J|A|C|S ARTICLES

Determination of the Free Energy Landscape of α-Synuclein Using Spin Label Nuclear Magnetic Resonance Measurements

Jane R. Allison,[†] Peter Varnai,[‡] Christopher M. Dobson, and Michele Vendruscolo*

Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, U.K.

Received June 9, 2009; E-mail: mv245@cam.ac.uk

Abstract: Natively unfolded proteins present a challenge for structure determination because they populate highly heterogeneous ensembles of conformations. A useful source of structural information about these states is provided by paramagnetic relaxation enhancement measurements by nuclear magnetic resonance spectroscopy, from which long-range interatomic distances can be estimated. Here we describe a method for using such distances as restraints in molecular dynamics simulations to obtain a mapping of the free energy landscapes of natively unfolded proteins. We demonstrate the method in the case of α-synuclein and validate the results by a comparison with electron transfer measurements. Our findings indicate that our procedure provides an accurate estimate of the relative statistical weights of the different conformations populated by α-synuclein in its natively unfolded state.

Introduction

An overall description of the conformational space of a protein is obtained through the characterization of the range of structures accessible to it, including those comprising the folded and unfolded states as well as the metastable states populated during the folding and misfolding processes.[1,2] The characterization of these states is particularly important because they often play crucial roles in the folding and misfolding processes.[1,3,4] Additionally, a wide range of proteins, many of which are involved in gene regulation and signal transduction, are being recognized as natively unfolded or as containing extended unstructured regions.[3] Since these states are formed by heterogeneous ensembles of structures,[5−14] describing them in terms of ensembles of conformations is essential.[15−17] Molecular dynamics simulations are capable of generating such conformational ensembles,[18] especially with the incorporation of experimental restraints, which augments the information that can be extracted from experimental measurements by providing atomic-level structural detail.[14,19−29] In using measurements derived from nuclear magnetic resonance (NMR) spectroscopy, however, it is important to account for the fact that the recorded values represent time- and ensemble-averages. To achieve this result, time- or ensemble-averaging can be performed within molecular dynamics simulations.[9,14,19–21,23,24,26–32] At each time-step, the time- or ensemble-averaged observables are compared with their corresponding experimental values, and an

[†] Current address: Laboratory of Physical Chemistry, ETH Zürich, CH-8093 Zürich, Switzerland.

[‡] Current address: University of Sussex, Department of Chemistry and Biochemistry, Brighton BN1 9QJ, U.K.

(1) Fersht, A. R. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*; W. H. Freeman & Co.: New York, 1999.
(2) Vendruscolo, M.; Dobson, C. M. *Phil. Trans. R. Soc. A* **2005**, *363*, 433–450.
(3) Dyson, H. J.; Wright, P. E. *Nat. Rev. Mol. Cell. Biol.* **2005**, *6*, 197–208.
(4) Chiti, F.; Dobson, C. M. *Annu. Rev. Biochem.* **2006**, *75*, 333–366.
(5) Kohn, J. E.; Millett, I. S.; Jacob, J.; Zagrovic, B.; Dillon, T. M.; Cingel, N.; Dothager, R. S.; Seifert, S.; Thiyagarajan, P.; Sosnick, T. R.; Hasan, M. Z.; Pande, V. S.; Ruczinski, I.; Doniach, S.; Plaxco, K. W. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 12491–12496.
(6) Gillespie, J. R.; Shortle, D. *J. Mol. Biol.* **1997**, *268*, 158–169.
(7) Klein-Seetharaman, J.; Oikawa, M.; Grimshaw, S. B.; Wirmer, J.; Durchardt, E.; Ueda, T.; Imoto, T.; Smith, L. J.; Dobson, C. M.; Schwalbe, H. *Science* **2002**, *295*, 1719–1722.
(8) Teilum, K.; Maki, K.; Kragelund, B. B.; Poulsen, F. M.; Roder, H. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 9807–9812.
(9) 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.
(10) Bertoncini, C. W.; Jung, Y. S.; Fernandez, C. O.; Hoyer, W.; Griesinger, C.; Jovin, T. M.; Zweckstetter, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 1430–1435.
(11) Dedmon, M. M.; Lindorff-Larsen, K.; Christodoulou, J.; Vendruscolo, M.; Dobson, C. M. *J. Am. Chem. Soc.* **2005**, *127*, 476–477.
(12) Choy, W. Y.; Forman-Kay, J. D. *J. Mol. Biol.* **2002**, *308*, 1011–1032.
(13) Smith, L. J.; Bolin, K. A.; Schwalbe, H.; MacArthur, M. W.; Thornton, J. M.; Dobson, C. M. *J. Mol. Biol.* **1996**, *255*, 494–506.
(14) Francis, C. J.; Lindorff-Larsen, K.; Best, R. B.; Vendruscolo, M. *Proteins* **2006**, *65*, 145–152.
(15) Mittag, T.; Forman-Kay, J. D. *Curr. Opin. Struct. Biol.* **2007**, *17*, 3–14.
(16) Vendruscolo, M. *Curr. Opin. Struct. Biol.* **2007**, *17*, 15–20.
(17) Eliezer, D. *Curr. Opin. Struct. Biol.* **2009**, *19*, 23–30.
(18) Daggett, V.; Fersht, A. R. *Nat. Rev. Mol. Cell. Biol.* **2003**, *4*, 497–502.
(19) Bonvin, A. M. J. J.; Boelens, R.; Kaptein, R. *J. Biomol. NMR* **1994**, *4*, 143–149.
(20) Kemmink, J.; Scheek, R. M. *J. Biomol. NMR* **1995**, *6*, 33–40.
(21) Bonvin, A. M. J. J.; Brunger, A. T. *J. Mol. Biol.* **1995**, *250*, 80–93.
(22) Vendruscolo, M.; Paci, E.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 14817–14821.
(23) Vendruscolo, M.; Paci, E.; Dobson, C. M.; Karplus, M. *J. Am. Chem. Soc.* **2003**, *125*, 15686–15687.
(24) Best, R. B.; Vendruscolo, M. *J. Am. Chem. Soc.* **2004**, *126*, 8090–8091.
(25) Lindorff-Larsen, K.; Best, R. B.; DePristo, M. A.; Dobson, C. M.; Vendruscolo, M. *Nature* **2005**, *433*, 128–132.
(26) Clore, G. M.; Schwieters, C. D. *J. Am. Chem. Soc.* **2004**, *126*, 2923–2938.
(27) Clore, G. M.; Schwieters, C. D. *Biochemistry* **2004**, *43*, 10678–10691.
(28) Clore, G. M.; Schwieters, C. D. *J. Mol. Biol.* **2006**, *355*, 879–886.
(29) Hess, B.; Scheek, R. M. *J. Magn. Reson.* **2003**, *164*, 19–27.

10.1021/ja904716h CCC: $40.75

energy penalty is added to enforce the agreement with the experimental information.[9,14,19–21,24,26,26,28–30,32,33] In this way, the simulations are biased toward structures that satisfy the restraints in a time- or ensemble-averaged manner. This type of simulations has been used to characterize the ensembles of conformations representing a variety of different protein states, including disordered, intermediate, transition and folded states.[2,26–28,33,34]

For globular proteins in their native or near-native states, a range of experimental techniques, including X-ray crystallography and NMR spectroscopy, provide structural information that can be used as restraints.[35] Disordered states, however, present additional challenges. For instance, nonsequential NOEs, which provide a rich source of information for determining folded structures, are seldom detected, as the internuclear distances are too large and variable.[6] In other cases, such as for ^3J couplings and secondary chemical shifts, averaging over the heterogeneous range of structures leads to a considerable reduction of the structural information that can be extracted from the data.

A particularly useful source of structural information about conformationally heterogeneous states of proteins is represented by paramagnetic relaxation enhancement (PRE) NMR spectroscopy.[6,8,9,14,36−41] This technique exploits the change in the relaxation rate of a nuclear spin induced by the presence of a distant paramagnetic group to infer the distance between the two centers. This effect is sensitive up to distances in the range 12−20 Å, making it useful for characterizing disordered states.[6,8,9,14,36] These long-range distances have been used as restraints in molecular dynamics simulations to determine structural ensembles for the disordered states of several proteins,[9,14,36,42] including α-synuclein,[10,11] a 140-residue natively unfolded protein.[43−45] This protein is capable of forming amyloid fibrils both *in vitro* and *in vivo*, and it is the primary constituent of the deposits observed in Parkinson's disease.[46−50] It has also been proposed that in its natively unfolded state, the C-terminus of α-synuclein forms transient intramolecular interactions with the N-terminus and with the central non-β-amyloid component (NAC) region (residues 61−95),[51] which is essential for aggregation,[52−59] thus reducing the overall aggregation propensity of this protein.[10,11,60,61]

In this work we incorporate PRE-derived distances as ensemble-averaged restraints in molecular dynamics simulations to characterize the natively unfolded state of α-synuclein. After a careful validation of the method that we used, we employ it to obtain a detailed description of the free energy landscape of α-synuclein, which provides a representation of the statistical weights of the variety of conformational states populated by this protein.

Methods

Simulation Methods. All simulations were carried out within the CHARMM simulation package,[62] modified to allow interatomic distance restraints to be applied as ensemble averages. The CHARMM19 polar hydrogen representation was used. Newton's equations of motions were employed and the temperature was controlled using the Nosé-Hoover thermostat. Bond lengths were constrained with the SHAKE algorithm,[63] allowing for an integration time-step of 2 fs.

Unrestrained Molecular Dynamics Simulations. We generated five different ensembles (UERg23, UERg20, USRg24, UST590, and RC) using molecular dynamics simulations in which no restraints derived from experimental measurements were applied. An explanation of the abbreviated names and the details of the simulation protocols are provided in Table 1. Ensembles UERg23 and UERg20 were used as reference ensembles (see below), from which PRE-like distances were back-calculated and used as

(30) Gsponer, J.; Hopearuoho, H.; Whittaker, S. B.-M.; Spence, G. R.; Moore, G. R.; Paci, E.; Radford, S. E.; Vendruscolo, M. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 99–104.

(31) Gsponer, J.; Hopearuoho, H.; Cavalli, A.; Dobson, C. M.; Vendruscolo, M. *J. Am. Chem. Soc.* **2006**, *128*, 15127–15135.

(32) Richter, B.; Gsponer, J.; Varnai, P.; Salvatella, X.; Vendruscolo, M. *J. Biomol. NMR* **2007**, *37*, 117–135.

(33) Vendruscolo, M.; Paci, E. *Curr. Opin. Struct. Biol.* **2003**, *13*, 82–87.

(34) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. *J. Am. Chem. Soc.* **1999**, *121*, 2337–2338.

(35) Brünger, A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J. S.; Kuszewski, J.; Nilges, M.; Pannu, N. S.; Read, R. J.; Rice, L. M.; Simonson, T.; Warren, G. L. *Acta Crystallogr. D* **1998**, *54*, 905–921.

(36) Kristjansdottir, S.; Lindorff-Larsen, K.; Fieber, W.; Dobson, C. M.; Vendruscolo, M.; Poulsen, F. M. *J. Mol. Biol.* **2005**, *347*, 1053–1062.

