JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Improvement of the Treatment of Loop Structures in the UNRES Force Field by Inclusion of Coupling between Backbone- and Side-Chain-Local Conformational States

Paweł Krupa,[†,‡,⊥] Adam K. Sieradzan,[†,⊥,*] S. Rackovsky,[‡,§,⊥] Maciej Baranowski,[∥] Stanisław Ołdziej,[∥] Harold A. Scheraga,[‡] Adam Liwo,[†] and Cezary Czaplewski[†]

[†]Faculty of Chemistry, University of Gdańsk, Wita Stwosza 63, 80-952 Gdańsk, Poland

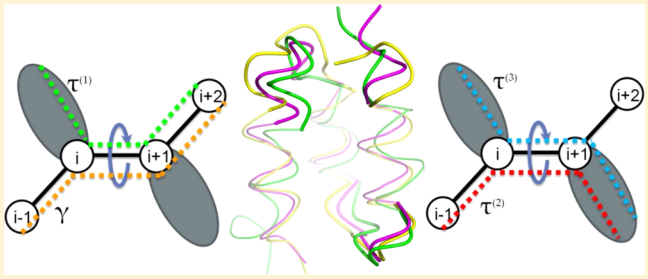
[‡]Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853-1301, United States

[§]Department of Pharmacology and Systems Therapeutics, The Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, New York, New York 10029, United States

[∥]Intercollegiate Faculty of Biotechnology, University of Gdańsk and Medical University of Gdańsk, Kładki 24, 80-922 Gdańsk, Poland

Supporting Information

ABSTRACT: The UNited RESidue (UNRES) coarse-grained model of polypeptide chains, developed in our laboratory, enables us to carry out millisecond-scale molecular-dynamics simulations of large proteins effectively. It performs well in *ab initio* predictions of protein structure, as demonstrated in the last Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP10). However, the resolution of the simulated structure is too coarse, especially in loop regions, which results from insufficient specificity of the model of local interactions. To improve the representation of local interactions, in this work, we introduced new side-chain-backbone correlation potentials, derived from a statistical analysis of loop regions of 4585 proteins. To obtain sufficient statistics, we reduced the set of amino-acid-residue types to five groups, derived in our earlier work on structurally optimized reduced alphabets [Solis, A. D.; Rackovsky, S. *Proteins: Struct., Func., Bioinf.*, **2000**, *38*, 149−164], based on a statistical analysis of the properties of amino-acid structures. The new correlation potentials are expressed as one-dimensional Fourier series in the virtual-bond-dihedral angles involving side-chain centroids. The weight of these new terms was determined by a trial-and-error method, in which Multiplexed Replica Exchange Molecular Dynamics (MREMD) simulations were run on selected test proteins. The best average root-mean-square deviations (RMSDs) of the calculated structures from the experimental structures below the folding-transition temperatures were obtained with the weight of the new side-chain-backbone correlation potentials equal to 0.57. The resulting conformational ensembles were analyzed in detail by using the Weighted Histogram Analysis Method (WHAM) and Ward's minimum-variance clustering. This analysis showed that the RMSDs from the experimental structures dropped by 0.5 Å on average, compared to simulations without the new terms, and the deviation of individual residues in the loop region of the computed structures from their counterparts in the experimental structures (after optimum superposition of the calculated and experimental structure) decreased by up to 8 Å. Consequently, the new terms improve the representation of local structure.



1. INTRODUCTION

Proteins play a central role in biology.[1] Knowledge of protein structure, dynamics, thermodynamics, and kinetics is crucial for understanding the processes in which proteins are involved. However, of over 500 000 protein sequences currently stored in the UniProt database,[2] only about 80 000 structures have been determined and deposited in the Protein Data Bank (PDB).[3] This discrepancy between the number of known protein sequences and the number of known structures demonstrates the need for the development of faster and more accurate methods to predict protein structures.

Currently, the most successful methods for protein−structure prediction are knowledge-based approaches, mostly comparative modeling and the fragment method developed by Baker and colleagues.[4−8] However, for proteins whose sequences are unrelated to those in structural databases, the reliability and accuracy of knowledge-based methods is low.[9] For such targets, physics-based modeling is likely to contribute to the solution of the structure-prediction problem, most likely when combined with knowledge-based approaches. Moreover, using physics-based simulations of protein folding enables us to predict not only protein structure[10] but also thermodynamic

Received: June 11, 2013
Published: September 3, 2013

properties, folding pathways, and kinetic parameters of protein folding.[11,12]

Apart from predicting protein structure, dynamics, and thermodynamics, molecular simulations of proteins are used to predict binding sites and thermodynamics for protein ligands. Moreover, methods based on molecular dynamics (MD) can provide a more dynamic view of receptor−ligand binding than docking[13] and make it possible to investigate conformational changes for both receptor and ligand during binding.[14] These approaches have been used in preliminary *in silico* experiments of drug development, successfully reducing the cost of designing new drugs.[15,16]

Despite advances in computational power, all-atom MD simulations are still too time-consuming to perform routine structure prediction by *ab initio* folding.[17] Using standard computer resources, *ab inito* folding simulations at the all-atom level can be performed only for the fastest-folding small proteins, such as the tryptophan cage (20 residues)[18] or the villin headpiece (35 residues).[19−21] All-atom simulations up to millisecond time scales, for up to 100-residue proteins can be performed using dedicated supercomputer resources such as ANTON,[22] but access to such resources is limited.

Another way to increase the time- and size-scale of simulations is to use coarse-grained models of polypeptide chains. Knowledge- and physics-based coarse-grained models of proteins have been developed since the 1970's for this purpose.[23,24] For the last two decades, we have been developing the physics-based UNited RESidue (UNRES) approach, a coarse-grained model. This coarse-grained force field has been derived rigorously[24−26] as a potential of mean force of polypeptide chains in water, in which secondary degrees of freedom have been averaged out. UNRES provides a 3−4 order of magnitude speedup with respect to all-atom molecular dynamics simulations in explicit water.[27,28] This speed up is achieved both by reducing the number of interactions and by not considering the fast-moving degrees of freedom in the equations of motion explicitly. These degrees of freedom are implicit in the potential of mean force and contribute to effective friction and random forces in the Langevin equations of motion.[27,28] Recently, we extended UNRES to treat the *trans*−*cis* isomerization of peptide groups[29] and D-amino-acid residues.[30] With the second extension, we simulated the folding and assembly pathway of a homotetrameric $\beta\beta\alpha$ (BBAT1) protein[31] and demonstrated the stability of the gramicidin D homodimer,[30] which has a double-helical DNA-like structure, each chain having an alternating D- and L-amino-acid-residue sequence.

UNRES performs well in protein-structure prediction, as assessed in Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiments.[32−35] In the recent CASP10 experiment, two predictions made with the use of UNRES were identified by the assessors as the best for the new-fold targets T0663 and T0740. We noticed, however, that, while UNRES performs well in finding the overall fold of protein fragments from 50 to 200 amino-acid residues, the accuracy is only 5−8 Å on average. We found that this feature is due mostly to imperfect modeling of loop regions and other regions with weakly defined secondary structure. This diminished accuracy (compared to the accuracy of representations of regions with well-defined secondary structure) arises from the nonspecificity of local interactions, in which only three basic residue types: glycine, proline, and alanine (which represents all nonglycine and nonproline residue types) are considered.[25,36,37] Therefore, in this work, we focused on the improvement of the specificity of local interactions. We introduced new terms that account for the coupling between backbone- and side-chain conformational states, which are represented as residue-pair-specific torsional potentials in the virtual-bond-dihedral angles involving side-chain centers. In this preliminary work, the new potentials were derived as statistical potentials. Introduction of the physics-based counterparts of these potentials will be the subject of our future work.

