J|A|C|S COMMUNICATIONS

Published on Web 03/03/2009

Toward an Accurate Determination of Free Energy Landscapes in Solution States of Proteins

Alfonso De Simone,[†] Barbara Richter,[†] Xavier Salvatella,[‡] and Michele Vendruscolo*,[†]

Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, U.K., and ICREA and Institute for Research in Biomedicine, Baldiri Reixac 10-12, 08028 Barcelona, Spain

Received November 6, 2008; E-mail: mv245@cam.ac.uk

Downloaded by NESLI CONSORTIA UK on July 8, 2009
Published on March 3, 2009 on http://pubs.acs.org | doi: 10.1021/ja8087295

In the 50 years since the first determination of the structures of proteins, our understanding of the states that they adopt in solution has enormously improved.[1] It is now well-established that proteins populate a wide variety of different states in solution, many of which are conformationally highly heterogeneous.[1−8] Even in their native states, proteins constantly undergo structural fluctuations on time scales ranging from picoseconds to seconds and beyond.[2−8] It is therefore indispensable to achieve an accurate description of the inherent protein dynamics in order to account for biologically relevant processes, including enzymatic activity[6,9,10] and the formation of biomolecular complexes.[8,11] This goal is challenging, especially with respect to the treatment of dynamical regions of proteins, such as loop or unstructured sections, which often play important biological roles. A powerful strategy for characterizing the structure and dynamics of proteins in solution is emerging from methods that combine NMR and molecular dynamics simulations.[5,8,12,13] Although these methods have offered encouraging results, it is still unclear whether they can provide ensembles of structures with the correct equilibrium statistical weights.

Here we address this fundamental question by showing that molecular dynamics simulations with ensemble-averaged restraints[5,12,13] serve as a very accurate tool for calculating the free energies associated with the equilibrium ensembles corresponding to the native states of proteins. We adopted an approach in which AMBER[14] molecular dynamics simulations of ubiquitin are used to generate a collection of structures, forming a reference ensemble representing the state of the protein in solution. NMR observables are then back-calculated from this ensemble and used as restraints in CHARMM[5,12] molecular dynamics simulations aimed at reconstructing the distribution of structures of the reference ensemble. Approaches using reference ensembles have proved to be very powerful in structural biology[12,15] since they allow for an objective cross-validation analysis in which the atomic coordinates of the conformations to be reconstructed are known exactly; thus, both the average structure and the structural heterogeneity obtained from the restrained simulations can be compared with great accuracy to those of the reference ensemble. To define a reference representation of the ubiquitin solution ensemble, we employed the AMBER99SB force field,[16] which has been shown to accurately reproduce the native-state dynamics of ubiquitin;[14] a comparison of the reference-ensemble-calculated and experimentally measured S^2 order parameters and residual dipolar couplings (RDCs) is presented in Figure S1 in the Supporting Information. As structural restraints, we employed RDCs,[17] which are particularly suitable for probing protein structure and dynamics with sensitivity up to the millisecond time scale.[8,13,18,19] We extracted 36 reference RDC data sets from the reference ensemble as the best fit of 36 RDCs that were recently reported.[8] These reference RDCs are thus compatible with the type of experimental data that can be measured with standard aligning media. We used only RDCs for NH bond vectors, which are the most commonly measured ones. The restrained ensemble was generated by adding to the CHARMM22 force field[20] a pseudoenergy term (see the Supporting Information) given by

$$E = \sum_i (D_i^{\mathrm{res}} - D_i^{\mathrm{ref}})^2 \qquad (1)$$

where the sum runs over the restrained (D_i^{res}) and reference (D_i^{ref}) ensemble-averaged RDCs. The restraints were imposed as averages over M replicas of the protein molecule[5,12,13] and employed in annealing cycles (see the Supporting Information for further details); here we used $M = 2$, 4, 8, and 16. The restrained ensemble was sampled at the end of each annealing cycle, where the system was allowed to relax at 300 K.

