# Deciphering Protein Dynamics from NMR Data Using Explicit Structure Sampling and Selection

Yiwen Chen,*† Sharon L. Campbell,* and Nikolay V. Dokholyan*

*Department of Biochemistry and Biophysics, †Department of Physics and Astronomy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

ABSTRACT Perhaps one of the most prominent realizations of recent years is the critical role that protein dynamics plays in many facets of cellular function. While characterization of protein dynamics is fundamental to our understanding of protein function, the ability to explicitly detect an ensemble of protein conformations from dynamics data is a paramount challenge in structural biology. Here, we report a new computational method, Sample and Select, for determining the ensemble of protein conformations consistent with NMR dynamics data. This method can be generalized and extended to different sources of dynamics data, enabling broad applicability in deciphering protein dynamics at different timescales. The structural ensemble derived from Sample and Select will provide structural and dynamic information that should aid us in understanding and manipulating protein function.

## INTRODUCTION

Protein dynamics is intimately linked to molecular function, as motions over a wide range of timescales contribute to molecular recognition, i.e., protein-ligand (1), protein-protein (2–4), protein-DNA interactions (5), and protein activity (such as enzyme catalysis (6,7)). Hence, the detailed characterization of protein dynamics is of great importance for elucidating protein function (8). Nuclear magnetic resonance (NMR) spectroscopy is uniquely suited for characterization of protein dynamics, as it can provide site-specific dynamic information over a wide range of timescales (9–11). However, the ability to obtain explicit structural information on conformational ensembles dynamically sampled by a protein of interest remains a difficult task. All-atom molecular dynamics (MD) (12) simulations have shown increased utility in providing atomic level information on the conformations that are sampled by a protein on a picosecond-nanosecond timescale (13), and have recently facilitated structural interpretation of dynamics data obtained from both backbone and side-chain NMR relaxation experiments (13,14).

$^1$H-$^{15}$N NMR relaxation experiments are commonly used to measure $^{15}$N T1, T2, and nuclear Overhauser enhancement for backbone NH resonances in proteins (9,11). These parameters can be expressed in terms of a spectral density function and often interpreted based on a model-free approach proposed by Lipari and Szabo (15) to obtain order parameters ($S^2$), which reflect the fluctuations of a backbone N-H bond vector due to its internal motion. $S^2$ has limiting values of zero and unity, corresponding to isotropic motion or complete rigidity, respectively (8,11). Additional parameters, such as the overall tumbling motion and correlation time for internal motion of the molecule, can be also derived. Although less commonly utilized, similar approaches can be employed to extract order parameters for C-C ($H_3$) bond vectors (9).

Although MD simulations have enabled explicit structural and dynamic elucidation of experimentally derived order parameters (16–20), the molecular mechanics force field used in all-atom MD simulations is an approximation of the more accurate quantum mechanic-based description of the interactions in proteins. Thus, the errors in the force field can give rise to inconsistencies between protein conformations derived from MD simulations and those generated by hybrid quantum mechanics/molecular mechanics simulations or those determined using experimental constraints (21). These errors can lead to biased description of protein dynamics and difficulties in reproducing experimentally derived order parameters (20). To better characterize the protein conformations generated from MD simulations, experimentally derived order parameters have been used in recent studies as constraints in MD simulations to bias sampling toward conformations that are more consistent with experimental data (14,22).

Here, we propose a new computational method, Sample And Select (SAS), that employs MD simulations (or other sampling methods) to determine conformational ensembles consistent with NMR dynamics data. We report the evaluation and validation of the SAS method using NMR-derived order parameters. Unlike other constraint-based methods (14,22,23), where constraint-driven forces such as NMR data are directly incorporated into the MD simulation, conformational sampling is completely decoupled from selection of experimentally consistent conformations in the SAS method. Thus, the SAS method enables integration of different sampling methods and dynamics data to obtain a better representation of the structural ensemble to be consistent with experimental data, as long as computationally and experimentally sampled dynamics are on a comparable timescale.

*Submitted January 9, 2007, and accepted for publication June 5, 2007.*

Address reprint requests to N. V. Dokholyan, Tel.: 919-843-2513; E-mail: dokh@med.unc.edu.

Editor: Arthur G. Palmer III.

0006-3495/07/10/2300/07 $2.00

doi: 10.1529/biophysj.107.104174

This method is particularly useful when dynamics data (e.g., residual dipolar coupling data (24,25)) is difficult to be directly incorporated as constraints in sampling methods such as MD simulation (12,26). Another advantage of the SAS method is that it allows an explicit detection of insufficiencies/errors in conformational sampling. For example, if a computationally generated ensemble that correlates well with the experimental data cannot be obtained, it can be inferred that the conformational sampling by this method is either insufficient or erroneous. This type of information may be especially useful for improving current molecular mechanics force fields (12) using NMR dynamics data.

## METHODS

### SAS method

When the SAS method is used in conjunction with NMR-derived order parameters, either an x-ray or low energy NMR solution structure is first used as the starting structure in all-atom nanosecond MD simulations, to generate a large ensemble of conformations for the protein of interest. Next, a Monte Carlo simulated annealing procedure is employed to select conformations most consistent with the NMR-derived order parameters. The selection procedure does not rely on any specific type of sampling method or experimental data.