2. METHODS

2.1. UNRES Representation of Polypeptide Chain. In the UNRES model[11,25,26,29,36−44] a polypeptide chain is represented as a sequence of α-carbon (C^α) atoms with attached united side chains (SC's) and united peptide (p's) groups positioned halfway between two consecutive C^α atoms. Only the united side chains and united peptide groups act as interaction sites, while the C^α atoms assist only in the definition of geometry (Figure 1). The effective energy function is defined as the restricted free energy (RFE) or the potential of mean force (PMF) of the chain constrained to a given coarse-grained conformation along with the surrounding solvent.[25,26,39] This

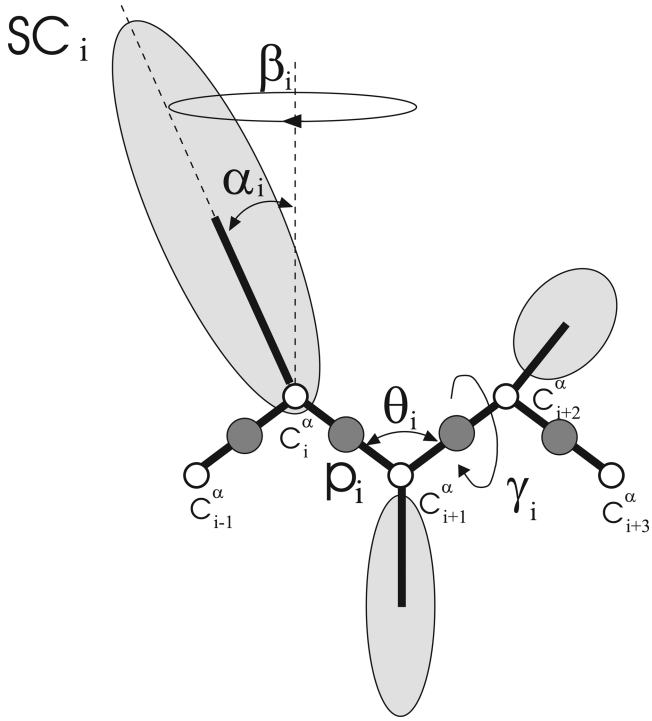


Figure 1. The UNRES model of polypeptide chains. The interaction sites are peptide-group centers (p), and side-chain centers (SC) attached to the corresponding α-carbons with different $C^\alpha\cdots$SC bond lengths, d_{SC}. The peptide groups are represented as dark gray circles and the side chains are represented as light gray ellipsoids of different size. The α-carbon atoms are represented by small open circles. The geometry of the chain can be described either by the virtual-bond vectors dC_i (from C_i^α to C_{i+1}^α), $i = 1,2,...,n-1$, and dX_i (from C_i^α to SC_i), $i = 2,...,n-1$, represented by thick lines, where n is the number of residues, or in terms of virtual-bond lengths, backbone virtual-bond angles $\theta_i, i = 1,2,....,n-2$, backbone virtual-bond-dihedral angles $\gamma_i, i = 1,2,....,n-3$, and the angles α_i and $\beta_i, i = 2,3,....,n-1$ that describe the location of a side chain with respect to the coordinate frame defined by C_{i-1}^α, C_i^α, and C_i^α, C_{i+1}^α.

dx.doi.org/10.1021/ct4004977 | J. Chem. Theory Comput. 2013, 9, 4620−4632

effective energy function, including the new terms introduced in this work, is expressed by eq 1.

$$U = w_{SC}\sum_{i<j} U_{SC_iSC_j} + w_{SCp}\sum_{i\neq j} U_{SC_ip_j} + w_{pp}^{VDW}\sum_{i<j-1} U_{p_ip_j}^{VDW} + w_{pp}^{el} f_2(T)\sum_{i<j-1} U_{p_ip_j}^{el} + w_{tor} f_2(T)\sum_i U_{tor}(\gamma_i) + w_{tord} f_3(T)\sum_i U_{tord}(\gamma_i, \gamma_{i+1}) + w_b\sum_i U_b(\theta_i) + w_{rot}\sum_i U_{rot}(\alpha_{SC_i}, \beta_{SC_i}) + w_{bond}\sum_i U_{bond}(d_i) + w_{corr}^{(3)} f_3(T) U_{corr}^{(3)} + w_{corr}^{(4)} f_4(T) U_{corr}^{(4)} + w_{corr}^{(5)} f_5(T) U_{corr}^{(5)} + w_{corr}^{(6)} f_6(T) U_{corr}^{(6)} + w_{turn}^{(3)} f_3(T) U_{turn}^{(3)} + w_{turn}^{(4)} f_4(T) U_{turn}^{(4)} + w_{turn}^{(6)} f_6(T) U_{turn}^{(6)} + w_{SC-corr} f_2(T)\sum_{m=1}^{3}\sum_i U_{SC-corr}(\tau_i^{(m)}) \quad (1)$$

where the U's are energy terms, θ_i is the backbone virtual-bond angle, γ_i is the backbone virtual-bond-dihedral angle, α_i and β_i are the angles defining the location of the center of the united side chain of residue i (Figure 1), and d_i is the length of the ith virtual bond, which is either a $C^\alpha \cdots C^\alpha$ virtual bond or $C^\alpha \cdots SC$ virtual bond. Each energy term is multiplied by an appropriate weight, w_x, and the terms corresponding to factors of order higher than 1 are additionally multiplied by the respective temperature factors which were introduced in our work[11] and which reflect the dependence of the first generalized-cumulant term in those factors on temperature, as discussed in refs 11 and 45. The factors f_n are defined by eq 2.

$$f_n(T) = \frac{\ln[\exp(1) + \exp(-1)]}{\ln\{\exp[(T/T_o)^{n-1}] + \exp[-(T/T_o)^{n-1}]\}} \quad (2)$$

where T_o = 300 K.

The term $U_{SC_iSC_j}$ represents the mean free energy of the hydrophobic (hydrophilic) interactions between the side chains, which implicitly contains the contributions from the interactions of the side chain with the solvent. The term $U_{SC_ip_j}$ denotes the excluded-volume potential of the side-chain–peptide-group interactions. The peptide-group interaction potential is split into two parts: the Lennard-Jones interaction energy between peptide-group centers ($U_{p_ip_j}^{VDW}$) and the average electrostatic energy between peptide-group dipoles ($U_{p_ip_j}^{el}$); the second of these terms accounts for the tendency to form backbone hydrogen bonds between peptide groups p_i and p_j. The terms U_{tor}, U_{tord}, U_b, U_{rot}, and U_{bond} are the virtual-bond-dihedral angle torsional terms, virtual-bond dihedral angle double-torsional terms, virtual-bond angle bending terms, side-chain rotamer, and virtual-bond-deformation terms; these terms account for the local properties of the polypeptide chain. The terms $U_{corr}^{(m)}$ represent correlation or multibody contributions from the coupling between backbone-local and backbone-electrostatic interactions, and the terms $U_{turn}^{(m)}$ are correlation contributions involving m consecutive peptide groups; they are, therefore, termed turn contributions. The multibody terms are indispensable for reproduction of regular α-helical and β-sheet structures.[25,39,46] The $U_{SC-corr}$ terms are newly introduced side-chain backbone correlation potentials; they are expressed as Fourier series in the $SC\cdots C^\alpha\cdots C^\alpha\cdots C^\alpha$ ($\tau^{(1)}$), $C^\alpha\cdots C^\alpha\cdots C^\alpha\cdots SC$ ($\tau^{(2)}$), and $SC\cdots C^\alpha\cdots C^\alpha\cdots SC$ ($\tau^{(3)}$) virtual-bond-dihedral angles (Figure 2).

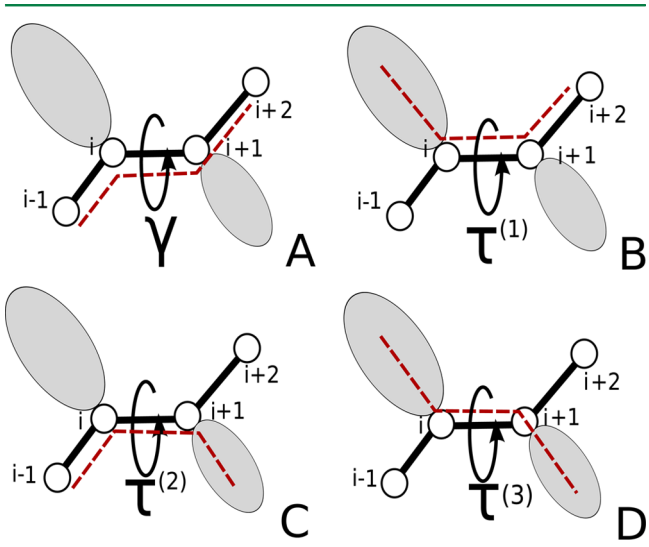


Figure 2. Illustration of backbone torsional angle γ (A) and side-chain backbone torsional angles: $\tau^{(1)}$ (B), $\tau^{(2)}$ (C), $\tau^{(3)}$ (D).

