Title
Using molecular dynamics for high resolution protein structure prediction

Author
Lee, Matthew Randolph

Publication Date
2001

Peer reviewed|Thesis/dissertation

Using Molecular Dynamics for High Resolution Protein Structure Prediction

by

Matthew Randolph Lee

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

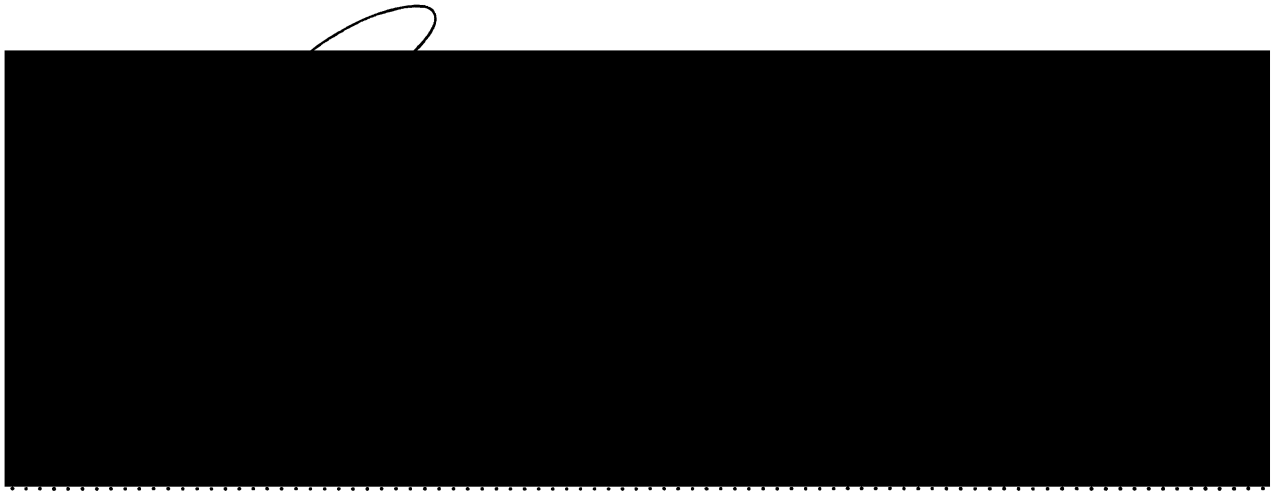
Pharmaceutical Chemistry

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA SAN FRANCISCO



Date University Librarian

Degree Conferred: .....

I dedicate this to my son Jared,

and

to my wife Christiane

and

to the lasting memory of Peter A. Kollman, who left us on May 25, 2001.

Acknowledgements

As I turn yet another page in my book of life, entering the workplace and finally leaving behind seemingly endless years of education, I look back on my graduate school years as one of the most enjoyable periods of my life, and I look forward to an exciting new life, feeling well equipped with the knowledge and experience I gained through my graduate school training. It seemed this day would never come, and now that it has, I would like to thank those influential individuals who helped me reach this point. First and foremost, I thank my Father in heaven for having filled my life with blessings, both at home and in my professional career. I thank my wife, Christiane, for always being there, for helping make life enjoyable on a shoestring budget, for all the routine chores like laundry and cooking, and most importantly, for providing being such a giving, nurturing, compassionate mother to our son Jared. I thank Jared for making me smile, for keeping me modest, and for tolerating my long hours away from home. To Toshio Kitazawa, I give my sincerest thanks for showing me how fun science can be, for mentoring me, for stimulating my quest to obtain a doctorate degree, and for being such a kind, gracious and honorable man. I thank my parents for raising me to be independent, for letting me choose my own paths and suffer my own consequences. I am also grateful to Carlos Simmerling for teaching me so much about the finer details of MD, to Irina Massova for keeping me motivated, to Ken Dill for sharing his always insightful wisdom, and to Jim Caldwell for helping me to laugh and keep things in perspective. Finally, I am deeply gratified to my Ph. D. research advisor, Peter Kollman, for providing such great mentorship, for always being available, for his uncanny ability to decipher tables, for

always being encouraging and in good spirits, and for having contributed so much to the field of computational chemistry.

Several chapters in this dissertation have been previously published:

Chapter 2 is a reprint of an article that appeared in **Proteins**.

Chapter 3 is a reprint of an article that appeared in **J. Am. Chem. Soc.**

Chapter 4 is an article that has been submitted for publication in **Structure**.

Chapter 5 is an article that has been submitted for publication in **J. Mol. Biol.**

Matthew Randolph Lee

ABSTRACT

We now live in the genomics era, where novel sequences abound, awaiting structural determination that will probably only ever be solved experimentally in a small fraction of these new targets, due to the time constraints of experimental methods. Thus, the allure of accurate, insightful protein structure prediction is greater now than it ever has been, but the leading-edge methods fall well short of providing useful predictions, unless there is a very high percentage of sequence identity. Amino acid sequences exhibit an enormously large number of possible conformations, leading to a dimensionality problem that can only be overcome by reducing the representation of the protein. Unfortunately, resolving the difficulty of dimensionality by simplifying the representation also limits the extent of accuracy that can be had. A logical answer to this predicament is to pass on structures obtained from an *ab initio* or comparative modeling protein structure prediction effort, which both are effective at dramatically reducing the number of allowable configurations, into a more accurate method such as molecular mechanics/dynamics, to move from low/medium resolution structure predictions to high resolution ones. This can be accomplished by simply more effective scoring of the large number of predictions that arise from the early stages, and by drawing the best predictions ever more closely to the native state. This thesis has been an exploratory effort, met with significant success, designed to evaluate the promise of using methods within molecular mechanics, molecular dynamics in particular, in the endgame of protein structure prediction for high resolution protein structure prediction.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 2.

Chapter 3.

Chapter 4.

Chapter 5.

LIST OF FIGURES

Chapter 2.

Chapter 3.

Chapter 4.

Chapter 5.

Chapter 1:

Introduction

Successful *ab initio* protein structure prediction remains one of the greatest challenges in computational chemistry, while protein structure prediction in general has become of much greater widespread interest due to the vast number of new sequences identified by genomics initiatives in recent years. A widely held belief, postulated by Anfinsen[1], is that the structure of every protein is governed by thermodynamics, that native structures lie at the global free energy minimum. Although examples are beginning to surface in which this postulate does not hold true, with the first published case being α-lytic protease[2], these are widely regarded as exceptions to the rule. Thus, a computational approach should be capable of both identifying and finding the native structure by monitoring the free energy in most proteins. However, there are three pragmatic reasons why the structure prediction community cannot take advantage of this premise (or discount it). To begin with, the free energy landscape presumably has a very uneven surface, meaning that structures with lower free energies do not necessarily have more native similarity. Second of all, the dimensionality of conformational space, coupled with limited sampling in modern computational methods, makes it difficult to explore many conformations over a practical timescale. The third reason, perhaps the most important of all, is that obtaining an accurate free energy function has remained a challenging unsolved problem.

Compared to NMR and crystallography experiments, methods for structure prediction produce structures with low levels of accuracy, lacking sufficient accuracy in the energy potential. To address the first and second practical reasons mentioned above, the ruggedness of the landscape and the dimensionality problem, *ab initio* approaches use simple, highly empiricized potential energy functions that bias the dominant forces thought to direct protein folding[3] and in some cases, additional database restraints[4,5] to

narrow down the number of acceptable conformations. A popular choice in these approaches is the use of a lattice model, which unrealistically separates the continuum of backbone dihedrals into only a few bins. Another common simplification is the omission of hydrogen atoms and the representation of side chains in a simpler form, such as a single particle. The main alternative to *ab initio* methods, comparative modeling, in its "classical" approach (for review, see Sanchez et al.)[6], creates a core from the structurally conserved region, then builds in the variable loop regions and lastly adds side chains, by inheritance if possible or from rotamer libraries with simple potential functions. The main focus of comparative modeling is to reduce the dimensionality issue by using as much knowledge-based information as possible. Both these approaches are designed with the intent of allowing a computer to rapidly screen out a relatively small number of native-like conformations and neither captures all of the physics involved with structure determination. In order to address the third pragmatic reason, the accuracy issue, these predicted structures can then be superimposed onto an all-atom molecular mechanics model for the endgame of a hierarchical protein structure prediction approach; by being more faithful to the structural detail and applying a more sophisticated potential energy function, the hope is that the structural and energetic description of the system will be improved to allow for high resolution protein structure prediction.

The aforementioned methods for protein structure prediction generate many tens to thousands of structures, which gives rise to three major objectives in the "endgame" or final stages of structure prediction: 1) correct identification of the native state, 2) enhancing the selection process and 3) refinement of best structure predictions, which are still not accurate enough for practical application.

Molecular dynamics (MD), which moves a molecular mechanics structure along the energy landscape according to Newtonian mechanics, can in principle with sufficient sampling be used for the structural refinement of *ab initio* and comparative modeling predictions. Chapter 3 represents the first successful attempt at using state-of-the art explicit solvent, restraint-free molecular dynamics simulations for refinement of the HP-36 villin headpiece and ribosomal S15 proteins, although we do not find similar success using this refinement method in Chapter 5 on any of 12 other proteins. Although the possibility of either our energy function being inadequate to improve structures or Anfinsen's thermodynamic hypothesis being incorrect can not be discounted, an inability to refine structures with standard state-of-the-art molecular dynamics simulations probably stems from the inability to sample sufficiently, as these dynamics simulations are now accurate enough to maintain the native structure with 1 to 2 Å Cα RMSD, well within the expected fluctuation range of native states under physiological conditions.[7] A number of ways of improving the sampling are currently in the literature, and include lowering energy barriers of the potential energy surface through mean-field approaches such as Locally Enhanced Sampling[8,9], filtering out the low energy motions as in Self-guided MD[10], and running the simulations with and implicit solvent, such as the Generalized Born approximation[11,12] or the Finite Difference Poisson Boltzmann method[13].

Being able to consistently refine predicted structures would be an extremely valuable aspect of the endgame itself, but theoretical work should also provide additional physical information that can be used to improve the *ab initio* and comparative modeling methods. For instance, MD simulations on fluids generate thermodynamic properties such as the specific heat, while those on small rigid molecules predict conformational

free energy differences. Similarly, it would be beneficial for protein structure refinement MD runs to not only produce a trajectory of conformations, but to also provide an estimate of the relative free energy for each one. This could serve not only as a more effective discriminator over the feeder methods that use more approximate potential functions, but also be used iteratively to improve the initial stage structure prediction methods.

Over the past decade, a number of studies have approached this difficult evaluation in various biomolecules using a continuum solvent model in which the total free energy is partitioned. This continuum solvent model, when used to post-process molecular dynamics simulations through the Molecular Mechanics-Poisson Boltzmann/Surface Area free energy (MM-PBSA), had been shown to give results that correlate well with experimental data when comparing nucleic acid configurations in solution[14]. In these nucleic acid systems, however, G_{elec} seems to account largely for the free energy differences, with the other contributions being relatively insensitive to configuration. This leads to only a single dominant term that requires "fine tuning". Still, this general type of approach had also been used successfully on other biomolecular systems as well. Eisenberg & McLachlan were able to reproduce, with high accuracy and using a very simple empirical solvation free energy, the experimental solvation free energies of transfer observed in amino acids[15]. Later, a more rigorous continuum solvent treatement, similar to MM-PBSA, was used to discriminate between folded and intentionally "misfolded" protein conformations that are structurally and energetically very different from each other[16,17].

Perhaps the most challenging application of the combined explict/continuum model MM-PBSA is to accurately compare the free energies of different solvated stable

protein conformers. Unlike nucleic acids, the stability in proteins is significantly affected by forces other than the electrostatics, most notably the hydrophobic effect and loss of configurational entropy[18]. Chapter 2 represents the first attempt at using an implicit solvent free energy function for discriminating between the native state and a partially folded, compact conformation. This nice result was further tested on a much more challenging set of small alpha proteins in Chapter 3, where through molecular dynamics simulations, we were able to not only refine structures as mentioned above, but also see a corresponding drop in our continuum solvent free energy. In Chapter 4, we investigated much larger decoy sets as well, and compare the quality of NMR and crystal structures, and provide stimulating results, which suggest that an unfolded state can be used as a reference point to identify the native without prior knowledge of any native information, the first of the three main goals in the endgame of structure prediction. In Chapter 5, we characterize the ability of our high resolution protein structure prediction tools to act as the endgame of a statistically meaningful set of proteins.

Given that genomics initiatives have fueled the more general interest in protein structure prediction, due to the ever growing disparity between known sequences and structures, a set of methods useful for the endgame of protein structure prediction should be 1) accessible to anyone who has an interest, rather than just people who specialize in the area, 2) capable of taking advantage of the rapidly growing computer power, and 3) applicable to large numbers of protein conformations. In Appendix A, I present the scripting programs I wrote as part of my Ph. D. that enabled me to automate those processes that previously required manual human intervention, from setup to analysis, which on a large scale set of proteins otherwise acts as the bottleneck. These scripts are an ideal way of maximizing efficient usage of the types of parallel architecture that are

becoming increasingly popular, namely the single processor Intel or AMD machines that are being networked together, by applying methods for the endgame of protein structure prediction that are useful in identifying the native state, ranking structure predictions, and refining some of the structure predictions further.

REFERENCES:

1. Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science* **181**(96), 223-230.

2. Derman, A. I. & Agard, D. A. (2000). Two energetically disparate folding pathways of alpha-lytic protease shave a single transition state. *Nat. Struct. Biol.* **7**(5), 394-397.

3. Chan, H. S. & Dill, K. A. (1991). Polymer principles in protein structure and stability. *Annu. Rev. Biophys. Biophys. Chem.* **20**(2), 447-490.

4. Skolnick, J., Kolinski, A. & Ortiz, A. R. (1997). MONSSTER: a method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.* **265**(2), 217-241.

5. Simons, K. T., Bonneau, R., Ruczinski, I. & Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins*(SUPP3), 171-176.

6. Sánchez, R., Pieper, U., Melo, F., Eswar, N., Martí-Renom, M. A., Madhusudhan, M. S., Mirkovi*c, N. & Sali, A. (2000). Protein structure modeling for structural genomics. *Nat. Struct. Biol.* **7 Suppl**(12), 986-990.

7. Brooks, C. L., 3rd, Karplus, M. & Pettitt, B. M. (1988). Proteins: a theoretical perspective of dynamics, structure and thermodynamics. *Advan. Chem. Phys.* **71**, 1-259.

8. Roitberg, A. & Elber, R. (1991). Modeling side chains in peptides and proteins: Application of the locally enhanced sampling and the simulated annealing methods to find minimum energy conformations. *J. Chem. Phys.* **95**(12), 9277-9287.

9. Simmerling, C., Lee, M. R., Ortiz, A. R., Kolinski, A., Skolnick, J. & Kollman, P. A. (2000). Combining MONSSTER and LES/PME to predict protein structure from amino acid sequence: Application to the small protein CMTI-1. *J. Am. Chem. Soc.* **122**(35), 8392-8402.

10. Wu, X. W. & Wang, S. M. (1998). Self-guided molecular dynamics simulation for efficient conformational search. *J. Phys. Chem. B* **102**(37), 7238-7250.

11. Dominy, B. N. & Brooks, C. L. (1999). Development of a generalized born model parametrization for proteins and nucleic acids. *J. Phys. Chem. B* **103**(18), 3765-3773.

12. Tsui, V. & Case, D. A. (2000). Molecular dynamics simulations of nucleic acids with a generalized born solvation model. *J. Am. Chem. Soc.* **122**(11), 2489-2498.

13. Luo, R., David, L. & Gilson, M. K. (submitted). Accelerating Finite Difference Poisson Boltzmann Calculations for Static and Dynamic Systems. *J. Comp. Chem.*

14. Srinivasan, J., Cheatham, T. E., Cieplak, P., Kollman, P. A. & Case, D. A. (1998). Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate - DNA helices. *J. Am. Chem. Soc.* **120**(37), 9401-9409.

15. Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature* **319**(6050), 199-203.

16. Novotný, J., Rashin, A. A. & Bruccoleri, R. E. (1988). Criteria that discriminate between native proteins and incorrectly folded models. *Proteins* **4**(1), 19-30.

17. Vorobjev, Y. N., Almagro, J. C. & Hermans, J. (1998). Discrimination between native and intentionally misfolded conformations of proteins: ES/IS, a new method for calculating conformational free energy that uses both dynamics simulations with an explicit solvent and an implicit solvent continuum model. *Proteins* **32**(4), 399-413.

18. Dill, K. A. (1985). Theory for the folding and stability of globular proteins. *Biochemistry* **24**(6), 1501-1509.

Chapter 2

On the Use of MM-PB/SA in Estimating the Free Energies of Proteins: Application to Native, Intermediates and Unfolded Villin Headpiece

Matthew R. Lee

Yong Duan

Peter A. Kollman

Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA 94143-0446, United States of America

Grant Sponsor: National Institutes of Health; Grant number: GM-0717
Grant Sponsor: National Institutes of Health; Grant number: GM-29072

Proteins (2000) **39**:309-316

ABSTRACT:

We investigate the stability of three different ensembles of the 36-mer villin headpiece subdomain, the native, a compact folding intermediate and the random coil. Structures were taken from a one μs molecular dynamics folding simulation and a 100 ns control simulation on the native structure. Our approach for each conformation is to first determine the solute internal energy from the molecular mechanics potential, then to add the change resulting from solvation (ΔG_{solv}). Explicit water was used to run the simulation and a continuum model was used to estimate ΔG_{solv} with the finite difference Poisson-Boltzmann model accounting for the polarization part and a linearly surface area-dependent term for the non-polar part. We leave out the solute vibrational entropy from these values but demonstrate that there is no statistical difference among the native, folding intermediate and random coil ensembles. We find the native ensemble to be ~26 kcal/mol more stable than the folding intermediate and ~39 kcal/mol more stable than the random coil ensemble. With an experimental estimate for the free energy of denaturation equal to 3 kcal/mol, we approximate the non-native degeneracy to lie between 10^{16} and 10^{25}. We also present a possible scheme for the mechanism of folding, first order exponential decay of a putative transition state, with an estimate for the $t_{½}$ of folding of ~1 μs.

INTRODUCTION:

Among the goals of computer simulations of protein systems are to understand the mechanism and kinetics of folding and to predict the correct native structure from the very large number of possibilities.

Because the time scale of protein folding, ranging from tens of microseconds to seconds, makes it currently prohibitive to study the entire mechanisms of folding using all-atom models with explicit solvent, simplified models have been used and have given exciting insights[1,2,3,4]. All atom models have been used to give insights into protein unfolding by raising the temperature to high values[5,6,7,8,9]. Also, advances in computer power have enabled studies on the early stages of the mechanism of protein folding[10,11], using all-atom, explicit solvent models. Progress has also been made in predicting protein structure from sequence[12,13], but there is still much work to be done. A crucial element in reaching the goal of predicting protein structures is the development of a method that can discriminate between the correct native structure and other alternatives. Because native protein structures at physiological temperatures are determined by their free energies, which consist of competing enthalpic and entropic parts, gas-phase energies alone are unlikely to be effective for such a purpose, even at the atomic level[14]. As a result, two general types of approaches have emerged for adding the entropy: knowledge-based and physical. The knowledge-based methods rely on comparison with properties of known proteins[15] taken from the protein structure databases. The physically based methods use functions from molecular mechanical force fields. Recently, we[16,17,18] and Hermans' group[19] proposed two similar physical methods and showed they were effective in comparing different structures of free energies of nucleic acids[16] and

proteins[19]. The challenge remains to try such functions on even more challenging decoy structures that come increasingly closer to the correct native structure. An interesting test case has been afforded us in this regard, as our simulation of the early phase of villin folding found a variety of structures including a metastable intermediate. We also have a control simulation on native villin (minimized average NMR structure[20]) which lasted 100 nanoseconds, approximately 10 times longer than any comparable simulation on a native protein.

We have applied the molecular mechanics-Poisson Boltzmann/surface area (MM-PB/SA) method developed by Srinivasan *et al.*[16] to the folding and native simulations of villin, a net 1.1 microseconds worth of structures. We find, encouragingly, that the native structure is calculated to have a noticeably more favorable free energy, 15-35 kcal/mole lower than all other structures, with the intermediate characterized by the lowest free energy found during the folding trajectory.

METHODS:

As previously reported[11], we ran molecular dynamics with the Cornell *et al.* all-atom force field[21], the TIP3P model for water, periodic boundary conditions and an 8 Å cutoff for all solute/solvent non-bonded interactions (with no cutoff for intra-solute interactions) to sample conformational space in the isothermal-isobaric ensemble. Energy calculations reported in this study were made every 100 ps, totaling 10,000 evaluations for the folding trajectory and 1,000 for the control simulation. We first approximated the free energy of each snapshot as the sum of two terms: the internal energy of the protein (E_{MM}) and a solvation free energy (ΔG_{solv}).

$$G_1 = E_{MM} + \Delta G_{solv} \quad (1)$$

E_{MM} is the sum of an internal strain energy (E_{int}), a van der Waals energy (vdW_{tot}), and an electrostatic energy (EEL_{tot}). E_{int} is the energy associated with vibration of covalent bonds and rotation of valence bond angles and torsional angles. vdW_{tot} and EEL_{tot} are further broken down into short range values, those that are within three covalent bonds (vdW_{1-4} and EEL_{1-4}), and long range values that are four or more covalent bonds apart (vdW_{NB} and EEL_{NB}).

The entropy of a given snapshot (S_{solute}), excluding conformational entropy, can be estimated by calculating the translational, rotational and vibrational partition functions with normal mode analysis on a Newton-Raphson minimization. Because configurational differences stem primarily from the latter, we will refer to this as the "vibrational" entropy. This is the most time-intensive part of the MM-PB/SA method on a per snapshot basis and we performed the vibrational entropy calculation on five conformers each of the native state, the metastable folding intermediate, and the denatured state.

Obtaining the solvation free energy from an implicit description of solvent as a continuum is advantageous because it affords a solvation potential that is a function only of the solute's geometry, as discussed and implemented by Srinivasan *et. al.*[18]:

$$\Delta G_{solv} = \langle \Delta G_{NP} \rangle + \langle \Delta G_{pol} \rangle \approx (\gamma \bullet SASA + b) + \langle \Delta G_{pol} \rangle \quad (2)$$

The non-polar solvation free energy (ΔG_{NP}) includes the (largely entropic) cost of creating a solute-sized cavity in solvent and the free energy of inserting the discharged solute into that cavity. Also referred to as the first solvation shell effects, this term has been found experimentally in hydrocarbons to be linearly related to the solvent accessible surface area (SASA), which is obtained from Sanner's MSMS algorithm[22] (probe radius = 1.400 Å). The γ coefficient is set to 5.42 cal/mol • Å^2 and b is set to 920 cal/mol. The electrostatic solvation free energy (ΔG_{pol}) is the cost of charging the discharged solute in the cavity. We adhered to the same Poisson-Boltzmann protocol as described by Srinivasan *et. al.*[18], which uses DelPhi[23] and most of its standard default parameters, together with PARSE atomic radii[24] and Cornell et al. charges[21], to calculate the electrostatic solvation free energy difference for the system between exterior dielectrics of 80 (solvent) and unity (gas phase) according to the position dependent electrostatic potential. One small difference in this usage of DelPhi is to use larger grid spacing of 0.5 Å, extending 20% beyond the edge of the solute. Additionally, we used fewer finite difference iterations (1000) for each (ΔG_{pol}) calculation, which was still amply sufficient as we found the values in this system to reach 90% convergence at around 50 iterations.

RESULTS & DISCUSSION:

The native structure has the lowest MM-PB/SA free energy estimate.

Figures 1 and 3a show the actual MM-PB/SA free energy data as a function of time from the folding and control simulations. As shown in Table 1, we predict the native villin headpiece conformation to be on average ~25 kcal/mol more stable than the lowest energy state encountered during the 1 μs folding simulation (15 kcal/mol at the

smallest gap). This non-native low energy state is, as previously reported, highly compact with a residence time of 160 ns[11]. In comparison, we predict the native conformation to be on average ~35 kcal/mol more stable than the unfolded state.

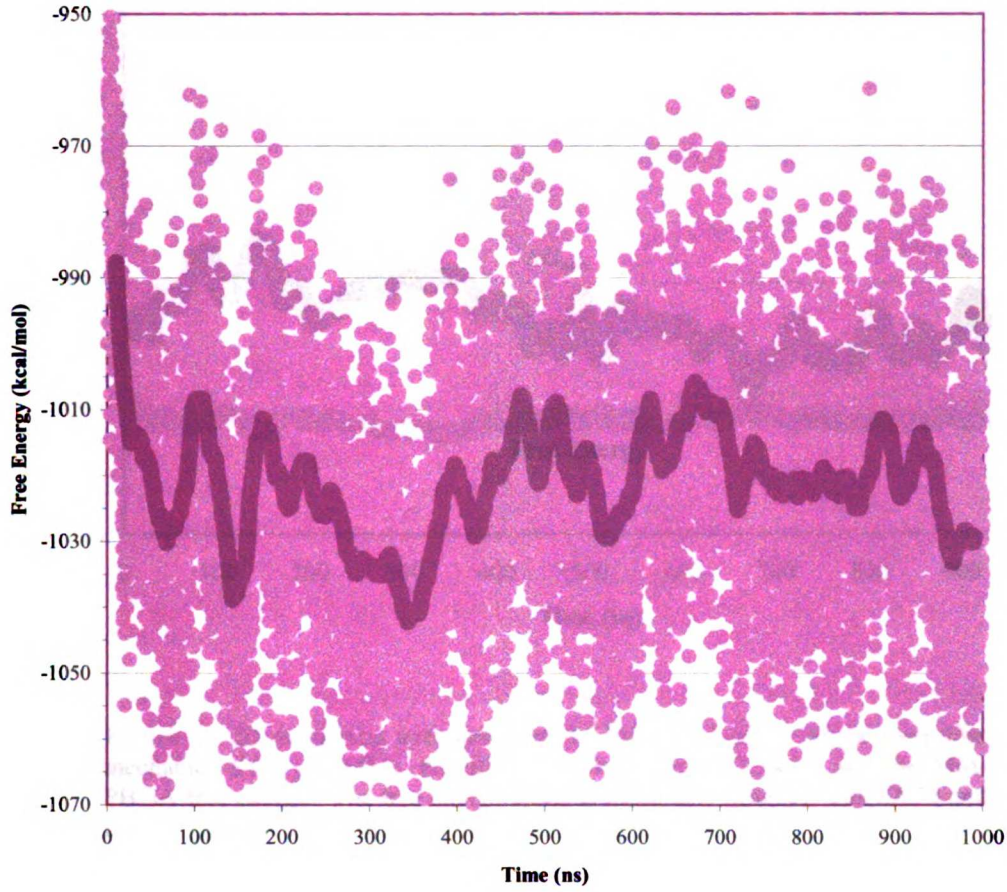


Figure 1. MM-PB/SA free energy from the folding trajectory. The free energy was calculated once every 100 ps, a total of 10,000 times for one μs of data. For each Poisson-Boltzmann calculation, 1000 iterations were used with grid spacing of 0.5 Å, PARSE radii and Cornell, *et al.* charges. The 20 ns running average of 100 ps time steps is shown as the darker solid line. The previously reported folding intermediate ensemble, lies between 240 and 400 ns. We refer the structures from 500 to 1000 ns as the "random coil ensemble" or the "unfolded state".

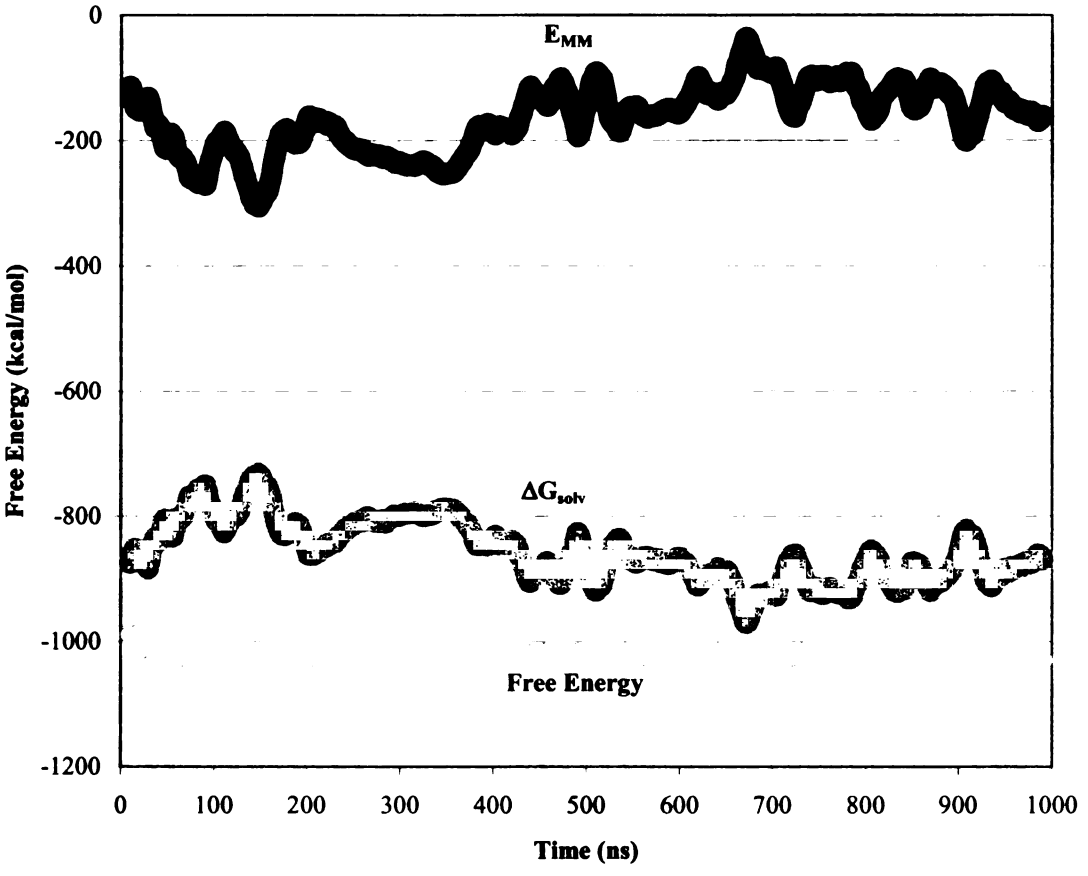


Figure 2. 20 ns running averages from the folding trajectory of the molecular mechanics energy (EMM), the solvation free energy (DGsolv) and the MM-PB/SA free energy. The free energy exhibits much less variation than EMM and DGsolv, and the latter two are strongly inversely related.