(37) Liang, B. Y.; Bushweller, J. H.; Tamm, L. K. *J. Am. Chem. Soc.* **2006**, *128*, 4389–4397.

(38) Iwahara, J.; Schwieters, C. D.; Clore, G. M. *J. Am. Chem. Soc.* **2004**, *126*, 5879–5896.

(39) Battiste, J. L.; Wagner, G. *Biochemistry* **2000**, *39*, 5355–5365.

(40) Donaldson, L. W.; Skrynnikov, N. R.; Choy, W. Y.; Muhandiram, D. R.; Sarkar, B.; Forman-Kay, J. D.; Kay, L. E. *J. Am. Chem. Soc.* **2001**, *123*, 9843–9847.

(41) Tang, C.; Iwahara, J.; Clore, G. M. *Nature* **2006**, *444*, 383–386.

(42) Marsh, J. A.; Forman-Kay, J. D. *J. Mol. Biol.* **2009**, *391*, 359–374.

(43) Eliezer, D.; Kutluay, E.; Bussell, R.; Browne, G. *J. Mol. Biol.* **2001**, *307*, 1061–1073.

(44) Weinreb, P.; Zhen, W.; Poon, A.; Conway, K.; Lansbury, P. T. *Biochemistry* **1996**, *35*, 13709–13715.

(45) Uversky, V. N. *J. Biomol. Struct. Dyn.* **2003**, *21*, 211–234.

(46) Spillantini, M. G.; Schmidt, M. L.; Lee, V. M.; Trojanowski, J. Q.; Jakes, R.; Goedert, M. *Nature* **1997**, *388*, 839–840.

(47) Baba, M.; Nakajo, S.; Tu, P. H.; Tomita, T.; Nakaya, K.; Lee, V. M.; Trojanowski, J. Q.; Iwatsubo, T. *Am. J. Pathol.* **1998**, *152*, 879–884.

(48) Goedert, M. *Nat. Rev. Neurosci.* **2001**, *2*, 492–501.

(49) Moore, D. J.; West, A. B.; Dawson, D. L.; Dawson, T. M. *Annu. Rev. Neurosci.* **2005**, *28*, 57–87.

(50) Bisaglia, M.; Mammi, S.; Bubacco, L. *FASEB J.* **2009**, *23*, 329–340.

(51) Ueda, T.; Fukushima, H.; Masliah, E.; Xia, Y.; Iwai, A.; Yoshimoto, M.; Otero, D.; Kondo, J.; Ihara, Y.; Saitoh, T. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 11282–11286.

(52) Giasson, B. I.; Murray, I. V. J.; Trojanowski, J. Q.; Lee, V. M. Y. *J. Biol. Chem.* **2001**, *276*, 2380–2386.

(53) Du, H. N.; Tang, L.; Luo, X. Y.; Li, H. T.; Hu, J.; Zhou, J. W.; Hu, H. Y. *Biochemistry* **2003**, *42*, 8870–8878.

(54) Madine, J.; Doig, A.; Middleton, D. *Biochem. Soc. Trans.* **2004**, *32*, 1127–1129.

(55) Bertoncini, C. W.; Rasia, R. M.; Lamberto, G. R.; Gonzalo, R.; Binolfi, A.; Zweckstetter, M.; Griesinger, C.; Fernandez, C. O. *J. Mol. Biol.* **2007**, *372*, 708–722.

(56) Tartaglia, G. G.; Pawar, A.; Campioni, S.; Chiti, F.; Dobson, C. M.; Vendruscolo, M. *J. Mol. Biol.* **2008**, *380*, 425–436.

(57) Rivers, R. C.; Kumita, J. R.; Tartaglia, G. G.; Dedmon, M. M.; Pawar, A.; Vendruscolo, M.; Dobson, C. M.; Christodoulou, J. *Protein Sci.* **2008**, *17*, 887–898.

(58) Wu, K.-P.; Kim, S.; Fela, D. A.; Baum, J. *J. Mol. Biol.* **2008**, *378*, 1104–1115.

(59) Sandal, M.; Valle, F.; Tessari, I.; Mammi, S.; Bergantino, E.; Musiani, F.; Brucale, M.; Bubacco, L.; Samori, B. *PLoS Biol.* **2008**, *6*, 99–108.

(60) Bernado, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R. W. H.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17002–17007.

(61) Rospigliosi, C. C.; McClendon, S.; Schmid, A. W.; Ramlall, T. F.; Barre, P.; Lashuel, H. A.; Eliezer, D. *J. Mol. Biol.* **2009**, *338*, 1022–1032.

(62) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.

(63) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.

(64) Paci, E.; Vendruscolo, M.; Dobson, C. M.; Karplus, M. *J. Mol. Biol.* **2002**, *324*, 151–163.

Table 1. Different Conformational Ensembles Discussed in This Work[a]

name	solvent	t	N_{rep}	$N_{\mathrm{structures}}$	T	$\langle R_g \rangle$	brief description
UERg23	EEF1[73]	400	20	23 675	540	23.2	ensemble for which $\langle R_g \rangle$ is set to be similar to that of a previously determined α-synuclein ensemble[11]
UERg20	EEF1[73]	320	16	64 000	505	20.5	compact ensemble used for validating the generality of the method
USRg24	SASA[65]	260	20	52 000	515	24.1	ensemble for which $\langle R_g \rangle$ is set to be similar to that of UERg23
UST590	SASA[65]	320	16	64 000	590	32.6	ensemble in which T is chosen to best reproduce UERg23
RC	vacuum	200	1	10 000	523	41.8	random coil ensemble
RSRg23	SASA[65]	288	24	57 600	590	23.2	ensemble obtained by using PRE-like distances calculated from UERg23
RSexp	SASA[65]	200	24	48 000	490	32.1	ensemble obtained by using experimentally derived PRE distances

[a] The names used to identify the ensembles are provided together with their main features. "U" and "R" refer to unrestrained or restrained simulations, respectively, and "RC" to the random coil ensemble. "E" and "S" refer to the type of implicit solvent model used (EEF1 and SASA, respectively); the remainder of the name indicates the distinguishing feature of the specific ensemble. "Solvent" specifies the implicit solvent model used, t indicates the total sampling time of the simulation in ns, N_{rep} specifies the number of independent replicas used in the sampling and $N_{\mathrm{structures}}$ the total number of structures collected during the simulations. T is the simulation temperature in Kelvin and $\langle R_g \rangle$ is the ensemble-averaged radius of gyration in Å. For RSexp, the $\langle R_h^{-1} \rangle^{-1}$ is given rather than the $\langle R_g \rangle$.

restraints in the simulations that we carried out for validating the method. USRg24 and UST590 were used to assess the information content of the restraints and the effects of restraining on the degree of expansion of the protein structures. RC represents a disordered polypeptide chain, in which only excluded volume interactions are taken into account; this ensemble was generated using CHARMM19 *in vacuo* with no electrostatic interactions and with the nonbonded interactions truncated to maintain only the repulsive part of the van der Waals potential.

For each ensemble, we began the simulations by heating the protein to the designated temperature followed by equilibration for 0.1 ns prior to the collection of structures. Structures were collected every 5 ps (2500 steps) except for the RC ensemble, for which they were saved and every 20 ps. This time interval was sufficient to ensure that subsequent structures were not correlated.

The UERg23 ensemble was filtered to increase the degree of residual structure (see Results) by selecting only those structures with more than 15 contacts between the NAC region (residues 61−95) and the C-terminus (residues 110−140). Two residues were considered to be in contact if their C_α atoms were within 8.5 Å of each other, following the definition of Dedmon et al.[11]

Molecular Dynamics Simulations with Ensemble-Averaged Restraints. In molecular dynamics simulations with ensemble-averaged restraints, multiple replicas are simulated in parallel.[9,11,14,19–21,24,26–29,31–33] A reaction coordinate (ρ) is defined as the sum of the difference between the calculated (f_l^{calc}) and the reference (f_l^{ref}) values of the observables[9]

$$\rho(t) = N_{\mathrm{restr}}^{-1} \sum_{l=1}^{N_{\mathrm{restr}}} \left(f_l^{\mathrm{ref}} - f_l^{\mathrm{calc}}(t)\right)^2 \tag{1}$$

where the index l runs over the N_{restr} restrained observables. In the case of PRE-derived distances, we define f_l^{calc} as

$$f_l^{\mathrm{calc}}(t) = d_{ij}^{\mathrm{calc}}(t) = \left(N_{\mathrm{rep}}^{-1} \sum_{k=1}^{N_{\mathrm{rep}}} r_{ij,k}^{-6}(t)\right)^{-1/6} \tag{2}$$

where $r_{ij,k}(t)$ is the distance between the C_α atom of residue i and the amide hydrogen of residue j calculated for replica k of the restrained ensemble at time t and N_{rep} is the number of replicas.

In the case of the tests with the reference ensembles, f_l^{ref} is taken as the ensemble-averaged distance d_{ij}^{ref}

$$d_{ij}^{\mathrm{ref}} = \left(N_{\mathrm{ref}}^{-1} \sum_{k=1}^{N_{\mathrm{ref}}} r_{ij,k}^{-6}\right)^{-1/6} \tag{3}$$

where N_{ref} is the number of structures in the reference ensemble. When experimental data were used, f_l^{ref} is the distance calculated as described below.

During the PRE-restrained molecular dynamics simulations, $d_{ij}^{\mathrm{calc}}(t)$ is allowed to vary freely between $d_{ij}^{\mathrm{ref}} - L$ and $d_{ij}^{\mathrm{ref}} + U$, where L and U are suitably chosen values of the lower (L) and upper (U) bounds (see below). Outside the square well, a harmonic penalty of the form

$$\frac{\alpha N_{\mathrm{rep}}}{2}(\rho(t) - \rho_0(t))^2 \tag{4}$$

is added to the energy if $\rho(t) > \rho_0(t)$, where

$$\rho_0(t) = \min[\rho(\tau)] \quad (0 \le \tau \le t) \tag{5}$$

and α is a force constant associated with the restraints.[64] In this way, as the simulation proceeds, the ensemble is progressively biased toward structures that satisfy the restraints.

Ensemble-restrained molecular dynamics simulations were carried out using the SASA[65] implicit solvation model. Multiple replicas were simulated in parallel, and an extra phase was included immediately after the heating stage during which α was increased from its starting value of 500 kcal mol^{-1} Å^{-2} to its final value (364,500 kcal mol^{-1} Å^{-2}) by a factor 3 every 10 ps. N_{rep}, L, and U were varied as discussed below. The structures collected at each point in time were pooled together into one ensemble prior to analysis.

Calculation of Distance Restraints: Experimental Distance Restraints. In a spin label NMR experiment, the ^{1}H−^{15}N heteronuclear single quantum coherence (HSQC) spectra of a spin-labeled protein are recorded with the spin label in its oxidized (paramagnetic) and reduced (diamagnetic) states. The effects due to the presence of a free electron are quantified by the intensity ratio, $I_{\mathrm{ox}}/I_{\mathrm{red}}$, which compares the intensity (height) of the cross-peaks in the oxidized (I_{ox}) and reduced (I_{red}) states. Distances were calculated from $I_{\mathrm{ox}}/I_{\mathrm{red}}$ as described previously,[11] but with some modifications as discussed below.

