ELSEVIER

State of the art in studying protein folding and protein structure prediction using molecular dynamics methods

M.R. Lee, Y. Duan,† P.A. Kollman

Department of Pharmaceutical Chemistry, University of CA, San Francisco, CA, USA

This study presents an overview of the state of the art in using molecular dynamics methods to simulate protein folding and in the end game of protein structure prediction. In principle, these methods should allow the highest level of detail possible and the highest accuracy, but they are limited by both the accuracy of the force field used in the simulation and the sampling possible in the available computer time. We describe current capabilities in running the simulations longer and more efficiently. © 2001 by Elsevier Science Inc.

INTRODUCTION

Among the major challenges in computational chemistry applied to biological systems are developing appropriate methodologies to: simulate the folding of proteins from the unfolded state to the correct native state; and correctly predict the native structure of a protein from its amino acid sequence. These challenges are related, in that predicting the native structure of a protein is a subset of simulating the folding of proteins from the unfolded state. However, using shortcuts to predict the final structure is currently far more feasible than folding the structure from first principles. In fact, the longest simulation to date of first-principle folding for a protein in an explicit box of water molecules is 1 μsec, as described in Duan and Kollman.[1] This required more than two months of computer time using 256 processors on a CRAY T3E. On the other hand, experimentally, the fastest folding protein appears to take 10–20 μsec to fold.[2] Currently, to simulate protein folding, researchers must use simpler than all-atom representations of the protein and solvent and pay the price of less accuracy in representating atomic interactions, or be content with the insights gained in following the folding processes up to the time scale possible with available computer resources. Predicting the native structure of a protein can be addressed by a wide variety of computational approaches represented in the research of scientists such as Baker,[3] Dill,[4] Shakhnovich,[5] Skolnick,[6] Scheraga,[7] and Levitt[8], as well as others too numerous to mention fully. These approaches, such as MONSSTER[6] or ROSETTA[3] often use simpler than the all-atom representations in their simulations. Current options include making the potential surface smoother and carrying out Monte Carlo simulations, e.g., MONSSTER, or using empirical heuristics from known protein structures to guide analysis, e.g., ROSETTA. The possible role for molecular dynamics simulations in these studies is in the "end game," where lower resolution structures found by the simpler approaches[3–8] are used to carry out molecular dynamics trajectories in explicit solvent using a combined molecular mechanics/Poisson Boltzmann-Surface Area (MM-PBSA) methodology to estimate the free energy of such structures. The term "end game" means the process of taking a structure that has the correct native state topology (sequence of helices, sheets, and turns) and is ~3–6Å RMSD from the native state and moving it significantly closer to the correct structure. This term also refers to the ability to distringuish correctly which of two structures is the lowest in free energy. We have used such an approach with the structures of Baker on villin and S15[9] and have also applied MM-PBSA to estimate the free energy along the μs villin trajectory.[10] We have also used a locally enhanced sampling approach[11] to successfully take Skolnick's MONSSTER[6] structure of the small 3-disulfide protein CMTI from 3.7Å RMSD relative to experimental structure to a significantly lower RMSD, improving qualitative features of the structure as well. Below we briefly describe these three studies.

RESULTS AND DISCUSSION

Molecular Dynamics Simulations on Villin

Until Yong Duan was able to parallelize more efficiently the computer program for molecular dynamics[12] and extensive

Corresponding author: P. Kollman, Department of Pharmaceutical Chemistry, University of CA, San Francisco, CA, 94143, USA

