JCTC Journal of Chemical Theory and Computation
Article
pubs.acs.org/JCTC

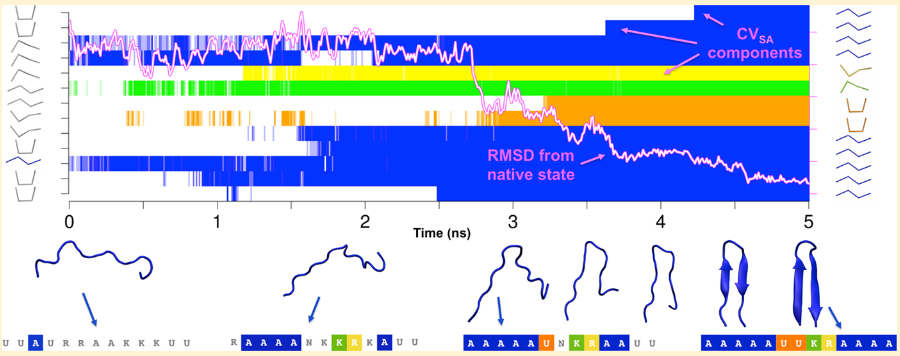
Using Local States To Drive the Sampling of Global Conformations in Proteins

Alessandro Pandini[†] and Arianna Fornili*[,‡]

[†]Department of Computer Science, College of Engineering, Design and Physical Sciences and Synthetic Biology Theme, Institute of Environment, Health and Societies, Brunel University London, Uxbridge UB8 3PH, United Kingdom

[‡]School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom

S *Supporting Information*



ABSTRACT: Conformational changes associated with protein function often occur beyond the time scale currently accessible to unbiased molecular dynamics (MD) simulations, so that different approaches have been developed to accelerate their sampling. Here we investigate how the knowledge of backbone conformations preferentially adopted by protein fragments, as contained in precalculated libraries known as structural alphabets (SA), can be used to explore the landscape of protein conformations in MD simulations. We find that (a) enhancing the sampling of native local states in both metadynamics and steered MD simulations allows the recovery of global folded states in small proteins; (b) folded states can still be recovered when the amount of information on the native local states is reduced by using a low-resolution version of the SA, where states are clustered into macrostates; and (c) sequences of SA states derived from collections of structural motifs can be used to sample alternative conformations of preselected protein regions. The present findings have potential impact on several applications, ranging from protein model refinement to protein folding and design.

INTRODUCTION

Conformational changes in proteins are often associated with protein–protein interactions,[1] ligand binding,[2] and posttranslational modifications.[3] They are at the basis of powerful mechanisms for functional regulation such as allostery,[4,5] and they can be fuelled by chemical reactions to produce large-scale mechanochemical motions in molecular motors.[6]

The structural and energetic characterization of conformational transitions is therefore of central interest in understanding protein function. Computational approaches such as molecular dynamics (MD) simulations offer a powerful way to investigate such processes with atomic resolution. However, the conformational transitions usually found in biologically relevant systems are beyond the time scale currently accessible to equilibrium MD simulations. Different methods have been developed to overcome this problem by accelerating the sampling of "rare events".[2,7−9]

Most of the available methods are based on the use of collective variables (CVs),[7] which define the part of the space where the sampling is enhanced and usually represent the progress of the system along the process of interest. A range of CVs have been used in the literature to describe protein conformational changes, ranging from general descriptors of the protein global shape to local geometric parameters specific to the process under examination.[10−18]

While the details of the global conformational landscape of a protein are uniquely determined by its amino acid sequence, it is now well-established that protein fragments tend to adopt recurring backbone conformations. In particular, coarse-grained and sequence-independent structural alphabets[19] (SAs) have been derived, which contain the minimal set of typical C$^{\alpha}$ conformations of protein fragments (local states or "letters") needed to reconstruct experimental protein structures with a high level of accuracy. SAs have been exploited in the past for a number of applications, including local structure[20] and flexibility[21] prediction, sequence-based structural comparison,[22] structure mining,[23−25] fold classification[26] and recognition,[27] and de novo prediction.[28,29] Recently they have been shown to correctly represent also the dynamical properties of proteins.[30]

Received: October 15, 2015

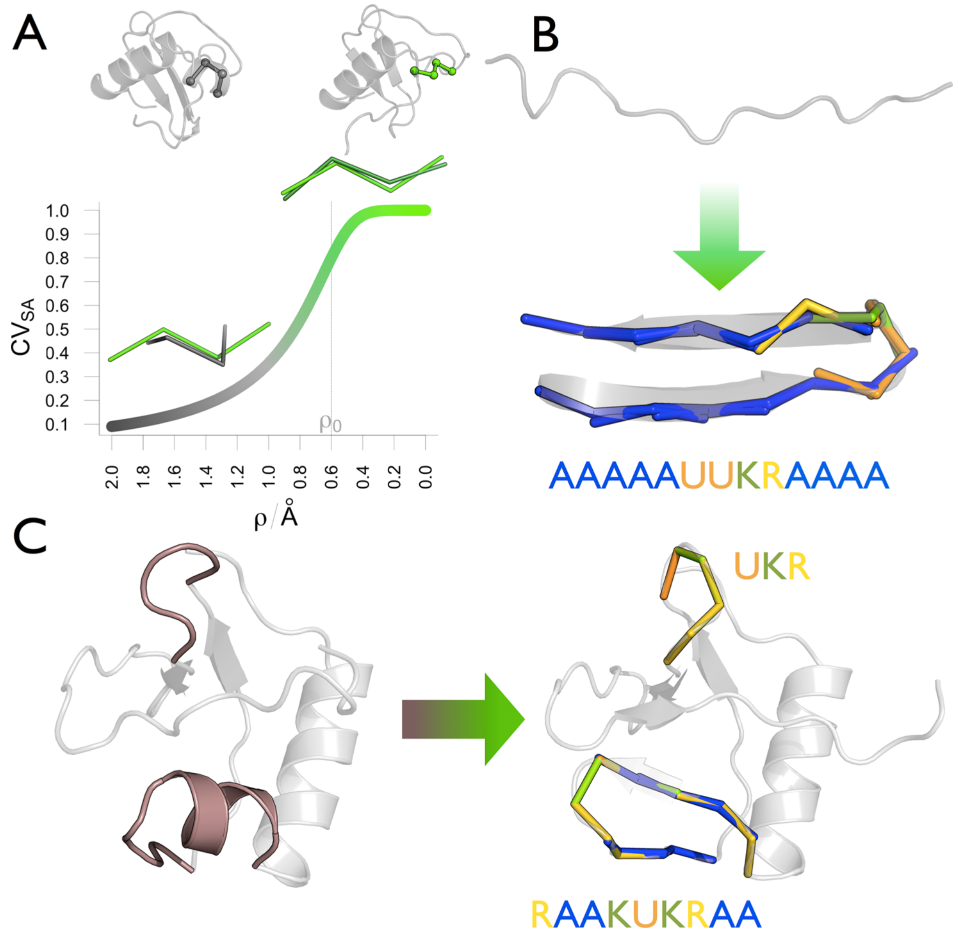


Figure 1. Schematic description of the CV_{SA} and its applications. (A) A plot of the switching function used in the definition of the CV_{SA} is reported, showing the dependence of the function on the C^{α} RMSD (ρ) of a fragment (gray to dark green licorice) from a reference SA state (green). Using the CV_{SA}, the sampling of local states can be biased for all fragments in a protein (B) or for fragments in selected regions (C).

Indeed, they have been used to analyze trajectories from MD simulations[31−33] and in particular to detect signal transmission in allosteric proteins.[32,34]

The main goal of the present work is to investigate if the a priori knowledge provided by SAs on the local states preferentially adopted in protein structures can be exploited to accelerate the exploration of protein conformations in MD simulations. In the following we use a SA-based CV (CV_{SA}) to guide the sampling of fragment conformations in small proteins toward predefined local states in metadynamics and steered MD (SMD) simulations.

We first address the question of whether enhancing the sampling of native local structures can accelerate the sampling of global native structures. We show that folded structures of small proteins can be reproducibly and efficiently recovered in short simulations using the SA letters that best represent the local native structures. We then reduce the extent of the information on the local native structure provided a priori by using a simplified or reduced version of the SA (rSA), where local states are replaced by macrostates defining fragment shapes instead of detailed structures. We show that in all cases the missing information required to get the global folded state is recovered by the MD sampling, which allows the fragments to adapt to their specific environment by adjusting their conformation and relative orientation.

Finally, we investigate possible ways to produce synthetic SA strings to guide the sampling when no information on the desired final state is available. In particular, we show that libraries of SA strings can be derived from databases of structural motifs and can be used to sample alternative conformations of parts of a protein. The resulting conformations can then be ranked a posteriori with a scoring function to identify the most native-like structures.