The energy-term weights were determined by force-field calibration to reproduce the structure and folding thermodynamics of selected training proteins.[11] As a training protein, the GA (protein G-related albumin-binding) module (an α protein; PDB code: 1GAB)[47] was used to optimize weights and side-chain–side-chain interaction parameters.[11]

2.2. Determination of Potentials of Mean Force from Statistical Analysis. The potentials for the virtual-bond-dihedral angles $\tau^{(1)}$, $\tau^{(2)}$, and $\tau^{(3)}$ that involve side-chain centers were determined from statistical analysis of loop fragments and other fragments with no defined secondary structure of 4585 proteins. These potentials are denoted as $U_{SC-corr;XY}(\tau^{(m)})$. The potentials for the backbone virtual-bond-dihedral angle γ have also been determined using the same method, for reference. These reference potentials are denoted as $U_{stor}(\gamma)$. The proteins where selected from the entire PDB database. The proteins with loop/unstructured regions shorter than three residues or those which contained unstructured regions only at the N- or at the C-terminus were not considered. Repeated loop fragments were removed; so, the database consisted only of unique sequences. The database consisted of 267 664 fragments for the $\tau^{(1)}$ and $\tau^{(2)}$ angles, 334 431 fragments for the $\tau^{(3)}$ angles, and 214 614 fragments for the γ (backbone virtual-bond-dihedral) angles. Selection of regions without well-defined secondary structure was motivated by the fact that these regions contain few long-range interactions that could have influenced the determined potentials. For this reason, it can be assumed that there is no need to subtract the contributions due to other interactions from the resulting PMF's.

In order to improve the statistical properties of the database, we relied on results from our previous work on the structurally optimized reduced amino-acid alphabet.[48] The 20 amino acids were divided into groups, which are listed in Table 1. The statistical potentials are denoted in eqs 4 and 5 as $U_{stor,XY}(\gamma)$ and $U_{SC-corr,XY}(\tau^{(m)})$ (where m = 1,2,3, depending on the type of virtual-bond-dihedral angles under consideration) and X and Y denote the residues at the first and second position on the central virtual bond. Each residue (X or Y) belongs to one of

dx.doi.org/10.1021/ct4004977 | J. Chem. Theory Comput. 2013, 9, 4620–4632

Table 1. Reduced Set of Amino-Acid-Residue Types[48]

type no.	amino acids type
1	G
2	D,N
3	A,E,H,K,Q,R,S,T
4	C,F,I,L,M,V,W,Y
5	P

the five types listed in Table 1. This gives 25 types of $U_{stor}(\gamma)$ potentials, 20 types of $U_{SC-corr,XY}(\tau^{(1)})$ and $U_{SC-corr,XY}(\tau^{(2)})$ potentials, and 16 types of $U_{SC-corr,XY}(\tau^{(3)})$ potentials. Glycine, which is one of the five amino-acid types, does not have a side chain, and therefore, the $U_{SC-corr,GlyY}(\tau^{(1)})$, $U_{SC-corr,XGly}(\tau^{(2)})$, $U_{SC-corr,GlyY}(\tau^{(3)})$, $U_{SC-corr,XGly}(\tau^{(3)})$, and $U_{SC-corr,GlyGly}(\tau^{(3)})$ are not defined.

For each type of potential, histograms of the appropriate virtual-bond-dihedral angle (γ or $\tau^{(m)}$,m = 1,2,3) were constructed. The data for the 4585 proteins were binned in 10° intervals. The dimensionless potentials of mean force were calculated from eq 3:

$$f_{XY}(\tau_i^{(m)}) = \ln(N_{XY,\max}^{(m)}) - \ln(N_{XY,i}^{(m)}) \tag{3}$$

where $N_{XY,i}^{(m)}$ is the number of counts of the mth type angle (γ, $\tau^{(1)}$, $\tau^{(2)}$ or $\tau^{(3)}$) for residue types X and Y in the ith bin, and $N_{XY,\max}^{(m)}$ is the largest number of counts over all bins, for a given type of angle and given types of residues. The dimensionless PMFs, in γ and $\tau^{(m)}$,m = 1,2,3, were subsequently fitted to the one-dimensional Fourier series (eqs 4 and 5) by the linear least-squares method.

$$U_{stor;XY}(\gamma) = a_o + \sum_{n=1}^{6} a_n \cos(n\gamma) + b_n \sin(n\gamma) \tag{4}$$

$$U_{SC-corr;XY}(\tau^{(m)}) = a_{mo} + \sum_{n=1}^{6} a_{mn} \cos(n\tau^{(m)}) + b_{mn} \sin(n\tau^{(m)}) \tag{5}$$

where X and Y are the five types of residues collected in Table 1; a_n, n = 1,2,...6; b_n, n = 1,2,...6; a_{mn}, m = 1,2,3 and n = 1,2,...6; b_{mn}, m = 1,2,3 and n = 1,2,...6 are coefficients of the Fourier expansions of $U_{stor;XY}(\gamma)$ and $U_{SC-corr;XY}(\tau^{(m)})$, respectively.

To estimate the added value of the coupling between the side-chain and backbone-local interactions [i.e., the part of the PMF not already included in the regular torsional contributions $U_{tor}(\gamma)$], the statistical side-chain correlation potentials involving side chains were compared with that in the backbone virtual-bond-dihedral angle γ. To accomplish this, we fitted each potential in the $\tau^{(i)}$ angle to that in the γ angle, with amplitude and phase shift as adjustable parameters. Fitting was accomplished by minimizing the target function $F_{XY}^{(m)}(A_o, \phi)$ given by eq 6.

$$F_{XY}^{(m)}(A_o, \phi) = \frac{1}{2\pi} \int_{-\pi}^{\pi} [U_{stor;XY}(\gamma_i) - A_o \times U_{SC-corr;XY}(\gamma + \phi^{(m)})]^2 d\gamma, \quad m = 1, 2, 3 \tag{6}$$

where $U_{stor}(\gamma)$ is the statistical backbone torsional potential defined by eq 4, $U_{SC-corr}$ is the statistical side-chain backbone correlation torsional potential under consideration, defined by eq 5, ϕ is the phase angle, and A_o is the amplitude-scaling factor. The target function was evaluated by a numerical quadrature with integration step dγ = 1° (π/180).

The quality of fitting can be assessed by calculating the RMSD between the $U_{SC-corr}$ and U_{stor} potentials (eq 7), and the extent to which $U_{SC-corr;XY}(m)$ and $U_{stor;XY}$ are related to each other can be assessed by calculating the correlation coefficients, as expressed by eq 8.

$$\mathrm{RMSD}_{XY}^{(m)} = \sqrt{\langle [U_{stor;XY}(\gamma) - A_o \times U_{SC-corr;XY}(\gamma + \phi^{(m)})]^2 \rangle} = \sqrt{F_{XY}^{(m)}} \tag{7}$$

$$r_{XY}^{2(m)} = \frac{\mathrm{cov}[U_{stor;XY}(\gamma) U_{SC\text{-}corr;XY}(\gamma + \phi^{(m)})]^2}{\mathrm{var}[U_{stor;XY}(\gamma)]\ \mathrm{var}[U_{SC\text{-}corr;XY}(\tau^{(m)})]} \tag{8}$$

where

$$\mathrm{cov}[U_{stor;XY}(\gamma) U_{SC-corr;XY}(\gamma + \phi^{(m)})] = \frac{1}{2\pi} \int_{-\pi}^{\pi} U_{stor;XY}(\gamma) U_{SC-corr;XY}(\gamma + \phi^{(m)}) d\gamma - \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} U_{stor;XY}(\gamma) d\gamma \int_{-\pi}^{\pi} U_{SC-corr;XY}(\tau^{(m)}) d\tau^{(m)} \tag{9}$$

$$\mathrm{var}[U_{stor;XY}(\gamma)] = \frac{1}{2\pi} \int_{-\pi}^{\pi} U_{stor;XY}(\gamma)^2 d\gamma - \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} U_{stor;XY}(\gamma) d\gamma \right)^2 \tag{10}$$

$$\mathrm{var}[U_{SC-corr;XY}(\tau^{(m)})] = \frac{1}{2\pi} \int_{-\pi}^{\pi} U_{SC-corr;XY}(\tau^{(m)})^2 d\gamma - \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} U_{SC-corr;XY}(\tau^{(m)}) d\gamma \right)^2 \tag{11}$$

are the covariance of $U_{stor;XY}(\gamma)$ and $U_{SC-corr;XY}(\tau^{(m)})$, and the variances of $U_{stor;XY}(\gamma)$ and $U_{SC-corr;XY}(\tau^{(m)})$, respectively, and $\langle ... \rangle$ denotes the average of a quantity.

