical (rotameric) values(29). The D106E substitution allows the catalytic carboxylic acid to be

placed within 2.5 Angstroms of the N7 nitrogen without significant distortion of the side chain. In contrast, the corresponding D163N-XMP structure shows a carboxylic acid-N7 distance of over 4 Angstroms.

PROFEC analysis

The PROFEC free energy estimation software(26) was used to suggest how GMP could be modified to yield a parasite-selective ligand. The free energy cost of adding a test particle (or potential modification) the C2 position of the base was evaluated for both the human-GMP and parasite-GMP trajectories. These two free energy "maps" were then combined to yield a picture of how GMP (or GMP-like ligands) could be modified to improve their parasite-selectivity, either by enhancing their affinity for the HGXPRTase, or by impairing their affinity for the human enzyme. Interestingly, there is a large region along the edge of the base by the N2 and N3 positions where our software suggests that modifications be made. This region is shown in Figure 7A. Since the parasite-GMP complex is much more solvent exposed along the N2/N3 edge of the base, it is not surprising that PROFEC picks this region for parasite-selective modifications. Additional atoms here would displace a water molecule in the parasite-GMP complex, while they would probably have a steric clash with D193 or K68 in the slightly better-packed human-GMP complex.

The recent computer-aided design of several parasite-selective ligands here at UCSF(5) allows us to test our observations. Somoza, et. al., found two lead compounds via computer database screening with the DOCK program(30). These two compounds, a phthalic anhydride-nitrobenzene and an indol-2-one-nitrobenzene, are shown in Figure

7B, superimposed upon GMP in their DOCKed conformations(31). Figure 7B also includes the PROFEC contour map showing where GMP could be modified to increase or decrease its parasite selectivity. Interestingly, the two compounds superimpose somewhat with the parasite-favoring region shown in Figure 7A. Moreover, the indol-2-one projects substantially into the region that we expect to be unfavorable for parasite selectivity, while the phthalic anhydride follows the parasite-selective contour much more closely. This only becomes significant when considered in the context of Table 5, which lists the IC50s and selectivity (IC50/IC50) of both compounds as reported by Somoza, et. al. Here we see that the indol-2-one is in fact relatively nonspecific, while the phthalic anhydride shows a greater than threefold preference for the parasite enzyme.

Discussion and Conclusions

In any molecular modeling study, one must be aware of the limitations of the model used and how they relate to the questions of interest. In our case, we were interested in the qualitative details of molecular recognition between HGPRTase/HGXPRTase and GMP or XMP. These two substrates only differ by the substitution of a carbonyl (xanthine, XMP) for an amino group (guanine, GMP) at the C2 position. In addition, the two enzymes we studied show only subtle structural differences in their C2-binding pockets. Given these observations, we chose a computational model that allowed us to simulate a number of enzyme-substrate complexes for a reasonable time, and took care to ensure that the representation was sufficient to maintain the structure of each complex relatively well. Since we were interested in the local structure and dynamics of each active site, we chose to model those regions in detail while holding

most of each enzyme fixed. While our calculations are relatively short (300 picoseconds each), they appear to be sufficient to highlight the differences between the enzyme-substrate complexes. In each case, enzyme-substrate interactions and geometries are well converged, with the possible exception of our D163N-XMP complex, which shows substantial structural deviations. Given the representation used, we do not expect that we have definitively divined the structure of each simulated complex, nor completely modeled the overall dynamics of either enzyme in solution. Our calculations have, however, allowed us to rationalize the ability of *T. foetus* to process xanthine, explain the effects of two mutations, and suggest the basis for the parasite-selectivity of different ligands.

With the above limitations in mind, our calculations have revealed a plausible model for the ability of the *T. foetus* HGXPRTase to process XMP. The residues of the C2 pocket, specifically D163/D193, show much higher fluctuations in the parasite enzyme when compared to the human enzyme. This is particularly significant given the lower overall RMS deviation of the parasite simulations from their starting structures. The parasite enzyme appears to be able to reorganize to accommodate both the C2 carbonyl of XMP and the C2 amino group of GMP, while the C2 pocket of the human enzyme is relatively "frozen" in a conformation that only recognizes GMP. In addition, the R155E parasite mutant shows reduced fluctuations in this region, compatible with its reduced ability to process xanthine – it too becomes "frozen" in the GMP-binding conformation. It is important to note (Figure 6) that neither the parasite/GMP or human/GMP crystal structures show this difference in their C2 pocket temperature factors. However, the most important observation is probably the difference in human-

XMP and parasite-XMP fluctuations, since this helps to explain why the parasite enzyme can recognize XMP while the human cannot. Since there are no crystal structures of either of these complexes, our simulations provide the only structural information of how xanthine and XMP are recognized.

In contrast to R155E, where our simulations show a stable structure with reduced fluctuations from the wild type enzyme, our D163N model is substantially distorted. The removal of a negative charge associated with the aspartic acid to asparagine mutation removes a salt bridge necessary for structuring R169. Without its salt-bridging partner, R169 is drawn toward the nearby negative charges of E102 and D103. This attraction distorts the structure of the complex. Most importantly, it significantly increases the distance between the N7 nitrogen of XMP and the catalytic carboxylic acid of aspartic acid D106. This model prompted the design and testing of a D106E/D163N double mutant. The increased length of the catalytic carboxylic acid at position 106 restores catalytic activity to the parasite enzyme. While we do not expect that our simulation precisely describes the structure of the D163N/XMP complex, our calculations are substantially validated by the efficacy of the D106E/D163N double mutant. Interestingly, the homologous residues of D163 and R169 form most of the magnesium ion binding site seen in the enzyme-base-PRPP structure solved by Focia, et. al.(11) This led them to conclude, like our study, that the D163N mutation exerts its effects through significant structural changes in the enzyme (distortion of the Mg++ binding site) rather than direct or water-mediated contacts with the base.

In addition to understanding the mechanistic basis of xanthine specificity and providing a model of the specific effect of some parasite mutants, our calculations have

shown how GMP could be modified to form a xanthine-selective substrate or inhibitor. These PROFEC calculations indicate a region along the edge of the base, near the C2 substituent and the N3 position, where added derivatives would be preferred in the parasite enzyme but unfavorable in the human. Interestingly, this region is occupied by the parasite-selective ligands recently developed by the Kuntz group at UCSF. As far as we are aware, this represents the first case where free energy extrapolation methods have been used to study species-selective ligand modifications. While PROFEC calculations only yield qualitative suggestions, our results here imply that they will be a useful tool in species-selective drug design.

Acknowledgements

JWP received support from a NSF predoctoral fellowship and the University of California Office of the Chancellor.

Table 1: Kinetic parameters for human HGPRTase and *T. foetus* HGXPRTase activity

Table 1a[a]					
Enzyme	Substrate	Km,app (μM)	Kcat (sec^{-1})	kcat/Km ($\mu M^{-1}sec^{-1}$)	Relative rate
Human	Hypoxanthine	1.9 +/- 0.3	8.4	2.7	1
Human	Xanthine	NA	NA	NA	NA
Human	Guanine	3.1 +/- 0.9	14.2	7.5	2.7
Table 1b[b]					
Enzyme	Substrate	Km (μM)	Kcat (sec^{-1})	kcat/Km ($\mu M^{-1}sec^{-1}$)	Efficiency
Parasite	Hypoxanthine	3.05 +/- 0.54	8.92 +/- 0.46	2.92	1
Parasite	Xanthine	6.08 +/- 0.81	4.82 +/- 0.8	0.78	0.26
Parasite	Guanine	2.4 +/- 0.74	2.48 +/- 0.24	1.03	0.35
Table 1c[c]					
Parasite D163N	Hypoxanthine	4.92 +/- 0.8	8.95 +/- 0.67	1.81	0.62
	Xanthine	>300	--	--	--
	Guanine	9.13 +/- 1.1	2.17 +/- 0.13	0.24	0.08
Parasite R155E	Hypoxanthine	2.06 +/- 0.51	10.22 +/- 0.8	4.96	1.69
	Xanthine	95.8 +/- 18.4	33.05 +/- 2.6	0.34	0.11
	Guanine	3.01 +/- 0.23	25.98 +/- 1.9	8.63	2.95

[a] data from (8)
[b] data from (9)
[c] data from (4)

Table 2: Distances associated with enzyme-substrate hydrogen bonds (in Angstroms)

Table 2			
Simulation/Structure	I157/I187 O – R2	D163/D193 O – R2	D163/D193 N – R2
Human-GMP (xray)	3.0	3.0	4.1
Human-GMP	3.5 +/- 0.4	3.3 +/- 0.4	4.0 +/- 0.5
Human-XMP	3.8 +/- 0.2	3.6 +/- 0.3	3.9 +/- 0.7
Parasite-GMP (xray)	3.3	3.2	3.8
Parasite-GMP	3.2 +/- 0.3	4.7 +/- 1.3	4.1 +/- 0.7
Parasite-XMP	3.6 +/- 0.2	4.0 +/- 0.4	4.2 +/- 0.7
R155E-XMP	3.8 +/- 0.2	3.4 +/- 0.2	2.1 +/- 0.4
D163N-XMP	3.8 +/- 0.2	3.4 +/- 0.3	2.0 +/- 0.3

Table 3: Distances between the catalytic D106/D136 and the substrate N7 nitrogen (in Angstroms)

Table 3			
Simulation	Minimum distance	Average	Maximum
Human-GMP	2.8	3.9	5.4
Human-XMP	3.0	4.4	6.0
Parasite-GMP	2.8	3.5	4.8
Parasite-XMP	2.9	3.5	4.7
R155E-XMP	2.9	3.6	5.0
D163N-XMP	3.0	4.2	6.0

Table 4: Kinetic parameters of the D106E/D163N double mutant

Table 4				
Enzyme	Substrate	Kcat (sec^{-1})	Km (μM)	Kcat/Km ($\mu M^{-1}sec^{-1}$)
Wild type	Xanthine	4.8 +/- 0.8	6.08 +/- 0.8	0.78
D163N	Xanthine	>300	-	-
D106E/D163N	Xanthine	0.39	21.5	0.18

Table 5: IC50s of lead compounds studied by Somoza, et. al.(5)

Table 5			
Compound	Parasite IC50 (μM)	Human IC50 (μM)	Ratio (Hum/Par)
Phthalic anhydride	300	>1000	>3
Indol-2-one	240	200	0.83

Bibliography

1. Chin, M. S., and Wang, C. C. (1994) *Mol Biochem Parasitol* **63**(2), 221-9
2. Eads, J. C., Scapin, G., Xu, Y., Grubmeyer, C., and Sacchettini, J. C. (1994) *Cell* **78**(2), 325-34
3. Somoza, J. R., Chin, M. S., Focia, P. J., Wang, C. C., and Fletterick, R. J. (1996) *Biochemistry* **35**(22), 7032-40
4. Munagala, N. R., and Wang, C. C. (1998) *Biochemistry* **37**(47), 16612-16619
5. Somoza, J. R., Skillman, A. G., Jr., Munagala, N. R., Oshiro, C. M., Knegtel, R. M., Mpoke, S., Fletterick, R. J., Kuntz, I. D., and Wang, C. C. (1998) *Biochemistry* **37**(16), 5344-8
6. Krenitsky, T., Papaioannou, R., and Elion, G. (1969) *Journal of Biological Chemistry* **244**(5), 1263-1270
7. Xu, Y., Eads, J., Sacchettini, J. C., and Grubmeyer, C. (1997) *Biochemistry* **36**(12), 3700-12
8. Keough, D. T., Ng, A. L., Winzor, D. J., Emmerson, B. T., and de Jersey, J. (1999) *Mol Biochem Parasitol* **98**(1), 29-41
9. Munagala, N. R., Chin, M. S., and Wang, C. C. (1998) *Biochemistry* **37**(12), 4045-51
10. Vos, S., Parry, R., Burns, M., deJersey, J., and Martin, J. (1998) *Journal of Molecular Biology* **282**(Oct 2), 875-889
11. Focia, P., Craig, S., and Eakin, A. (1998) *Biochemistry* **37**(49), 17120-17127
12. Case, D. A., Pearlman, D. A., Caldwell, J. W., Cheatham, T. E., Ross, W. S., Simmerling, C. L., Darden, T. A., Merz, K. M., Stanton, R. V., Cheng, A. L.,