The present findings indicate that MD simulations and knowledge-based SAs can be effectively combined to enhance the exploration of the conformational space of proteins. The proposed CV_{SA} has a wide range of applications, ranging from protein model refinement to protein folding and design.

■ RESULTS

The results presented in this work are derived using a new CV based on the 25-letter SA M32K25[30] (CV_{SA}). To define a CV_{SA}, the structure of the protein is first partitioned into four-residue fragments of C^{α} atoms. The CV_{SA} is then calculated as a sum of single-fragment terms (eq 1 in Methods), each consisting of a switching function related to the deviation of the fragment from a predefined reference letter extracted from the SA (Figure 1A). The overall CV_{SA} measures the number of fragments matching the structure of their corresponding reference SA letters. The fragments used to define a CV_{SA} can overlap, and they can cover the whole protein or only selected regions.

In the following we investigate if, when combined with an enhanced sampling technique, the CV_{SA} can be used to guide the folding of the entire protein to its native state (Figure 1B) and to sample alternative conformations in selected regions (Figure 1C).

Folding 3D Structures from SA Strings. In this section we apply the CV_{SA} to fold two small fast-folders, the β-hairpin of the GB1 protein, and the Trp-cage mini-protein. These are small proteins that are known to adopt a stable fold in solution with relatively short folding times (6 μs for the GB1 β-hairpin[35] and 4 μs for the Trp-cage[36]), so that they are often used as model systems to study protein folding. As expected from their folding rates, equilibrium MD simulations in the nanosecond time scale are usually not suitable to observe folding events for these two proteins,[37] and either enhanced sampling[11,38,39] or long (~10−100 μs) MD simulations[40,41] have been used in the past.

In this section we use the CV_{SA} combined with the enhanced sampling method metadynamics[42] (Methods). The reference SA letters are determined from the local-fit encoding (Methods) of the experimental native structure, where each SA letter is the best match of a fragment in its native conformation (Table 1). CV_{SA} contains all the overlapping fragments in the protein. In the local-fit mode used to assign the letters, no information on adjacent fragments is used to determine the letter of a given fragment. The sampling along the CV_{SA} is then biased in metadynamics simulations (Methods), so that structures with CV_{SA} values in the whole accessible range $[0, {CV_{SA}}^{max}]$ are explored, where ${CV_{SA}}^{max}$ is equal to the number of fragments included in the CV_{SA}.

Table 1. M32K25 SA and rSA Strings Used for the Folding and Refinement Simulations

Folding		
GB1 β-hairpin	DBAALVWOQABAB	
GB1 β-hairpin (rSA)	AAAAAUUKRAAAA	
Trp-Cage	UUUUUVUPFXVTPRIHF	
Trp-Cage (rSA)	UUUUUUUKKUUUKRAAK	
KUNg1 β-hairpin (rSA)	AAAAKUNAAAA	
KUNg2 β-hairpin (rSA)	AAKUNAA	
TR464 Refinement	L1[a]	L2[a]
Best Model	RUUU**UUO**RGML[b]	L**UUO**R[d]
Target	SRAD**KVP**QICM	J**UOS**N
Best Model (rSA)	RUUU**UUK**RAAA	A**UUK**R
Target (rSA)	RRAA**KUK**RAAA	K**UKR**N
BlindL2a (rSA)	RAA**KUK**RAA[c]	**UKR**[e]
BlindL2b (rSA)	RAA**KUK**RAA	**KUN**
BlindL2c (rSA)	RAA**KUK**RAA	**NKU**

[a] Fragments covering the central loop residues are indicated in bold. [b] The fragments include residues 18−31. [c] The fragments include residues 19−30. [d] The fragments include residues 38−45. [e] The fragments include residues 39−44.

The CV_{SA} defined in this way contains information on the local structure of the native state but not on the global folded structure. Indeed, using the CV_{SA} differs from providing the coordinates of the folded conformation as reference structure in that (a) the native conformations of single fragments are used, with no information on the relative arrangement of different fragments, and (b) discrete SA states are used to represent fragment conformations, so that the CV_{SA} does not contain the actual native fragment structures, but the SA letters closest to them.

Folded β-hairpin conformations, i.e., conformations with C^{α} RMSD from the experimental native structure ($RMSD_{nat}$) ≤ 2 Å, were sampled multiple times within a 100 ns metadynamics simulation (Figure 2A) starting from an extended conformation, with a minimum $RMSD_{nat}$ of 0.4 Å. Considering that the CV_{SA} was defined using reference SA letters from the native structure, folded conformations are expected to have high CV_{SA} values. Indeed, almost all low-$RMSD_{nat}$ structures belong to the high-CV_{SA} ensemble, defined as composed of structures with $CV_{SA} \geq {CV_{SA}}^{max} - 2$ (blue points in Figure 2A). Moreover, after filtering the trajectory for high-CV_{SA} structures, a significant enrichment was found in folded conformations, going from a percentage of folded structures of 6% for the overall trajectory to 39% for the CV_{SA}-filtered one (Table 2). Metadynamics simulations of the Trp-cage showed a similar behavior. Folded structures were sampled with $RMSD_{nat}$ as low as 0.5 Å (Figure 2C), and high-CV_{SA} ensembles were composed for the 27% of native-like structures (Table 2). As expected, for both proteins no folded conformations were found in control unbiased MD simulations of the same length (Table 2).

Non-native structures with high-CV_{SA} values were also sampled (blue points above the dashed lines in Figure 2A,C). In these conformations the single fragments match their reference SA letter, but their global arrangement differs from the native one. Hairpin-like conformations were found where terminal segments are in a β-strand conformation, but they are too distant to form a β-sheet, resulting in a percentage of native contacts as low as 68% (Supporting Information Figure S1A). This is consistent with the fact that the CV_{SA} does not contain a bias toward nonlocal native contacts, so that non-native global arrangements of native local states can be sampled during the simulation.

In summary, the present results show that providing local structural information can enhance the folding toward the global native structure of small proteins. Using the CV_{SA} promotes the sampling of structures with the desired sequence of discrete local states (or strings of letters). During the simulation, different relative arrangements of these fragments are explored, with a significant fraction of folded global structures. In the following section we further reduce the information needed a priori to build the CV_{SA} by introducing a simplified version of the structural alphabet.

Generating SA Strings: Simplified Alphabet. The extent to which the target local structures are known before the simulation can vary significantly, and in some cases it might be limited to a prediction of the shape of the fragment. It is thus convenient to have a reduced set of SA letters representing more general fragment states (macrostates). To this end, a cluster analysis was performed on the 25 letters of the M32K25 SA (Methods). The resulting six clusters (Figure 3), named after the letter of their representative, describe fragment states that differ mainly for the pseudotorsion around the two central C^{α} atoms (Supporting Information Table S1).

SA encodings can be translated into rSA encodings by replacing each letter with the corresponding cluster representative (Table S1). For example, the GB1 β-hairpin SA string

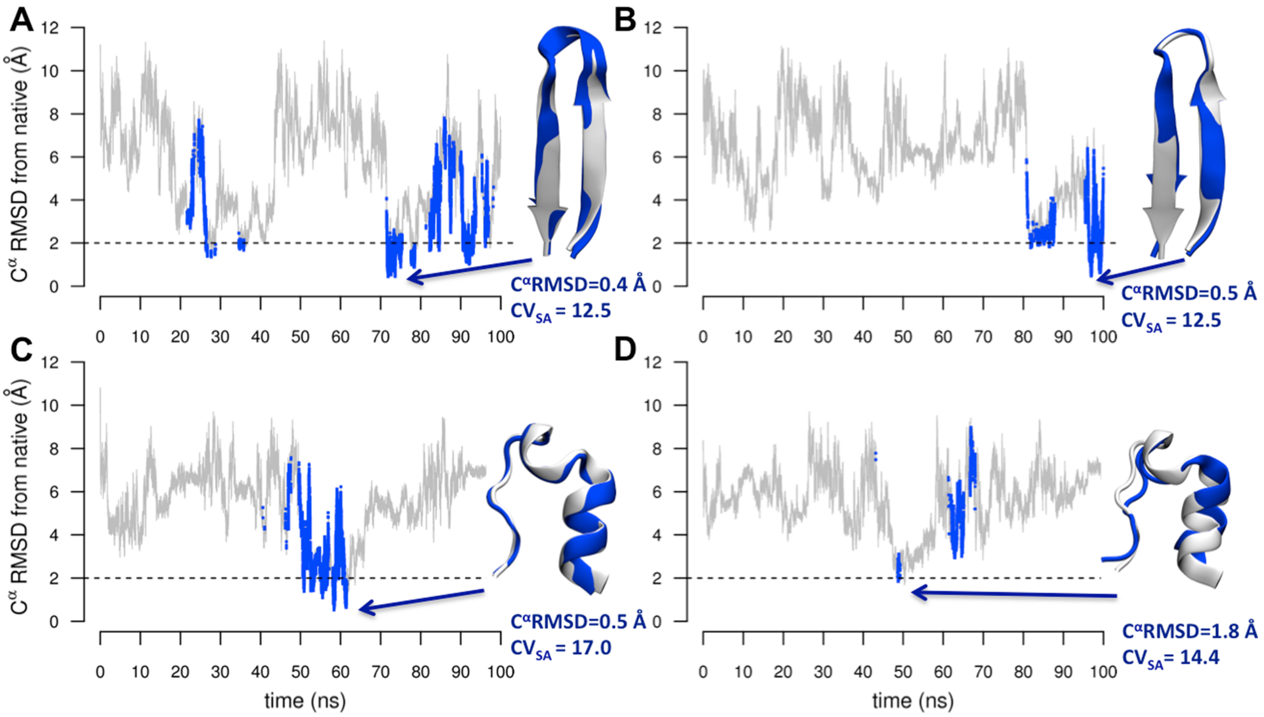


Figure 2. Time evolution of the C^α RMSD from the experimental native structure of the GB1 β-hairpin (upper panels) and the Trp-cage mini-protein (lower panels) during metadynamics simulations. The CV_{SA} was defined using the SA (left panels) and rSA (right panels) encodings of the experimental structures. High-CV_{SA} points are colored in blue. For each panel, the experimental structures (white cartoon) are superimposed to the best matching high-CV_{SA} structure (blue).

