Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102
Volume 23, Issue Number 1, (2005)

An Hypothesis Paper

An Independent Method for the Analysis of Protein Folding Kinetics from All-atom Molecular Dynamics Simulations

Neelan J. Marianayagam[§,*]
Andrew G. Brown
Sophie E. Jackson

Department of Chemistry
University of Cambridge
Cambridge CB2 1EW
United Kingdom

[§]Present address:
Department of Bioengineering
University of California
Berkeley, California 94720, USA

http://www.jbsdonline.com

Abstract

We propose a method for extracting useful kinetic information from all-atom molecular dynamics simulations of protein folding. By calculating the time correlation functions between the evolution of different structural properties during the course of the simulation we can determine the endpoint of the reaction and the mechanism by which it occurs. As a test of our method we use thermal denaturation simulations on a 76 residue protein, ubiquitin. The method we present should be used in combination with current techniques for analyzing molecular dynamics trajectories.

Introduction

For several years, protein folding has been studied *in silico* using high temperature unfolding simulations (1-2). An equilibrated crystal structure is subject to high temperatures, which cause unfolding of the polypeptide chain. From this unfolding trajectory, the folding pathway is gleaned by implementing the principle of detailed balance. However, with the increase in computational power due to both faster processors, more sophisticated clusters and distributed computer environments, it is now possible to directly simulate the protein folding reaction using all-atom models in implicit solvent (3-4). In these cases, the simulations are started from an extended conformation and fold to their native states within computationally feasible timescale. At present these simulations are only possible for small proteins or substructures of larger proteins, however, as computational power increases these techniques will undoubtedly be extended to more complex molecules. As such, it is essential to have the tools to extract the correct kinetic information from these simulations. In order to extract proper kinetic parameters from these simulations, it is important to identify the appropriate endpoint to the reaction. Previous simulation studies have used a change in C-α root mean square deviation (rmsd) from the crystal structure as a criterion for finding the native state (an rmsd of within 2 Å is considered correctly folded) (3-4). The progress of the folding reaction is monitored by the relaxation of the rmsd from a high value in the unfolded state to within 2 Å of the crystal structure, which is considered the folded state. However, recent analysis of crystal structures shows that this reasoning may not be accurate since the crystal structures themselves are just a snapshot of a highly degenerate state (5). In this study, we propose an independent and unambiguous method for determining the endpoint of these simulations, by utilizing the time correlation function formalism. Many all-atom molecular dynamics simulations for protein unfolding monitor the progress of the reaction by probing two or more distinct structural properties over time (6). One example of this is monitoring the variation in rmsd and solvent-accessible surface area (sasa) (7). In this study, we propose investigating the correlation of these structural properties over time by calculating their cross-correlation functions.

Phone: +1 510 486 5077
Fax: +1 510 486 6488
Email: neelanm@yahoo.com

Time correlation functions have been used to investigate many systems ranging from the condensed phase to the evolutionary dynamics of proteins (8-10). We show that the time correlation functions for structural properties probed during the course of molecular dynamics simulations can give insight into protein folding kinetics.

As a test case, we use thermal unfolding simulations of ubiquitin, a 76-residue protein with a mixed α/β structure. High temperature molecular dynamics simulations have been used to probe the folding/unfolding pathway of this protein (7). We explored the folding kinetics by probing two distinct structural properties over the time course of the simulation Cα rmsd and hydrophobic solvent accessible surface area (sasa) (7). In this work, we calculate time correlation functions for the evolution of these properties during the course of the simulations. In the unfolded state, these values are highly uncorrelated while in the folded state these values have a high degree of correlation. We give recommendations for how this method could be easily applied to room temperature folding simulations. We propose that this type of analysis will provide an independent method for determining the endpoint of a folding simulation and should be used in conjunction with the current methods of analysis (4).

Methods

The methodology for the 7 all-atom molecular dynamics simulations of ubiquitin have been described elsewhere (7). In this section we describe the methods employed in the correlation-function analysis. In this study we carry out the analysis by averaging over the 7 trajectories used in reference 7, this was done so that direct comparisons could be made to the differing methods of analysis. It will be shown that the analysis presented here substantiates the conclusions drawn in reference 7 as to the folding mechanism of ubiquitin. However to check the consistency of our method, we carried out averages over three of the trajectories and these results were similar to the results calculated from all 7 simulations (data not shown).