We used here a set of 478 PRE-derived distances described previously to characterize the natively unfolded state of α-synuclein.[11] In addition, we considered an additional set of 117 distances obtained from a new spin label positioned at N122; the experimental procedure used was the same as before.[11] Thus, we employed a total of 595 PRE-derived distance restraints (i.e., 4.25 restraints on average per residue).

When enforcing PRE-derived distance restraints, there is a degree of tolerance toward violations in the ensemble-averaged back-calculated distances d_{ij} (eq 2) at each point in time because a contribution toward ρ (eq 1) is only made if $d_{ij}^{\mathrm{calc}}(t)$ is outside a square well-defined by $d_{ij}^{\mathrm{ref}} - L$ and $d_{ij}^{\mathrm{ref}} + U$. The exact value of L and U was varied during the testing phase as discussed in Results.

The nature of the equations used to compute a distance from $I_{\mathrm{ox}}/I_{\mathrm{red}}$[6,39] implies that for high $I_{\mathrm{ox}}/I_{\mathrm{red}}$, a small change in $I_{\mathrm{ox}}/I_{\mathrm{red}}$ results in a large change in the calculated distance. We therefore treated the distance that would be calculated if $I_{\mathrm{ox}}/I_{\mathrm{red}} = 0.85$ (d_{ij}^{max}) as the longest reliable distance. $d_{ij}^{\mathrm{ref}} > d_{ij}^{\mathrm{max}}$ were used as "negative"

(65) Ferrara, P.; Apostolakis, J.; Caflisch, A. *Proteins* **2002**, *46*, 24–33.

restraints[37,39] by assigning only a lower bound corresponding to $d_{ij}^{\max} - L$. In addition, values $I_{\text{ox}}/I_{\text{red}} < 0.15$ may be unreliable,[37,39] as any experimental uncertainty is large relative to the size of the measured $I_{\text{ox}}/I_{\text{red}}$. The distance that would be calculated if $I_{\text{ox}}/I_{\text{red}} = 0.15$ ($d_{ij}^{\min}$) was therefore the shortest distance to be given both lower and upper boundaries. Any $d_{ij}^{\text{ref}} < d_{ij}^{\min}$ was assigned only an upper bound corresponding to $d_{ij}^{\min} + U$.

Calculation of Distance Restraints: Reference Distance Restraints. To test the method that we used in this work, a set of reference distance restraints, d_{ij}^{ref}, were calculated from the reference ensembles UERg23 and UERg20 to be analogous to the distances that would be obtained from a spin label NMR experiment. Eight residues evenly distributed along the α-synuclein sequence were designated to be "spin-labeled". The distance between the C_α atom of the spin-labeled residues *i* and the amide hydrogens on all nonadjacent residues *j* were calculated from and averaged over the N_{ref} structures comprising each reference ensemble. Distances between atoms belonging to the same or sequentially adjacent residues were removed to give 1000 restraints in total, similar to the maximum number of distances that is typically determined experimentally. A "free" data set consisting of a further 1000 distances was also calculated by considering a second set of eight spin-labeled residues; these distances were used for cross-validation. A r^{-6} averaging was used to simulate the averaging inherent in a spin label NMR experiment. Each reference distance restraint was assigned a lower bound, *L*, and upper bound, *U*, in a similar manner as for experimental data.

Analysis and Comparison of Ensembles of Structures. Calculation of R_g and R_h. The radius of gyration, R_g, was calculated from the positions of the heavy atoms of each structure. When experimental restraints were used, the ensemble-averaged hydrodynamic radius, $\langle R_h^{-1}\rangle^{-1}$, was computed using the program HYDROPRO[66] with default settings for comparison with the experimental value. A phenomenological relationship between R_g and R_h, which was parametrized by linear regression,[9] was used to convert the calculated R_g for each structure into an R_h. The harmonic mean was computed to reflect the averaging inherent in the experimental measurement of R_h by NMR diffusion experiments.[67]

Calculation of *Q* Factors and of *S* Factors. The agreement between the average values of a given observable in the reference ensemble or measured experimentally (f_l^{ref}) and those back-calculated from and averaged over the ensemble obtained using restrained molecular dynamics simulations (f_l^{calc}) was quantified with a quality factor[68]

$$Q = \frac{\left(\sum_{l=1}^{N_{\text{obs}}} (f_l^{\text{ref}} - f_l^{\text{calc}})^2\right)^{1/2}}{\left(\sum_{l=1}^{N_{\text{obs}}} (f_l^{\text{ref}})^2\right)^{1/2}} \tag{6}$$

where N_{obs} is the number of observables; low values of *Q* indicate good agreement.

To quantify the agreement between two distributions we used the distance[32]

$$s_l = \sum_{m=1}^{N_{\text{bins}}} |p_{m,l}^{\text{ref}} - p_{m,l}^{\text{calc}}| \tag{7}$$

where N_{bins} is the number of bins into which the histograms were divided and $p_{m,l}$ is the normalized probability of finding a particular observable in bin *m* of histogram *l*. s_l ranges from 0 to 2, with low values representing similar histograms. Summation over all N_{obs} histograms quantifies the overall agreement of two ensembles in terms of distance distributions

$$S = N_{\text{obs}}^{-1} \sum_{l=1}^{N_{\text{obs}}} s_l \tag{8}$$

The values of s_l and *S* depend on the bin width, the ideal value of which depends in turn on the width of the distributions being compared. A bin width of 1 Å was found to be suitable for the wide range of distance and R_g distributions encountered. Using the same bin width for all distributions allows comparisons of the s_l values computed from different pairs of atoms in the same and different ensembles. It is also a prerequisite for combining the various s_l into an overall *S* value.

The statistical error in the *Q* and *S* values was estimated by randomly splitting the data into two sets and computing a *Q* or *S* value for each set with respect to the reference ensemble. The splitting was repeated 10 times such that 20 different *Q* or *S* values were collected. The standard deviation of these values was taken as the statistical error. The convergence of the ensemble-restrained molecular dynamics simulations was checked by comparing *Q* or *S* values calculated separately for the first and second halves of the simulation and monitoring R_g throughout time (data not shown).

Calculation of s_l Maps. The s_l values for all N_{obs} distance distributions were plotted in two dimensions according to the pairs of residues for which the distribution is defined. The MATLAB (The MathWorks, Inc.) `griddata` function was used to interpolate between the nonuniformly spaced points for which s_l values were computed.

Definition of Residual Contact Probability Maps. The residual contact probability (RCP)[11] is defined as $-\ln(p_{ij}^{calc}/p_{ij}^{ref})$, where p_{ij}^{calc} and p_{ij}^{ref} are the probabilities of contact formation in the ensemble of interest and in the random coil ensemble, respectively. The pseudoenergy value is smoothed over a window of seven residues to account for the concentration of distance restraints around the spin-labeled residues.

Definition of Distance Comparison Maps. Distance comparison (DC) maps were created by plotting the root-mean-square distance (rmsd) between two residues, *i* and *j*, normalized by the rmsd computed from the random coil ensemble

$$\text{DC}_{ij} = \frac{\left\langle d_{ij}^{\text{calc}2}\right\rangle^{1/2}}{\left\langle d_{ij}^{\text{rc}2}\right\rangle^{1/2}} \tag{9}$$

Here, $\langle d_{ij}^{\text{calc}^2}\rangle^{1/2}$ and $\langle d_{ij}^{\text{rc}^2}\rangle^{1/2}$ are defined as

$$\left\langle d_{ij}^2\right\rangle^{1/2} = \left(N_{struct}^{-1} \sum_{k=1}^{N_{struct}} d_{ij,k}^2\right)^{1/2} \tag{10}$$

where N_{struct} is the number of structures in the calculated or random coil ensemble. Similar results were obtained if $\langle d_{ij}^{\text{rc}^2}\rangle^{1/2}$ was computed using an equation that predicts the rmsd between two residues with sequence separation N_{sep} for a random flight chain with excluded volume and dihedral angles taken from a PDB coil library.[69] The normalization by $\langle d_{ij}^{\text{rc}^2}\rangle^{1/2}$ is important because it removes the dependence of the inter-residue distance on the sequence separation, allowing pairs of residues with different sequence separations and also proteins of different lengths to be compared. The DC map was not smoothed and was plotted as discrete points using the MATLAB `imagesc` function.

Detection of Correlations between Distance Distributions. The correlation between two distance distributions was investigated by computing s_p values (analogous to s_l values, eq 7) to quantify the similarity between the 2D distance histograms $p(r_{\text{AB}}, r_{\text{AC}})$ and $p(r_{\text{AB}})*p(r_{\text{AC}})$, where r_{AB} and r_{AC} are the distances between the C_α atoms of residues A and B or A and C, respectively. A, B, and C

[66] Garcia de la Torre, J.; Huertas, M. L.; Carrasco, B. *Biophys. J.* **2000**, *78*, 719–730.

[67] Wilkins, D. K.; Grimshaw, S. B.; Receveur, V.; Dobson, C. M.; Jones, J. A.; Smith, L. J. *Biochemistry* **1999**, *38*, 16424–16431.

[68] Bax, A. *Protein Sci.* **2003**, *12*, 1–16.

[69] Zhou, H. X. *Biophys. J.* **2002**, *83*, 2981–2986.

came from a set of 10 residues spaced approximately 14 positions apart along the 140-amino-acid sequence. We chose a set of residues that were not involved in either the experimental or reference distance restraints so that the identification of correlations was not complicated by the direct influence of a restraint on the interatomic distances. High values of s_p indicate that r_{AB} and r_{AC} are correlated, that is, $p(r_{AB}, r_{AC}) \neq p(r_{AB}) * p(r_{AC})$.

The s_p values were displayed as 2D contour plots of s_p with respect to B and C for each value of A. Because A, B, and C each take only 10 values, the MATLAB (The MathWorks, Inc.) `interp2` function was used to linearly interpolate the s_p values in 2 dimensions, giving an estimated s_p value for all possible BC combinations for each of the chosen A.

Free Energy Landscapes. Free energy landscapes were obtained for each ensemble by plotting the 2D histogram of $p(R_g, \text{SASA})$ according to

$$F(R_g, \text{SASA}) = -\ln p(R_g, \text{SASA}) \qquad (11)$$

where $p(R_g, \text{SASA})$ is the joint probability distribution of the R_g and the solvent exposed surface area (SASA).

Results

Our aims in this work are first to establish a general method for using distance information derived from spin label NMR measurements as ensemble-averaged restraints in molecular dynamics simulations to determine ensembles of conformations representing disordered states of proteins and then to use this method to determine the free energy landscape of α-synuclein in its monomeric state under physiological conditions. Thus we first demonstrate that molecular dynamics simulations with ensemble-averaged distance restraints are capable of reproducing a known ensemble of conformations with the correct statistical weights. We then apply the method using experimentally derived distances for α-synuclein and further validate the results by a comparison with distance distributions derived from electron transfer (ET) measurements,[70] which were not used as restraints in the calculations.