### Order parameter calculation

The angular correlation function (15,20) for a given bond vector between two atoms is defined as

$$C_{\mathrm{I}}(\tau) = \langle P_2(\vec{\mu}(t) \times \vec{\mu}(t+\tau))\rangle_{\mathrm{t}}, \qquad (1)$$

where $\vec{\mu}(t)$ is the time-dependent unit vector pointing along the same direction of bond vector, $P_2(x) = 1/2(3x^2 - 1)$ is the second Legendre polynomial, and $\langle \ldots \rangle_{\mathrm{t}}$ denotes the average over time. The order parameter $S^2$ (15,20) is defined as the long-time limit of the angular correlation function

$$S^2 = \lim_{\tau \to +\infty} C_{\mathrm{I}}(\tau). \qquad (2)$$

The order parameter $S^2$ can be calculated as the following ensemble average from simulations (20),

$$S^2 = \frac{3}{2}(\langle x^2\rangle^2 + \langle y^2\rangle^2 + \langle z^2\rangle^2 + 2\langle xy\rangle^2 + 2\langle xz\rangle^2 + 2\langle yz\rangle^2) - \frac{1}{2}, \qquad (3)$$

where $x$, $y$, and $z$ are the components of the unit bond vector of the same bond along three Cartesian axes, and $\langle \ldots \rangle$ denotes the average over an ensemble of conformations. Equations 2 and 3 are equivalent only when the simulations are sufficiently long or adequately sample conformational space. Before order parameter calculations, each conformation within the ensemble is superimposed onto the starting structure based on all $C_\alpha$ atoms to remove translational and rotational degrees of freedom.

### All-atom MD simulation

All-atom MD simulations were performed on five proteins: TNfn3 (27), $\beta$ARK1 PH domain (17), ubiquitin, lysozyme (28), and eglin c (16) (Table 1) in explicit water (TIP3P model) using the ff99 force field and AMBER 8 package (29). The lowest energy NMR solution structure was used for MD simulations of the $\beta$ARK1 PH domain and eglin c, whereas x-ray structures were employed for ubiquitin, lysozyme, and TNfn3. Since order parameters of eglin c were obtained on a mutant that contains a Phe-to-Trp mutation at position 10, the same mutation was made for eglin c in our simulation. The following protocol was employed for all five proteins.

**TABLE 1 Comparison of protein conformational consistency determined by unconstrained MD and the SAS method with NMR-derived backbone order parameters**

| Name (length in aa) | PDB code | Unconstrained MD | SAS |
|---|---|---|---|
| Eglin c (70) | 1EGL | 0.3–0.5 | 0.96–0.98 |
| Ubiquitin (76) | 1UBQ | 0.6–0.8 | 0.99 |
| TNfn3 (90) | 1TEN | 0.2–0.5 | 0.98–0.99 |
| $\beta$ARK1 PH domain (119) | 1BAK | 0.6–0.8 | 0.97–0.98 |
| Lysozyme (129) | 1JEF | 0.6–0.7 | 0.99 |

The numbers shown in the third and fourth columns are the Pearson correlation coefficients between calculated and NMR-derived values.

As indicated above, the experimentally determined protein structure was used as the starting structure for the simulations (Table 1). Counterions were added to neutralize the system based on the initial net charge of the system. Before the MD simulations, a 1000-step energy minimization procedure was employed on the protein using harmonic restraints with a spring constant of 500 kcal/(mol × Å$^2$) and then another 1000-step energy minimization was employed using similar restraints but on water molecules rather than the protein. These two rounds of minimization were then followed by a series of minimizations with harmonic restraints of decreasing spring constants of 500, 200, 100, 50, and 10 kcal/(mol × Å$^2$) on the protein. Before the equilibration simulation, a 100-ps MD simulation was performed on the protein using an initial temperature of 0.5 K and a final temperature of 300 K with harmonic restraints of spring constant 500 kcal/(mol × Å$^2$). The bond length (only for bonds involving H atoms) was fixed by the SHAKE algorithm (30) and constant temperature maintained by the weak-coupling algorithm (31). After the simulation reaches equilibrium, the last 3-ns trajectory was used to generate the structural ensembles. In total, 5000 structures were collected for further selection through a Monte Carlo simulated annealing procedure (see below). Using the same protocol, we generated two more trajectories starting from different randomized initial velocities for each protein.

### Select a conformational ensemble most consistent with experimental data

For each trajectory, we use a Monte Carlo (MC) simulated annealing (32) procedure to select a fixed number ($N$) of conformations as a representative ensemble from the 5000 conformations by minimizing the following object function $\chi^2$,

$$\chi^2 = \frac{1}{L}\sum_{\mathrm{i}=1}^{\mathrm{L}}(S^2_{\mathrm{i,cal}} - S^2_{\mathrm{i,exp}})^2, \qquad (4)$$

where $S^2_{\mathrm{i,cal}}$ and $S^2_{\mathrm{i,exp}}$ are simulated and experimentally derived order parameters of the $i^{\mathrm{th}}$ bond vector, respectively. $L$ is the total number of experimentally derived order parameters in the dataset. The simulated order parameters are calculated as ensemble averages over $N$ selected conformations. We start each MC simulation by first randomly selecting $N$ conformations as the initial ensemble. An individual MC move in our simulation consists of a random swapping between previously selected and unselected conformations. If after the swapping move, the new object function $\chi^2_{\mathrm{new}}$ is no larger than the old one $\chi^2_{\mathrm{old}}$, we accept the move; otherwise, we accept the move with probability $P = e^{(\chi^2_{\mathrm{old}} - \chi^2_{\mathrm{new}})/\mathrm{T}}$, where $T$ is an effective temperature of the system (32). We start the simulation at a high temperature $T_0$, where the acceptance ratio of MC moves is $>0.95$, and decrease the system temperature in the following way until reaching a low temperature $T_{\mathrm{f}}$, where the acceptance ratio of Monte Carlo is $<1 \times 10^{-5}$:

