JCTC Journal of Chemical Theory and Computation

ARTICLE

pubs.acs.org/JCTC

# Subtle Monte Carlo Updates in Dense Molecular Systems

Sandro Bottaro,[*,†,∥] Wouter Boomsma,[*,†,‡,∥] Kristoffer E. Johansson,[§] Christian Andreetta,[§] Thomas Hamelryck,[§] and Jesper Ferkinghoff-Borg[*,†]

[†]Department of Electrical Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark

[‡]Department of Astronomy and Theoretical Physics, Lund University, Lund, Sweden

[§]Department of Biology, University of Copenhagen, Copenhagen, Denmark

Supporting Information

**ABSTRACT:** Although Markov chain Monte Carlo (MC) simulation is a potentially powerful approach for exploring conformational space, it has been unable to compete with molecular dynamics (MD) in the analysis of high density structural states, such as the native state of globular proteins. Here, we introduce a kinetic algorithm, CRISP, that greatly enhances the sampling efficiency in all-atom MC simulations of dense systems. The algorithm is based on an exact analytical solution to the classic chain-closure problem, making it possible to express the interdependencies among degrees of freedom in the molecule as correlations in a multivariate Gaussian distribution. We demonstrate that our method reproduces structural variation in proteins with greater efficiency than current state-of-the-art Monte Carlo methods and has real-time simulation performance on par with molecular dynamics simulations. The presented results suggest our method as a valuable tool in the study of molecules in atomic detail, offering a potential alternative to molecular dynamics for probing long time-scale conformational transitions.

## 1. INTRODUCTION

The conformational flexibility of molecules plays a central role in many important biological processes, including signaling, catalysis, regulation, and aggregation.[1−4] Structural and dynamical information of the conformational changes associated with these processes can be partly extracted from spectroscopic techniques, such as X-ray diffraction or nuclear magnetic resonance experiments (NMR).[4] Molecular simulations serve as an ideal complement to these techniques, by allowing the conformational variation to be studied at a detailed atomic level.

Molecular simulations, however, are faced with two main challenges: the design of an accurate energy function[5] and the construction of a sampling strategy capable of efficiently exploring the conformational space.[6] In the all-atom physical potentials usually employed in protein simulations, the energy landscape is rugged and complex due to the presence of a large number of protein−protein and protein−solvent interactions. For these systems, molecular dynamics (MD) is commonly considered the technique of choice.

The alternative approach to molecular simulation, Markov chain Monte Carlo (MC), has the potential to explore the energy landscape more rapidly than MD. In particular, the transitions between consecutive microstates in an MC simulation are not required to follow the dynamics of the system (i.e Newton's law). Using a scheme to accept/reject proposed updates to the chain, it is possible to generate conformations according to the Boltzmann distribution associated with the system. While the MC approach does not provide explicit real-time information, it allows for a rapid exploration of conformations separated by high-energy barriers (i.e., long time-scales) and thereby an efficient thermostatistical characterization of the system. This makes the Monte Carlo method extremely well suited for large scale simulations of, for instance, protein aggregation[7] or for exploring the conformational space of intrinsically disordered proteins,[8] both of which are still mostly intractable using MD. However, for the exploration of dense systems, where even small variations in the degrees of freedom (e.g., dihedral angles) are likely to introduce collisions in the molecule, MC simulations often perform poorly. In an attempt to alleviate this problem, many MC procedures extend their kinetics with so-called *local moves*, which produce subtle deformations in a small segment of the protein chain, while keeping the positions of all atoms outside the segment fixed.

The geometrical issues behind the local move problem were first studied by Go̅ and Scheraga in 1970.[9] On the basis of these considerations, Theodorou and co-workers developed the *concerted rotation* MC-move by working out the necessary requirements for detailed balance (the central condition to ensure Boltzmann distributed sampling).[10] The method works with seven adjacent dihedral angles along the chain. One of these angles is turned by a random amount, and the values of the six remaining angles are determined by numerically solving a set of equations, resulting in a new closed chain structure. Several variants of this original approach have been proposed.[11,12] The most recent is the CRA method,[13] in which increased efficiency was obtained by including bond angle variations[14] and imposing a locality constraint to raise the probability of chain closure, a technique originally introduced in the context of semilocal moves.[15]

Several alternative formulations of the local move problem have been proposed. The *configurational bias* method is based on the idea of regrowing a segment of a chain one atom at a time.[16,17] This approach has been extended with various look-ahead and

**Received:** September 14, 2011
**Published:** December 19, 2011

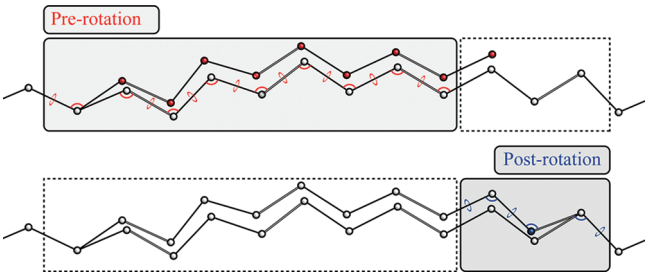


**Figure 1.** Illustration of the concerted rotation method. During prerotation, new values for the angles shown in red (light gray box) are proposed for a small segment of the chain, introducing a break of the chain. The role of the postrotation step (dark gray box) is then to find the necessary compensating changes in the six remaining degrees of freedom, labeled in blue, in order to return to a closed state of the chain.

biasing strategies to decrease the number of rejected growth attempts.[18,19] More recently, robotics-inspired methods have been employed to perform local backbone deformations[20] and to characterize the flexibility of protein loops.[21] As another alternative, an off-lattice version of the *crankshaft* move has been proposed.[22] The method consists of a rigid rotation of a chain segment around the axis defined by the $C_\alpha$ atoms delimiting the segment. An extension of this approach, the *backrub* move,[23,24] has led to successful applications in the context of both protein design and modeling.[25]

The crankshaft/backrub move stands out from the remaining methods for its simplicity and ease of implementation. However, the kinetics produced by this move are limited to hinge-like motions, which can potentially reduce the rate with which the move can decorrelate a structure. The remaining local move methods all introduce a break in chain-connectivity. This forces the methods to treat the placement of a subset of the atoms as a special case in order to maintain a closed chain, thereby introducing an asymmetry in the degrees of freedom involved in the move. Using Boltzmann factors or constrained proposals, it is possible to control the local geometry for the initial, stochastic part of the move. However, the final closure step will typically introduce unfavorable local structure in the chain, leading to an elevated rejection rate.