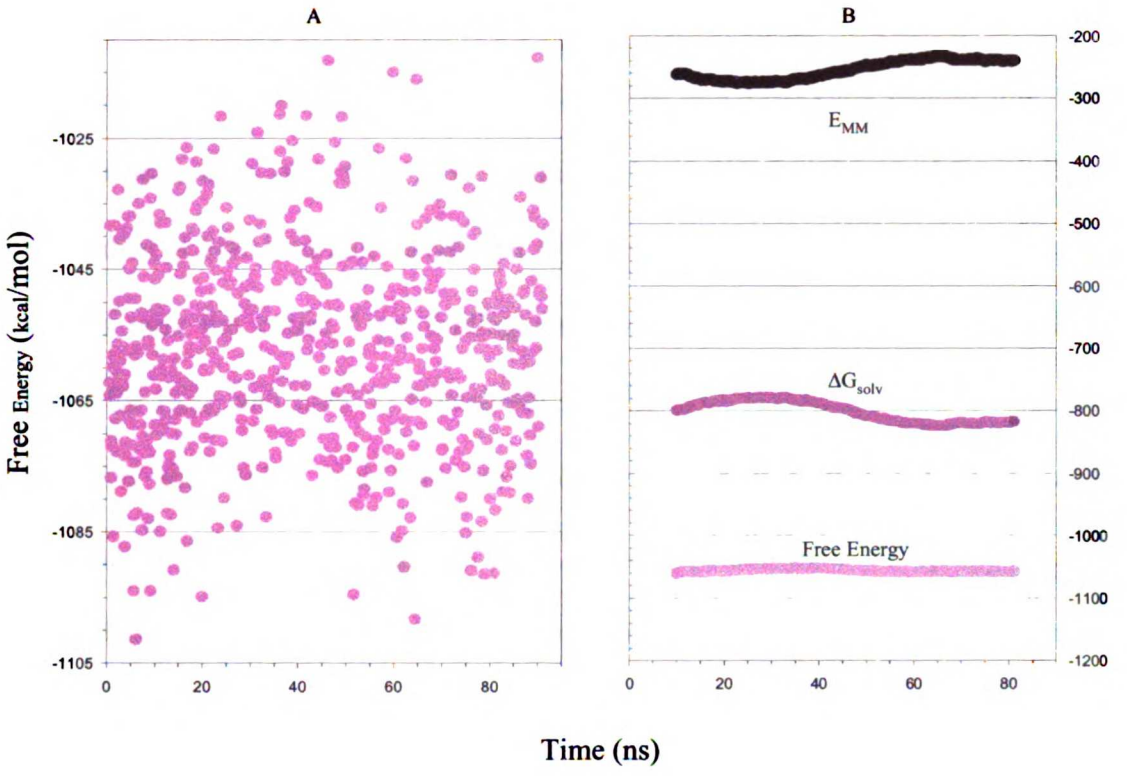


Figure 3. Control simulation results showing (A) actual MM-PB/SA data and (B) 20 running averages. The 20 ns running average of the free energy remains relatively constant at –1058 kcal/mol..

The folding trajectory roughly obeys Boltzmann statistics, according to MM-PB/SA.

In our previous work, we further characterized the folding trajectory by a clustering method, using a limit of 3 Å main-chain RMSD from the cluster's average, and found 30 marginally stable states that were populated with ~500 or more of the 50,000 total coordinate sets[11]. The relationship between the natural log of the cluster population and the MM-PB/SA free energy appears to be a reasonably well behaved Boltzmann distribution, with a correlation coefficient of -0.54. We do not expect a perfect inverse

relationship because kinetic barriers distort the Boltzmann relationship in a non-ergodic trajectory and because the MM-PB/SA free energy is not completely accurate.

Electrostatics are the major source of fluctuation, but not a good predictor of G_1

As can be seen in Figures 2 and 3b, ΔG_{solv} and E_{MM} each exhibit much more fluctuation than their sum, G_1. Over a typical 10 ns period, ΔG_{solv} and E_{MM} will each oscillate over a 300 kcal/mol range and G_1 over a 50 kcal/mol range. The standard deviations over the entire 1 μs folding trajectory are 66.5 kcal/mol for ΔG_{solv}, 73.7 kcal/mol for E_{MM}, and only 17.6 kcal/mol for G_1. The reason for such a disparity in variances is that ΔG_{solv} and E_{MM} are strongly inversely related with a correlation coefficient of -0.97; large changes of ΔG_{solv} are always accompanied by approximately equal and opposite changes in E_{MM}. This inverse relationship can be explained by looking at their dependency on their individual electrostatic components. E_{MM} has a correlation coefficient of 0.95 with its electrostatic term, the intra-protein Coulombic energy ($E_{MM\text{-}eel}$), and ΔG_{solv} one of 1.00 with its electrostatic term, the cost of charging the solute (G_{pol}). The correlation between $E_{MM\text{-}eel}$ and G_{pol} is also strong with a coefficient of -0.97. As intra-solute electrostatic interactions are formed, $E_{MM\text{-}eel}$ and resultantly E_{MM} decrease, while electrostatic interactions between solute and solvent are broken, and resultantly G_{pol} and ΔG_{solv} increase. Thus, the solvation free energy and gas phase energy are inversely related because their preponderant terms are themselves inversely related. The causal factors for fluctuation of ΔG_{solv} and E_{MM} are their electrostatic terms, while the causal factor for fluctuation in G_1 is the total electrostatics for the solvated system (EEL_{tot}), the sum of $E_{MM\text{-}eel}$ and G_{pol}.

Given that electrostatics provide the major source of fluctuation in the solvated protein system, a separate issue remains as to whether or not EEL_{tot} dictates the general trend of G_1. We find EEL_{tot} over the course of the folding trajectory to have a correlation coefficient of only 0.30 with G_1, and the sum of the remaining non-electrostatic terms in the system (non-EEL) one of 0.77 with G_1. In addition, EEL_{tot} in the folding trajectory has a much smaller standard deviation (16.2 kcal/mol) than non-EEL (27.3 kcal/mol). It is not the delicate balance between the sum of strongly opposing terms, G_{pol} and $E_{MM\text{-}eel}$, that relates best to our estimate of the total free energy. Rather, it is the sum of all other terms not associated with electrostatics that drives the shape of the G_1 trajectory. This is not to suggest that forces created by electrostatic interactions are a small contribution to the sum of all forces acting on a protein, that they do not drive the motion of the protein. What the variances in the distributions of EEL_{tot} and non-EEL show are that the sum of electrostatics is much more constant, and what the correlation coefficients with G_1 show are that the sum of non-electrostatics is more responsible for changes in the free energy.

Table 1. Summary of Energies from MD Simulations on the Villin Headpiece

Ensemble	<vdW>	<eel>	<E_{int}>	<E_{MM}>	<ΔG_{pol}>	<ΔG_{NP}>	<ΔG_{solv}>	<G_1>	<$T \cdot S_{gas}$>
Native (1 to 100 ns)	-104.6 (11.1)	-731.9 (31.5)	582.0 (16.1)	-254.5 (35.7)	-821.9 (32.2)	18.7 (0.7)	-803.2 (31.9)	**-1057.7 (15.9)**	455.94 (3.8)
Folding Intermediate (240 to 400 ns)	-105 (14.9)	-720.1 (30.1)	600.6 (17.6)	-224.6 (38.0)	-824.6 (32.1)	17.7 (0.9)	-806.8 (31.6)	**-1031.5 (15.7)**	450.12 (2.0)
Unfolded (500 to 1000 ns)	-65.6 (17.5)	-660.9 (50.9)	598.3 (16.8)	128.2 (61.4)	-912.9 (55.7)	21.7 (1.7)	-891.2 (54.5)	**-1019.2 (17.2)**	455.35 (5.6)

NOTES: 1) values are given in kcal/mol with standard deviations in parenthesis
2) n = 5 for T•Sgas
3) G1 = EMM + DGsolv

Compact structures have better long range van der Waals contacts.

As seen in Table 1, we consider the 100 ns native simulation as a single ensemble and have broken the folding trajectory into two further ensembles, folding intermediate and the unfolded (the last half microsecond of the trajectory). The one energy component that is similar for the native and intermediate states (~ -105 kcal/mol) and significantly more favorable than in the ensemble of unfolded structures (~ -66 kcal/mol) is vdW_{tot}.

The P value from a two-tailed student's t-test between average values from native and folding intermediate is statistically insignificant (> 0.05). In contrast, P values, even after the Bonferroni correction for multiple group comparison, between native and unfolded and between intermediate and the unfolded are highly significant (< 0.0001). However, when looking at EEL_{tot}, the internal strain energy (E_{int}), and the energy of the non-polar first solvation shell effects (ΔG_{NP}), none of the 3 pairwise comparisons of the three ensembles is significantly different for any of the three energies. In addition, we find vdW_{1-4} to show virtually no fluctuation in any of our simulations, that the variance found in vdW_{tot} is essentially identical to that of vdW_{NB}, which implies that it is the long-range van der Waals interactions (4 or more covalent bonds apart) that are more favorable in the native and intermediate states. This is reasonable since these two ensembles are more compact and more favorable van der Waals interactions would be a rational causal factor that they might share in common.

Although the above shows that the two similarly compact native and folding intermediate states have dispersion energies (vdW_{NB}) that are similarly favorable over the less compact unfolded state, this does not imply that all states with native compactness will necessarily have dispersion energies as favorable as the native state. It is possible that the most highly compact structures will not have well packed interiors and therefore higher than native dispersion energies. In this case, the native-like dispersion energies in the folding intermediate were accomplished only at the expense of internal strain energies (see below).

A more statistically meaningful way to associate van der Waals interactions with compactness is to look at the correlation coefficients between the vdW_{NB} and some

parameter that estimates degree of compactness, because this weights the relationship at every snapshot as opposed to comparing three group averages, which can only reveal if the relationship is direct or inverse. The most common measure for degree of compactness is the radius of gyration (R_γ), which calculates the root mean square deviation of each atom in a molecule from the center of mass. The correlation coefficient over the entire folding simulation is 0.82 and that from the control simulation is 0.67, suggesting that compactness and long-range dispersive forces are indeed related, albeit more so in the less compact non-native structures. However, even when looking only at the most compact region of the folding simulation, the folding intermediate, the correlation coefficient is still much higher than the value in the native structures, 0.79. Thus, although the two similarly compact states have similarly favorable dispersion energies, the relationship between R_γ and vdW_{NB} is substantially lower in the native state. Among compact states, there can be a larger distribution of correlation coefficients than among unfolded structures. Perhaps this lesser degree of correlation in the native state can be explained if very well-packed protein sidechains hinder deviations in the compactness from being accompanied by changes in VDW_{NB}. The hydrophobic core in the native state will have very favorable van der Waals contacts, and hence a reasonably constant vdW_{tot} that will likely be less sensitive to the protein's periodic expansion and relaxation than that of the folding intermediate and unfolded ensemble whose sidechains have more freedom to reorient themselves.

Table 2. Summary of Internal Strain Energies from MD Simulations on the Villin Headpiece[1]

Ensemble	Bond	Angle	Dihedral	E_{int}
Native	104.5	274.7	202.8	**582.0**
(1 to 100 ns)	(8.1)	(11.8)	(8.6)	**(16.1)**
Folding Intermediate	106.7	290.7	203.6	**601.1**
(240 to 400 ns)	(8.6)	(12.5)	(10.4)	**(15.8)**
Unfolded	105.9	285.9	206.7	**598.5**
(500 to 1000 ns)	(8.2)	(12.5)	(9.4)	**(15.5)**
(Intermediate) - (Native)	2.2	16.0	0.8	**19.0**
(Unfolded) - (Intermediate)	-0.8	-4.8	3.1	**-2.6**

[1] values are given in kcal/mole, with standard deviations in parenthesis

The native state has less internal strain than other compact structures.

We use the same type of comparisons between native and other compact structures as we did between compact and non-compact structures in the previous section. Referring back to Table 1, one can compare the energy terms between native and folding intermediate ensembles. G_I is ~27 kcal/mol lower and E_{int} ~19 kcal/mol lower in the native ensemble than in the folding intermediate. The difference seen in the averages of E_{int} is highly significant with a Bonferroni-corrected P value < 0.0001. None of the other energy terms (EE_{tot}, vdW, and ΔG_{NP}) are significantly different between the native and folding intermediate states.

By these group comparisons, it appears that E_{int} has the greatest association with G_I in the native and little association in all other ensembles. Again, correlation coefficients provide more information. In the unfolded ensemble (500 ns – 1 μs), the coefficient between E_{int} and G_I was only 0.30, suggesting that they are relatively independent of one another. In the control simulation on the native ensemble, we observe

a correlation coefficient of 0.73 and in the compact folding intermediate ensemble, a coefficient of 0.26. These data suggest that biggest source of disparity between the native and other low R_γ states in this study is E_{int}, the internal strain energy. Table 2 takes a more detailed look at the E_{int} and finds that the major source of difference is the angle term.

Entropic contributions to the MM-PB/SA method.

Table 1 shows that vibrational entropy does not differ by much among the native, compact and unfolded states and that the $T \bullet S_{solute}$ term does not appear to be any more or less favorable in any of the states. The P values for the three pairwise comparisons are all greater than 0.05 and thus not statistically different. This is in agreement with the findings of Hermans when comparing folded and misfolded protein structures, using the harmonic approximation from the covariance matrix of the positional fluctuations during the dynamics trajectory,[19] and of our group when comparing different forms of nucleic acids[16,17,18]. Neither method is particularly accurate, but both show that the $T \bullet S_{solute}$ term is comparable for various similar "structures" of small proteins.

The native and random coil intrinsic "vibrational" entropies are similar but it is the entropy associated with the hydrophobic effect that is represented in the ΔG_{NP} term. As expected, this part of the solvation free energy is least favorable in the unfolded states which also has the highest R_γ. This ΔG_{NP} term makes the unfolded states, from 500 ns to 1 us, about 3 kcal/mol less favorable than the native ($P < 0.0001$). However, ΔG_{NP} in the folding intermediate is about 1 kcal/mol more favorable than in the native ensemble ($P < 0.0001$). As should be expected with a simple linear relation, the fact that the folding

intermediate is (statistically) significantly more favorable than the native (albeit by only 1 kcal/mol), shows that the limitations of this term arise from the uncertainty in the function, not from sample size.

Estimating the conformational entropy of the denatured state.

The free energy of denaturation (ΔG_{denat}) for small proteins has been estimated to fall between 5 and 10 kcal/mol[25,26,27,28]. MacKnight linearly extrapolated a series of guanidine hydrochloride (GuHCl) denaturations of villin to 0 M GuHCl at pH = 5.4 and 4° C and estimates ΔG_{denat} to be 3.3 (±0.4) kcal/mol. This may at first seem to be inconsistent with our ΔG_I between native and unfolded villin estimates. However, until now we have been considering the free energies of the individual snapshots and not the entropy associated with considering all the unfolded states as a conformational ensemble, the conformational entropy. In fact, the experimental ΔG_{denat} can be used in conjunction with ΔG_I to estimate the effective conformational degeneracy of this ensemble. Assuming a Boltzmann distribution of the two-state model, the number of individual denatured conformations, i.e. the degeneracy of the denatured state (Ω_{denat}), can be estimated as follows:

$$P(denat)/P(native) = e^{(-\Delta G_{denat}/RT)} = \Omega_{denat}/\Omega_{nat} \bullet e^{(-\Delta G_I / RT)} \quad (3)$$

where ΔG_I is an average effective G_I difference between native and denatured states. Assuming the degeneracy of the native state (Ω_{nat}) is unity,

$$\Omega_{denat} = e^{[(\Delta G_1 - \Delta G_{denat})/RT]} \quad (4)$$

$$S_{conf} = R\ln\Omega_{denat} = (\Delta G_1 - \Delta G_{denat})/T \quad (5)$$

$$\ln\Omega_{denat} = \Delta\Delta G/RT \quad (6)$$

The total conformational entropy associated with the degeneracy of the denatured state (S_{conf}) and hence the log of Ω_{denat} are directly proportional to $\Delta\Delta G$, the difference between protein stability, ΔG_{denat}, and the effective G_1 difference between a typical native and denatured snapshot, ΔG_1.

From looking at the free energy during the microsecond trajectory, it is not clear whether the free energy difference between the native conformation and the folding intermediate (~26.2 kcal/mol) should be used as ΔG_1 or if instead that between the native and the average non-compact states (~38.5 kcal/mol) is more appropriate. If the compact states are all of approximately similar free energy and together represent the dominating configuration of the denatured state, then we would use a ΔG_1 = 26.2 kcal/mol, giving us the smallest estimate for $\Delta\Delta G$ of 22.9 kcal/mol. This translates to a lower bound degeneracy estimate on the order of $3.8 \bullet 10^{16}$. If on the other hand, the non-compact random coil configurations are Boltzmann-weighted far more than the compact ones, then we would approximate ΔG_1 = 38.5 kcal/mol. This leads to upper bound estimates of $\Delta\Delta G$ = 35.2 kcal/mol and of $\Omega_{denat} = 3.0 \bullet 10^{25}$. The value for Ω_{denat} can then be converted into another interesting value, an average number of degrees of freedom per

residue (y): $y^{36} = \Omega_{denat}$. Our range of estimates for Ω_{denat} corresponds to a range of y values from 2.9 to 5.1, which is in qualitative agreement with Dill's estimates[29] for y.

Estimating the Free Energy of Unfolding.

In the previous section, we attempted to use the experimental ΔG_{denat} together with our ΔG_1 in order to estimate the degeneracy of the non-native state. Alternatively, we can use ΔG_1 together with other degeneracy estimates to obtain a free energy of unfolding and compare that directly with the experimental value. Karplus estimates that in a 27-mer small protein, there is a mixture of 10^{10} "semi-compact globule" conformations and 10^{16} random coil conformations[25], which correspond to y values of 2.3 and 3.9, respectively, and Ω_{denat} values for a 36-mer of $2.2 \bullet 10^{13}$ and $2.2 \bullet 10^{21}$, respectively. By splitting the denatured state probability shown in equation (3) into a sum of two probabilities, again assuming Ω_{nat} is unity, Karplus's estimates for the degeneracy of compact ($\Omega_{compact}$) and random coil (Ω_{RC}) states can be used together with the respective ΔG_1 predictions in this study ($\Delta G_{1,compact}$ and $\Delta G_{1,RC}$),.

$$\begin{aligned} P(denat)/P(native) &= P(compact)/P(native) + P(random\ coil)/P(native) \\ &= \Omega_{compact} \bullet e^{(-\Delta G_{1,compact} / RT)} + \Omega_{RC} \bullet e^{(-\Delta G_{1,RC} / RT)} \end{aligned} \qquad (7)$$

This results in a ΔG_{denat} of 7.7 kcal/mol with 89% of the denatured state being a compact structure and 11% random coil; thus, the total error is 4.4 kcal/mol.

Estimating the folding kinetics.

Figure 4 attempts to separate the noise of the folding trajectory by looking at the 200 ns running average (thick line). This plot captures the folding intermediate and suggests that a transition state might lie around the 700 ns mark. If this is the case and villin continues to undergo first order exponential decay towards its native state without encountering any further kinetic traps, extrapolating out the smoothed plot leads to a half time for folding of 1.05 μs ($k_f = 6.6 \bullet 10^5$), leading to a total time from the denatured state to 90% "folded" of 4.2 μs.

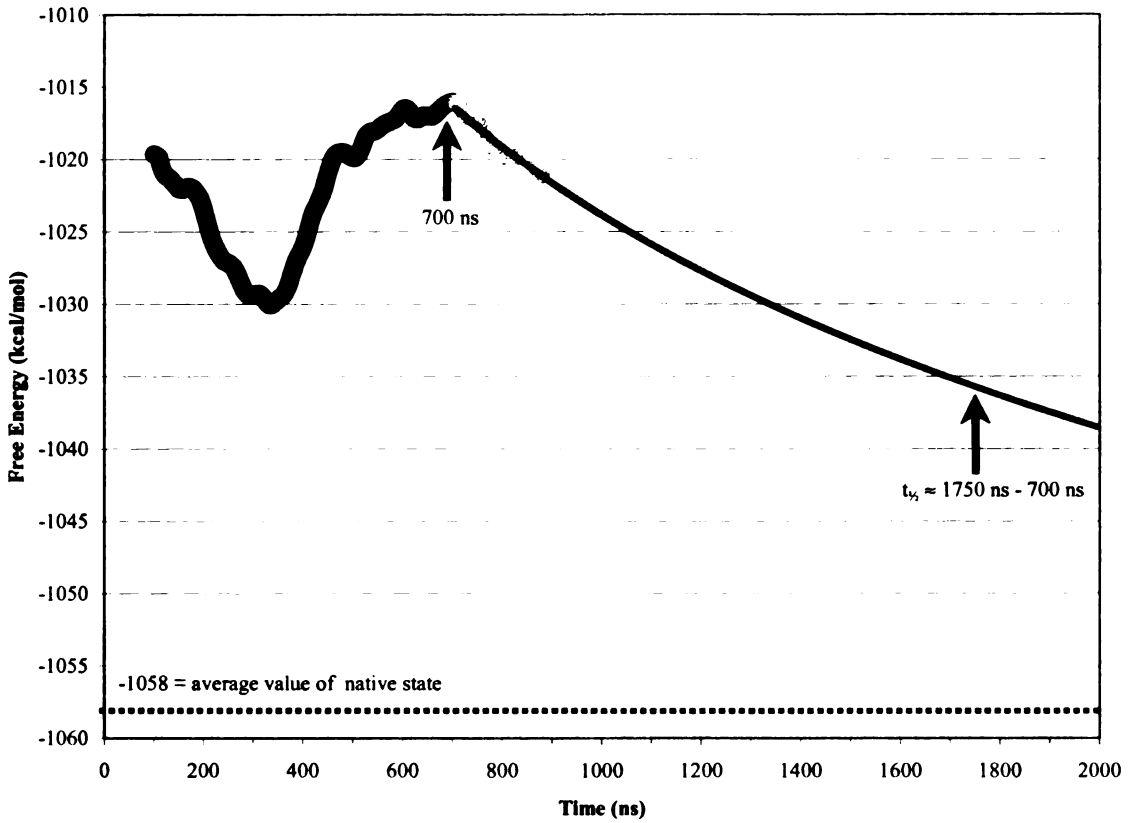


Figure 4. Control simulation results showing (A) actual MM-PB/SA data and (B) 20 running averages. The 20 ns running average of the free energy remains relatively constant at –1058 kcal/mol..

Plaxco *et. al.*[30] have summarized the intrinsic folding rates for a set of 12 non-homologous, simple, single domain proteins and looked at their relationship with size, stability and topology of the proteins. They found that size and stability have weak or non-existent relationships with $\ln(k_f)$, but that the relative contact order (CO), which reflects the relative amount of local and non-local contacts in a protein's native structure, shows a strong inverse correlation (R = -0.81). CO is the average sequence separation distance between all non-hydrogen atoms that are within 6 Å, normalized by the sequence length. CO for the average, minimized, NMR villin structure[20] comes out to 11.0 %. Our

estimate for villin, from the extrapolation of the folding trajectory, of $\ln(k_f) = 13.4$ is consistent with the data presented by Plaxco *et. al.*[30]; adding this data actually increases the magnitude of the correlation coefficient between CO and $\ln(k_f)$ to -0.84. Although the extrapolation leads to strong agreement with the estimate that would arise from villin's CO, there are a number of assumptions made, the two most important are the following.

First, the extrapolation in Fig. 4 assumes that folding will proceed without going through any further metastable intermediates, such as the one found between 240 – 400 ns. In our previous paper,[11] it was concluded, based on villin's CO of 11.0 % and the Plaxco/Baker CO least squares line, that villin folds on a 10-100 μs time scale. There, we suggested that the villin headpiece may continue to fall into metastable intermediates until it found one that was close enough to the more stable native structure to allow it to reach the native state by further subtle readjustments of the structure. Again approximating the 90% "folded" state as native, this range of folding times (10-100μs) would correspond to a range for the half time of folding between 3.3 and 33.3 μs, each of which also increases the correlation between CO and $\ln(k_f)$ to -0.84 and -0.85, respectively. At this point, it is not clear which picture is correct.

Secondly, although the native structure is of significantly lower free energy according to the combined molecular mechanical/continuum model (MM-PB/SA) than anything found in the folding trajectory so far, it is not known at this point how closely this free energy model can reproduce the "true" native global free energy minimum of villin.

CONCLUSIONS:

The recent completion of a one microsecond folding simulation has allowed us to demonstrate that the MM-PB/SA method can successfully identify the native conformation from other compact structures in a small, single domain protein, the villin headpiece. As the folding trajectory formed the very compact intermediate, the biggest change in energy was a drop in the dispersion energy to levels as favorable as in the native state, and during this simulation we found a high correlation between the dispersion energy and Rγ. However, the folding intermediate was only able to accomplish such favorable van der Waals contacts at the expense of exhibiting more internal strain, particularly in the angle term, which was the key term more favorable in the native state than in the intermediate.

The differences in MM-PB/SA free energies of villin between native and the non-native structures, combined with the estimated free energy of unfolding, leads to an estimate of the conformational degeneracy in the non-native state between 10^{16} and 10^{25} or an average number of conformations per residue between 2.9 and 5.1. Smoothing the energies over a large window leads to an apparent transition state for folding at 700 ns in the trajectory. If one assumes no further kinetic traps, our estimate is that it may take an additional 3.5 μs to fold villin from this point.

REFERENCES:

[1] Dill KA ,Chan HS. From Levinthal to pathways to funnels. Nat. Struct. Biol. 1997; 4:10-19.

[2] Dill KA, Bromberg S, Yue K, Feibig KM, Yee DP, Thomas PD & Chan HS. Principles of protein folding--a perspective from simple exact models. Protein Sci. 1994; 4:561-602.

[3] Skolnick, J; Kolinski, A; Ortiz, AR. MONSSTER: a method for folding globular proteins with a small number of distance restraints. J. Mol. Biol. 1997; 265:217-41.

[4] Kolinski A, Skolnick J. Lattice Models of Protein Folding, Dynamics and Thermodynamics. Austin: R. G. Landes Company; 1996.

[5] Alonso DO, Daggett V. Molecular dynamics simulations of protein unfolding and limited refolding: characterization of partially unfolded states of ubiquitin in 60% methanol and in water. J. Mol. Biol. 1995; 247:501-20.

[6] Alonso, DO; Daggett, V. Molecular dynamics simulations of hydrophobic collapse of ubiquitin. Protein Sci. 1998; 7:860-74.

[7] Li A, Daggett V. Molecular dynamics simulation of the unfolding of barnase: characterization of the major intermediate. J. Mol. Biol. 1998; 275:677-94.

[8] Tirado-Rives J, Orozco M & Jorgensen WL. Molecular dynamics simulations of the unfolding of barnase in water and 8 M aqueous urea. Biochemistry, 1997; 36:7313-29.

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

[10] Duan Y, Wang L, Kollman PA. The early stage of folding of villin headpiece subdomain observed in a 200-nanosecond fully solvated molecular dynamics simulation. Proc. Natl. Acad. Sci. USA 1998; 95:9897-9902.

[11] Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 1998; 282:740-744.

[12] Moult J, Hubbard T, Bryant SH, Fidelis K & Pedersen JT. Critical assessment of methods of protein structure prediction (CASP): round II. Proteins 1997; Suppl, 1:2-6.

[13] Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999; 34:82-95.

[14] Novotný J, Bruccoleri R, Karplus M. An analysis of incorrectly folded protein models. Implications for structure predictions. J. Mol. Bio. 1984; 177:787-818.

[15] Martin ACR, MacArthur MW, Thornton JM. Assessment of comparative modeling in CASP2. Proteins 1997; Suppl, 1:14-28.

[16] Srinivasan J, Miller J, Kollman PA, Case DA. Continuum solvent studies of the stability of RNA hairpin loops and helices. J. Biol. Struct. Dyn.1998; 16:671-82.

[17] Cheatham TE 3rd, Srinivasan J, Case DA & Kollman PA. Molecular dynamics and continuum solvent studies of the stability of polyG-polyC and polyA-polyT DNA duplexes in solution. J. Biol. Struct. Dyn. 1998; 16:265-80.

[18] Srinivasan J, Cheatham TE 3rd, Cieplak P, Kollman PA & Case DA. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate - DNA helices. J. Am. Chem. Soc. 1998; 120:9401-9409.

[19] Vorobjev YN, Almagro JC, Hermans J. Discrimination between native and intentionally misfolded conformations of proteins: ES/IS, a new method for calculating conformational free energy that uses both dynamics simulations with an explicit solvent and an implicit solvent continuum model. Proteins 1998; 32:399-413.

[20] McKnight CJ, Doering DS, Matsudaira PT, Kim PS. A thermostable 35-residue subdomain within villin headpiece. J. Mol. Biol. 1996; 260:126-134.

[21] Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J. Am. Chem. Soc. 1995; 117:5179-5197.

[22] Sanner MF, Olson AJ, Spehner JC. Reduced surface: an efficient way to compute molecular surfaces. Biopolymers 1996; 38:305-20.

[23] Gilson MK, Honig B. Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. Proteins 1998; 4:7-18.

[24] Sitkoff D, Sharp KA, Honig B. Accurate calculation of hydration free energies using macroscopic solvent models. J. Phys. Chem. 1994; 98:1978-1988.

[25] Sali A, Shakhnovich E, Karplus M. How does a protein fold? Nature 1994; 369:248-251.

[26] Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins 1995; 21:167-195.

[27] Onuchic JN, Wolynes PG, Luthey-Schulten Z, Socci ND. Toward an outline of the topography of a realistic protein-folding funnel. Proc. Natl. Acad. Sci. USA 1995; 92:3626-3630.

[28] Pande VS, Grosberg AY, Tanaka T. On the theory of folding kinetics for short proteins. Fold. Des. 1997;2:109-114.

[29] Dill KA. Theory for folding and stability of globular proteins. Biochemistry 1985; 24:1501-1509

[30] Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J. Mol. Biol. 1998; 277:985-994.

Chapter 3

2.1 and 1.8 Å <Cα RMSD> Structure Predictions on Two Small Proteins, HP-36 and S15

Matthew R. Lee[1]

David Baker[2]

Peter A. Kollman[1]

[1] Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA 94143-0446, United States of America

[2] Department of Biochemistry, University of Washington, Seattle, Washington 98195, United States of America

J. Am. Chem. Soc. (2001) **123**:1040-1046

ABSTRACT:

On two different small proteins, the 36-mer villin headpiece domain (HP-36) and the 65-mer structured region of ribosomal protein (S15), several model predictions from the *ab initio* approach Rosetta were subjected to molecular dynamics simulations for refinement. After clustering the resulting trajectories into conformational families, the average Molecular Mechanics-Poisson Boltzmann (MM-PBSA) free energies and C_α RMSD's were then calculated for each family. Those conformational families with the lowest average free energies also contained the best C_α RMSD structures (1.4 Å for S15 and for HP-36 core) and the lowest average C_α RMSD's (1.8 Å for S15 and 2.1 Å for HP-36 core). For comparison, control simulations starting with the two experimental structures were very stable, each consisting of a single conformational family, with an average C_α RMSD of 1.3 Å for S15 and 1.2 Å for HP-36 core (1.9 Å over all-residues). In addition, the average free energies rank correlate strongly with the average C_α RMSD's (r_s = 0.77 for HP-36 and r_s = 0.83 for S15). Molecular dynamics simulations combined with the MM-PBSA free energy function provide a potentially powerful tool for the protein structure prediction community in allowing for both high resolution structural refinement and accurate ranking of model predictions. With all the information that genomics is now providing, this methodology may allow for advances in going from sequence to structure.

INTRODUCTION:

While the concerted effort in genomics rapidly uncovers a vast number of new gene sequences, the gap between known sequences and structures grows ever larger, thereby increasing the usefulness and interest in meaningful structural information that non-experimental methods can provide. There are two important challenges in protein structure prediction.