The square of the correlation coefficient, r^2 of eq 8, is the fraction of the explained variance. When applied to comparing the $U_{SC-corr;XY}$ and the various $U_{SC-corr;XY}(\tau^{(m)}+\phi)$ profiles, the values of r^2 show the extent to which the variation of the corresponding dihedral angles $\tau^{(m)}$ involving side chains can be explained by the variation of the backbone dihedral angle γ.

2.3. Testing UNRES with the New Potentials. The UNRES force field with the new $U_{SC-corr}$ potentials was tested with a set of small proteins (37−76 residues) used by us previously[11] for assessing the performance of previous versions of the UNRES force field: the recombinant B domain (FB) of staphylococcal protein A (PDB code: 1BDD)[49](α-helical structure), apo calbindin D9k from *Bos taurus* (PDB code: 1CLB)[50](α-helical structure), the LysM domain from *E. coli* (PDB code: 1E0G)[51] (α + β structure), the Fbp28 WW domain from *Mus musculus* (PDB code: 1E0L)[28,52] (β structure), the GA module (PDB code: 1GAB)[47] (α-helical structure), the DFF-C domain of DFF45/ICAD from *Homo sapiens* (PDB code: 1KOY)[53](α-helical structure), the POU-specific domain from *Homo sapiens* (PDB code: 1POU)[54] (α-helical structure), and the purine repressor (PurR) DNA-binding domain from *E. coli* (PDB code: 1PRU)[55] (α-helical

dx.doi.org/10.1021/ct4004977 | J. Chem. Theory Comput. 2013, 9, 4620−4632

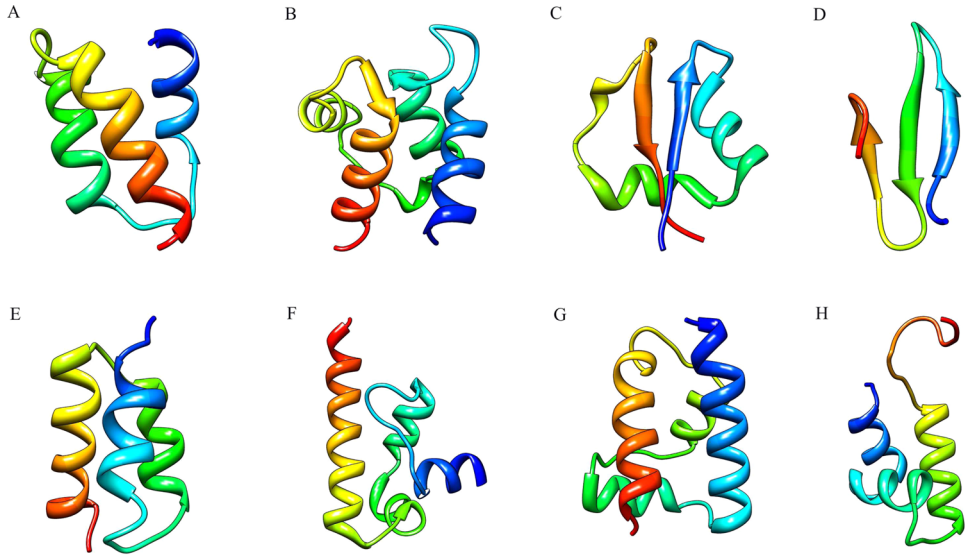


Figure 3. Cartoon representation of truncated experimental structures of tested proteins: (A) 1BDD, (B) 1CLB, (C) 1E0G, (D) 1E0L, (E) 1GAB, (F) 1KOY, (G) 1POU, (H) 1PRU. The chains are colored from blue (N-terminus) to red (C-terminus).

structure). The highest sequence identity between the test proteins was 24.32%, as determined by the ClustalW2 program,[56] with average identity of 9.69%. For completeness, the test proteins were present in the database from which the potentials were derived; however, because these eight proteins constituted less than 0.2% of the database and only loop and other unstructured fragments were taken (about 24% for 1KOY which has the largest loop regions out of the test proteins), this does not introduce any significant bias toward the test proteins. Weakly defined N- and C-terminal fragments of the test proteins have been truncated from the structures determined by NMR. The truncated experimental structures are shown in Figure 3.

For each test protein, Multiplexed Replica Exchange Molecular Dynamics (MREMD)[57,58] simulations were performed. MREMD is a generalization of the regular Replica Exchange Molecular Dynamics (REMD) method,[59−61] in which not one but several trajectories are run at a given temperature, which increases the efficiency of sampling.[58] The method is intrinsically paralellizable, as implemented in the UNRES model, and it scales up to 75% with over 4000 CPUs.[26] A total of 40 trajectories were run at 20 temperatures (2 trajectories per temperature), the temperatures being T = 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300, 310, 320, 330, 340, 350, 360, 370, 385, and 400 K. Fifty million steps with a length of 0.1 mtu (4.89 fs)[62] were run for every test protein, which is equivalent to about 0.49 ms real physical time (the time scale in UNRES MD simulations is distorted because of the omission of fast-moving degrees of freedom[27,28]). The Berendsen thermostat[63] with the coupling parameter τ = 48.9 fs was used to maintain constant temperature. Despite the fact that the Berendsen thermostat does not provide canonical sampling,[64] it performs well in MREMD simulations with UNRES.[58] For each protein, the simulations were started from the extended structure. The variable time step (VTS) algorithm[27] was used to integrate the equations of motion.

MREMD simulations were carried out with the following three values of the weights of the $U_{SC-corr}$ terms: $w_{SC-corr}$ = 0.0 (no contribution from the correlation terms), $w_{SC-corr}$ = 0.57, and $w_{SC-corr}$ = 1.00. The value $w_{SC-corr}$ = 0.57 was selected to correspond closely to the "average" value of RT (where R is the universal gas constant and T is the absolute temperature) corresponding to the conversion of the dimensionless potentials of mean force derived from statistical analysis (eq 3) to kcal/mol. The typical temperatures of the X-ray and NMR experiments to determine protein structure range from 277 K (RT = 0.60) to 303 K (RT = 0.55), for which an average value is 0.57. The weights of all other energy terms remained as in the force field calibrated with the 1GAB protein.[11]

3. RESULTS AND DISCUSSION

3.1. Potentials of Mean Force from Statistical Analysis. The plots of 21 sample statistical potentials are shown in Figure 4, while those of the remaining 35 potentials are shown in Figure S1 of the Supporting Information. The Fourier coefficients of the analytical fits to the potentials (eq 5) are collected in Tables S1, S2, and S3 of the Supporting Information.