In the present study, we demonstrate that this problem constitutes one of the primary bottlenecks in current MC simulations of dense protein systems. We present a novel and efficient solution, in which the geometrical constraints are naturally incorporated in a proposal distribution. This leads to a concerted-rotation type Monte Carlo move, *CRISP* (Concerted Rotations Involving Self-consistent Proposals), which effectively proposes closed structures with user-controlled variations of all involved degrees of freedom. We demonstrate the correctness of the method and assess its efficiency by estimating the correlation time associated with the move. The results demonstrate that CRISP significantly outperforms the current state of the art MC methodologies. We proceed with a study of the native ensemble of ubiquitin. A comparison to X-ray and NMR experimental data shows that our improved sampling strategy enables us to cover the entire known conformational fluctuation spectrum of ubiquitin in solution, including several experimentally confirmed conformational switches. In addition to the clear performance improvement over existing MC methods, we demonstrate that our method has comparable real-time performance to MD on this system.

## 2. RESULTS

**2.1. Method Overview.** For simplicity, we will present the method in the context of protein molecules, although the basic principles apply to other chain molecules (Text S1 and S2, Supporting Information). A typical parametrization used in protein simulations is one with flexible dihedral angles and bond angles, but with bond lengths fixed. Given this parametrization, the local move problem can be phrased as follows: propose new values for all dihedral and bond angles in a region of a protein chain so that any atom position outside the region remains untouched. The requirement of chain integrity imposes a strong dependency among the degrees of freedom. From this perspective, the local move problem is essentially a matter of finding the cross-correlation between the degrees of freedom that fulfills the geometrical constraints given by the protein representation. In this paper, we will demonstrate how a probability distribution can be constructed that takes these dependencies into account. A natural framework for these considerations is that of the *concerted rotation*. [Note that it could also be formulated as a *configurational bias* move with the bias given by the derived probability distribution.] In the concerted rotation approach, a move is divided into a stochastic *prerotation* step followed by a deterministic *postrotation* step. During prerotation, new angles are proposed for a small segment of the chain, introducing a break of the chain. The postrotation step then closes the chain by finding the necessary compensating changes in the six postrotational degrees of freedom (Figure 1).

The derivation of our desired probability distribution is based on two observations. First, we note that given the described molecular chain representation, an exact, analytical solution for the postrotation problem can be derived (Text S1). This means that for any given value of the prerotated degrees of freedom, the resulting postrotation values can be determined with high efficiency and robustness. This solution represents a great advantage over other concerted-rotation methods by avoiding the tedious numerical resolution of a system of six equations in six unknowns.

The second step is the realization that the analytical solution allows us to express the coupling between pre- and postrotation as a linear transformation, which enables the construction of a probability distribution that controls both pre- and postrotational degrees of freedom as well as the necessary chain-closure constraints. To our knowledge, this is a novel mathematical description of the 40-year-old chain closure problem. Unlike previous approaches, it makes it possible, to first order, to directly sample closed chains. Since the complete derivation is quite involved, we only highlight the main features here and refer the reader to the Supporting Information for details (Text S2−S4).

To illustrate the nature of the procedure, we consider a local move where $n$ degrees of freedom $\bar{\chi} = (\chi_1 ... \chi_n)$ of the chain backbone are modified, leading to a new conformation $\bar{\chi}'$. Angular variations $\delta\bar{\chi} = \bar{\chi}' - \bar{\chi}$ are drawn from a multivariate Gaussian distribution

$$p(\delta\bar{\chi}) \propto \exp\left( -\frac{1}{2} \delta\bar{\chi}^T \lambda \mathbf{C}_n \delta\bar{\chi} \right) \quad (1)$$

where the scalar parameter $\lambda$ specifies the degree of locality, with increasing $\lambda$ leading to smaller changes. $\mathbf{C}_n$ is an $n$-dimensional diagonal matrix introduced with the purpose of scaling, by a factor $k$, the allowed variations of bond and $\omega$ dihedral angles

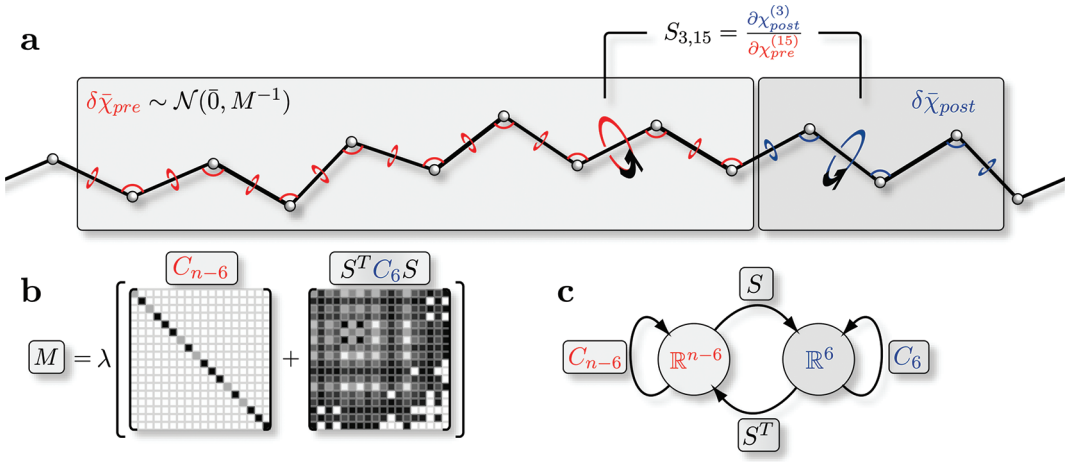


**Figure 2.** Graphical representation of the proposal probability distribution used in CRISP moves. (a) Angular variations are drawn from a normal distribution with mean $\bar{0}$ and covariance $\mathbf{M}^{-1}$. The **S** matrix couples the prerotational degrees of freedom (red) to the postrotational ones (blue). In the example shown in the figure, the matrix element $S_{3,15}$ reports the variation of the third postrotational angle upon a change in the prerotational degree of freedom at position 15. (b) **M** is a sum of a diagonal matrix $\mathbf{C}_{n-6}$ that controls the variations of the prerotational degrees of freedom, and $\mathbf{S}^T\mathbf{C}_6\mathbf{S}$. This last, nondiagonal matrix operates on $\delta\bar{\chi}_{\mathrm{pre}}$ as shown in c: the **S** matrix first reports the changes $\delta\bar{\chi}_{\mathrm{post}}$ upon the variation $\delta\bar{\chi}_{\mathrm{pre}}$. The variations of postrotational angles, shown in blue, are then properly constrained via $\mathbf{C}_6$, and $\mathbf{S}^T$ maps back the postrotational changes to the prerotational, $n-6$-dimensional space. $\lambda$ is a free parameter controlling the overall size of the move.