Vincent, J. J., Crowley, M., Ferguson, D. M., Radmer, R. J., Seibel, G. L., Singh, U. C., Weiner, P. K., and Kollman, P. A. (1997) *University of California, San Francisco*

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

14. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1996) *J Amer Chem Soc* **118**(9), 2309-2309

15. Bayly, C. I., Cieplak, P., Cornell, W. D., and Kollman, P. A. (1993) *J. Phys. Chem.* **97**(40), 10269-10280

16. Frisch, M., Trucks, G., Schlegel, H., Gill, P., Johnson, B., Robb, M., Cheeseman, J., Keith, T., Petersson, G., Montgomery, J., Raghavachari, K., Al-Laham, M., Zakrzewski, V., Ortiz, J., Foresman, J., Peng, C., Ayala, P., Chen, W., Wong, M., Angres, J., Replogle, E., Gomperts, R., Martin, R., Fox, D., Binkley, J., Defrees, D., Baker, J., Stewart, J., Head-Gordon, M., Gonzales, C., and Pople, J. (1995), 94 D.3 Ed., Gaussian, Inc., Pittsburgh, PA

17. Eksterowicz, J. (1997)

18. Saenger, W. (1984) *Principles of Nucleic Acid Structure*. Springer Advanced Texts in Chemistry (Cantor, C. E., Ed.), Springer-Verlag, New York

19. Jorgensen, W. L., Chandrasekhar, J., Madura, J., Impey, R. W., and Klein, M. L. (1983) *J. Chem. Phys.* **79,** 926

20. Berendsen, H. J. C., Potsma, J. P. M., van Gunsteren, W. F., DiNola, A. D., and Haak, J. R. (1984) *J. Chem. Phys.* **81,** 3684-3690

21. McCammon, J. A., and Harvey, S. C. (1987) *Dynamics of proteins and nucleic acids*, Cambridge University Press, Cambridge

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

23. Luty, B. A., Tironi, I.G., van Gunsteren, W.F. (1995) *J. Chem. Phys* **,** 3014-3021

24. Cheatham, T. E., III, Miller, J. L., Fox, T., Darden, T. A., and Kollman, P. A. (1995) *J. Amer. Chem. Soc.* **117**(14), 4193-4194

25. Garcia-Vilorca, M., and Kollman, P. A. (1998) *International Journal of Quantum Chemistry (submitted)*

26. Radmer, R. J., and Kollman, P. A. (1998) *J Comput Aid Molec Design* **12**(3), 215-227

27. Fersht, A. R. (1985) *Enzyme Structure and Mechanism*, W. H. Freeman and Co., New York

28. Xu, Y., and Grubmeyer, C. (1998) *Biochemistry* **37**(12), 4114-24

29. Dunbrack, R. L., and Karplus, M. (1993) *J Mol Biol* **230**(2), 543-574

30. Ewing, T. J. A., and Kuntz, I. D. (1997) *J Comput Chem* **18**(9), 1175-1189

31. Oshiro, C.

Figure Legends

Figure 1: Schematic of the reaction carried out by the human and parasite phosphoribosyltransferases.

Figure 2: Superimposition of the human-GMP and parasite-GMP active sites from their crystal structures. The parasite complex is colored by atom type, and the human complex is displayed in cyan. Key protein residues and substrate atoms mentioned in the text are labeled, with the human residue number in parentheses. Several enzyme-substrate hydrogen bonds are highlighted for both complexes. The additional salt bridge between D193 and K68 in the human structure is also displayed.

Figure 3: Carbon-alpha root mean square deviation of each simulation from its starting structure. Complexes with GMP are solid lines; complexes with XMP are graphed as long dashed lines.

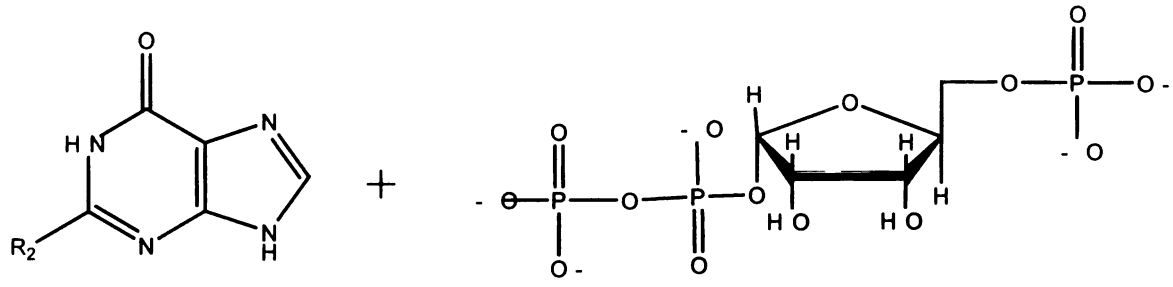
Figure 4: Superimposition of the human-GMP crystal structure (colored by atom) with the human-GMP (cyan) and human-XMP (yellow) structures averaged over each simulation. The enzyme-substrate hydrogen bonds shown in Figure 2 are also displayed for the x-ray coordinates, for reference.

Figure 5: Superimposition of the parasite-GMP crystal structure (colored by atom) with the parasite-GMP (cyan) and parasite-XMP (yellow) structures averaged over each simulation. The enzyme-substrate hydrogen bonds shown in Figure 2 are also displayed for the x-ray coordinates, for reference.

Figure 6: Protein backbone atomic fluctuations (B-factors) for each set of coordinates. These are only plotted for the residues near the C2 pocket; 153-173 in the parasite and 183-203 in the human enzyme. The parasite numbering is used in this figure and the human enzyme data has been translated to superimpose corresponding residues. For each residue, B-factors are listed for the N, CA, C, and O atoms of the protein backbone.

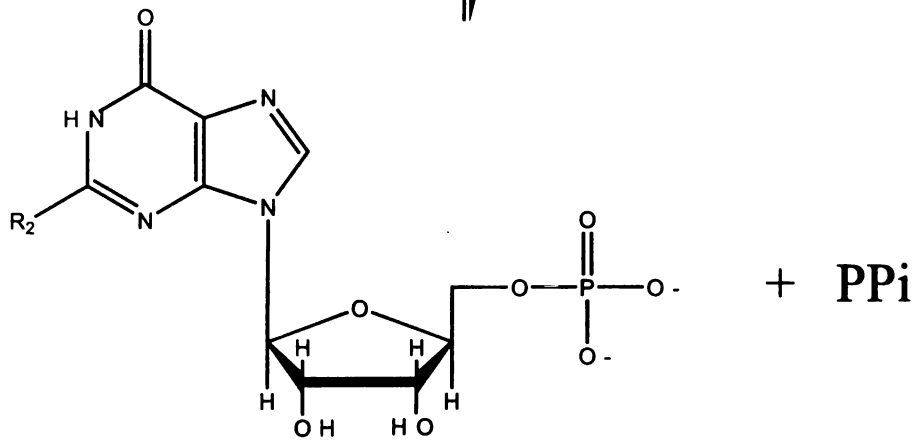
Figure 7a: PROFEC contour map of parasite selectivity. A slice through the contour map of suggestions provided by the PROFEC software is shown superimposed on the parasite-bound structure of GMP. The green contours correspond to regions where modifications are predicted to enhance parasite selectivity by –2.0 kcal/mol. The black contours correspond to the 0.0 kcal/mol contour, while the red contour is unfavorable by +1.0 kcal/mol. Since this is a contour map, addition of heavy atom modifications to GMP within the black contour is expected to increase parasite selectivity, while derivatives that place heavy atoms beyond the black contour or in the red regions are expected to show decreased parasite selectivity.

Figure 7b: PROFEC/DOCK comparison. The DOCKed binding conformations of the phthalic anhydride (yellow) and indol-2-one (cyan) compounds found by Somoza, et. al. are superimposed on the parasite-bound structure of GMP and the same PROFEC map used above. Again, the contour levels are green (-2.0 kcal/mol), black (0.0 kcal/mol) and red (+1.0 kcal/mol), while several thicknesses of the map have been drawn to aid comparison. Both compounds place several atoms in the favorable region suggested by PROFEC. In addition, the indol-2-one projects significantly beyond the first black contour and into the unfavorable region, while the phthalic anhydride only projects slightly past this region. This is compatible with the higher parasite-selectivity seen for the phthalic anyhydride.