The first challenge is to generate higher resolution structure predictions, especially when sequence identity is low. The most recent community-wide Critical Assessment of Structure Prediction experiment, CASP III, serves as the best forum to evaluate the current state of protein structure prediction. Of the "*ab initio* targets", defined as those having no close structural relatives in the PDB, results were promising in that for roughly half of the easy to medium difficulty targets, approximately 60% of the predictions were successful in obtaining the correct architectures[1]. However, to be useful for contributing to a greater understanding of function or for experimental design, much more than the correct architecture must be in place, a deficiency in nearly every CASP III 3D coordinate prediction of *ab initio* targets. Of the 12 *ab initio* targets that had more than two α-helices, not a single prediction of those with > 60 % coverage (the percentage of target residues that was modeled) had a C_α RMSD over all modeled residues of less than 7.0 Å; the vast majority were well over 10.0 Å away. Because of the enormously complex energy landscape of proteins, the number of local minima must be reduced by *ab initio* or comparative methods in order to obtain a good set of predictions in a reasonable amount of time. The approach of the Rosetta protein folding algorithm is to work from the bottom up, first modeling local structure and then performing tertiary

assembly. The effect of simplifying the energy landscape, however, is that the native state can no longer be as readily discriminated from among the *ab initio* predictions. Bringing these predictions to the realm of molecular mechanics introduces much of the physics back into the system, resulting in a more accurate free energy landscape. Ever since accurate methods for treatment of long range electrostatics, such as particle mesh Ewald[2], have been included in molecular dynamics simulations, simulations on experimental structures of biomolecules have remained within 1 to 2 Å RMSD[3,4], while those on non-native structures steadily drift into new conformational families (this work, unpublished results Duan and Kollman[5] and Alonso & Daggett[6]), suggesting that the native states are indeed at the global free energy minimum of a molecular mechanics representation. Thus, if conformational space could be exhaustively explored in a molecular dynamics or Monte Carlo simulation, the native state should be capable of being found. Moreover, in the interest of protein structure prediction, if the energy landscape is globally convex as is widely believed, extended dynamics simulations should be able to drive non-native conformations down the free energy gradient closer to the native state.

The second important challenge is to be able to more accurately rank the large number of structure predictions that emerge, even within a single prediction method on any given protein. Due to the necessary limitations of the community wide experiment, only 5 or fewer 3D coordinate predictions per group were submitted. Hence, an inability to accurately rank the native structural quality of predictions in the absence of an experimental structure for any prediction method will usually preclude the best predictions from being identified. Without a standard for comparing coordinates, scoring functions together with physically meaningful (and often subjective) measures like

compactness and the numbers of surface-exposed hydrophobic residues and unpaired buried polar residues can sometimes identify good conformations, but are rarely if ever able to identify those predictions that most closely resemble the native state. Therefore, the need remains for a highly accurate free energy function that can capture the same subtle differences that allow nature to guide a protein to its native conformation in order to help identify the best predictions in an unbiased way. Such a free energy function may also help to reveal the relative importance of underlying forces involved in protein stability, another deficiency highlighted by the assessors of CASP III. Vorobjev *et al.*[7] were the first to apply a physics-based effective free energy potential involving gas phase internal energy calculations combined with implicit solvent on a limited set of native and intentionally "misfolded" proteins. After generating conformational ensembles with explicit solvent molecular dynamics on 9 of the 22 pairs of native and misfolded proteins created by Holm & Sander[8] (the EMBL set), then calculating the average free energy of the ensembles, they found the native to always be more favorable. Lazardis & Karplus[9] later demonstrated that their effective free energy can discriminate native structures from a more extensive series of misfolded structures, including the entire EMBL set, and the decoy set of Park & Levitt[10]. We recently applied an effective free energy potential, Molecular Mechanics-Poisson Boltzmann/Surface Area (MM-PBSA), to HP-36 in which we correctly ranked the native structure, an early stage "on-pathway" folding intermediate and an ensemble of unfolded conformers, with physically meaningful relative differences[11]. As previously discussed[9,11], an advantage of these physics-based methods is that, due to the difference in conformational entropy between the unfolded and native states, the energy not only favors the native state, but must be of appreciable size. This sizeable gap should be directly related to the number of residues, as larger

proteins have more degrees of freedom and thus a greater degeneracy of the unfolded state.

In the current study, we meet both of the challenges of protein structure prediction in the context of two small proteins. We run extended molecular dynamics simulations that lead to higher resolution structure predictions in both cases. We also demonstrate how robust the MM-PBSA method is in distinguishing a small handful of "off-pathway" *ab intio* model predictions from one another and from the native configuration, and evaluate its ability to identify any forces among the predictions that might account for some having more native quality than others.

METHODS:

Rosetta.

Rosetta builds protein structures from fragments with similar amino acid sequences using a fragment insertion-simulated annealing method for searching conformational space and a simple side chain centroid based energy/scoring function which favors hydrophobic burial, strand pairing, and other low resolution features of native protein structures. Structures were generated for the two sequences studied here with the method used for the Rosetta predictions in the CASP3 structure prediction experiment (Proteins suppl3, 1999), except that homologues of the two proteins were excluded from the fragment libraries. For HP-36, sidechains were added in using the backbone-dependent library of SCWRL[12].

Molecular dynamics.

We ran production-phase molecular dynamics with a 2.0 fs timestep under the isothermal-isobaric ensemble (300 K and 1 atm) with the Cornell et al. all-atom force field[13], the TIP3P[14] model for water, periodic boundary conditions, the particle mesh Ewald method (PME)[2] for electrostatics, a 10 Å cutoff for Lennard-Jones interactions, and the use of SHAKE[15] for restricting motion of all covalent bonds involving hydrogen, all within the AMBER 5.0 suite of programs[16]. 2816 TIP3P water molecules were added around HP-36 and 3000 around S15 in order to end up with a buffer of about 10 Å from the edge of the periodic box, resulting in box sizes approximately 90,000 $Å^3$ for HP-36 and 160,000 $Å^3$ for S15. Temperature was maintained by the Berendsen coupling algorithm[17] using separate τ coupling constants of 1.0 for the protein and solvent and pressure was maintained with isotropic molecule-based scaling[17], also with a τ coupling constant of 1.0. The PME grid spacing was ~1.0 Å and was interpolated on a cubic B-spline, with direct sum tolerance set to 10^{-5}. We removed the net center of velocity every 100 ps to correct for the small energy drains that result from the use of SHAKE, discontinuity in the potential energy near the Lennard Jones cutoff value and constant pressure conditions.

For equilibration, we first minimized the solutes with the steepest descent method for the first 500 steps, followed by the conjugate gradient method until the RMS of the Cartesian elements of the gradient was less than 0.4 kcal/mol•Å. Water molecules alone were then minimized in the same way until the RMS was less than 0.1 kcal/mol•Å and then slowly heated, while allowing them to move unrestrained for 25 ps (with a 1 fs timestep) in order to fill in any vacuum pockets. The solute atoms alone were then minimized in the presence of ever decreasing positional restraints, thereby allowing them

to slowly feel the forces of the now equilibrated waters, until the positional restraints reached zero. Finally, a temperature ramp was used to gradually raise the temperature of the whole system over 20 ps up to 300 K.

In order to cluster the molecular dynamics trajectories, we defined conformational families as being those with C_α RMSD's of less than 2.0 Å from the cluster average. In cases where the value was greater than 2.0 Å from any cluster, we placed them in the most representative conformational family, with every structure being a member of a single family. We analyzed the trajectories using AMBER 5.0, Procheck[18] and UCSF MidasPlus[19]. Simulations were run on Origin200's at UCSF and on the Origin2000 at the National Center for Supercomputing Applications.

Post-processing the energy of the trajectory data.

Coordinates from the trajectory were saved every 5 ps with the MM-PBSA calculation evaluation performed on each of them. The MM-PBSA free energy of each snapshot is approximated as the sum of two terms: the internal energy of the protein (E_{MM}) and a solvation free energy (ΔG_{solv}).

$$G_1 = E_{MM} + \Delta G_{solv} \qquad (1)$$

E_{MM} is the sum of an internal strain energy (E_{int}), a van der Waals energy (vdW), and an electrostatic energy (EEL). E_{int} is the energy associated with vibration of covalent bonds and rotation of valence bond angles and torsional angles. vdW and EEL are further broken down into short range values, those that are within three covalent bonds ($vdW_{1\text{-}4}$ and $EEL_{1\text{-}4}$), and long range values (vdW_{NB} and EEL_{NB}).

The entropy of a given snapshot, which will loosely be referred to as the "vibrational" entropy, can be estimated by calculating the translational, rotational and vibrational partition functions with normal mode analysis on a Newton-Raphson minimization ($T \bullet S_{solute}$). This, however, is the most time-intensive part of the MM-PBSA method on a per snapshot basis. Given the results in our previous study[11], where we found this term to be indistinguishable among the native state, the folding intermediate and the unfolded state of HP-36, we did not perform this calculation in the current study.

Obtaining the solvation free energy from an implicit description of solvent as a continuum is advantageous because it affords a solvation potential that is a function only of the solute's geometry, as discussed and implemented by Srinivasan et al.[20], thereby making it computationally tractable. In contrast, calculating the entire free energy from the explicit solvent is very impractical. It would require a very costly potential of mean force calculation since the simulations on different conformations have little overlap in phase space and the partition function of the system including explicit waters would take an extremely long time to calculate, largely due to the fact the water structures do not converge.

$$\Delta G_{solv} = \langle \Delta G_{solv,NP} \rangle + \langle \Delta G_{solv,elec} \rangle \approx (\gamma \bullet SASA + b) + \langle \Delta G_{solv,elec} \rangle \quad (2)$$

The non-polar solvation free energy ($\Delta G_{solv,NP}$) includes the (largely entropic) cost of creating a solute-sized cavity in solvent and the free energy of inserting the discharged solute into that cavity. Also referred to as the first solvation shell effects, this term has been found experimentally in hydrocarbons to be linearly related to the solvent accessible surface area (SASA), which is obtained from Sanner's MSMS algorithm[21] (probe radius

= 1.400 Å). The γ coefficient is set to 5.42 cal/mol • $Å^2$ and b is set to 920 cal/mol. The electrostatic solvation free energy ($\Delta G_{solv,elec}$) is the cost of charging the discharged solute in the cavity. We adhered to the same Poisson-Boltzmann protocol as described by Srinivasan et al.[20], which uses DelPhi[22] and most of its standard default parameters, together with PARSE atomic radii[23] and Cornell et al. charges[13], to calculate the electrostatic solvation free energy difference for the system between exterior dielectrics of 80 (solvent) and unity (gas phase) according to the position dependent electrostatic potential. One small difference in this current application of DelPhi is to use larger grid spacing of 0.5 Å, extending 20% beyond the edge of the solute. Additionally, we used fewer finite difference iterations (1000) for each ($\Delta G_{solv,elec}$) calculation, which was still amply sufficient as we found the values in this system to reach 90% convergence at around 50 iterations.

RESULTS & DISCUSSION:

Rosetta results on HP-36 and S15.

The Rosetta method, as previously described[24], rapidly generates ~1000 structure predictions with centroid sidechains in a matter of hours. The four HP-36 models chosen for this study, labeled 17, 18, 54 and 60, ranged in global similarity to the experimental structure from 2.76 to 8.47 Å C_α RMSD (Table 1). These four were selected as they were centers of the four most highly populated clusters from the initial 1000 Rosetta predictions. The five S15 models, labeled 0, 43, 112, 156 and 471, ranged from 2.14 to 8.06 C_α RMSD (Table 1). For this protein, we screened the 100 best scoring Rosetta models for those with a C_α RMSD < 4.5 Å and selected the three with the best Rosetta

scores (471, 43, and 156); we also selected the two with the best scores (0 and 112) without regards to RMSD. Although the best Rosetta predictions are very good, they are among a larger number of less impressive predictions and the correlation between RMSD and Rosetta score is rather poor; with S15, the best Rosetta scoring conformations had RMSD's of 8.06 and 7.27 Å. This demonstrates the difficulty in blindly selecting the best predictions, even from a method as promising as Rosetta.

For comparison, it may prove useful to look at results on a similar target at CASP III. The CASP III target closest in difficulty to the two proteins investigated in this work was a medium-difficulty "*ab initio* target", the 89-mer protein HDEA, which like the 36-mer HP-36 and the 65-mer structured region of S15 has three alpha helices. A non-*ab initio* threading method from the Bryant group yielded perhaps the best prediction at CASP III for HDEA, which modeled only 54 % of the target residues and had a C_α RMSD of 5.85 Å over those residues, although the model submitted as first by the Bryant group[25] had a much higher C_α RMSD of 10.76 Å. Of the more difficult cases in which most or all of the target residues were modeled, the *ab initio* work of the Scheraga group[26] came up with the best prediction, a model with 100 % coverage and a C_α RMSD of 7.27 Å, while the model submitted as their first had a C_α RMSD of 8.94 Å. Again, the two challenges of protein structure prediction can be seen from the CASP III results of HDEA, where the best predictions 1) still had very high RMSD's and 2) were not the predictions submitted as first.

Simulations on the native structures.

The characteristics of HP-36 and S15 make them good candidates for *ab intio* structure prediction. Because part of our goal was to improve the resolution of structure

predictions, which entails an extended amount of computer time, we chose to study proteins containing the simplest non-trivial topology, which according to the results of CASP III appear to be small alpha proteins containing three secondary structural elements, like HDEA. HP-36 forms three small helices packed together in a novel architecture[27] with the NMR structure (1vii) having much lower B-factors over the core residues 6-33 (with the N-terminal residue 41 renumbered as one). The 86-mer S15 forms 4 helices in the X-ray structure (1a32)[28], although the first 21 N-terminal residues including the N-terminal helix are very disordered and not included in our model structures, with residue 22 renumbered as residue one. In addition to having the same general topology as HDEA, they are reasonably sized and have enough of a hydrophobic core and secondary structure to make them thermostable at room temperature.

Simulations of the experimental structures were carried out as a basis for comparison. Minimization, solvation and equilibration were required prior to the production-phase simulations, leading to small deviations (< 1 Å C_α RMSD) from the experimental coordinates. During the subsequent control simulations of the equilibrated HP-36 NMR structure, the all-residue C_α RMSD was, on average, 1.90 Å away from the NMR structure with a standard deviation of 0.29 Å (Figure 2A); over the core region, the average C_α RMSD was 1.20 Å with a standard deviation of 0.16 Å (Figure 2B). The difference in these C_α RMSD's is consistent with the distribution of experimental B-factors. Those with the highest B-factors exhibited the most fluctuation. The corresponding control simulation on S15 led to an all-residue C_α RMSD of 1.26 Å from the X-ray structure with a standard deviation of 0.21 Å (Figure 3).

Through clustering the trajectories, we found that both control simulations consisted of a single family, which demonstrates good stability of the native states in our

simulation. This implies that at room temperature, there is not enough thermal energy to overcome a kinetic barrier if the experimental structure should happen to lie outside the global free energy minimum (see discussion below on HP-36), or that the actual global minimum is the same as that resulting from our molecular mechanics energy potential.

Table 1. A summary of the molecular dynamics results

Model	Rosetta Score	Cα RMSD (Å) all residues[1]		Cα RMSD (Å) core region[2]		% Native Contacts[3]	% Native Helical Content[4]	ΔG_{tot}[5] (kcal/mol)	
		init.	avg.	init.	avg.			avg.	SD
HP-36									
$17_{(0\text{-}735)}$	-24.9	5.40	5.18	3.18	2.89	68.5	83.8	35.5	15.2
$18_{(0\text{-}270)}$	-29.5	3.17	3.52	2.70	3.27	73.4	80.6	15.5	14.7
$18_{(270\text{-}1600)}$			2.78		2.14	77.6	80.6	-1.2	16.2
$54_{(0\text{-}960)}$	-27.1	2.76	3.19	2.07	2.87	70.2	89.4	15.2	14.7
$60_{(0\text{-}935)}$	-30.3	8.47	8.41	6.07	6.58	58.2	78.1	15.3	14.4
$Native_{(0\text{-}3000)}$		0.00	1.90	0.00	1.20	90.9	87.7	0.0	15.7
S15									
$0_{(0\text{-}855)}$	45.1	7.27	7.56			72.8	94.2	46.7	18.4
$43_{(0\text{-}200)}$	66.5	4.40	4.87			74.5	90.7	62.1	24.0
$43_{(200\text{-}775)}$			5.09			75.4	90.7	40.8	16.1
$112_{(0\text{-}775)}$	44.1	8.06	9.03			68.3	90.7	52.8	24.8
$156_{(0\text{-}760)}$	78.6	2.14	2.18			87.3	96.3	34.1	18.9
$471_{(0\text{-}500)}$	66.0	2.81	1.81			85.4	96.3	30.5	18.5
$471_{(500\text{-}960)}$			2.86			82.7	96.3	31.0	20.5
$Native_{(0\text{-}1000)}$		0.00	1.26			92.8	96.3	0.0	44.4

The trajectories were clustered, giving rise to conformational families for some of the models. All values except for the initial RMSD's and Rosetta scores are average values over the dynamics.

[1] The S15 all residue RMSD excludes the less ordered N-terminal 21 residues, where the average mainchain temperature factor in the X-ray structure is 40.4, and spans the remaining 65 amino acids, where the average mainchain temperature factor is 25.5.

[2] The HP-36 core region comprises residues 6-33, where the average mainchain B-value in the NMR structure is 0.68, compared to 1.53 outside the core.

[3] A contact is defined by any two residues containing atoms ≤ 3.5 Å apart. There were 89 native contacts in 1vii (HP-36) and 221 in 1a32 (S15).

[4] Residues were assigned as helical if they fell within the core helical region of the Ramachandran map according to Procheck and were contiguous with at least two other helical residues. A total of 20 residues were helical in 1vii and 54 in 1a32.

[5] The average is relative to the native's average. Only 18(270-1600) had an average value comparable to the native's with $P = 0.31$.

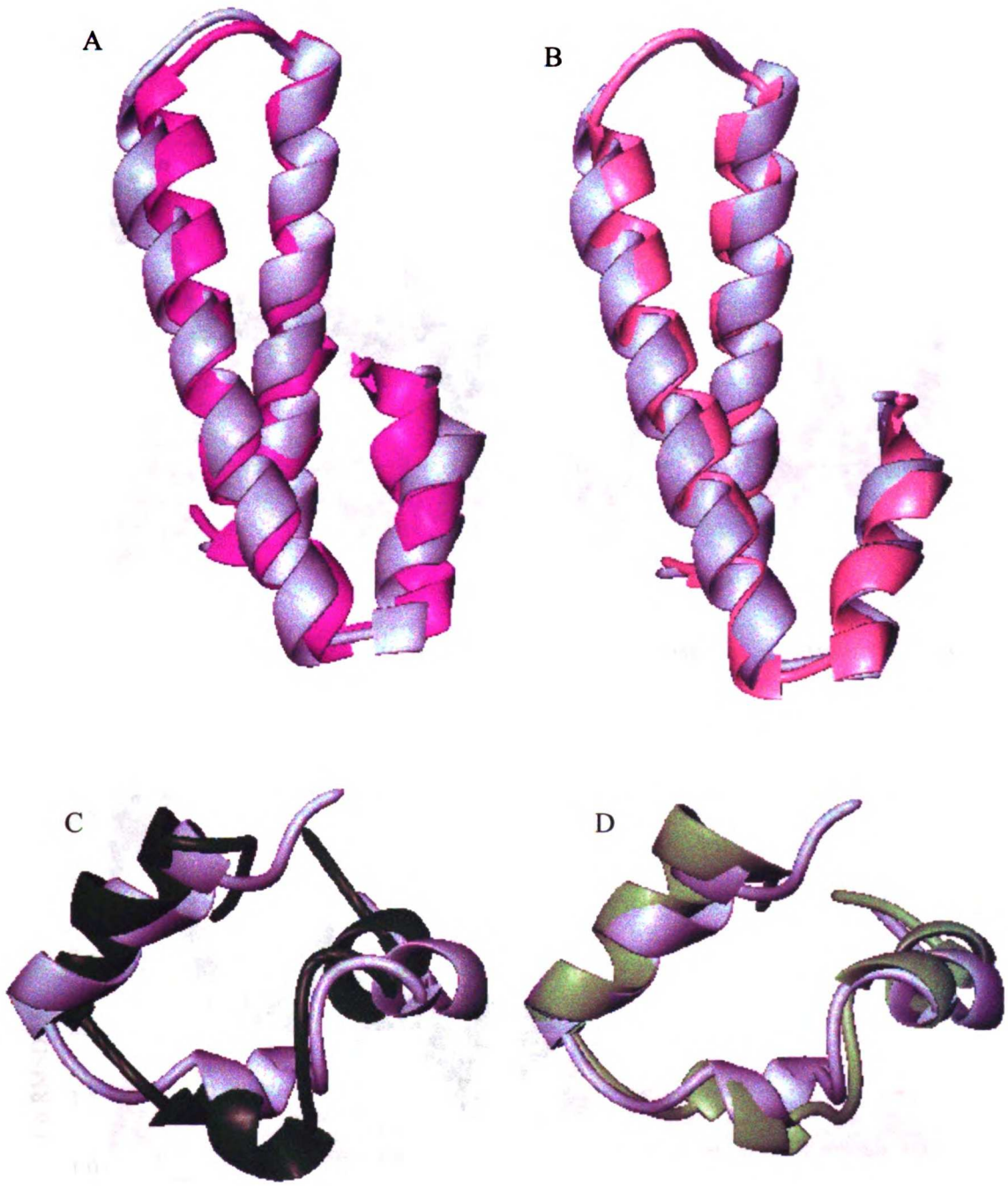


Figure 1. Cartoon diagram comparisons of the experimental structures (shown in gray) with the best ab initio predictions in this study. S15 from the simulation of Rosetta model 156 at 0 ps in magenta (*a*) and at 750 ps, the lowest C_α RMSD structure (1.39 Å), in pink (*b*). HP-36 from the simulation of Rosetta model 18 at 0 ps in dark green (*c*) and at 1250 ps, the lowest core C_α RMSD structure (1.41 Å), in light green (*d*).

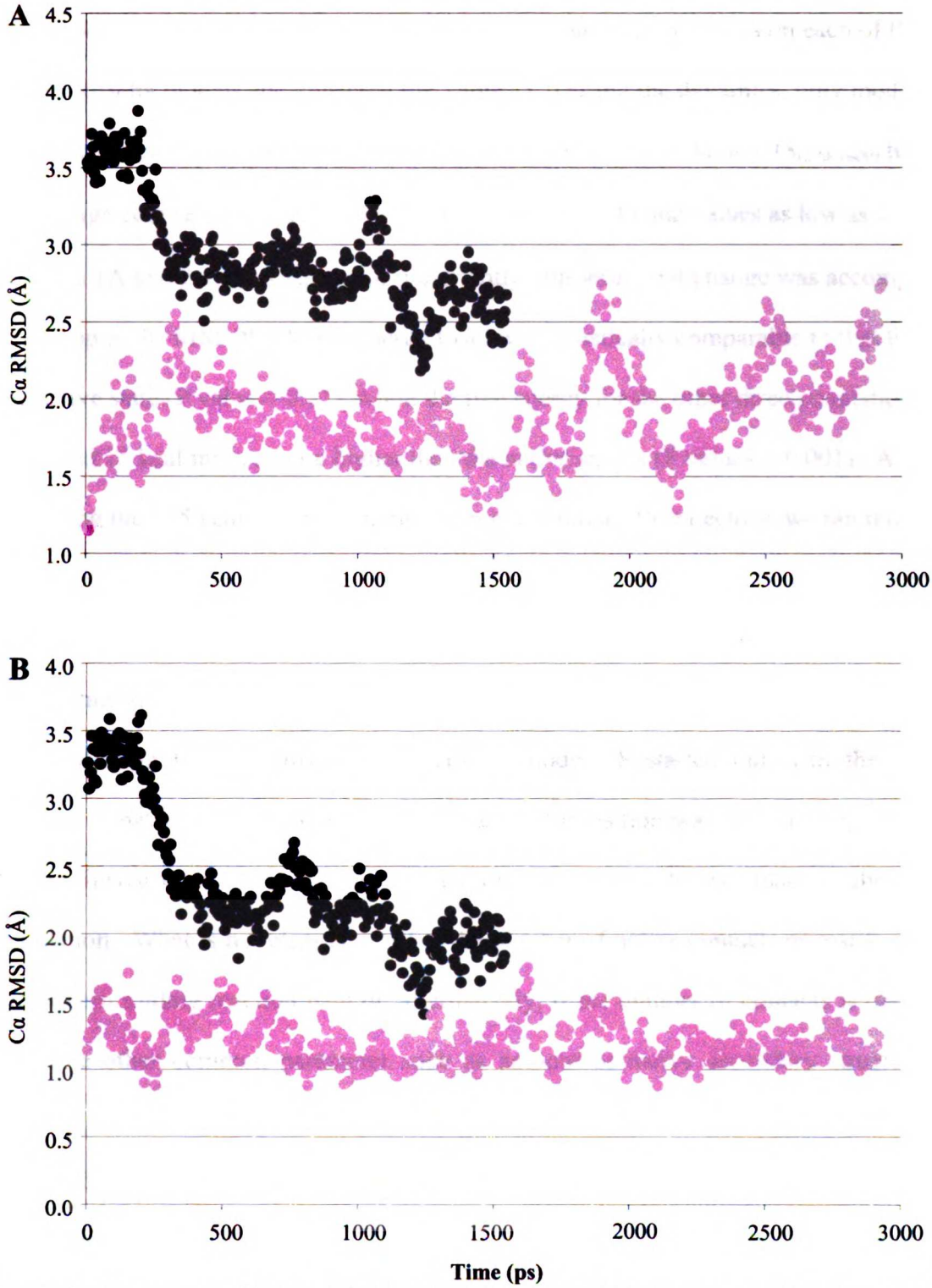


Figure 2. Timecourse of the C_α RMSD of HP-36 vs. the NMR structure, resulting from molecular dynamics simulations in explicit water, starting with the NMR structure (•) or Rosetta model 18 (•). (A) shows the C_α RMSD over all-residues and (B) shows the C_α RMSD over the core region (6-33).

Simulations on Rosetta HP-36 predictions.

We ran approximately one nanosecond of molecular dynamics on each of the HP-36 Rosetta models and clustered the Table 1). During the dynamics, only model 18 underwent a conformational transition (Figure 2), with the new family $18_{(270\text{-}1600)}$ having an average core region C_α RMSD of 2.14 Å (SD = 0.25 Å) and values as low as 1.41 Å (Figures 1A and 2B). Perhaps most importantly, this structural change was accompanied by a drop in the MM-PBSA free energy to a level statistically comparable to that found in the native state (P value = 0.31), while the free energy for the other three simulations remained 15 kcal/mol or more higher than the native state's (P values < 0.001). After observing the ~15 kcal/mol free energy drop in the model 18 trajectory, we ran it out about 50% longer than the others' and did not find any additional structural or energetic changes, which would agree with the structure having a free energy comparable to the native state's.

Among the four Rosetta predictions, model 18 started out with the greatest number of native contacts and the conformational transition was also accompanied by a further increase in native contact formation, although still less than in the control simulation. What is not clear is whether the number of native contacts primarily dictates the protein folding reaction path or, alternatively, if the number of contacts is dependent on some other common parameter such as amount of native secondary structure that primarily governs the reaction path. If the number of native contacts is the major independent parameter in the folding reaction, then the lack of structural improvement in the other three models may have been due to the inability to increase the number of native contacts in the one nanosecond time range.

In the one μs folding simulation of HP-36 by Duan & Kollman[5], secondary structure differed markedly between the native control simulation and every non-native structure, as the simulation was started from an extended state with no secondary structure. Here however, due to the nature of the Rosetta method, all four of the model structures had very reasonable secondary structure, not appreciably different from the control simulation and the structures showed a very poor correlation between the percentage of native helical formation and RMSD. Thus, when comparing compact structures, the amount of native secondary structure is not as good a measure of progress towards the native free energy basin as the number of native contacts.

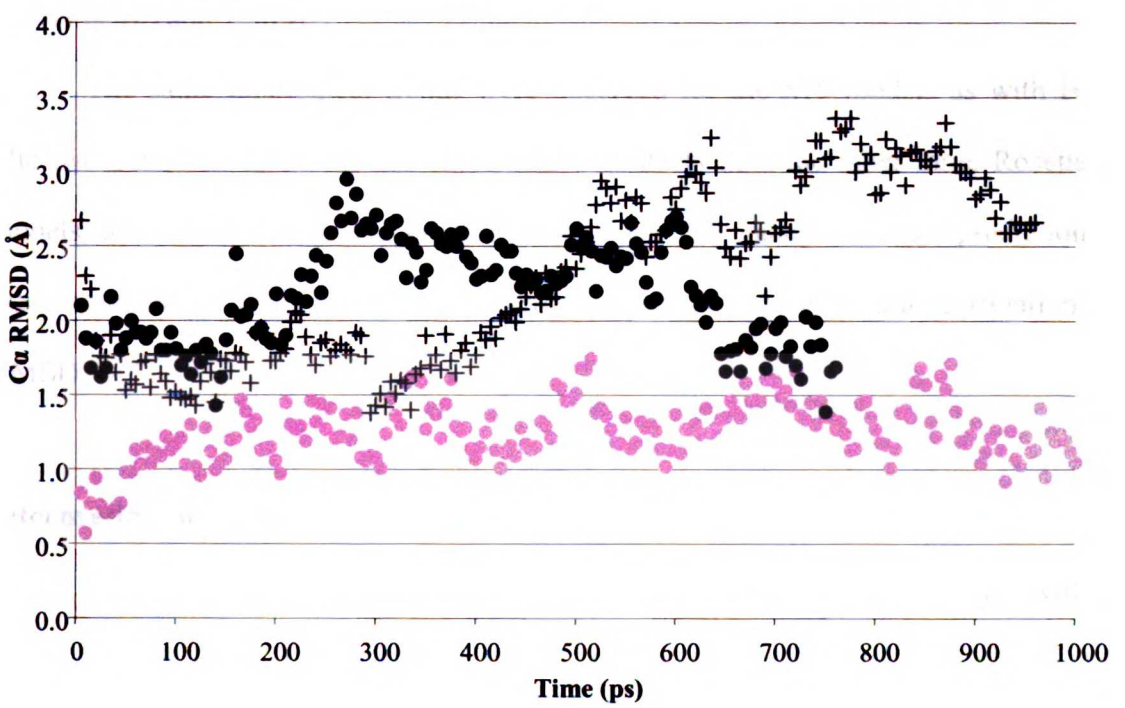


Figure 3. Timecourse of the C_α RMSD of S15 vs. the X-ray structure, resulting from molecular dynamics simulations in explicit water, starting with X-ray structure (•), Rosetta model 156 (•) or Rosetta model 471 (+).

Simulations on Rosetta S15 predictions.

Simulations on the 5 Rosetta S15 models were also carried out for close to one nanosecond (Table 1). Two of these trajectories contained a conformational transition, 43 and 471, neither of which was associated with any improvement. Unlike what we found with HP-36, none of these 7 structural families possesses an average free energy comparable to the native state's (P values < 0.001), although the free energies of the five models did deviate from one another, with the two most energetically favorable models, 156 and 471, also containing the best structures. As can be seen in Figure 3, the average C_α RMSD's of 156 and of 471, prior to its conformational transition (0-500 ps), were 2.18 (SD = 0.34 Å) and 1.81 Å (SD = 0.26 Å), respectively, with minimum values of 1.39 (Fig 1*b*) and 1.38 Å.

The same topological trends were observed for the S15 models as with HP-36. The two best, 156 and 471, had more native contacts than the other three Rosetta S15 models, and still less than the control simulation. Secondary structure prediction was again universally good for all the Rosetta models and showed little correlation with RMSD.

Interpreting the energies.