$$T_{\mathrm{i}+1} = \alpha T_{\mathrm{i}},\ \alpha = \left(\frac{T_{\mathrm{f}}}{T_0}\right)^{\frac{1}{\mathrm{n}}}(i = 1, 2 \ldots n). \tag{5}$$

In the equation above, $n$ is the total number of temperature decreases from $T_{\mathrm{f}}$ to $T_0$ in each simulation, and typically is 100. At each temperature $T_{\mathrm{i}}$ ($i = 1, 2 \ldots n$), we perform $1 \times 10^7$ MC moves. When the MC simulation started from a different initial random ensemble, the final ensembles have >90% overlap with each other, indicating good convergence of MC simulations. For all the calculations involved in this study, we fix the conformation number $N$ to 40. To evaluate the dependence of the results on the number of conformers selected by the SAS method, we plot, in the case of ubiquitin, the minimal value of the object function $\chi^2$ as a function of $N$. We find that the minimal values are similar to each other when $N$ is equal to 40, 60, 80, or 100. There is an increase of the minimal value of the object function $\chi^2$ when $N$ increases (see Supplementary Material), suggesting that there is a small subset of conformations in a given trajectory which best fits the experimentally derived order parameter.

To evaluate the effect of sampling on the determined conformations, all conformations generated by three independent simulations ($3 \times 5000 = 15{,}000$ conformations) are combined and the same MC procedure used to select $N$ conformations that best fit the experimentally derived order parameters. The resultant ensembles yield better correlations between calculated and experimentally derived values than the ones generated using individual simulations (see Results and Discussion).

## Complete-linkage clustering

To perform clustering of structures in this study, we define the distance between two structures as their backbone ($C_\alpha$ atoms) root-mean-square-deviation (RMSD). The smaller the RMSD from each other, the greater is the similarity between two structures. The complete-linkage clustering (33) proceeds by first finding the two entities that have the minimal distance between them. By joining those two entities into a cluster, the minimal distance between two entities is searched, and those entities that have already been clustered as a single unit is taken. This process is repeated until there are no more entities to cluster. In complete-linkage clustering, the distance between two clusters is defined as the maximal distance between any two members from the two clusters. All the clustering in this work is performed using the program OC (34).

## Entropy calculation

The distributional entropy (35) of selected conformations over the structure clusters is defined in as

$$I = -\sum_{\mathrm{i}=1}^{\mathrm{N}} P_{\mathrm{i}} \ln P_{\mathrm{i}}, \tag{6}$$

where $P_{\mathrm{i}}$ is the probability of finding a selected structure in the $i^{\mathrm{th}}$ cluster, $N$ is the total number of clusters and the sum is over all clusters. We use $\theta_{\mathrm{obs}} = I_{\mathrm{obs}}/I_{\mathrm{ran}}$ as a measure to quantify how structurally diverse the representative conformations are in conformational space. $I_{\mathrm{obs}}$ is the observed entropy of the distribution of representative conformations over the structure clusters, and $I_{\mathrm{ran}}$ is the entropy if the conformations are randomly distributed. When is $\theta_{\mathrm{obs}}$ close to unity, the selected conformations are uniformly distributed in the conformational space; when $\theta_{\mathrm{obs}}$ is close to zero, the selected conformations are dominantly populated by just one cluster.

## RESULTS AND DISCUSSION

We first compared the SAS to unconstrained MD methodology using five different proteins ranging from 70 to 129 amino acids (16,17,27,28,36,37) for which backbone order parameters are available. In contrast to SAS and constraint-based methods, unconstrained MD simulations do not make use of experimental constraints. We find that for all five proteins, the SAS method yields higher correlations between calculated and experimentally derived order parameters (Table 1). For example, the Pearson correlation coefficients between the order parameters calculated from unconstrained MD simulations and NMR-derived backbone order parameters for the third fibronectin type III domain of human tenascin (TNfn3), are only between 0.2 and 0.5 due to the fact that the $S^2$ values for some residues are much lower in the simulations than the ones derived from experiments. This issue has been recognized to be a common problem with the AMBER force field (38) (Fig. 1 *a*). The correlation coefficients are comparable to those obtained from earlier studies of TNfn3 using unconstrained MD simulations of similar lengths but using a different force field (14). In contrast, order parameters calculated from SAS-generated conformational ensembles have a correlation coefficient of >0.98 with NMR-derived $S^2$ values (Fig. 1 *b*). The significantly increased correlation indicates that by incorporating the experimentally derived information, the SAS method is able to generate a better representation of the structural ensemble on a picosecond-nanosecond timescale relative to unconstrained MD simulations.

