J|A|C|S
COMMUNICATIONS

Published on Web 08/28/2002

All-Atom Structure Prediction and Folding Simulations of a Stable Protein

Carlos Simmerling,*, † Bentley Strockbine,† and Adrian E. Roitberg*,‡

Center for Structural Biology and Department of Chemistry, State University of New York - Stony Brook, Stony Brook, New York 11794, and Quantum Theory Project and Department of Chemistry, University of Florida, Box 118435, Gainesville, Florida 32611

Received June 20, 2002

Recent advances in computer technology and new modeling techniques have facilitated simulations of peptide folding in atomic detail.[1] Here, we present results from all-atom, fully unrestrained folding simulations for a stable protein with nontrivial secondary structure elements and a hydrophobic core.

"Trpcage", a 20 residue sequence optimized by the Andersen group at University of Washington,[2] is currently the smallest protein displaying two-state folding properties. The size of this construct and presence of protein-like features mark the design of this miniprotein as a significant milestone.[3] We agree with suggestions[2] made by Neidigh et al. that this provides an ideal model system for folding simulations. Our theoretical work was assessed by the Andersen group prior to the release of the experimentally determined coordinates. Compared over the well-defined regions of the experimental structure, our prediction has a remarkably low 0.97 Å C_α root-mean-square-deviation (rmsd) and 1.4 Å for all heavy atoms. The simulations suggest additional features that are consistent with experiments but not evident in the NMR-derived structures.

We initiated our simulations using only the trpcage TC5b[2] amino acid sequence (N^{20}LYIQWLKDGGPSSGRPPPS^{39}), with an extended initial conformation built by the LEaP module of AMBER version 6.0.[4] All molecular dynamics (MD) simulations were fully unrestrained and carried out in the canonical ensemble using the SANDER module, which we modified to improve performance on the Linux/Intel PC cluster that was used for all calculations. The ff99 force field[5] was employed, with the exception of ϕ/ψ dihedral parameters which were refit[6] (see Supporting Information) to improve agreement with ab initio relative energies[7] of alanine tetrapeptide conformations. Parameters were not fit to data for the trpcage. Solvation effects were incorporated using the Generalized Born model,[8] as implemented[9] in AMBER.

Folding of peptides of similar length have been simulated when the experimental structure is available to determine convergence.[10] When an experimentally determined structure is not available, it is difficult to evaluate the quality of conformations sampled during a simulation. The convergence of predictions from multiple simulations is a reasonable approach to identify a "folded" state, but this can be misleading if the protein is not completely structured at the temperature of interest. We decided to also monitor potential energy (including solvation free energy) during the simulations. MD simulations of 100 ns were performed at 300 K, but all were kinetically trapped on this time scale, showing strong dependence on initial conditions and failing to converge to similar conformational ensembles. We therefore increased the temperature to 325 K. The potential energy as a function of time during this simulation is shown in Figure 1a. A decrease of approximately 40 kcal/mol is seen over the course of ~10 ns, after which no further improvement is noted. Two independent simulations converged to essentially identical families of structures after 5 and 20 ns.

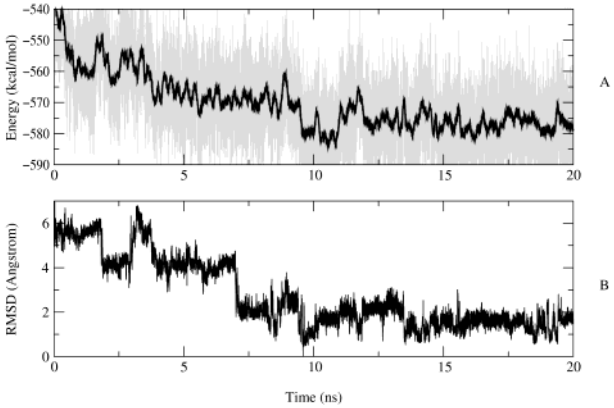

Figure 1. (A) Potential energy of the trpcage as a function of time during MD. The solid line is a running average over 10 ps. (B) Backbone rmsd during the same MD, compared to the lowest-energy conformation.