Like both of the native states, the HP-36 low energy state $18_{(270\text{-}1600)}$ remained stable for over 1 ns. In contrast, one of the two low energy S15 states (model 471) that was still ~30 kcal/mol higher than the native level shifted after ~500 ps into a separate family where the free energy was not statistically different from the initial family's and the geometric similarity to the experimental structure was noticeably diminished. Since the free energy of $471_{(270\text{-}1600)}$ was ~30 kcal/mol higher, it is not unexpected that it unlike

the native would transition into another state. These observations reflect the non-linear relationship between C_α RMSD and G_{tot} that one would expect even from a funnel-shaped energy landscape; structures having similar free energies may differ significantly in terms of their geometries, particularly so the higher they are in free energy. Thus, the Spearman rank (r_s) correlation coefficient is more appropriate for this relationship than the Pearson product-moment correlation coefficient, which is relevant for linear relationships between two variables. Figure 2 shows the relationship between C_α RMSD and ΔG_{tot} for the two proteins investigated in this work. As mentioned in the Introduction, because the conformational entropy and thus ΔG_{tot} (which does not account for S_{conf}) are dependent on the number of residues[11], the strength of the relationship should be looked at separately for the two proteins. For S15, $r_s = 0.83$ (n=8) and for HP-36, $r_s = 0.77$ (n=6). Statistically, there is a strong association for both S15 and HP-36 between C_α RMSD and ΔG_{tot}; given their sample sizes, both of their r_s values exceed the critical level for rejecting the null hypothesis of no relationship with $P < 0.002$. It should also be noted that apart from HP-36 model 18, which may be an alternative global minimum (see below), the smallest relative free energy value seen is 15 kcal/mol in the 36-mer HP-36 and 30 kcal/mol in the 65-mer S15, further corroborating the hypothesis that the energy gap between native state and any non-native state is directly related to the size of the protein.

A benefit of using the physics-based MM-PBSA free energy as a scoring function is that individual force contributions can be readily examined and compared among the successful and unsuccessful model predictions. Our data here (Table 2) and previously[11] suggests that van der Waals interactions are what primarily set apart the native state from the non-native states, which likely can only be properly achieved by precise packing of

the sidechains. The S15 model simulations all had van der Waals energies of 30 or more kcal/mol higher than the native, with this term also being the dominant component separating the two best MM-PBSA scorers, 156 and 471, from the native. With HP-36, none of the four predictions achieved the native van der Waals energy, although with model 18, the conformational change was associated with a sharp drop in the total electrostatics energy that was enough to compensate for the less favorable van der Waals energies to allow for a total free energy equal to that of the native state. While the van der Waals energy correlates best with RMSD, model 54 has a more favorable van der Waals energy than the second conformation of model 18; however, the total MM-PBSA still favors the latter, and the native state still has best van der Waals energy among all the HP-36 conformational states.

Table 2. Comparing the energy components

Model	ΔE_{strain}[1]	ΔE_{vdW}[2]	$\Delta\Delta G_{solv,NP}$[3]	ΔG_{elec}[4]	ΔG_{tot}
HP-36					
$17_{(0\text{-}735)}$	13.43	6.26	-0.70	16.53	35.52
$18_{(0\text{-}270)}$	8.80	9.45	0.03	-2.74	15.54
$18_{(270\text{-}1600)}$	4.11	8.30	0.34	-13.98	-1.22
$54_{(0\text{-}960)}$	6.24	3.35	-0.44	6.05	15.20
$60_{(0\text{-}935)}$	-4.80	11.21	0.73	8.12	15.25
$Native_{(0\text{-}3000)}$	0.00	0.00[5]	0.00[5]	0.00	0.00
S15					
$0_{(0\text{-}855)}$	-12.36	38.20	-0.02	15.71	46.73
$43_{(0\text{-}200)}$	4.57	40.50	1.10	10.78	62.14
$43_{(200\text{-}775)}$	1.84	35.38	1.74	-3.40	40.76
$112_{(0\text{-}775)}$	-0.70	41.61	1.28	5.38	52.77
$156_{(0\text{-}760)}$	1.25	32.46	1.74	-6.55	34.09
$471_{(0\text{-}500)}$	-6.11	33.38	1.63	-3.61	30.50
$471_{(500\text{-}960)}$	-1.71	30.66	0.87	-3.78	30.98
$Native_{(0\text{-}1000)}$	0.00	0.00[6]	0.00[6]	0.00	0.00

All values are in kcal/mol, are averages for the structural family, and are relative to the native states

[1] internal strain energy associated with bond, angle and dihedral motions away from their reference values

[2] intra-protein Lennard Jones potential energy

[3] non-polar contribution to the solvation free energy

[4] sum of intra-protein Coulombic energy and electrostatic element of the solvation free energy

[5] absolute values for HP-36 E_{vdW} and $\Delta G_{solv,NP}$ are -113.3 and 18.2 kcal/mol, respectively

[6] absolute values for S15 E_{vdW} and $\Delta G_{solv,NP}$ are -255.9 and 29.5 kcal/mol, respectively

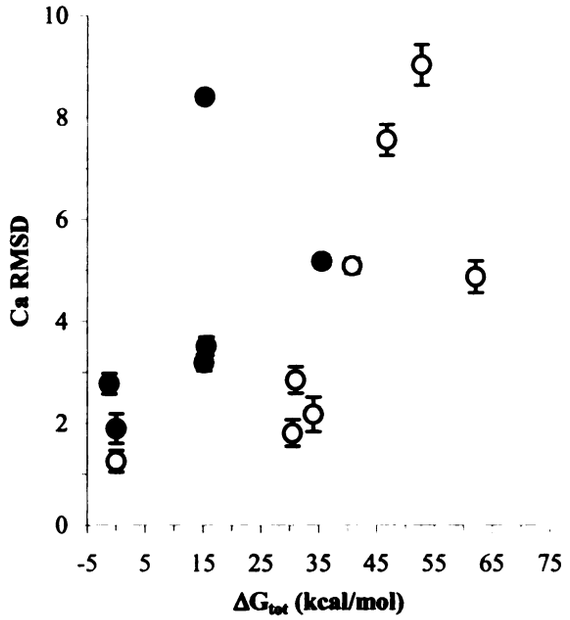


Figure 4. Plot showing correlation between average values C_α RMSD and ΔG_{tot} for HP-36 (•) and S15 (○), with each data point representing a separate conformational family.

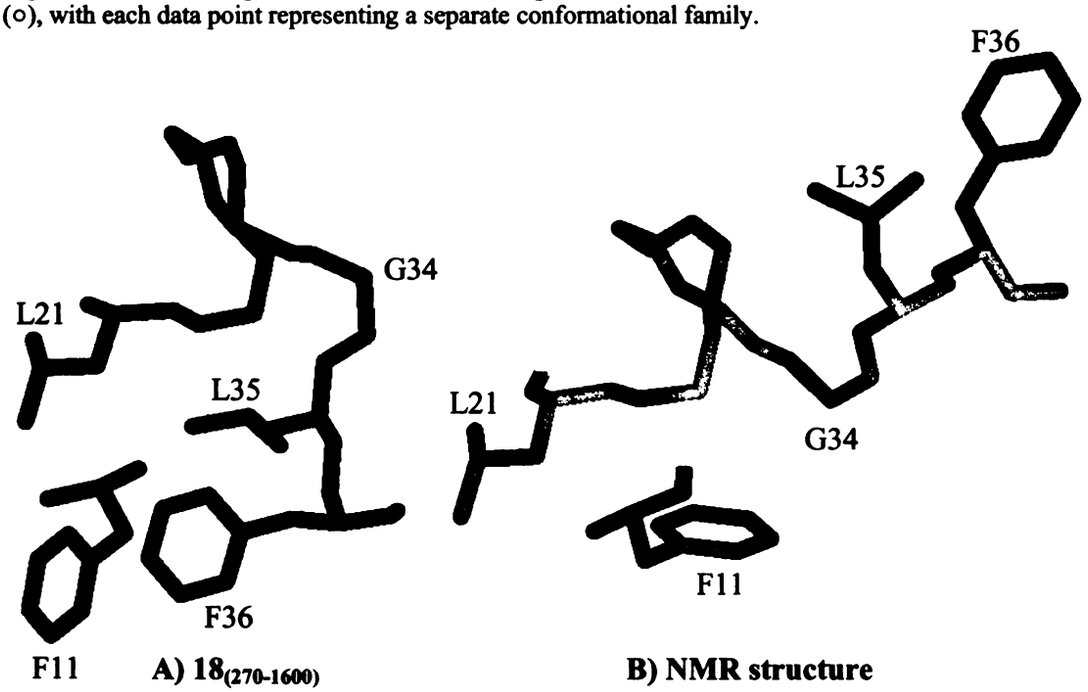


Figure 5. Illustration of the C-termini, residues 21-36, demonstrating the region of greatest geometric disparity between the average structure from the $18_{(270\text{-}1600)}$ low energy state (A) and the NMR structure (B). For clarity, only the hydrophobic sidechains are shown together with the backbone N, CA, and C atoms. For reference, phenylalanine 11 is shown as well. Solvent lies to the right of glycine 34, with the two hydrophobic residues on the NMR structure being solvent exposed.

That HP-36 $18_{(270\text{-}1600)}$ and the native state lie at the same free energy level is rather intriguing. Table 2 suggests that while their total free energies are similar, the native state forms better van der Waals interactions and has a poorer overall charge distribution, which more specifically arises from weaker a solute-solvent electrostatic interaction (data not shown). We find that the degree of charge burial is higher in model 18 than in the native structure; perhaps the Poisson equation is not sufficiently penalizing model 18 for its charge burial, which could possibly explain why our calculations show it having a better solute-solvent electrostatic energy. However, it is also possible that the NMR structure does have worse electrostatics than model 18. Figure 3 depicts the C-termini of both states, the region where they differ most. Particularly interesting is how in the NMR structure the two hydrophobic endmost residues L35 and F36, which happen to be the mostly highly disordered monomers, are almost completely solvent exposed, forming a separate miniature hydrophobic cluster. In contrast, the average structure from the $18_{(270\text{-}1600)}$ low energy state has the L35 and F36 sidechains packed against the core of the protein, with the polar backbone atoms instead being solvent exposed. Given these topologies, we believe it is likely that the NMR structure may not be the single most energetically favorable conformation and can find no structural basis for why $18_{(270\text{-}1600)}$ should not have a free energy as favorable as that from the native state. Perhaps prior to expression of the final two C-terminal residues, a highly stable core that includes several hydrophobic interactions locks the protein into a kinetic trap. At this point, we do not know how much of the difference in ΔG_{elec} is real and how much of it is artifactual.

Efficiency.

In each of the conformational families containing the equilibrated initial structure, the average free energies and C_α RMSD's from every 10^{th} ps over the first 150 ps (n = 15) give good agreement with the averages taken from every 5^{th} ps over the entire window (Table 3). With the method described in this work, one can pragmatically rank 5 to 10 small protein structure predictions using two SGI R10000 processors in about one month by running 150 ps of molecular dynamics on each model prediction. With a dedicated 64-node SGI Origin, one can conceivably rank ~150-300 structures in one month by running in coarse grain parallel, although the human intervention associated with this kind of setup would lead to a considerable slowdown. If one instead seeks to accomplish structural refinement, such as that found with some of the Rosetta model predictions in this work, simulations much longer than 150 ps may be necessary. To carry out one ns of simulation time, as we did for each of the model predictions in this study, one can expect to spend upwards of one month of computer time on a single SGI R10000 processor per model conformation of a small protein.

Table 3. Statistical Efficiency

	P value[1]
HP-36	MM-PBSA
$17_{(0\text{-}735)}$	0.43
$18_{(0\text{-}270)}$	0.87
$54_{(0\text{-}960)}$	0.87
$60_{(0\text{-}935)}$	0.09
$\text{Native}_{(0\text{-}300}$	0.05
S15	
$0_{(0\text{-}855)}$	0.90
$43_{(0\text{-}200)}$	0.78
$112_{(0\text{-}775)}$	0.35
$156_{(0\text{-}760)}$	0.09
$471_{(0\text{-}500)}$	0.34
$\text{Native}_{(0\text{-}100}$	0.79

[1] The P values are for comparison of averages that result from post-processing either a) the first 150 ps every 10^{th} ps or b) the entire initial conformational family every 5^{th} ps.

There are two ways to increase the efficiency of sampling. First, replacing the inclusion of explicit waters during the dynamics simulation with a continuum solvent model such as the Generalized Born or the Analytical Continuum Electrostatic potential[29] should allow many more structures to be examined with the same computational expense. Secondly, one can use Locally Enhanced Sampling (LES)[30] in the molecular dynamics trajectory, which we have found can drive the structure to more native like values more quickly[31].

CONCLUSIONS:

As the genome projects continue to unravel novel gene sequences, successful protein structure prediction has more potential application now than ever before. If

enough atomic detail can be reliably predicted, in particular at the active and allosteric sites, better understanding of function can be achieved without the time consuming process of experimentally determining the structure. As CASP III has shown, however, the structure prediction community must still make significant advances before this goal can be realized, especially on sequences that have low sequence identity and on "*ab initio* targets", those with no structural relatives in the PDB. The hierarchical method presented here, to combine an *ab initio* method like Rosetta with molecular dynamics and MM-PBSA, seems to be promising for enabling more accurate protein structure predictions because the final stage is capable of both accurately ranking models and further refining them. We suggest methods like this may allow for a significant advance in CASP IV compared to CASP III predictions and should ultimately be useful in helping to generate accurate structures from the myriad of new sequences stemming from the genome projects.

Beginning with the Rosetta algorithm and ending with all-atom molecular dynamics simulations, we take sequence information of two small proteins and find structures that lie only 1.4 Å C_α RMSD from the experimental structures. These geometrically best conformations are members of conformational families that have both 1) the lowest average C_α RMSD and 2) the most favorable average MM-PBSA free energy among all non-native states. The single energy component that relates best to both RMSD and total free energy is the van der Waals term, which is the only term consistently more favorable in the native than in all other states. While it has been suggested that electrostatics are important in separating misfolded decoys from native structures, the present work that includes highly native-like decoys is consistent with our previous study[11] on protein stability in suggesting that electrostatics have a poor

correlation with the MM-PBSA free energy, which itself rank correlates well with the C_α RMSD.

While we show in this work that molecular dynamics can sometimes within hundreds of picoseconds lead to structural refinement of some model predictions of small proteins, future work is required to show how general this result is. Although we believe that molecular dynamics will generally guide proteins to lower free energies, simulations for limited amount of time will not always be capable of overcoming barriers, resulting in refinement of only some structures, as we found with HP-36 and S15. If longer simulations lead to ever decreasing free energies, as we suggest, then the more extended the simulation, the greater the probability is of refining low resolution structure predictions. As computers become ever more powerful, allowing one to run longer simulations, standard molecular dynamics as well as a number of other methods, such as Locally Enhanced Sampling[30] and self-guided molecular dynamics[32], can be used to more readily find new structures and MM-PBSA will help in evaluating if they are lower in energy.

REFERENCES:

1. Orengo, C.A.; Bray, J.E.; Hubbard, T. ; LoConte L.; Sillitoe I. *Proteins* **1999** Suppl 3, 149-179.

2. Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, 98, 10089-10092.

3. Fox, T.; Kollman, P.A. *Proteins* **1996**, 25, 315-334.

4. Cheatham, T.E., III; Kollman, P.A. *J. Mol. Bio.* **1996**, 259, 434-444.

5. Duan, Y.; Kollman, P.A. *Science* **1998**, 282, 740-744.

6. Alonso, D.O.V.; Daggett, V., *Protein Sci.* **1998**, 7, 860-874.

7. Vorobjev, Y.N.; Almagro, J.C.; Hermans, J. *Proteins* **1998**, 32, 399-413.

8. Holm, L.; Sander, C. *J. Mol. Biol.* **1992**, 225, 93-105.

9. Lazardis, T.; Karplus, M. *J. Mol. Bio.* **1999**, 288, 477-487.

10. Park, B.H.; Levitt, M. *J. Mol. Bio.* **1996**, 258, 367-392.

11. Lee, M.R.; Duan, Y.; Kollman, P.A. *Proteins* **2000**, 309-316.

12. Bower, M.J.; Cohen, F.E.; Dunbrack Jr., R. L. *J. Mol. Biol.* **1997**, 267, 1268-82.

13. 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. Am. Chem. Soc.* **1995**, *117*, 5179-5197.

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

15. Ryckaert, J.P.; Ciccotti, G.; Berendsen, H.J.C. *J. Comp. Phys.* **1977**, 23, 327-341.

16. Case, D. A.; Pearlman, D. A.; Caldwell, J. A.; 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. *AMBER 5.0* **1997**, *University of California, San Francisco.*

17. Berendsen, H.J.C.; Postma, J.P.M.; van Gunsteren, W.F.; DiNola, A.; Haak, J.R. *J. Chem. Phys.* **1984**, 81, 3684-3690.
18. Laskowski, R.A.; MacArthur, M.W.; Moss, D.S.; Thornton, J.M. *J. Appl. Cryst.*, **26** 1993, 283-291.
19. Ferrin, T.E., Huang, C.C.; Jarvis, L.E.; Langridge, R. *J. Mol. Graph.* **1988**, 6, 13-27.
20. Srinivasan, J.; Cheatham 3rd, T.E.; Cieplak P.; Kollman P.A.; Case, D.A. *J. Am. Chem. Soc.* **1998**, 120, 9401-9409.
21. Sanner, M.F.; Olson A.J.; Spehner, J.C. *Biopolymers* **1996**, 38, 305-20.
22. Gilson, M.K.; Honig, B. *Proteins* **1998**, 4, 7-18.
23. Sitkoff, D.; Sharp, K.A.; Honig, B. *J. Phys. Chem.* **1994**, 98, 1978-1988.
24. Simons, K.T.; Bonneau, R.; Ruczinksi, I.; Baker, D. *Proteins* **1999**, Suppl 3, 171-176.
25. Panchenko, A.; Marchler-Bauer, A.; Bryant, S.H. *Proteins* **1999**, Suppl 3, 133-140.
26. Lee, J.; Liwo, A.; Ripoll, D.R.; Pillardy, J.; Scheraga, H.A. *Proteins* **1999**, Suppl 3, 204-208.
27. McKnight, C.J.; Matsudaira, P.T. ; Kim, P.S. *Nat. Struct. Biol.* **1997**, 4, 180-184.
28. Clemons, W.M.; Davies, C.; White, S.W.; Ramakrishnan, V. **1998**, 6, 429-438.
29. Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, 100, 1578-1599.
30. Simmerling, C.; Elber, R. *J. Am. Chem. Soc.* **1994**, 116, 2534-2547.
31. Simmerling, CL; Lee, M.R.; Ortiz, A.O.; Kolinski, A.; Kollman, P.A. *J. Am. Chem. Soc.* **2000**, 122, 8392-8402.
32. Wu, X.W.; Wang, S.M. *J. Phys. Chem. B* **1998**, 102, 7238-7250.

Chapter 4

Free energy calculations highlight differences in accuracy between X-ray and NMR structures and add value to protein structure prediction

Matthew R. Lee

Peter A. Kollman*

Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA 94143-0446, United States of America

Submitted to Structure

Background: While X-ray crystallography structures of proteins are considerably more reliable than those from NMR spectroscopy, it has been difficult to assess the inherent accuracy of NMR structures, particularly the side chains.

Results: For 15 small single-domain proteins, we use a molecular mechanics/dynamics based free energy approach to investigate native, decoy and fully extended alpha conformations. Decoys are all less energetically favorable than native in 9 of the 10 X-ray structures and in none of the 5 NMR structures, but short 150 ps molecular dynamics simulations on the experimental structures cause them to have the lowest predicted free energy in all 15 proteins. In addition, a strong correlation exists ($r^2 = 0.86$) between the predicted free energy of unfolding, from native to fully extended, and the number of residues.

Conclusions: This work suggests that the approximate treatment of solvent used in solving NMR structures can lead NMR model conformations to be less reliable than crystal structures. This conclusion was reached because of the considerably higher calculated free energies and the extent of structural deviation during aqueous dynamics simulations of NMR models compared to those determined by X-ray crystallography. Also, the strong correlation found between protein length and predicted free energy of unfolding in this work suggests, for the first time, that a free energy function can allow for identification of the native state based on calculations on an extended state and in the absence of an experimental structure.

INTRODUCTION:

While methods for experimental determination of protein structure have had an enormous impact on the study of molecular action, protein design, and interpretation of chemical, kinetic or thermodynamic experiments, they are often quite challenging. Elucidation of a protein structure by X-ray crystallography demands a supersaturated concentration, which can usually only be achieved upon addition of agents that compete with the protein for water. These foreign agents and packing effects of crystallization itself can induce structural defects; while this artifactual information is reported with the structure in known instances, it is not possible to realize all of the errors caused by these model-specific systematic limitations. Another pitfall of crystallography occurs on segments having very low or non-existent electron densities, which presumably contain highly disordered atoms that are in motion and thus difficult to detect in the time scale of crystallography. Additionally, in some instances of low to medium resolution structures, oxygen atoms cannot be distinguished from nitrogen atoms. While these kinds of gross inaccuracies are relatively uncommon, other smaller deviations almost certainly exist in all X-ray structures, due to differences between the crystal environment, which is only 50% aqueous by volume, and the natural surroundings; this fundamental difference between crystal and native structures, as well as the non-static nature of proteins, creates an average atomic uncertainty of around 0.5 Å in structures, with the best data.

In comparison, protein structures solved by nuclear magnetic resonance (NMR) are completely solvated, free of the constraints of a crystal lattice, allowing for better description of the inherent flexibility, in surroundings much closer to what the native protein actual experiences. However, despite the more realistic environment that NMR

structures present, they are inherently less reliable than X-ray data because there is much less experimental data per atom than in crystallography. Differences among the various models of an NMR ensemble are usually much greater than 1 Å, often 2 Å. Determination of a protein structure using NMR involves a refinement process, usually starting with a randomly generated conformation that satisfies some local distance constraints, and proceeding with a sampling protocol that attempts to satisfy as many NOEs as possible, until a point is inevitably reached, where the structure is incapable of being improved further. While more NOEs generally allow for more accurate structures, shortcomings of the refinement stage are what preclude greater precision in the method. Likely the most severe approximation usually made during the refinement stage of NMR structure determination is an inaccurate representation of the solvent, which can prevent finding a better solution with lower positional uncertainty, even if the refinement stage were capable of exploring every possible conformation, due to systematic errors in the energy potential. Generating tens of structures with low average RMSDs compared to the mean structure does not necessarily imply accuracy, only that there is less uncertainty of the ensemble not having satisfied the NOEs on the energy surface used to describe the biomolecule. In the vast majority of NMR structures, inclusion of solvent effects is done by using a distance dependent dielectric constant in the Coulomb energy, and is thus not very accurate.

As a step towards understanding some of the qualitative differences between NMR and X-ray structures, we investigated the Molecular Mechanics-Poisson Boltzmann/Surface Area (MM-PBSA) free energies of X-ray and NMR structures, before and after short, computationally inexpensive molecular dynamics simulations, in

comparison to large sets of decoy conformations, on a total of 15 small, single-domain proteins. Sets of decoys for 8 proteins came from the "Rosetta All Atom Decoy Set" [1, 2] and 7 from the Park & Levitt 4-state reduced decoy set[3, 4]. While it is widely believed that the native structure lies at the global free energy minimum[5], which would satisfy the demands of thermodynamics, alpha-lytic protease has recently emerged as an exception, with the native state exhibiting a half life of unfolding on the order of 1 year[6]. Nevertheless, the native state should, in the majority of cases, obey macroscopic thermodynamics and lie at the global minimum, irrespective of whether the native structure has been solved by X-ray crystallography or NMR spectroscopy. At the very least, the native state should have a free energy substantially lower than unfolded and poorly folded conformations. This work suggests that NMR structures can benefit significantly from short aqueous molecular dynamics simulations and that free energy calculations can be used to identify the native state in the absence of an experimental structure.

RESULTS:

Decoys compared to crystal structures.

The 4-state reduced decoy set[4, Homepage, 1999 #34] consists of ~650 conformations for 7 proteins, with each conformer differing from the native conformation at 10 specific dihedral angles, that always lie in regions between or at the ends of secondary structure elements. Each dihedral may adopt only one of 4 possible discrete values, leading to an exhaustive enumeration of 1,048,576 (4^{10}) possible conformations per protein, of which ~650 were physically reasonable after removing those with steric

conflicts and unreasonably extended chains. Thus, the decoys for any given protein differ only in their tertiary structure but cover a wide range of native similarity. Three of these proteins are purely alpha and the other four are mixed alpha/beta, with the native counterpart being an X-ray structure in all seven cases.

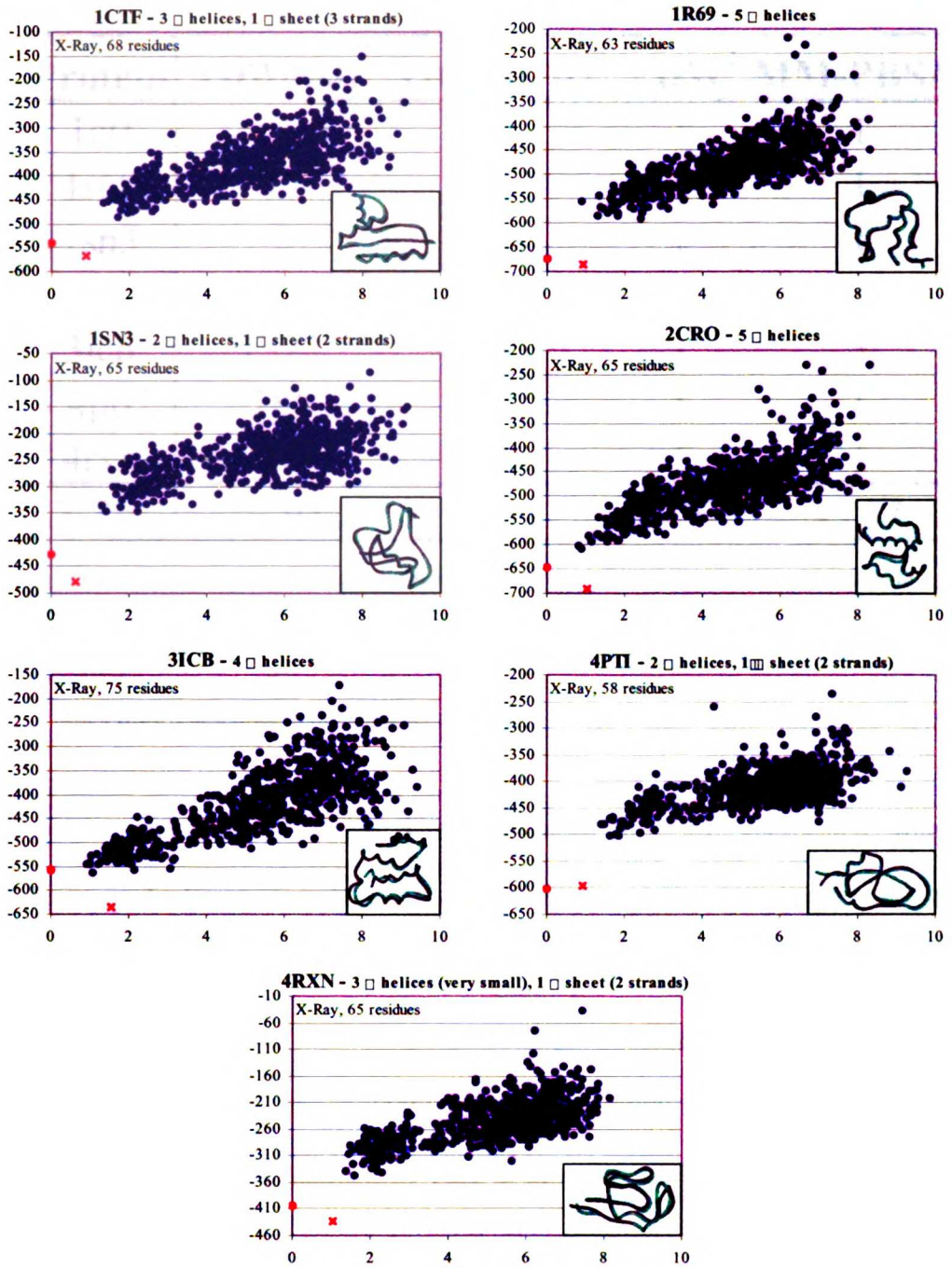


Figure 1. Single-point minimization MM-PBSA (y-axes) vs. Cα RMSD (x-axes) on crystal structures and on decoys containing native secondary structure (Park & Levitt 4-state reduced set). Each blue dot represents a single decoy. There were approximately 650 decoys for each of the 7 proteins. Red circles are minimized X-ray crystal structures. Red exes are crystal structures that have been minimized after 150 ps of molecular dynamics in explicit solvent. Figure inlays contain an overlap of the crystal structure before (gray tube) and after (cyan tube) dynamics.

Table 1. X-ray rank among Park & Levitt set

Protein	VDW(MJ)[1]	VDW(MJ)12[1]	VDW[2]	MM-PBSA[3]
1ctf	2	1	1	1
1r69	77	1	1	1
1sn3	2	1	1	1
2cro	160	1	1	1
3icb	1327	3	1	3
4pti	286	4	1	1
4rxn	49	3	1	1
<Z>	-3.95	-3.98	-3.92	-3.57

[1] VDW(MJ) is a distance dependent contact potential and VDW(MJ)12 is the same, but with a sharper repulsive term. The results of these energy functions are taken from Park et al. 1997.

[2] VDW is the attractive dispersion energy between non-bonded atoms in the MM-PBSA calculation

[3] MM-PBSA is described in the Methods

For each protein in the 4-state reduced set, we performed single-point minimization calculations (see Methods) on all the decoys, on the initial crystal structure, and on a 150 ps snapshot from an explicit solvent molecular dynamics simulation that started with the minimized crystal structure. Figure 1 shows the resulting MM-PBSA free energies as a function of Cα RMSD. This free energy function does better than any of the 18 scoring functions studied by Park et al. (1997)[7] and at least as well as other recently reported physically based functions that have successfully examined this same decoy set[8-10]. The crystal structures, shown as the gray tube diagrams on the pictorial inlays and represented by the red solid circles, have lower, more favorable free energies than all of the decoys in 6 out of the 7 proteins, with the crystal structure coming out 3rd

best among 654 decoys on 3icb, albeit even for this protein, the best structure with MM-PBSA had a Cα RMSD of only ~1 Å from the native.

The Z-score (see Methods) has been widely used to evaluate the goodness of a protein structure scoring function[11], but good Z-scores only imply that the native structure receives a much better score than the average score of all the conformers in the decoy set. Table 1 shows the X-ray rank results of two distance dependent contact potential energy functions from Park et al. (1997)[7] that were among the four best (in terms of average Z-scores of all the proteins in the 4-state reduced set) out of the 18 functions investigated, alongside the X-ray rank results from MM-PBSA and its van der Waals component alone (VDW). While the average Z-scores are comparable in each of the four, VDW(MJ) clearly does a relatively poor job in picking out the crystal structure as best. Our VDW correctly identifies all 7 crystal structures, MM-PBSA identifies 6 out of 7, VDW(MJ)12 predicts 4 out of 7 correctly, and VDW(MJ) does not predict any correctly. These X-ray rank results indicate that energy functions, which result in good Z-scores, are not necessarily good at correctly identifying the native fold.