To characterize the structural diversity and the distribution of the representative backbone conformations generated by

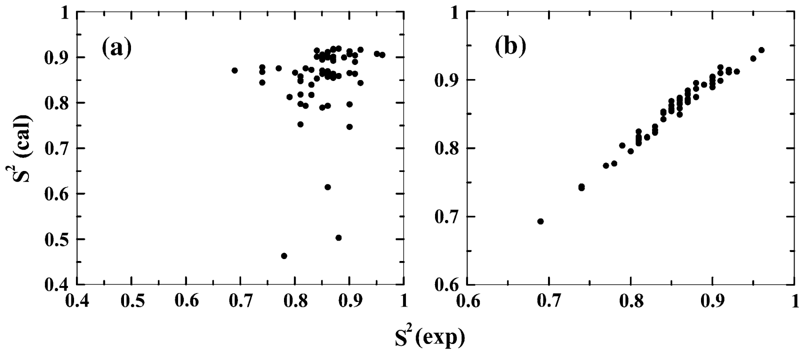

FIGURE 1 Comparisons of calculated and experimental backbone order parameters for TNfn3. (*a*) Calculation is performed using unconstrained MD ($r = 0.20$) and (*b*) calculation is performed by SAS ($r = 0.99$). Both results are obtained from one (same) of the three 3-ns MD trajectories.

the SAS method over the entire conformational space sampled by unconstrained MD, we perform complete-linkage clustering of the structures (Methods) generated from unconstrained MD simulations based on their mutual structural similarity and map each of the representative conformations onto the resultant clusters. To quantify the structural diversity of the representative conformations, we designed an entropy-based measure (35), the distribution entropy ratio ($\theta_{\mathrm{obs}} = I_{\mathrm{obs}}/I_{\mathrm{ran}}$), in which $I_{\mathrm{obs}}$ is the observed distributional entropy of representative conformations over the structure clusters and $I_{\mathrm{ran}}$ is the entropy if the conformations were randomly distributed (see Methods). The closer $\theta_{\mathrm{obs}}$ is to unity, the more uniform the selected states are distributed in conformational space. On the other hand, when $\theta_{\mathrm{obs}}$ is close to zero, the selected conformations are predominantly populated by just one cluster. Interestingly, we observe generally high $\theta_{\mathrm{obs}}$ (mostly between 50 and 88%) for the representative conformations at different levels of cluster hierarchies (Fig. 2), indicating that large structural diversity exists within the determined conformations.

To further evaluate the SAS methodology, both backbone and side-chain order parameters for three proteins, ubiquitin (36,37), eglin c (16), and TNfn3 (27), were employed in MD simulations. Methyl side-chain order parameters describe internal motions of C-C (H3) vectors contained within methyl containing amino-acid side chains and are intimately related to the side-chain rotameric motions (39,40). Consistent with our analysis, the SAS method yields better correlations between calculated and NMR-derived order parameters compared to unconstrained MD for these three data sets. For example, in the case of ubiquitin, the Pearson correlation coefficient between calculated order parameters using unconstrained MD and NMR-derived values are only 0.5–0.8 (Fig. 3 *a*), while the Pearson correlation coefficient between the SAS method and NMR-derived $S^2$ values has a much higher correlation coefficient of 0.91–0.97 (0.86–0.91 for backbone and 0.90–0.97 for side-chain order parameters) (Fig. 3 *b*). Yet, the performance of the SAS method is comparable to that of constraint-based methods (14,22) (Table 2).

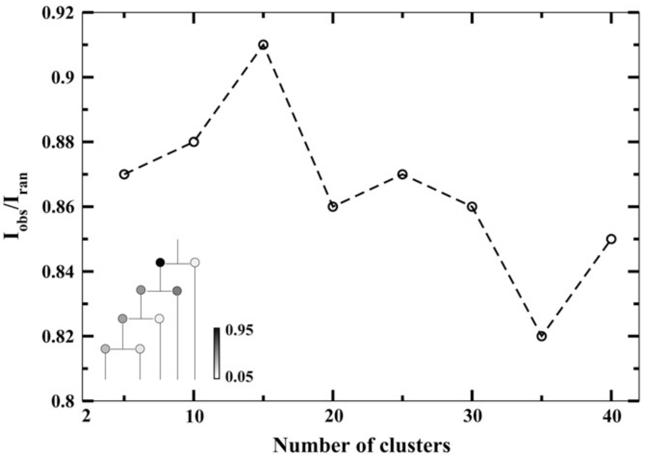

FIGURE 2 The distribution entropy ratio (see Methods), a measure of the structural diversity of the selected conformations in the conformational space is shown as a function of the number of clusters at each hierarchy of the dendogram. As the number of clusters is a monotonic function from the top (one cluster) to the bottom (each cluster contains one conformation) of the dendogram, it is a natural indicator of the hierarchies in the dendogram. The grayscale diagram in the bottom-left corner is used to show the top layers of the dendogram where each node represents the cluster and the grayscale of the node represents the probability that the selected conformations are within that cluster. The result is obtained from one of the three 3-ns MD simulations of TNfn3.

To cross-validate the SAS method, the side-chain scalar three-bond N-C$_\alpha$-C$_\beta$-C$_\gamma$ ($^3J_{\mathrm{N-C}_\gamma}$) or C′-C$_\alpha$-C$_\beta$-C$_\gamma$ ($^3J_{\mathrm{C'-C}_\gamma}$) couplings were back-calculated from the determined conformations of ubiquitin, as described previously (14,22), and compared with experimental values determined independently from order parameters. We observe good agreement (Pearson correlation coefficient of 0.85–0.90) between calculated and experimental scalar-coupling values, providing further validation of the SAS method.