relative to $\varphi$,$\psi$ angles:

$$C_{ii} = \begin{cases} k & \text{if } i \text{ is a bond or } \omega \text{ dihedral angle} \\ 1 & \text{if } i \text{ is a } \varphi \text{ or } \psi \text{ dihedral angle} \end{cases} \quad (2)$$

Due to the deterministic nature of the chain closure problem, the values of the six postrotational degrees of freedom $\bar{\chi}_{\mathrm{post}} = (\chi^{(1)}_{\mathrm{post}} \ldots \chi^{(6)}_{\mathrm{post}})$ are determined by the remaining $n-6$ prerotational angles via our analytical solution. To first order, this allows us to express the variation of the six postrotational angles as a function of the prerotational variation, $\delta\bar{\chi}_{\mathrm{post}} = \mathbf{S}\delta\bar{\chi}_{\mathrm{pre}}$, where **S** is a $6 \times (n-6)$ matrix (Text S2). This information can be directly embedded in the proposal distribution of eq 1. As demonstrated in Text S3, the proposal distribution can now be written as

$$p(\delta\bar{\chi}_{\mathrm{pre}}) \propto \exp\left(-\frac{1}{2}\delta\bar{\chi}^T_{\mathrm{pre}}\lambda(\mathbf{C}_{n-6} + \mathbf{S}^T\mathbf{C}_6\mathbf{S})\delta\bar{\chi}_{\mathrm{pre}}\right)$$
$$= \exp\left(-\frac{1}{2}\delta\bar{\chi}^T_{\mathrm{pre}}\mathbf{M}\delta\bar{\chi}_{\mathrm{pre}}\right) \quad (3)$$

Figure 2 illustrates the construction of the proposal function in eq 3. Angular variations for the prerotational degrees of freedom are drawn from a Gaussian distribution with mean $\bar{0}$ and covariance $\mathbf{M}^{-1}$ (Figure 2a). **M** is a sum of two terms (Figure 2b): the scaling diagonal matrix $\mathbf{C}_{n-6}$, acting on the prerotational angles, and $\mathbf{S}^T\mathbf{C}_6\mathbf{S}$. The latter, nondiagonal matrix carries the correlations between pre- and postrotational angles arising from the fixed bond-lengths constraint and from the restrictions given by the stereochemistry of the protein backbone. In other words, it operates on $\delta\bar{\chi}_{\mathrm{pre}}$ as depicted in Figure 2c: The **S** matrix first reports the compensating changes $\delta\bar{\chi}_{\mathrm{post}}$ upon the variation $\delta\bar{\chi}_{\mathrm{pre}}$. The postrotational variations are then properly constrained via $\mathbf{C}_6$, and $\mathbf{S}^T$ finally maps the changes back to the $n-6$ dimensional space of the prerotation.

A proof of correctness of the first order approximation and a study of the range of its effectiveness is presented in Figures S1 and S2. We check the validity of the MC procedure, outlined in Text S4, by demonstrating detailed balance (Figure S3). The approach is further validated by demonstrating that, for long simulations on a small system, MC and MD methods produce comparable ensembles (Figure S4 and Table S1).

We proceed by establishing the performance of CRISP relative to two successful local move methods from the literature: the CRA concerted rotation method (CRA)[13] and a detailed balance version of the crankshaft-based *backrub* method (CRANKSHAFT)[24] (see Materials and Methods).

**2.2. Controlled Variations.** The main motivation for introducing local kinetics into a simulation is to increase sampling efficiency in dense systems, since nonlocal moves tend to propose a high number of self-colliding structures. However, while local moves successfully reduce the rate of self-collision, they are faced with a different problem: due to the strong chain-closure constraints, local moves will often introduce unfavorable values to a subset of the involved degrees of freedom, leading to an increase in the rate of rejection. We illustrate this problem using the CRA concerted rotation move. The method is constructed around the idea of limiting the movement of the end point of the prerotation (the breakpoint), in order to increase the probability of finding a solution for the postrotation. It is evident from Figure 3 that this strategy creates an imbalance in the move: constraining the displacement of the breakpoint is not sufficient to avoid significant fluctuations of the postrotational angles. In this case, the effect does not represent a significant problem for dihedral angles, but a typical change of 5° for *all* postrotational bond angles is dramatic, considering that the experimentally observed distribution width is ~2.7° for such degrees of freedom (Figure S5).[26] Other local move methods suffer from similar problems: in crankshaft-type moves, the bond angles surrounding the pivotal points will be subject to large fluctuations, while concerted rotation methods that do not include bond angles typically involve large jumps in dihedral angle values.[10,11] In contrast to existing methods, Figure 3 demonstrates that CRISP displays identical variations in pre- and postrotational degrees of freedom, *de facto* eliminating the asymmetry introduced by the

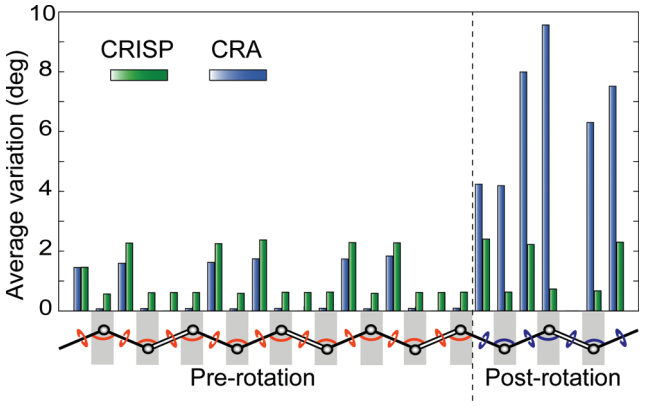


**Figure 3.** Average angular variations based on $5 \times 10^4$ attempted CRISP and CRA updates on ubiquitin. Each bar corresponds to the average variation of the degree of freedom shown in the chain below the histogram. In the prerotation (red angles), similar average angular variations are proposed by both methods. During postrotation (blue angles), large angular variations are introduced by CRA, due to the lack of a strategy controlling these degrees of freedom. Conversely, no imbalance between pre- and postrotation is observed when using CRISP.

chain-closure constraint. Note that the difference between bond angle and dihedral variations is user-defined (see eq 2).