Table 2. Assessing predictive value of energy functions

Protein	MM-PBSA		VDW	
	r_s [1]	Z^2	r_s [1]	Z^2
1ctf	0.77	-2.47	-0.18	-3.36
1r69	0.55	-3.88	-0.27	-5.01
1sn3	0.52	-4.57	-0.32	-3.97
2cro	0.66	-3.03	-0.03	-4.57
3icb	0.75	-1.86	0.13	-3.58
4pti	0.44	-5.21	-0.37	-3.27
4rxn	0.65	-4.00	-0.48	-3.66
AVG	0.62	-3.57	-0.22	-3.92

[1] r_s is the Spearman rank correlation coefficient between Cα RMSD and the energy; it is more meaningful than the standard Pearson product-moment correlation coefficient in non-parametric relationships that are not linearly related.

[2] Z score is the number of standard deviations separating the energy of the native conformation from the average energy of the entire set. (see Methods)

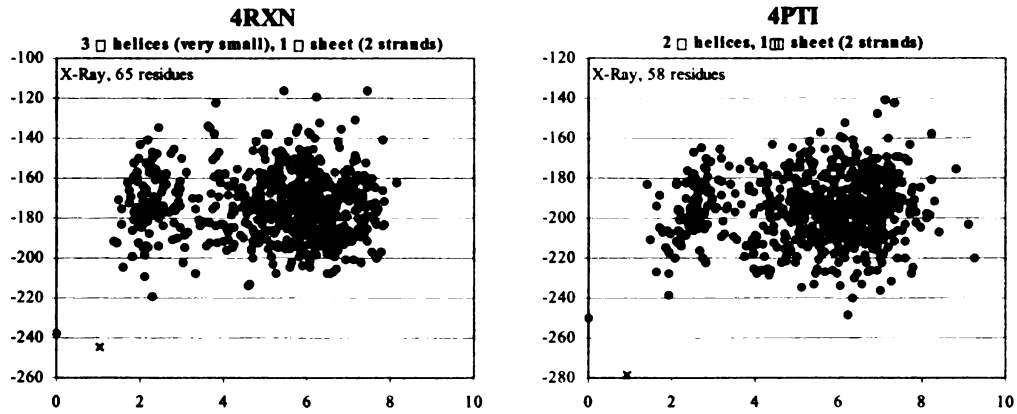


Figure 2. Single-point minimization VDW (y-axes) vs. Cα RMSD (x-axes) on crystal structures and on decoys containing native secondary structure (Park & Levitt 4-state reduced set). Only two representative proteins are shown, demonstrating the lack of a relationship between native similarity and van der Waals energies, despite identification of native fold from all decoys. Red circles are minimized X-ray crystal structures. Red exes are crystal structures that have been minimized after 150 ps of molecular dynamics in explicit solvent.

The Z-score also does not address strength of relationship between native similarity and the scoring function. Instead, correlation coefficients provide a far more direct criterion for establishing strength of association between two variables, and are thus more fitting for judging predictive value of a scoring function for structure prediction. For parametric samples, in which both variables are normally distributed on an interval scale, implying a linear relationship, the standard (Pearson product-moment) correlation coefficient (r) is most appropriate, but for non-linear relationships on an ordinal scale, in which one or both variables are not normally distributed, the Spearman rank correlation coefficient (r_s) is most appropriate. In a Boltzmann distribution, conformations are weighted exponentially, according to their free energies, $P(i) = \exp(-\Delta E_i/RT)$, where ΔE_i is the difference in free energies between two states, *i* and some reference, such as the native state. Because there is no reason to not believe that the vast majority of proteins obey microscopic thermodynamics, one should expect that protein conformations roughly populate in a Boltzmann distribution, rather than a Gaussian distribution. Thus, to evaluate the strength of association between any variable and a free energy like MM-PBSA, r_s is more appropriate than r. For predictive value in protein structure prediction, a strong correlation with native similarity is highly desired, so we evaluated the Spearman rank correlation between MM-PBSA and Cα RMSD in Table 2, which shows a reasonably good correlation, slightly better than that reported by Gatchell et al.[10] and Dominy & Brooks[9]. While Table 1 indicates the lack of association between good Z-scores and the ability to correctly identify the native fold, Table 2 shows that good Z-scores do not imply good predictive value. Although the VDW potential did

slightly better than MM-PBSA in terms of Z-score, it has no meaningful relationship with Cα RMSD, as illustrated in Figure 2 on two representative proteins and quantified in Table 2. 8 Å conformations have the same VDW energy as 2 Å structures

Table 3. Pearson product-moment correlation coefficient between Cα RMSD and MM-PBSA

Protein	Cα RMSD bin		
	0 - 2.5	2.5 - 5.0	> 5.0
1ctf	0.62	0.38	0.36
1r69	0.63	0.45	0.37
1sn3	0.66	0.38	0.23
2cro	0.72	0.37	0.41
3icb	0.48	0.55	0.38
4pti	0.84	0.38	0.32
4rxn	0.54	0.47	0.34
AVG	0.64	0.43	0.34

The Cα RMSD bins contain every structure in the decoy set within the specified values. The linear relationship between Cα RMSD and MM-PBSA is strongest in the bin of close structures.

While some have suggested that there is no physical requirement for a relationship between free energy and native similarity[12], Dill & Chan introduced the widely accepted view of a funnel-shaped free energy landscape[13] to describe proteins, where the native state has the lowest free energy and the more distant the native similarity, the less favorable the free energy. If the free energy landscape is indeed globally convex, the relationship between native similarity and free energy should be approximately linear for only those conformations immediately surrounding the native state, and the further structures lie from the native state, the less linear the relationship is,

until finally on the level surface of the funnel, where native similarity is very low, there is no relationship at all. We investigated this by separating the 4-state reduced decoy sets into three bins of native similarity: close (< 2.5 Å), medium (2.5 - 5.0 Å) and distant (> 5 Å). For each similarity bin, we evaluated the Pearson product-moment correlation coefficient, which again is the strength of relationship between two variables that are *linearly* related. As expected, the close structures showed the greatest degree of linear association between Cα RMSD and MM-PBSA ($r = 0.64$), with the distant structures showing only a slight tendency, and medium structures falling in between. These results suggest, together with the rank correlation results in Table 2, that the free energy and native similarity are related on an ordinal scale, with that relationship becoming increasingly linear as native similarity increases.

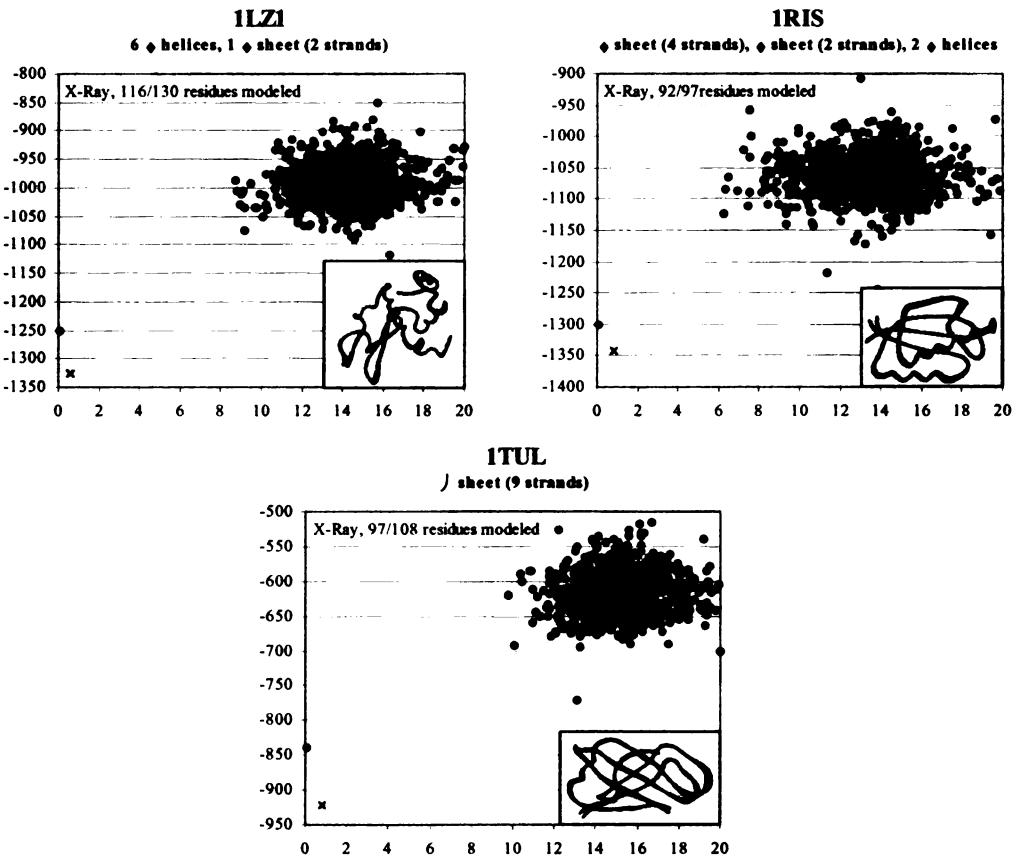


Figure 3. Single-point minimization MM-PBSA (y-axes) vs. Cα RMSD (x-axes) on crystal structures and on decoys with no secondary structural restrictions (Rosetta *ab initio* all atom decoy set). Note these decoys were generated in a structure prediction effort without information of the native structure. Red circles are minimized X-ray crystal structures. Red exes are crystal structures that have been minimized after 150 ps of molecular dynamics in explicit solvent. Figure inlays contain an overlap of the crystal structure before (gray tube) and after (cyan tube) dynamics.

The Rosetta "All Atom Decoy Set" consists of 1000 decoy conformations for each protein, with each conformer generated by the Baker group in the same manner as that used for the community-wide Critical Assessment of Structure Prediction experiment from a 1998 (CASP III)[2]. Three of the eight that we investigated from this decoy set had crystal structures. In contrast to the 4-state reduced set, the Rosetta set usually does not populate the low Cα RMSD regions very well, which should lead to a limited relationship at best between functions with good predictive value and Cα RMSD among

these decoys, because as discussed earlier, the linear correlation falls off beyond the 5 Å mark (Table 3). Furthermore, because the structures in this set differ from one another immensely more than they do in the 4-state reduced set where 10 dihedral angles are the only degrees of freedom, the noise of the energy should much greater. Thus, we cannot hope to distinguish 8 Å from 15 Å structures, even with a free energy function that were entirely precise. All that can be hoped for in this Rosetta decoy set is the ability to distinguish native from everything else, which MM-PBSA effectively accomplishes (Figure 3).

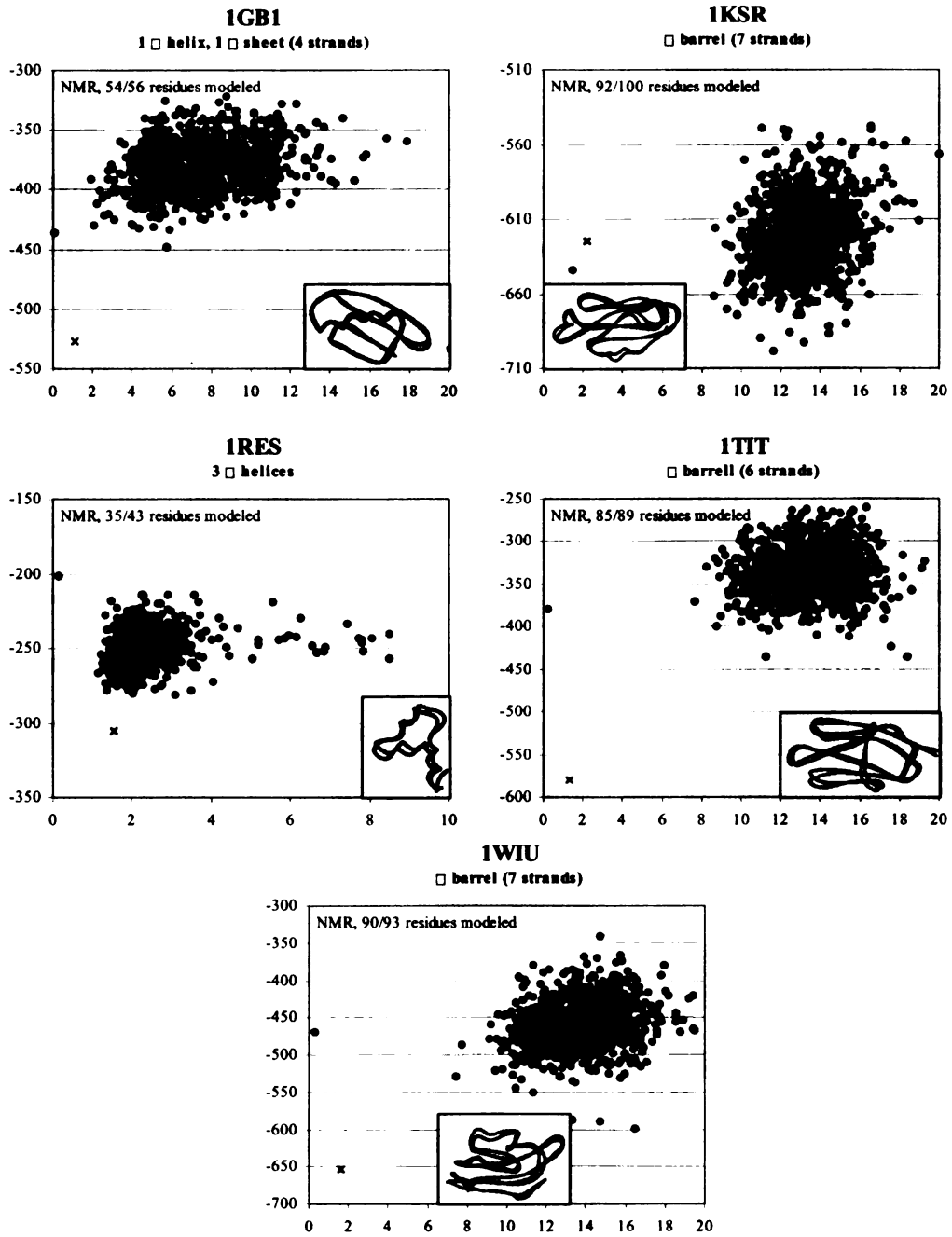


Figure 4. Single-point minimization MM-PBSA (y-axes) vs. Cα RMSD (x-axes) on NMR structures and on decoys with no secondary structural restrictions (Rosetta *ab initio* all atom decoy set). Red exes are NMR structures that have been minimized after 150 ps of molecular dynamics in explicit solvent. Figure inlays contain an overlap of the crystal structure before (gray tube) and after (cyan tube) dynamics.

Decoys compared to NMR structures.

The five other proteins that we investigated in the Rosetta decoy set had NMR structures. What clearly distinguishes these 5, shown in Figure 4, from the 10 sets with crystal structures, is that the minimized NMR structures (red closed circles) do not have the lowest free energies in any of the proteins. For 1gb1, there is a 6 Å decoy lower in free energy, a 16 Å one for 1ksr, a 9 Å one for 1res, an 18 Å one for 1tit, and a 17 Å one for 1wiu. We believe that this arises presumably because of the unsophisticated refinement methods used to solve the NMR structures, as discussed above. While these reported NMR structures are undoubtedly in the correct global structural fold, these single-point minimization MM-PBSA results suggest that the NMR structures are nowhere near the bottom of the native energy basin, that minimization alone is insufficient to overcome the numerous bad contacts, bond lengths, valence bond and dihedral angles, which additively can lead to many tens to hundreds of kcal/mol penalties, with only minor perturbations to the correct native topology and structure, in terms of RMSD.

Effect of molecular dynamics.

Figures 1, 3 and 4 also show the effect of molecular dynamics on the native structure compared to single-point minimization MM-PBSA calculations. Experimental structures that have undergone 150 ps of molecular dynamics, followed by minimization, are shown as the cyan tube diagrams on the inlays and represented by the red exes. These native 150 ps snapshots have the best free energies in all 10 of the X-ray examples (7 from the 4-state reduced model and 3 from Rosetta), including 3icb where the minimized

X-ray structure ranked 3rd best, and the best free energies in four of the five NMR examples where none of the minimized NMR structures ranked best. In the 150 ps snapshot of 3icb, the region that deviated most from the crystal was one of two Ca^{2+} binding loops in the protein. While the 3icb deposited pdb structure contains hetero-atom records for two Ca^{2+} ions, the structures in the decoy set do not, so to be consistent with the decoy set and have a level playing field, we removed these divalent cations from the crystal structure prior to evaluating the single-point minimization MM-PBSA, creating a locally unfavorable hole in the system, which was filled in the 150 ps snapshot. Because hetero-atoms are not included in structure predictions, it is appropriate to remove them from the experimental structures as well, when trying to evaluate a scoring function's ability to pick out the native conformation. This leaves crystal structures with locally unfavorable regions, where the missing hetero-atoms may have been involved in stabilizing the protein, creating an artifactual energy penalty of the native structure, that can be compensated for with a short 150 ps dynamics simulation.

Table 4. Free energy improvement of X-ray structures by molecular dynamics

protein	Δ MM-PBSA	Δ strain[1]	Δ VDW	Δ solv_NP[2]	Δ EEL_tot[3]	Cα RMSD	Resolution
1lz1	-73.4	-0.4	-12.1	1.1	-62.0	0.59	1.5
1ris	-40.3	6.8	-34.3	0.8	-44.6	0.79	2.0
1tul	-83.0	-10.4	-36.5	1.0	-37.2	0.84	2.2
1ctf	-25.2	-8.6	-15.2	0.7	-2.0	0.89	1.7
1r69	-9.8	0.4	-3.5	0.4	-7.1	0.93	2.0
1sn3	-50.8	-16.5	-25.8	0.5	-8.9	0.64	1.2
2cro	-43.1	-19.6	-16.9	0.5	-7.1	1.04	2.4
3icb	-79.8	-27.1	-0.9	1.3	-53.2	1.54	2.3
4pti	5.5	37.1	-28.6	0.3	-3.3	0.93	1.5
4rxn	-29.1	-17.2	-6.4	0.2	-5.7	0.93	1.2
AVG	-42.9	-5.5	-18.0	0.7	-23.1	0.97	1.8

The differences are between single point calculations on the initial structure as well as on the 150 ps snapshot of the dynamics simulation. They are not as precise as ensemble average calculations, which are not possible as the minimum requirement for a statistically meaningful ensemble average is 15 snapshots over 150 ps.

[1] The internal strain energy results from deviations away from reference values in bond length, angle and dihedral terms.

[2] The non-polar solvation energy accounts for the cost of solvating a discharged solute.

[3] The total electrostatics energy is the sum of intra-solute Coulombic energies and solute-solvent electrostatic energies.

Table 4 numerically summarizes the single-point minimization data of the X-ray structures, before and after molecular dynamics, for MM-PBSA and each of its four components. 9 out of the 10 crystal structures had a more favorable free energy after the dynamics simulations, with the 150 ps snapshots having moved 0.97 Å on average from their initial conformation and being 43 kcal/mol on average more favorable. Only the 4pti crystal structure, which was already 100 kcal/mol more favorable than the best decoy, and incidentally whose crystal structure did not contain any hetero-atoms other than water molecules, did not experience an improvement. These substantial improvements in free energy and ~1.0 Å movement away from the crystal structure

results from 1) the absence of hetero-atoms included in the X-ray crystal, 2) differences between our more representative aqueous solution and the crystal surroundings, and perhaps 3) inaccuracies in the force field.

Table 5. Free energy improvement of NMR structures by molecular dynamics

protein	Δ MM-PBSA	Δ strain[1]	Δ VDW	Δ solv_NP[2]	Δ EEL_tot[3]	Cα RMSD
1gb1	-91.2	-7.9	-38.2	0.4	-45.6	1.12
1ksr	19.7	44.5	106.4	3.5	-134.7	2.18
1res	-103.8	-26.2	-56.5	-0.1	-21.0	1.56
1tit	-200.8	-45.7	-76.8	-0.2	-78.1	1.32
1wiu	-184.7	-69.3	-68.7	2.0	-48.7	1.65
AVG	-112.1	-20.9	-26.8	1.1	-65.6	1.57

See footnotes to Table 4.

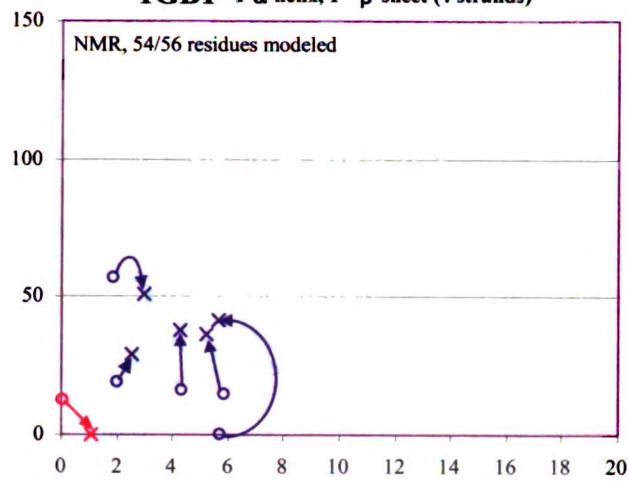
Table 5 shows the same results as Table 4, but for the 5 NMR examples. After 150 ps, the NMR structures, none of which had the most favorable single-point minimization MM-PBSA, moved 60% further (<Cα RMSD> = 1.57 Å), on average, from their starting structures than did the crystal structures. They also experienced a much greater decrease in free energy, 112 kcal/mol on average, with only the 150 ps 1ksr NMR snapshot not showing an improvement over the initial NMR structure, and thus not becoming more favorable than its Rosetta decoy set.

The incorrect ranking of the 1ksr conformations stems not from flaws in MM-PBSA, but rather from using it to compare single-point calculations on minimized structures. Although it is a rapid and thus desirable calculation, there are at least three reasons why this single-point minimization MM-PBSA method cannot be expected to

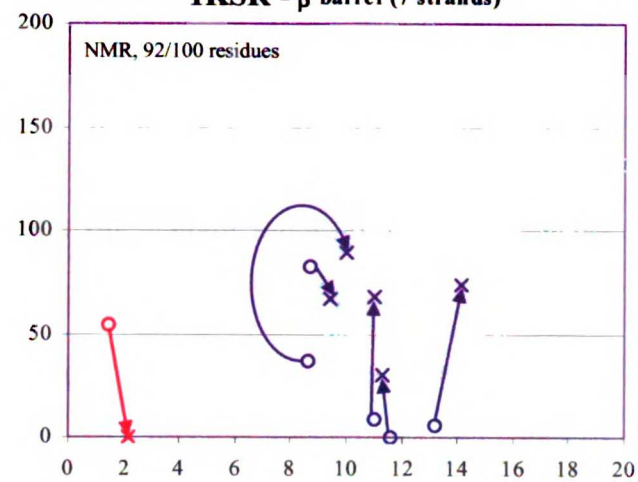
succeed in all cases. First, minimizations effectively remove temperature, thereby altering the balance between enthalpy and entropy and thus changing the free energy surface as well. Second, minimizations have difficulty escaping local minima, which can over penalize conformations experiencing locally unfavorable energies like those solved by NMR. Third, the MM-PBSA values fluctuate considerably, with standard deviations on the order of 20-30 kcal/mol. Therefore, to obtain a more accurate MM-PBSA value, a statistically sufficient ensemble of molecular dynamics trajectories should be generated, with comparison of the resulting ensemble-averages. An ensemble, which samples conformational space at 300 K, does not overly weight enthalpic contributions, can much more readily alleviate locally unfavorable interactions to escape local minima, and provides enough data to generate meaningful ensemble-averages that can be compared by using *t* tests to evaluate the significance of differences.

Figure 5. (shown on following page) Effect of using ensemble-averages on MM-PBSA (y-axes) vs. Cα RMSD (x-axes). Circles show the same single-point minimization MM-PBSA results as Figure 4, but on a relative scale. Exes are the ensemble-averages from 150 ps of molecular dynamics simulation, starting from the conformation represented by the open circle that the arrow originates from.

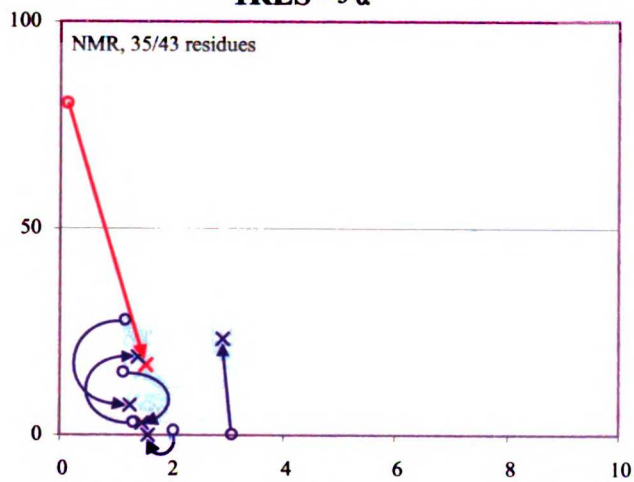
1GB1- 1 α helix, 1 β sheet (4 strands)



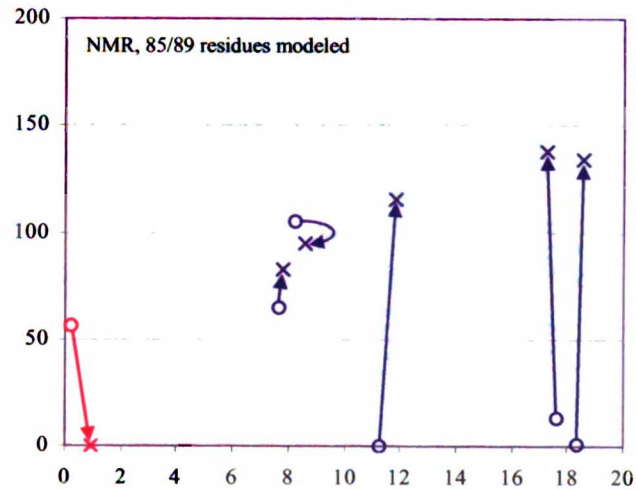
1KSR - β barrel (7 strands)



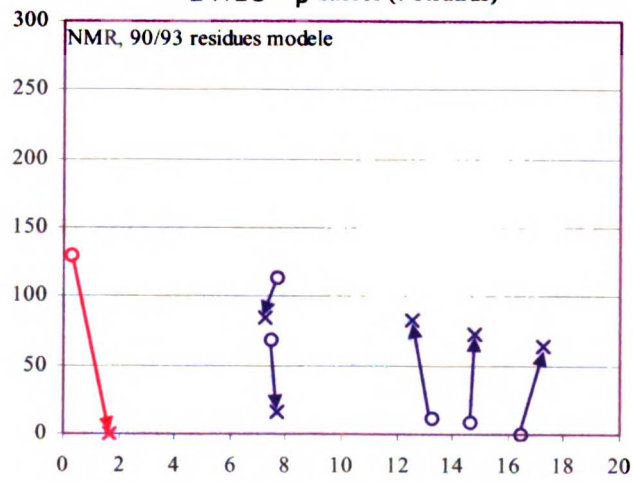
1RES - 3 α



1TIT - β barrell (6 strands)



1WIU - β barrel (7 strands)



Thus, for 1ksr, as well as for each of the other NMR examples, we generated 6 ensembles: one from the NMR structure, two from the Rosetta decoys with the lowest RMSD, and three from the Rosetta decoys with the lowest single-point minimization MM-PBSAs. The open circles in Figure 5 show the single-point minimization MM-PBSA of all the decoys and the initial NMR structure, relative to the most favorable conformation, with the red ones being the NMR structure and the dark blue ones being the 5 decoys selected for molecular dynamics. (Note that the energies are relative in Figure 5 and absolute in Figure 4.) The exes in Figure 5 are the resulting ensemble-average MMPBSA values, relative to the best. The arrows map initial snapshot to its corresponding ensemble average. Upon comparing the ensemble-averages, we find that the native state now has the most favorable MM-PBSA free energy in every protein, except 1res, where the lowest free energy has only a 1.5 Å Cα RMSD from the NMR determined structure. It is also particularly noteworthy that this approach shows the native structure to be most stable for 1ksr, where MD followed by minimization (red ex if Figure 4) did not lead to the NMR structure being most stable. To be sure, the MD average structure analysis was only done on 6 candidates, rather than the 1000 in the entire decoy set, albeit we picked the lowest energy and lowest RMSD ones from the original minimization analysis as our decoys.

Size dependence of the free energy of unfolding.

As the whole allure of protein structure prediction rests in its potential to determine structures faster than experimental methods, an often overlooked requirement is that the predictor have an absolute means of knowing when the native state has been

found. A scoring function that has a high correlation between score and native similarity, when applied to a database of structure predictions, can only identify the lowest scoring conformation, which it predicts to have the most native similarity, but it cannot determine if this best scoring structure is native or not. Consequently, we investigated the possibility of using an extended state as the reference, rather than the native state, for our average-ensemble MM-PBSA free energy, as Chiche et al.[14] did using the Eisenberg and McLachlan SFE solvation energy[15]. However, we used an all α helical structure for technical reasons (see Methods) as the extended reference, and we also added hydrogen atoms to sulfur atoms of cysteine residues involved in disulfide bonds of the native structure. We find, as shown in Figure 6, that among the 15 proteins studied in this work, a strong correlation exists ($r^2 = 0.86$) between the size of a protein, in terms of the number of residues, and its Δ(MM-PBSA) average-ensemble free energy, in going from native state to fully extended state that is entirely alpha. Because the absolute average-ensemble MM-PBSA of a fully extended helical state for any protein can always be simulated, this correlation implies that one can come up with an expected absolute average-ensemble MM-PBSA value for the native state, based only on the number of residues, thereby providing an absolute check for identifying the native state.

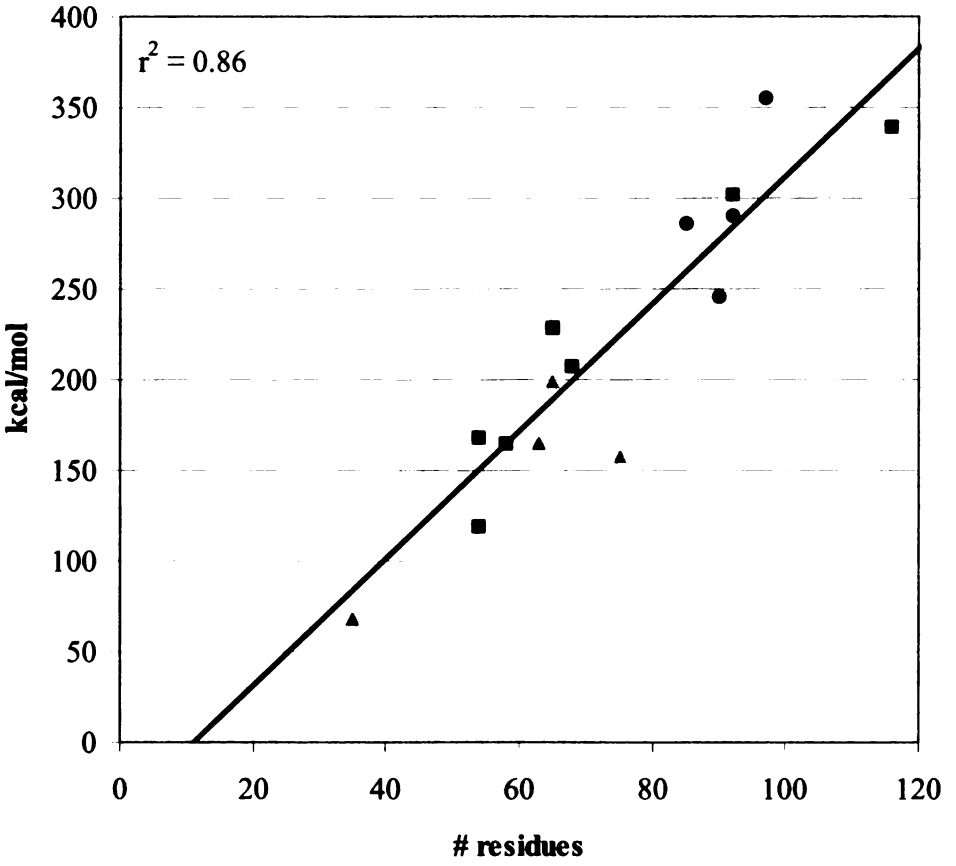


Figure 6. Size dependence of $\Delta(\text{MM-PBSA})_{\alpha\text{-nat}}$. Data on the 15 proteins in this work, 4 alpha proteins (yellow triangles), 4 beta proteins (blue circles) and 7 mixed proteins (green squares), shows a strong relationship between the number of residues and the free energy of unfolding. The X-intercept is 10 residues, suggesting that the most favorable conformation for peptides of this size may be all α-helical.