Generation of Reference Ensembles. To test and validate the strategy that we follow to carry out molecular dynamics simulations with ensemble-averaged distance restraints, we apply the "test of the reference ensemble".[32,71,72] In this test, a reference ensemble of conformations is generated by molecular dynamics simulations. Then reference restraints are obtained from this ensemble by averaging a set of interatomic distances chosen to be equivalent to those generated by typical experimental measurements. Finally the reference ensemble is reconstructed by using the reference restraints in ensemble-restrained molecular dynamics simulations. This type of validation strategy has a long history[71] and is similar to the one that we described for establishing the MUMO method[32] and more recently to prove that native free energy landscapes can be calculated with high accuracy.[72]

The test of the reference ensemble has the advantage that all aspects of the state to be characterized are known. There are many advantages in testing a computational method in this way. Problems related to possible inaccuracies in the experimental data and in the translation of experimental NMR signals into structural restraints are avoided.[32,71,72] Moreover, the ensembles produced using ensemble-restrained molecular dynamics simulations can be compared to the reference ensembles from which the restraints were calculated in terms of both averages and distributions. This aspect is important because ensembles of conformations are best described in terms of distributions, whereas experimental observables are both time- and ensemble-averages. Thus testing the method using reference data provides a unique opportunity carry out cross-validation using quantities that report on both the averages (Q values, see Methods) and the distributions (S values, see Methods).

The generation of a realistic reference ensemble is, however, by itself a computational challenge. The use of explicit solvent models does not allow sufficient sampling to encompass the time scale of motion in disordered states, which can extend to milliseconds or more. Implicit solvent models provide faster alternatives but tend to favor the compact structures characteristic of globular proteins, at least in the case of those that we have tested and when the simulations are carried out at room temperature.

To generate less compact structures, one possibility is to raise the temperature in the simulations to shift the balance between the energy and the entropy of the protein to favor the sampling of more disordered states. The disadvantage of this procedure is that high energy conformations are generated. Since the purpose here, however, is to generate only a reference ensemble, this problem may be ignored. Indeed, we are not concerned at this stage about whether our reference ensemble is an exact reflection of the ensemble of structures sampled by native α-synuclein under experimental conditions. Our aim is to establish a computational procedure that is capable of reproducing a known reference ensemble and can therefore be used with confidence to reconstruct an unknown ensemble. We will also present below an alternative procedure based on the prediction of experimental measurements not used as restraints in the simulations.

We generated two reference ensembles of α-synuclein structures, UERg23 (unrestrained in EEF1 implicit solvent with $\langle R_g \rangle$ ~23 Å) and UERg20 using unrestrained molecular dynamics simulations with the EEF1[73] implicit solvent model (Table 1). The simulation temperature for UERg23 was chosen so that the $\langle R_h^{-1} \rangle^{-1}$ was close to that of the previously obtained ensemble of α-synuclein structures[11] and the experimental value for α-synuclein in solution[74] (26.6 ± 0.5 Å). Because experimental data and previous simulations suggested that α-synuclein in solution has a tendency to form contacts between the C-terminus and the central NAC region,[11] only structures with more than 15 contacts between these two regions were included in the reference ensemble. This selection process reduced the $\langle R_h^{-1} \rangle^{-1}$ by about 1 Å but did not markedly change other ensemble-averaged quantities (data not shown). The UERg20 ensemble contains more compact conformations ($\langle R_g \rangle$ ~20 Å) and was not filtered.

Minimization of Over-Restraining and Under-Restraining. Disordered states comprise heterogeneous ensembles of structures.[15,16] Consequently, it is not appropriate to enforce restraints upon a single replica, since a single structure compatible with all of the restraints is unlikely to be representative of the structures actually present and may even be physically impossible to obtain.[21,28,75] This problem, sometimes referred to as

[70] Lee, J. C.; Gray, H. B.; Winkler, J. R. *J. Am. Chem. Soc.* **2005**, *127*, 16388–16389.

[71] Kuriyan, J.; Petsko, G.; Levy, R. M.; Karplus, M. *J. Mol. Biol.* **1986**, *190*, 227–254.

[72] De Simone, A.; Richter, B.; Salvatella, X.; Vendruscolo, M. *J. Am. Chem. Soc.* **2009**, *131*, 3810–3811.

[73] Lazaridis, T.; Karplus, M. *Proteins* **1999**, *35*, 133–152.

[74] Morar, A. S.; Olteanu, A.; Young, G. B.; Pielak, G. J. *Protein Sci.* **2001**, *10*, 2195–2199.

[75] Zagrovic, B.; van Gunsteren, W. F. *Proteins* **2006**, *63*, 210–218.

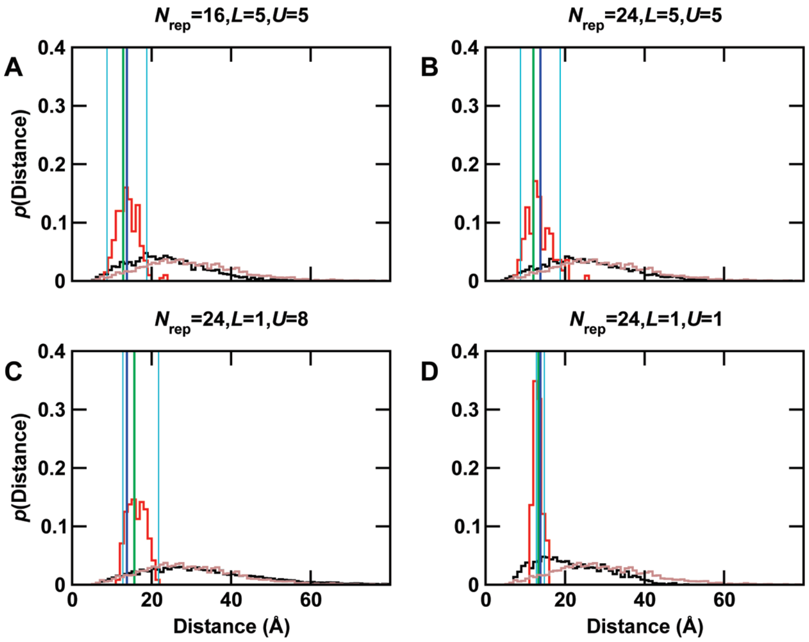

Figure 1. Analysis of the effects of changing N_{rep}, L, and U on the distance distributions. The distribution of the distances $r_{ij,k}(t)$ for each PRE-restrained ensemble sampled over all time-points and all replicas is in black. The distribution of ensemble averaged distances $d_{ij}(t)$ over all time points is in red. The distribution of distances $r_{ij,k}^{\mathrm{ref}}$ calculated from UERg23 is in gray. Also shown are the overall time- and ensemble-average calculated from each PRE-restrained ensemble, d_{ij}^{calc}, in green, and from UERg23, d_{ij}^{ref}, in blue, and the lower (L) and upper (U) bounds (cyan). The number of replicas, L, and U are shown above each graph.

over-restraining,[16] arises because experimental data derived from NMR spectroscopy are ensemble-averages over a large number of molecules in solution and time-averages over the length of the experiment.[16,32]

To avoid over-restraining, multiple copies of the molecule can be simulated in parallel, with the restraints imposed on the observables averaged over all replicas.[16] The number of replicas, however, cannot become too large because as this number grows, the experimental information quickly becomes insufficient to define the structures of all of the replicas, a problem known as overfitting, or under-restraining.[16,32] The number of replicas must therefore be carefully chosen so as to simultaneously avoid under-restraining and over-restraining.[16,32]

In a recent study, Ganguly and Chen[76] suggested that the use of PRE-derived distance restraints to characterize disordered states of proteins is an underdetermined problem unless a large number of distance restraints is used or additional sources of information are also considered. The simulations upon which they base this conclusion were carried out using 61 distance restraints for a 56-residue protein. In our study we used 1000 reference distance restraints (about 7 per residue) or 595 experimental distance restraints (about 4.25 per residue) to characterize the ensemble of structures populated by a 140-residue protein, both of which are greater than the empirical recommendation of 4 restraints per residue made by Ganguly and Chen.[76] Moreover, their results were generated using symmetric bounds of ±5 Å around the PRE-derived distances, whereas we have shown here that asymmetric bounds perform significantly better in the case of r^{-6} averaging. In addition, we also used the experimental R_g value to direct the conformational sampling, thus employing a further source of structural information. We therefore conclude that the procedure that we describe here meets the stringent criteria proposed by Ganguly and Chen[76] to avoid the generation of underdetermined ensembles.

A standard means of determining the optimal number of replicas is cross-validation.[21,77−79] Typically, about 20% of the data are excluded from the working data set (the restraints). Reproduction of these free data provides a more stringent test than satisfaction of the restraints. This is because, unlike satisfaction of the restraints, which generally improves with more replicas, reproduction of the free data becomes worse. This type of cross-validation is particularly effective in identifying the appearance of under-restraining, but it may not detect over-restraining.[16] For example, compact conformations of Δ131Δ satisfy this cross-validation test, even if they are over-restrained.[14]

Choice of Distance Bounds. The averaging of the inverse sixth power of the distance makes the PRE-derived distance restraints particularly sensitive to conformations in which the unpaired electron in the spin label is close to an amide hydrogen. As a consequence, the fewer replicas there are, the greater the proportion of replicas that must contain short distances to satisfy the restraint at each point in time (Figure 1A and B). This effect results in narrow distributions of distances, which contain mostly short distances close to the r^{-6} average, and ultimately, in ensembles of structures that are too compact (Table 2). Accordingly, despite carrying out the molecular dynamics simulations with ensemble-averaged restraints at temperatures at which the $\langle R_g \rangle$ of an unrestrained ensemble matched that of the corresponding reference ensemble, we found that upon application of reference distance restraints the $\langle R_g \rangle$ decreased, even with 32 replicas (Table 2).

(76) Ganguly, D.; Chen, J. H. *J. Mol. Biol.* **2009**, *390*, 467–477.

(77) Brünger, A.; Clore, G. M.; Gronenborn, A.; Saffrich, R.; Nilges, M. *Science* **1993**, *261*, 328–331.

(78) Burling, F. T.; Weiss, W. I.; Flaherty, K. M.; Brunger, A. T. *Science* **1996**, *271*, 72–77.

(79) Brünger, A. *Nature* **1992**, *355*, 472–475.