**2.3. Simulation Efficiency.** The optimal way to quantify sampling efficiency is by measuring the correlation time associated with a given kinetic algorithm,[27] which represents the number of MC steps separating two independent samples. The correlation time allows us to compare the efficiency of the different methods and to establish the optimal values of the two free parameters of CRISP. Since obtaining converged estimates of the correlation time requires extensive simulations, we consider the equilibrium fluctuations around the stable helical state of the small peptide $Ala_{14}$.[13]

In Table 1, we report the correlation times in MC steps for the energy, $\tau_e$, and the average correlation time, $\bar{\tau}_d$, of the 20 central dihedral angles when using CRISP, CRA, and CRANKSHAFT moves as described in the Materials and Methods. There is a difference in the dimension of configurational space explored by the three methods which should be taken into account when comparing the correlation times. Given the correlation time, the size of explored space can be estimated by calculating the standard deviation $\sigma$ of the distribution for the energy and for dihedral angles (Table 1). CRISP shows a dramatic improvement of a factor of 15−20 in sampling efficiency compared to CRA as a consequence of the more appropriate treatment of the geometrical problem. Furthermore, the CRANKSHAFT move explores a conformational and energetic space which is ∼30% smaller than the one covered by CRISP for each degree of freedom, at a computational cost that is 2−3 times larger.

The two free parameters of CRISP, $k$ and $\lambda$, were optimized with respect to the correlation time (Figure S6). In our experience, this setting is not sensitive to the type of protein being simulated, and all simulations in our study therefore use these values.

**2.4. The Native Ensemble of Ubiquitin.** While the $Ala_{14}$ system is useful for the calculation of correlation times, it is not representative of the structural heterogeneity observed in native globular proteins. We therefore extend our analysis with a study of ubiquitin. Ubiquitin is a key to several cellular signaling networks[28,29] and is recognized by a broad variety of proteins

**Table 1. Correlation Time $\tau$ in MC Steps and Standard Deviation $\sigma$ of the Distribution for CRA, CRANKSHAFT, and CRISP Moves Calculated over 5 Independent Runs**

| | $\bar{\tau}_d$ ($10^3$ steps) | $\bar{\sigma}_d$ (deg) | $\tau_e$ ($10^3$ steps) | $\sigma_e$ (kcal/mol) |
|---|---|---|---|---|
| CRA | 13.2 ± 1.1 | 9.31 | 18.2 ± 3.3 | 3.04 |
| CRANKSHAFT | 1.9 ± 0.4 | 7.68 | 2.9 ± 0.9 | 2.42 |
| CRISP | 0.78 ± 0.01 | 10.67 | 0.85 ± 0.05 | 3.39 |

with high specificity. Furthermore, this protein is well characterized by NMR[30−33] and has been used extensively as a model system in previous computational approaches.[34−37] We use this as a model system for a comparison of CRISP to existing MC sampling algorithms, to MD, and to stochastic dynamics[38] (SD) simulations.

Structural fluctuations around the native state can be expressed as root mean squared fluctuations (RMSF), which measure the amplitude of movements of individual atoms around their equilibrium positions. The RMSF values generally grow with the simulation time, converging when the neighborhood of the local energy minima is exhaustively explored. We captured the global time evolution of this process by considering the sum of the individual RMSF values for all $C_\alpha$ atoms in the chain (cumulative RMSF). Although not as rigorous as the correlation time estimation, this procedure is useful to evaluate and compare the efficiency of different sampling techniques in the vicinity of the native state. In Figure 4, we show the simulation-time evolution of the cumulative $C_\alpha$ RMSF for the different methods. For each method, we report the average cumulative RMSF over 10 simulations performed at $T$ = 300 K, starting from a relaxed state of the human ubiquitin X-ray structure (1UBQ). The MD/SD simulations covered 10 ns using the exact same force field and conditions (see Materials and Methods). For visualization purposes, the $x$ axis for MD/SD is scaled to match the CPU time of the MC methods on the same machine.

Since detailed balance is fulfilled for all simulations (Figures S2 and S4), we expect the fluctuations of the different methods to eventually converge to the same levels. The observed differences thus reflect the degree of ergodicity obtained by the various sampling methods within the given simulation time. For CRA and CRANKSHAFT, the cumulative RMSF saturates at around 40 Å with a similar convergence time (Figure 4). The SD simulations are considerably faster but saturate approximately at the same level. Our CRISP method clearly outperforms the competing MC methodologies: the cumulative RMSF quickly crosses the 40 Å barrier, saturating at the same level as the MD simulations (∼50 Å). To further investigate the nature of these fluctuations, Figure 5 shows the converged RMSF profile per $C_\alpha$ atom for the different simulation methodologies. The fluctuations produced by CRISP are remarkably similar to MD, while the RMSF profiles of SD, CRA, and CRANKSHAFT are consistently lower.

As an experimental reference, we present the RMSF of two NMR-derived ensembles: MUMO (PDB code 2NR2[35]) and EROS (PDB code 2K39[33]), selected to represent the variation of experimentally based ensembles reported in the literature (Figure 6b). The fluctuations obtained with CRISP and MD are in good agreement with the experimental data. Specifically, the large variability observed in the $\beta_1-\beta_2$ loop, the C-terminal region of $\alpha_1$, and the $\beta_3-\beta_4$ loop cover the main conformational variability observed in X-ray ubiquitin complexes, as represented

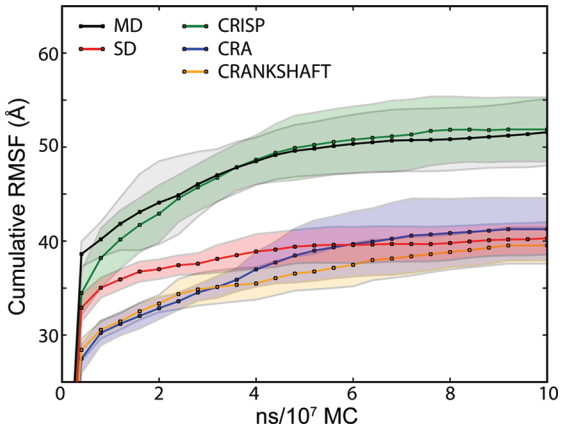


**Figure 4.** Time evolution of the cumulative $C_\alpha$ RMSF relative to MD simulations (in black/gray) compared with SD and with MC simulations using CRISP, CRA, and CRANKSHAFT as local moves around the native state of protein ubiquitin. The shaded regions show the standard deviation based on 10 independent runs.

by the zinc finger ubiquitin-binding domain of isopeptidase T (2G45) and the conjugating enzyme (E2) binding domain (1AAR)[33] (Figure 6c–d).