To assess the added value of the new potentials preliminarily [with respect to the regular $U_{tor}(\gamma)$ potentials of eq 1], we determined how much the $U_{SC-corr,XY}(\tau^{(m)})$, m = 1,2,3, potentials are correlated with the statistical backbone $U_{stor,XY}(\gamma)$ potentials determined in this work. To accomplish this, we minimized $F_{XY}^{(m)}$ defined by eq 6 and then calculated the correlation coefficients, $r_{XY}^{2(m)}$, eq 8, between the $U_{SC-corr,XY}(\tau^{(m)})$ profiles and the $U_{stor,XY}(\gamma)$ profiles, after applying the phase-shift angle ϕ and scaling coefficients A_o, obtained from eq 6 by least-squares fitting of $U_{SC-corr,XY}(\tau^{(m)})$ to $U_{stor,XY}(\gamma)$. The squares of the correlation coefficients (the extent of the explained variances) are displayed graphically in Figure 5. The values of the RMSDs between the $U_{SC-corr,XY}(\tau^{(m)})$ and $U_{stor,XY}(\gamma)$ profiles, squares of the correlation coefficients, the scaling factors A_o, and the phase angles ϕ of eq 6 are summarized in Table S4 of the Supporting Information.

It can be seen that most of the $U_{SC-corr,XY}(\tau^{(2)})$ potentials correlate very well with the corresponding $U_{stor,XY}(\gamma)$ potentials (Figure 5B). The fraction of the explained variance is about 0.9 or greater for most of the pairs of amino-acid types (Table 1) except for the pairs consisting of type 2 (asparagine and aspartic acid) with type 1 (glycine), type 2, and type 5 (proline),

dx.doi.org/10.1021/ct4004977 | J. Chem. Theory Comput. 2013, 9, 4620−4632

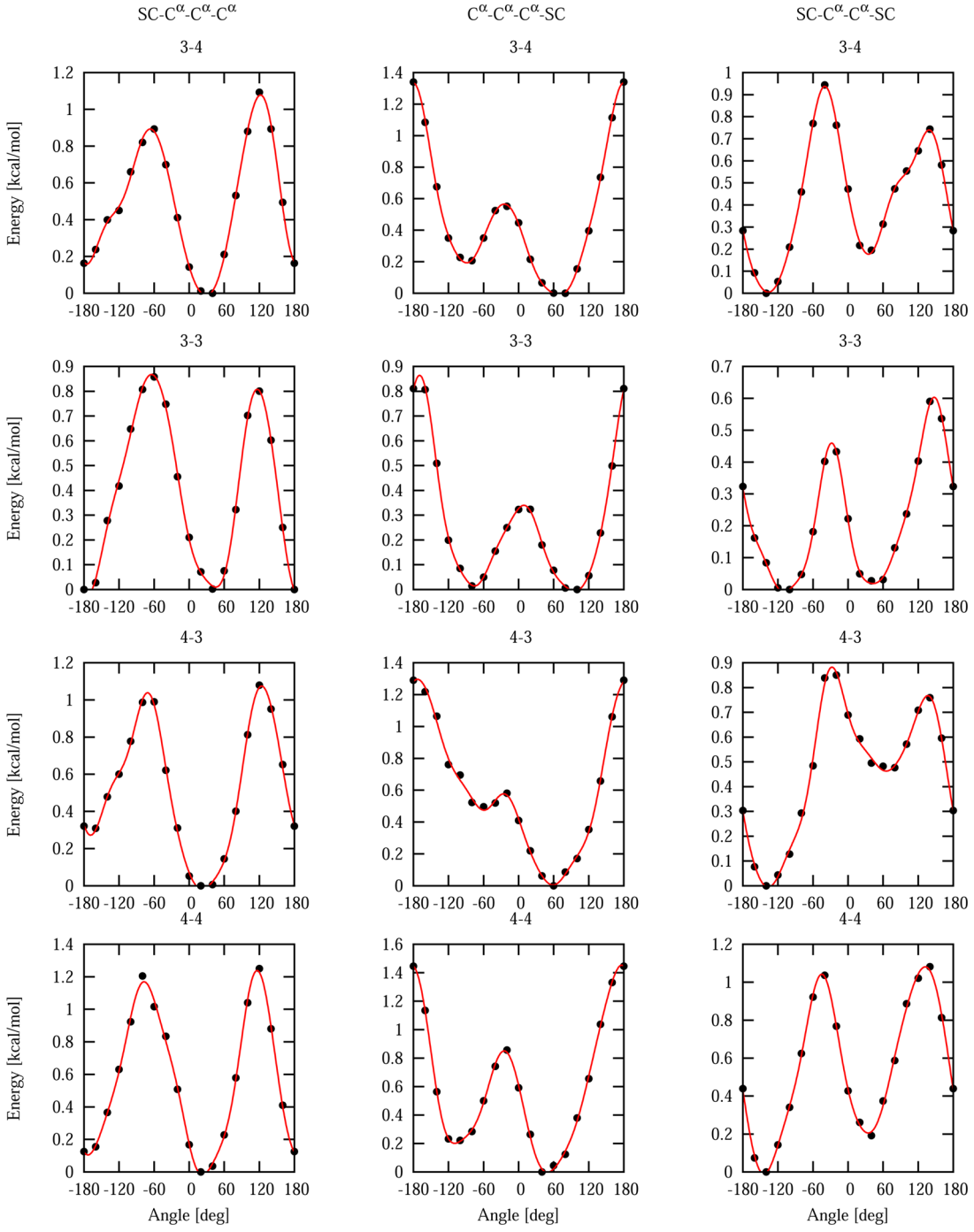


Figure 4. Plots of the 21 sample side-chain backbone correlation potentials $U_{XY}(\tau^{(m)})$, m = 1,2,3, where X and Y each denote one residue type of the reduced amino-acid alphabet (see Table 1), calculated from the statistical analysis of loop structures. The values of X and Y are stated above each plot. Black circles represent the values of the dimensionless PMFs calculated from histograms (eq 3) and red lines represent one-dimension Fourier series fits (eq 5) to the PMF values.

respectively (green or yellow-green squares in the bottom row of the graph in Figure 5B). For example, $r_{44}^{2(2)}$ = 0.953, $r_{45}^{2(2)}$ = 0.919, and $r_{55}^{2(2)}$ = 0.861, contrary to $r_{22}^{2(2)}$ = 0.557 and $r_{52}^{2(2)}$ = 0.587 (Table S4 of the Supporting Information). The correlation is also weaker (with the square of the correlation coefficient of about 0.8) for pairs involving type 2 with type 3 (small and charged residues) and type 5 with type 3 (the deep-sky-blue squares in the second row of Figure 5B).

The $U_{SC-corr,XY}(\tau^{(1)})$ (Figure 5A) and $U_{SC-corr,XY}(\tau^{(3)})$ profiles (Figure 5C) correlate weaker with the $U_{stor;XY}(\gamma)$. Because $U_{SC-corr;XY}(\tau^{(1)})$ involves the side chain of residue X and $U_{SC-corr;XY}(\tau^{(2)})$ involves that of residue Y (cf. the definitions of the angles τ in Figure 2), it can be concluded that the local conformational states of the side chain of the first residue are less dependent on the local backbone states compared to those of the side chain of the second residue in a pair (i and i+1 of Figure 2A). This is understandable because the second side chain in a pair can overlap with the peptide group on the axis defining the virtual-bond-dihedral angle γ to a greater extent than the first one; this can be judged from the conformational-