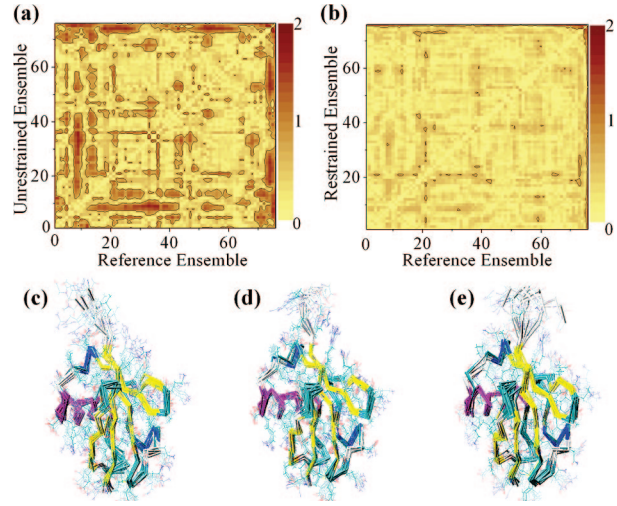

Figure 1. Analysis of the distributions of interatomic distances in the restrained and unrestrained ensembles. The **S** matrix[12] was calculated on the distance distributions of backbone N atoms. (a) Comparison of the unrestrained ensemble (CHARMM22 force field[20]) with the reference ensemble (AMBER99SB force field[16]). (b) Comparison of the restrained ensemble with the reference ensemble. Darker colors indicate regions of lower structural similarity; these regions are mapped on the ubiquitin structures for illustrative purposes by connecting lines in Figure S2. Data are shown for the case of eight replicas; **S** results for ensembles generated by employing 2, 4, 8, and 16 replicas are shown in Tables S1 and S2 and Figure S3. (c, d, e) Representations of the reference ensemble, the unrestrained ensemble, and the restrained ensemble, respectively; secondary-structure elements are shown in yellow (β strands) and purple (α helices).

We first assessed the ability of the restrained simulations to reproduce the structural heterogeneity of the reference ensemble by employing the **S**-matrix method[12] (Figure 1a,b). Since ubiquitin

[†] University of Cambridge.
[‡] ICREA and Institute for Research in Biomedicine.

10.1021/ja8087295 CCC: $40.75 © 2009 American Chemical Society

has 76 residues, **S** is here a 76 × 76 matrix in which each entry s_{ij} represents the difference between the distributions P^{ref} and P^{res} of the distance between the backbone N atoms of residues i and j in the reference and restrained ensembles, respectively:

$$s_{ij} = \sum_k |P^{ref}_{ij,k} - P^{res}_{ij,k}| \quad (2)$$

where k runs over the bins used to plot the distance distributions P^{ref} and P^{res} (see Figure S6). The **S** matrix provides an accurate characterization of both the local and global structural similarity of two protein ensembles. Since each distribution is normalized to 1, the s_{ij} values range from 0 to 2. An s_{ij} value of 0 corresponds to identical distributions, whereas a value of 2 corresponds to completely nonoverlapping distributions. We used the value $s_{ij} = 0.7$ to characterize similar distributions (see Figure S6).

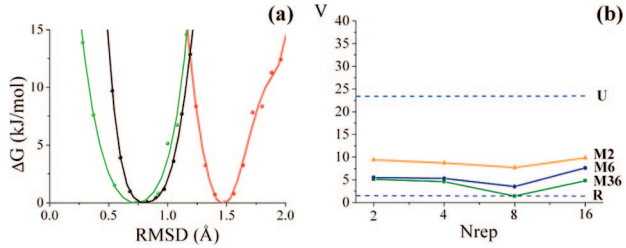

Figure 2. (a) Comparison of the free-energy landscapes of the reference ensemble (green), the unrestrained ensemble (red) and the eight-replica RDC-restrained ensemble (black). Free energies were calculated as a function of the root-mean-square deviation (rmsd) between C^{α} atoms from a representative structure of the reference ensemble (see also Figure S5) using the formula $\Delta G = -kT \log[H(\text{rmsd})]$, where H(rmsd) is the histogram of rmsd values sampled during the simulations. (b) Analysis of the quality (V) of the restrained ensemble as a function of the number of replicas (N_{rep}) used. V is defined as the percentage of pairwise distances with $s_{ij} > 0.7$ (see Table S2). The unrestrained ensemble (U) has the largest value of V, while the reference ensemble (R) has the lowest value, which is essentially identical to that of the restrained ensemble with eight replicas and 36 alignment media (M36); the results with six (M6) and two (M2) alignment media show almost the same quality as those obtained with 36 alignment media, suggesting that fairly accurate reconstructions are also obtained with small numbers of alignment media.

The **S** matrix calculated for the unrestrained and reference ensembles reveals a significant diversity in the two types of sampling, with many regions presenting $s_{ij} > 0.7$ (Figure 1a). This result is not surprising, considering the different parametrizations of the CHARMM22 and AMBER99SB force fields. When the restraining term on the RDC values is added to the CHARMM22 force field (Figure 1b), most of the regions that in the comparison of unrestrained ensembles presented $s_{ij} > 0.7$ are found to have much lower s_{ij} values. This level of accuracy is within statistical errors, as it is comparable to that found in comparing the two halves of the reference ensemble itself (Table S2 and Figure S4).

The results presented here indicate that the combination of NMR restraints with a force field enables the accurate reconstruction of all 2850 distinct distance distributions between pairs of backbone N atoms of ubiquitin. Since the **S** matrix is a stringent measure of similarity between two ensembles,[12] we conclude that the restrained ensemble accurately reconstructs the reference ensemble. In order to verify whether the free energy of the restrained ensemble also closely reproduces that of the reference ensemble, we projected the free energy on the root-mean-square deviation (rmsd) between Cα atoms using a representative structure of the reference ensemble (Figure 2a). The free-energy landscape of the reference ensemble itself has a minimum at 0.75 Å (Figure 2a, green line). In contrast, we found a free-energy minimum at 1.45 Å for the unrestrained CHARMM ensemble (Figure 2a, red line). Thus, the use of restraints enabled an almost complete recovery of the reference free energy, with a minimum at 0.80 Å (Figure 2a, black line). Although the free-energy landscapes in Figure 2a were calculated using 36 alignment media, the use of a smaller number of alignment media (see the Supporting Information) did not compromise the quality of the reconstructed ensembles (Figure 2b) and free energies (Figure S5); the best results were obtained in all cases using eight replicas (Figure 2b and Tables S1 and S2).