base
R_2 = H: H
R_2 = NH_2: G
R_2 = O: X

α-PRPP



+ PPi

nucleoside monophosphate
R_2 = H: HMP
R_2 = NH_2: GMP
R_2 = O: XMP

Figure 1

Figure 2

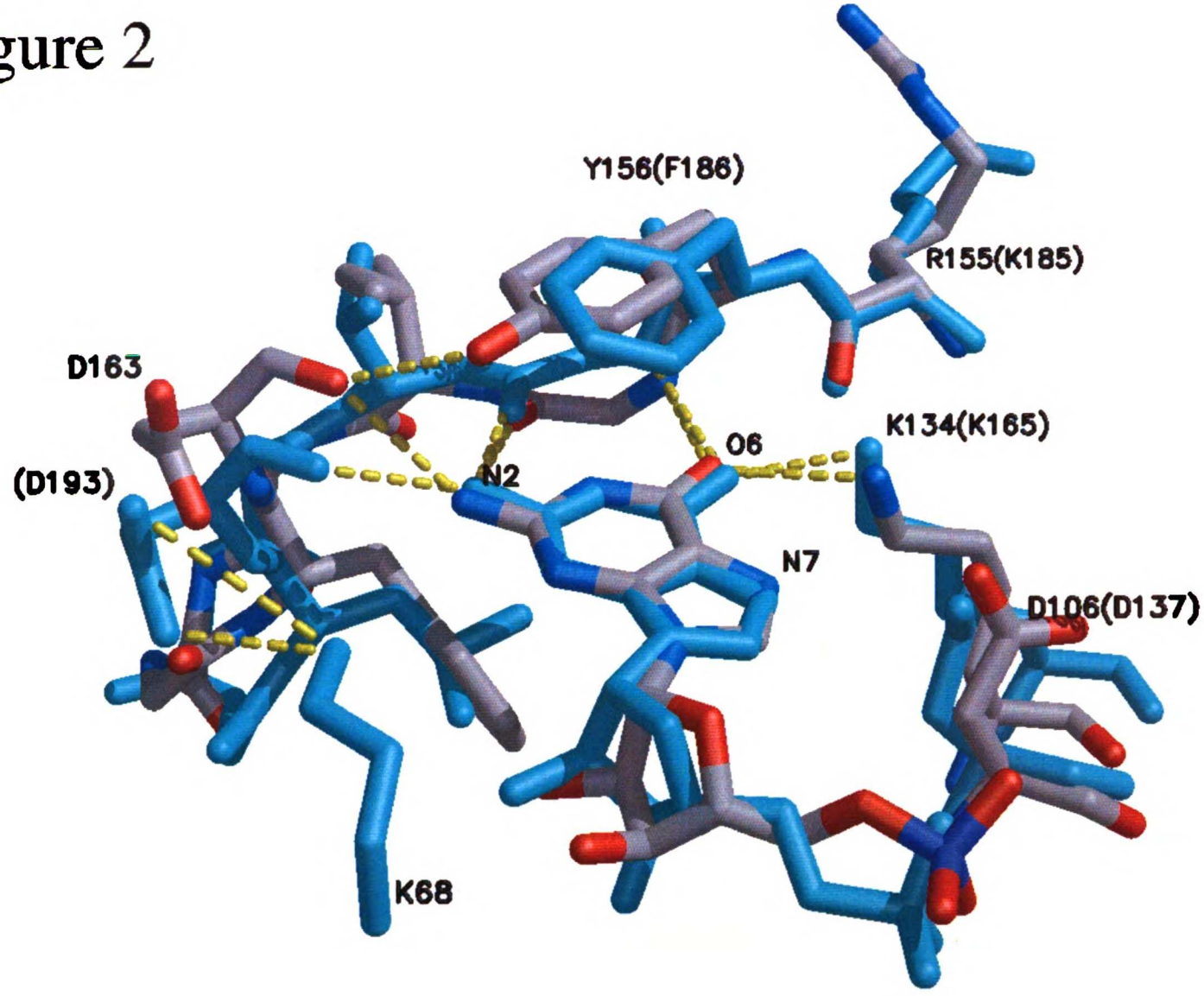
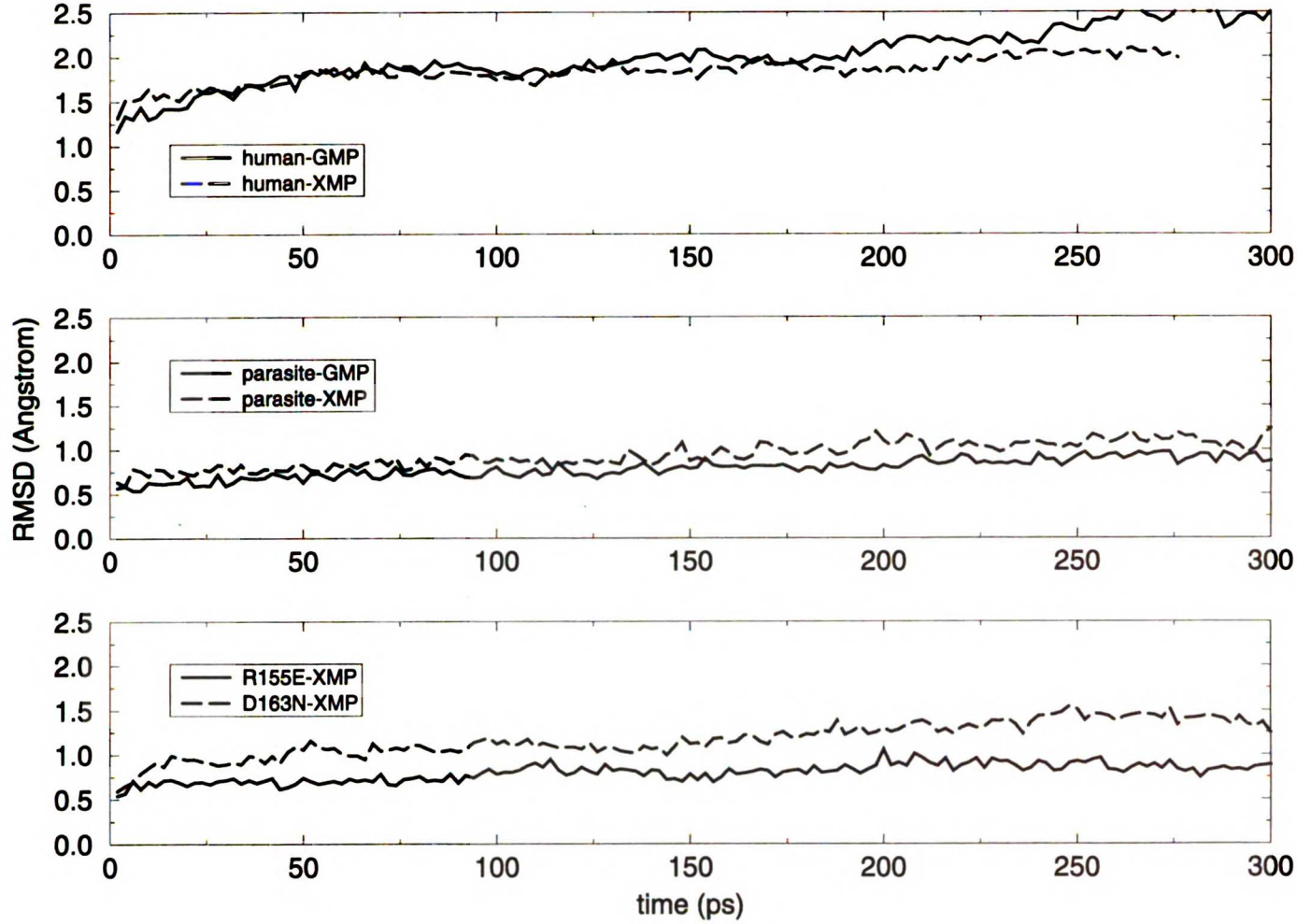