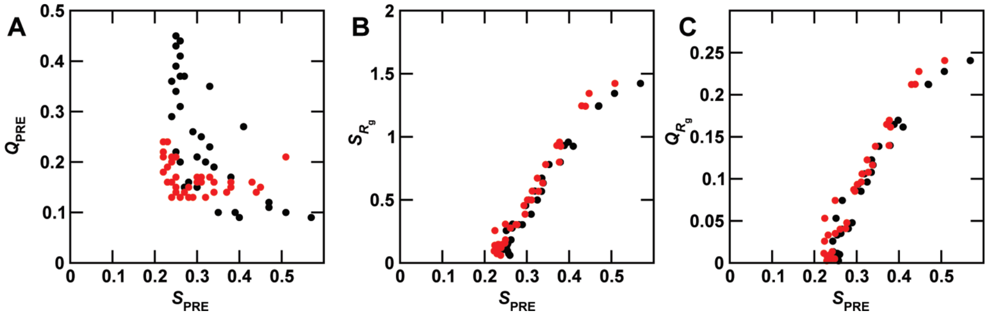

Figure 2. Relationship between the similarity of two distributions (S) and the agreement of the corresponding average values (Q). Correlation between the S_{PRE} values and (A) the Q_{PRE} values, (B) the S_{R_g} values, and (C) the Q_{R_g} values. In each plot, the working data set is shown in black and the free data set is in red. Each point represents a different ensemble-restrained simulation using distance restraints calculated from UERg23 (or UERg20) and a particular combination of N_{rep}, L, and U.

Table 2. Systematic Study of the Effects of Variation of N_{rep}, L, and U on the Quality of Reconstructed Ensembles[a]

N_{rep}	L	U	$\langle R_g \rangle$ (Å)	Q_{R_g}	S_{R_g}	$Q_{w\mathrm{PRE}}$	$Q_{f\mathrm{PRE}}$	$S_{w\mathrm{PRE}}$	$S_{f\mathrm{PRE}}$
16	5	5	18.3	0.21	1.24	0.11	0.16	0.47	0.43
24	5	5	19.3	0.17	0.96	0.09	0.14	0.39	0.37
32	5	5	20.0	0.14	0.78	0.10	0.17	0.35	0.34
16	1	1	13.8	0.41	2.0	0.11	0.16	0.66	0.54
24	1	1	17.6	0.24	1.42	0.09	0.21	0.57	0.51
32	1	1	17.9	0.23	1.34	0.10	0.15	0.51	0.45
16	1	8	19.9	0.14	0.80	0.17	0.15	0.38	0.38
24	1	8	21.2	0.09	0.45	0.15	0.13	0.30	0.29
32	1	8	21.9	0.06	0.33	0.15	0.13	0.28	0.28

[a] The $\langle R_g \rangle$, Q, and S values assess the similarity between the reference ensemble UERg23 and that generated by molecular dynamics simulations with ensemble-averaged distance restraints, varying the number of replicas (N_{rep}) and the lower (L) and upper (U) bounds as shown. Q_{R_g} and S_{R_g} refer to the R_g, $Q_{w\mathrm{PRE}}$ and $S_{w\mathrm{PRE}}$ to the working PRE distance restraints, and $Q_{f\mathrm{PRE}}$ and $S_{f\mathrm{PRE}}$ to the free PRE distances.

We therefore investigated how to increase the range of structures accessible to the ensemble of structures at each point in time other than explicitly increasing the number of degrees of freedom by increasing N_{rep}. The smaller L and U are, the closer $d_{ij}^{\mathrm{calc}}(t)$ is to d_{ij}^{ref} at each step in the simulation (Figure 1B and D). Altering N_{rep} does not directly control the range of distances contributing to $d_{ij}^{\mathrm{calc}}(t)$ (i.e., the width of the distribution of distances $r_{ij,k}$ at each time-point, t). We found, however, that a wider range of distances are sampled at each point in time if N_{rep} is large (Figure 1A versus B). Increasing L and U is another indirect means of allowing more variation in $d_{ij}^{\mathrm{calc}}(t)$. Over many time-points, this variation results in a wider range of distances being sampled for a given N_{rep}. Thus, increasing the tolerance to instantaneous fluctuation in the ensemble-averaged observables compensates for the effects of using fewer replicas, thus enabling the use of a smaller number of replicas and decreasing the problem of overfitting.

In this work we exploit this equivalence to reproduce distance distributions. In the simplest case of a uniform distribution, for the time- and ensemble-averages with fewer replicas and larger L and U to be equivalent to those obtained with more replicas and smaller L and U, the $d_{ij}^{\mathrm{calc}}(t)$ over multiple time-steps must be evenly distributed within $d_{ij}^{\mathrm{ref}} - L$ and $d_{ij}^{\mathrm{ref}} + U$. If L and U are equal, the range of $d_{ij}^{\mathrm{calc}}(t)$ collected over all time points would then be evenly distributed around d_{ij}^{ref}, so that the r^{-6} average calculated from the overall distribution of $r_{ij,k}^{\mathrm{calc}}$, pooled over all N_{rep} replicas and all time-points t, will be smaller than the imposed restraint d_{ij}^{ref} because approximately half of the $r_{ij,k}$ lie between d_{ij}^{ref} and $d_{ij}^{\mathrm{ref}} - L$, and these small $r_{ij,k}$ have a disproportionately large influence on d_{ij}^{calc}. If the tolerance is to be used to compensate for using fewer replicas, then L and U must be chosen such that the overall distribution of $r_{ij,k}^{\mathrm{calc}}$ contains a smaller proportion of short distances. This can be done by choosing L to be less than U, thus favoring $d_{ij}^{\mathrm{calc}}(t) > d_{ij}^{\mathrm{ref}}$ at the expense of $d_{ij}^{\mathrm{calc}}(t) < d_{ij}^{\mathrm{ref}}$.

We tested a range of different combinations of L and U with 16, 24, and 32 replicas, using reference PRE-like distance restraints calculated from UERg23. The key results are summarized in Table 2. At the simulation temperature of $T = 515$ K, the $\langle R_g \rangle$ of an unrestrained ensemble generated using the same implicit solvent model as was used for the ensemble-restrained molecular dynamics simulations (USRg24) is similar to that of UERg23, thus any difference in the $\langle R_g \rangle$ between the restrained ensemble and UERg23 can be attributed to the effects of the restraining method. We found that with $L = 1$ and $U = 8$ we obtained the desired effect without leaving the upper bound so large that it ceased to act as a restraint. Such unbalanced bounds are reminiscent of those used for NOEs, which are also subject to r^{-6} (or in some cases, r^{-3}) averaging.

Cross-Validation with Multiple Observables. An important point reflected in our results is that it is possible to reproduce the average values of interatomic distances even if their distributions are poorly reproduced. Indeed, we found little correlation between the overall Q_{PRE} and S_{PRE} values for a given ensemble (Figure 2A). Thus the optimal conditions for the reproduction of the r^{-6} -averaged PRE-derived distances are not the same as for the reproduction of their underlying distributions. This situation can occur because many different distributions can give rise to the same average[80] (Figure 3). By contrast, two similar distributions can have a different average. This is because different types of average report on different aspects of the underlying distribution. A linear average, for instance, lies near the center of the distribution, whereas a highly

[80] Bürgi, R.; Pitera, J.; van Gunsteren, W. F. *J. Biomol. NMR* **2001**, *19*, 305–320.

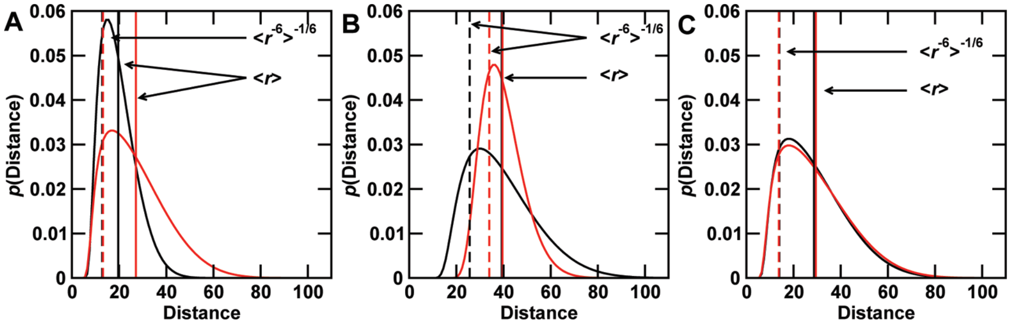

Figure 3. Examples of distance distributions and their averages. (A) Distributions having equal nonlinear averages but different linear averages. (B) Distributions having equal linear averages but different nonlinear averages. (C) Distributions having equal linear and nonlinear averages.

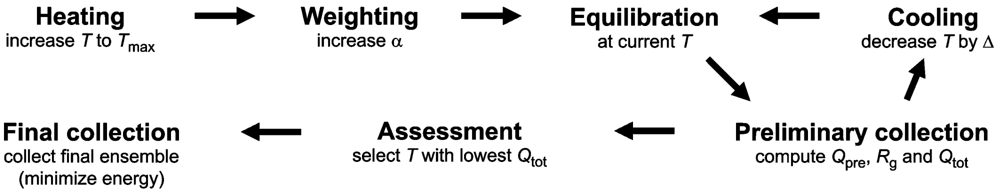

Figure 4. Protocol for carrying out molecular dynamics simulations of disordered states of proteins with ensemble-averaged distance restraints. We set $N_{\mathrm{rep}} = 24$, $L = 1$, and $U = 8$. The molecules are first heated to $T_{\max}$ (in our case 750 K) in 50 K increments, and then the force constant, α, is increased to a value that is sufficiently high that the restraints are satisfied but not so high as to cause large changes in the energy (we used $\alpha = 364, 500$ kcal mol Å^{-2}). The next three steps form a loop in which, after an equilibration phase in which the temperature is held constant, a set of structures is collected for the cross-validation test, before the temperature is lowered by $\Delta T = 25$ K and the process is repeated. We found that the 1920 structures (80 per replica) collected at each temperature were sufficient to obtain reliable estimates of Q_{R_g} and Q_{PRE} and thus Q_{tot}. At the temperature at which Q_{tot} is optimized, further structures are collected. Finally, the energy of these structures may be minimized in a more accurate solvation model such as GB/SA or explicit water at 300 K to eliminate any unfavorable local conformations.

nonlinear average such as the r^{-6} average lies toward the edge and is most influenced by the outliers in the distribution (Figure 3). Thus the poor correlation between S_{PRE} and Q_{PRE} is caused by nonmatching left-hand sides of the distributions (data not shown). The lack of correlation between Q_{PRE} and S_{PRE} constitutes a challenging problem since S values are normally known only for reference ensembles. Strategies that use experimental data can therefore often only use Q values for validation. In the following we present a way to overcome these difficulties.

If it is possible to use multiple experimental techniques to measure different types of average for the same or related observables, then these can be combined to give more information about the shape of the underlying distribution. Proof of this principle was given by Choy et al.,[81] who were able to obtain a more precise description of the R_g distribution of an unfolded protein when they fitted the distribution function to both the $\langle R_g^2\rangle^{1/2}$ obtained from SAXS and the $\langle R_h^{-1}\rangle^{-1}$ obtained from PFG-NMR simultaneously.