dx.doi.org/10.1021/ct4004977 | J. Chem. Theory Comput. 2013, 9, 4620−4632

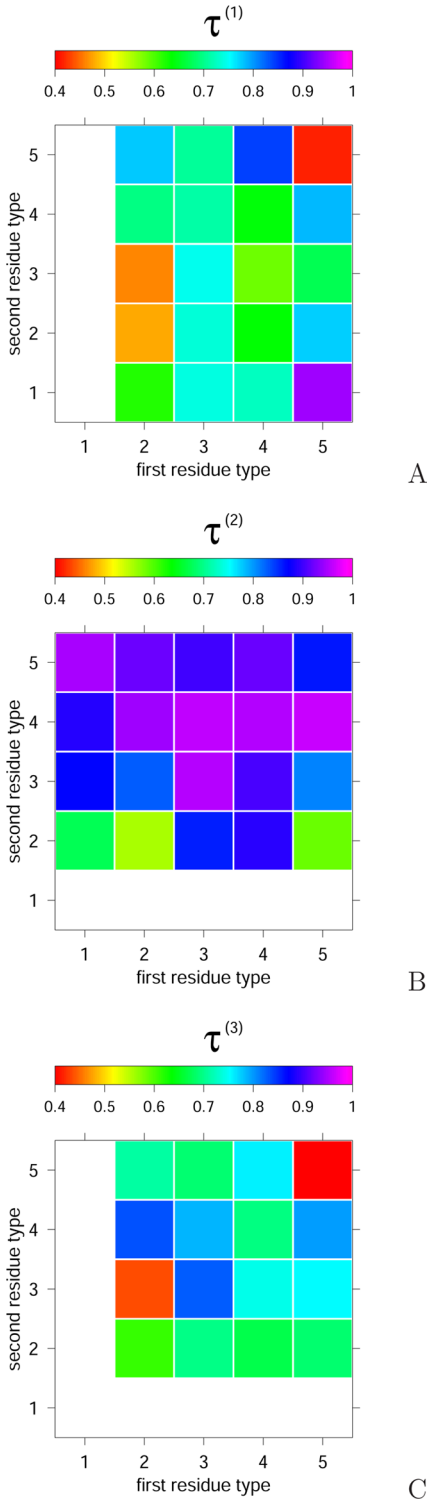


Figure 5. Color maps of the square of the correlation coefficients (r^2_{XY} of eq 8), between $U_{\mathrm{stor};XY}(\gamma)$ and $U_{\mathrm{SC-corr};XY}(\tau^{(m)})$ of eq 6, where, in eq 6, $\tau^{(m)} = \gamma + \phi^{(m)}$ for all pairs of amino-acid types (Table 1). (A) $\tau^{(1)}$, (B) $\tau^{(2)}$, (C) $\tau^{(3)}$. The color scales are on top of each panel. Blank areas correspond to pairs of types for which the respective $\tau^{(m)}$ torsional angle is undefined (if Gly is the first residue of $\tau^{(1)}$, the second residue of $\tau^{(2)}$, or one of the residues of $\tau^{(3)}$).

energy maps of terminally blocked amino-acid residues,[65] which imply a greater dependence on the backbone angle φ (which is defined by the C′−N−C$^{\alpha}$−C′ atoms and thus pertains to the first peptide group of the residue) than on the backbone angle ψ (which is defined by the N−C$^{\alpha}$−C′−N atoms and thus pertains to the second peptide group). It should be noted that the regular torsional potentials, $U_{\mathrm{tor};XY}(\gamma)$, also depend more on the type of the second residue (Y) than on that of the first one (X).[30,37,66,67]

The relatively weak correlation of the $U_{\mathrm{SC-corr},XY}(\tau^{(2)})$ potentials with the $U_{\mathrm{stor};XY}(\gamma)$ potentials when the second residue is Asp or Asn (type 2), and the first one is polar or charged (type 3), can be explained in terms of the interactions between the polar or charged groups and the backbone; these interactions will distort the side-chain centers from their "average" positions with respect to the backbone. The smallest fraction of the explained variance (0.59 or less) is observed for pairs of residues, both of which are of type 2 (Asp and Asn) or both are of type 5 (Pro) and a type 2 with type 3 for $U_{\mathrm{SC-corr},XY}(\tau^{(1)})$ (Figure 5A). Additionally, the correlation is relatively weak when the first residue in an $U_{\mathrm{SC-corr},XY}(\tau^{(2)})$ profile is of type 2 (Figure 5B) and the second one is of type 1 or 3.

For $U_{\mathrm{SC-corr},XY}(\tau^{(3)})$ the correlation is the weakest for the 2−3 and the 5−5 pairs (Figure 5C).

The phase shifts $\phi^{(m)}$ of eq 6 provide some insight into the relationship between the $U_{\mathrm{SC-corr},XY}(\tau^{(m)})$ and $U_{\mathrm{stor};XY}(\gamma)$ potentials. Their values are visualized as bar diagrams in Figure 6. It can be seen that the general trend of the phase shift is from

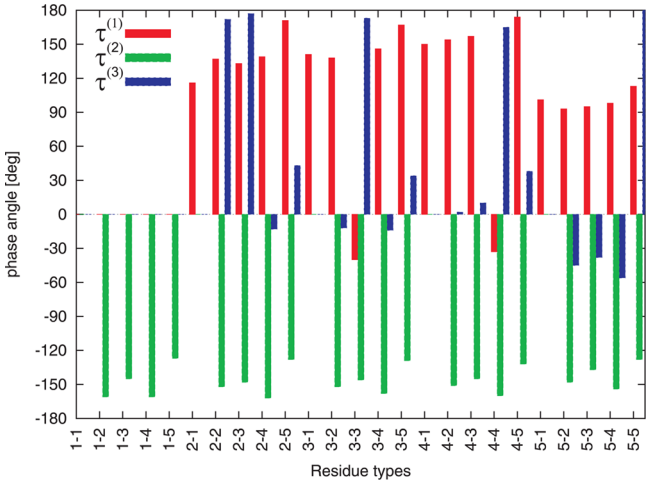


Figure 6. Bar diagram of the phase-shift angles ϕ of eq 6. Residue types of Table 1 (separated by hyphen) are on the abscissa. Colors correspond to torsional angle types: $\tau^{(1)}$ (red), $\tau^{(2)}$ (green), and $\tau^{(3)}$ (blue). It should be noted that the $\tau^{(m)}$ angles involving the glycine residue can be undefined (see the legend of Figure 5).

about −150° to about −120° for $U_{\mathrm{SC-corr},XY}(\tau^{(2)})$ and from about 90° to about 170° for $U_{\mathrm{SC-corr},XY}(\tau^{(1)})$ (for proline). These values of $\phi^{(m)}$ are roughly the values resulting from the geometry of L-amino-acid residues. For $U_{\mathrm{SC-corr},XY}(\tau^{(3)})$, the phase shift should be approximately zero, because the projections of the side chains on the plane perpendicular to the C$^{\alpha}$···C$^{\alpha}$ virtual-bond axis are rotated by approximately the same angle with respect to the projections of the terminal C$^{\alpha}$···C$^{\alpha}$ backbone virtual-bond axes, unless there is substantial coupling between the backbone-local and side-chain-local conformational states. However, this is not always the case. In particular, for type 2, the phase angle is often 180°, which suggests that the polar groups of aspartic acid and asparagine are strongly involved in interactions with the backbone or the

dx.doi.org/10.1021/ct4004977 | J. Chem. Theory Comput. 2013, 9, 4620−4632

neighboring side chains (the blue bars corresponding to the phase angle for $\tau^{(3)}$ in Figure 6).

The potentials described in this section were implemented in the UNRES package available free of charge from www.unres.pl.

3.2. Performance of the New Potentials in *Ab Initio* Simulations of Protein Structure. As mentioned in the Methods section, we ran MREMD simulations of the benchmark proteins with $w_{SC-corr}$ = 0.0, 0.57, and 1.0. For each run, the heat-capacity curves were calculated during the progress of simulations and monitored for convergence, as described in our earlier work.[11,58,68] Simulations were terminated when the heat-capacity curves calculated from at least two consecutive simulation periods superposed on each other with the maximum difference not exceeding 0.05 kcal/(mol·K). Sample convergence plots for 1BDD are shown in Figure 7. The fastest convergence of the heat capacity curves was achieved for $w_{SC-corr}$ = 0 (Figure 7A) and the slowest for $w_{SC-corr}$ = 1 (Figure 7C). This feature could result from the fact that the new $U_{SC-corr}$ potentials introduce more local conformational states, resulting in the appearance of additional local minima in the energy surface. We also noted that introduction of the new potentials results in additional heat-capacity peaks; this is understandable because UNRES with new terms has not yet been optimized to reproduce thermodynamic properties of proteins. This optimization will be carried out with the maximum-likelihood method.