Large fluctuations in the MD and CRISP simulations are also observed in the $\alpha_1$ N-cap region around GLU24 and in the $\beta_4$–$\alpha_2$ loop around GLY53. It is worth noticing that residual fluctuations in these regions are directly linked to a more subtle conformational switch consisting of the flipping of the ASP52/GLY53 amide plane (Figure 6e). This movement exposes the backbone CO of ASP52 to the exterior of the molecule, while the backbone NH of GLY53 forms an internal hydrogen bond with the side chain of GLU24, which changes rotamer state due to this interaction. This switch has recently been observed in a crystal structure of monomeric human ubiquitin,[39] and the flexibility of these residues was hypothesized to play a role when ubiquitin binds with deubiquinating enzymes. Notably, this backbone transition is not observed in the 10 ns SD or in the CRA/CRANKSHAFT simulations.

We stress that any such *in silico* observation is in principle a consequence of the applied force field, not the sampling method used. However, the fact that these transitions are not seen by all of the simulation methods within the same time frame using the same force field points to the importance of an efficient sampling strategy.

We analyze the different ensembles in detail by comparing the probability distributions of both dihedral backbone angles and $C_\alpha$ positions produced by the different methodologies. In Table 2, we report the Jensen-Shannon divergence (JSD) of the $\varphi/\psi$ distributions between a reference ensemble $\overline{\text{MD}}$ and the individual trajectories (Figure S7). The $\overline{\text{MD}}$ ensemble is constructed using the samples from all conducted MD simulations. MD serves as a meaningful reference, as it represents the broadest and most complete representation of the near-native dynamics among the existing methods. The results show that CRISP and SD simulations reproduce the dihedral backbone distributions of $\overline{\text{MD}}$ with greater accuracy compared to CRA and CRANKSHAFT. The same scenario is observed when considering the Kullback–Leibler divergence (KLD) between $C_\alpha$ position distributions (Table 2), a recently proposed alternative measure of ensemble similarity.[40]

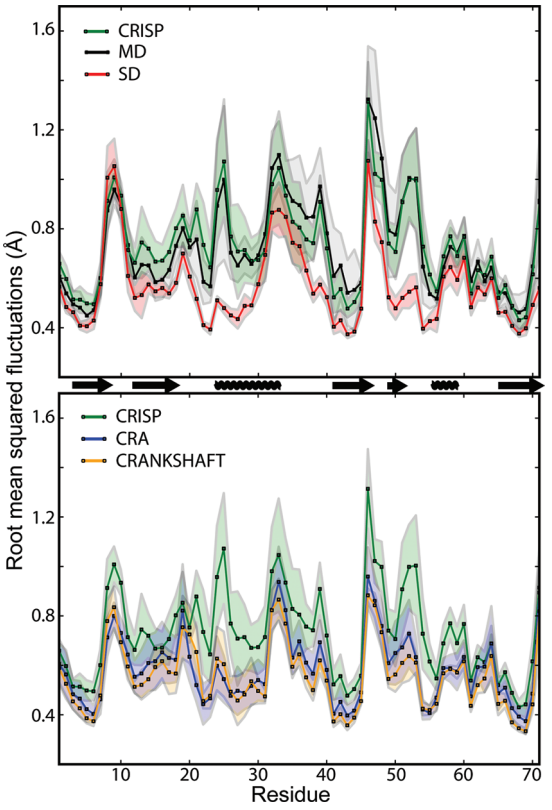


**Figure 5.** Ubiquitin RMSF from 10 ns MD simulations (in black/gray) compared to SD and to $9 \times 10^7$ long MC runs using CRISP, CRA, and CRANKSHAFT. The shaded regions show the standard deviation based on 10 independent runs.

## 3. DISCUSSION

The efficiency of Markov chain Monte Carlo protein simulations relies heavily on the kinetic algorithm used to probe the various possible conformational states of the molecule. In particular, densely packed systems typically require the presence of a move which restricts itself to modifying atom positions within a short stretch of the molecule. Designing such moves is a nontrivial task, due to the complex interdependencies among bond and dihedral angles that arise from constraining the end point of the modified stretch to be fixed. Our present study introduces a novel technique for incorporating these interdependencies directly into a proposal distribution. The resulting local move, CRISP, displays significant performance improvements compared to existing state-of-the-art MC methods.

It should be noted that efficient side-chain dynamics is an important prerequisite for obtaining the presented MC results. In particular, in the often dense hydrogen bond networks characterizing native proteins, one should ensure that an MC simulation includes side-chain moves that can break and form these bonds independently. We discuss this issue in greater detail in the Materials and Methods and in Figure S8. In addition, when conducting a local backbone move, the corresponding displacement of rigidly attached side-chain atoms can lead to self-collisions in the chain, and thus an elevated rejection rate. This consideration is of great importance especially for long side-chains at high densities. The presented move could therefore potentially be improved by including constraints from side-chain