That $\Delta(\text{MM-PBSA})_{\alpha\text{-nat}}$ relates linearly to the size of a protein is not a coincidence, and can be simply rationalized. The MM-PBSA free energy does not account for conformational entropy (S_{conf}); it predicts the intrinsic free energy of a particular snapshot, without including the effect of other degenerate structures residing at

the same energy level, which would effectively lower the relative free energy of this ensemble of near degenerate structures by increasing S_{conf}. In other words, $\Delta G_u = \Delta(\text{MM-PBSA})_{u\text{-}nat} - T \cdot S_{conf,u}$. If we use the expression for Boltzmann's law, $S_{conf,i} = R \cdot \ln(\Omega_i)$, where Ω_i is the number of degenerate structures at a given energy level i and assume for the unfolded state that $\Omega_i = y^n$, with y representing the average number of conformations per residue and n the number of residues, $S_{conf,u}$ can be assumed to be directly proportional to n, $S_{conf,u} = n \cdot R \cdot \ln(y)$. In view of the empirical fact that ΔG_u remains relatively insensitive to protein size, $\Delta(\text{MM-PBSA})_{u\text{-}nat}$ should be roughly equal to $T \cdot S_{conf,u}$ and thus also proportional to the number of residues. If one finally assumes that the MM-PBSA of the α helical state is representative of the MM-PBSA of other individual members of the unfolded state, $\Delta(\text{MM-PBSA})_{\alpha\text{-}nat} \approx \Delta(\text{MM-PBSA})_{u\text{-}nat}$.

A final interesting observation from Figure 6 is that the regression line has an X-intercept of 10 residues. This suggests that peptides of 10 amino acids or fewer prefer the α-helical conformation over any other. For peptides so small, hydrophobic clusters, which are likely critical for compact conformations, would be marginally stable at best. Furthermore, a collapsed structure would probably have less favorable van der Waals interactions than the repeating (i to $i + 4$) attractions found in an α helix. Another possibility for interpreting the far left end of Figure 6 is that the linear relationship adopts a much smaller slope for very small peptides.

DISCUSSION:

High resolution X-ray crystallography structures have an average atomic uncertainty on the order of 0.5 Å. Interestingly, 150 ps snapshots from molecular

dynamics simulations on crystal structures had lower single-point MM-PBSA free energies than the initial crystal structures in 9 out of the 10 cases, with only 4pti not benefiting energetically from molecular dynamics. While the initial structures were already more favorable than entire ensembles of decoys in all but 3icb, the 150 ps snapshots had a better single-point MMPBSA values in all 10 cases. In addition to having more favorable predicted free energies, 150 ps snapshots moved, on average, less than 1 Å from their initial coordinates; these limited coordinate shifts may have been due to our removal of hetero-atoms, due to adverse effects caused by packing artifacts or other defects in the crystal structure, or due to the intrinsic tendency of proteins to breathe.

The average atomic uncertainty of NMR structures is difficult to quantify. While a popular idea is to evaluate the average deviation from a central average structure on an "NMR ensemble", this does not account for systematic uncertainties caused by inaccuracies of the energy surfaces being used to refine the structure, or for the inability to sample sufficiently during the refinement. 150 ps snapshots from simulations on NMR structures showed much greater improvements in free energy and much greater movement, over 1.5 Å, compared to their initial structures, than in the X-ray examples. All 5 of the NMR models were less favorable than a significant number of decoy structures, and 4 of the 5 150 ps snapshots had a markedly improved free energy, to levels significantly below the best scoring decoys. However, the more accurate method for evaluating free energies, ensemble-average MM-PBSA, favors the native state in all 5 of the NMR examples. The larger structural shifts and drops in predicted free energies for NMR, than for X-ray structures, is consistent with the greater uncertainty in NMR

structures. Moreover, this work suggests that short explicit solvent molecular dynamics simulations can correct, at least in part, for the errors introduced during the standard in vacuo refinement protocol of NMR structure solution.

MM-PBSA provides meaningful, physically-based insight into relative free energies of proteins[16, 17], as do a few other energy functions[8-10], but an important finding of this work is that it presents the first look at using this kind of free energy to determine whether a protein structure prediction is of native quality, *sans* the actual experimental structure. We find that a strong correlation exists between the size of a protein and its MM-PBSA free energy of unfolding, from native state to an all alpha-helical state ($r^2 = 0.86$).

Biological Implications:

A critical step for making use of the now abundant genomic information is having accurate three dimensional protein structures, with X-ray crystallography and NMR spectroscopy currently being the two methods that can be used to determine these structures. However, although crystal structures are well known to be more accurate than NMR models, it has been challenging to asses the inaccuracies in the NMR models that are obtained through refinement of NOE constraints. The present work suggests that short room temperature molecular dynamics simulations with accurate treatment of solvent effects and long range electrostatics, which are dramatically more computationally accessible than they were only 5 years ago, are important for escaping locally trapped, energetically unfavorable geometries that are inherent in NMR models.

While protein structure prediction methods are still not at the point where they can be used in place of experimental methods, if they are ever to reach that lofty goal, they must be capable of more than just generating the native structure, for these methods always generate a multitude of models. Structure predictors must also be able to 1) identify which among the scores of generated conformations are most native like and 2) know if the best structures are actually in the native state or not. Molecular mechanics free energy functions like MM-PBSA, that include implicit solvation free energies and are physically grounded, perform better than statistical and empirical functions at ranking structure predictions. We also show in this work that MM-PBSA can be used, together with an alpha helix extended state, to accurately predict when a protein conformation is in the native state without any *a priori* native state information, such as tertiary contacts or secondary structure. This method is based only on the protein length and the difference in free energy between a given conformation and the alpha extended conformation.

EXPERIMENTAL PROCEDURES:

The AMBER 5 suite of programs[18] was used for all molecular mechanics simulations. The Cornell et al. all-atom force field[19] (parm94) was used for simulations and the parm96 force field[20], which differs only in the φ, ψ torsional potentials of the peptide unit, was used in the MM-PBSA free energy analysis.

Minimization

We used a single minimization protocol on all protein conformations: steepest descent for the first 10 cycles, followed by conjugate gradient until the RMS of the Cartesian elements of the potential energy gradient fell below 0.4 kcal/mol·Å. Minimizations were carried out in the gas phase, using a distance dependent dielectric constant of $4r_{ij}$ and a cutoff for all non-bonded interactions of 25 Å.

Molecular dynamics

We ran all production-phase molecular dynamics simulations with a 2.0 fs time step under the isothermal-isobaric ensemble (300 K and 1atm) with explicit solvent, using the TIP3P model[21] for water, periodic boundary conditions, the particle mesh Ewald (PME) method[22] for electrostatics, a 10 A cutoff for Lennard-Jones interactions, and the use of SHAKE[23] for restricting motion of all covalent bonds involving hydrogen atoms. Water molecules were added around the proteins using a 10 Å buffer from the edge of the periodic box. The temperature and pressure were maintained by the Berendsen coupling algorithm using a τ coupling constants of 1.0. PME grid spacing was ~1.0 Å and was interpolated on a cubic B-spline, with the direct sum tolerance set to 10^{-5}. We removed the net center of velocity every 100 ps to correct for the small energy drainage, that results from the use of SHAKE, discontinuity in the potential energy near the Lennard Jones cutoff value, and constant pressure conditions.

For equilibration, we solvated the minimized structures, minimized the water molecules alone until the RMSD was < 0.1 kcal/mol·Å and then slowly heated, while

allowing the water to move unrestrained for 25 ps (with a 1.0 fs time step) in order to fill any vacuum pockets.

MM-PBSA

Coordinates from a trajectory were saved every 5 ps, and the MM-PBSA calculation evaluated on each of them. The MM-PBSA free energy of each snapshot is approximated as the sum of two terms, using an interior dielectric constant of 4: the internal energy of the protein (E_{MM}) and a solvation free energy (ΔG_{solv}). E_{MM} is the sum of an internal strain energy (E_{int}), a van der Waals energy (VDW), and an intra-solute electrostatic energy (EEL). ΔG_{solv} consists of the cost of submerging a discharged solute in solvent (Δsolv_NP) and the subsequent cost of adding the charges back to the solute (Δsolv_eel). Δsolv_NP is approximated as being linearly related to the solvent accessible surface area (SASA): 5.42*SASA + 920 cal/mol. We adhered to the same Poisson-Boltzmann protocol as first described by Srinivasan et al.,[24] which used DelPhi II[25] and most of its standard default parameters, together with PARSE atomic radii and Cornell et al. charges, to calculate Δsolv_eel. The entropy of a given snapshot, which is mostly vibrational, can be calculated with normal mode analysis on a Newton-Raphson minimization. This, however, is the most time-intensive part of the MM-PBSA method on a per-snapshot bases. Given the results in our previous study[16], where we found this term to be indistinguishable among the native state, the folding intermediate, and the unfolded state of HP-36, we did not perform this calculation in the current study. For a more detailed discussion of the MM-PBSA method, see the review by Kollman et al.[26].

Single-point minimization and ensemble-average calculations

When comparing the experimental and 150 ps structures with all the structures in the decoy set, we took each individual structure, performed minimization and evaluated MM-PBSA, using only a single value for the reported MM-PBSA, which we refer to as single-point minimization values. For the ensemble-average values, we took the average of every 10^{th} ps over a 150 ps molecular dynamics simulation, as we previously showed that this protocol provides the least expensive, yet statistically sufficient protocol for evaluating an ensemble-average MM-PBSA[17].

NMR structures

When using the term "the NMR structure", we are referring to model 1 in each of the NMR ensembles. We used this as the representative for simulation purposes, as it is more physically realistic than an average structure. The RMSDs, however, are always calculated in reference to the average NMR structure, as it is most representative of the various geometries of the ensemble.

Fully extended conformations

In order to create a fully extended chain for our reference state, we selected all alpha-helical conformations, because they were computationally efficient and well behaved. The other alternative, an extended beta strand, experiences bends in the rod wherever a proline resides, preventing the extended state from being remaining linearly shaped, and leading to water box sizes that are immensely larger than those for the all alpha-helical conformations. Flat-well restraints on the backbone φ and ψ torsion angles

were used to keep the backbone in a helical conformation, with no energy penalty for $-180° < \varphi < -60°$ and $-60° < \psi < -30°$, a parabolic side extending ±20° with a 30 kcal/mol·rad^2 force constant, and linear sides, with slopes at the outer edge of the parabolas, extending beyond that.

Z-score

The Z-score of a given value among a sample, Z_i, expresses how many standard deviations value *i* is away from the average value of the sample. Negative Z-scores mean the value is less than the average. For example, in the 4-state reduced decoy set, a Z-score of -2.0 for a crystal structure would means that the crystal structure has an energy that is 2.0 standard deviations lower than the average, which for a perfectly Gaussian distribution would mean that the native is more favorable than 97.5% of all the decoys.

Acknowledgements

We gratefully acknowledge research support from NIH (GM-29072) and supercomputer time through the NSF-supported NCSA center as well as the Advanced Biomedical Computing Center at NCI-Frederick.

REFERENCES:

1. Baker Laboratory Homepage, University of Washington, Department of Biochemistry, http://depts.washington.edu/bakerpg/ (accessed August 2000)
2. Simons, K.T., Bonneau, R., Ruczinski, I. & Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins*, pp. 171-176.
3. Levitt Laboratory Homepage, Stanford University, Department of Structural Biology, http://dd.stanford.edu/ (accessed August 2000)
4. Park, B. & Levitt, M. (1996). Energy Functions That Discriminate X-Ray and near-Native Folds from Well-Constructed Decoys. *J. Mol. Biol.* **258**, pp. 367-392.
5. Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. *Science* **181**, pp. 223-230.
6. Derman, A.I. & Agard, D.A. (2000). Two energetically disparate folding pathways of alpha-lytic protease shave a single transition state. *Nat. Struct. Biol.* **7**, pp. 394-397.
7. Park, B.H., Huang, E.S. & Levitt, M. (1997). Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.* **266**, pp. 831-846.
8. Lazaridis, T. & Karplus, M. (2000). Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* **10**, pp. 139-145.
9. Dominy, B.N.B., C.L., III (in press). Identifying Native-Like Protein Structures Using Physics-Based Potentials. *J. Comp. Chem.*

10. Gatchell, D.W., Dennis, S. & Vajda, S. (2000). Discrimination of near-native protein structures from misfolded models by empirical free energy functions. *Proteins* **41**, pp. 518-534.

11. Bryant, S.H. & Altschul, S.F. (1995). Statistics of Sequence-Structure Threading. *Curr. Opin. Struct. Biol.* **5**, pp. 236-244.

12. Lazaridis, T. & Karplus, M. (1999). Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.* **288**, pp. 477-487.

13. Dill, K.A. & Chan, H.S. (1997). From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **4**, pp. 10-19.

14. Chiche, L., Gregoret, L.M., Cohen, F.E. & Kollman, P.A. (1990). Protein Model Structure Evaluation Using the Solvation Free Energy of Folding. *Proc. Natl. Acad. Sci. USA* **87**, pp. 3240-3243.

15. Eisenberg, D. & McLachlan, A.D. (1986). Solvation energy in protein folding and binding. *Nature* **319**, pp. 199-203.

16. Lee, M.R., Duan, Y. & Kollman, P.A. (2000). Use of MM-PB/SA in estimating the free energies of proteins: Application to native, intermediates, and unfolded villin headpiece. *Proteins* **39**, pp. 309-316.

17. Lee, M.R., Baker, D. & Kollman, P.A. (2001). 2.1 and 1.8 angstrom average C-alpha RMSD structure predictions on two small proteins, HP-36 and S15. *J. Am. Chem. Soc.* **123**, pp. 1040-1046.

18. Case, D.A., Pearlman, D.A., Caldwell, J.A., 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. (1997). AMBER 5.0, University of California, San Francisco: San Francisco.

19. 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. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **117**, pp. 5179-5197.

20. Kollman, P., Dixon, R., Cornell, W., Fox, T., Chipot, C. & Pohorille, A. (1997). The development/application of a 'minimalist' organic/biochemical molecular mechanic force field using a combination of *ab initio* calculations and experimental data. In *Computer Simulation of Biomolecular Systems* (Wilkinson, P., Weiner, P. & Van Gunsteren, W., eds.), vol. 3. pp. 83-96, Elsevier.

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

22. Darden, T., York, D. & Pedersen, L. (1993). Particle mesh Ewald: an N.log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, pp. 10089-10092.

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

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

25. Sharp, K.A., Nicholls, A. & Sridharan, S. (1998). Delphi, II ed, Columbia University: New York, NY 10032.

26. Kollman, P.A., Massova, I., Reyes, C., Kuhn, B., Huo, S.H., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D.A. & Cheatham, T.E. (2000). Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.* **33**, pp. 889-897.

Chapter 5:

Molecular dynamics in the endgame of protein structure prediction

Matthew R. Lee[1]

Jerry Tsai[2]

David Baker[2]

Peter A. Kollman[1]*

[1] Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA 94143-0446, United States of America

[2] Department of Biochemistry, University of Washington, Seattle, Washington 98195, United States of America

submitted to J. Mol. Biol.

ABSTRACT:

In order to effectively sample as much conformational space as possible, methods for protein structure prediction make necessary simplifications that also prevent them from being as accurate as desired. Thus, the idea of feeding them, hierarchically, into a more accurate method that samples less effectively was introduced a decade ago, but has not met with more than limited success in a few isolated instances. Ideally, the final stages should be able to identify the native state, show a good correlation with native similarity in order to add value to the selection process, and refine the structures even further. In this work, we explore the possibility of using state-of-the-art explicit solvent molecular dynamics and implicit solvent free energy calculations to accomplish all three of those objectives on 12 small, single-domain proteins, 4 each of alpha, beta and mixed topologies. We find that this approach is very successful in ranking the native and also enhances the structure selection of predictions generated from the Rosetta method.

INTRODUCTION:

Approaches for predicting three-dimensional protein structure based on amino acid sequence, ranging from *ab initio* to comparative modeling, all make considerable approximations in order to contend with the otherwise intractable number of possible conformations. Commonly in *ab initio* methods, a simplified energy potential is used together with a reduced representation of the protein, in which case side chains are often represented by centroids, hydrogen atoms are usually omitted, and only a few discrete torsional angles are allowed. Comparative modeling methods also rely on many of the same approximations, albeit primarily on the non-homologous regions. These simplifications, while beneficial in that they filter out the majority of unrealistic and improbable structures, limit the degree of accuracy that can be obtained. Even over the homologous regions of a comparative modeling effort, the exact native structure of any sequence inevitably differs from its nearest structural neighbor template, particularly in localized areas that may allow for small global superposition differences, despite large, local deviations that can not be corrected without a more accurate representation of the protein and the energy potential, as well as sufficient sampling.

The solution for overcoming the limiting simplifications is not to remove them from the outset, but rather to add in the detail when necessary, because introducing a higher level of accuracy to the energy potential makes for a more rugged surface that is more difficult to sample, thereby restricting the distance in conformational space that can be sampled on a practical timescale. Thus, the current structure prediction methods must draw the tertiary structure sufficiently close to the correct structure, within a "radius of

convergence", before all-atom detail with a continuum torsional space may be capable of improving them further.

The first attempts at using all-atom models as the final stage of a hierarchical approach took place a decade ago[1,2], before control simulations were capable of maintaining the native state, at a time when the computer power required for even short simulations was very demanding. These investigators applied their methods to the GCN4 leucine zipper, which has a very simple coiled coil homodimer topology consisting of two 33-mer monomers, of which all 33 residues were α-helical. In the end, they obtained ~1 Å backbone RMSD structures, but only with the help of α-helical constraints applied to every residue. While having brought forth the enticing idea of hierarchical protein structure prediction, these studies were only successful because they knew the correct structure to begin with and used native constraints to severely reduce the conformational search. Samudrala et al. later attempted a hierarchical approach by building all-atom models from a subset of off-lattice predictions on a set of 13 proteins and applying minimization alone[3], leading to the correct global topology in 6 of the cases. However, in this study, it was not demonstrated and is unlikely that the final stage of this hierarchical effort added any value to the initial off-lattice models, since minimization affords extremely limited conformational sampling at best. More recently, with advances in simulation methods, most notably being accurate means for treatment of long-range electrostatics[4] that allows for maintenance of native protein structures[5], our group (Simmerling et al.) used an enhanced sampling protocol called Locally Enhanced Sampling[6], which has been shown to lower energy barriers using a mean-field approach, that drove a 3.7 Å 29-mer protein structure with an incorrectly packed beta sheet to a 2.2

Å conformation with the correct topology[7]. Even more recently, we ran nanosecond state-of-the-art molecular dynamics simulations, with accurate long-range electrostatics and explicit solvent, on initial structure predictions for the 36-mer HP-36 and the 65-mer S15 alpha proteins, not only improving some of the model predictions to sub-2.0 Å Cα RMSD structures, but also demonstrating that the highest resolution models also had the best predicted Molecular Mechanics-Poisson Boltzmann (MM-PBSA) free energies among a handful of other models with less native similarity[8].

In the current work, we further explore the promise of using explicit solvent molecular dynamics simulations together with MM-PBSA for the endgame of structure predictions on 12 other small single-domain proteins, 4 alpha, 4 beta and 4 mixed. The three main objectives are: 1) identification the native state, 2) improved filtering over the previous stage by providing better correlation with native similarity and 3) refinement of the structures.

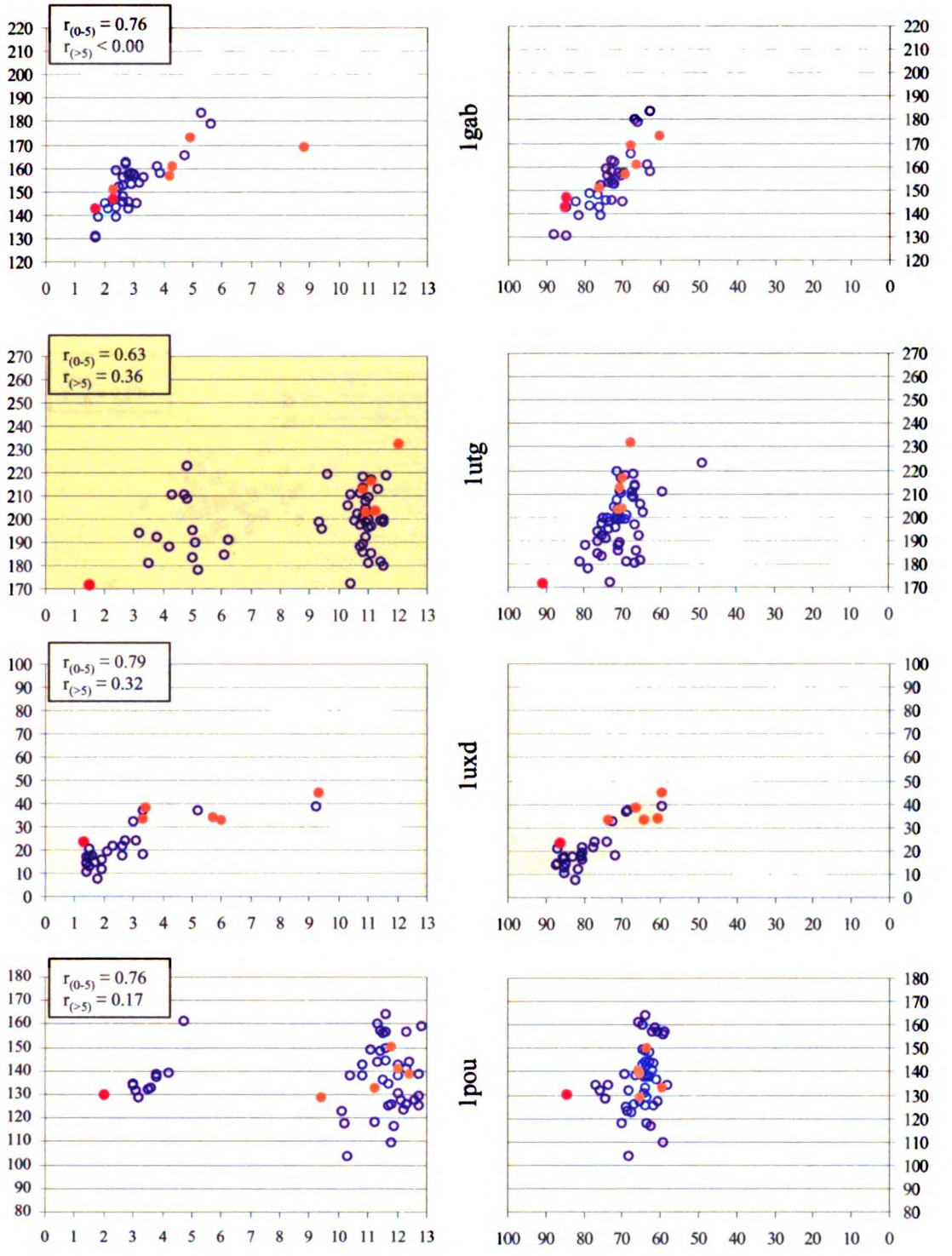


Figure 1. Alpha proteins. For each protein, the ensemble-average MM-PBSA was plotted as a function of either Cα RMSD (left panels) or Q, % of native contacts, (right panels) for every conformational family with a life of more than 100 ps. The conformational families containing the native state are illustrated as closed red circles, those containing the 5 most favorable Rosetta initial structures as close orange circles, and those having started from the remaining Rosetta structures as open blue circles.

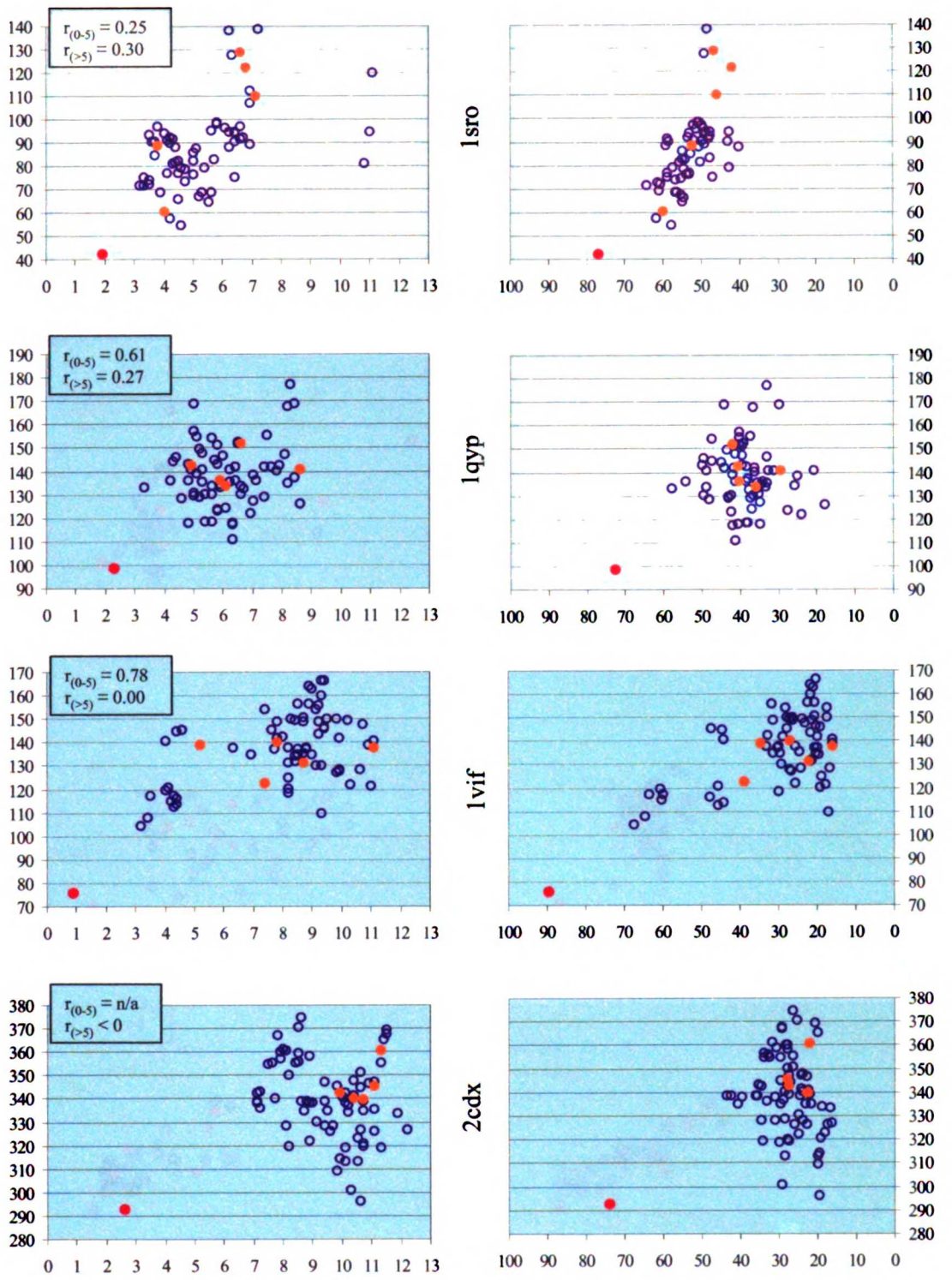


Figure 2. Beta proteins. See Figure 1 captions.

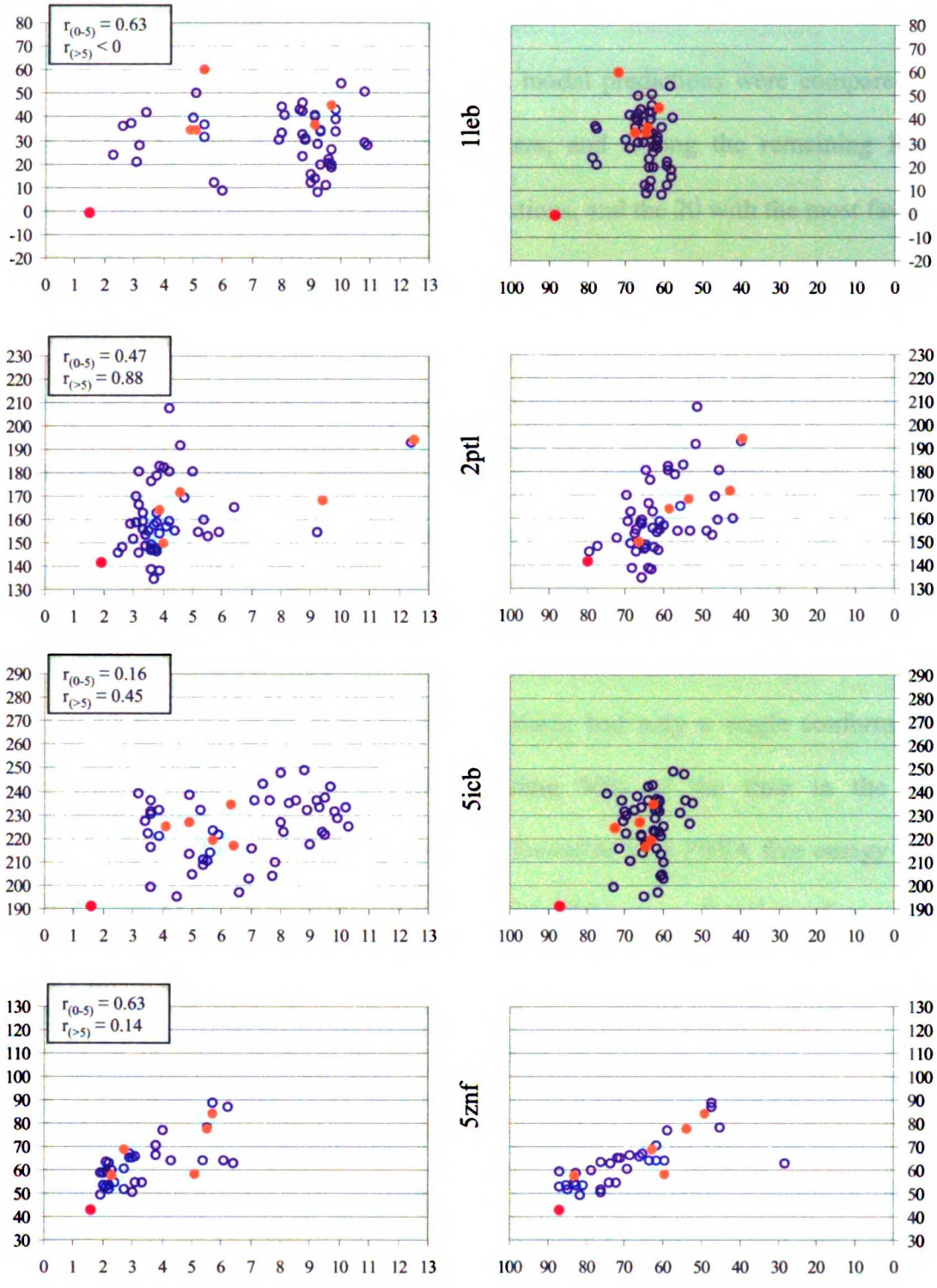


Figure 3. Mixed proteins. See Figure 1 captions.