In this work we adopt a similar strategy, but with one difference. In our case we use observables that report on two different aspects of the structure, namely, the PRE-derived distances and R_g or, in the case of experimental data, $\langle R_h^{-1}\rangle^{-1}$. In order for the information contained in the r^{-6} -averaged PRE-derived distances and the linearly averaged R_g to be combined, the first requirement is that the distributions of each type of observable are correlated. We find that this is indeed the case since ensembles for which S_{PRE} are low also have low S_{R_g} (Figure 2B). In fact, we find that Q_{R_g} is also highly correlated with S_{PRE} (Figure 2C), most likely because the linearly averaged R_g is most sensitive to the center of the distribution, and the correlation of the S values indicates that the distributions of the two types of observable are broadly similar. These observations lead us to the conclusion that cross-validation against different types of average, which report on different aspects of the underlying distribution, provides a better measure of whether the distributions and the ensembles are correct. Thus when we use experimental data, we refer to $Q_{\mathrm{tot}} = Q_{R_g} + Q_{\mathrm{PRE}}$ to determine when we have reconstructed the ensemble correctly.

Table 3. Reproduction of the UERg23 Ensemble Using the Method Described in This Work[a]

T (K)	$\langle R_g\rangle$ (Å)	Q_{R_g}	S_{R_g}	$Q_{w\mathrm{PRE}}$	$Q_{f\mathrm{PRE}}$	Q_{tot}	$S_{w\mathrm{PRE}}$	$S_{f\mathrm{PRE}}$
500	20.4	0.12	0.635	0.15	0.15	0.42	0.31	0.31
525	21.5	0.07	0.381	0.16	0.15	0.38	0.31	0.31
550	22.5	0.03	0.215	0.16	0.16	0.35	0.30	0.29
575	23.0	0.01	0.111	0.17	0.18	0.36	0.29	0.28
590	**23.2**	**0.00**	**0.061**	**0.17**	**0.15**	**0.32**	**0.26**	**0.25**
600	23.4	0.01	0.069	0.17	0.17	0.35	0.29	0.28
625	23.6	0.02	0.094	0.18	0.19	0.39	0.29	0.27
650	24.1	0.04	0.143	0.19	0.21	0.44	0.28	0.28
675	24.4	0.05	0.196	0.19	0.24	0.48	0.29	0.29
700	24.5	0.06	0.211	0.20	0.25	0.51	0.30	0.29
725	24.4	0.05	0.243	0.22	0.25	0.52	0.30	0.29
750	24.7	0.06	0.284	0.21	0.28	0.55	0.30	0.30

[a] The $\langle R_g\rangle$, Q, and S values quantify how well UERg23 ($R_g = 23.2$ Å) is reproduced using the method illustrated in Figure 4. Q_{R_g} and S_{R_g} refer to the R_g, $Q_{w\mathrm{PRE}}$ and $S_{w\mathrm{PRE}}$ to the working PRE distance restraints, and $Q_{f\mathrm{PRE}}$ and $S_{f\mathrm{PRE}}$ to the free PRE distances. The results for the optimal T are in bold type.

General Protocol for Molecular Dynamics Simulations of Disordered States of Proteins with Ensemble-Averaged Distance Restraints from PRE Experiments. As part of the systematic study described here we have designed a protocol applicable to any type of disordered state for which both PRE measurements and a measure of the global size, such as the R_g or R_h are available (Figure 4). After the heating and equilibration

[81] Choy, W. Y.; Mulder, F. A. A.; Crowhurst, K. A.; Muhandiram, D. R.; Millett, I. S.; Doniach, S.; Forman-Kay, J. D.; Kay, L. E. *J. Mol. Biol.* **2002**, *316*, 101–112.

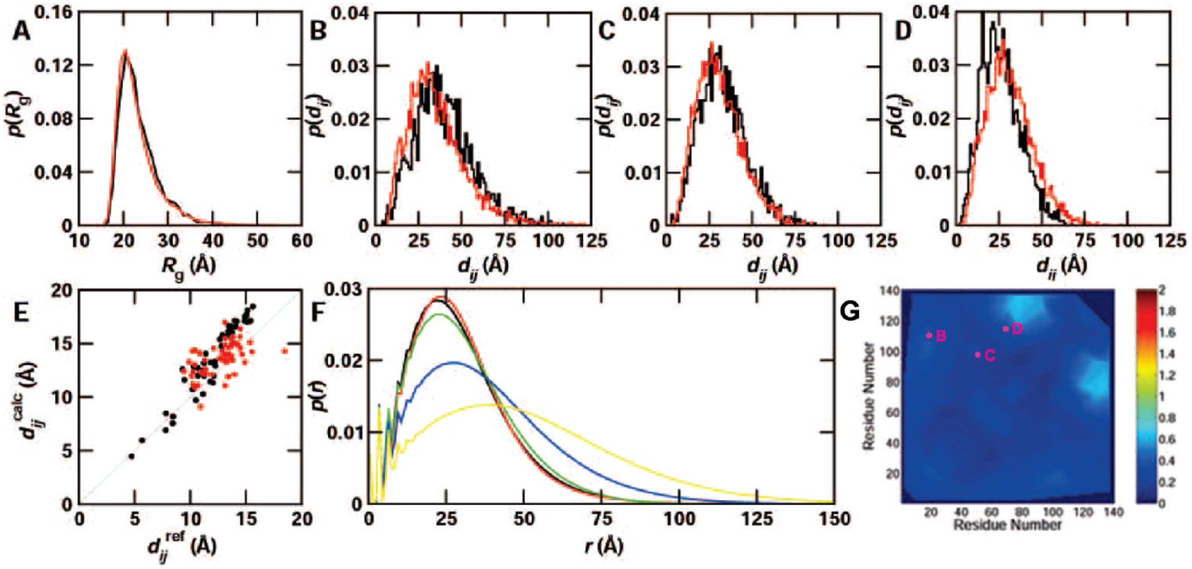

Figure 5. Comparison of the reconstructed ensemble (RSRg23) with the corresponding reference ensemble (UERg23). (A) R_g distributions. (B−D) Three examples of PRE distance distributions for residues: (B) 17−110, (C) 51−98, and (D) 68−115. (E) Scatter plot of PRE distances. For the distributions, UERg23 is shown in black and RSRg23 in red. In the scatter plot, the working data set is in black and the free data set in red. (F) Comparison of $p(r)$ for (black) UERg23, (red) RSRg23, (green) USRg24, (blue) UST590, and (yellow) RC. (G) s_l map showing the overall agreement between UERg23 and RSRg23 in terms of distributions. The locations of the pairs of residues for which the distributions are plotted in panels B−D are indicated on panel G.

phases (see Methods), we carry out an iterative process to determine the optimal simulation temperature.

In using this protocol to reproduce UERg23, we find that at $T = 590$ K there is a clear minimum in the Q and S values (Table 3). We designate the ensemble produced at this T RSRg23. At the global level, the $\langle R_g \rangle$ of RSRg23 is perfectly matched to that of the reference ensemble. Moreover, the reproduction of the distance distributions is particularly good (Table 3 and Figure 5B−D); only three distance distributions are shown in Figure 5, as it is not practical to examine all of them individually in this way. Instead, we created an s_l map, which provides a means of visualizing the quality of the reconstruction of all interatomic distance distributions (Figure 5G). The color of each point represents how similar the distance distributions between the residues indicated on the axes are for UERg23 and RSRg23. Even for pairs of residues near the region of highest s_l values (e.g., D), the distance distributions are in fact quite similar (Figure 5D).

In addition to the S_{PRE} value, a graphical representation of how well the distributions were reproduced throughout is provided by the overall pairwise distance distribution function, $p(r)$. This is one of the few experimentally accessible distribution functions, obtained by taking the sine Fourier transform of the SAXS scattering profile of a protein in solution.[82,83] The experimental $p(r)$ includes contributions from all pairs of interatomic distances within a macromolecule. Here, it was approximated by considering only $C_\alpha-C_\alpha$distances to reduce the computational cost. This approximation is justified because we only compare reference $p(r)$ distributions.

As well as comparing the $p(r)$ for UERg23 and RSRg23, we also consider the $p(r)$ of RC, USRg24, and UST590 to show the degree of difference in $p(r)$ to be expected between different ensembles and thus emphasize how similar our reference and reconstructed ensembles are. For small values of r, the $p(r)$ of all of the ensembles overlay, with two well-defined peaks at about 4 and 7 Å, respectively, corresponding to residues close together in sequence (Figure 5F). Thereafter, the $p(r)$ for RC is considerably flatter and broader than the $p(r)$ of the other ensembles, as is expected given its much larger $\langle R_g \rangle$ (Table 1). The much broader and flatter $p(r)$ of UST590 compared to that of RSRg23 reveals the extent of the compaction effects induced by the application of PRE-derived distance restraints. The $p(r)$ of UERg23 and USRg24 are similar. However the $p(r)$ of RSRg23 provides an even closer match to that of UERg23, indicating that the application of PRE-derived distance restraints provides additional information not present in the effective energy function defined by the force field and solvent model.

To test the generality of the protocol, we also reconstructed a more compact reference ensemble, UERg20, using 24 replicas with $L = 1$ and $U = 8$. We were again able to accurately reproduce the reference ensemble in terms of distributions and averages (data not shown), indicating that this method is applicable to different types of ensembles. The optimal T in each case depended on the broadness of the ensemble being reproduced and the compactness of the structures, being lower for narrower ensembles and more compact structures.

On the Definition of an Ensemble of Structures. The structure of a protein is readily defined in terms of the positions of its atoms. By contrast, it is much more difficult to define the ensemble of structures representing its thermal fluctuations. Such an ensemble contains many structures, thus it is impractical to provide the atomic coordinates of all of them together with their statistical weights.

One way of defining ensembles of conformations is in terms of the probability distributions of interatomic distances. In this definition, two ensembles are equal if all the probability distributions are equal. In terms of S values (see Methods), equal ensembles result in $S = 0$. This definition becomes ambiguous, however, in the presence of correlated motions as then two ensembles may differ even if all their distance distributions are

[82] Bilsel, O.; Matthews, C. R. *Curr. Opin. Struct. Biol.* **2006**, *16*, 86–93.
[83] Svergun, D. I.; Koch, M. H. J. *Rep. Prog. Phys.* **2003**, *66*, 1735–1782.