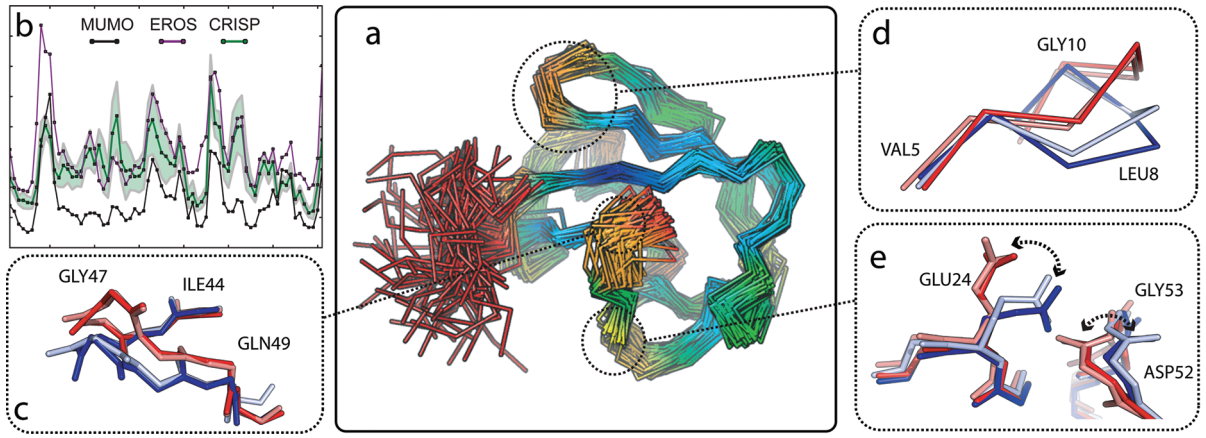


**Figure 6.** Structural ensemble of ubiquitin obtained with CRISP. (a) Backbone trace of 90 random samples from 10 9 × $10^7$ MC-iteration-long simulations using CRISP moves. The residue's color varies from blue (RMSF = 0.5 Å) to green and yellow to red (RMSF > 1.3 Å). (b) RMSF profile from CRISP simulations and from the NMR ensembles EROS and MUMO. (c) The VAL5−LYS11 stretch of the crystal structure 1AAR (dark red) and 2G45 (dark blue). The closest MC samples to the crystal structures are shown in light red and light blue. (d) The ILE44−LEU50 loop of 1AAR (dark red) and 2G45 (dark blue). (e) The ASP52/GLY53 conformational switch. With respect to the crystal structure of 1UBQ (in dark red), this amide-plane flipped state is coupled to the side chain movement of GLU24 and is recurrently found in crystal structures of complexes with deubiquinating enzymes, including 2G45 (in dark blue). Both flipped (light blue) and unflipped (light red) states are explored in our MC sampling.

**Table 2. Comparison between the $\overline{\text{MD}}$ Ensemble and MD, SD, CRISP, CRANKSHAFT, and CRA Simulations[a]**

| | JSD ($\varphi/\psi$) | KLD ($C_\alpha$) |
|---|---|---|
| MD | 0.07 ± 0.04 | 174 ± 148 |
| CRISP | 0.09 ± 0.05 | 300 ± 110 |
| SD | 0.10 ± 0.06 | 530 ± 57 |
| CRA | 0.13 ± 0.07 | 1084 ± 115 |
| CRANKSHAFT | 0.17 ± 0.09 | 2265 ± 719 |

[a] Averages and standard deviations of the JS divergence in $\varphi/\psi$ space and of the KL divergence of the $C_\alpha$ positions are calculated over 10 runs.

interactions into the derived probability distribution, thus merging backbone and side-chain dynamics into a single move.

One of the goals of our study was to investigate the relative sampling efficiency of Monte Carlo versus molecular dynamics. As has been observed before,[41] our results demonstrate that, in dense environments, molecular dynamics provides a more efficient sampling algorithm than previously described state-of-the-art Monte Carlo methods, presumably due to the high information content provided by the gradient in this scenario. However, this does not seem an inherent limitation of the Monte Carlo method. Our current study indicates, as previously demonstrated for smaller systems,[42,43] that MC can provide the same level of accuracy and efficiency as MD, given that an efficient sampling strategy is used. This result therefore suggests that MC simulations can be reliably employed not only in less dense scenarios (such as intrinsically disordered proteins,[8] protein aggregation,[7] or flexible protein loops[21]) but also for a general *in silico* characterization of flexibility and dynamics of compact molecular systems.

## 4. MATERIALS AND METHODS

**4.1. Molecular Representation.** We used a full-atom representation of proteins with fixed bond lengths and flexible dihedral and bond angles. Although flexibility in bond angles is sometimes omitted, it has been shown to increase sampling efficiency.[13,14,44] The peptide dihedral angle $\omega$ is included as it is known experimentally to vary at least at the level of bond angles.

**4.2. Simulation Setup.** The molecular dynamics and stochastic dynamics simulations were conducted using the molecular-modeling package TINKER 5.1.[45] As described in previous studies,[46] stochastic dynamics at constant temperature $T$ = 300 K models the viscous drag of water (frictional coefficient 91 $ps^{-1}$). Constant temperature $T$ = 300 K molecular dynamics simulations were run using the Beeman integration method. Bond lengths were constrained with the RATTLE algorithm, allowing time steps of 2 fs. MC simulations were conducted using the standard Metropolis-Hastings Monte Carlo scheme at physiological temperature ($T$ = 300 K). Note that several techniques (such as replica exchange[48] or multicanonical ensembles[49]) can be used to enhance the sampling relative to standard MD or Metropolis-Hastings MC simulations. In the present study, for comparative purposes and simplicity, we limited ourselves to direct sampling from the canonical ensemble.

**4.3. Force Field.** All simulations in this study were conducted using the $OPLS_{aa}$[50] potential in combination with the generalized Born/surface area implicit solvent model GB/SA.[51] Despite the limitations of implicit solvent models, this combination has been widely used and successfully applied to identify the native state of a large set of proteins[52] and for folding simulations.[46,43,47] Note that the CRISP method is not necessarily limited to implicit solvent simulations. Several examples exist of fruitful combinations of Monte Carlo sampling using explicit solvents models.[53,54]

The MC implementation of the $OPLS_{aa}$+GB/SA force field followed that of the Tinker software package and was verified to reproduce the same energy values as this package. The MD and MC results reported in the paper are therefore directly comparable, both in terms of energetics and computational time. We acknowledge that both methods could be optimized further using for instance hardware specific implementations.

**4.4. Correlation Times.** The MC simulations on $Ala_{14}$ were conducted using the standard Metropolis-Hastings Monte Carlo scheme at physiological temperature ($T$ = 300 K), using only local internal moves (i.e., the N- and C-terminal were kept fixed

during simulation). For CRISP and CRA, the move length was set to five residues. The free parameter $\lambda$, which determines the overall size of the CRISP and CRA moves, was tuned for optimality in the context of the correlation times of $Ala_{14}$ (Figure S6). In the CRANKSHAFT move, the number of residues involved in each update was randomly chosen in the range 2−12, as reported in the original description of backrub.[24] Fixed-length CRANKSHAFT moves of length 5 were also attempted but were found to lead to dramatically inferior performance.