Figure 3: CA RMSD



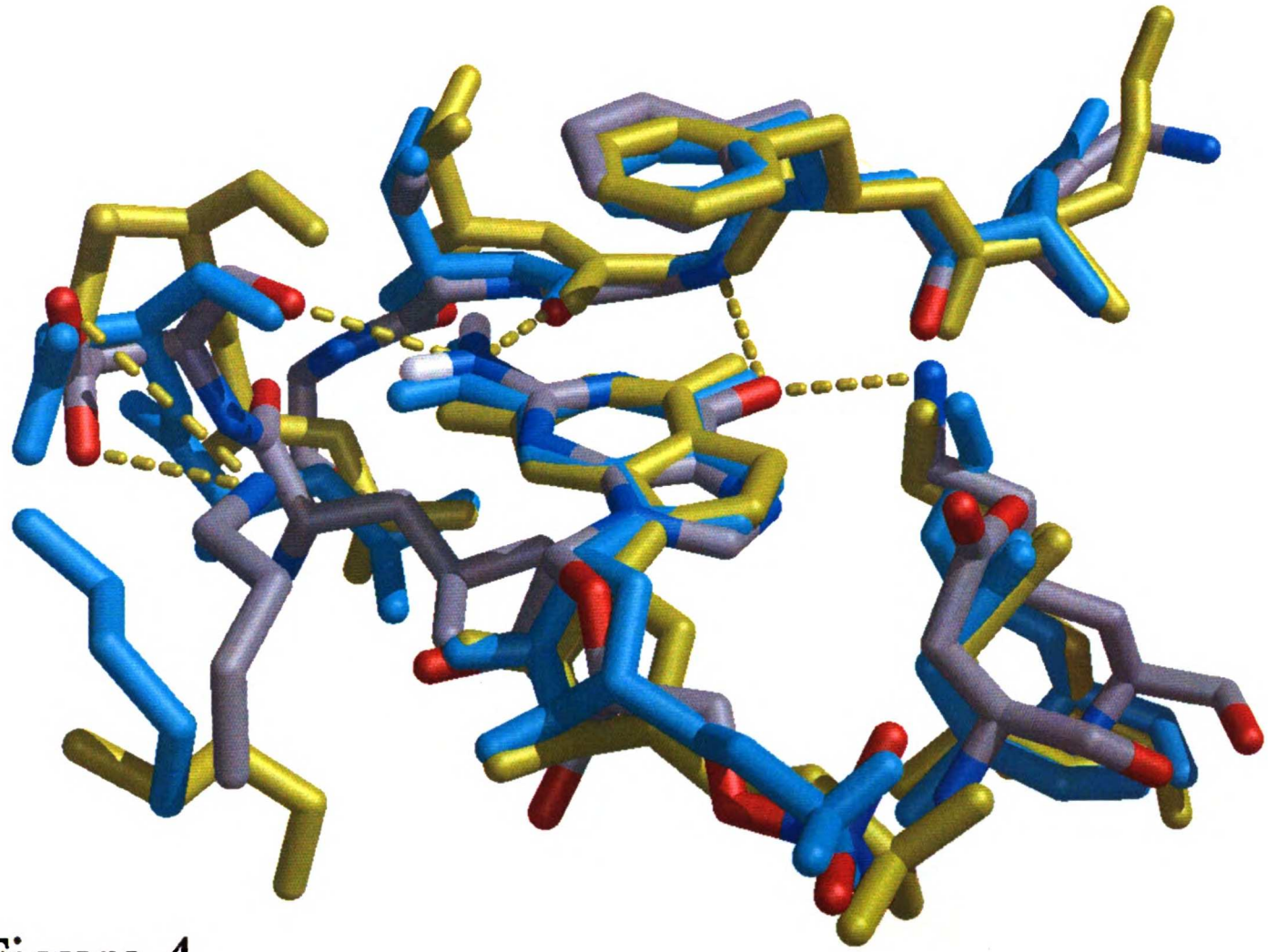


Figure 4

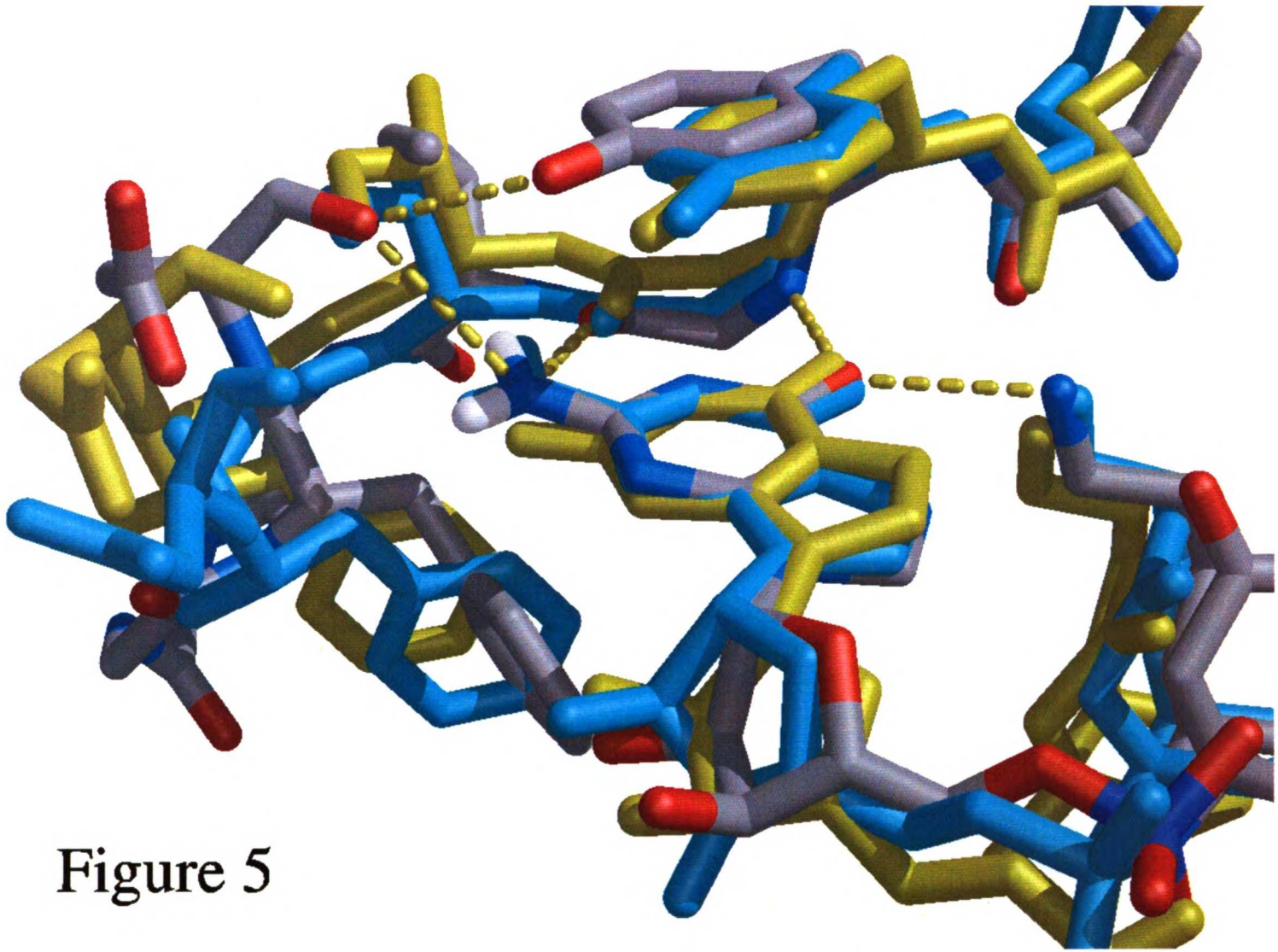


Figure 5

Figure 6

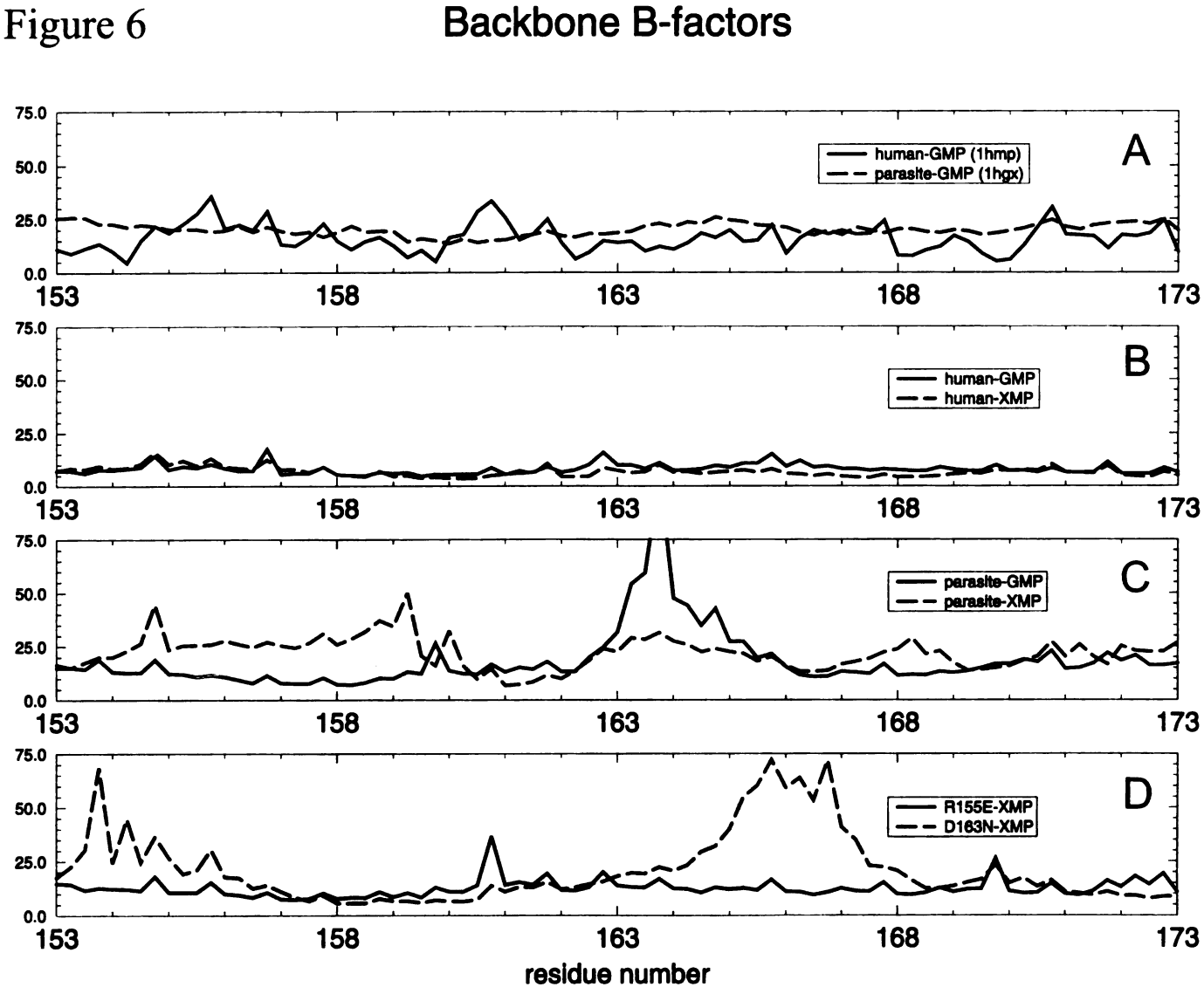

Backbone B-factors
75.0
50.0
25.0
0.0
153
158
163
168
173
human-GMP (1hmp)
parasite-GMP (1hgx)
A
human-GMP
human-XMP
B
parasite-GMP
parasite-XMP
C
R155E-XMP
D163N-XMP
D
residue number

Figure 7

Chapter 10: Future perspectives

The future of CMC/MD

The CMC/MD software that I developed for my thesis is scheduled to be included in the 6.0 release of the AMBER molecular mechanics/molecular dynamics package[1]. It will most likely be included as a compile-time option for the SANDER molecular dynamics program, since the potential masking associated with the CMC/MD steps requires considerable computational overhead. At present, it is not compatible with SANDER's Particle Mesh Ewald (PME)[2] method of calculating long-range electrostatic contributions. Ewald sums, by their very nature, calculate the long-range electrostatic terms for each atom in the system in a single calculation. This collective nature is at odds with the mixed system – some surrounding residues, one real MC residue, several "ghost" MC residues – simulated with CMC/MD. For a CMC/MD calculation with nine MC residues, nine Ewald sums are required for every force evaluation or every Monte Carlo step. As a single PME calculation is still a 20-50% additional cost versus the corresponding cutoff calculation, our prospective nine residue CMC/MD-PME calculation would be 180-450% slower than its cutoff counterpart. The judicious use of some approximations might reduce the number of Ewald sums necessary. For example, if PME is used for the forces and energies of the "real" system but a cutoff is used to treat the "ghost" MC residues, the cost of a more-accurate electrostatic representation is reduced somewhat. MC steps could be based on the cutoff interaction energy (cheap) or the full PME interaction energy (more accurate but computationally more expensive). The above approach would be very suitable for comparing the association of similarly-

charged ligands to a highly charged system – netropsin and analogs binding to DNA, for example.

Mean-field dynamics, Representative dynamics

The ability of the CMC/MD software to perform molecular dynamics with a number of similar ligands occupying the same position in a single protein active site offers some new opportunities in free energy estimation. Instead of doing CMC/MD, with its accumulating populations of different chemical species, one can imagine running a single simulation that samples coordinates for many ligands. Post-processing of the coordinates from this simulation could be used in a perturbation method [3] or linear interaction energy estimation method [4] to estimate the free energies of many ligands from a single simulation, and do so more rapidly than CMC/MD. Unlike CMC/MD, however, these approaches will not serve to rapidly sort or rank ligands based on their binding free energies. In addition, they do not correspond to any proper thermodynamic ensemble. I have implemented two types of dynamics using the CMC/MD software that will be used for these types of simulations. It should be noted that these methods are similar in spirit to the "ensemble dynamics" facility provided by the SPASMS molecular simulation package[5].

The first method, called "mean-field dynamics" is also somewhat related to the Locally Enhanced Sampling (LES) method of Elber, et. al.[6] With LES, a portion of the simulated system (like the ligand) is divided into N equivalent copies. Each copy interacts in a mean-field way with the remainder of the system – forces and energies are divided by 1/N. It has been formally proven that global minima of the LES system

correspond to global minima of the real (non-LES) system, prompting its use in optimization calculations, like simulated annealing of small peptides or protein side chain loops[7]. In our "mean-field" dynamics, N ligands are simulated in a binding site. Each ligand feels the full interaction with the protein, but the protein only feels 1/N of the normal interaction with each ligand. Hopefully, this will allow each ligand to sample numerous low-energy binding conformations while preventing the protein from "organizing" around one ligand to the exclusion of others. The input specification for mean-field dynamics is detailed in Appendix 4.

The second method, "representative dynamics". is somewhat simpler than mean-field dynamics. From the family of N ligands, a single "representative" ligand is selected. A calculation is carried out with all N ligands and the receptor, where the "representative" ligand interacts fully with the receptor. The receptor feels the interaction with the representative ligand, but does not feel the forces from any of the other ligands. The remaining ligands, however, feel the full force of their interaction with the receptor. "Representative dynamics" is exactly equivalent to a CMC/MD calculation with forces on the ghost ligands but where the "real" ligand is never changed – i.e. no Monte Carlo steps are carried out. Again, the input specification for representative dynamics is detailed in Appendix 4.

The future of free energy calculations

Interestingly, as my thesis has progressed, interest in free energy calculations for binding predictions has waned somewhat in the computational chemistry community. To some extent, it appears to be perceived as a "solved problem" – traditional free energy

methods like FEP and TI are assumed to give the correct answer for any two-ligand comparison, provided computer power is available to run the calculation for a sufficiently long time. The presence of multiple conformational minima separated by significant free energy barriers, whether in the ligand or in the receptor, is probably the major limiting factor in such calculations today (Chapter 7, references [8,9]). Substantial advances have been made in the comparison of charged and neutral systems. Both Resat & McCammon and Archontis, et. al. have shown that an extended thermodynamic cycle involving a continuum electrostatics calculation (essentially a Born correction) in addition to a thermodynamic integration step permits quantitative free energy calculations that compare neutral and net charged species[10,11].

A second force working against interest in free energy calculations for binding predictions is the explosive adoption of combinatorial chemistry methods and the associated high-throughput screening methodologies. The ability to rapidly (and cheaply) synthesize and test thousands to hundreds of thousands of compounds has changed the role of the computational chemist in a pharmaceutical context. Instead of providing insight to explain and extend structure-activity relationships derived from traditional medicinal chemistry compound series, or make single compound suggestions for synthesis, computational chemists are now expected to take an active role in library design, monomer selection, and data analysis from high-throughput screening runs. Much of "computational chemistry" in these contexts consists of statistical analysis using large datasets and experimental design, making use of computational chemistry tools such as conformational analysis and pharmacophore mapping in conjunction with methods from other fields, including information theory and statistics. However, these

alternative advances primarily impact lead discovery, and there is a recognition that the "solved problem" of pairwise free energy calculations provides a powerful tool for the subsequent optimization of a lead compound, especially given the increased availability of high-resolution structural information.