Perhaps not too surprising, the correlation coefficients between calculated and NMR-derived values of both backbone and side-chain order parameters show a noticeable dependence on the effective sampling time (Table 2). For example, with TNfn3, when any one of the three 3-ns MD trajectories is used for determining the conformations, the correlation coefficients between calculated and NMR-derived values increase to 0.80–0.84. However, when all three 3-ns MD trajectories are used, the correlation coefficient increases

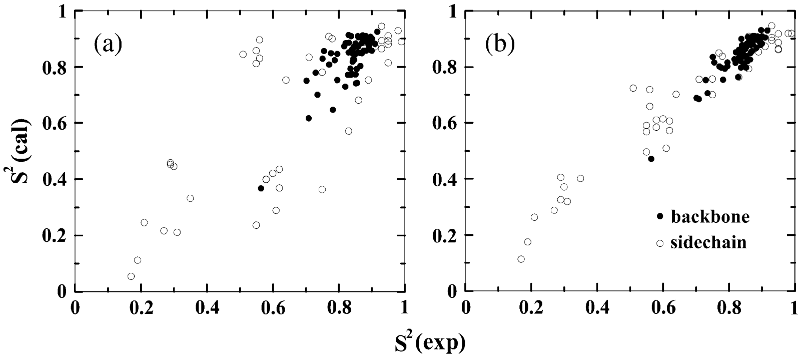

FIGURE 3 Comparisons of calculated and experimentally derived order parameters for ubiquitin are shown (backbone, *solid circle*; side chain, *open circle*). (*a*) Calculation is performed using unconstrained MD ($r = 0.82$) and (*b*) calculation is performed by the SAS method ($r = 0.97$). Both results are obtained from one (same) of the three 3-ns MD trajectories.

**TABLE 2 Comparison of protein conformational consistency determined by unconstrained MD, constraint-based MD (14,22) and the SAS method with experimental backbone and side-chain order parameters**

| | Unconstrained | Constraint-based | SAS | |
|---|---|---|---|---|
| Name | MD | MD | Individual | Combined |
| Eglin c | 0.4–0.7 | N/A | 0.90–0.94 | 0.96 |
| Ubiquitin | 0.5–0.8 | 0.98 | 0.91–0.97 | 0.97 |
| TNfn3 | 0.3–0.6 | >0.9 | 0.80–0.84 | 0.92 |

The numbers shown are the Pearson correlation coefficients between calculated and experimentally derived values. *Individual* indicates that only one of the 3-ns MD trajectories is used in the calculation, whereas *Combined* indicates that all three MD trajectories were used in the calculation.

to 0.92. A similar trend is observed for eglin c (Table 2). These results suggest that insufficient sampling in MD does influence the conformations determined by the SAS method. Therefore, in practice, it is important to generate as structurally diverse an ensemble of conformations as possible for selection, to alleviate the effect of the insufficient sampling. On the other hand, the dependence of the simulation results on the effective sampling time can be beneficial, since a longer effective sampling time only increases the consistency between the determined conformations and experimentally derived dynamics data (Table 2). This feature allows the SAS method to take full advantage of the increasing computational power and more efficient sampling methods, such as MD, that employ implicit solvent (41,42).

The SAS method shows some similarity to methods employed in an earlier study (43), in which the authors determined the unfolded state of the N-terminal SH3 domain of the *Drosophila* signal transduction protein drk (drkN SH3) using NMR data including nuclear Overhauser enhancement restraints, J-coupling constants, and chemical shifts. However, distinct assumptions are made and different computational strategies are used in these two methods. For example, the SAS method assumes that only a subset of the computationally sampled conformations contributes to the experimentally measured observables and the conformations in the selected subset have equal contributions. In contrast, Choy and Forman-Kay (43) assume that all of the sampled conformations have nonnegligible contributions to the experimentally measured properties, and a calculated weight is assigned to each conformation. Predicated on this assumption, they found that only a subset of structures (<60) dominates the contribution to the experimental observables among the 60,000 computationally generated unfolded structures of the drkN SH3 domain, which is consistent with the assumption in SAS method. This result suggests that despite different underlying assumptions and computational strategies, both methods are consistent with each other.

For ubiquitin, we employed the PROCHECK (44) program to evaluate the stereochemical quality of individual conformations selected by the SAS method (Table 3). PROCHECK (44) provides a detailed evaluation of the stereochemistry of a protein structure. The evaluation is based on statistics (45) derived from a set of high-resolution x-ray structures with a resolution of 2.0 Å or better from the Protein DataBank (46). We find that individually selected conformations by the SAS method have comparable stereochemical quality with other structures sampled by MD, indicating that SAS method does not produce an ensemble compatible with the order parameter at the price of selecting conformations with nontypical stereochemical quality in the original ensemble (Table 3). It is also important to note that since the SAS method decouples conformational sampling and selection, it does not improve the stereochemical quality of individual sampled conformations. Instead, it improves the overall representation at the ensemble level to be more consistent with experimental data.