Table 2. Fraction of Native-like Structures in Metadynamics and SMD Simulations of the GB1 β-Hairpin and the Trp-Cage Mini-protein

	$p^{Nat}{}_{M}$ (%)[d]	$p^{Nat}{}_{FiltCVsa}$ (%)[e]	$p^{Nat}{}_{SMD}$ (%)[f]
GB1 β-hairpin (SA)[a]	5.9	39.4	
GB1 β-hairpin (rSA)[b]	2.0	37.5	36.0
GB1 β-hairpin (unbiased MD)[c]	0.0		
Trp-cage (SA)[a]	3.3	26.5	
Trp-cage (rSA)[b]	0.2	0.4	33.3
Trp-cage (unbiased MD)[c]	0.0		

[a] CV_{SA}-biased simulation with SA encoding. [b] CV_{SA}-biased simulation with rSA encoding. [c] Unbiased MD simulation. [d] Percentage of structures in the whole metadynamics trajectory with $RMSD_{nat} \leq 2$ Å. [e] Percentage of structures in the metadynamics high-CV_{SA} ensemble with $RMSD_{nat} \leq 2$ Å. [f] Percentage of productive SMD trajectories (final structure with $RMSD_{nat} \leq 2$ Å).

"DBAALVWOQABAB" becomes "AAAAAUUKRAAAA" (Table 1), where the turn-encoding letters are easily recognizable as being sandwiched between two stretches of strand-encoding "A" letters.

When going from the SA to the rSA encoding, the maximum RMSD between each fragment experimental structure and its corresponding letter for the two proteins increases from 0.4 to 1 Å, thus reducing the extent of information on the fragment target states contained in the rSA-based CV_{SA}. Nevertheless, metadynamics simulations showed that folded conformations can still be recovered for both proteins (Figure 2, panels B and D), with overall $RMSD_{nat}$ values as low as 0.5 Å (β-hairpin) and 1.8 Å (Trp-cage). Filtering the trajectories for high-CV_{SA} structures produced an enrichment in folded structures for the β-hairpin from 2 to 38% (Table 2). A less extensive sampling of high-CV_{SA} structures and hence of folded states was instead observed for the Trp-cage (Table 2).

The performance of the rSA-based CV_{SA} was also tested by using a different enhanced sampling technique, the steered MD[43] (SMD; see Methods), where the CV_{SA} was steered from the starting value to its maximum value $CV_{SA}{}^{max}$ in a predetermined amount of time and in multiple replicas. The folded state was recovered at the end of the simulation in about one-third of the runs (productive runs) for both molecules (Figure 4 and Table S2), confirming that the rSA-based CV_{SA} contains sufficient information to reach the folded state for both proteins.

Since the steering bias was applied on the overall CV_{SA}, the single fragments were free to reach the reference structure at different times (Figure S2). In the productive runs, the folding mechanism mainly resembled a zipper mechanism,[37] where the contacts between the β-strands started to form from the turn toward the termini (Figure S2).

In the nonproductive trajectories, the final structures reached the $CV_{SA}{}^{max}$ maximum value, but they were trapped in non-native conformations where the N- and C-terminal segments were differently aligned (Figure 4, orange structural ensembles) and a smaller fraction of native contacts was recovered (Figure

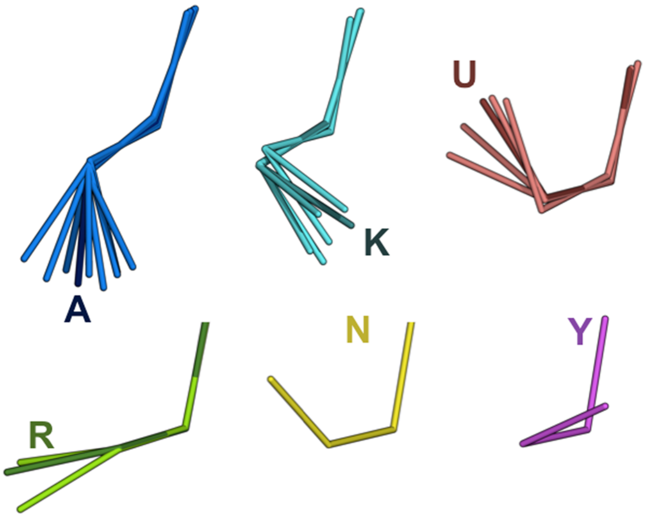


Figure 3. Licorice representation of the six clusters of M32K25 letters. The cluster representatives forming the reduced alphabet rSA are labeled and shown in darker colors.

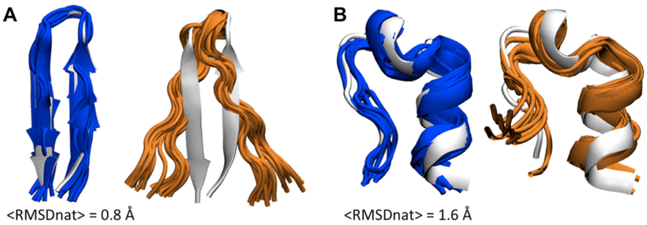


Figure 4. Productive (blue) and nonproductive (orange) SMD runs of the GB1 β-hairpin (A) and the Trp-cage mini-protein (B). Final SMD structures (colored cartoon) are superimposed onto the experimental structures (white cartoon). Low-work relaxed structures (Supporting Information) are reported. A SMD run is defined productive if $\mathrm{RMSD}_{\mathrm{nat}} \leq 2$ Å at the end of the simulation. The average value of $\mathrm{RMSD}_{\mathrm{nat}}$ is also reported for productive runs.

S3, orange lines) compared to the productive trajectories. The relaxation to the native state was hindered by the early formation of non-native contacts (Figure S4).

Interestingly, we observed that part of the nonproductive SMD trajectories can be filtered out without using information on the target native state by calculating the work needed to drive the overall transition for each trajectory. Indeed, the trajectories associated with low-work transformations (Supporting Information) were found to have a higher success rate in producing a correctly folded structure.

The present results show that the reduced representation of the SA can be effectively used to sample folded conformations of small proteins within both the metadynamics and the SMD frameworks. The generation of the rSA fragment macrostates is a first step toward the definition of artificial sequences of fragment states, which do not require a priori knowledge of the final structure. In the next section we show that recurring motives can be identified in the rSA encoding of experimental loop structures, indicating that libraries of predefined strings can be used to define the $\mathrm{CV}_{\mathrm{SA}}$ reference states.

Generating SA Strings: Recurrence of SA Motifs in Loops. Loops, defined as nonrepetitive structural units connecting regular secondary structures, have been shown to adopt recurring conformations by different classification schemes.[19] The existence of supersecondary structural motifs has been exploited in the past for both function and structure prediction.[19,44] In this section, we analyze the loops contained in the ArchDB database[44] (Supporting Information) to show that a large number of loop structures can be encoded by a small number of recurring rSA strings or rSA motifs. In the following we focus on β-hairpins of length 4 (ArchDB.BN4.100, Supporting Information), which are one of the most populated loop types in ArchDB,[44] but the analysis can be easily extended to any loop type. For each β-hairpin, the structure of the loop plus the first residue from each flanking β-strand was encoded into three-letter strings with the rSA.

Of the 56 possible combinations of six letters in strings of length 3, only 26 rSA motifs were observed in the 3314 loop structures of the ArchDB.BN4.100 data set (Table S3). Interestingly, 97% of all the structures fall in the first five rSA motifs (Figure 5). All together, these motifs cover 20 of the original 21 ArchDB subclasses (Supporting Information). The structures of the five rSA motifs are significantly different from each other, with an average intermotif RMSD of 2.2 Å (Table S4).

