# 2.1 and 1.8 Å Average $C_\alpha$ RMSD Structure Predictions on Two Small Proteins, HP-36 and S15

**Matthew R. Lee,[†] David Baker,[‡] and Peter A. Kollman*,[†]**

*Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California 94143-0446, Department of Biochemistry, University of Washington, Seattle, Washington 98195*

*Received August 23, 2000*

**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/surface area (MM−PBSA) free energies and alpha carbon ($C_\alpha$) RMSDs were then calculated for each family. Those conformational families with the lowest average free energies also contained the best $C_\alpha$ RMSD structures (1.4 Å for S15 and HP-36 core) and the lowest average $C_\alpha$ RMSDs (1.8 Å for S15, 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 $C_\alpha$ 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' ranks (Spearman rank, $r_s$) correlate well with the average $C_\alpha$ RMSDs ($r_s = 0.77$ for HP-36, $r_s = 0.83$ 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 of 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 nonexperimental 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, which is 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 $C_\alpha$ RMSD over all modeled residues of <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, which results 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 1−2 Å RMSD,[3,4] but those on non-native structures steadily drift into new conformational families (this work, unpublished results, Duan and Kollman,[5] and Alonso and Daggett[6]), which suggests 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

* To whom correspondence should be addressed. Fax: (415) 502-1411. E-mail: pak@cgl.ucsf.edu.

[†] University of California San Francisco.

[‡] University of Washington.

[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. Biol.* **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.

10.1021/ja003150i CCC: $20.00
Published on Web 01/20/2001

emerge, even within a single prediction method on any given protein. Due to the necessary limitations of the community-wide experiment, only five or fewer 3D coordinate predictions per group were submitted. Hence, in the absence of an experimental structure, an inability to accurately rank the native structural quality of predictions will usually preclude the best predictions from being identified from any prediction method. Without a standard for comparing coordinates, scoring functions together with physically meaningful (and often subjective) measures, such as 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 that are 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 and Sander[8] (the EMBL set), then calculating the average free energy of the ensembles, they found the native to be always more favorable. Lazaridis and 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 and 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; it also must be of appreciable size. This sizable gap should be directly related to the number of residues, because larger proteins have more degrees of freedom and, thus, a greater degeneracy of the unfolded state.

In the current study, we met both of the challenges of protein structure prediction in the context of two small proteins. We ran extended molecular dynamics simulations that led to higher resolution structure predictions in both cases. We also demonstrated 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 we evaluated 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, side chains were added using the backbone-dependent library of SCWRL.[12]

**Molecular Dynamics.** We ran production-phase molecular dynamics with a 2.0 fs time step 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 were added around S15 in order to end up with a buffer of about 10 Å from the edge of the periodic box, which resulted in box sizes of 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 $\tau$ coupling constants of 1.0 for the protein and solven,t and pressure was maintained with isotropic molecule-based scaling,[17] also with a $\tau$ coupling constant of 1.0. The 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 drains that resulted 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, using 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 <0.4 kcal/mol·Å. Water molecules alone were then minimized in the same way until the RMS was <0.1 kcal/mol·Å and then slowly heated, while allowing them to move unrestrained for 25 ps (with a 1.0 fs time step) 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 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.

To cluster the molecular dynamics trajectories, we defined conformational families as being those with $C_\alpha$ RMSDs of <2.0 Å from the cluster average. In cases where the value was >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 Origin200s at UCSF and on the Origin2000 at the National Center for Supercomputing Applications.

**Postprocessing the Energy of the Trajectory Data.** Coordinates from the trajectory were saved every 5 ps, and the MM−PBSA calculation evaluation was performed on each of them. The MM−PBSA free energy of each snapshot ($G_{tot}$) is approximated as the sum of two terms: the internal energy of the protein ($E_{MM}$) and a solvation free

[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] Lazaridis, T.; Karplus, M. *J. Mol. Biol.* **1999**, *288*, 477−487.

[10] Park, B. H.; Levitt, M. *J. Mol. Biol.* **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, R. L., Jr. *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. Comput. Phys.* **19**77, *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*; University of California, San Francisco: San Francisco, CA; **1997**.

[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. Crystallogr.* **26 1993**, 283−291.