While direct industrial focus on free energy calculations has faded somewhat, the present robustness of traditional free energy methods, coupled with developments like CMC/MD, has opened new academic arenas for free energy calculations. The ability of free energy calculations to decompose contributions to relative free energies from different phases (vacuum versus solution or folded versus unfolded states, for example) lets them provide unique insights into fundamental physical processes. Nonadditivities in partitioning behavior; the competition between hydrophobicity, local strain, and internal entropy in peptide solvation; the role of conformational entropy and the "macrocyclic effect" in ligand binding – all these are fertile areas for new free energy calculations.

References:

1)Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E.; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Ferguson, D. M.; Radmer, R. J.; Seibel, G. L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. *University of California, San Francisco* **1997**.

2)Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089-10092.

3)Radmer, R. J.; Kollman, P. A. *J Comput Chem* **1997**, *18*, 902-919.

4)Aqvist, J.; Medina, C.; Samuelsson, J. E. *Protein Eng* **1994**, *7*, 385-391.

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

6)Elber, R., Karplus, M. *Journal of the American Chemical Society* **1990**, *112*, 9161-9175.

7)Simmerling, C. L., Elber, R. *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 3190-3193.

8)Ota, N., Brunger, A. *Theor Chem Acc* **1997**, *98*, 171-181.

9)Simmerling, C., Fox, T., Kollman, P.A. *Journal of the American Chemical Society* **1998**, *120*, 5771-5782.

10)Resat, H.; McCammon, J. *Journal of Chemical Physics* **1998**, *108*, 9617-9623.

11)Archontis, G.; Simonson, T.; Moras, D.; Karplus, M. *Journal of Molecular Biology* **1998**, *275*, 823-846.

Appendix 1: Pseudocode for the CMC/MD algorithm

Control statements are in *ITALICS*; important variable names are UPPERCASE; time, temperature, energy and forces are highlighted in **BOLD**.

Read standard molecular dynamics input
    Topology file (potential function, connectivity)
    Dynamics input (temperature coupling, timestep, simulation length, constraints or restraints, cutoff, dielectric, number of runs, run length)
Initialize CMC/MD variables to default values
Read CMC/MD input
 FILENAMES
    MCFIL (file for Monte Carlo statistics)
    MCRST (file for restart/reinitialization of CMC/MD)
 CONTROL VARIABLES
    IMCDO (main CMC/MD control variable)
    IMCRST (initialize/restart CMC/MD)
    IMCSOL (control variable for solvation offset, adaptive CMC/MD)
    IMCCP (control variable for multicopy CMC/MD)
    IMCNS (Monte Carlo step frequency)
    IMCINT (control variable for inclusion of intramolecular energies)
    EMCIF (floating point multiplier for intramolecular CMC/MD forces felt by MC residues)
    EMCNF (floating point multiplier for nonbonded CMC/MD forces felt by MC residues)
    EMCSF (floating point multiplier for nonbonded CMC/MD forces felt by non-MC residues)
    EMCOFE (floating point multiplier for 1-4 electrostatic energy included in MC energy)
    EMCOFN (floating point multiplier for 1-4 nonbonded energy included in MC energy)
    IMCFIL (frequency of updates to Monte Carlo history file MCFIL)
 MONTE CARLO VARIABLES (block format)
    IMCFRS (residue number of first MC residue)
    IMCLST (residue number of last MC residue)
    IMCCUR (residue number of current MC residue)
    IMCNS (Monte Carlo step frequency)
    Loop over MC residues, for each read
        EMCREO (solvation or adaptive offset for each MC residue)
        DMCWIN (width of trial move probability window for this residue, read if IMCCP > 2)
        IMCSET (set-membership for this residue, read if IMCCP > 3)
IF IMCDO is not -1, set up Monte Carlo calculation
    Loop over atoms
        Assign IMCMSK(I) for each atom.
            IF the atom is in a non-MC residue,
                IMCMSK(I) = 0
            IF the atom is in a MC residue,
                IMCMSK(I) = residue number – IMCFRS + 1
    Loop over MC residues
        IF DMCWIN has not been read, assign DMCWIN for this residue
            DMCWIN(I) = 1/N * (residue number – IMCFRS + 1)
        Initialize EMCRES(I), MC energy accumulator for each residue
        Initialize IMCCNT(I), number of MC counts for each residue
        Initialize PMCRES(I), Boltzmann probability accumulator for residue I
        Initialize DMCRES(I), Boltzmann probability accumulator for residue I when EMCREO offset is subtracted
    IF IMCRST = 1, read MCRST input file

Read IMCCNT, DMCRES, and EMCREO for each residue

BEGIN Enter main dynamics loop

R(T), T Calculate energies and forces; bonds, angles, dihedrals, vdW and electrostatics

Bonds

Loop over I,J bonded pairs

Calculate bond interaction

Look up potential function masking in bonds table

Multiply forces on atoms I, J by mask terms

Add bond energy to total energy if appropriate

Add bond energy to EMCRES(MC) if appropriate

Angles

Loop over I,J,K angle triples

Calculate angle interaction

Look up potential function masking in angle table

Multiply forces on I,J,K by mask terms

Add angle energy to total energy if appropriate

Add angle energy to EMCRES(MC) if appropriate

Dihedrals

Loop over I,J,K,L dihedral quadruplets

Calculate dihedral interaction

Look up potential function masking in dihedral table

Multiply forces on I,J,K,L by mask terms

Add dihedral energy to total energy if appropriate

Add dihedral energy to EMCRES(MC) if appropriate

Loop over I,L 1-4 interaction pairs

Calculate 1-4 vdW and electrostatic interaction

Look up potential function masking in dihedral table

Multiply forces on I,J,K,L by mask terms

Add 1-4 energy to total energy if appropriate

Add 1-4 energy to EMCRES(MC) if appropriate

Van der Waals/Electrostatics

Loop over I,J nonbonded pairs

Calculate nonbonded vdW and electrostatic interaction

Look up potential function masking in nonbond table

Multiply forces on I,J by mask terms

Add nonbonded energy to total energy if appropriate

Add nonbonded energy to EMCRES(MC) if appropriate

NOTE: Force masking is also applied to the molecular virial used to calculate the pressure – the system pressure is calculated considering only the "real" (imccur) MC residue.

R(T), F(R,T), T

if T is an integer multiple of IMCNS, enter Monte Carlo routine

calculate "Boltzmann probabilities" for each residue:

calculate exp_value(I) = exp(-EMCRES(I)/kT)

sum up exp_values

PMCRES(I) = PMCRES(I) + exp_value(I)/sum

calculate "corrected Boltzmann probabilities" for each residue

calculate exp_value(I) = exp(-(EMCRES(I)-EMCREO(I))/kT)

sum up exp_values

DMCRES(I) = DMCRES(I) + exp_value(I)/sum

add 1 to total # of Boltzmann probability accumulations

select trial move – generate random number in the range {0,1}

loop over residues

IF random number < DMCWIN(I)
test MC acceptance
ELSE
check next residue
update IMCCUR based on accept/reject of trial move
add 1 to Monte Carlo history for the IMCCUR residue
exit main Monte Carlo routine
IF **T** < **Tmax**, dynamics continues
given the position and force on each atom at time **T**, carry out a leapfrog integration step:
acceleration (**T**) = force(**T**)/mass
velocity (**T** + **dT**/2) = velocity (**T** – **dT**/2) + acceleration (**T**) * **dT**
position (**T** + **dT**) = position (**T**) + velocity (**T** + **dT** /2) * **dT**
T = **T** + **dT**
R(T+dT), T+dT Return to beginning of main dynamics loop (*BEGIN*)
ELSE (**T** = **Tmax**)
Number of runs = number of runs + 1
Write out molecular dynamics restart information
Write out Monte Carlo restart information (MCRST)
IF IMCSOL > 1 do adaptive CMC/MD
Calculate relative free energy (DGRES(I))of each MC residue (vs. IMCFRS),
based on IMCCNT (IMCSOL = 2) or DMCRES (IMCSOL = 3).
If IMCSOL = 2, correct this free energy for the current EMCREO offsets
Set new EMCREO = -DGRES(I)
Write new EMCREO to MDOUT, MCFIL
Re-initialize IMCCNT(I)'s to zero
Re-initialize PMCWIN(I)'s to zero
EMCREO(I)'s and IMCCUR are the only elements of the MC history
carried to the next run.
ELSE
Write out Monte Carlo acceptance probabilities (MCWTRA)
IF number of runs < max
Return to beginning of main dynamics loop (*BEGIN*)
IF number of runs = max
End simulation (*END*)

END End of main dynamics loop

Lookup tables for potential function masking:

Nonbonded (van der Waals and electrostatics) lookup (subroutine MCMSK)

Atom I	Atom J	Force I	Force J	Etot	Emcres if IMCINT = 0	Emcres if IMCINT = 1	Notes
0	0	1.0	1.0	1.0	---	---	2 non-MC atoms always see each other
IMCCUR	IMCCUR	1.0	1.0	1.0	---	IMCCUR	The IMCCUR residue is part of the "real" system & adds to total energy and virial.
MC-1	MC-1	1.0	1.0	0.0	---	MC	2 MC atoms in the same residue always see each other
MC-1	MC-2	0.0	0.0	0.0	---	---	2 MC atoms in different residues never see each other
MC-2	MC-1	0.0	0.0	0.0	---	---	(symmetric)
IMCCUR	0	1.0	1.0	1.0	IMCCUR	IMCCUR	The IMCCUR residue is part of the "real" system & adds to total energy and virial.
0	IMCCUR	1.0	1.0	1.0	IMCCUR	IMCCUR	(symmetric)
MC-1	0	EMCNF	0.0	0.0	MC	MC	MC "ghost" feels surroundings if EMCNF > 0
0	MC-1	0.0	EMCNF	0.0	MC	MC	(symmetric)

Bond lookup table (subroutine MCMSK)

Atom I	Atom J	Force I	Force J	Etot	Emcres if IMCINT = 0	Emcres if IMCINT = 1	Notes
0	0	1.0	1.0	1.0	---	---	2 non-MC atoms always see each other
IMCCUR	IMCCUR	1.0	1.0	1.0	---	IMCCUR	The IMCCUR residue is part of the "real" system & adds to total energy and virial.
MC-1	MC-1	1.0	1.0	0.0	---	MC	2 MC atoms in the same residue always see each other
MC-1	MC-2	0.0	0.0	0.0	---	---	2 MC atoms in different residues never see each other
MC-2	MC-1	0.0	0.0	0.0	---	---	(symmetric)
IMCCUR	0	1.0	1.0	1.0	---	IMCCUR	The IMCCUR residue is part of the "real" system & adds to total energy and virial.
0	IMCCUR	1.0	1.0	1.0	---	IMCCUR	(symmetric)
MC-1	0	EMCIF	0.0	0.0	---	MC	MC "ghost" feels surroundings if EMCNF > 0
0	MC-1	0.0	EMCIF	0.0	---	MC	(symmetric)

Note that 2 different MC residues should never be covalently bonded to one another (rows 4 & 5, above). Also, bonds never add to the MC interaction energies (EMCRES) unless intra-perturbed group terms are included in the calculation (IMCINT = 1).