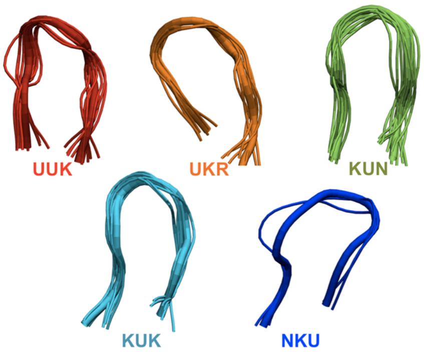


Figure 5. Structures of the most populated rSA motifs for β-hairpin loops of length 4 in the ArchDB.BN4.100 database. The cluster representatives of the loop structures belonging to each motif are shown, with the most populated cluster highlighted as a thicker tube.

The existence of a small number of recurring rSA motifs can be exploited to build $\mathrm{CV}_{\mathrm{SA}}$ variables that are not tailored to a specific experimental structure as was done in the previous sections. As we will see in the next section, such $\mathrm{CV}_{\mathrm{SA}}$ can be used to dictate the general shape of a loop, while the specific conformation is determined by the amino acid sequence and/or the environment of the loop.

Folding Different Amino Acid Sequences Using the Same rSA String. In this section we show that simulations based on the same rSA string can correctly reproduce differences in loops that, while having a similar shape, present structural differences due to their different amino acid composition.

We identified two groups of β-hairpin loops with similar shape (encoded by the rSA string "KUN") but with differences in the specific structure that correlate with differences in their amino acid sequence (Supporting Information).

A structural superimposition (Figure 6A) shows that KUN loops are clustered in two groups (KUNg1 and KUNg2), with KUNg2 structures (green) featuring a more bent conformation around the residue in position p4 compared to KUNg1 (blue), a different orientation of the p4 CO bond (licorice), and significantly different Ramachandran plots at the p4 and p5 positions (arrows in Figure 6D). A sequence alignment shows

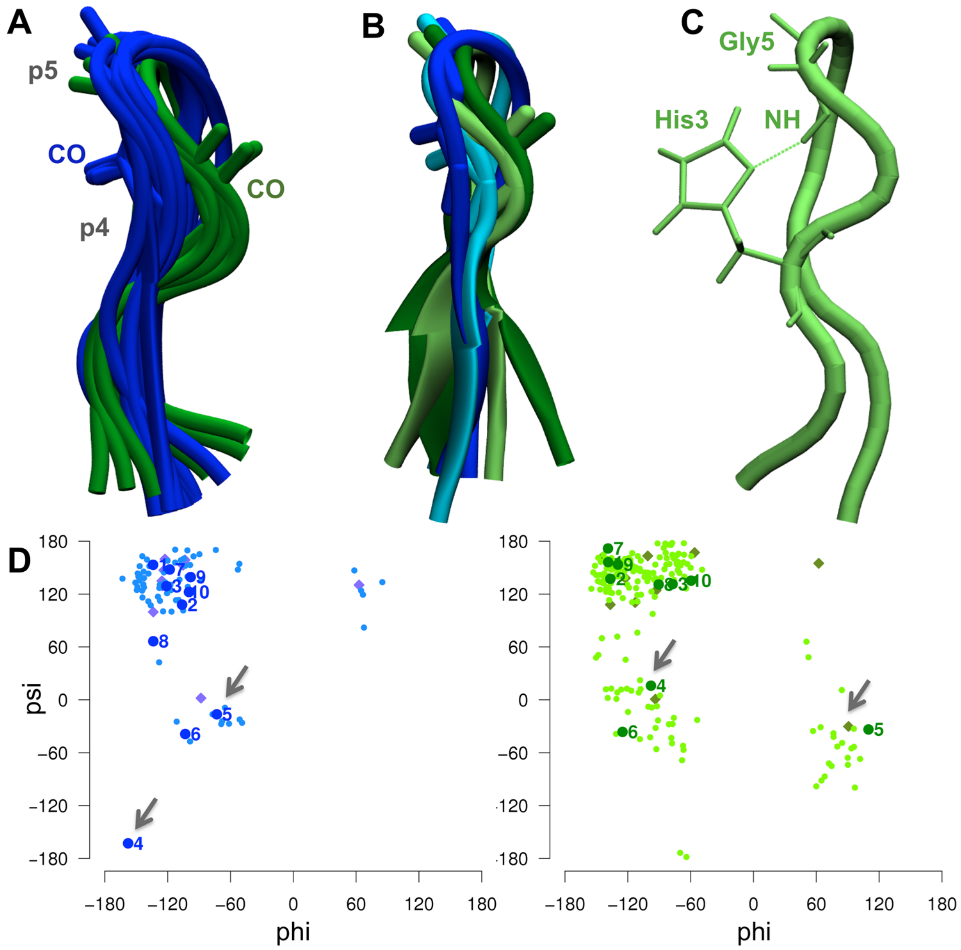


Figure 6. Comparison of KUNg1 and KUNg2 *β*-hairpins. (A) Superimposition of experimental structures of KUNg1 (blue) and KUNg2 (green) *β*-hairpins. (B) Superimposition of experimental (dark colors) and simulated (light colors) structures of KUNg1 (blue) and KUNg2 (green) *β*-hairpins. The simulated structures were extracted from the high-CV$_{SA}$ metadynamics ensembles with a cluster analysis. The backbone CO bonds of the residues in positions p4 and p5 are represented as licorice. (C) His3-Gly5 hydrogen-bonded interaction in the KUNg2 metadynamics structure. (D) Ramachandran plots of experimental (large numbered circles), metadynamics (diamonds), and SMD (small circles) folded structures for KUNg1 (left panel) and KUNg2 (right panel).

that all KUNg2 loops have a Gly in position p5 (Figure S5A), which allows for the larger bending of the loop.

Two sequences representative of the two groups were selected. CV$_{SA}$-biased MD simulations were run to fold each amino acid sequence into its corresponding *β*-hairpin starting from an extended conformation. The same rSA encoding was used for the two *β*-hairpins (Table 1).

In both cases, the native state was recovered in metadynamics and SMD simulations (Table S5), with a percentage of native-like structures ranging from 4−10% (metadynamics) to 36−72% (SMD). Remarkably, even if the same rSA encoding was used for both amino acid sequences, the simulated folded structures reproduce the differences between the experimental ones. Indeed, KUNg2 simulated structures (Figure 6B, light green) are more bent at p4 than KUNg1 (light blue) and the simulated conformations of each amino acid sequence are more similar to the experimental structure with the same sequence than to the other one (Table S6). Correspondingly, the Ramachandran plots of simulated structures show differences in the distribution of φ and ψ angles that parallel the differences between the experimental structures, in particular for residues in positions p4 and p5 (arrows in Figure 6D).

A possible explanation of the conformation adopted by KUNg2 at position p4 can be found by looking at the nearby side chains. Indeed, the bent loop arrangement in KUNg2 allows the Gly5 backbone NH group to be on the same side of the His3 side chain, favoring the formation of a His3-Gly5 hydrogen bond both in metadynamics (Figure 6C) and SMD (Figure S5B) simulations.

The present results indicate that the CV$_{SA}$ based on rSA motifs contains sufficient information to guide the folding toward the correct general shape of the loop backbone while at the same time allowing for adjustments to obtain sequence-specific structures. Side chains are not included in the definition of CV$_{SA}$, but they are explicitly taken into account during the simulation by the all-atom force field and they can modulate the loop conformation via direct side chain−backbone interactions.

Real Life Application: Protein Model Refinement. In this section we use the CV$_{SA}$ and the library of loop motifs to refine a protein model. Differently from the previous sections, only part of the structure is considered in the definition of the CV$_{SA}$, while the rest is left unbiased during the simulation.

The protein to be refined was selected from the targets of the Refinement category of the CASP8 exercise (Supporting Information). The best model generated in the normal prediction exercise failed to correctly describe two regions L1 (residues 19−30) and L2 (residues 39−44) (Figure 7A), indicated by CASP organizers as problematic and showing deviations from the experimental structure > 4.0 Å (Table 3).