E-mail address: pak@cgl.ucsf.edu

†Current address: Department of Chemistry, University of Delaware, DE, USA

Journal of Molecular Graphics and Modelling 19, 146–149, 2001

655 Avenue of the Americas, New York, NY 10010

1093-3263/01/$–see front matter
PII S1093-3263(00)00126-1

dedicated time was available, it made no sense to use all-atom simulations to attempt to fold proteins starting from unfolded states. That is because one could only run trajectories of length in the range of 10 nsec, too short to see any interesting events. Instead, researchers raised the temperature and carried out unfolding simulations[13] to describe folding by analyzing the trajectories backward. However, it is not clear how relevant the simulations at higher temperature are to processes at room temperature or that there is no hysteresis. The chance to get dedicated time on a CRAY T3D, which has an architecture with very rapid and low latency communication between processors, promised by the Pittsburgh Supercomputer Center, inspired Duan to rewrite the molecular dynamics code in AMBER to try to get a high level of parallelism. He succeeded in achieving a speedup of 170× on 256 processors, albeit at the cost of not including the particle mesh Ewald (PME)[14] approach to rigorously represent long-range electrostatic interactions. Thus, the code used no cutoff for protein–protein interactions and a residue-based cutoff for protein–water and water–water interactions. This approach should be adequate for relatively hydrophobic proteins, but would not work well in maintaining the native structure for nucleic acids. After running a 100-nsec simulation control on the native structure to insure that the protocol (i.e., nonbonded cutoffs) still maintained a stable native structure, the temperature was raised to 1000K and the simulation was run for 1 nsec at constant volume; this resulted in a protein with minimal secondary structure or native contacts. After equilibrating the water to bring it back to 300K, the simulation began with this thermally unfolded state of the protein and was extended to 200 nsec. After this we contacted CRAY research (subsequently named SGI/CRAY) to see if we could have considerable time on their corporate T3E computer, which is 4 times faster than the T3D. Howard Pritchard and John Carpenter of SGI/CRAY facilitated our use of the T3E, which enabled the simulation to be carried out to 1 μsec,[2] which was approximately two orders of magnitude longer than any previous simulation of a protein in a periodic box of water molecules. The simulation did not reach the native structure, nor should it have in this time scale. It did go from essentially zero helix to 60% in the 50–100-nsec range, in good agreement with previous experiments on helical proteins, which suggest that helix formation occurs in the 10–100-nsec range. Hydrophobic stabilization, as measured by the Eisenberg-McLachlan solvation free energy,[15] suggested that hydrophobic collapse was occurring on the same time scale as helix formation. The other noteworthy feature of our simulation was that the fluctuations in the RMSD and radius of gyration of the structure were considerable throughout the trajectory, with the exception of the period between 240 and 400 nsec (and a few other shorter periods), where metastable intermediates formed. This shows that proteins don't collapse and stay collapsed while searching for the native structure; rather, they continually collapse and expand as they search for a more stable conformation. The metastable intermediate was the structure closest to native reached during the simulation. Interestingly enough, based on our subsequent MM-PBSA analysis, it was the lowest in free energy reached during the folding trajectory. We calculated the MM-PBSA free energy for the entire folding trajectory, as well as the 100-nsec native trajectory.[10] Encouragingly, the free energies found in the native trajectory were about 30 kcal/mole more stable than the average from the folding trajectory. This supports the idea that the simulation did not fail to reach the native state because it had found a lower energy structure; rather, it had not sampled sufficiently long to reach this native state. As noted above, the metastable intermediate was the lowest in free energy throughout the trajectory, leading to a correlation between RMSD and MM-PBSA free energy. Since then, we have run two additional 500-nsec trajectories on villin starting from different initial geometries. These showed a similar hydrophobic collapse and helix formation in the 10–100-nsec time scale as well as the formation of some metastable intermediate structures, although none as stable as the one found in the original trajectory.[16] We also ran a folding simulation starting with an extended protein chain. Within a few nsec, the protein had formed the "U" shape characteristic of both the native and thermally unfolded one. Based on this result, we surmise that a fully extended chain is clearly a very high energy structure