To evaluate the uniqueness of the determined structural ensembles, we first compare the residue-wise backbone RMSD profile (based on C$\alpha$ atoms) of the ensembles determined from different trajectories with respect to the native structure. In the case of ubiquitin, we find that the residue-wise backbone RMSD profiles of different ensembles are quite distinct from each other, indicating that the backbone configuration of the ensemble is not uniquely determined by fitting a given set of order parameters (Fig. 4). We then compare the distributions of the side-chain dihedral angle $\chi_1$ of different ensembles and find that the side-chain configuration is also not uniquely determined (Fig. 5). In addition, we find that the side-chain configurations determined by the SAS method are different from the ones determined using restrained MD in earlier studies (17). For example, in the ubiquitin ensemble determined by restrained MD, there are conformations in which the $\chi_1$ of Ile-13 is positive (22), while

**TABLE 3 A comparison is made between the stereochemical quality of individual conformation selected by the SAS method from one trajectory for ubiquitin when both backbone and side-chain order parameter are used, and all conformations generated in the corresponding trajectory**

| Ensemble/% | Rama-core | Rama-disall | Bond length | Bond angle | Bad contacts |
|---|---|---|---|---|---|
| All structures | 83.2 ± 3.5 | 0 | 82.4 ± 1.9 | 75.4 ± 2.0 | 0 |
| SAS-selected | 82.5 ± 3.0 | 0 | 82.8 ± 1.9 | 75.6 ± 1.9 | 0 |

The comparison is made by using PROCHECK (44) program and five stereochemical parameters are used: Rama-core refers to the percentage of the backbone dihedral angles ($\varphi$, $\psi$) in a given protein structure that are in the most favorable region in the Ramachandran plot; Rama-disall refers to the percentage of the backbone dihedral angles in the forbidden region of Ramachandran plot; bond length and bond angle refer to the percentage of the bond length and angle values that are in the most favorable range; bad contacts refer to the number of the bad contacts, as defined by nonbonded heavy atoms at a distance of $\leq$2.6 Å, in a structure.

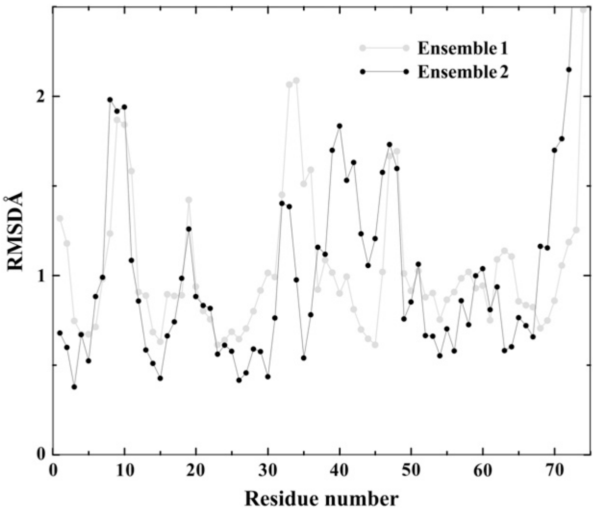

FIGURE 4 For two ensembles determined from different trajectories, the backbone RMSD (based on Cα atoms) with respect to the native structure of ubiquitin are plotted as a function of residue number.

the Ile-13 $\chi_1$ in the conformations determined by the SAS method are negative (Fig. 5 *b*). These results indicate that there are different sets of conformations that fit comparably well with a given set of order parameters. One possible solution to this issue of nonuniqueness of the determined ensemble may be to use other types of constraints besides the order parameter data in the ensemble determination.

Extraction of S2 from the NMR experimental data is often complicated by several factors including anisotropic rotation (including anisotropic internal motion), additional pathways other than dipolar coupling that differentially affect T2 (i.e., conformational transitions, sample heterogeneity) relative to T1 relaxation, variations in sample conditions, unknown random and systematic errors, effective N-H or C-H bond differences, CSA variations, and motional coupling (13). As such, it is important to recognize that the derivation of S2 from experimental data may not always directly relate to the geometric description of the motional amplitude of a bond vector. Therefore, great cautions should be taken to interpret the determined ensemble by either SAS method or other restrained-MD methods.

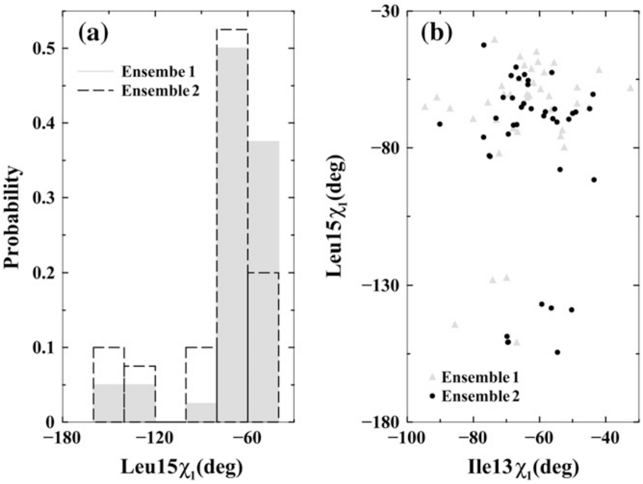

FIGURE 5 The distribution of the side-chain dihedral $\chi_1$ for Leu-15 and Ile-13 in ubiquitin is compared between two ensembles determined from different trajectories. (*a*) Histograms of the side-chain dihedral $\chi_1$ for Leu-15 are plotted; (*b*) scatter plots of the side-chain dihedral $\chi_1$ for Leu-15 and Ile-13 are shown.