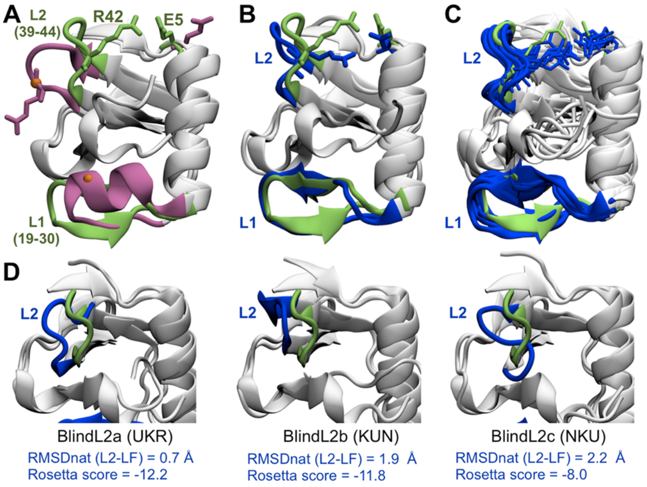


Figure 7. Refinement of the TR464 target from CASP8. (A) Superimposition of the L1 (residues 19−30) and L2 (residues 39−44) regions for the starting unrefined model (pink) and the target experimental structure (green). The rest of the structures is represented as white cartoon. Residues E5 and R42 are highlighted as licorice. A E5-R42 salt bridge interaction is formed only in the experimental structure. (B) Superimposition of the L1 and L2 regions in the target structure (green) and in a representative structure from metadynamics (blue). The E5-R42 interaction is recovered in the metadynamics simulation (blue licorice). (C) Superimposition of the L1 and L2 regions in final structures from productive SMD runs (blue cartoon). The E5-R42 salt bridge is recovered in all of the productive SMD runs. (D) Comparison of the structures with the best Rosetta score (blue) from the high-CV_{SA} metadynamics ensembles obtained using the BlindL2a (left), BlindL2b (center), and BlindL2c (right) encodings for L2. The Rosetta score is reported together with the L2 $RMSD_{nat}$ calculated in local fit.

The rSA encoding sequences of L1 and L2 in the best model and in the target experimental structure (Table 1) show that L1 changes from an $\alpha-\beta$-hairpin with the central turn in a "UUK" conformation to a "KUK" β-hairpin, while the central turn in L2 changes from "UUK" to "UKR". Interestingly, these three turn encodings are among the top five motifs for length-4 loops in β-hairpins described before (Table S3).

Enhanced sampling simulations were first run with a CV_{SA} based on the rSA target encodings of L1 and L2 (Target rSA in Table 1) to test if local information can be used to guide the refinement. A single CV_{SA} coordinate was used containing the fragments of both regions. The distance between the structures sampled during the simulations and the target conformation was measured by calculating the C^{α} RMSD for the overall structure ($RMSD_{nat}$) and separately for the two regions L1 ($RMSD_{nat}$(L1)) and L2 ($RMSD_{nat}$(L2)).

Structures significantly closer to the target structure than the starting model were sampled during 100 ns long metadynamics simulations. Both $RMSD_{nat}$(L1) and $RMSD_{nat}$(L2) were reduced to ≤2 Å in ∼21% of the CV_{SA}-filtered trajectory ($p^{Nat}_{FiltCVSA}$ in Table 3), with deviations for the single regions as low as 0.5 (L1) and 1.5 Å (L2) for the best metadynamics structure (Figure 7B). Similarly, SMD simulations with rSA encoding improved the starting model in 18% of the runs, with $RMSD_{nat}$ values as low as 1.7 Å (Figure 7C and Table S7). Control unbiased simulations failed to significantly improve the starting model. Indeed, no structures were found where both L1 and L2 were in a native-like conformation (unbiased MD p^{Nat} in Table 3).

The analysis of the time evolution of $RMSD_{nat}$ for metadynamics (Figure S6B, upper panel) and SMD runs (Figure S7A,B) indicates that L2 is the last loop to adopt the native conformation, suggesting that its rearrangement is the difficult step in the model refinement.

To test the performance of CV_{SA} in a real life situation where the target state is not known, multiple metadynamics simulations were run using "blind" rSA encodings for the L2 loop (BlindL2a−c in Table 1). L2 encodings were chosen among the top five motifs for length 4 loops in β-hairpins identified in the previous section (Figure 5). The BlindL2a encoding "UKR" coincides with the native Target rSA encoding used in the previous calculations.

The blind trajectories were filtered for high CV_{SA} values ($CV_{SA} \geq CV_{SA}^{max} - 2$), and the resulting structures were rescored using the Rosetta scoring function[45,46] (Methods). The best scoring structure was found in the BlindL2a simulation (Table S8), with a $RMSD_{nat}$(L2) of 0.7 Å when using only L2 for the best fit superposition (local fit or LF) and of 3.0 Å when calculated using the whole protein structure (global fit or GF). Indeed, the L2 conformation in the best CV_{SA}-refined structure (blue in Figure 7D, left panel), while featuring a residual translational displacement from the experimental structure (green), has a very good match to the experimental shape. Remarkably, 13 of the top 20 scoring structures are from the BlindL2a simulation (gray rows in Table S8).

The best structures from the other two blind simulations differed significantly from the experimental L2 conformation (BlindL2b and BlindL2c in Figure 7D), with $RMSD_{nat}$(L2)

Table 3. Refinement of the CASP8 Target TR464

	p^{Nat} (%)[b]	$p^{Nat}_{FiltCVsa}$ (%)[c]	$RMSD_{nat}$ (Å)[d]	$RMSD_{nat}$(L1) (Å)[e]	$RMSD_{nat}$(L2) (Å)[e]
Best Model			2.94	4.00 (3.22)	4.35 (2.02)
Best Refined Model			2.23	1.57 (1.38)	2.38 (0.49)
Metadyn (Target)[a]	2.1	21.4	2.00	0.53 (0.27)	1.50 (0.30)
unbiased MD	0.0		2.32	2.86 (2.45)	2.48 (1.64)

[a] CVSA metadynamics simulation run using the Target rSA string. [b] Percentage of structures in the whole trajectory with both $RMSD_{nat}$(L1) and $RMSD_{nat}$(L2) ≤ 2 Å. [c] Percentage of structures in the high-CVSA ensemble with both $RMSD_{nat}$(L1) and $RMSD_{nat}$(L2) ≤ 2 Å. [d] C^{α} RMSD from the native structure calculated over the whole structure. Minimum values observed during the simulation are reported for metadynamics and unbiased MD. [e] C^{α} RMSD from the native structure calculated over only L1 and L2. Minimum values observed during the simulation are reported for metadynamics and unbiased MD. Values in parentheses are calculated in local fit (only the L1 or L2 structures are used for the best fit superposition instead of the whole structure).

DOI: 10.1021/acs.jctc.5b00992

values of 2.2 (BlindL2b) and 1.9 Å (BlindL2c) in the local fit. Interestingly, they have also poorer scores according to Rosetta (Table S8). Thus, rescoring the structures with Rosetta allows BlindL2a structures to be identified as the closest to the native state without using information on the experimental structure.

To summarize this section, we showed how the CV_{SA} coupled with libraries of rSA loop motifs can be effectively used to sample multiple alternative loop conformations and, when combined with a knowledge-based rescoring potential, to refine protein models.

■ DISCUSSION

The occurrence of recurring local structures in proteins has inspired several approaches to structure prediction and design. Large libraries of sequence-dependent fragments have been successfully employed in fragment-assembly strategies such as Rosetta.[45−47] Alternatively, small sets of coarse-grained sequence-independent fragments were used as structural alphabets,[30,48−50] often coupled with machine learning predictors.[20] Recently, it was also shown that local structural changes observed in MD conformational ensembles can be analyzed in terms of changes of SA letters without loss of information. This was particularly true for the M32K25 SA, which has been derived from conformational attractors, i.e., regions in the fragment conformational space that are highly populated by experimental structures.[30] The use of this SA turned out to be particularly effective in investigating allosteric proteins.[32]

All these data indicate that SAs can provide a compact and reliable representation of the most populated conformational states of protein fragments, recapitulating the structural features of a large number of experimental structures.[30] The central idea of the present work is to exploit the experimental information distilled in a SA to accelerate the exploration of protein conformations in MD simulations. We thus introduced a SA-based collective variable (CV_{SA}) to control the match between simulated and SA fragment conformations. Combining the CV_{SA} with either metadynamics or steered MD techniques, it was possible to bias the sampling of fragment conformations toward experimentally preferred local states. While SAs have been used in the past for postprocessing MD trajectories,[31−33] this is the first time to our knowledge that they are used to enhance the sampling during an MD simulation.

The use of the CV_{SA} allows the introduction of knowledge-based elements in the simulation without loss of generality. Since the CV_{SA} is based on local states, no assumption is required on nonlocal contacts and thus on the relative arrangement of the fragments. Moreover, the SA fragments contain only C^{α} atoms, so that they can be used with any amino acid sequence and no a priori information is needed on side chain structures.

CVs based on secondary structures have been used in the past either to accelerate the folding of small proteins to their native states[13] or to explore the space of their accessible folds.[51] Performances comparable to the CV_{SA} were obtained when folding the GB1 β-hairpin with metadynamics simulations.[13] However, these CVs are based on canonical structures of blocks of secondary structure elements, and for β structures they contain pairs of β-strands.[13] The CV_{SA} differ from these in that (a) it does not require nonlocal information on specific arrangements of secondary structure elements and (b) it can be used to describe regions with irregular structure.