To make the analysis of the calculated conformational ensembles and selection of candidate conformations close to that used in the CASP exercises, the set of conformations was always divided into five clusters, by using Ward's minimum-variance method.[69] The fractions of the conformations that belong to the respective clusters were calculated by summing the probabilities of their member conformations calculated with the use of WHAM,[70] as described in our earlier work.[11] These fractions are defined by eq 12.

$$f_i(T_a) = \sum_{k \in I} w_k(T_a) \tag{12}$$

with

$$w_k(T_a) = \frac{\exp(\omega_k - U_k/RT_a)}{\sum_k \exp(\omega_k - U_k/RT_a)} \tag{13}$$

where $f_i(T_a)$ is the fraction of the conformations of the ith clusters at absolute temperature T_a selected to carry out the analysis, I is the set conformations that belong to the ith cluster, $w_k(T_a)$ is the weight of the kth conformation obtained in the MREMD projected onto the absolute temperature T_a, U_k is the UNRES energy of the kth conformation, ω_k is the logarithm of the weighting factor of the kth conformation determined by WHAM,[11] and R is the universal gas constant. The summation in the denominator of eq 13 extends to all conformations obtained in a given MREMD run. For each protein, T_a was chosen as a temperature below that of the major heat-capacity peak; the values of T_a are summarized in Table 2. Further, for each protein, we defined the ensemble of native-like conformations as the sum of clusters for which the cluster-average RMSD was less than or equal to 0.1 Å per amino-acid residue. We also defined the fractions of conformations that belong to the clusters with the lowest root-mean-square deviation from the experimental structures for all proteins studied except 1E0L (for which an α-helical hairpin instead of a three-stranded antiparallel β-sheet was obtained in all simulations). These fractions are collected in Table 2.

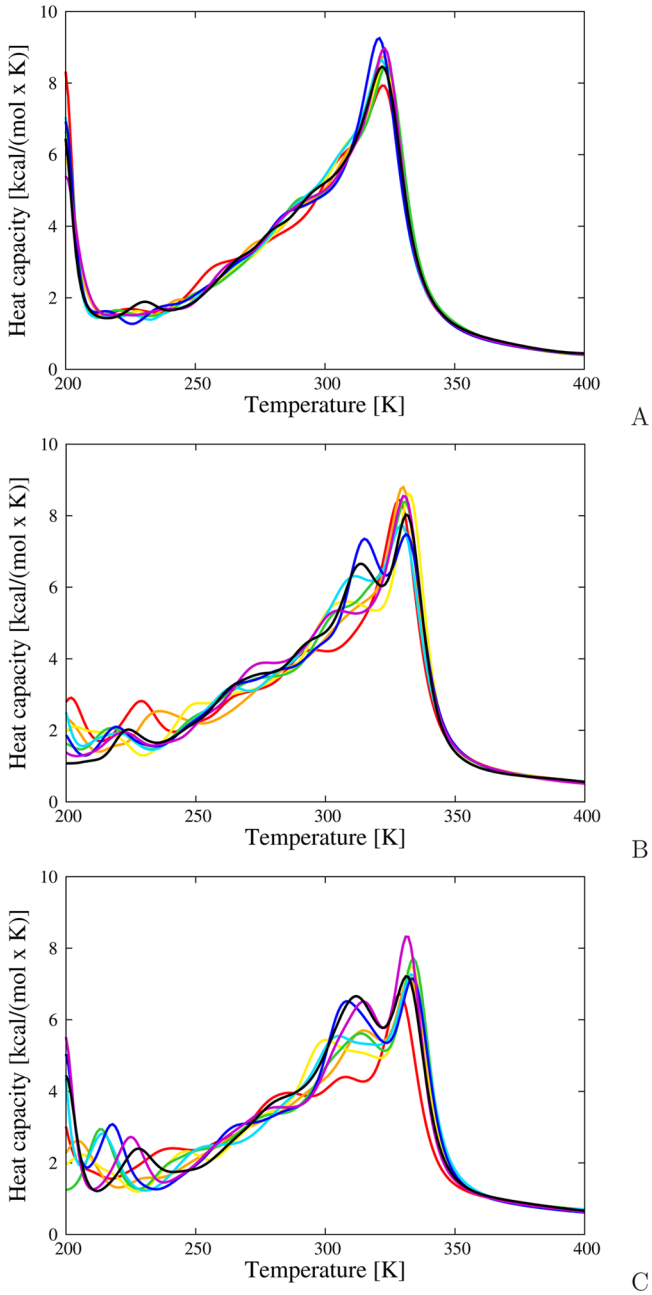


Figure 7. Convergence plots of the heat capacity of 1BDD for $w_{SC-corr}$ = 0 (A), $w_{SC-corr}$ = 0.57 (B), and $w_{SC-corr}$ = 1.0 (C). Different colors denote heat-capacity curves for consecutive windows of the MREMD simulation, for the range from 10 000 to 50 000 000 MD steps divided into 8 equal windows. Red, window 1; orange, window 2; yellow, window 3; green, window 4; cyan, window 5; blue, window 6; purple, window 7; black, window 8.

We used the C^α root-mean-square deviation (C^α RMSD) as a measure of the agreement between the calculated and the experimental structures. Because MREMD gives conformational ensembles, we mostly used ensemble-averaged RMSDs. The four RMSD values that we analyzed are defined by eqs 14−17, respectively.

$$\langle \rho \rangle (T_a) = \sum_i \rho_i w_i(T_a) \tag{14}$$

dx.doi.org/10.1021/ct4004977 | J. Chem. Theory Comput. 2013, 9, 4620−4632

Table 2. Fraction of the Cluster of Conformations with the Lowest RMSD and Fraction of Native-Like Conformations for Different Values of $w_{SC-corr}$

		lowest-RMSD cluster			native-like conformations		
		$w_{SC-corr}$					
protein	T_a [K][a]	0.0	0.57	1.0	0.0	0.57	1.0
		fraction					
1BDD	280	0.051	0.272	0.104	1.000	0.956	0.684
1CLB	210	0.235	0.280	0.480	0.482	0.496	0.481
1E0G	270	0.010	0.329	0.449			
1GAB	280	0.274	0.538	0.658	0.886	0.909	0.658
1KOY	290	0.376	0.140	0.252			
1POU	210	0.155	0.380	0.470		0.380	
1PRU	235	0.146	0.660	0.324			
average		0.170	0.340	0.470			

[a]The temperature at which the averages were calculated.

$$\langle\rho\rangle_{clust}^{min}(T_a) = \min_I \sum_{i\in I} \rho_i w_i(T_a) \quad (15)$$

$$\rho^{min} = \min_i \rho_i \quad (16)$$

$$\langle\rho\rangle_{clust}^{nat}(T_a) = \sum_{i\in N} \rho_i w_i(T_a) \quad (17)$$

where ρ_i is the C^α RMSD of the ith conformation and all other symbols are defined by eq 13 and the adjacent text.

The values of $\langle\rho\rangle(T_a)$, $\langle\rho\rangle_{clust}^{min}(T_a)$, and ρ^{min} are shown, for all proteins, in Figure 8A–C as bar diagrams. The values of $\langle\rho\rangle_{clust}^{nat}(T_a)$ are shown in Figure 8D, for 1BDD, 1CLB, 1GAB, and 1POU; the clusters of conformations of the other proteins do not meet the RMSD criterion of native-likeness (less than 0.1 Å average RMSD per amino-acid residue). The left (empty), middle (slanted checker pattern), and right (filled) bars correspond to calculations with $w_{SC-corr}$ = 0, $w_{SC-corr}$ = 0.57, and $w_{SC-corr}$ = 1.0, respectively.

It can be seen that, with $w_{SC-corr}$ = 0.57 (slanted-checker pattern bars), the average RMSD ($\langle\rho\rangle(T_a)$ of eq 14 in Figure 8A), the RMSD over the lowest-RMSD cluster ($\langle\rho\rangle_{clust}^{min}(T_a)$ of eq 15 in Figure 8B), and the lowest RMSD (ρ^{min} of eq 16 in Figure 8C), averaged over all proteins (the rightmost bars in panels A–C of Figure 8), decrease by 0.20 Å, 0.47 Å, and 0.66 Å, respectively, with respect to the force field without the new terms ($w_{SC-corr}$ = 0). It should be noted that the RMSDs over the sets of native-like conformations ($\langle\rho\rangle_{clust}^{nat}(T_a)$ of eq 17 in Figure 8D) could be compared only for 1BDD, 1CLB, 1GAB, and 1POU because no sufficiently native-like conformations were obtained for 1E0G, 1E0L, 1KOY, and 1PRU, and only $w_{SC-corr}$ = 0.57 produced native-like conformations for 1POU.