Angle lookup table (subroutine MCAMSK)

Atom I	Atom J	Atom K	Force I	Force J	Force K	Etot	Emcres if IMCINT =		Notes
							0	1	
0	0	0	1.0	1.0	1.0	1.0	---	---	non-MC atoms always see each other
IMCCUR	IMCCUR	IMCCUR	1.0	1.0	1.0	1.0	---	IMCCUR	The IMCCUR residue is part of the "real" system & adds to total energy and virial.
MC-1	MC-1	MC-1	1.0	1.0	1.0	0.0	---	MC	MC atoms in the same residue always see each other
MC-1	MC-2	MC-3	0.0	0.0	0.0	0.0	---	---	Any mix of MC atoms in different residues never see each other
MC-1	MC-2	0	0.0	0.0	0.0	0.0	---	---	Any mix of MC atoms in different residues and "real" atoms feel no angle interactions
IMCCUR 0 0 IMCCUR IMCCUR 0	0 IMCCUR 0 IMCCUR 0 IMCCUR	0 0 IMCCUR 0 IMCCUR IMCCUR	1.0	1.0	1.0	1.0	---	IMCCUR	Any mix of IMCCUR atoms and real atoms acts as part of the "real" system & adds to the total energy. . .
MC	0	0	EMCIF	0.0	0.0	0.0	---	MC	MC "ghosts" feel surroundings if EMCIF > 0; real atoms never feel forces due to angles involving "ghost" atoms
0	MC	0	0.0	EMCIF	0.0	0.0	---	MC	
0	0	MC	0.0	0.0	EMCIF	0.0	---	MC	
MC	MC	0	EMCIF	EMCIF	0.0	0.0	---	MC	
MC	0	MC	EMCIF	0.0	EMCIF	0.0	---	MC	
0	MC	MC	0.0	EMCIF	EMCIF	0.0	---	MC	

Dihedral lookup table (subroutine MCPMSK)

Atom I	Atom J	Atom K	Atom L	Force I	Force J	Force K	Force L	Etot	Emcres if IMCINT =		Notes
									0	1	
0	0	0	0	1.0	1.0	1.0	1.0	1.0	---	---	non-MC atoms always see each other
IMCCUR	IMCCUR	IMCCUR	IMCCUR	1.0	1.0	1.0	1.0	1.0	---	IMCCUR	The IMCCUR residue is part of the "real" system & adds to total energy and virial.
MC-1	MC-1	MC-1	MC-1	1.0	1.0	1.0	1.0	0.0	---	MC	MC atoms in the same residue always see each other
MC-1	MC-2	MC-3	MC-4	0.0	0.0	0.0	0.0	0.0	---	---	Any mix of MC atoms in different residues never see each other.
MC-1	MC-2	MC-3	0	0.0	0.0	0.0	0.0	0.0	---	---	Any mix of MC atoms in different residues and real atoms never interact.
IMCCUR 0 0 0 IMCCUR IMCCUR IMCCUR 0	0 IMCCUR 0 0 IMCCUR IMCCUR 0 IMCCUR	0 0 IMCCUR 0 IMCCUR 0 IMCCUR IMCCUR	0 0 0 IMCCUR 0 IMCCUR IMCCUR IMCCUR	1.0	1.0	1.0	1.0	1.0	---	IMCCUR	Any mix of IMCCUR atoms and real atoms acts as part of the "real" system & adds to the total energy. . .not all are enumerated
MC	0	0	0	EMCIF	0.0	0.0	0.0	0.0	---	MC	MC "ghosts" feel surroundings if EMCIF > 0; real atoms never feel the associated force.
0	MC	0	0	0.0	EMCIF	0.0	0.0	0.0	---	MC	
0	0	MC	0	0.0	0.0	EMCIF	0.0	0.0	---	MC	
MC	MC	0	MC	EMCIF	EMCIF	0.0	EMCIF	0.0	---	MC	
MC	0	MC	MC	EMCIF	0.0	EMCIF	EMCIF	0.0	---	MC	
0	MC	MC	MC	0.0	EMCIF	EMCIF	EMCIF	0.0	---	MC	

Not all atom combinations are enumerated in the table, for readability. General rules are listed in the Notes section of the table.

1-4 interaction (van der Waals and electrostatics) lookup (subroutine MCMSK)

Atom I	Atom L	Force I	Force L	Etot	Emcres if IMCINT = 0	Emcres if IMCINT = 1	Notes
0	0	1.0	1.0	1.0	---	---	2 non-MC atoms always see each other
IMCCUR	IMCCUR	1.0	1.0	1.0	---	IMCCUR	The IMCCUR residue is part of the "real" system & adds to total energy and virial.
MC-1	MC-1	1.0	1.0	0.0	---	MC	2 MC atoms in the same residue always see each other
MC-1	MC-2	0.0	0.0	0.0	---	---	2 MC atoms in different residues never see each other
MC-2	MC-1	0.0	0.0	0.0	---	---	(symmetric)
IMCCUR	0	1.0	1.0	1.0	---	IMCCUR	The IMCCUR residue is part of the "real" system & adds to total energy and virial.
0	IMCCUR	1.0	1.0	1.0	---	IMCCUR	(symmetric)
MC	0	EMCIF	0.0	0.0	---	MC	MC "ghost" feels surroundings if EMCNF > 0
0	MC	0.0	EMCIF	0.0	---	MC	(symmetric)

Nonbonded (van der Waals and electrostatics) lookup for mean-field dynamics with N residues
(subroutine MCMSK, IMCDO = -2)

Atom I	Atom J	Force I	Force J	Etot	Notes
0	0	1.0	1.0	1.0	2 non-MC atoms always see each other
MC-1	MC-1	1.0	1.0	1/N	2 MC atoms in the same residue always see each other
MC-1	MC-2	0.0	0.0	0.0	2 MC atoms in different residues never see each other
MC-2	MC-1	0.0	0.0	0.0	(symmetric)
MC	0	EMCNF	EMCSF	1/N	MC "ghost" feels surroundings if EMCNF > 0
0	MC	EMCSF	EMCNF	1/N	(symmetric)

Note that for mean-field dynamics there is no single "real" MC residue – IMCCUR is meaningless. Instead, all MC residues add to the total system energy, but this quantity is not formally well-defined. EMCNF is usually set to 1.0 – each MC atom feels the full force of surrounding atoms. The MC energy accumulators (EMCRES[n]) are similarly meaningless. EMCSF is usually set to 1/N – each atom of the surroundings feels 1/N of the normal interaction with any MC atom.

Appendix 2: Description of CMC/MD inputs and specifications, including topology files and coordinate issues

Since it was built as a modification of the SANDER molecular dynamics package, the input files for CMC/MD are generally similar to those described for SANDER in the AMBER 5.0 manual.

PRMTOP ("topology file"): The CMC/MD topology should contain one of each Monte Carlo residue of interest (guests; sidechains) in addition to the receptor or solvent. For a host:guest, solvation, or protein:ligand calculation, all of the guests should be treated as one "molecule" in order to avoid pressure artifacts. Though they are all part of the same "molecule", the ligands should still not be covalently connected. A somewhat different topology is required for CMC/MD on peptides or protein side chains. This topology is described in detail in Chapter 5, where we present our calculations on dipeptides and T4 lysozyme. All the MC residues are part of the single polymer molecule, which includes non-MC and MC regions. The first MC residue (N) is covalently connected to the preceding residue (N-1) as usual. The remaining MC residues are specified in order but not covalently connected to one another. The final MC residue (M) is connected to the subsequent normal residue (M+1) as usual. Covalent crosslinks are then specified between each MC residue and the N-1 and M+1 normal residues. This way, each MC residue is connected to the normal residues that flank the MC region in exactly the same way.

INPCRD ("input coordinates"): Input coordinates for the CMC/MD run should again contain data for each CMC/MD residue, in addition to the remainder of the system. Typically, the CMC/MD residues will all be roughly superimposed in the binding pocket. For the HIV-RT:TIBO work, each ligand was minimized in the presence of the receptor, and these ligand coordinates were combined with a single copy of HIV-RT to make the CMC/MD input coordinates. One caveat; since the nonbonded pairlist subroutine was not modified, the CMC/MD residues cannot overlap exactly. If this occurs, a divide-by-zero error will result when the interatomic distances are calculated between precisely overlapped atoms. Note that this is only an issue with the initial PDB input (coordinates specified to 10^-3 Angstrom). Due to the additional precision of the internal and external (restart, coordinate file, etc.) AMBER coordinate specifications, this is not an issue once dynamics has started.

MDIN: Several additional control variables were added to the main SANDER namelist. In addition, CMC/MD requires a short stretch of block formatted input for input that varies from one MC residue to the next (like solvation offsets, etc.). Each control variable is detailed below, along with the block input:

IMCDO (main CMC/MD control variable)

= -2 "Mean-field" dynamics. CMC/MD input is read from the MDIN file, but the current MC residue is ignored. Instead, all MC residues are treated as "ghosts" with

fractional interactions with the surroundings. Monte Carlo steps are not carried out, and statistics are not accumulated.

= -1 Default. CMC/MD is turned off, normal SANDER behavior

= 0 Equilibration. CMC/MD input is read from the MDIN file, the current MC residue is set to IMCCUR, and the potential function is masked appropriately. Monte Carlo steps are not carried out, and statistics are not accumulated.

= 1 CMC/MD, ghost forces on. CMC/MD input is read from the MDIN file, etc. Monte Carlo steps are carried out every IMCNS steps. EMCNF is set to 1.0, EMCIF to 1.0, and EMCSF to 0.0.

= 2 CMC/MD, ghost forces off. CMC/MD input is read from the MDIN file, etc. Monte Carlo steps are carried out every IMCNS steps. EMCNF is set to 0.0, EMCIF to 1.0, and EMCSF to 0.0.

= 3 CMC/MD, general case. CMC/MD input is read from the MDIN file, etc. Monte Carlo steps are carried out every IMCNS steps. EMCNF, EMCIF, and EMCSF are taken from the namelist input.

IMCRST (initialize/restart CMC/MD)

= 0 No restart. All CMC/MD arrays, history are initialized to zero. At the end of each run, a MC restart file is written to the file specified by MCRST

= 1 Restart. Prior CMC/MD history, including IMCCUR and solvation offsets, are read from MCRST before the calculation begins. Again, at the end of each run, MCRST is overwritten with the current MC history.

IMCSOL (control variable for solvation offset, adaptive CMC/MD)

= 0 Default. No solvation offsets are read from the MC block data.

= 1 Offsets. Solvation offsets are read from the MC block data and used to bias the MC sampling.

= 2 Adaptive CMC/MD. Free energies for the adaptive offsets are calculated based on the number of MC counts accumulated for each residue. Adaptive offsets are updated at the end of every NSTLIM dynamics steps (1 dynamics run)

= 3 Adaptive CMC/MD. Free energies for the adaptive offsets are calculated based on the "Boltzmann" probabilities of each residue. Adaptive offsets are updated at the end of every NSTLIM dynamics steps (1 dynamics run). IMCSOL = 3 provides better convergence of the free energy offsets for unfavorable (high free energy) states.

IMCCP (control variable for multicopy CMC/MD)

= 0 Default. Each MC residue is treated as an independent entity, with a trial probability of 1/N

= 1 MC Copies. The trial probability windows for each MC residue are read from the block MC input. The value specified in the MC input is the upper bound of the trial move range for that residue, on the range {0,1}. If there are M copies of a residue, each should be assigned a trial move range of (1/N)*(1/M). Note that there can be different numbers of copies for different MC residues.