Use of MM-PBSA in Analysis of Predicted Protein Structures

Encouraged by the reasonableness of the MM-PBSA results for the villin example, we applied it to low-resolution structures with the idea that running molecular dynamics simulations on such structures, after adding all the atoms and putting the molecules in periodic boxes of water, would both drive the structure closer to native and allow the evaluation of the relative free energy of candidate structures. We have carried out three sets of studies to examine this question. In the first set of studies,[9] we used structures from David Baker on both villin and another protein (S15); after taking representative structures from five families of structures described by Baker's ROSETTA, we carried out nsec-length trajectories on each and analyzed the results with MM-PBSA. In both cases, we had significant conformational changes occurring during the molecular dynamics, some of which improved the agreement between calculated and observed RMSD. We also found that the native structure was lowest in free energy, but that the ones closest in RMSD tended to be the next in free energy. These results encourage the use of MM-PBSA in the end game of protein folding. In a second set of studies, we collaborated with the Sali group[17] to carry out shorter molecular dynamics trajectories (150 psec) on one correct model structure and a second reasonable homology model structure, each from 13 proteins, ranging in size from ~60 to ~400 residues. The goal here was to see if MM-PBSA could successfully reproduce which of the two structures was lowest in free energy. This was a blind study, in that we did not know the answer when the calculations were done and we correctly predicted the native in all 13 proteins. However, we are not yet in a position to say, if we do not have the native structure, whether the best MM-PBSA structure is actually the native or just lower in energy. We are investigating ways to try to answer that question. Finally, we are collaborating with the Baker group[18] on some of his ab initio targets for CASP4 contest in 2000 for protein structure prediction, taking his best candidates for the folded structure and trying to rank them with MM-PBSA. We do not yet know whether doing the MM-PBSA analysis will "add value" to these predicted structures. If one of them is within 3–4Å of native, based on our results on villin and S15,[9] we have a good chance of identifying it as best. But if all are 5Å or more from the correct structure, then our chances of identifying the best are less probable.

Use of Locally Enhanced Sampling to Improve Predicted Protein Structures

Can molecular dynamics be used to improve the structure of low-resolution predicted protein structures? We have been successful in doing this in at least one case, using an approach called locally enhanced sampling (LES), which is a mean-field approach to simulations.[19] Part of the system consists of multiple copies of the same part of the topology that do not interact with each other; the other regions see an average interaction from these copies. We have taken the MONSSTER[6] structure of the small, 3-disulfide protein CMTI and have carried out both conventional molecular dynamics and dynamics with LES on the initial structure, using 7 multiple-copy groups of 4 residues each in the LES calculation (the C-terminal group had 5 residues, for a total of 29 in the protein).[11] Whereas the single-copy MD did not change significantly in 2 nsec, the LES simulation dramatically improved the topology of both the C-terminal beta sheet and an internal 3–10 helix. Use of PME with LES was critical in fixing up packing of the β sheet, which contained a high density of charged residues. The final RMSD for the LES simulation was 2.2Å, compared with an initial 3.7Å. A single-copy MD simulation starting with native led to an RMSD of 1.4Å and was 1.8Å from the LES predicted structure. Thus, in this one case, we have shown the utility of LES/MD in protein structure refinement for the end game of protein folding. In aqueous simulations, where most of the atoms are water, using LES costs only 20% more time. In a related study refinement of a RNA hairpin loop,[20] the LES simulation reached the correct structure in 200 psec, whereas the single copy had not done so after 7 nsec; thus, in that application, LES improved the sampling by at least a factor of 35.

SUMMARY

Molecular dynamics simulations can be useful in understanding the initial stages of protein folding, which requires considerable computer power and high levels of parallelism to get to an interesting time scale using conventional molecular dynamics. Thanks to excellent work on parallel MD by Yong Duan and generous gifts of computer time by the Pittsburgh Supercomputer Center and SGI/CRAY, a milestone in protein folding simulations has been achieved, with an increase in simulation length of two orders of magnitude. A second exciting development was the realization that a combined explicit/continuum solvent approach, MM-PBSA,[21] can usefully estimate the relative free energies of complex molecules in solution. We have applied this to analyze the free energies in our villin folding trajectory and predict the lowest free energy structures from a set of known targets[9] or to aid in predicting unknown targets.[18] Finally, LES has been shown to have considerable potential in aiding in the refinement of protein structures. If the initial structure is close enough to the true one, LES can lower energy barriers compared with single-copy molecular dynamics and aid in improving predicted structures for proteins.

ACKNOWLEDGEMENTS

Peter Kollman is grateful to the NIH (GM-29072) for research support and to the NRAC program of the NSF for computer time.