[19] Ferrin, T. E.; Huang, C. C.; Jarvis, L. E.; Langridge, R. *J. Mol. Graph.* **1988**, *6*, 13−27.

**Table 1.** Summary of the Molecular Dynamics Results[a]

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

[a] The trajectories were clustered, giving rise to conformational families for some of the models. All values except for the initial RMSDs and Rosetta scores are average values over the dynamics. [b] 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. [c] The HP-36 core region comprises residues 6−33, where the average mainchain *B* value in the NMR structure is 0.68, as compared to 1.53 outside the core. [d] A contact is defined as any two residues containing atoms ≤ 3.5 Å apart. There were 89 native contacts in 1vii (HP-36) and 221 in 1a32 (S15). [e] 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; 54, in 1a32. [f] The average $G_{\text{tot}}$ is relative to the native's average. Only 18 (270−1600) had an average value comparable to the native's value, with $P = 0.31$.

energy ($\Delta G_{\text{solv}}$).

$$G_{\text{tot}} = E_{\text{MM}} + \Delta G_{\text{solv}} \tag{1}$$

$E_{\text{MM}}$ is the sum of an internal strain energy ($E_{\text{int}}$), a van der Waals energy (vdW), and an electrostatic energy (EEL). $E_{\text{int}}$ 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 ($\text{vdW}_{1-4}$ and $\text{EEL}_{1-4}$), and long-range values ($\text{vdW}_{\text{NB}}$ and $\text{EEL}_{\text{NB}}$).

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 ($TS_{\text{solute}}$). 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 because 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 that the water structures do not converge.

$$\Delta G_{\text{solv}} = <\Delta G_{\text{solv,NP}}> + <\Delta G_{\text{solv,elec}}> \approx (\gamma \cdot \text{SASA} + \text{b}) + <\Delta G_{\text{solv,elec}}> \tag{2}$$

The nonpolar solvation free energy ($\Delta G_{\text{solv,NP}}$) 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 $\gamma$ coefficient is set to 5.42 cal/mol·Å$^2$, and b is set to 920 cal/mol. The electrostatic solvation free energy ($\Delta G_{\text{solv,elec}}$) is the cost of charging the discharged solute into 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 a 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 $\Delta G_{\text{solv,elec}}$ calculation, which was still amply sufficient, because we found the values in this system reached 90% convergence at ∼50 iterations.

## Results and Discussion

**Rosetta Results on HP-36 and S15.** The Rosetta method, as previously described,[24] rapidly generates ∼1000 structure predictions having centroid side chains 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 $C_\alpha$ RMSD (Table 1). These four were selected because 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 $C_\alpha$ RMSD (Table 1). For this protein, we screened the 100 best-scoring Rosetta models for those with a $C_\alpha$ 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 consideration of 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

[20] Srinivasan, J.; Cheatham, T. E., III; 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.

best Rosetta scoring conformations had RMSDs 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 α-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 $C_\alpha$ RMSD of 5.85 Å over those residues, although the model submitted as first by the Bryant group[25] had a much higher $C_\alpha$ 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 $C_\alpha$ RMSD of 7.27 Å, whereas the model submitted as their first had a $C_\alpha$ 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 RMSDs 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 nontrivial topology, which according to the results of CASP III, appears 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 residue one). The 86-mer S15 forms four helices in the X-ray structure (1a32),[28] although the first 21 N-terminal residues including the N-terminal helix are very disordered and are 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, which led to small deviations (<1 Å $C_\alpha$ RMSD) from the experimental coordinates. During the subsequent control simulations of the equilibrated HP-36 NMR structure, the all-residue $C_\alpha$ 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 $C_\alpha$ RMSD was 1.20 Å, with a standard deviation of 0.16 Å (Figure 2B). The difference in these $C_\alpha$ RMSDs 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 $C_\alpha$ 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.