Qualitatively, it would appear that sasa and rmsd are good reaction coordinates with which to follow the folding reaction since they probe distinct elements of structure. However, to quantify their suitability we calculate generalized correlation coefficients for sasa and rmsd (11).

$$c_{RS} = \frac{\left\langle \left(R_i - \langle R \rangle\right)\left(S_i - \langle S \rangle\right) \right\rangle}{\left(\left\langle \left(R_i - \langle R \rangle\right)^2 \right\rangle \left\langle \left(S_i - \langle S \rangle\right)^2 \right\rangle \right)^{0.5}} \qquad [1]$$

where R_i and S_i are the values for rmsd and sasa, respectively, at the *i*th time point of the simulation and quantities within '<>' represent average values. Suitable reaction coordinates should have high values of c indicating that they change concertedly through the course of the simulation (13). We calculate an average value of c_{RS} over the seven simulations and find a correlation coefficient of 0.84, indicating that sasa and rmsd follow each other well through the course of the simulation.

We then calculate correlation functions between these two properties using the following equation (11, 12):

$$B(t) = \left\langle R(t')S(t' + t) \right\rangle \qquad [2]$$

where $R(t')$ is the value at time t' in the simulation and $S(t' + t)$ is the value of S at some other time point in the simulation. The average in Eq. [2] is taken over multiple time origins. We also calculate autocorrelation functions for R and S (data not shown).

Results and Discussion

In Figure 1 we show plots of the normalized cross-correlation functions of R and S from three of the simulations. Figure 2 shows cross-correlation function averaged

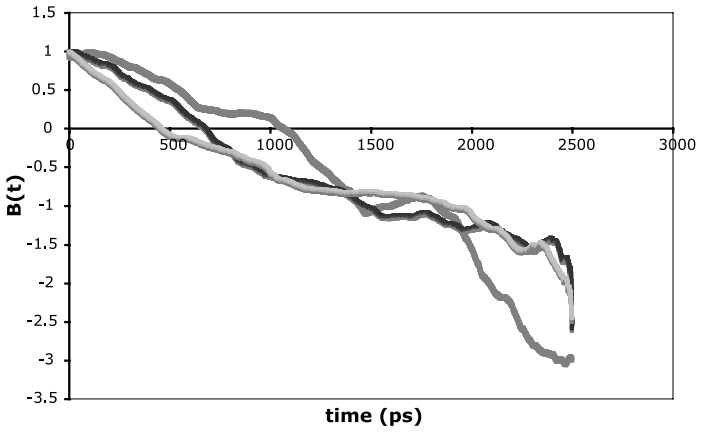

Figure 1: Time correlations functions from three of the simulations.

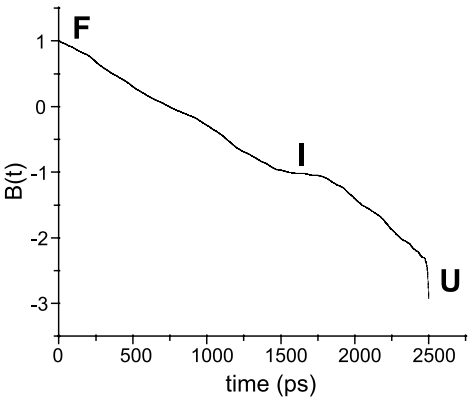

Figure 2: Average correlation function over all seven thermal unfolding simulations.

over the seven simulations (the ensemble average). The species populated on the pathway are shown in bold beginning with the folded state (**F**), intermediate state (**I**), and unfolded state (**U**). The time points for the population of **I** and **U** are consistent with those from our previous analysis (7). As the simulations progress there is loss of correlation, the time correlation function achieves negative values towards the end indicating the two structural properties become uncorrelated. It is interesting to note that the average correlation coefficient calculated using Eq. [1] shows high degree of correlation between R and S. It is worth pointing out the differences between Eq. [1] and [2]. The correlation coefficient shows how well the two properties follow each other throughout the course of the simulation, that is do the properties change in a concerted manner (13). Equation [1] is essentially a method of determining the suitability of the reaction coordinate (13). Equation [2] is a measure of the dynamics of the system, as time progresses do the quantities remain close to their initial values or is there some deviation?

There are two distinct phases in the correlation functions calculated in this work. An initial relaxation to 1500 ps and then a second, faster relaxation to yield negative values of B. This indicates two distinct kinetic processes and a possible change in rate determining step which points to the existence of an intermediate state. The accumulation of an intermediate on the folding pathway of this protein has been shown by both simulation and experiment (7, 14). As the unfolded state is populated, the values of *R* and *S* become highly uncorrelated, indicating the high degree of structural heterogeneity in the unfolded polypeptide chain. The autocorrelation functions for *R* and *S* do not show this type of behavior (data not shown), indicating this type of analysis is suitable to two distinct structural properties.