**4.5. Ubiquitin.** All MC simulations on ubiquitin were conducted using the standard Metropolis-Hastings Monte Carlo scheme at $T$ = 300 K. The move-set was composed as follows: 20% local moves, 75% single side-chain moves, and 5% pivot moves. Two different types of single side-chain moves were used: with weight 2/3, samples were drawn from the Dunbrack backbone independent rotamer library[55] (compensating for the bias introduced), while the remaining 1/3 consisted of local side-chain moves (see below). For the pivot moves, new values for the $\varphi$ and $\psi$ values of a single, randomly chosen residue were drawn from a Gaussian distribution with zero mean and $\sigma = 1°$.

**4.6. Side-Chain Sampling.** In order to obtain an efficient side-chain sampling, we included a semilocal side-chain move in our move set. Inspired by the biased Gaussian step,[15] this move consists of updating the $\chi$ side-chain angles with a constraint toward small displacement of atoms involved as acceptors or donors in hydrogen bonds (Figure S8). This type of move was necessary in order to enable small adjustment of the side-chains without breaking the dense network of noncovalent interactions and was found to greatly facilitate both backbone and side-chain transitions.

It is important to note that all MC methods in our comparison share the same set of Monte Carlo moves for the side chains. It is thus the combination of improved backbone dynamics and efficient side-chain dynamics that gives rise to the increased fluctuations observed with CRISP.

**4.7. RMSF Calculations.** For each MC/MD simulation, samples were dumped every $2 \times 10^4$ MC steps/4 ps and superimposed on the crystal structure 1UBQ, excluding the highly fluctuating terminal residues 71−76. The $C_\alpha$ RMSF of each ensemble was calculated as the root mean squared deviation from the mean position.

**4.8. CRANKSHAFT.** The CRANKSHAFT move[24] includes an optimized placement of $C_\beta$ and $H_\alpha$ atoms, which does not fulfill detailed balance and is therefore omitted in our implementation.

**4.9. Jensen-Shannon and Kullback−Leibler Divergence.** The average Jensen-Shannon divergence $\langle \mathrm{JSD}(\overline{\mathrm{MD}}||X)\rangle$ between the reference ensemble $\overline{\mathrm{MD}}$ and the ensemble produced by the method $X$ was calculated as

$$\langle \mathrm{JSD}(\overline{\mathrm{MD}}||X)\rangle = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{N}\sum_{j=1}^{N}\mathrm{JSD}(p^j_{\overline{\mathrm{MD}}}(\varphi,\psi)||p^j_{Xi}(\varphi,\psi)) \quad (4)$$

where the index $i$ runs over the 10 simulations, $j$ runs over the residues, and $p$ is the $\varphi,\psi$ probability distribution estimated using a binning procedure.

The Kullback−Leibler ensemble distance measure was calculated as described in the original reference.[40] In short, assuming the ensembles to be modeled as multivariate normal distributions, it is possible to find a closed-form expression for the KL divergence, which has a direct interpretation in terms of similarity between ensembles. This measure was averaged over 10 runs. All the samples are aligned to the crystal structure 1UBQ prior to the analysis.

**4.10. Availability.** The CRISP method is implemented as part of the Phaistos software package, freely available under the GNU General Public License v3.0 at sourceforge.net/projects/phaistos.

## ■ ASSOCIATED CONTENT

**Supporting Information.** Text S1: Analytical solution for chain closure. Text S2: First order approximation expressing the six postrotational degrees of freedom as a function of the prerotational ones. Text S3: Full expression for the matrix **M** in eq 3. Figure S1: Validation of the first order approximation presented in Text S2. Figure S2: Range of validity of the first order approximation presented in Text S2. Text S4: Outline of the Monte Carlo algorithm. Figure S3: Demonstration of detailed balance for CRISP moves. Figure S4: Angular distribution obtained from Monte Carlo runs using CRISP and from molecular dynamics simulations on alanine 5. Table S1: Comparison between average collective variables calculated using SD and different MC methodologies. Figure S5: Probability distribution and cumulative probability distribution of energy jumps for different local Monte Carlo moves. Figure S6: Dependence of the correlation time over the choice of the parameter $\lambda$ of eq 3. Figure S7: Jenson-Shannon divergence for the $\varphi/\psi$ distribution of ubiquitin between MD and SD, CRISP, CRA, and CRANKSHAFT moves. Figure S8: Snapshots from a Monte Carlo simulation on ubiquitin, illustrating a locked side-chain conformation. This material is available free of charge via the Internet at http://pubs.acs.org/.

## ■ AUTHOR INFORMATION

### Corresponding Author

*E-mail: sbo@elektro.dtu.dk; wouter.boomsma@thep.lu.se; jfb@elektro.dtu.dk.

### Author Contributions

‖W.B. and S.B. contributed equally to this work

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENT

We thank our colleagues Jes Frellsen, Tim Harder, Kasper Stovgaard, and Mikael Borg for useful discussions and Jacobus J. Boomsma and Kresten Lindorff-Larsen for valuable comments and suggestions. S.B is supported by Radiometer. W.B and K.E.J are funded by the Danish Council for Independent Research (FNU, 272-08-0315 and FTP, 274-08-0124, respectively). C.A. and T.H. acknowledge the Danish Program Commission on Nanoscience (NaBiIT, 2106-06-0009).

## ■ REFERENCES

(1) Eisenmesser, E.; Millet, O.; Labeikovsky, W.; Korzhnev, D.; Wolf-Watz, M.; Bosco, D.; Skalicky, J.; Kay, L.; Kern, D. *Nature* **2005**, *438*, 117–121.

(2) Chiti, F.; Dobson, C. *Nat. Chem. Biol.* **2008**, *5*, 15–22.

(3) Nevo, R.; Stroh, C.; Kienberger, F.; Kaftan, D.; Brumfeld, V.; Elbaum, M.; Reich, Z.; Hinterdorfer, P. *Nat. Struct. Mol. Biol.* **2003**, *10*, 553–557.