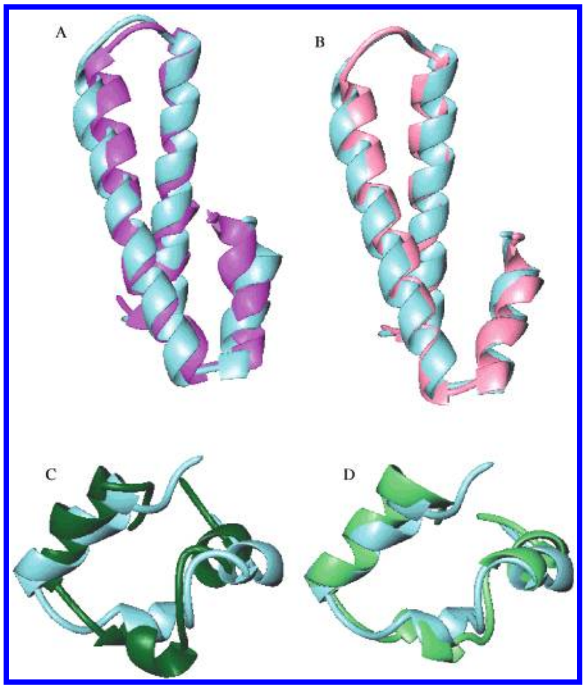

**Figure 1.** Cartoon diagram comparisons of the experimental structures (shown in gray) with the best ab initio predictions in this study. S15 from the simulation of Rosetta model 156 at 0 ps in magenta (a) and at 750 ps, the lowest $C_\alpha$ RMSD structure (1.39 Å), in pink (b). HP-36 from the simulation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

**Simulations on Rosetta HP-36 Predictions.** We ran approximately 1 ns of molecular dynamics on each of the HP-36 Rosetta models and clustered the results in Table 1. During the dynamics, only model 18 underwent a conformational transition (Figure 2), with the new family $18_{(270-1600)}$ having an average core region $C_\alpha$ 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 ($P = 0.31$), while the free energy for the other three simulations remained 15 kcal/mol or more higher than the native state's ($P < 0.001$). 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's having a free energy comparable to that of the native state.

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

[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.

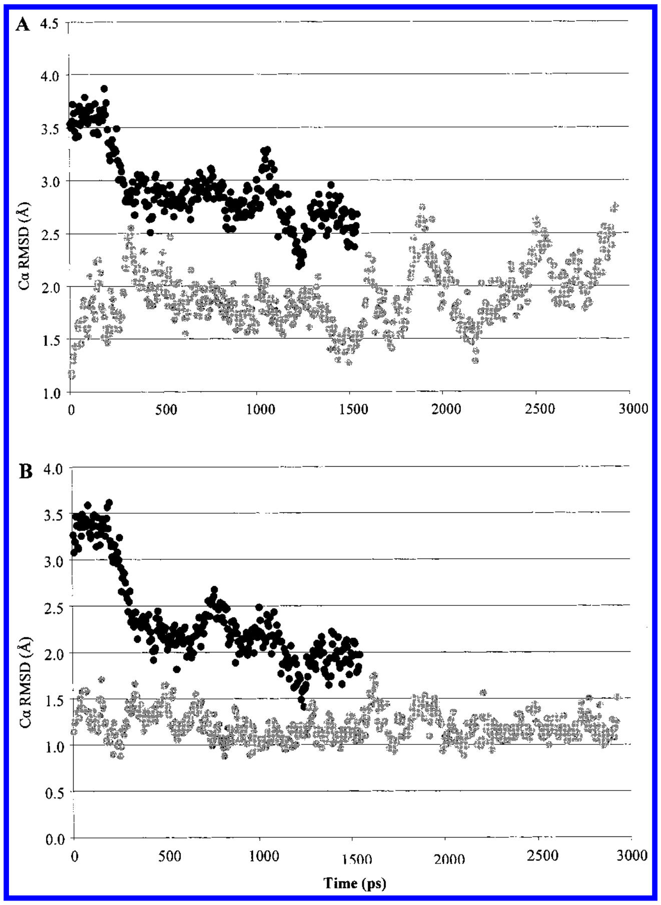

**Figure 2.** Timecourse of the $C_\alpha$ RMSD of HP-36 vs the NMR structure, resulting from molecular dynamics simulations in explicit water, starting with the NMR structure (gray circles) or Rosetta model 18 (black circles). (A) shows the $C_\alpha$ RMSD over all residues and (B) shows the $C_\alpha$ RMSD over the core region (6−33).