We assigned this family as our "folded" state, and selected the snapshot with the lowest potential energy across the simulation as our representative structure. In Figure 1b, we show the backbone rmsd *relative to this structure* during the course of the same simulation from Figure 1a. A clear correlation between energy and rmsd is present; the energy plateau is reached at the same time as the convergence to the final structure family with rmsd values of ~1−2 Å. The simulation was extended to 50 ns, and no significant change in energy or rmsd profiles was observed.

Since folding was not reversible during these simulations, we performed a 20-ns simulation at 400 K which showed extensive sampling of conformations with rmsd values from <1.0 to 7 Å; even under these conditions the "native" family was transiently located on six different occasions and was the lowest energy sampled, although it comprised only 3% of all structures at this elevated *T*. These provide additional evidence that the 325 K simulations are not trapped in high-energy basins.

On the basis of this analysis, the low-energy structure was given to the experimental group as our prediction prior to the release of the coordinates of their family of 38 NMR-based models. The NMR-based coordinates are now available (PDB code 1L2Y), and the similarity of the NMR models to our low-energy snapshot is quite remarkable (Figure 2). NMR and theoretical structures share all of the following characteristics: residues 21−27 form a short α-helix, a single turn of 3_{10} helix is present at residues 30−33, and the rest of the chain wraps back along the helical axis toward the N terminus of the chain. The indole ring of Trp25 forms the center of a hydrophobic core, flanked by the side chains of Tyr22, Leu26, and two extrahelical nonneighboring prolines (31 and 37).

* To whom correspondence should be addressed. E-mail: carlos.simmerling@sunysb.edu.
† State University of New York - Stony Brook.
‡ University of Florida.

10.1021/ja0273851 CCC: $22.00

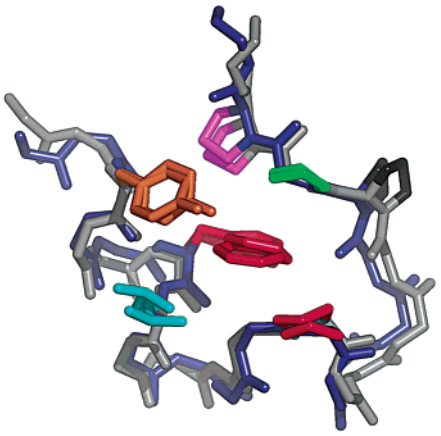

Figure 2. Trpcage heavy-atom structures: low-energy MD (blue backbone) and NMR (gray backbone). Only side chains forming the trp cage are shown, using the same color scheme as in ref 1.

The Pro_3 triplet exhibits a polyproline II helix (which is the first nativelike element established during the simulations, reducing the entropic penalty for formation of the cage), with the central Pro37 forming part of the cage. Two unusual intramolecular hydrogen bonds are present, between the Trp25 indole NHϵ1 and the backbone carbonyl of residue $i + 10$ (Arg35), and between Gly30 HN and the carbonyl of residue i - 5 (Trp25). After folding, both are highly populated in MD (92 and 75%, respectively).

Neglecting the first and last residues and three side chains (all poorly defined in the NMR models, discussed below), the heavy-atom rmsd between the experimental model and our low-energy structure is a remarkably low 1.4 Å. Due to the large fluctuations observed using the continuum solvent, we carried out refinement of our model using 2-ns, 300-K MD in explicit water. This resulted in further improvement, and the average over the final 500 ps has a heavy-atom rmsd of only 1.1 Å compared to the NMR model.

In analogy to calculations reported[2] for the experimental model, we performed ring current shift calculations for our structure using SHIFTS[11] 4.1 (http://www.scripps.edu/case). The correspondence between the chemical shift deviations (CSDs) for theoretical and NMR-based models is excellent, with root-mean-square error of 0.22 ppm and correlation coefficient of 0.99 for the two data sets. These include the highly stereospecific CSDs for Gly30 Hα ($-3.43/-0.96$ for model 1 and $-3.00/-0.54$ for our structure), but we have excluded the outlier Pro37, which is in close contact with Trp25. The experimental structures assigned these prolines in the down pucker based on NOE restraints,[12] and result in a Hβ3 shift of 0.34 ppm. During our simulations however, the down and up puckers are nearly equally populated with rapid sub-ns exchange, and representative structures give Hβ3 shifts of -0.22 and 1.22 ppm, respectively. This issue highlights the interplay between simulation and experiment, and further analysis is underway.