REFERENCES

1 Duan, Y., and Kollman, P. A. Pathways to a folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* 998, **282**, 740–744
2 Burton, R. E., Huang, G.S., Daugherty, M.A., Caldera, T.L., and Oas, T. G. The energy landscape of a fast-folding protein mapped by Ala→Gly substitutions. *Nature Struc. Biol.* 1997, **4**, 305–310
3 Simons, K.T., Bonneau, R., Ruczinski, I., and Baker, D. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins: Structure, Function Genetics* 1999, **Suppl. 3**, 171–176
4 Ishikawa, K., Yue K., and Dill, K. Predicting the structures of 18 peptides uding Geocore. *Protein Sci.* 1999, **8**, 716–721
5 Shimada, J., Ishchenko, A.V., and Shakhnovich, E.I. Analysis of knowledge-based protein-ligand protentials using a self-consistent method. *Protein Sci.* 1999, **9**, 765–775
6 Ortiz, A.R., Kolinski, A., Rotkiewicz, P., Ilkowski, B., and Skolnick, J. Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins: Structure, Function and Genetics* 1999, **Suppl. 3**, 177–185
7 Lee, J., Liwo, A., Ripoli, D.R., Pillardy, J., and Scheraga, H.A. Calculation of protein conformation by blobal optimization of a potential energy function. *Proteins: Structure, Function and Genetics* 1999, **Suppl. 3**, 204–208
8 Samudrala, R., Xiu, Y., Huang, E., and Levitt, M. Ab inito protein structure prediction using a combined hierarchical approach. *Proteins: Structure, Function and Genetics* 1999, **Suppl. 3**, 194–198
9 Lee, M., Baker, D., and Kollman, P.A. Getting 2.1Å and 1.8Å C-alpha RMSD Predictions on Two Small Proteins, HP-36 and S15. *J. Amer. Chem. Soc.* 2001, **123**, 1040–1046
10 Lee, M.R., Duan, Y., and P.A. Kollman. Use of MM-PBSA in estimating the free energies of proteins: Application to native, intermediates and unfolded villin headpiece. *Proteins: Structure, Function and Genetics* 2000, **39**, 309–316
11 Simmerling, C., Lee, M.R., Kolinski, A., Ortiz, A., Skolnick, J., and Kollman, P.A. Combining MONSSTER and LES/PME to predict protein structure from amino acid sequence: Application to the small protein CMTI-1. *J. Amer. Chem. Soc.* 2000, **122**, 8392–8402
12 Duan, Y. Unpublished results
13 Alonso, D.O.V., Alm, E., and Daggett, V. Characterization of the unfolding pathway of the cel-cycle p135UC1 by molecular dynamics simulations: Implications for domain swapping. *Structure Folding Design* 2000, **8**, 101–110
14 Essmann, U., Perara, L., Berkowitz, M., Darden, T., Lee, H., and Pedersen, L.G. A smooth particle mesh Ewald method. *J. Chem. Phys.* 1995, **103**, 8577–8593
15 Eisenberg, D., and Maclachlan, A. Solvation energy in protein folding and binding. *Nature* 1986, **319**, 199–203
16 Duan, Y. Further unpublished results on villin.
17 Lee, M.R., Sali, A., and Kollman, P.A. Unpublished
18 Lee, M.R., Baker, and Kollman, P.A. CASP4 predictions 2000, Asilomar, CA.
19 Simmerling, C., and Elber, R. Hydrophobic collapse in a cyclic hexapeptide-computer simulations of CHDLIC and CAAAAC in water. *J. Amer. Chem. Soc.* 1994, **116**, 2534–2547

20 Simmerling, C., Miller, J.L., and Kollman, P.A. Combined locally enhanced sampling and particle mesh Ewald as a strategy to locate the experimental structure of a non-helical nucleic acid. *J. Amer. Chem. Soc.* 1998, **120**, 7149–7155

21 Srinivasan, J., Cheatham, T.E., Cieplak, P., Kollman, P.A., and Case, D.A. Continuum solvent studies of the stability DNA, RAN and phosphoramidate-DNA helices. *J. Amer. Chem. Soc*. 1998, **120**, 9401–9409