The performance of the CV_{SA} was first tested on the folding of peptides and mini-proteins. In all cases, conformations with a CV_{SA} value equal or close to $CV_{SA}{}^{max}$ were sampled during metadynamics simulations starting from unfolded conformations. In these high-CV_{SA} ensembles, all the fragments were in their target SA state or close to it, indicating that the structures had a local native-like conformation. Remarkably, a significant portion of each ensemble (27−40%) had also a global native-like conformation. In the SMD runs, the CV_{SA} was explicitly steered to its maximum value, so that high-CV_{SA} states were ensured to be sampled in a fixed amount of time. Similarly to metadynamics, the ensemble of high-CV_{SA} conformations composed by the final SMD structures had a significant proportion of native-like conformations, with percentages up to 72%.

Folded structures were thus observed when the fragments, in addition to being in the correct local state, had also a native-like arrangement, with native-like interfragment contacts. No information was directly provided on nonlocal native contacts, but folded structures were successfully formed during the relatively short simulations performed here. The bias on the CV_{SA} ensured an increased probability of observing native-like local structures, which in turn increased the probability of finding them in a global native-like arrangement compared to an unbiased simulation.

While the local states used to perform these calculations were derived from the target global structures, experimental and predicted information on the local structure of a protein may be available even in the absence of its global structure. For example, the experimental secondary structure composition can be derived from CD spectra,[52] while the sequence of secondary structure elements can be usually predicted with a high level of accuracy.[53−55] For irregular regions, libraries of structural motifs are available,[44] while protein regions involved in conformational changes can be identified by hydrogen/deuterium exchange mass spectrometry.[56]

By construction, high CV_{SA} values can be used to discriminate conformations with a local native structure among those generated during the simulation, but these are not necessarily globally folded. To fold larger and more complex systems than the ones considered in this work in a comparable amount of time, additional information on interfragment contacts would need to be provided. This information could be derived from inter-residue contact prediction, for example using recently proposed methods based on coevolution.[57] On the other end, when no contact information is provided a priori, the possibility to sample multiple global arrangements compatible with the same sequence of local states could be exploited for protein design methods. Indeed, simulations using the same reference SA string with different amino acid sequences would show how the spectrum of different interfragment arrangements is modulated by the primary structure of the protein.

The form chosen for the CV_{SA} (eq 1) allows for some degree of structural variability also in the local structure of high-CV_{SA} ensembles. Indeed, the conformation of a fragment is not required to exactly match its reference SA letter to contribute significantly to the CV_{SA}, but small adjustments in the local structure are possible if energetically favored. This behavior is regulated by the ρ_0 parameter (eq 1), which defines the tolerance on the fragment deviation from the reference letter in the switching function. In most of the calculations, a value of 0.6 Å was used, which is comparable to the cluster radius in the macrostates of the reduced alphabet (rSA). When using a CV_{SA} with the rSA encoding, the requirement for the fragment is thus to match a macrostate, with freedom to adapt to any of the SA states that compose it.

Coupling the MD sampling with the reduced version of the alphabet rSA instead of the full SA has the advantage that less information on the system needs to be provided before the simulation. Indeed, even when using the same reference rSA string for different amino acid sequences, the structural differences of the experimental structures are still recovered during the MD simulation. While the rSA-based CV_{SA} provides information on the sequence of local macrostates, during the simulation the fragments can adopt the states that best match their specific chemical environment.

The reduced complexity of rSA strings can be exploited for the generation of guess reference strings for the CV_{SA} without detailed knowledge of the desired final structure. In particular, secondary structure predictors[53−55] can be used for regular structures, while loop encodings can be extracted from loop databases. Indeed, we showed that a large part of the loops in experimental β-hairpin structures can be described by a reduced number of rSA strings. This analysis can be easily extended to other types of loops,[44] generating a comprehensive library of rSA motifs.

The use of rSA sequence libraries was exemplified with the refinement of a protein model. Alternative rSA sequences, generated from the most populated motifs for β-hairpins, were used to refine the L2 loop in the CASP8 target TR464. No information on the final state was used to define the CV_{SA}. L2 structures closer to the native state than the starting model were identified by rescoring with Rosetta. In particular, the ensemble generated with the native rSA motif could be identified as the best one because of the higher proportion of its conformations in the top ranking structures after rescoring.

The results discussed above suggest that the combination of CV_{SA}-based MD sampling with rSA libraries is a promising approach for the development of protein refinement methods. In particular, the increasing availability of parallel computing resources could be exploited to test a large number of rSA strings, which might be needed to be considered when multiple regions are involved in the refinement.

The development of refinement methods that can systematically improve the quality of protein models has proven to be particularly challenging so far, and it is still an area of ongoing work. Structure prediction relying on knowledge-based approaches seems to have reached a plateau in their accuracy,[58] and different refinement strategies are now required to achieve effective improvements. The combination of knowledge-based prediction with physics-based methods looks particularly promising. Indeed, an MD-based method was found for the first time to be the best performing in recent rounds of CASP.[58,59] Key factors in this success were the coupling of MD with knowledge-based elements and the use of an averaging procedure over multiple parallel trajectories to enhance the structure sampling and to generate an enrichment of native-like versus non-native features.[60] Another example of these types of approaches is the use of distance maps taken from high-resolution experimental structures as restraints in MD simulations.[59]

Sampling remains critical especially if the refinement requires large energy barriers to be overcome.[58] In these cases, enhanced sampling, as performed in our study, is necessary. Moreover, CASP11 results demonstrated that loops tend to be more challenging to refine. To this end, a CV_{SA}-based strategy is particularly suitable since it uses tunable local biases for loop regions.

Enhanced sampling methods were used in this work mainly to accelerate the sampling of high-CV_{SA} conformations and not to derive energetic or kinetic data. Snapshots from metadynamics simulations were rescored with an external scoring function for model refinement, while work values in SMD simulations were only used to prefilter candidate structures. However, provided that appropriate simulation lengths and postprocessing protocols are used, it is in principle possible to extract direct energetic information from CV_{SA}-biased trajectories and derive free energy changes for the processes involved, ranging from free energy changes associated with loop rearrangements or changes in secondary structure to folding energies. Moreover, when combined with suitable global CVs, the CV_{SA} could be used to investigate the relative kinetics of formation of secondary and tertiary elements[61,62] in the folding of proteins, for example when comparing diffusion−collision and nucleation−condensation pathways.[63] Following recent successful examples where secondary structure-based CVs were combined with NMR chemical shifts to study denatured states[64] and IDPs,[65] the CV_{SA} could in principle be coupled with CVs containing specific experimental information on unfolded or denatured states. In this context, using the CV_{SA} is particularly suitable since it is able to describe both regular and irregular local structures.

As a final remark, the CV_{SA} can be used in any CV-based enhanced sampling approach, including hybrid approaches mixing CVs with REMD-like methods, such as parallel tempering[11] or bias exchange metadynamics,[66] where exchanges are allowed between replicas that use different CVs. The second approach would be particularly suitable to explore alternative local structure arrangements by using multiple CV_{SA} with different SA strings.

In conclusion, we showed that, by enhancing the sampling of local states from a structural alphabet, it is possible to recover the global native state in MD simulations of small proteins. This finding is robust against approximate definitions of local states. Moreover, we showed how artificial sequences of SA states from libraries of recurring SA motifs can be used to generate alternative conformations of protein regions. Biasing the sampling of local states has a wide range of potential applications, going from protein design to the study of conformational changes based on large local rearrangements such as hinged motions, secondary structure transitions, or loop remodelling.

■ METHODS

Full details on methods, simulation setup and data analyses can be found in the Supporting Information.

Structural Alphabet and Structure Encoding. A SA is a collection of prototypical backbone conformations adopted by short fragments in protein structures, where each letter represents a fragment conformational state. In this work, we used the M32K25 SA, composed of 25 representative fragments of four consecutive C^{α} atoms.[30] A protein structure can be encoded into a SA string by progressively labeling each overlapping four-residue fragment (i.e., residues 1−4 for fragment 1, 2−5 for fragment 2, and so on) from the N-term to the C-term of the protein with a SA letter (A−Y), so that conformation of a protein of N residues is encoded into a structural string of length $N - 3$. In this work, the labeling of a

DOI: 10.1021/acs.jctc.5b00992

fragment is performed by identifying the SA letter that has the minimum root-mean-square deviation (RMSD) from the fragment (local-fit encoding). No information from adjacent fragments is used, so that letters encoding consecutive fragments are assigned independently from each other, allowing for a nonexact match in the overlapping region.

Local macrostates were defined by clustering the 25 letters of M32K25 (Supporting Information). The representatives of the six resulting clusters define a reduced version of the M32K25 SA (rSA).