= 2 MC Sets. The set-memberships and trial probability windows for each MC residue are read from the block MC input. Copies of the same chemical species (i.e.

side chain rotamers) should be assigned to the same set. IMCCP = 2 is required for adaptive CMC/MD with MC copies.

IMCNS (Monte Carlo step frequency)

= 0 Default. No MC steps are carried out.

= n MC steps are carried out every N dynamics steps. IMCNS should never be less than NSNB, the nonbonded pairlist update.

IMCINT (control variable for inclusion of intramolecular energies)

= 0 Default. No intramolecular terms contribute to the CMC/MD energies. By intramolecular terms, we mean all terms (bond, angle, dihedral, 1-4, nonbond and electrostatic) between atoms in the same MC residue or between a MC atom and atoms of the surroundings. If IMCINT = 0, only nonbond and electrostatic energies between MC atoms and the surroundings are included in the calculated MC free energy.

= 1 All inter- and intra-molecular terms contribute to the free energy differences calculated by CMC/MD. IMCINT = 1 is necessary for the calculation of free energy differences that involve significant contributions from strain or intra-perturbed group interactions.

EMCIF (floating point multiplier for intramolecular CMC/MD forces felt by MC residues)

= 1.0 Default. EMCIF should always be set to 1.0, and is included only for debugging purposes. This preserves bond, angle, and dihedral interactions and geometries within MC residues and between MC residues and atoms of the surroundings they are covalently bonded to.

EMCNF (floating point multiplier for nonbonded CMC/MD forces felt by MC residues)

= 0.0 Default. Only the IMCCUR MC residue feels forces exerted by the surroundings. Ghost MC residues behave as though they were in the gas phase.

= 1.0 Full ghost forces. All ("real" and "ghost") MC residues feel forces exerted by the surrounding non-MC residues. This will often yield unacceptably large forces and velocities (with correspondingly unstable dynamics) if there is significant van der Waals overlap between "ghost" MC residues and the surroundings. Smaller values of EMCNF (0.1, 0.01) may mitigate this problem somewhat.

EMCSF (floating point multiplier for nonbonded CMC/MD forces felt by non-MC residues)

= 0.0 Default. The surroundings never feel forces corresponding to the "ghost" MC residues.

> 0.0 Mean-field dynamics. For mean-field dynamics, this value is typically set to 1/N, where N is the number of mean-field MC residues.

EMCOFE (floating point multiplier for 1-4 electrostatic energy included in MC energy)

= 1.0 Default.

= 0.0 1-4 electrostatic interactions are excluded from the MC energy. This option was originally used as a debugging aid.

EMCOFN (floating point multiplier for 1-4 nonbonded energy included in MC energy)

= 1.0 Default

= 0.0 1-4 nonbonded interactions are excluded from the MC energy. Similar to EMCOFE, above.

IMCFIL (frequency of updates to Monte Carlo history file MCFIL)

= 99999 Default. Effectively no updates to MC history file MCFIL.

= n MCFIL is updated every n dynamics steps. MCFIL provides a useful history of the CMC/MD calculation that can be used for convergence graphs, etc.

MONTE CARLO VARIABLES (block format)

Line 1: (3I5) IMCFRS, IMCLST, IMCCUR

IMCFRS (residue number of first MC residue)

IMCLST (residue number of last MC residue)

Self-explanatory. The MC residues are required to be a contiguous stretch of residues in the topology file.

IMCCUR (residue number of current MC residue)

IMCCUR denotes the current (or "real") MC residue for the purposes of potential function masking and trial move acceptance/rejection. It is read from the block MC input in MDIN, but can be superseded by data from MCRST if this is a restarted MC calculation (IMCRST = 1).

Line 2: (I5) IMCNS

IMCNS (Monte Carlo step frequency)

Monte Carlo moves will be attempted every IMCNS dynamics steps.

Lines 3 to 3 + (IMCLST-IMCFRS+1): per-residue input

I5,F8.3,F14.11,I5: I, EMCREO(I), PMCWIN(I), IMCSET(I)

I: number of the MC residue (IMCFRS = 1, etc.). Not used, placeholder for readability.

EMCREO(I): Umbrella sampling offset for residue I.

Could be solvation offset or adaptive offset.

PMCWIN(I): Upper bound of MC trial move range for residue I. The lower bound of the trial move range is specified by considering the upper bound of the N-1 residue, or zero in the case of the first MC residue. Only read if IMCCP > 0.

IMCSET(I): Chemical species set to which residue I belongs. All copies of a given species should be assigned to the same set. Different species should be assigned to different sets. Only read if IMCCP > 1, and only necessary for adaptive CMC/MD with residue copies.

MCRST: Typically, the user will not have to prepare their own MCRST file. The first CMC/MD run should be started with IMCRST = 0, and no MCRST file will be read. This run will generate a MCRST file that summarizes the Monte Carlo history during the calculation. Subsequent (restarted) CMC/MD runs should use IMCRST = 1 and specify

the correct MCRST input file. To simplify input and output, the MCRST file is read at the beginning of the run, but that same filename is used for the output of the next MCRST file. Care should be taken that important MCRST files are not accidentally overwritten. You can also directly use the restart file from an adaptive CMC/MD run in a non-adaptive calculation. The MCRST file is a FORTRAN formatted file, with the following specifications:

Line 1: (3I5) IMCFRS, IMCLST, IMCCUR

IMCFRS and IMCLST are compared with the values from MDIN to ensure that this restart file is correct for the current MC topology. IMCCUR, the current MC residue, overrides the value read from MDIN.

Line 2: (2I10) IMCTOT, IMCTOB

IMCTOT is the total number of MC steps tried so far. IMCTOB is the same, and a legacy from prior code.

Lines 3 to 3 + (IMCLST-IMCFRS + 1):

(I5,I10,2F14.4) I, IMCCNT(I), DMCRES(I), EMCREO(I)

I: MC residue number; again IMCFRS = 1.

IMCCNT(I): Number of times I was the "real" (IMCCUR) residue after a MC step. In other words, the accumulated population of species I.

DMCRES(I): Accumulated "Boltzmann" probability of species I.

EMCREO(I): Umbrella sampling offset for residue I.

Overrides the value read from MDIN.

OUTPUTS

When a CMC/MD run is started, the input parameters are written out in the standard output (MDOUT). In addition, MC data are reported periodically as the trajectory progresses. Every time the MDINFO file is updated, the current MC data are reported, including the current MC residue, the accumulated number of counts per residue, and the accumulated "Boltzmann" probability data. Along with the MDINFO data, the CMC/MD statistics file (MCFIL) is updated every IMCFIL steps. Two kinds of data are written to this file. For non-adaptive calculations, each time the file is updated one line is added for each MC residue. This line specifies the current timestep, MC residue number, number of MC counts, current MC energy (EMCRES), and the difference between this energy and EMCRES for the first MC residue. These latter two are reported for debugging, analysis, or post-processing purposes. The format used is

(i8,i5,i10,2F14.4) NSTEP,I,IMCCNT(I),EMCRES(I),EMCRES(I) - EMCRES(1)

and one line is printed for each MC residue.

In an adaptive calculation, the biasing potentials are also printed to this file each time they are adapted.
The following format is used:

("OFF:",I7,10F14.10) IMCTOT, EMCREO[I, I;1->n]

where IMCTOT is the total number of MC steps thus far and EMCREO() are the biasing offsets of each residue (or set if IMCCP = 2). The MCFIL output file can thus be searched for the "OFF:" string to produce a history of the adaptive offsets.

Finally, the MCRST file described above is updated at the end of each MD "run", when the usual MD coordinate and velocity restart file is being written.

SAMPLE FILES

Sample link.in input file for CMC/MD in the host-guest system from Chapter 3, showing the MC residues as a single molecule, separate from the receptor(s):

```
rebek host dimer + 9 solutes + CHCl3 solvent

HOS    0host.res
ME     0ME.res
ENE    0ENE.res
FMT    0FMT.res
DFM    0DFM.res
TFM    0TFM.res
CF4    0CF4.res
MCL    0MCL.res
DCM    0DCM.res
CLF    0CLF.res
CL3    0chcl3.res

DU
  1  0  0  0  0
all nine guests
O  0  0  1  3  1
ME 2*** ENE *** FMT *** DFM *** TFM *** CF4 ***
MCL *** DCM *** CLF

host monomer 1
O  0  0  1  3  1
HOS 2

host monomer 2
O  0  0  1  3  1
HOS 2

QUIT
```

Sample link.in file for CMC/MD on a peptide. Note the additional bonds that need to be specified between the CMC/MD residues and the surrounding polymer.

```
MCMD dipeptide with PARM94; ala, val, ser
--placeholder line  (formerly dbase name)

DU
  0  1  1  1  0
 MCMD dipeptide
O  1   0   1   3   0
ACE 2ALA *** VAL *** VAL *** VAL ***
SER *** SER *** SER NME

  2  9C  N    0
  3  9C  N    0
  4  9C  N    0
  5  9C  N    0
  6  9C  N    0
  7  9C  N    0
  1  3C  N    0
  1  4C  N    0
  1  5C  N    0
  1  6C  N    0
  1  7C  N    0
  1  8C  N    0

QUIT
```

Sample MDIN file for CMC/MD in a simple (2-solute) system

```
298K NPT  SHAKE
 &cntrl
 IREST =   0, IMIN =   0,  NRUN  =  1,
 NTX   =   1, NSTLIM=  50000,
 NTB   =   2, NTP  =   1,
 TEMPI = 0.0, TEMP0 =300.0,
 DT    =  0.0010, NTT = 1,
 NTC   =   1, NTF  =   1,
 NSNB  =   1, CUT  =   9.0, SCEE = 1.2, IDIEL = 1,
 NTPR  =  5000, NTWX =  50000,
 IBELLY = 0,
 IMCDO = 1,
 &end
  1   2   2   1
```

Sample adaptive CMC/MD input (MDIN) for the HIV/RT-TIBO system – adapting on single residues only.

```
 HIV RT complex with water and 10 TIBO derivates.  bonds SHAKEn.
 mcmd run w/ solvation offset, imcfrc = 0
 &cntrl imin = 0, ibelly = 1,
  nrun = 15, nstlim = 2000, dt = 0.0015, nsnb = 20 ,
 temp0 =300.0, ntt = 1, tautp = .2, tauts = .2,
  ntc = 3, ntf = 3, ntwx = 5000, irest = 0, ntx = 1, idiel=1,
  cut = 9.0, scnb = 2.0, scee = 1.2, ivcap=1, matcap=7801,
  ntpr=500, init =3, tol=0.05, irest=0,ntr=1,
  imcdo = 1, imcsol = 3, imcfil = 20, trand = 500, ig = 71277,
  vlimit = 20.0, imcrst=0
 &end
constrained resiudues
2.0
RES 365 374
END
END
flexible residues in belly
RES 37 38  40 58  87 87  90 91  94 94
RES 97 98  101 119  124 125  127 128
RES 131 131  135 135 147 147  149 149  150 167
RES 187 187  190 190  217 223  242 242
RES 248 249  270 272  341 347
END
tibo derivatives
RES 365 374
END
counter ion
RES 375 375
END
flexible water
RES 795 1301
END
END
  365  374  369    1
   20
    1   0.00
    2   0.00
    3   0.00
    4   0.00
    5   0.00
    6   0.00
    7   0.00
    8   0.00
    9   0.00
   10   0.00
```