Consistent with the NMR-based models, the charged terminal residues sample multiple conformations during the simulations. The side chains of residues Leu21, Lys27, and Arg35 are also not well defined in the NMR-based models. Correspondence between model variation and simulation flexibility is also observed for both Leu21 and Lys27. In contrast, while Arg35 shows large model variation for nearly all χ dihedral angles, this region exhibits relatively small fluctutations during the simulation. Closer examination of the MD data revealed that Arg35 participates in a solvent-exposed salt bridge with the γ-carboxyl group of Asp28. The pairing was stable but transiently lost on multiple occasions in both continuum and explicit solvents. In this case the simulations likely provide the more reliable picture; a lack of NOEs and absence of prochiral assignments for Arg25 β and γ protons may have led to the poor convergence[12] of the NMR-based models. In fact, creation of this salt bridge was the motivation for mutating to these residues during trpcage design, and pH titration experiments show a large stability dependence coupled to Asp28 protonation.[2] The only remaining significant difference between our model and that determined by NMR is in the side chain of Leu26; we are currently investigating this issue.

One native simulation unfolded, resulting in loss of all elements of the hydrophobic core except a Trp25-Pro31 pair. A reduction in distances between the indole ring and Gly30/Pro31 is observed, consistent with experimental evidence for more negative CSDs at this *T*. Due to the complex nature of the unfolded ensemble, further simulation and analysis are warranted.

Experimental data also suggests that a 16-residue sequence obtained from truncation of the C-terminal PPPS in trpcage does not significantly populate a single fold;[2] a 40-ns simulation of this construct did not converge to any single structure, further strengthening the hypothesis that the cage motif contributes to the stability of this protein. A more detailed analysis of these complex trajectories will be presented elsewhere.

While the CASP competitions[13] offer the opportunity for verifiable blind predictions of protein structure, we undertook this study due to the creation of the small and unusually stable miniprotein. The simulations we have described did not include any structural or other experimental data for the trpcage but still converged to a highly similar family of conformations. In addition, our simulations suggest plausible structural details beyond those available from NMR models, such as the Asp-Arg salt bridge. This demonstrates that MD simulations have reached the point where accurate structure refinement and prediction through direct simulation are not only becoming possible but may soon also be routine enough to contribute significantly to our understanding of the factors that determine folding. Extension to larger systems is a challenging step for the future.

Acknowledgment. We are grateful to Niels Andersen for useful discussions evaluating our prediction prior to the release of the NMR-based coordinates and for helping bridge the size gap between theory and experiment. Funding for this research was provided by the AMDeC Foundation of New York City, through its "Tartikoff/Perelman/EIF fund for Young Investigators in Women's Cancers", NIH GM6167803 and the University of Florida.

Supporting Information Available: Force field parameters (PDF). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Shea, J. E.; Brooks, C. L. *Annu. Rev. Phys. Chem.* **2001**, *52*, 499−535.
(2) Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. *Nat. Struct. Biol.* **2002**, *9*, 425−30.
(3) Gellman, S. H.; Woolfson, D. N. *Nat. Struct. Biol.* **2002**, *9*, 408−10.
(4) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E., III; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Tsui, V.; Radmer, R. J.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G. L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. *AMBER 6*; University of California, San Francisco, 1999.
(5) Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049−1074.
(6) Strockbine, B. A.; Simmerling, C. L. Unpublished Data.
(7) Beachy, M. D.; Chasman, D.; Murphy, R. B.; Halgren, T. A.; Friesner, R. A. *J. Am. Chem. Soc.* **1997**, *119*, 5908−5920.
(8) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.
(9) Tsui, V.; Case, D. A. *Biopolymers* **2000**, *56*, 275−291.
(10) Zagrovic, B.; Sorin, E. J.; Pande, V. *J. Mol. Biol.* **2001**, *313*, 151−169.
(11) Osapay, K.; Case, D. A. *J. Am. Chem. Soc.* **1991**, *113*, 9436−9444.
(12) Andersen, N. H. Personal Communication.
(13) Moult, J.; Fidelis, K.; Zemla, A.; Hubbard, T. *Proteins: Struct., Funct., Genet.* **2001**, 2−7.

JA0273851