## CONCLUSIONS

We present a new computational method, termed SAS, for determining the ensemble of protein conformations consistent with NMR dynamic data. The SAS method offers a flexible framework that allows incorporation of different types of computational sampling methods (e.g., Monte Carlo, coarse-grained MD (47–49)) and NMR dynamics data (e.g., residual dipolar coupling (24,25)) to obtain a better representation of the structural ensemble to be consistent with experimental data.

## SUPPLEMENTARY MATERIAL

To view all of the supplemental files associated with this article, visit www.biophysj.org.

We thank Dr. Feng Ding for helpful discussions. We thank Dr. Andrew Lee for kindly providing order parameter data for eglin c. We thank Dr. Yibing Wu for reading and providing critical feedback on the manuscript. We also thank Daud Cole for reading the manuscript.

This work is supported in part by the Muscular Dystrophy Association grant No. MDA3720 and the American Heart Association grant No. 0665361U and the North Carolina Biotechnology Center grant No. 2006-MRG-1107 to N.V.D.

The authors have declared that no conflicts of interest exist.

## REFERENCES

1. Steinbach, P. J., A. Ansari, J. Berendzen, D. Braunstein, K. Chu, B. R. Cowen, D. Ehrenstein, H. Frauenfelder, J. B. Johnson, and D. C. Lamb. 1991. Ligand binding to heme proteins: connection between dynamics and function. *Biochemistry*. 30:3988–4001.
2. Dixon, R. D., Y. Chen, F. Ding, S. D. Khare, K. C. Prutzman, M. D. Schaller, S. L. Campbell, and N. V. Dokholyan. 2004. New insights into FAK signaling and localization based on detection of a FAT domain folding intermediate. *Structure*. 12:2161–2171.
3. Yan, J., Y. Liu, S. M. Lukasik, N. A. Speck, and J. H. Bushweller. 2004. CBFβ allosterically regulates the Runx1 Runt domain via a dynamic conformational equilibrium. *Nat. Struct. Mol. Biol.* 11:901–906.
4. Chen, Y., and N. V. Dokholyan. 2006. Insights into allosteric control of vinculin function from its large scale conformational dynamics. *J. Biol. Chem.* 281:29148–29154.
5. Bracken, C., P. A. Carr, J. Cavanagh, and A. G. Palmer III. 1999. Temperature dependence of intramolecular dynamics of the basic leucine zipper of GCN4: implications for the entropy of association with DNA. *J. Mol. Biol.* 285:2133–2146.

6. Daniel, R. M., R. V. Dunn, J. L. Finney, and J. C. Smith. 2003. The role of dynamics in enzyme activity. *Annu. Rev. Biophys. Biomol. Struct.* 32:69–92.

7. Kern, D., and E. R. Zuiderweg. 2003. The role of dynamics in allosteric regulation. *Curr. Opin. Struct. Biol.* 13:748–757.

8. Wand, A. J. 2001. Dynamic activation of protein function: a view emerging from NMR spectroscopy. *Nat. Struct. Biol.* 8:926–931.

9. Kay, L. E. 1998. Protein dynamics from NMR. *Nat. Struct. Biol.* 5(Suppl):513–517.

10. Palmer III, A. G. 1993. Dynamic properties of proteins from NMR spectroscopy. *Curr. Opin. Biotechnol.* 4:385–391.

11. Palmer III, A. G. 2001. NMR probes of molecular dynamics: overview and comparison with other techniques. *Annu. Rev. Biophys. Biomol. Struct.* 30:129–155.

12. Karplus, M., and J. A. McCammon. 2002. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* 9:646–652.

13. Case, D. A. 2002. Molecular dynamics and NMR spin relaxation in proteins. *Acc. Chem. Res.* 35:325–331.

14. Best, R. B., and M. Vendruscolo. 2004. Determination of protein structures consistent with NMR order parameters. *J. Am. Chem. Soc.* 126:8090–8091.

15. Lipari, G., and A. Szabo. 1982. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *J. Am. Chem. Soc.* 104:4546–4559.

16. Hu, H., M. W. Clarkson, J. Hermans, and A. L. Lee. 2003. Increased rigidity of eglin c at acidic pH: evidence from NMR spin relaxation and MD simulations. *Biochemistry.* 42:13856–13868.

17. Pfeiffer, S., D. Fushman, and D. Cowburn. 2001. Simulated and NMR-derived backbone dynamics of a protein with significant flexibility: a comparison of spectral densities for the $\beta$ARK1 PH domain. *J. Am. Chem. Soc.* 123:3021–3036.

18. Prabhu, N. V., A. L. Lee, A. J. Wand, and K. A. Sharp. 2003. Dynamics and entropy of a calmodulin-peptide complex studied by NMR and molecular dynamics. *Biochemistry.* 42:562–570.

19. Pang, Y., M. Buck, and E. R. Zuiderweg. 2002. Backbone dynamics of the ribonuclease binase active site area using multinuclear $^{15}$N and $^{13}$C NMR relaxation and computational molecular dynamics. *Biochemistry.* 41:2655–2666.

20. Chatfield, D. C., A. Szabo, and B. R. Brooks. 1998. Molecular dynamics of staphylococcal nuclease: Comparison of simulation with $^{15}$N and $^{13}$C NMR relaxation data. *J. Am. Chem. Soc.* 120:5301–5311.