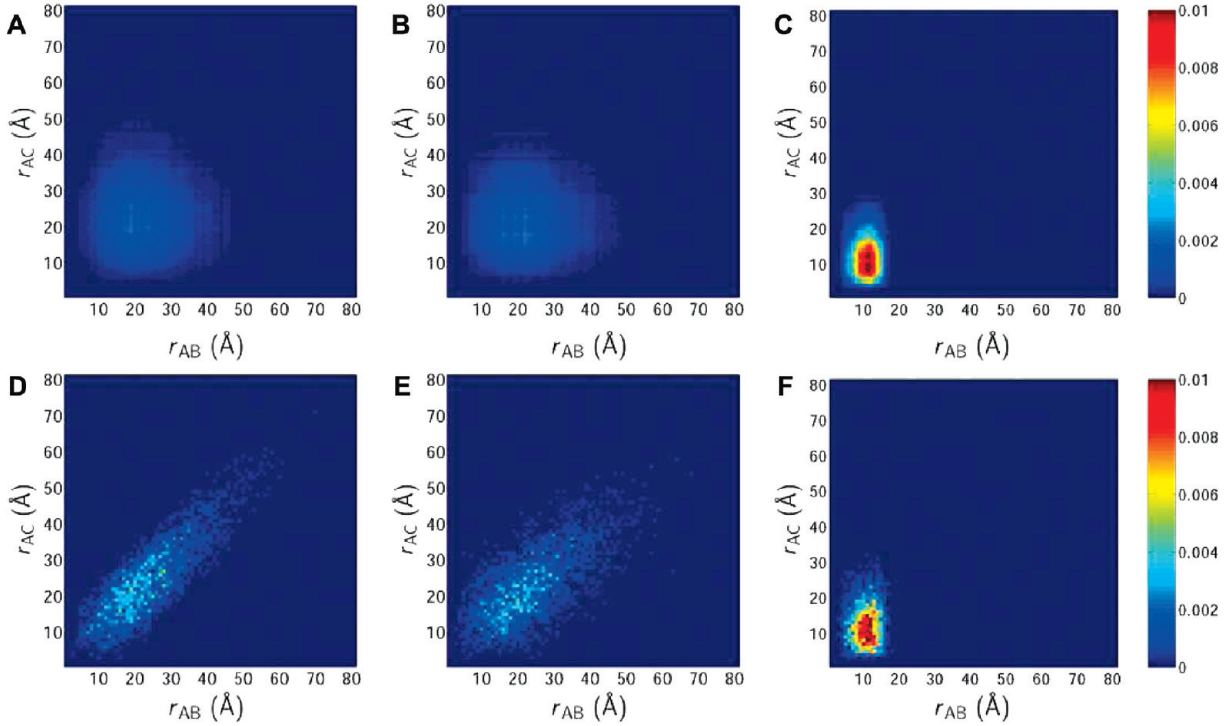

Figure 6. Identification of the presence of correlated motions in an ensemble. (A−C) Products of the distance distributions $p(r\mathrm{AB})*p(r\mathrm{AC})$. (D−F) Joint probability distributions $p(r\mathrm{AB}, r\mathrm{AC})$ for UERg23. The triplets of residues used: (A and D) A = 1, B = 105, C = 116 (s_l = 0.86); (B and E) A = 1, B = 71, C = 130 (s_l = 0.50); (C and F) A = 82, B = 71, C = 130 (s_l = 0.23).

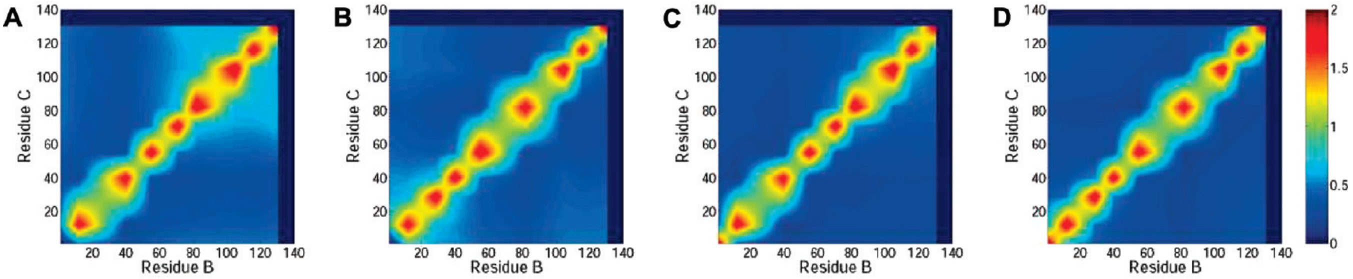

Figure 7. Representation of the s_p values, which we used to quantify the agreement between $p(r\mathrm{AB}, r\mathrm{AC})$ and $p(r\mathrm{AB})*p(r\mathrm{AC})$. (A and C) A = 29; (B and D) A = 71 for (A and B) the UERg23 ensemble and (C and D) the RSRg23 ensemble.

equal. This possibility is illustrated in Figure 6, where the joint distributions, $p(r\mathrm{AB}, r\mathrm{AC})$, and the product of the distributions, $p(r\mathrm{AB})* p(r\mathrm{AC})$, are plotted for three different combinations of pairs of atoms exhibiting a range of s_l values. In the presence of correlations, the joint distribution exhibits an elongated shape, whereas the product remains roughly spherical.

To investigate the presence of correlations in our ensembles, we used s_p values to quantify the similarity between the 2D histograms of $p(r\mathrm{AB}, r\mathrm{AC})$ and $p(r\mathrm{AB})*p(r\mathrm{AC})$ and then plotted these s_p values as 2D contour plots (Figure 7). In UERg23, there are few correlations, other than those arising from the persistence length of the polypeptide chain. RSRg23 is essentially identical in this respect. The lack of correlations means that in this case, if all of the probability distributions are the same, as demonstrated by the low s_l values when comparing these two ensembles, then they can be considered to be equal.

Free Energy Landscapes. Free energies calculated from molecular dynamics simulations with experimentally derived restraints are different from those computed from unrestrained simulations as a result of the addition of the restraint energy potential to bias the conformational search. Since the restraint term can be viewed as a correction to the underlying force field, an estimate of the free energy can be made by computing the population with respect to a given coordinate or set of coordinates.[72] In this way, we generated free energy landscapes for the reference ensemble UERg23, an unrestrained ensemble generated using the same implicit solvent model as when the restraints are applied and containing structures of a similar size to the reference ensemble (USRg24), the reconstructed ensemble RSRg23 and the random coil ensemble (Figure 8). The very good agreement between the free energy landscapes of UERg23 and RSRg23, in combination with the very different free energy landscapes of URERg24 and RC shows that adding a penalty energy function based on distance restraints to the existing force field is able to alter the resulting ensemble of structures to match that from which the restraints were derived.

Application of the Method Using Experimentally Derived PRE Distances. So far in this work we have developed and discussed a protocol for performing ensemble-restrained molecular dynamics simulations that is capable of accurately reconstructing disordered state ensembles. We now apply this procedure to characterize the natively unfolded protein α-sy-

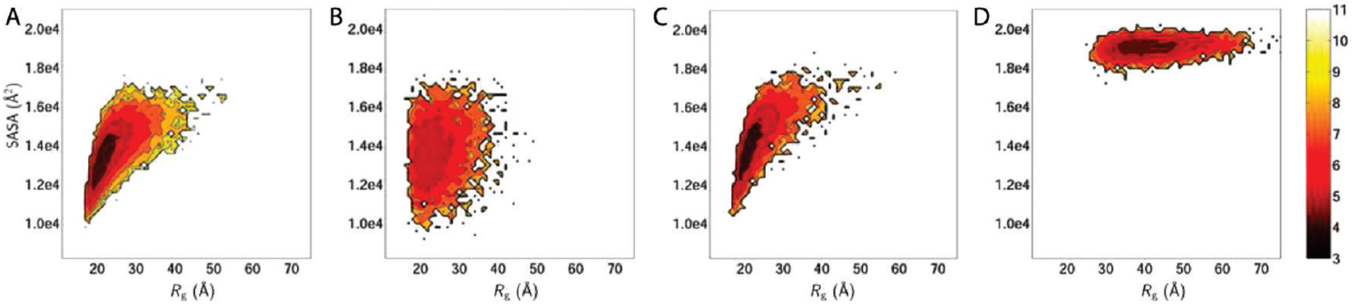

Figure 8. Comparison of the free energy landscapes of various ensembles. (A) UERg23, (B) USRg24, (C) RSRg23, and (D) RC. The free energy is defined as $F(R_g, \text{SASA}) = -\ln p(R_g, \text{SASA})$. For comparison, we report in Supporting Information the free energy landscape as a function of R_g and the end-to-end distance R_{ee}.

nuclein using distances derived from PRE experiments. This task has been carried out previously using a similar method[11] but with some important differences.

First, the previous ensemble of structures of α-synuclein was biased to have an overall size similar to that of α-synuclein in D_2O at 298 K ($\langle R_h^{-1}\rangle^{-1} = 26.6$ Å).[74] The PRE measurements, however, were carried out in phosphate buffer with 100 mM NaCl at 283 K. Subsequent measurements in Mes-Na buffer with 100 mM NaCl at 288 K revealed that α-synuclein is more expanded in the presence of salt ($\langle R_h^{-1}\rangle^{-1} = 32.0$ Å).[84] We therefore used a value of $\langle R_h^{-1}\rangle^{-1} = 32.0$ Å to calculate Q_{R_h}. In addition, we used a further 119 distances obtained from a spin label positioned at residue N122 that were not available previously. We chose the simulation temperature according to our general protocol (Figure 4). We only included distances calculated from $0.15 < I_{ox}/I_{red} < 0.85$ (for which both L and U were applied) for the calculation of Q_{PRE}, as the distances calculated from the experimental PRE data are not reliable for I_{ox}/I_{red} outside these bounds. We found that Q_{PRE} changes very little with T (Table 4), highlighting how insensitive the r^{-6} average is to the nature of the underlying distribution. Q_{R_h} therefore provided the greatest contribution to Q_{tot}, which is the means by which we chose the optimal simulation temperature (490 K).

The ensemble of structures that we determined is characterized by a broad R_g distribution (Figure 10A), indicative of the wide range of different structures populated by α-synuclein. It is wider and shifted toward larger values of R_g relative to the previously obtained ensemble, in keeping with the larger $\langle R_h^{-1}\rangle^{-1}$. However, the R_g distribution of RC contains larger R_g values and is broader than that of either of the ensembles produced using ensemble-restrained molecular dynamics simulations,

Table 4. Selection of the Simulation Temperature Using Experimental Restraints for α-Synuclein[a]

T (K)	$\langle R_h^{-1}\rangle^{-1}$ (Å)	Q_{R_h}	Q_{wPRE}	Q_{fPRE}	Q_{tot}
475	31.3	0.018	0.18	0.21	0.41
490	**32.1**	**0.006**	**0.19**	**0.20**	**0.40**
500	32.6	0.021	0.19	0.22	0.42
525	33.2	0.042	0.19	0.20	0.43
550	33.7	0.056	0.19	0.20	0.45
575	34.2	0.069	0.20	0.19	0.46
600	34.3	0.085	0.20	0.20	0.49
625	34.5	0.081	0.20	0.20	0.48
650	34.6	0.085	0.21	0.20	0.50
675	34.8	0.092	0.21	0.20	0.50
700	34.9	0.094	0.21	0.21	0.51

[a] The Q values quantify how well the experimental $\langle R_h^{-1}\rangle^{-1}$ (32.0 Å) and the PRE distances for α-synuclein are reproduced using the method illustrated in Figure 4. Q_{R_h} refers to the $\langle R_h^{-1}\rangle^{-1}$, Q_{wPRE} to the working data set (80% of the PRE distances), and Q_{fPRE} to the free data set (remaining 20%). The results for the reconstructed ensemble at the selected T are in bold type.