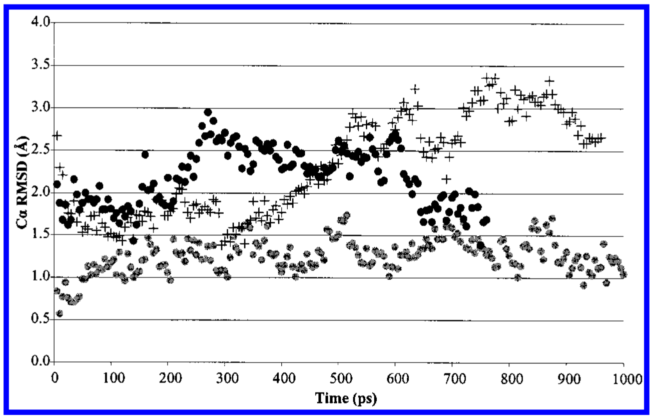

**Figure 3.** Timecourse of the $C_\alpha$ RMSD of S15 vs the X-ray structure, resulting from molecular dynamics simulations in explicit water, starting with X-ray structure (gray circles), Rosetta model 156 (black circles) or Rosetta model 471 (+).

folding reaction, then the lack of structural improvement in the other three models may have been due to their inability to increase the number of native contacts in the 1 ns time range.

In the one *μ*s folding simulation of HP-36 by Duan and Kollman,[5] secondary structure differed markedly between the native control simulation and every non-native structure, because 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 so good a measure of progress toward the native free energy basin as the number of native contacts.

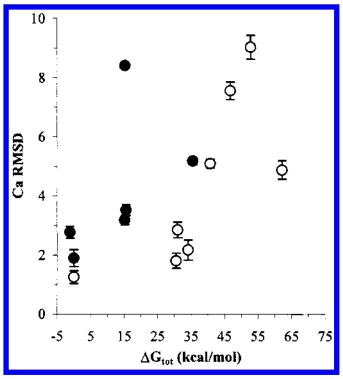

**Figure 4.** Plot showing correlation between average values $C_\alpha$ RMSD and $\Delta G_{\text{tot}}$ for HP-36 (●) and S15 (○), with each data point representing a separate conformational family.

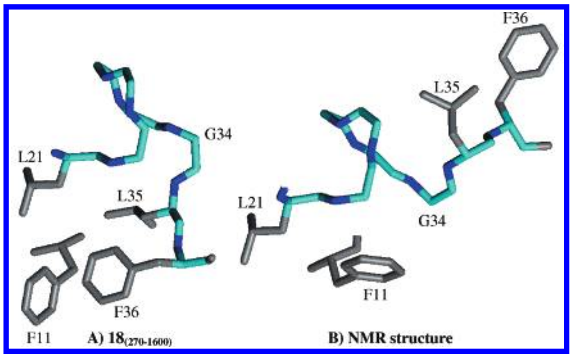

**Figure 5.** Illustration of the C-termini, residues 21−36, demonstrating the region of greatest geometric disparity between the average structure from the $18_{(270-1600)}$ low-energy state (A) and the NMR structure (B). For clarity, only the hydrophobic side chains are shown, together with the backbone 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.

**Simulations on Rosetta S15 Predictions.** Simulations on the five Rosetta S15 models were also carried out for close to 1 ns (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 seven structural families possesses an average free energy comparable to that of the native state ($P < 0.001$), 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 $C_\alpha$ RMSDs of 156 and of 471, prior to their 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 (Figure 1b) and 1.38 Å.

The same topological trends were observed for the S15 models as for 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 of the Rosetta models and showed little correlation with RMSD.