Sample MDIN for adaptive CMC/MD with multiple copies (from the ALA-VAL-SER case above):

```
# mc/md using sander
 &cntrl
  init = 4, irest=1, ntx=7,
  nrun=50, nstlim=50000,
  nsnb=10, tempi=300.0, temp0=300.0,
  ntt=-5000, trand = 0, tautp = 0.2, tauts = 0.2,
  dtemp = 20.0,
  ig = 71277, ictcor = 0, vlimit = 20.0,
  ntc=2, ntf=2, tol = 0.00001, dt = 0.001,
  idiel=1, cut=8.0, scee = 1.2,
  ntb=2, ntp = 1, npscal = 1,
  nmropt=1, iftres=1,
  ntpr=2500, ntwx = 10000,
  ibelly = 0, ntr = 0,
  imcdo = 3, imcsol = 1, imcfil = 1000, imccp = 2,
  imcrst = 0, EMCIF = 1.0, EMCNF = 0.0, IMCINT = 1 ,
  emcofe = 1.0d0, emcofn = 1.0d0,
 &end
 &wt
     type = 'REST', value1 = 1.0,
 &end
 &wt
     type = 'END',
 &end
DISANG = rmc2.rest
   2   8   4   3
  100
   1   0.000  0.333333333   1
   2  16.000  0.444444444   2
   3  16.000  0.555555555   2
   4  16.000  0.666666666   2
   5   8.000  0.777777777   3
   6   8.000  0.888888888   3
   7   8.000  0.999999999   3
```

Sample non-adaptive CMC/MD input for a calculation with multiple side-chain copies (T4 lysozyme):

```
equil/test of t4 topology
 &cntrl
  TIMLIM = 360000,
  IREST = 0,
  NTX = 1, INIT = 3,
  NRUN = 9, NSTLIM= 25000, DT = .001,
  NTB = 0, NTP = 0, PRES0 = 1.0, NPSCAL = 1,
  NTT =  -2500,
  TAUTP = 0.2,  TAUTS = 0.2, TEMPI = 298.0,
  TEMP0 =  298.0, DTEMP = 10.0, TRAND = 0,
  NTC = 2, TOL = 0.000005,
  NTF = 2, VLIMIT = 20.0,
  CUT = 8.0, IFTRES = 1,
  IDIEL = 1, SCNB = 2.0, SCEE = 3.3,
  NSNB = 20, NTNB = 1,
  NDFMIN = 0, NTCM = 0, NSCM = -1,
  IBELLY = 0, NMROPT = 1, NTR = 1,
  IVCAP = 0, FCAP = 1.5,
  NTPR = 1000, NTWX = 12500, NTXO = 1,
  imcdo = 3, imcsol = 1, imcfil = 1000, imccp = 2,
  imcrst = 0, EMCIF = 1.0, EMCNF = 0.0, IMCINT = 0 ,
  emcofe = 1.0d0, emcofn = 1.0d0,
 &end
 &wt
     type = 'REST', value1 = 1.0,
 &end
 &wt
     type = 'END',
 &end
DISANG = ../setup/cal_chi1_chi2.rest
veenstra cavity-constraint groups
10.0
FIND
CA * * *
N * * *
C * * *
SEARCH
RES 1 97
RES 100 101
RES 103 105
RES 107 110
RES 112 113
RES 115 115
RES 122 125
RES 188 193
RES 196 196
RES 203 212
END
END
  133  181  133   3
  100
    1   0.000  0.125000000    1
```

2	-7.200	0.166666667	2
3	-7.200	0.208333333	2
4	-7.200	0.250000000	2
5	-3.300	0.291666667	3
6	-3.300	0.333333333	3
7	-3.300	0.375000000	3
8	-4.400	0.388888889	4
9	-4.400	0.402777778	4
10	-4.400	0.416666667	4
11	-4.400	0.430555556	4
12	-4.400	0.444444444	4
13	-4.400	0.458333333	4
14	-4.400	0.472222222	4
15	-4.400	0.486111111	4
16	-4.400	0.500000000	4
17	-1.200	0.513888889	5
18	-1.200	0.527777778	5
19	-1.200	0.541666667	5
20	-1.200	0.555555556	5
21	-1.200	0.569444444	5
22	-1.200	0.583333333	5
23	-1.200	0.597222222	5
24	-1.200	0.611111111	5
25	-1.200	0.625000000	5
26	-7.000	0.638888889	6
27	-7.000	0.652777778	6
28	-7.000	0.666666667	6
29	-7.000	0.680555556	6
30	-7.000	0.694444444	6
31	-7.000	0.708333333	6
32	-7.000	0.722222222	6
33	-7.000	0.736111111	6
34	-7.000	0.750000000	6
35	-5.300	0.763888889	7
36	-5.300	0.777777778	7
37	-5.300	0.791666667	7
38	-5.300	1.205555556	7
39	-5.300	1.219444444	7
40	-5.300	1.233333333	7
41	-5.300	1.247222222	7
42	-5.300	1.261111111	7
43	-5.300	1.275000000	7
44	-11.500	1.295833333	8
45	-11.500	0.916666667	8
46	-11.500	0.937500000	8
47	-11.500	0.958333333	8
48	-11.500	0.979166667	8
49	-11.500	1.000000000	8

Appendix 3: Message-Passing Parallel Pseudocode for the CMC/MD algorithm

CMC/MD was implemented in a Message Passing (MPI) parallel version of SANDER from AMBER 5.0. MPI is a general framework for writing parallel computer programs that are easily transportable across architectures. MPI SANDER uses a relatively straightforward master/slave parallelization scheme. A single "master" node is responsible for I/O and the main flow of control of the program. Parallelization is restricted to computationally intensive parts of the calculation, like the evaluation of forces and energies. The work in these routines is divided among several processors, synchronized by the master node. Information flows between processors by explicit function calls to message passing routines from the MPI library. The MPI implementation of CMC/MD involves only small modifications to the MPI SANDER parallel code, described below.

The master node reads all of the molecular dynamics input, including the CMC/MD input, and broadcasts it to all of the slave nodes. This broadcast includes all variables necessary for masking the potential function The potential energy evaluation is split among nodes, with each node calculating the interactions for a few residues. Each processor accumulates MC interaction energies (EMCMSK(I)) independently. After the energy evaluation is completed by all processors, the master node collects and sums these accumulated values for each MC residue. The master node then carries out the trial move generation, Metropolis Monte Carlo, accumulation and reporting of the Monte Carlo history. Once the MC move is carried out, the new value of IMCCUR is broadcast by the master to all the slave nodes. The master node also handles writing of the MCRST file at the end of each MD run. This is all shown in the following pseudocode; "MASTER" refers to a task done only by the master node; "EVERY" denotes tasks carried out by both master and slave nodes, and "MPI" is used to show interprocessor communications. It may help to refer to the pseudocode in Appendix 1 for comparison.

```
MASTER initiates SANDER calculation
MASTER reads input file (MDIN), topology, input coordinates/velocities,
        MC restart file if necessary (MCRST)
MASTER initializes variables, based on input files
MPI MASTER broadcasts control variables to slave nodes
BEGIN main dynamics loop
MPI barrier to synchronize all nodes
MPI MASTER broadcasts coordinates of every particle to slave nodes
EVERY node initializes its copy of EMCMSK(I) = 0
EVERY node calculates its subset of bonds, angles, dihedrals, 1-4's and nonbonds
        During this evaluation, the potential is masked as necessary and interaction
        energies are accumulated to each node's EMCMSK(I)
MPI barrier to synchronize all nodes when energy evaluation is complete
MPI collection to calculate total system energy, forces on each atom from individual
        nodes' copies
MPI collection to calculate total EMCMSK(I)'s from individual nodes' copies
MPI barrier to synchronize all nodes prior to Monte Carlo step
```

MASTER enters main MC routine, carries out Metropolis MC, etc.
MPI MASTER broadcasts new MC residue (IMCCUR) to slave nodes
MPI barrier to synchronize all nodes prior to integration
EVERY node carries out its fraction of the integration step
MPI MASTER collects coordinates for every particle from individual nodes
MASTER does any necessary output, control variables, etc.
END end of main dynamics loop;
    usual control statements for return to beginning of loop or
    end of simulation, as necessary. . .

Appendix 4: Description of CMC/MD inputs for mean-field and ensemble dynamics.

Sample MDIN file for mean-field dynamics in the host-guest system from Chapter 3. For mean-field dynamics, IMCDO is set to –2 (mean-field dynamics, no MC steps); EMCNF is set to 1.0 ("ghosts" feel full forces from the receptor), and EMCSF is set to 1/N, where N is the number of MC residues (the surroundings feel 1/N of the normal interaction with each MC residue).

```
Mean-field MC/MD of host+9 guests in chloroform, production
 &cntrl
  imin = 0, maxcyc = 0, nsnb = 25,
  init = 4, ntx = 7, IREST = 1,
  ntc = 2, ntf = 2, tol = 0.00001,
  ntb = 2, ntp = 1, comp = 108.6, npscal = 1,
  idiel = 1, cut = 12.0, scee = 1.2,
  nrun = 1, nstlim = 1000, dt = .001,
  ntt = 1, temp0 = 300.0, dtemp = 20.0,
  trand = 1000, ig = 77752,
  ntpr = 50, ntwx = 50,
  ictcor = 1, vlimit = 20.0,
  nmropt = 1,
  imcdo = -2,
  EMCIF = 1.0, EMCNF = 1.0, EMCSF = 0.111,
  IMCINT = 1, EMCOFE = 1.0, EMCOFN = 1.0,
  imcsol = 0, imccp = 0, imcfil = 500, imcrst = 0,
 &end
 &wt
         type = 'REST', value1 = 0.5,
 &end
 &wt
         type = 'END',
 &end
DISANG = open.dist2
   1   9   1   -2
 500
   1  0.000
   2  0.000
   3  0.000
   4  0.000
   5  0.000
   6  0.000
   7  0.000
   8   0.000
   9   0.000
```

Sample MDIN file for representative dynamics in the same host-guest system. For representative dynamics, IMCDO is set to 0 (no Monte Carlo steps), EMCNF to 1.0 (ghosts feel the full force of the surroundings) and EMCSF to 0.0 (the surroundings feel no forces from the ghosts).

```
representative dynamics of host+9 guests in chloroform, production
 &cntrl
  imin = 0, maxcyc = 0, nsnb = 25,
  init = 4, ntx = 7, IREST = 1,
  ntc = 2, ntf = 2, tol = 0.00001,
  ntb = 2, ntp = 1, comp = 108.6, npscal = 1,
  idiel = 1, cut = 12.0, scee = 1.2,
  nrun = 1, nstlim = 1000, dt = .001,
  ntt = 1, temp0 = 300.0, dtemp = 20.0,
  trand = 1000, ig = 77752,
  ntpr = 50, ntwx = 50,
  ictcor = 1, vlimit = 20.0,
  nmropt = 1,
  imcdo = 0,
  EMCIF = 1.0, EMCNF = 1.0, EMCSF = 0.0,
  IMCINT = 1, EMCOFE = 1.0, EMCOFN = 1.0,
  imcsol = 0, imccp = 0, imcfil = 500, imcrst = 0,
 &end
 &wt
         type = 'REST', value1 = 0.5,
 &end
 &wt
         type = 'END',
 &end
DISANG = open.dist2
   1   9   1   0
 500
   1  0.000
   2  0.000
   3  0.000
   4  0.000
   5  0.000
   6  0.000
   7  0.000
   8  0.000
   9  0.000
```

NOTE: guest 1 (specified by IMCCUR) is the only one that exerts forces on the host