We fit the correlation functions to an exponential decay process. The data fit best to double exponential functions (R = 0.95). This type of analysis can serve a twofold purpose. First, it gives insight into a possible folding mechanism, the presence of two kinetic phases would provide evidence for an intermediate species being populated on the folding pathway. Second, rate constants for the formation of both intermediate and the native state can also be obtained. These can be compared to the experimental rates.

The analysis detailed above can be extended to other structural properties. In Table I we list pairs of structural properties that can be used to monitor folding, their expected values in the folded and unfolded state, expected correlation coefficients and expected values of the correlation functions in unfolded and folded states. We can now generalize a procedure for applying the correlation function formalism to room temperature protein folding simulations: I. Pick two distinct structural properties that vary with time during the simulation. It is important to choose properties that probe different aspects of structure. In this study we chose rmsd (which looks at secondary structure) and hydrophobic sasa, which is a probe of tertiary structure.

II. Calculate correlation coefficients to see if the structural properties change concertedly through the course of the simulation. III. Calculate the cross correlation functions and plot these against time. These correlation functions should provide an appropriate end-point to the simulation, upon convergence. IV. Analyze the functional form of the plots: are they homogeneous or are there multiple phases? This type of analysis will aid in the determination of a folding mechanism and fitting to the appropriate functions will yield rate constants. We do not extract rate constants in our test case since simulations were conducted at high temperatures and, therefore, cannot be compared to experimental rate constants measured at significantly lower temperatures. V. Compare conclusions extracted from the time correlation formalism to those gained from the traditional methods of analysis.

Table I
Expectation values for the time correlation functions of various properties used to probe folding.

Properties	Values in unfolded state	Values in folded state	Correlation coefficients	B(t) Unfolded	B(t) Folded
R_g/Number of Native contacts	high/low	low/high	uncorrelated	1	<0
rmsd/R_g	high/high	low/low	correlated	<0	1
rmsd/number of hydrogen bonds	high/low	low/high	uncorrelated	1	<0

Conclusions

In this work, we have presented an independent method for determining the end-point of all-atom molecular dynamics protein folding simulations. The technique involves utilization of the relaxation of the time-correlation function of structural properties that are used to follow folding. We validate this method by calculating time correlation functions of rmsd and sasa from high temperature unfolding simulations of ubiquitin. The techniques we implement here can be easily extended to the robust folding simulations that are being carried out on the massive distributed computing clusters (4). This method should be used in conjunction with the other techniques to extract the most reliable information possible from these folding simulations. While the current techniques for extracting kinetic information from the simulations yield experimentally verifiable rate constants for small systems, it is expected that for larger proteins there will greater uncertainties due to the higher degree of degeneracy in the crystal structures (5). The method presented here will be invaluable in reducing this ambiguity.

Acknowledgements

The authors would like to thank the Welton Foundation for funding during the course of this work. N.J.M. would like to thank Dr. Robert Best and Abhirami Ratnakumar for helpful discussions.

References and Footnotes

1. A. Li and V. Daggett. *Proc Natl Acad Sci USA 91*, 10430 (1994).
2. J. E. Shea and C. L. Brooks, 3rd. *Annu Rev Phys Chem 52*, 499 (2001).
3. C. D. Snow, H. Nguyen, V. S. Pande, and M. Gruebele. *Nature 420*, 102 (2002).
4. V. S. Pande, I. Baker, J. Chapman, S. P. Elmer, S. Khaliq, S. M. Larson, Y. M. Rhee, M. R. Shirts, C. D. Snow, E. J. Sorin, and B. Zagrovic. *Biopolymers 68*, 91 (2003).
5. M. A. DePristo, P. I. de Bakker, and T. L. Blundell. *Structure (Camb) 12*, 831 (2004).
6. T. Lazaridis and M. Karplus. *Science 278*, 1928 (1997).
7. N. J. Marianayagam and S. E. Jackson. *Biophys. Chem. 111*, 159 (2004).
8. H. C. Andersen. *J. Phys. Chem. B 106*, 8326 (2002).
9. G. Arya, E. J. Maginn, and H. Chang. *J. Chem. Phys. 113*, 2079 (2000).
10. N. V. Dokholyan and E. I. Shakhnovich. *J Mol Biol 312*, 289 (2001).
11. M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford University Press (1987).
12. R. Zwanzig. *Annu Rev Phys Chem 16*, 67 (1965).
13. R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich. *J. Chem. Phys. 108* 334 (1998).
14. H. M. Went, C. G. Benitez-Cardoza, and S. E. Jackson. *FEBS Lett 567*, 333 (2004).

Date Received: March 8, 2005

Communicated by the Editor Ramaswamy H Sarma