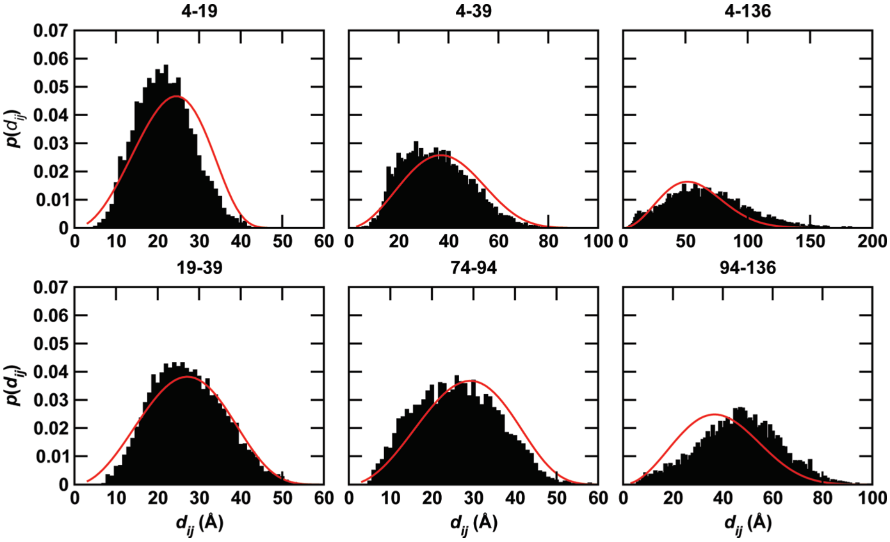

Figure 9. Comparison of the distance distributions from experimental electron transfer (ET) data with those calculated here from the ensemble of α-synuclein structures. The red lines are the fit of the worm-like chain model to experimental electron transfer (ET) data,[70] and the black bars show the distributions calculated from the ensemble of α-synuclein structures obtained using ensemble-restrained molecular dynamics simulations with experimental distance restraints; residue pairs are shown above each graph.

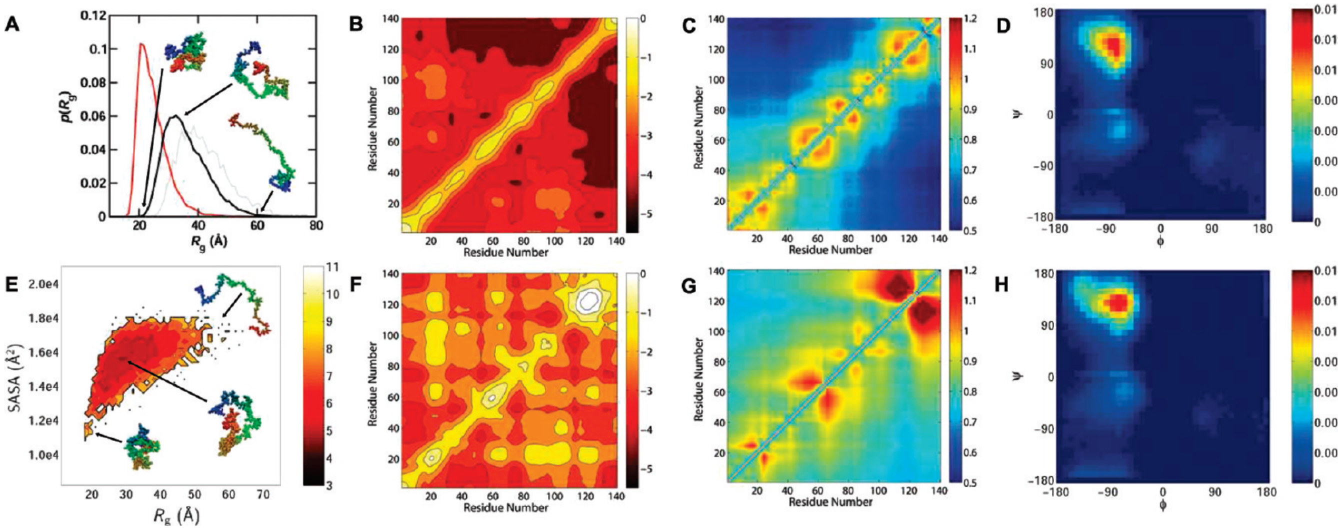

Figure 10. Analysis of the ensemble of α-synuclein structures obtained using experimental PRE distances as ensemble-averaged restraints in molecular dynamics simulations. (A) R_g probability distribution for (black) the ensemble of α-synuclein structures obtained here, (red) the ensemble previously obtained with a smaller value of R_g,[11] and (gray) a random coil ensemble. (B and F) RCP maps for the previous[11] (B) and present (F) α-synuclein ensembles. (C and G) DC maps for the previous[11] (C) and present (G) α-synuclein ensembles. (E) Free energy landscape of α-synuclein, where the free energy is defined as $F(R_g, \text{SASA}) = -\ln p(R_g, \text{SASA})$. (D and H) Ramachandran plots of the backbone dihedral angles averaged over the most helical portion of the N-terminus identified experimentally, residues 6−37 (D) and the C-terminus, residues 103−140 (H).

indicating that the size and range of structures accessible to α-synuclein in solution is reduced relative to a random coil.

To validate our ensemble of α-synuclein structures, we calculated distributions of inter-residue distances and compared them to the distributions fitted to the results of electron transfer (ET) measurements.[70] This is an important test of the ability of our method to reproduce distributions as well as averages. It should be noted, however, that the nature of the fitted distributions depends on the fitting method used. Nevertheless, our calculated distributions are in reasonably good agreement with those determined experimentally (Figure 9), thus providing additional support for the validity of our ensemble and our general method.

The nature of the structures comprising the ensembles obtained previously and here are summarized by residual contact probability (RCP)[11] and distance comparison (DC) maps (Figure 10B and C). RCP maps, which were used in our previous characterization of α-synuclein,[11] show the probability of occurrence of distances shorter that 8.5 Å, whereas DC maps are sensitive to the position of the center of the inter-residue distance distribution and thus the most frequently observed inter-residue distances.

As expected, as a result of the more expanded structures sampled, the pseudoenergies shown in the RCP maps are lower and the values in the DC map are larger for the present ensemble than for the previous ensemble. The RCP map of the previous α-synuclein ensemble[11] and the PRE data[10,11] suggest preferential contact formation between the C-terminus and the central NAC region. These contacts are still present but less pronounced in the ensemble of structures generated here, whereas the contact probability between residues 1−30 and the remainder of the sequence remain similar. In addition, measures sensitive to longer distances, such as the filtering method used by Bernardo et al.[60] and the DC measure used here, show that the distances between the N- and C-termini are also short relative to those expected for a random coil. These shorter distances are not so apparent in the RCP maps as the large sequence separation between the residues involved precludes the existence of many distances less than 8.5 Å. Interestingly, the shortest inter-residue distances involve the region around residue 120. This result is in agreement with a range of experimental data, including chemical shifts, R_1 and R_2 relaxation rates, heteronuclear NOEs[55,85] and in particular RDCs,[10] all of which contain anomalies in this region.

Other features of the DC maps also correlate with experimental data. A DC value near 1.0 for residues close together in sequence in the N-terminus may be indicative of residual helical structure, as the rmsd for an α-helix and a random coil are indistinguishable for separations of up to eight residues.[86] This observation is in agreement with the experimental data for α-synuclein in solution, which suggest some helical propensity in the N-terminus.[43,55] The larger DC values in the C-terminus, which could correspond to either β or PPII structure,[86] are most likely attributable to β structure, as the experimental data for α-synuclein does not indicate a tendency to form PPII structure.[55] Ramachandran plots of the ensemble-averaged dihedral angle preferences confirm that the region of the N-terminus identified as being most helical experimentally has a slightly greater propensity to form helical structure than the C-terminus (Figure 10 E and F).

The relatively short distances between the N- and C-termini, which are likely to have an electrostatic origin given the opposite charges of the termini, may play a role in preventing aggregation under normal solution conditions by reducing the accessibility of the central hydrophobic NAC region. Such an effect would be in keeping with experimental data suggesting that the aggregation properties of α-synuclein are mediated by a subtle interplay between charged residues distributed throughout the sequence.[52−54,57,57,87−90]

The free energy landscape of α-synuclein as a function of R_g and SASA (Figure 10 E) shows that α-synuclein mostly populates conformations whose R_g and SASA are between 20% and 50% greater than those of compact globular states. The most populated region of the free energy landscape includes structures with low R_g but high SASA, which may be advantageous for increasing the binding surfaces of α-synuclein within the crowded cellular environment. This feature has been suggested to provide a functional advantage for a number of natively unfolded proteins[3,91] but may also increase the risk of aggregation.

Conclusions

We have presented a general procedure for generating free energy landscapes of disordered states of proteins using molecular dynamics simulations with ensemble-averaged distance restraints derived from spin label NMR measurements. The validity of this approach was demonstrated by a systematic investigation in which known reference ensembles were reconstructed with high accuracy in terms of both averages and distributions of a range of observables. We then presented an application of the method to calculate the free energy landscape of the natively unfolded state of α-synuclein, which was validated by showing an excellent agreement with distance distributions obtained from electron transfer experiments. The method that we have presented is general and can be applied to highly heterogeneous states of proteins for which spin label NMR measurements have been carried out.

Acknowledgment. This work was supported by the Woolf Fisher Trust (J.R.A.), the Leverhulme Trust (P.V., C.M.D., M.V.), the Wellcome Trust (P.V., C.M.D.), the European Molecular Biology Organisation (M.V.), and the Royal Society (M.V.).

Supporting Information Available: Free energy landscape as a function of R_g and the end-to-end distance R_{ee}. This material is available free of charge via the Internet at http://pubs.acs.org.

JA904716H

(84) Binolfi, A.; Rasia, R. M.; Bertoncini, C. W.; Ceolin, M.; Zweckstetter, M.; Griesinger, C.; Jovin, T. M.; Fernandez, C. O. *J. Am. Chem. Soc.* **2006**, *128*, 9893–9901.

(85) Bussell, R.; Ramlall, T. F.; Eliezer, D. *Protein Sci.* **2005**, *14*, 862–872.

(86) Zagrovic, B.; Pande, V. *Nat. Struct. Biol.* **2003**, *10*, 955–961.

(87) Biere, A. L.; Wood, S. J.; Wypych, J.; Steavenson, S.; Jiang, Y. J.; Anafi, D.; Jacobsen, F. W.; Jarosinski, M. A.; Wu, G. M.; Louis, J. C.; Martin, F.; Narhi, L. O.; Citron, M. *J. Biol. Chem.* **2000**, *275*, 34574–34579.

(88) Antony, T.; Hoyer, W.; Cherny, D.; Heim, G.; Jovin, T. M.; Subramaniam, V. *J. Biol. Chem.* **2003**, *278*, 3235–3240.

(89) Fernandez, C.; Hilty, C.; Wider, G.; Guntert, P.; Wuthrich, K. *J. Mol. Biol.* **2004**, *336*, 1211–1221.

(90) Murray, I. V. J.; Giasson, B. I.; Quinn, S. M.; Koppaka, V.; Axelsen, P. H.; Ischiropoulos, H.; Trojanowski, J. Q.; Lee, V. M. Y. *Biochemistry* **2003**, *42*, 8530–8540.

(91) Gunasekaran, K.; Tsai, C. J.; Kumar, S.; Zanuy, D.; Nussinov, R. *Trends Biochem. Sci.* **2003**, *28*, 81–85.