In summary, we have presented evidence that the inclusion of ensemble-averaged NMR restraints into molecular dynamics simulations is a strategy that successfully guides the sampling of conformational space toward the ensemble of structures populated by a protein in solution, even if the force field used is not exact. Although here we have used RDCs as restraints and presented the case of ubiquitin, a protein that exhibits rather limited fluctuations in its native state, the method that we have discussed is general and can be implemented for a wide range of NMR observables, including nuclear Overhauser effects,[21] paramagnetic relaxation enhancements,[22,23] J couplings[5] (including those through hydrogen bonds[24]), and S^2 order parameters.[5,12] With a careful choice of the observables to be restrained, this type of approach should be capable of providing an accurate representation of the dynamics of proteins in solution under a variety of different conditions.

Acknowledgment. We acknowledge support by EMBO (A.D.S., M.V.), the Leverhulme Trust (X.S., M.V.), and the Royal Society (M.V.).

Supporting Information Available: Materials and methods, Tables S1 and S2, and Figures S1−S6. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Fersht, A. R. *Structure and Mechanism in Protein Science*; W. H. Freeman: New York, 1999.
(2) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598–1603.
(3) Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646–652.
(4) Mittermaier, A.; Kay, L. E. *Science* **2006**, *312*, 224–228.
(5) Lindorff-Larsen, K.; Best, R. B.; DePristo, M. A.; Dobson, C. M.; Vendruscolo, M. *Nature* **2005**, *433*, 128–132.
(6) Boehr, D. D.; McElheny, D.; Dyson, H. J.; Wright, P. E. *Science* **2006**, *313*, 1586–1587.
(7) Vendruscolo, M.; Dobson, C. M. *Science* **2006**, *313*, 1638–1642.
(8) Lange, O. F.; Lakomek, N. A.; Fares, C.; Schroder, G. F.; Walter, K. F. A.; Becker, S.; Meiler, J.; Grubmuller, H.; Griesinger, C.; de Groot, B. L. *Science* **2008**, *320*, 1471–1475.
(9) Eisenmesser, E. Z.; Millet, O.; Labeikovsky, W.; Korzhnev, D. M.; Wolf-Watz, M.; Bosco, D. A.; Skalicky, J. J.; Kay, L. E.; Kern, D. *Nature* **2005**, *438*, 117–121.
(10) Henzler-Wildman, K. A.; Thai, V.; Lei, M.; Ott, M.; Wolf-Watz, M.; Fenn, T.; Pozharski, E.; Wilson, M. A.; Petsko, G. A.; Karplus, M.; Hubner, C. G.; Kern, D. *Nature* **2007**, *450*, 838–844.
(11) Gsponer, J.; Christodoulou, J.; Cavalli, A.; Bui, J. M.; Richter, B.; Dobson, C. M.; Vendruscolo, M. *Structure* **2008**, *16*, 736–746.
(12) Richter, B.; Gsponer, J.; Varnai, P.; Salvatella, X.; Vendruscolo, M. *J. Biomol. NMR* **2007**, *37*, 117–135.
(13) Clore, G. M.; Schwieters, C. D. *J. Am. Chem. Soc.* **2004**, *126*, 2923–2938.
(14) Showalter, S. A.; Bruschweiler, R. *J. Am. Chem. Soc.* **2007**, *129*, 4158–4159.
(15) Kuriyan, J.; Petsko, G.; Levy, R. M.; Karplus, M. *J. Mol. Biol.* **1986**, *190*, 227–254.
(16) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712–725.
(17) Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111–1114.
(18) Salvatella, X.; Richter, B.; Vendruscolo, M. *J. Biomol. NMR* **2008**, *40*, 71–81.
(19) Bouvignies, G.; Bernado, P.; Meier, S.; Cho, K.; Grzesiek, S.; Bruschweiler, R.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13885–13890.
(20) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.
(21) Bonvin, A. M. J. J.; Boelens, R.; Kaptein, R. *J. Biomol. NMR* **1994**, *4*, 143–149.
(22) Tang, C.; Iwahara, J.; Clore, G. M. *Nature* **2006**, *444*, 383–386.
(23) Lindorff-Larsen, K.; Kristjansdottir, S.; Teilum, K.; Fieber, W.; Dobson, C. M.; Poulsen, F. M.; Vendruscolo, M. *J. Am. Chem. Soc.* **2004**, *126*, 3291–3299.
(24) Gsponer, J.; Hopearuoho, H. I.; Cavalli, A.; Dobson, C. M.; Vendruscolo, M. *J. Am. Chem. Soc.* **2006**, *128*, 15127–15135.

JA8087295

Downloaded by NESLI CONSORTIA UK on July 8, 2009
Published on March 3, 2009 on http://pubs.acs.org | doi: 10.1021/ja8087295