(4) Boehr, D.; Nussinov, R.; Wright, P. *Nat. Chem. Biol.* **2009**, *5*, 789–796.
(5) Ponder, J.; Case, D. *Adv. Protein Chem.* **2003**, *66*, 27–85.
(6) Liwo, A.; Czaplewski, C.; Oldziej, S.; Scheraga, H. A. *Curr. Opin. Struct. Biol.* **2008**, *18*, 134–139.
(7) Li, D.; Mohanty, S.; Irbäck, A.; Huo, S. *PLoS Comput. Biol.* **2008**, *4*, e1000238.
(8) Mao, A.; Crick, S.; Vitalis, A.; Chicoine, C.; Pappu, R. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 8183.
(9) Gō, N.; Scheraga, H. *Macromolecules* **1970**, *3*, 178–187.
(10) Dodd, L.; Boone, T.; Theodorou, D. *Mol. Phys.* **1993**, *78*, 961–996.
(11) Hoffmann, D.; Knapp, E. *Eur. Biophys. J.* **1996**, *24*, 387–403.
(12) Dinner, A. *J. Comput. Chem.* **2000**, *21*, 1132–1144.
(13) Ulmschneider, J.; Jorgensen, W. *J. Chem. Phys.* **2003**, *118*, 4261–4271.
(14) Bruccoleri, R.; Karplus, M. *Macromolecules* **1985**, *18*, 2767–2773.
(15) Favrin, G.; Irbäck, A.; Sjunnesson, F. *J. Chem. Phys.* **2001**, *114*, 8154–8158.
(16) Frenkel, D.; Mooij, G.; Smit, B. *J. Phys.: Condens. Matter* **1992**, *4*, 3053–3076.
(17) Escobedo, F. J.; de Pablo, J. J. *J. Chem. Phys.* **1995**, *102*, 2636–2652.
(18) Vendruscolo, M. *J. Chem. Phys.* **1997**, *106*, 2970–2976.
(19) Chen, Z.; Escobedo, F. *J. Chem. Phys.* **2000**, *113*, 11382–11392.
(20) Coutsias, E.; Seok, C.; Jacobson, M.; Dill, K. *J. Comput. Chem.* **2004**, *25*, 510–528.
(21) Nilmeier, J.; Hua, L.; Coutsias, E.; Jacobson, M. *J. Chem. Theory Comput.* **2011**, *7*, 1564–1574.
(22) Betancourt, M. *J. Chem. Phys.* **2005**, *123*, 174905–174905−07.
(23) Davis, I.; Arendall, W., III; Richardson, D.; Richardson, J. *Structure* **2006**, *14*, 265–274.
(24) Smith, C.; Kortemme, T. *J. Mol. Biol.* **2008**, *380*, 742–756.
(25) Lauck, F.; Smith, C.; Friedland, G.; Humphris, E.; Kortemme, T. *Nucleic Acids Res.* **2010**, *38*, W569–W575.
(26) Engh, R.; Huber, R. *Acta Crystallogr., Sect. A* **1991**, *47*, 392–400.
(27) Frenkel, D.; Smit, B. Appendix D. In *Understanding molecular simulation: from algorithms to applications*, 2nd ed.; Academic Press: San Diego, CA, 2002; pp 525−532.
(28) Hershko, A.; Ciechanover, A. *Annu. Rev. Biochem.* **1998**, *67*, 425–479.
(29) Hicke, L.; Schubert, H.; Hill, C. *Nat. Rev. Mol. Cell. Biol.* **2005**, *6*, 610–621.
(30) Tjandra, N.; Feller, S.; Pastor, R.; Bax, A. *J. Am. Chem. Soc.* **1995**, *117*, 12562–12566.
(31) Cornilescu, G.; Marquardt, J.; Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 6836–6837.
(32) Chou, J.; David, A.; Bax, A. *J. Am. Chem. Soc.* **2003**, *125*, 8959–8966.
(33) Lange, O.; Lakomek, N.-A.; Fares, C.; Schroder, G.; Walter, K.; Becker, S.; Meiler, J.; Grubmuller, H.; Griesinger, C.; de Groot, B. *Science* **2008**, *320*, 1471–1475.
(34) Maragakis, P.; Lindorff-Larsen, K.; Eastwood, M.; Dror, R.; Klepeis, J.; Arkin, I.; Jensen, M.; Xu, H.; Trbovic, N.; Friesner, R. *J. Phys. Chem. B* **2008**, *112*, 6155–6158.
(35) Richter, B.; Gsponer, J; Varnai, P.; Salvatella, X; Vendruscolo, M. *J. Biomol. NMR* **2007**, *37*, 117–135.
(36) Lindorff-Larsen, K.; Best, R.; DePristo, M.; Dobson, C.; Vendruscolo, M. *Nature* **2005**, *433*, 128–132.
(37) Nederveen, A.; Bonvin, A. *J. Chem. Theory Comput.* **2005**, *1*, 363–374.
(38) Paterlini, M.; Ferguson, D. *Chem. Phys.* **1998**, *236*, 243–252.
(39) Huang, K. Y.; Amodeo, G. A.; Tong, L.; McDermott, A. *Protein Sci.* **2011**, *20*, 630–639.
(40) Lindorff-Larsen, K.; Ferkinghoff-Borg, J. *PLoS One* **2009**, *4*, e4203.
(41) Yamashita, H.; Endo, S.; Wako, H.; Kidera, A. *Chem. Phys. Lett.* **2001**, *342*, 382–386.
(42) Jorgensen, W.; Tirado-Rives, J. *J. Phys. Chem.* **1996**, *100*, 14508–14513.
(43) Ulmschneider, J.; Ulmschneider, M.; Di Nola, A. *J. Phys. Chem. B* **2006**, *110*, 16733–16742.
(44) Karplus, M. *Methods Enzymol.* **1986**, *131*, 283–307.
(45) Ponder, J.; Richards, F. *J. Am. Chem. Soc.* **1987**, *8*, 1016–1024.
(46) Snow, C.; Qiu, L.; Du, D.; Gai, F.; Hagen, S.; Pande, V. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 4077–4082.
(47) Voelz, V.; Bowman, G.; Beauchamp, K.; Pande, V. *J. Am. Chem. Soc.* **2010**, *132*, 1526–1528.
(48) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
(49) Ferkinghoff-Borg, J. *Eur. Phys. J. B* **2002**, *29*, 481–484.
(50) Jorgensen, D. S.; Maxwell, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *18*, 11225–11236.
(51) Di Qiu Shenkin, F. P.; Hollinger, P. S.; Still, W. C. *J. Phys. Chem. A.* **1997**, *101*, 3005–3014.
(52) Chopraa, C. M.; Summab, G.; Levitt, M. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 20239–20244.
(53) Jiang, L.; Kuhlman, B.; Kortemme, T.; Baker, D. *Proteins* **2005**, *58*, 893–904.
(54) Schymkowitz, J.; Rousseau, F.; Martins, I.; Ferkinghoff-Borg, J.; Stricher, F.; Serrano, L. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 10147.
(55) Dunbrack, R. L.; Cohen, F. E. *Protein Sci.* **1997**, *6*, 1661–1681.