RESULTS & DISCUSSION:

Conformational families

For each of the 12 proteins, 30 Rosetta model predictions were compared: the centers of the 5 most highly populated clusters, and among the remaining Rosetta predictions, the 5 with the best Cα RMSD predictions, and the 20 with the most favorable Rosetta energy scores. We equilibrated each of these Rosetta models and the experimental structures in a box of TIP3P[9] water with a 10 Å buffer and ran one ns production phase trajectories, for a total of 372 explicit solvent one ns simulations. After having clustered the resulting trajectories, using a 2.5 Å Cα RMSD cutoff (see Methods), we observe that the Rosetta model predictions had an average of 1.8 conformational families over the course of the nanosecond simulation; more specifically, the alpha proteins averaged 1.5, the beta proteins 2.4, and the mixed proteins 1.6. In comparison, 11 of the 12 trajectories on experimental structures had only a single conformational family, with the lone exception 1gab spending 90% of the time in the initial conformational family that had a slightly more favorable MM-PBSA free energy (Table 1). The ensemble-average values for the MM-PBSA as a function of two native similarity metrics, Cα RMSD on the left and percentage of native contacts on the right, are plotted for each conformational family in Figures 1-3.

Table 1. Native state stability

	protein[1]	residues	<RMSD>$_{init}$[2]	<RMSD>$_{2nd}$	<Q>$_{init}$[2]	<Q>$_{2nd}$
alpha	1gab	47	1.7	2.3	85.4	84.8
	1utg	62	1.5		91.1	
	1uxd	43	1.3		86.2	
	1pou	70	2.0		84.4	
beta	1sro	66	1.9		76.9	
	1qyp	42	2.3		72.3	
	1vif	48	0.9		89.8	
	2cdx	54	2.6		74.1	
mixed	1leb	63	1.5		88.5	
	2ptl	60	1.9		79.8	
	5icb	72	1.6		87.0	
	5znf	25	1.6		86.9	

[1] 1utg, 1vif and 5icb are X-ray crystal structures. The remaining 9 are NMR structures.

[2] In the NMR cases, the average NMR structure was used as the reference.

Native states

Table 1 shows that, for the most part, control simulations led to very stable native states having average Cα RMSD's under 2.0 Å and on average a percentage of native contacts (Q-values) greater than 80%. Among the exceptions, the NMR model for 1pou seemed to have a <Cα RMSD> on the high end, although it still had a very good Q-value of 84%. The three beta proteins with NMR structures, 1sro, 1qyp and 2cdx, had RMSDs on the high end as well as Q-values on the low end, when being compared to their respective average NMR structures. For 2cdx, the one with the greatest deviation from the NMR models, the snapshots from the 1 ns trajectory showed an average pair-wise Cα RMSD of 1.36 Å from one another with a standard deviation of 0.36 Å, and consisted of a single conformational family. Similarly, for 1sro and 1qyp, the average pair-wise RMSDs were 1.42 and 1.32 Å, with standard deviations of 0.41 and 0.43 Å, respectively, and they too populated single conformational families throughout their simulations.

These findings are in agreement with a separate study[10], in which we suggest that the approximate treatment of solvent used in solving NMR structures causes them to be less reliable than crystal structures.

MM-PBSA parameters

In the MM-PBSA free energy method, there are a few parameters that one cannot derive from "first principles." Perhaps the greatest difficulty lies in deciding what interior dielectric constant (ε_{int}) is most appropriate. On the one hand, because the atomic point charges in our force field have been derived based on high level quantum mechanical charges with a dielectric constant of 1, we may be justified in using $\varepsilon_{int} = 1$. On the other hand, the experimental dielectric constant in proteins is ~4. Thus, the choice of ε_{int} may be system dependent, with larger dielectric constants than 1 likely to be appropriate in some instances.

Another uncertainty is in deciding which parameters to use for describing the backbone torsional potentials. Because the original Cornell et al. force field (PARM94)[11], which was parameterized on a set of dipeptides, was shown to slightly favor alpha helical conformations on a training set of tetrapeptides[12], the torsions for phi and psi had been modified in response to high level *ab initio* calculations on the alanine tetrapeptide, which led to a significantly better agreement between molecular mechanical and quantum mechanical relative free energies on the tetrapeptide training set, giving rise to the PARM96 force field. However, it is still not clear that one is more generally the better choice for proteins, particularly in the post-processing stage of MM-PBSA calculations.

The non-polar component of the solvation free energy is a third area in which one may explore multiple values or functional forms. In principle, this term should account for all of the non-electrostatic contributions associated with solvating a molecule, primarily including the entropically unfavorable cost of cavity formation and the always attractive dispersion interactions between solute and solvent. Because it has been reasoned that both of the primary factors involved in this term are roughly proportional to the solvent-accessible surface area (SASA), as found in alkanes, MM-PBSA and other methods that calculate the solvation free energy with a continuum solvent[13-15] use a small *positive* linear γ coefficient to scale this term as a function of SASA, which assumes that the relative weighting of the unfavorable cavitation is stronger than that of the attractive dispersion. An alternative approach, long used by Cramer's and Truhlar's groups, has been to calculate atomic surface tensions that depend on properties such as atom type and nearest-neighbor recognition,[16] which does not always lead the non-polar solvation free energy to be positive. Recently, Pitera and van Gunsteren[17] demonstrated the importance of considering all solute-solvent van der Waals interactions (VDW), including those buried in the protein interior, indicating that solvent excluded volume may be more appropriate than surface area in relating to the favorable aspect of non-electrostatic solvation free energies. While we continue to make the linear approximation, we explore the effect of using different γ coefficients.

Previous studies applying MM-PBSA to binding free energies[18] and relative free energies of stability on proteins[8,19] have been successful using values between 1 and 4 for ε_{int}, the PARM96 force field, and a γ coefficient between 5 and 7 cal/mol·Å^2. Figures 1-3 graphically depict the results using our standard values: $\varepsilon_{int} = 4$, PARM96 and $\gamma = 5.42$

cal/mol·Å^2 and Tables 2-4 show the effects of changing ε_{int}, the dihedral component of the force field, and γ on the ability to rank the native structure and on the strength of relationship with Cα RMSD, which we discuss below.

Table 2. Native rank

	protein	VDW[1]	eel_tot[2]	MM-PBSA standard set[3]	$\varepsilon_{int} = 1$[4]	PARM94[5]	$\gamma = 54.2$[6]	n
alpha	1gab	9	25	7 (2.08 Å)	12 (2.42 Å)	8 (2.14 Å)	10 (2.24 Å)	39
	1utg	4	18	1	4 (9.00 Å)	1	2 (10.40 Å)	55
	1uxd	2	35	23 (1.80 Å)	27 (1.90 Å)	10 (1.56 Å)	17 (1.70 Å)	36
	1pou	18	1	17 (11.01 Å)	1	10 (10.24 Å)	1	53
beta	1sro	2	37	1	1	4 (4.27 Å)	1	70
	1qyp	1	65	1	1	2 (8.60 Å)	1	71
	1vif	1	16	1	1	1	1	73
	2cdx	8	53	1	1	32 (10.16 Å)	3 (10.45 Å)	77
mixed	1leb	1	48	1	4 (7.07 Å)	3 (7.95 Å)	1	53
	2ptl	1	46	5 (3.70 Å)	13 (3.84 Å)	4 (3.67 Å)	1	54
	5icb	1	49	1	9 (5.74 Å)	4 (6.33 Å)	1	53
	5znf	9	26	1	1	1	1	38
weighted avg.[7]		4.46	36.23	4.11	5.11	7.20	2.69	

in parenthesis are the average RMSD's of the conformational families having lower energies than the native.

[1] van der Waals energy

[2] total electrostatic energy, using $\varepsilon_{int} = 4$: intra-solute Coulombic + $\Delta G_{solv,pol}$

[3] standard set is $\varepsilon_{int} = 4$, PARM96, and $\gamma = 5.42$ cal/mol·Å^2

[4] standard set, except for ε_{int}

[5] standard set, except for the force field

[6] standard set, except for γ

[7] weighted according to n (see Methods)

Native rank

Table 2 shows the native rank of the conformational families containing the equilibrated experimental structures, according to its van der Waals (VDW) and total electrostatics (eel_tot) components, and according to MM-PBSA, using various permutations of the three parameters mentioned above. With the standard set, MM-PBSA predicts the native family as most energetically favorable in 8 of the 12 proteins

(Table 2). In the alpha proteins 1gab and 1uxd, the average Cα RMSD of conformational families lower in free energy is only 2.08 and 1.80 Å, respectively, with the lowest 1gab energy Rosetta structures having as good a RMSD as the NMR model, and with one having even more native contacts (from the average NMR structure) than the NMR model. In the mixed protein 2ptl, only three of the 54 conformational families had a lower predicted free energy, all three of which were low RMSD structures. Only in the case of the 4-helix bundle 1pou do the standard parameters decidedly fail, where nearly half of the 53 conformational families scored better than native. VDW alone performs worse than the standard set MM-PBSA, predicting the native family as most favorable in only 5 of the same 8 that the standard set MM-PBSA did and no others. Interestingly, eel_tot predicts the native as best in only a single instance, 1pou, the protein that the standard set had the most difficulty with, and otherwise ranks very inadequately. Along those lines, using the lower $\varepsilon_{int} = 1$ allows MM-PBSA to correctly rank native in 1pou, while really only worsening three others, the alpha protein 1utg and the two mixed proteins 1leb and 5icb. The PARM94 force field, which has been suggested to unduly favor alpha helices[12], performs similarly to PARM96 on the alpha proteins, but worse on the beta and mixed proteins. Finally, amplifying the γ coefficient, which would more heavily weight the unfavorable cavitation term's dependence on SASA, also allows MM-PBSA to correctly rank 1pou and 2ptl, but slightly upsets the correct ranking of 1utg and 2cdx, with a net effect of ranking a bit better than the standard set.

Table 3. Strength of association with Cα RMSD

	protein	VDW[1]	eel_tot[2]	MM-PBSA standard set[3]	MM-PBSA $\varepsilon_{int} = 1$[4]	MM-PBSA PARM94[5]	MM-PBSA $\gamma = 54.2$[6]	n
alpha	1gab	0.77	-0.07	0.77	0.53	0.82	0.81	35
	1utg	0.64	-0.31	0.63	0.02	0.49	0.60	11
	1uxd	0.78	-0.52	0.74	0.49	0.76	0.77	31
	1pou	0.82	-0.40	0.76	0.75	0.67	0.34	11
beta	1sro	0.08	0.05	0.25	0.29	-0.04	0.22	35
	1qyp	0.57	-0.35	0.61	0.62	0.70	0.65	15
	1vif	0.81	0.12	0.78	0.78	0.77	0.80	14
	2cdx							0
mixed	1leb	0.50	-0.90	0.63	0.47	0.31	0.44	8
	2ptl	0.46	-0.05	0.39	0.35	0.41	0.47	46
	5icb	0.24	-0.27	0.16	-0.04	0.23	0.20	17
	5znf	0.48	-0.08	0.63	0.63	0.47	0.58	29
weighted avg.[7]		0.53	-0.18	0.55	0.44	0.49	0.54	

These values represent the Pearson product-moment correlation coefficients among families < 5 Å from the experimental structure.

[1] van der Waals energy

[2] total electrostatic energy: intra-solute Coulombic + $\Delta G_{solv,pol}$

[3] standard set is $\varepsilon_{int} = 4$, PARM96, and $\gamma = 5.42$ cal/mol·Å^2

[4] standard set, except for ε_{int}

[5] standard set, except for the force field

[6] standard set, except for γ

[7] weighted according to n (see Methods)

Correlation with native similarity.

In order for any energy function to be useful for structure prediction, it must exhibit a good association with native similarity, not just correctly rank the native structure among a set of decoys. Moreover, in a successful hierarchical approach, the final stage must be more effective at correlating with native similarity than the initial structure prediction methods. In this study, we examine the linear correlation coefficient between Cα RMSD and the various energies as above, but only for structural families that were less than 5 Å from the experimental structures. We impose this 5 Å limit, because we suggest, in a separate work[10], based both on the notion of a globally convex free energy landscape and on data of large decoy sets, that the relationship between Cα RMSD and an effective free energy such as MM-PBSA is only linear near the native state, that the relationship disappears beyond 5 Å Cα RMSD.

Part of the reason why the relationship of our free energy with RMSD falls off beyond a certain point stems from RMSD not being the best way of describing native similarity. For instance, a single hinge motion between two domains can lead to very large RMSD values, despite native similarity being otherwise very high. Thus, Q-values, the percentage of formed native contacts, provide another way of judging how similar a given conformation is to the reference point, and should not lose their association with a free energy as readily as RMSD. As can be seen from Figures 1-3, the relationship of MM-PBSA is roughly more well behaved with respect to Q than RMSD in three (1gab, 1utg and 1uxd) of the four alpha proteins, in two (1sro and 1vif) of the four beta proteins, and in all four of the mixed proteins.

In Table 3, we summarize our findings, which show that the standard parameter set MM-PBSA correlates with Cα RMSD as well as any of the other terms or MM-PBSA parameter permutations. Somewhat surprisingly, VDW does as well as the much more computationally demanding entire effective free energy function itself, even though it did not rank native as well. Eel_tot shows virtually no correlation, which causes the MM-PBSA with $\varepsilon_{int} = 1$ to have a lesser association. The PARM94 force field performs similarly to PARM96 and the higher γ coefficient seem to have no effect on the strength of association between MM-PBSA and Cα RMSD.

Table 4. Ability to filter decoys

		Rosetta[1]		MM-PBSA[2]		VDW[2]	
	protein	init.[3]	avg.[4]	init.[3]	avg.[4]	init.[3]	avg.[4]
alpha	1gab	4.56	4.90	2.45	2.08	1.97	1.98
	1utg	10.87	11.20	8.36	8.32	9.96	9.80
	1uxd	5.31	5.52	1.82	1.52	1.70	1.58
	1pou	11.45	11.34	10.88	11.08	10.88	11.06
beta	1sro	6.12	5.66	4.34	4.56	4.72	4.48
	1qyp	6.51	6.42	5.78	5.86	6.30	6.58
	1vif	8.18	8.04	4.76	4.92	6.46	6.46
	2cdx	10.03	10.68	9.56	10.26	9.64	10.14
mixed	1leb	6.68	6.84	7.44	7.88	8.30	8.54
	2ptl	6.43	6.86	3.48	3.46	3.50	3.92
	5icb	5.33	5.48	5.96	5.86	5.62	5.44
	5znf	4.26	4.26	1.58	2.38	1.78	2.48
	avg.	7.14	7.27	5.53	5.68	5.90	6.04

[1] The average values of the 5 most highly populated Rosetta structures

[2] The average values of the 5 lowest energy structures

[3] Initial Cα RMSDs

[4] Ensemble-average Cα RMSDs from first conformational families if more than one
Note that this is an average of an average.

Although it is useful to have a scoring function that relates to native similarity, the more relevant issue, in the context of hierarchical protein structure prediction, is whether or not MM-PBSA provides a better filter than Rosetta at selecting the most promising predictions. Because Rosetta does not rely entirely on its energy score in identifying its most favored conformations, it is not appropriate to calculate its correlation coefficient with Cα RMSD. Instead, to evaluate whether or not MM-PBSA or VDW is advantageous over Rosetta in scoring its predictions, we compare the <Cα RMSD> of the best 5 conformations in Table 4, this being the centers of the 5 most highly populated clusters generated from Rosetta and the 5 lowest energy structures according to MM-PBSA or VDW alone. Under each of the three scoring functions, for each protein, we show the average values of the 5 deemed best, from both their initial and their ensemble-

average Cα RMSDs. As can be seen from Table 4 and from Figures 1-3, MM-PBSA (using the standard parameter set) improves the structure selection process, as does VDW. Among the alpha proteins, three of the four benefited from MM-PBSA, with the initial structures being on average 2 to 3 Å better than the 5 chosen by Rosetta. Additionally, for 1gab and 1uxd, the molecular dynamics simulations improved the RMSD even further by an additional ~ 0.5 Å. In the beta proteins, the selection process also benefited from the more accurate MM-PBSA in all four proteins, albeit marginally with 2cdx, which was only 0.5 Å better. Finally, among the mixed proteins, the selection process improved substantially in half the cases, with 2ptl and 5znf structures having Cα RMSDs that were roughly 3 Å lower, but those from 1leb and 5icb were marginally worse by one and 0.5 Å, respectively. Using the VDW energy alone as the scoring function allowed for the same qualitative areas of improvement in filtering, although the extent of improvement was slightly less. In summary, the average initial Cα RMSD among the 5 chosen by Rosetta was 7.14 Å, that from VDW was 5.90 Å, and that from MM-PBSA was 5.53 Å.

Refinement

The final objective in the endgame of hierarchical protein structure prediction entails improving the native quality of the initial predictions. As we found in our previous work, in which we refined two small alpha proteins[8], there are two aspects of refinement: 1) relaxation to allow for very small domain shifts and correction of locally unfavorable geometries, which have minimal barriers and occur within 50 ps of molecular dynamics time, due to the more accurate free energy surface in molecular

mechanics and 2) transitions over energy barriers into new conformational families that may have more favorable free energies and more native similarity. Thus to isolate the two potentialities, each trajectory was clustered into conformational families with a 2.5 Å cutoff as mentioned above.

Table 5. Relaxation of initial conformations

		0 - 2.5 Å			2.5 - 5.0 Å			> 5.0 Å		
	protein	init[1]	<RMSD>$_{init}$[2]	n	init[1]	<RMSD>$_{init}$[2]	n	init[1]	<RMSD>$_{init}$[2]	n
alpha	1gab	2.26	2.23	9	3.35	3.11	20	8.10	8.80	1
	1utg			0	3.88	4.42	5	10.36	10.52	25
	1uxd	1.86	1.70	20	3.25	3.38	6	6.73	7.00	3
	1pou	2.30	3.00	1	3.68	3.63	4	11.43	11.57	25
beta	1sro			0	4.13	4.18	19	7.47	7.20	10
	1qyp			0	3.99	4.54	8	6.49	6.57	21
	1vif			0	3.62	4.02	5	9.02	8.93	23
	2cdx			0			0	9.10	9.61	29
mixed	1leb	2.50	2.45	2	3.35	3.58	4	8.12	8.31	25
	2ptl	2.46	3.22	7	3.62	3.84	22	9.03	9.27	3
	5icb			0	3.81	4.02	4	7.53	7.62	24
	5znf	1.50	2.37	21	3.84	4.50	8	6.50	6.40	1
weighted avgs.[3]	alpha	1.99	1.90	30	3.45	3.40	35	10.61	10.78	54
	beta			0	4.02	4.25	32	8.22	8.36	83
	mixed	1.79	2.57	30	3.66	3.97	38	7.87	8.02	53
	all	1.89	2.24	60	3.70	3.86	105	8.80	8.95	190

Conformational families for each protein are grouped into 3 bins based on their initial Cα RMSD.
Values reported in this table are the mean values among all members in the bin.
[1] initial Cα RMSD
[2] ensemble-average of the initial conformational family
[3] weighted according to n (see Methods)

For looking at possible refinement in the form of relaxation, we examine the initial Cα RMSDs in comparison to the average RMSDs of the very first conformational family (Table 6). We further split the data into close, medium and distant bins, 0 – 2.5 Å from the experimental structure, 2.5 – 5.0 Å, and > 5.0, respectively, because we believe that the closer the structure is to the native to begin with, the greater the likelihood that conformational changes will be favorable. On average, only the relaxation of structures

in the close and medium bins of the alpha proteins improved the native similarity, but only slightly.

Table 6. Transitions from initial conformations

		0 - 2.5 Å			2.5 - 5.0 Å			> 5.0 Å		
	protein	$\langle RMSD \rangle_{init}$ [1]	$\langle RMSD \rangle_{2nd}$ [2]	n	$\langle RMSD \rangle_{init}$ [1]	$\langle RMSD \rangle_{2nd}$ [2]	n	$\langle RMSD \rangle_{init}$ [1]	$\langle RMSD \rangle_{2nd}$ [2]	n
mixed	1leb			0	3.20	3.40	1	8.59	8.58	19
	2ptl	3.20	3.55	4	3.86	4.00	14	10.95	10.86	3
	5icb			0	4.16	4.16	5	7.60	7.75	19
	5znf	2.77	3.63	2	5.43	5.88	3			0
alpha	1gab	2.65	3.25	2	3.43	3.83	3	8.80	9.00	1
	1utg			0	4.42	4.26	5	10.55	10.47	15
	1uxd	2.13	2.23	3	3.40	3.30	1	9.30	9.20	1
	1pou	3.00	3.10	1	3.67	3.90	3	11.73	11.27	16
beta	1sro			0	4.24	4.48	14	6.80	6.78	9
	1qyp			0	4.80	4.96	7	6.52	6.78	18
	1vif			0	4.01	4.14	5	8.92	8.87	20
	2cdx			0			0	9.44	9.67	24
weighted avg.[3]		2.75	3.15	12	4.17	4.32	61	8.88	8.90	145

Trajectories for each protein that underwent a structural transition are grouped into 3 bins based on their initial Cα RMSD.
Values reported in this table are the mean values among all members in the bin.

[1] ensemble-average of the initial conformational family

[2] ensemble-average of the 2nd conformational family

[3] weighted according to n (see Methods)

Not all of the trajectories contained more than a single conformational family, but among those that did, Table 7 compares the ensemble-average Cα RMSDs of the initial and second conformational families, again further split into similarity bins. While we did not see any bin in which conformational changes led to more native families, we would only expect this to happen with any regularity in the close similarity bin, where there very few transitions that are probably not statistically relevant.

CONCLUSIONS:

While in principle, progressive improvement in detail should allow for more accurate protein structure prediction, this has not been shown to be the case, except in a

few isolated proteins, largely due to the very short radius of convergence afforded by the more accurate, detailed methods such as molecular mechanics that include all-atom accuracy along with energy potentials based on first principles, and due to the structure predictions not being of high enough resolution, lying outside of the radius of convergence. Improvements in the initial stages of protein structure prediction by a small handful of methods, most notably the Rosetta method, and development of accurate methods for evaluating the relative free energies of stability, have provided us with the opportunity to demonstrate a successful hierarchical collaboration.

Native rank, filtering and refinement are the three main objectives for the endgame of protein structure prediction. Among the 12 proteins, each with a distinctive topology, the methods presented in this work handle the first of these goals very well, correctly placing native as first in 8 of the examples. In the remaining four, the lower energy structures had average Cα RMSDs of only ~2 Å in two of the proteins, which is considered to within the narrow range of natural fluctuation around the native state under physiological conditions[20], and 3.7 in another. In the fourth protein, using either a lower interior dielectric constant of unity or a higher γ coefficient for the non-polar solvation free energy lead to the corrected native rank. The methods presented in this study also perform adequately as a filtering mechanism in an absolute sense, and substantially better than Rosetta in a relative sense. The third objective, despite our previously reported success on the HP-36 villin headpiece and ribosomal S15 protein[8], is one in which we do not succeed; this does not imply that molecular dynamics made structures worse, only that it did not improve them.

That we were unable to really refine the best structures came as somewhat of a disappointment, but not entirely as a surprise. The nature of refinement found in both S15 and HP-36 included small helical domain shifts into more tightly packed structures. While we do not believe that our energy function is inadequate to improve the structures, we do feel that perhaps one nanosecond explicit solvent simulations are too short for more systematic refinement of close structures. In order to overcome this limitation, apart from trying to simply run longer simulations, methods that improve the sampling may provide the solution. Locally Enhanced Sampling was effective on CMTI, which has 3 disulfide bridges over 29 residues, as mentioned above, but application of this approach on proteins less stable than the disulfide-rich CMTI led to unstable control simulations on the native structure (unpublished results), presumably due to the additional entropy of the method which altered the free energy surface. But because Locally Enhanced Sampling still stands out as a promising method for overcoming large energy barriers, particularly when used locally rather than globally, one might envision application of this mean-field approach directed at those regions with greater known uncertainty in the beginning stages of a hierarchical structure prediction, such as the intervening sequences between predicted secondary structural elements. Alternatively, implicit solvent molecular dynamics simulations[21-23] provide another potential approach for improving sampling, both in terms of the length of simulation that can be accomplished and in terms of the more rapid conformational changes that accompany the absence of solvent viscosity.

Apart from the lack of success in refinement aspect, the methods presented in this work still performed admirably in ranking the native and selecting better structures than

Rosetta. With the automation software used in this work, along with the increasingly greater computational power that continues to emerge, the methods described in this work for the final stages of structure prediction are much more accessible to the structure prediction community than only a few years ago.

METHODS:

The AMBER 5 suite of programs[24] was used for all molecular mechanics simulations. The PARM94 all-atom force field[11] was used for the molecular dynamics simulations and both the PARM94 and PARM96[25] force fields, the latter of which differs only in the φ, ψ torsional potentials of the peptide unit, was used in the MM-PBSA free energy analysis.

Molecular dynamics

We ran all production-phase molecular dynamics simulations with a 2.0 fs time step under the isothermal-isobaric ensemble (300 K and 1atm) with explicit solvent, using the TIP3P model[9] for water, periodic boundary conditions, the particle mesh Ewald (PME) method[4] for electrostatics, a 10 A cutoff for Lennard-Jones interactions, and the use of SHAKE[26] for restricting motion of all covalent bonds involving hydrogen atoms. Water molecules were added around the proteins using a 10 Å buffer from the edge of the periodic box. The temperature and pressure were maintained by the Berendsen coupling algorithm using a τ coupling constants of 1.0. PME grid spacing was ~1.0 Å and was interpolated on a cubic B-spline, with the direct sum tolerance set to 10^{-5}. We removed the net center of velocity every 100 ps to correct for the small energy drainage, that

results from the use of SHAKE, discontinuity in the potential energy near the Lennard Jones cutoff value, and constant pressure conditions.

For equilibration, we solvated the minimized structures, minimized the water molecules alone until the RMSD was < 0.1 kcal/mol·Å and then slowly heated, while allowing the water to move unrestrained for 25 ps (with a 1.0 fs time step) in order to fill any vacuum pockets.

To cluster the molecular dynamics trajectories, we defined conformational families as being those with Cα RMSDs of <2.5 Å from the first structure in the family, with the first snapshot lying >2.5 Å from the first member of the initial family deemed as the first structure of the 2nd conformational family. On those families that were not populated for 100 or more ps, we did not calculate ensemble-averages and did not consider them in any of the results we report in this study.

MM-PBSA

Coordinates from a trajectory were saved every 5 ps, and the MM-PBSA calculation evaluated on each of them. The MM-PBSA free energy of each snapshot is approximated as the sum of two terms, using an interior dielectric constant of 4: the internal energy of the protein (E_{MM}) and a solvation free energy (ΔG_{solv}). E_{MM} is the sum of an internal strain energy (E_{int}), a van der Waals energy (VDW), and an intra-solute electrostatic energy (EEL). ΔG_{solv} consists of the cost of submerging a discharged solute in solvent (Δsolv_NP) and the subsequent cost of adding the charges back to the solute (Δsolv_eel). Δsolv_NP is approximated as being linearly related to the solvent accessible surface area (SASA): γ*SASA + 920 cal/mol. We adhered to the same Poisson-

Boltzmann protocol as first described previously[13], which used DelPhi II[27] and most of its standard default parameters, together with PARSE atomic radii and Cornell et al. charges[11], to calculate Δsolv_eel. The entropy of a given snapshot, which is mostly vibrational, can be calculated with normal mode analysis on a Newton-Raphson minimization. This, however, is the most time-intensive part of the MM-PBSA method on a per-snapshot bases. Given the results in our previous study[19], where we found this term to be indistinguishable among the native state, the folding intermediate, and the unfolded state of HP-36, we did not perform this calculation in the current study. For a more detailed discussion of the MM-PBSA method, see the review by Kollman et al.[18].

NMR structures

When using the term "the NMR structure", we are referring to model 1 in each of the NMR ensembles. We used this as the representative for simulation purposes, as it is more physically realistic than an average structure. The RMSDs, however, are always calculated in reference to the average NMR structure, as it is most representative of the various geometries of the ensemble.

Q-values

A contact is defined as any two residues containing atoms ≤ 3.5 Å apart. A contact map is generated for the actual experimental structure of X-ray crystals and for the average NMR structure of NMR ensembles. The Q-value represents the percentage of contacts in the native contact map that are also found in the conformation being

evaluated, with the exact same topologies being required in both the reference and target configurations.

Weighted averages

In Tables 2, 3, 5 and 6, we report a weighted average according to n, which is calculated as follows:

$$weighted\ avg. = \frac{\sum_{i=1}^{N} i \bullet n_i}{N}$$

n_i is the number of samples in ensemble i, and N is the total number of samples among all ensembles in the bin.

Automation

The bottleneck in running molecular dynamics simulations and MM-PBSA calculations on a single protein conformation lies in the computer time. However, when dealing with a larger number, human intervention and data analysis takes over that role. For this work, in which we simulated nanosecond length simulations, analyzed, and post-processed the MM-PBSA on 372 different structures, a set of programs with the Perl scripting language was written to automate the process (Appendix A in thesis by M.R. Lee[28]). These programs allow for implementation of a standard set of flags for running the simulations and scale with complete efficiency up to the number of computer processors available, by running simulations in coarse grain parallel. The majority of simulations in this work were run on 6 separate 4-processor Compaq Alpha ES40 machines, which when combined with the automation software, allowed for 24 independent simulations to be running simultaneously.

REFERENCES:

1. Vieth, M., Kolinski, A., Brooks, C. L., 3rd & Skolnick, J. (1994). Prediction of the folding pathways and structure of the GCN4 leucine zipper. *J. Mol. Biol.* **237**(4), 361-367.

2. Nilges, M. & Brünger, A. T. (1991). Automated modeling of coiled coils: application to the GCN4 dimerization region. *Protein Eng.* **4**(6), 649-659.

3. Samudrala, R., Xia, Y., Huang, E. & Levitt, M. (1999). Ab initio protein structure prediction using a combined hierarchical approach. *Proteins* **Suppl 3**(3), 194-198.

4. Darden, T., York, D. & Pedersen, L. (1993). Particle mesh Ewald: an N.log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**(12), 10089-10092.

5. Fox, T. & Kollman, P. A. (1996). The application of different solvation and electrostatic models in molecular dynamics simulations of ubiquitin: how well is the X-ray structure "maintained"? *Proteins* **25**(3), 315-334.

6. Roitberg, A. & Elber, R. (1991). Modeling side chains in peptides and proteins: Application of the locally enhanced sampling and the simulated annealing methods to find minimum energy conformations. *J. Chem. Phys.* **95**(12), 9277-9287.