21. Hu, H., M. Elstner, and J. Hermans. 2003. Comparison of a QM/MM force field and molecular mechanics force fields in simulations of alanine and glycine ''dipeptides'' (Ace-Ala-Nme and Ace-Gly-Nme) in water in relation to the problem of modeling the unfolded peptide backbone in solution. *Proteins.* 50:451–463.

22. Lindorff-Larsen, K., R. B. Best, M. A. Depristo, C. M. Dobson, and M. Vendruscolo. 2005. Simultaneous determination of protein structure and dynamics. *Nature.* 433:128–132.

23. Lindorff-Larsen, K., R. B. Best, and M. Vendruscolo. 2005. Interpreting dynamically averaged scalar couplings in proteins. *J. Biomol. NMR.* 32:273–280.

24. Bax, A. 2003. Weak alignment offers new NMR opportunities to study protein structure and dynamics. *Protein Sci.* 12:1–16.

25. Mittermaier, A., and L. E. Kay. 2006. New tools provide new insights in NMR studies of protein dynamics. *Science.* 312:224–228.

26. Dokholyan, N. V., S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich. 1998. Discrete molecular dynamics studies of the folding of a protein-like model. *Fold. Des.* 3:577–587.

27. Best, R. B., T. J. Rutherford, S. M. Freund, and J. Clarke. 2004. Hydrophobic core fluidity of homologous protein domains: relation of side-chain dynamics to core composition and packing. *Biochemistry.* 43:1145–1155.

28. Buck, M., J. Boyd, C. Redfield, D. A. MacKenzie, D. J. Jeenes, D. B. Archer, and C. M. Dobson. 1995. Structural determinants of protein dynamics: analysis of $^{15}$N NMR relaxation measurements for main-chain and side-chain nuclei of hen egg white lysozyme. *Biochemistry.* 34:4041–4055.

29. Pearlman, D. A., D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham, S. Debolt, D. Ferguson, G. Seibel, and P. Kollman. 1995. AMBER, a package of computer programs for applying molecular mechanics, normal-mode analysis, molecular-dynamics and free-energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* 91:1–41.

30. Ryckaert, J. P., G. Ciccotti, and H. J. C. Berendsen. 1977. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. *J. Comput. Phys.* 23:327–341.

31. Berendsen, H. J. C., J. P. M. Postma, W. F. Vangunsteren, A. Dinola, and J. R. Haak. 1984. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81:3684–3690.

32. Liu, J. S. 2001. Monte Carlo Strategies in Scientific Computing. Springer, New York.

33. Everitt, B. S., S. Landau, and M. Leese. 2001. Cluster Analysis. Oxford University Press, Oxford.

34. Barton, G. J. 2002. OC—A Cluster Analysis Program. University of Dundee, Scotland, UK.

35. Cover, T. M., and J. A. Thomas. 1991. Elements of Information Theory. Wiley, New York.

36. Lee, A. L., P. F. Flynn, and A. J. Wand. 1999. Comparison of $^{2}$H and $^{13}$C NMR relaxation techniques for the study of protein methyl group dynamics in solution. *J. Am. Chem. Soc.* 121:2891–2902.

37. Tjandra, N., S. E. Feller, R. W. Pastor, and A. Bax. 1995. Rotational diffusion anisotropy of human ubiquitin from $^{15}$N NMR relaxation. *J. Am. Chem. Soc.* 117:12562–12566.

38. Hornak, V., R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. 2006. Comparison of multiple AMBER force fields and development of improved protein backbone parameters. *Proteins.* 65:712–725.

39. Best, R. B., J. Clarke, and M. Karplus. 2004. The origin of protein sidechain order parameter distributions. *J. Am. Chem. Soc.* 126:7734–7735.

40. Hu, H., J. Hermans, and A. L. Lee. 2005. Relating side-chain mobility in proteins to rotameric transitions: insights from molecular dynamics simulations and NMR. *J. Biomol. NMR.* 32:151–162.

41. Ding, F., N. V. Dokholyan, and E. Shakhnovich. 2006. Emergence of protein fold families through rational design. *PLoS Comput. Biol.* 2:e85.

42. Feig, M., and C. L. Brooks III. 2004. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struct. Biol.* 14:217–224.

43. Choy, W. Y., and J. D. Forman-Kay. 2001. Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *J. Mol. Biol.* 308:1011–1032.

44. Laskowski, R. A., M. W. Macarthur, D. S. Moss, and J. M. Thornton. 1993. PROCHECK—a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* 26:283–291.

45. Morris, A. L., M. W. MacArthur, E. G. Hutchinson, and J. M. Thornton. 1992. Stereochemical quality of protein structure coordinates. *Proteins.* 12:345–364.

46. Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein DataBank. *Nucleic Acids Res.* 28:235–242.

47. Ding, F., and N. V. Dokholyan. 2005. Simple but predictive protein models. *Trends Biotechnol.* 23:450–455.

48. Ding, F., S. V. Buldyrev, and N. V. Dokholyan. 2005. Folding Trp-cage to NMR resolution native structure using a coarse-grained protein model. *Biophys. J.* 88:147–155.

49. Khare, S. D., F. Ding, and N. V. Dokholyan. 2003. Folding of Cu, Zn superoxide dismutase and familial amyotrophic lateral sclerosis. *J. Mol. Biol.* 334:515–525.