**Interpreting the Energies.** Like both of the native states, the HP-36 low-energy-state $18_{(270-1600)}$ 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

**Table 2.** Comparison of the Energy Components[a]

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

[a] All values are in kcal/mol, are averages for the structural family, and are relative to the native states. [b] Internal strain energy associated with bond, angle, and dihedral motions away from their reference values. [c] Intraprotein Lennard−Jones potential energy. [d] Nonpolar contribution to the solvation free energy. [e] Sum of intraprotein Coulombic energy and electrostatic element of the solvation free energy. [f] Absolute values for HP-36 $E_{\text{vdW}}$ and $\Delta G_{\text{solv,NP}}$ are −113.3 and 18.2 kcal/mol, respectively. [g] Absolute values for S15 $E_{\text{vdW}}$ and $\Delta G_{\text{solv,NP}}$ are −255.9 and 29.5 kcal/mol, respectively.

shifted after ∼500 ps into a separate family in which the free energy was not statistically different from that of the initial family, and the geometric similarity to the experimental structure was noticeably diminished. Because the free energy of $471_{(270-1600)}$ was ∼30 kcal/mol higher, it is not unexpected that it, unlike the native, would transition into another state. These observations reflect the nonlinear relationship between $C_\alpha$ RMSD and $G_{\text{tot}}$ 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 4 shows the relationship between $C_\alpha$ RMSD and $\Delta G_{\text{tot}}$ for the two proteins investigated in this work. As mentioned in the Introduction, because the conformational entropy and, thus, $\Delta G_{\text{tot}}$ (which does not account for $S_{\text{conf}}$) 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 good association for both S15 and HP-36 between $C_\alpha$ RMSD and $\Delta G_{\text{tot}}$. Given their sample sizes, the $r_s$ value for S15 exceeds the critical level for rejecting the null hypothesis of no relationship with $P < 0.02$ and the $r_s$ value for HP-36 exceeds that for a $P < 0.2$. 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, which further corroborates the hypothesis that the energy gap between the 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 sets apart the native state from the non-native states, which likely can only be properly achieved by precise packing of the side chains. All of the S15 model simulations had van der Waals energies ≥30 kcal/mol higher than the native state, with this term also being the dominant component separating the two best MM−PBSA scorers, 156 and 471, from the native state. 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 large 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. Although 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 the best van der Waals energy among all of the HP-36 conformational states.

The fact that HP-36 $18_{(270-1600)}$ and the native state lie at the same free energy level is rather intriguing. Table 2 suggests that although 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 a weaker 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 5 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 most highly disordered monomers, are almost completely solvent-exposed, thus forming a separate miniature hydrophobic cluster. In contrast, the average structure from the $18_{(270-1600)}$ low-energy state has the L35 and F36 side chains 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 $18_{(270-1600)}$ 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 $\Delta G_{\text{elec}}$ 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 $C_\alpha$ RMSDs from every 10th ps over the first 150 ps ($n = 15$) give good agreement with the averages taken from every 5th ps over the entire window (Table 3). With the method described in this work, one can pragmatically rank 5−10 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 to 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 1 ns of simulation time, as we did for each of the model predictions in this study, one can expect to spend upward 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[a] |
|---|---|
| **HP-36** | |
| $17_{(0-735)}$ | 0.43 |
| $18_{(0-270)}$ | 0.87 |
| $54_{(0-960)}$ | 0.87 |
| $60_{(0-935)}$ | 0.09 |
| $\text{native}_{(0-3000)}$ | 0.05 |
| **S15** | |
| $0_{(0-855)}$ | 0.90 |
| $43_{(0-200)}$ | 0.78 |
| $112_{(0-775)}$ | 0.35 |
| $156_{(0-760)}$ | 0.09 |
| $471_{(0-500)}$ | 0.34 |
| $\text{native}_{(0-1000)}$ | 0.79 |

[a] The *P* values are for comparison of MM-PBSA averages that result from postprocessing either the first 150 ps every 10th ps or the entire initial conformational family every 5th 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. Second, 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

Because 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 that methods such as this may allow for a significant advance in CASP IV, as 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 took sequence information of two small proteins and found structures that lie only 1.4 Å $C_\alpha$ RMSD from the experimental structures. These geometrically best conformations are members of conformational families that have both the lowest average $C_\alpha$ RMSD and 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 that is consistently more favorable in the native state than in all other states. Although 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] of protein stability in suggesting that electrostatics have a poor correlation with the MM−PBSA free energy, the rank of which correlates well with the $C_\alpha$ RMSD.

Although 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 a 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 there 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.

**Acknowledgment.** P.A.K. is pleased to acknowledge research support from the NIH (GM-29072) and for supercomputer time through the NSF-supported centers (NCSA and SDSC).

JA003150I

[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, C. L.; 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.