CV$_{SA}$ Definition. CV$_{SA}$ is defined as the number of four-residue fragments f_i in the protein with C$^\alpha$ RMSD ρ from a preassigned SA letter X_i (reference state) within a given cutoff ρ_0:

$$\mathrm{CV}_{\mathrm{SA}} = \sum_{i=1}^{N_{\mathrm{frag}}} \frac{(1 - (\rho(f_i, X_i)/\rho_0)^n)}{(1 - (\rho(f_i, X_i)/\rho_0)^m)} \tag{1}$$

Each term in the sum is a differentiable function[13] switching from 1 ($\rho \ll \rho_0$, f_i very close to the reference state X_i) to 0 ($\rho \gg \rho_0$, f_i very far from X_i), where n and m are user-defined parameters that modulate the switching rate. It follows that the maximum value that a CV$_{SA}$ can adopt (CV$_{SA}{}^{\max}$) corresponds to the number of fragments used for its definition. The RMSD ρ is calculated on the positions of C$^\alpha$ atoms. The CV$_{SA}$ was implemented in a modified version of PLUMED[10] 1.3. The resulting user interface for the CV$_{SA}$ definition is flexible, and any number of fragments can be used, spanning the entire protein structure or parts of it. The sequence of X_i letters defines the CV$_{SA}$ reference string. The CV$_{SA}$ section of sample PLUMED input files is reported in the Supporting Information for the GB1 β-hairpin (Appendix S1) and TR464 (Appendix S2). Different criteria can be used for the string assignment, as it is shown in Results. The CV$_{SA}$ can be used in combination with any CV-based enhanced sampling method implemented in PLUMED. The patch files used to implement the CV$_{SA}$ can be downloaded from https://afornililab.wordpress.com/software or http://people.brunel.ac.uk/~csstaap2/software.html.

MD Simulations. The GROMACS 4.5.5 program[67] was used to prepare the initial system coordinates and to run the simulations. The Amber99SB*-ILDN[68] force field was used for all the simulations. Enhanced sampling simulations were performed by coupling GROMACS with PLUMED-1.3.[10] In the metadynamics folding simulations, a CV describing a global property was used in addition to the CV$_{SA}$. In SMD simulations, the steering was performed with harmonic restraints moving at constant velocity.

Tools for Trajectory Postprocessing and Analysis. MD trajectories were analyzed with the GROMACS 4.5.5 tools[67] and with in-house scripts in R (available from https://afornililab.wordpress.com/software or http://people.brunel.ac.uk/~csstaap2/software.html). The bio3d[69] R package was used for coordinate manipulation and for the analysis of the ArchDB database.[44] The encoding of experimental structures and MD trajectories with the M32K25 SA was performed with GSATools.[34] Structures from the blind loop refinement of the TR464 CASP target were independently rescored with Rosetta[45−47] (ver. 3.3).

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00992.

Experimental procedures, figures reporting time evolution of contacts, RMSD and CV$_{SA}$, final structures from SMD simulations and work value distributions, and tables reporting SA clustering data, fractions of native-like structures, rSA motifs in ArchDB, Rosetta scores for TR464, composition of the simulated systems, and number and length of SMD runs, and appendices with PLUMED input files (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: a.fornili@qmul.ac.uk.

Funding

This work was supported by the British Heart Foundation (Grant FS/12/41/29724 to A.F.). We also acknowledge the support of the HPC-EUROPA2 project (Project No. 228398) with the support of the European Commission—Capacities Area—Research Infrastructures.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We gratefully acknowledge the computer resources, technical expertise, and assistance provided by the Red Española de Supercomputación. We also thank Cristian Micheletti for insightful comments and continued support throughout the work and Michael Sadowski and Katarzyna Maksimiak for useful discussions.

■ REFERENCES

(1) Nussinov, R.; Ma, B. Protein dynamics and conformational selection in bidirectional signal transduction. *BMC Biol.* **2012**, *10*, 2.

(2) Fornili, A.; Giabbai, B.; Garau, G.; Degano, M. Energy Landscapes Associated with Macromolecular Conformational Changes from Endpoint Structures. *J. Am. Chem. Soc.* **2010**, *132*, 17570−17577.

(3) Autore, F.; Pagano, B.; Fornili, A.; Rittinger, K.; Fraternali, F. In silico phosphorylation of the autoinhibited form of p47(phox): insights into the mechanism of activation. *Biophys. J.* **2010**, *99*, 3716−3725.

(4) Changeux, J.-P. 50 years of allosteric interactions: the twists and turns of the models. *Nat. Rev. Mol. Cell Biol.* **2013**, *14*, 819−11829.

(5) del Sol, A.; Tsai, C.-J.; Ma, B.; Nussinov, R. The origin of allosteric functional modulation: multiple pre-existing pathways. *Structure* **2009**, *17*, 1042−1050.

(6) Veigel, C.; Schmidt, C. F. Moving into the cell: single-molecule studies of molecular motors in complex environments. *Nat. Rev. Mol. Cell Biol.* **2011**, *12*, 163−176.

(7) Laio, A.; Gervasio, F. L. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep. Prog. Phys.* **2008**, *71*, 126601.

(8) Noé, F.; Fischer, S. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154−162.

(9) Garcia, A. E.; Onuchic, J. N. Folding a protein in a computer: an atomic description of the folding/unfolding of protein A. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 13898−13903.

(10) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun.* **2009**, *180*, 1961−1972.

(11) Bussi, G.; Gervasio, F. L.; Laio, A.; Parrinello, M. Free-energy landscape for beta hairpin folding from combined parallel tempering and metadynamics. *J. Am. Chem. Soc.* **2006**, *128*, 13435−13441.

(12) Branduardi, D.; Gervasio, F. L.; Parrinello, M. From A to B in free energy space. *J. Chem. Phys.* **2007**, *126*, 054103.

(13) Pietrucci, F.; Laio, A. A collective variable for the efficient exploration of protein beta-sheet structures: application to SH3 and GB1. *J. Chem. Theory Comput.* **2009**, *5*, 2197−2201.

(14) Batista, P. R.; Pandey, G.; Pascutti, P. G.; Bisch, P. M.; Perahia, D.; Robert, C. H. Free Energy Profiles along Consensus Normal Modes Provide Insight into HIV-1 Protease Flap Opening. *J. Chem. Theory Comput.* **2011**, *7*, 2348−2352.

(15) Kalgin, I. V.; Caflisch, A.; Chekmarev, S. F.; Karplus, M. New Insights into the Folding of a β-Sheet Miniprotein in a Reduced Space of Collective Hydrogen Bond Variables: Application to a Hydrodynamic Analysis of the Folding Flow. *J. Phys. Chem. B* **2013**, *117*, 6092−6105.

(16) Huang, D.; Caflisch, A. Evolutionary Conserved Tyr169 Stabilizes the β2-α2 Loop of the Prion Protein. *J. Am. Chem. Soc.* **2015**, *137*, 2948−2957.

(17) Zheng, W.; De Sancho, D.; Hoppe, T.; Best, R. B. Dependence of internal friction on folding mechanism. *J. Am. Chem. Soc.* **2015**, *137*, 3283−3290.

(18) Sutto, L.; Gervasio, F. L. Effects of oncogenic mutations on the conformational free-energy landscape of EGFR kinase. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 10616−10621.

(19) Offmann, B.; Tyagi, M.; de Brevern, A. G. Local protein structures. *Curr. Bioinf.* **2007**, *2*, 165−202.

(20) Joseph, A. P.; De Brevern, A. G. From local structure to a global framework: recognition of protein folds. *J. R. Soc., Interface* **2014**, *11*, 20131147.

(21) de Brevern, A. G.; Bornot, A.; Craveur, P.; Etchebest, C.; Gelly, J.-C. PredyFlexy: flexibility and local structure prediction from sequence. *Nucleic Acids Res.* **2012**, *40*, W317−W322.

(22) Léonard, S.; Joseph, A. P.; Srinivasan, N.; Gelly, J.-C.; De Brevern, A. G. mulPBA: an efficient multiple protein structure alignment method based on a structural alphabet. *J. Biomol. Struct. Dyn.* **2014**, *32*, 661−668.

(23) Guyon, F.; Camproux, A.-C.; Hochez, J.; Tufféry, P. SA-Search: a web tool for protein structure mining based on a Structural Alphabet. *Nucleic Acids Res.* **2004**, *32*, W545−8.

(24) Pandini, A.; Bonati, L.; Fraternali, F.; Kleinjung, J. MinSet: a general approach to derive maximally representative database subsets by using fragment dictionaries and its application to the SCOP database. *Bioinformatics* **2007**, *23*, 515−516.

(25) Regad, L.; Saladin, A.; Maupetit, J.; Geneix, C.; Camproux, A.-C. SA-Mot: a web server for the identification of motifs of interest extracted from protein loops. *Nucleic Acids Res.* **2011**, *39*, W203−W209.

(26) Le, Q.; Pollastri, G.; Koehl, P. Structural Alphabets for Protein Structure Classification: A Comparison Study. *J. Mol. Biol.* **2009**, *387*, 431−50.