7. Simmerling, C., Lee, M. R., Ortiz, A. R., Kolinski, A., Skolnick, J. & Kollman, P. A. (2000). Combining MONSSTER and LES/PME to predict protein structure from amino acid sequence: Application to the small protein CMTI-1. *J. Am. Chem. Soc.* **122**(35), 8392-8402.

8. Lee, M. R., Baker, D. & Kollman, P. A. (2001). 2.1 and 1.8 angstrom average C-alpha RMSD structure predictions on two small proteins, HP-36 and S15. *J. Am. Chem. Soc.* **123**(6), 1040-1046.

9. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**(2), 926-935.

10. Lee, M. R. & Kollman, P. A. (submitted). Free energy calculations highlight differences in accuracy between X-ray and NMR structures and add value to protein structure prediction. *Structure*.

11. 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. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **117**(19), 5179-5197.

12. Beachy, M. D., Chasman, D., Murphy, R. B., Halgren, T. A. & Friesner, R. A. (1997). Accurate ab initio quantum chemical determination of the relative energetics of peptide conformations and assessment of empirical force fields. *J. Am. Chem. Soc.* **119**(25), 5908-5920.

13. Srinivasan, J., Cheatham, T. E., Cieplak, P., Kollman, P. A. & Case, D. A. (1998). Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate - DNA helices. *J. Am. Chem. Soc.* **120**(37), 9401-9409.

14. Still, W. C., Tempczyk, A., Hawley, R. C. & Hendrickson, T. (1990). Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112**(16), 6127-6129.

15. Dominy, B. N. B., C.L., III. (in press). Identifying Native-Like Protein Structures Using Physics-Based Potentials. *J. Comp. Chem.*

16. Giesen, D. J., Hawkins, G. D., Liotard, D. A., Cramer, C. J. & Truhlar, D. G. (1999). A universal model for the quantum mechanical calculation of free energies of solvation in non-aqueous solvents (vol 98, pg 85, 1997). *Theor. Chem. Acc.* **101**(4), 309.

17. Pitera, J. W. & van Gunsteren, W. F. (2001). The importance of solute-solvent van der Waals interactions with interior atoms of biopolymers. *J. Am. Chem. Soc.* **123**(13), 3163-3164.

18. Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S. H., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D. A. & Cheatham, T. E. (2000). Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.* **33**(12), 889-897.

19. Lee, M. R., Duan, Y. & Kollman, P. A. (2000). Use of MM-PB/SA in estimating the free energies of proteins: Application to native, intermediates, and unfolded villin headpiece. *Proteins* **39**(4), 309-316.

20. Brooks, C. L., 3rd, Karplus, M. & Pettitt, B. M. (1988). Proteins: a theoretical perspective of dynamics, structure and thermodynamics. *Advan. Chem. Phys.* **71**, 1-259.

21. Luo, R., David, L. & Gilson, M. K. (submitted). Accelerating Finite Difference Poisson Boltzmann Calculations for Static and Dynamic Systems. *J. Comp. Chem.*

22. Dominy, B. N. & Brooks, C. L. (1999). Development of a generalized born model parametrization for proteins and nucleic acids. *J. Phys. Chem. B* **103**(18), 3765-3773.

23. Tsui, V. & Case, D. A. (2000). Molecular dynamics simulations of nucleic acids with a generalized born solvation model. *J. Am. Chem. Soc.* **122**(11), 2489-2498.

24. Case, D. A., Pearlman, D. A., Caldwell, J. A., 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. (1997). AMBER 5.0. University of California, San Francisco, San Francisco.

25. Kollman, P., Dixon, R., Cornell, W., Fox, T., Chipot, C. & Pohorille, A. (1997). The development/application of a 'minimalist' organic/biochemical molecular mechanic force field using a combination of *ab initio* calculations and experimental data. In *Computer Simulation of Biomolecular Systems* (Wilkinson, P., Weiner, P. & Van Gunsteren, W., eds.), Vol. 3, pp. 83-96. Elsevier.

26. Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. C. (1977). Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**(3), 327-341.

27. Sharp, K. A., Nicholls, A. & Sridharan, S. (1998). Delphi II edit. Columbia University, New York, NY 10032.

28. Lee, M. R. (2001). Using Molecular Dynamics for High Resolution Protein Structure Prediction. Ph.D. in Pharmaceutical Chemistry, University of California San Francisco.

Chapter 6:

Conclusions & Future Direction

CONCLUSIONS:

The work presented in this dissertation is the start of a bridge building between computationally demanding molecular mechanical simulations and genomic data. I demonstrate the usefulness of molecular dynamics simulations in refining structures, and of MM-PBSA in identifying the native structure, both relative to large sets of decoys, and in an absolute manner without *a priori* native state information by using an α-helical extended state as the reference point. Particularly important along the lines of working with genomic data, I also have written several scripting programs to allow for routine implementation of these methods for the endgame, the final stages, of protein structure prediction, that are capable of fully utilizing the type of increased computer power in modern architectures. Again the three main objectives of the endgame are:

1) identification of the native state
2) being able to judge native similarity by having a score that correlates well with native similarity in order to enhance the structure selection process
3) refinement of the best structure predictions.

What we learn specifically from Chapter 2 is that MM-PBSA allows for discrimination between the native state and a compact folding intermediate state, the start of an answer to objective 1. I showed that the lack of conformational entropy in MM-PBSA causes the energy gap between the native state, which has a very low degeneracy, and any other non-native state, which coexists at that energy level with many other structures, to be a sizeable margin that grows as the size of the protein increases, which led the potential for satisfying objective 1 more generally.

In Chapter 3, I found that I was able successfully meet objective 3, by markedly improving a few of the Rosetta structure predictions that were within a radius of convergence. I showed that state-of-the-art explicit solvent molecular dynamics can be useful for drawing structures closer to the native state as hypothesized. The most interesting result out of this work, however, was that the MM-PBSA energy gap principle learned from Chapter 1 was consistent with this work, and that refinement from a 3 Å to a 2 Å structure (which is arguably a member of the native state[1]) was accompanied by substantial energy drop, thereby providing another work in which we successfully met objective 1.

Among thousands of decoy structures investigated in Chapter 4, I found a reasonably good correlation with RMSD, thereby satisfying in part objective 2, and I also was the first to demonstrate that linear relationships between RMSD and an effective free energy exist only near the native state, which suggests that for satisfying objective 2, accurate physically-based scoring functions may only be of use among decoys less than ~5 Å from the native. Another interesting observation that came from this work is one of interest to the entire field of structural biology, that NMR models benefit markedly from short explicit solvent, restraint-free molecular dynamics simulations. While crystal structures were most favorable in every instance, NMR structures were significantly *less* favorable in every instance, and short 150 ps simulations corrected for this by ranking native as most favorable in each instance, by comparing at the ensemble-averages. Thus, this lends further support to applicability of MM-PBSA for meeting objective 1, but shows that it depends on having an accurate representative of the native state to begin with, which does not necessarily follow experimental structure determination. Alternatively, I demonstrate an unorthodox approach for meeting objective 1, that an

extended state calculation can be used as a reference point to give an absolute indicator of how good a particular free energy is, which is different and more useful than just knowing that a conformation has a lower free energy than all other decoys in its database set.

Finally, I show in Chapter 5 that MM-PBSA works well for meeting objectives 1 and 2 in a larger, more statistically significant set of 12 proteins. Not only did MM-PBSA correlate reasonably with RMSD, but it also significantly improved the selection process, compared to Rosetta. I also found, however, a single instance in which the standard set of MM-PBSA parameters failed in the native rank objective, that was corrected for by either using a lower dielectric constant or a higher γ coefficient of the non-polar solvation term.

FUTURE DIRECTION:

While I successfully accomplished the refinement objective on two small alpha proteins in Chapter 3, I was unable to reproduce similar results on any of the Rosetta predictions of the 12 proteins studied in Chapter 5. Thus, in future attempts to meet objective 3, simulation protocols that sample more effectively[1-3], as discussed in Chapter 1, may be useful.

The lack of successfully ranking the native in 1pou from the work in Chapter 5 suggests that a more accurate method for treating the non-polar solvation term may be warranted in future MM-PBSA development, particularly on protein stability studies, where the hydrophobic effect, long considered to be a dominant force in protein folding[4] is theoretically accounted for by this term. Because the solvent excluded volume (SEV) has been demonstrated to relate better to the attractive component of non-polar solvation

free energy than solvent accessible surface area (SASA)[5], one might consider splitting the γ·SASA term into a γ_1·SASA and a γ_2·SEV and redoing the parameterization, where γ_1 would be positive and γ_2 would be negative.

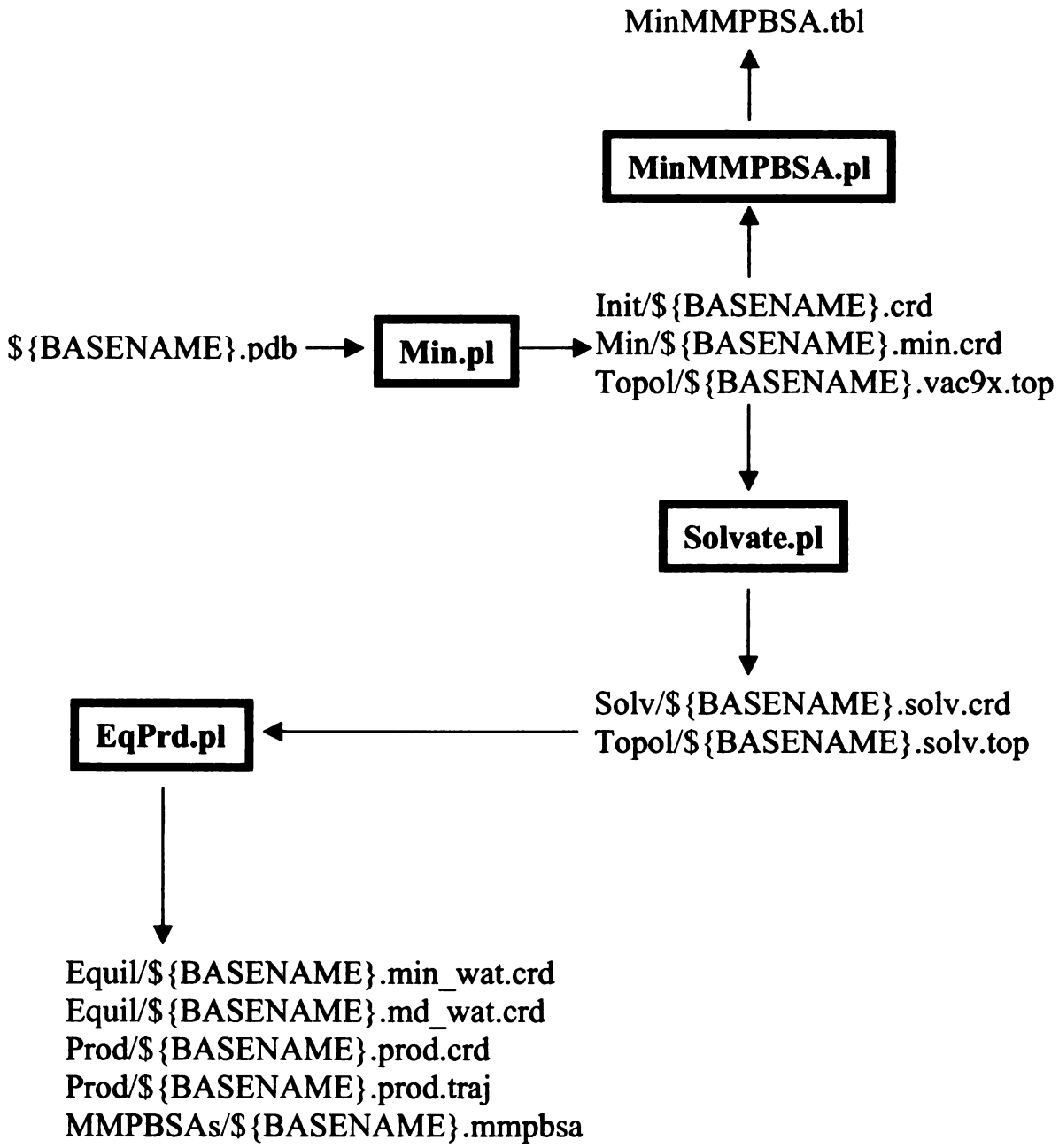
REFERENCES:

1. Brooks, C. L., 3rd, Karplus, M. & Pettitt, B. M. (1988). Proteins: a theoretical perspective of dynamics, structure and thermodynamics. *Advan. Chem. Phys.* **71**, 1-259.
2. Lazaridis, T. & Karplus, M. (1999). Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.* **288**(3), 477-487.
3. Tsui, V. & Case, D. A. (2000). Molecular dynamics simulations of nucleic acids with a generalized born solvation model. *J. Am. Chem. Soc.* **122**(11), 2489-2498.
4. Dill, K. A. (1985). Theory for the folding and stability of globular proteins. *Biochemistry* **24**(6), 1501-1509.
5. Pitera, J. W. & van Gunsteren, W. F. (2001). The importance of solute-solvent van der Waals interactions with interior atoms of biopolymers. *J. Am. Chem. Soc.* **123**(13), 3163-3164.

Appendix A:

Automating AMBER molecular dynamics simulations & MM-PBSA free energy calculations

OVERVIEW


MinMMPBSA.tbl
MinMMPBSA.pl
Init/${BASENAME}.crd
${BASENAME}.pdb
Min.pl
Min/${BASENAME}.min.crd
Topol/${BASENAME}.vac9x.top
Solvate.pl
Solv/${BASENAME}.solv.crd
Topol/${BASENAME}.solv.top
EqPrd.pl
Equil/${BASENAME}.min_wat.crd
Equil/${BASENAME}.md_wat.crd
Prod/${BASENAME}.prod.crd
Prod/${BASENAME}.prod.traj
MMPBSAs/${BASENAME}.mmpbsa

Min.pl

OVERVIEW: This program will creates AMBER coordinate and topology files for every *.pdb file found in the working directory, using **tleap**. It will also create individual "*batch scripts*" to minimize each structure and "*job scripts*" that will launch each of the individual "batch scripts", according to the type of queuing system specified by the user.

USAGE: Min.pl -sander [*0*] -tleap [*0*] -leaprc [*0*] -caps [*N*] -hetatm [*N*] -hyd [*N*] -sscut [*2.5*] -qtyp [*0*] -ncpus [*8*] -nsets [*2*] -tlim [*5*] -mreq [*10*] [*myleap*] [*toponly*] [*septop*] [*help*]

flags for finding necessary files

sander: Pathname of **sander** executable; default of 0 will search for "sander" in $path list.

tleap: Pathname of **tleap** executable; default of 0 will search for "tleap" in $path list.

leaprc: Pathname of a non-standard leaprc file; default is to create a standard one.

flags for how to run the minimization

maxcyc: Maximum # of cycles allowed

drms: Maximum rmsd of the Cartesian elements of the potential energy gradient

ntmin: 0 = steepest descent for 10 steps, then conjug. grad.
1 = steepest descent for ncyc steps, then conjug. grad.
2 = only steepest descent (only allowed with "-drms 0")

ncyc: Used only with "-ntmin 1"

flags for assigning topology and how pdb files are parsed

caps: Whether or not to use terminal protecting groups NME and ACE. Use "Y" or "y", otherwise charged terminals are default.

hetatm: Whether or not to read HETATM records from pdb files. Use "Y" or "y", otherwise ignoring HETATM's is default.

hyd: Whether or not to read hydrogen atoms from pdb files. Use "Y" or "y", otherwise ignoring (non-bb) hydrogens is default.

sscut: Cutoff to use for determining disulfide bonds.

flags for creating the job script

qtyp: Type of queue system (unicos, LSF, NQS or default of 0 for none)

ncpus: # cpu's that will be used in each set by job script

nsets: # sets that will be run by each job script

tlim: Estimated time needed (hours) for each individual minimization

mreq: Estimated memory required by each individual minimization

Note that tlim and mreq are ignored if qtyp is 0 (no queuing system).

flags for changing the default file output behavior

[myleap] Use this flag if you only want to create the leap input files for each ${BASENAME}.pdb found in "./Init/" or in "./".

[toponly] Use this flag if you only want to create the topology files for each ${BASENAME}.pdb

[septop] Use this flag to create separate topology files for each pdb structure.

Otherwise, hard links will be created for those with the same number of atoms.

help flag
[help] Print this USAGE message

FILE INPUT:

1) ${BASENAME}.pdb

Min.pl will sequentially search for all files ending with the suffix ".pdb" in the working directory and dynamically assign the root of the filename to the variable ${BASENAME}. (For example, for the file "abc.pdb", ${BASENAME} is "abc".) For each pdb file found, a set of files, whose nomenclature is based on the current ${BASENAME}, will be created as described below.

FILE OUTPUT:

1) Init/${BASENAME}.crd
2) Topol/${BASENAME}.vac94.top
   Topol/${BASENAME}.vac96.top
   Topol/${BASENAME}.vac99.top
   (tl.o$BASENAME if **tleap** fails on $BASENAME)

and the following unless "toponly" is used on the command line:

3) Run.min.$BASENAME (the batch scripts)
4) myleaprc.add.${BASENAME}
   (DO NOT DELETE - used by **Solvate.pl** module)
5) QMin.n (the job scripts)

Each QMin.n job script will contain ncpus*nsets number of batch scripts and should be submitted to the appropriate queue (or run interactively if qtyp is 0) and will produce the following files for each batch script:

1) Min/${BASENAME}.min.crd
2) Min/${BASENAME}.min.out
3) Min/${BASENAME}.min.info

Min.pl creates Init/, Topol/, and Min/ directories and organizes the output files accordingly.

DETAILED EXPLANATION OF FLAGS:

flags for finding necessary files

Min.pl will, by default, search through the path to try and locate executables. If these executables can not be found in the path, or if the user wishes to designate an alternative binary, the entire pathname should be used as the flag's argument (i.e. "-sander /usr/bin/sander ").

The standard leaprc file is compatible with all standard protein and nucleic acid residues, as well as all residues found in the standard AMBER force field. If a pdb file contains non-standard residues, such as an unnatural amino acid or an organic compound, the user will need to:

1) Create LEaP library files (*.off* or *.lib* suffix) and AMBER parameter files for every non-standard residue found in the pdb file; see LEaP documentation or the streptavidin tutorial on the AMBER web page:
   http://www.amber.ucsf.edu/amber/tutorial/streptavidin/index.html

2) Create a **tleap** source file that contains the command to load in the LEaP library and parameter files. The following approach is suggested.
   a) Run the "**Mkleaprc.pl**" program, which will create a **tleap** source file called "leaprc".
   b) Edit leaprc by adding the appropriate lines to the bottom:
      loadOff LEaP_input_file(s)
      loadAmberParams LEaP_param_file(s)
   c) Rename "leaprc" to something unique like, like "uniq_leaprc"

3) Use the uniquely named leaprc file as the argument for the "-leaprc" flag of **Min.pl**.

flags for how to run the minimization
The default protocol for minimizing is using steepest descent for 500 steps followed by conjugate gradient until the gradient of the Cartesian elements is below 0.4 kcal/mol/Å (or if 50,000 steps are run before this is satisfied), all in a distance dependent dielectric ($\varepsilon = 4r_{ij}$). Refer to the SANDER documentation for how to change the minimization protocol using the $maxcyc, $drms, $ntmin, and $ncyc flags.

flags for assigning topology and how pdb files are parsed
-caps, -hetatm, and -hyd: Pretty self-explanatory. By default, charged termini will be used and all HETATM and non-backbone hydrogen atom records are ignored from the pdb files. (The hydrogen atoms will be built-in by LEaP.)

-sscut: Threshold distance (in Å) between any two SG atoms used for assigning disulfide bridges. If any CYS (or CYX) residue is close enough to form a disulfide with more than one other cysteine residue, only the closest residue will be cross-linked. The standard value of 2.5 Å has a reasonable physical basis, but larger values may be desired for *ab initio* or comparative modeling protein structure predictions to lock down suspected or known disulfide bridges.

*WARNING ON RUNNING **Min.pl** MORE THAN ONCE ON A PDB FILE*
If **Min.pl** is run and disulfides are assigned, the involved CYS residues are renamed to CYX in the pdb file and any HG atoms will be discarded. When can this be a potential pitfall? If **Min.pl** is run again with a more stringent threshold (shorter -sscut argument), in which now fewer disulfides are assigned, the pdb file will incorrectly contain CYX residues, that are not part of a disulfide bridge and still lack HG atoms, because the more liberal previous run already converted the

CYS to a CYX. To revert back to CYS nomenclature of all cysteine residues, the program **"CYXtoCYS.pl"** can be run without any command line arguments, which will work on all pdb files in the working directory.

flags for creating the job script

As noted above in the OUTPUT FILES section, **Min.pl** creates two kinds of scripts, batch and job. A batch script contains all the information to run a minimization on single structure and a job script will run multiple batch scripts in coarse grain parallel series.

-qtyp: What kind of queuing job script to create. The unicos option is creates a standard NQS job script, except that the memory units are words rather than bytes. The default option will create a c-shell job script with no queuing information.

-ncpus: Number of processors (AKA "threads", or "CPU's") on which to run the job script in coarse grain parallel.

-nsets: Number of coarse grain parallel sets to run in series.

-tlim and -mreq: These flags are used in conjunction with -ncpus and -nsets to determine the total time and memory limits that will be used for the job script.

These flags are not the most intuitive and can best be understood by example. Suppose there are 20 pdb files in the working directory. The user wishes to run **Min.pl** on a machine with 4 open processes, so the -ncpus argument is set to 4. Due to time constraints of the batch system, each job can only run 2 sets of 4, so the -nsets argument is set to 2. Three QMin.n job scripts will be created: QMin.0 (8 batch scripts), QMin.1 (8 batch scripts), & QMin.2 (4 batch scipts). In QMin.0 and QMin.1, the first set of 4 will run in the background and once the job script is finished waiting for all 4 batch scripts of the first set to finish, the second set of 4 is launched and waited on; once the second set is complete, the job script is finished. The user approximates that each batch script will take 6 minutes and 5 MB of RAM, so -tlim is 0.1 and -mreq is 5. The job script will designate a total limit of 1 hour (8*0.1 = 0.8 hours, then rounded up) and 20 MB (5*4).

flags for changing the default file output behavior

myleap: The **Solvate.pl** program builds from the leaprc files created by **Min.pl** and if these files are accidentally removed, the "myleap" keyword can be entered on the **Min.pl** command line to just create the myleaprc.add.${BASENAME} files (without spending time using **tleap** to generate AMBER files).

toponly: If you only desire to create AMBER coordinate and topology files, without the batch and job scripts, use this command line keyword.

septop: If one has multiple conformations of a single biomolecule, it is desirable to create hard links of the vacuum topology files in order to save disk space (particularly when 5 or more conformations of the same are being evaluated).

Solvate.pl

OVERVIEW: For every Min/*.min.crd file found, this program will create solvated AMBER coordinate and topology files. It uses **tleap** to submerge the structure in a TIP3P water box with a 10 Å buffer on each edge.

USAGE: Solvate.pl [tleap] [ambpdb] [scwrl] [help] -p [94] -sscut [2.5]

flags for finding necessary files

tleap: Pathname of tleap executable; default of 0 will search for "tleap" in $path search.

ambpdb: Pathname of ambpdb executable; default of 0 will search for "ambpdb" in $path search.

scwrl: Pathname of scwrl executable, if scwrl sidechains are to be used. The default of 0 is to NOT use scwrl sidechains.

flags for run behavior

top: which FF to use for explicit solvent MD; default is parm94 Use "96" or "99" as flag options for the other FF's.

sscut: If scwrl sidechains are added, this flag is the cutoff distance for assigning SS bonds of the newly added side chains.

help flag

[help] Print this usage message

FILE INPUT:

1) Min/${BASENAME}.min.crd
2) Min/${ BASENAME }.min.out
3) Topol/${ BASENAME }.vac9x.top
4) ./myleaprc.add.${ BASENAME }

Solvate.pl will search for all Min/*.min.crd files and dynamically assign the ${BASENAME} variable in the same way that **Min.pl** does with *.pdb files. The min.out files are checked to ensure that the minimization completed properly.

FILE OUTPUT:

1) Topol/${ BASENAME }.solv.top
2) Solv/${ BASENAME }.solv.crd

Solvate.pl creates the Solv/ directory and puts all the solvated AMBER coordinate files there and all the solvated AMBER topology files in Topol/.

DETAILED EXPLANATION OF FLAGS:

flags for finding necessary files
The documentation of Min.pl applies to the -sander and -ambpdb flags.

-scwrl: This argument contains the pathname of the **scwrl** executable, if the user should desire to replace the existing side chains with those from the **scwrl** backbone-dependent rotamer library (Bower, M.J., Cohen, F.E. & Dunbrack, R.L., Jr. (1997) *J. Mol. Biol.* **267**, 1268-1282). Otherwise, the default argument of 0 will leave current side chains intact.

If the initial pdb structure did not contain side chains, **Min.pl** used tleap to build them in very crudely and it is highly recommended that you use **scwrl** to replace them.

NOTE: The option to build side chains appears in **Solvate.pl** rather than in **Min.pl** because **scwrl** fails very often in structures that have not been minimized.

-sscut: Like in **Min.pl**, the threshold distance (in Å) between any two SG atoms used for assigning disulfide bridges, but on side chains that were built in by **scwrl**. This flag can not be set unless the -scwrl flag is also set.

EqPrd.pl

OVERVIEW: For every Solv/*.crd file found, this program creates individual batch scripts to 1) equilibrate water atoms, 2) run production phase molecular dynamics and 3) optionally calculate the MM-PBSA free energies. Like **Min.pl**, it will also create batch and job scripts.

USAGE: EqPrd.pl [sander] [nice] [scwrl] -ncpus [8] -nsets [2] -ps [150] -tlim [5] -mreq [10] -qtyp [0] -equil [short] [trajonly] [help]

flags for finding necessary files

sander: Pathname of sander executable; defualt of 0 will search for "sander" in $path list.

flags for the batch scripts

ps: how long (ps) production phase will run

equil: "none", "short" or "long"
- none - no equilibration; creates only production batch scripts (*assumes Equil/*.md_wat.crd and Topol/*solv.top files exist*)
- short - 1) minimize H20 2) temp. ramp MD on H20 only
- long - 3) minimize entire system 4) temp. ramp on whole system

scwrl: Whether or not scwrl sidechains were added by Solvate.pl. Unless "Y" or "y" is argument, this flag is not set. If it is set, batch scripts will minimize solute again prior to water minimization

mmpbsa: Whether or not to calculate MM-PBSA on the production phase trajectory. Unless "N" or "n" is argument, MM-PBSA will be included

flags for creating the job script

qtyp: type of queue system (unicos, LSF, NQS or default of 0 for none)

ncpus: # cpu's that will be used by batch job

nsets: # sets that will be run by batch job

nice: The amount (between 0 and 20) to nice batch scripts in the job script The higher the number the lower the priority

tlim: Time needed (hours) for each minimization

mreq: Memory needed for each minimization

help flag

[help] Print this usage message

FILE INPUT:

1) Solv/${BASENAME}.solv.crd
2) Topol/${BASENAME}.solv.top

EqPrd.pl will search for all Solv/*solv.crd files and dynamically assitn the ${BASENAME} variable in the same way the **Min.pl** does with *.pdb files.

FILE OUTPUT:

Run.$BASENAME

QEqPrd.n files will then be generated,
each of which should be submitted to the queue
and will produce the following files for each $BASENAME:

1) Equil/${BASENAME}.min_wat.crd & Equil/${BASENAME}.md_wat.crd
2) Equil/${BASENAME}.min_wat.out.Z & Equil/${BASENAME}.md_wat.out.Z
3) Equil/${BASENAME}.min_wat.info & Equil/${BASENAME}.md_wat.info
4) Equil/${BASENAME}.md_wat.traj.Z
   NOTE: if longeq is specified, *.minall.* and *.tramp.*
   output files will also be generated.
5) Prod/${BASENAME}.prod.crd
6) Prod/${BASENAME}.prod.out.Z
7) Prod/${BASENAME}.prod.info
8) Prod/${BASENAME}.prod.traj.Z
9) ${BASENAME}.mmpbsa (energies from MMPBSA.pl and RMSD's from ptraj)

EqPrd.pl creates the Solv/ directory and puts all the solvated AMBER coordinate files there and all the solvated AMBER topology files in Topol/.

DETAILED EXPLANATION OF FLAGS:

flags for finding necessary files
Refer to the documentation of Min.pl.

flags for the batch scripts

-ps: This flag will specify how long to run the molecular dynamics production phase. The default length of 150 ps has been shown to give sufficient sampling for a meaningful <MM-PBSA>.

-equil: The user may select from three possible equilibration schemes:

none: If solvated AMBER coordinate and topology files already exist, this option may be used, but only after the files have been named appropriately; i.e. the coordinate file must be named "Solv/${BASENAME}.solv.crd" and the topology file "Topol/${BASENAME}.solv.top".

short: This option is the default and is recommended for most applications. It consists merely of minimizing and equilibrating the water molecules with a linear

temperature ramp from 5 to 300 K over 25 ps with a 1 fs time step. Assuming the initial solute itself has only gone through minimization prior to being submerged in a water box, this short equilibration scheme can potentially blow up during the subsequent production phase. At the first step of MD, the solute is essentially at 0 K (due to the minimization) and the solvent is at 300K (resulting from temperature ramp); the system as a whole is coupled to a heat bath of 300 K. Thus, this short scheme will lead to a sudden jolt of kinetic energy being added to the solute during the production run and thus can lead to failure. However, this short scheme saves a significant amount of time, which is desired when dealing with many conformations, has been found to be sufficient for most small protein conformations, in that the initial native structures generally remain under 1.5 Å from their starting point.

long: To more carefully equilibrate the system and better ensure that the starting structure will not move too far during the first hundreds of ps of the production phase, the user might want to implement this long equilibration scheme. After equilibrating the water molecules, the entire system will be minimized, followed by MD with a linear temperature ramp from 5 to 300 K over 20 ps with a 2 fs timestep.

scwrl: If **scwrl** sidechains were added during **Solvate.pl**, they should be minimized before the solvent is equilibrated; setting this flag will have the batch scripts do so.

mmpbsa: The batch scripts will by default launch MM-PBSA on the resulting production phase trajectories. If the user desires to perform free energy post-processing separately or not at all, this flag must be unset ("-mmpbsa N" on the command line of **EqPrd.pl**);

flags for creating the job script
Refer to the documentation of Min.pl.

-nice: This flag will place a "nice +n" before each batch script in the job script.

MinMMPBSA.pl

OVERVIEW: For every Min/*.min.crd file found, **MinMMPBSA.pl** will evaluate the MM-PBSA and print the energy information in a table, with the following column titles: structure, vdw, Solv_NP, EEL_tot, G94, G96, G99, and optionally CA_r and NatCon.

USAGE: MinMMPBSA.pl -ref [0] -ncpus [1] [help]

FILE INPUT:

1) Min/${BASENAME}.min.crd
2) Topol/${BASENAME}.vac9x.top

FILE OUTPUT:

1) MinMMPBSA.tbl

DETAILED EXPLANATION OF FLAGS:

ref: If a protein pdb file is entered here as the command line argument, **MinMMPBSA.pl** will, in addition to calculating the energies on each snapshot, calculate the CA RMSD (using ptraj) and % of native contacts (using ContactReader and ContactMap) between each snapshot and this reference conformation. The reference pdb file must be consistent with the topology file.

ncpus: The number of processors on which to run this in coarse-grain parallel.