(27) Ghouzam, Y.; Postic, G.; de Brevern, A. G.; Gelly, J.-C. Improving protein fold recognition with hybrid profiles combining sequence and structure evolution. *Bioinformatics* **2015**, *31*, 3782−3789.

(28) Thévenet, P.; Rey, J.; Moroy, G.; Tufféry, P. De novo peptide structure prediction: an overview. *Methods Mol. Biol.* **2015**, *1268*, 1−13.

(29) Deschavanne, P.; Tufféry, P. Enhanced protein fold recognition using a structural alphabet. *Proteins: Struct., Funct., Genet.* **2009**, *76*, 129−137.

(30) Pandini, A.; Fornili, A.; Kleinjung, J. Structural alphabets derived from attractors in conformational space. *BMC Bioinf.* **2010**, *11*, 97.

(31) Fornili, A.; Pandini, A.; Lu, H.-C.; Fraternali, F. Specialized Dynamical Properties of Promiscuous Residues Revealed by Simulated Conformational Ensembles. *J. Chem. Theory Comput.* **2013**, *9*, 5127−5147.

(32) Pandini, A.; Fornili, A.; Fraternali, F.; Kleinjung, J. Detection of allosteric signal transmission by information-theoretic analysis of protein dynamics. *FASEB J.* **2012**, *26*, 868−881.

(33) Craveur, P.; Joseph, A. P.; Esque, J.; Narwani, T. J.; Noel, F.; Shinada, N.; Goguet, M.; Leonard, S.; Poulain, P.; Bertrand, O.; Faure, G.; Rebehmed, J.; Ghozlane, A.; Swapna, L. S.; Bhaskara, R. M.; Barnoud, J.; Téletchéa, S.; Jallu, V.; Cerny, J.; Schneider, B.; Etchebest, C.; Srinivasan, N.; Gelly, J.-C.; de Brevern, A. G. Protein flexibility in the light of structural alphabets. *Front. Mol. Biosci.* **2015**, *2*, 20.

(34) Pandini, A.; Fornili, A.; Fraternali, F.; Kleinjung, J. GSATools: analysis of allosteric communication and functional local motions using a Structural Alphabet. *Bioinformatics* **2013**, *29*, 2053−2055.

(35) Muñoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. Folding dynamics and mechanism of beta-hairpin formation. *Nature* **1997**, *390*, 196−199.

(36) Qiu, L.; Pabit, S. A.; Roitberg, A. E.; Hagen, S. J. Smaller and faster: the 20-residue Trp-cage protein folds in 4 micros. *J. Am. Chem. Soc.* **2002**, *124*, 12952−12953.

(37) Best, R. B.; Mittal, J. Microscopic events in β-hairpin folding from alternative unfolded ensembles. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 11087−11092.

(38) Bolhuis, P. G. Kinetic pathways of beta-hairpin (un)folding in explicit solvent. *Biophys. J.* **2005**, *88*, 50−61.

(39) Marinelli, F.; Pietrucci, F.; Laio, A.; Piana, S. A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations. *PLoS Comput. Biol.* **2009**, *5*, e1000452.

(40) Snow, C. D.; Zagrovic, B.; Pande, V. S. The Trp cage: folding kinetics and unfolded state topology via molecular dynamics simulations. *J. Am. Chem. Soc.* **2002**, *124*, 14548−14549.

(41) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How fast-folding proteins fold. *Science* **2011**, *334*, 517−520.

(42) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562−12566.

(43) Isralewitz, B.; Gao, M.; Schulten, K. Steered molecular dynamics and mechanical functions of proteins. *Curr. Opin. Struct. Biol.* **2001**, *11*, 224−230.

(44) Bonet, J.; Planas-Iglesias, J.; Garcia-Garcia, J.; Marin-Lopez, M. A.; Fernandez-Fuentes, N.; Oliva, B. ArchDB 2014: structural classification of loops in proteins. *Nucleic Acids Res.* **2014**, *42*, D315−9.

(45) Leaver-Fay, A.; Tyka, M.; Lewis, S. M.; Lange, O. F.; Thompson, J.; Jacak, R.; Kaufman, K.; Renfrew, P. D.; Smith, C. A.; Sheffler, W.; Davis, I. W.; Cooper, S.; Treuille, A.; Mandell, D. J.; Richter, F.; Ban, Y.-E. A.; Fleishman, S. J.; Corn, J. E.; Kim, D. E.; Lyskov, S.; Berrondo, M.; Mentzer, S.; Popović, Z.; Havranek, J. J.; Karanicolas, J.; Das, R.; Meiler, J.; Kortemme, T.; Gray, J. J.; Kuhlman, B.; Baker, D.; Bradley, P. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **2011**, *487*, 545−574.

(46) Rohl, C. A.; Strauss, C. E. M.; Misura, K. M. S.; Baker, D. Protein structure prediction using Rosetta. *Methods Enzymol.* **2004**, *383*, 66−93.

(47) Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **1997**, *268*, 209−225.

(48) Micheletti, C.; Seno, F.; Maritan, A. Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *Proteins: Struct., Funct., Genet.* **2000**, *40*, 662−674.

(49) de Brevern, A. G.; Etchebest, C.; Hazout, S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins: Struct., Funct., Genet.* **2000**, *41*, 271−287.

(50) Camproux, A. C.; Gautier, R.; Tufféry, P. A hidden markov model derived structural alphabet for proteins. *J. Mol. Biol.* **2004**, *339*, 591−605.

(51) Cossio, P.; Trovato, A.; Pietrucci, F.; Seno, F.; Maritan, A.; Laio, A. Exploring the universe of protein structures beyond the Protein Data Bank. *PLoS Comput. Biol.* **2010**, *6*, e1000957.

(52) Greenfield, N. J. Using circular dichroism spectra to estimate protein secondary structure. *Nat. Protoc.* **2007**, *1*, 2876−2890.

(53) Wang, Z.; Zhao, F.; Peng, J.; Xu, J. Protein 8-class secondary structure prediction using conditional neural fields. *Proteomics* **2011**, *11*, 3786−3792.

(54) Buchan, D. W. A.; Minneci, F.; Nugent, T. C. O.; Bryson, K.; Jones, D. T. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.* **2013**, *41*, W349−57.

(55) Drozdetskiy, A.; Cole, C.; Procter, J.; Barton, G. J. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* **2015**, *43*, W389−W394.

(56) Konermann, L.; Pan, J.; Liu, Y.-H. Hydrogen exchange mass spectrometry for studying protein structure and dynamics. *Chem. Soc. Rev.* **2011**, *40*, 1224−1234.

(57) de Juan, D.; Pazos, F.; Valencia, A. Emerging methods in protein co-evolution. *Nat. Rev. Genet.* **2013**, *14*, 249−261.

(58) Feig, M.; Mirjalili, V. Protein Structure Refinement via Molecular-Dynamics Simulations: What works and what does not? *Proteins: Struct., Funct., Genet.* **2015**, DOI: 10.1002/prot.24871.

(59) Zhang, J.; Liang, Y.; Zhang, Y. Atomic-Level Protein Structure Refinement Using Fragment-Guided Molecular Dynamics Conformation Sampling. *Structure* **2011**, *19*, 1784−1795.

(60) Mirjalili, V.; Noyes, K.; Feig, M. Physics-based protein structure refinement through multiple molecular dynamics trajectories and structure averaging. *Proteins: Struct., Funct., Genet.* **2014**, *82* (S2), 196−207.

(61) Peter, E. K.; Pivkin, I. V.; Shea, J.-E. A kMC-MD method with generalized move-sets for the simulation of folding of α-helical and β-stranded peptides. *J. Chem. Phys.* **2015**, *142*, 144903.

(62) Potestio, R.; Micheletti, C.; Orland, H. Knotted vs. unknotted proteins: evidence of knot-promoting loops. *PLoS Comput. Biol.* **2010**, *6*, e1000864.

(63) Nickson, A. A.; Wensley, B. G.; Clarke, J. Take home lessons from studies of related proteins. *Curr. Opin. Struct. Biol.* **2013**, *23*, 66−74.

(64) Camilloni, C.; Vendruscolo, M. Statistical Mechanics of the Denatured State of a Protein Using Replica-Averaged Metadynamics. *J. Am. Chem. Soc.* **2014**, *136*, 8982−8991.

(65) Granata, D.; Baftizadeh, F.; Habchi, J.; Galvagnion, C.; De Simone, A.; Camilloni, C.; Laio, A.; Vendruscolo, M. The inverted free energy landscape of an intrinsically disordered peptide by simulations and experiments. *Sci. Rep.* **2015**, *5*, 15449.

(66) Piana, S.; Laio, A. A bias-exchange approach to protein folding. *J. Phys. Chem. B* **2007**, *111*, 4553−4559.

(67) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845−854.

(68) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 1950−1958.

(69) Grant, B. J.; Rodrigues, A. P. C.; ElSawy, K. M.; McCammon, J. A.; Caves, L. S. D. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* **2006**, *22*, 2695−2696.