A commonly used measure of the similarity of a predicted protein structure to the corresponding experimental structure is the Global Distance Test (GDT_TS),[71] which is the percentage of residues of the largest set, not necessarily continuous, deviating from the experimental structure by no more than a specified distance cutoff. The GDT_TS bar plots, constructed for the proteins studied at the 6.0 Å cutoff, are shown in Figure 8E. It can be seen that the GDT_TS score increases after including the new $U_{SC-corr}$ terms for 1CLB, 1E0G, 1GAB, and 1POU with $w_{SC-corr}$ = 0.57 (checker-patterned bars in Figure 8E), which confirms the results from the RMSD analysis.

Based on the data collected in Table 2, it can be seen that the ensembles of 1BDD and 1GAB (three-α-helix-bundle proteins) contain predominantly native-like structures, irrespective of the introduction of the new potentials. For 1CLB, the ensemble consists of about 50% of native-like structures, irrespective of $w_{SC-corr}$. For other proteins, both the fraction of the structures that belong to the lowest-RMSD cluster and the fraction of the native-like structures is the greatest for $w_{SC-corr}$ = 0.57 (Table 2). A major difference occurs for 1POU, for which native-like structures are obtained only with $w_{SC-corr}$ = 0.57; Figure 8A–C demonstrates that there is a dramatic drop of the ensemble-average RMSD, the average RMSD of the lowest-RMSD cluster, and the lowest RMSD obtained with $w_{SC-corr}$ = 0.57 with respect to those obtained with $w_{SC-corr}$ = 0. A small decrease of the RMSD, for the structures obtained with $w_{SC-corr}$ = 0.57 (middle checker-patterned bars in Figure 8A–C) with respect to those obtained without the new terms, is observed for 1GAB. Introducing the new terms increased the average RMSD (Figure 8A) for 1CLB (two EF-hand motifs; four α-helices) but, with $w_{SC-corr}$ = 0.57, the RMSD over the native-like cluster and the lowest RMSD decrease (Figure 8B–D). The decrease of the ensemble-average RMSDs is caused by about doubling the size of the lowest-RMSD clusters (Table 2, 0.170 to 0.340). For 1KOY, 1POU (complex α-helical folds with long loops or unstructured regions), and 1E0G (an $\alpha + \beta$) protein, introduction of the new terms, with $w_{SC-corr}$ = 0.57, results in a significant drop of all three RMSD values (Figure 8A–C); however, the calculated structures are sufficiently close to the experimental structure only for 1POU (Figure 8D). No improvement is observed for 1PRU (a complex α-helical structure with long unstructured regions) and for 1E0L (a three-stranded antiparallel β-sheet structure) in Figure 8A–C.

The tests performed here have also demonstrated that an exaggerated contribution ($w_{SC-corr}$ = 1.0) from the new terms deteriorates the quality of simulated structures. This feature probably results from the fact that interactions stabilizing regular secondary structures are underrepresented when the weight of the new terms is too large.

To determine what parts of the calculated structures were most improved by introducing the new terms, we analyzed plots of the deviations of the C^α atoms of the mean structures corresponding to the most native-like clusters from those of the experimental structures, as functions of residue number in the sequence. The deviations were calculated at optimal super-

dx.doi.org/10.1021/ct4004977 | J. Chem. Theory Comput. 2013, 9, 4620–4632

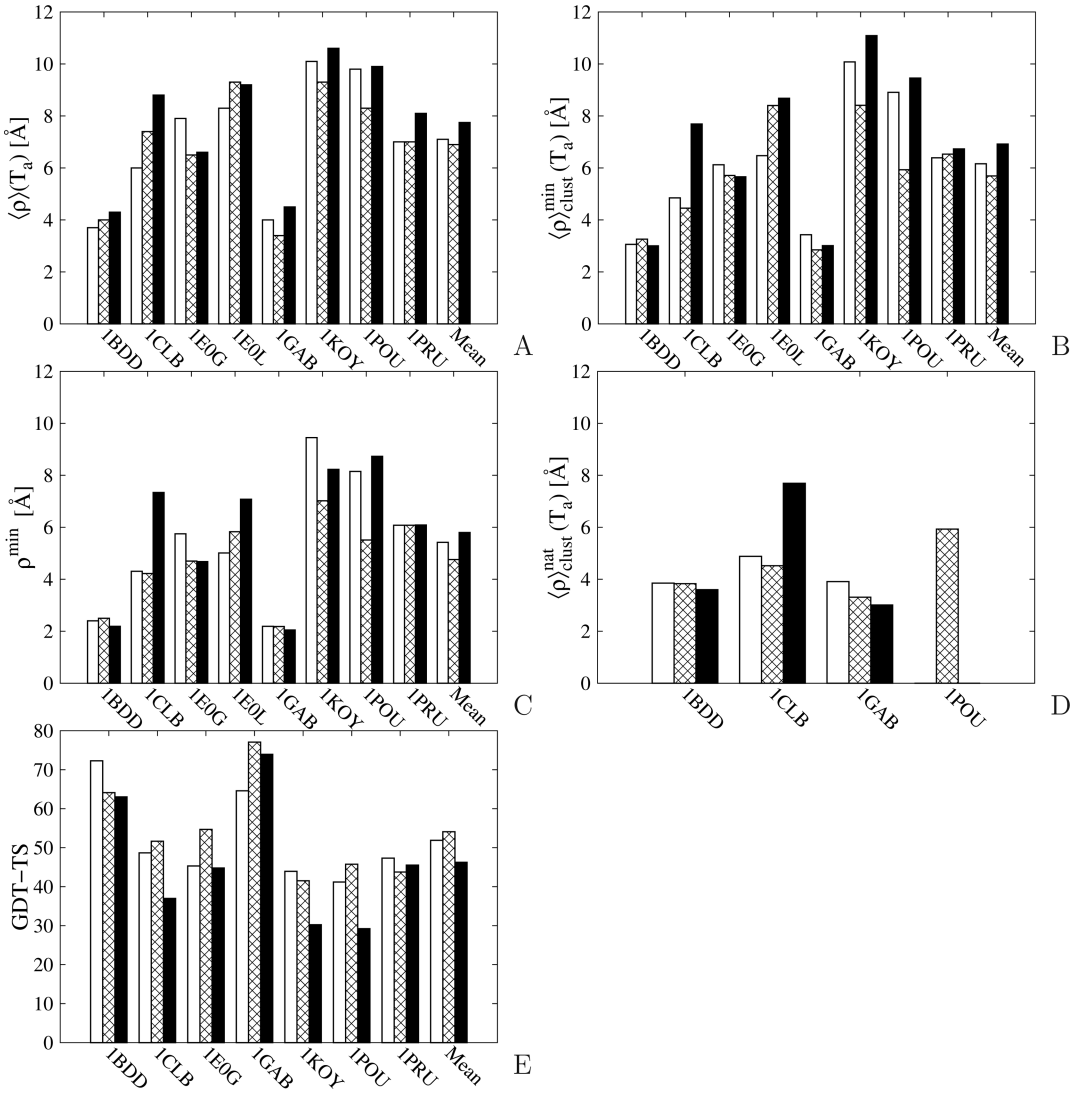


Figure 8. (A–D) Bar diagrams of the ρ values for all 8 test proteins and the ρ values averaged over all proteins shown on the right of each diagram, for $w_{SC-corr}$ = 0 (white bars), 0.57 (slanted-checker pattern bars), 1.0 (black bars), respectively. For each protein, panel A shows the RMSD averaged over the conformational ensemble generated during the MD run [$\langle\rho\rangle(T_a)$ of eq 14], panel B shows the minimum of the cluster-averaged RMSD [$\langle\rho\rangle^{min}_{clust}(T_a)$ of eq 15], panel C shows the lowest RMSD obtained during the corresponding MREMD run [ρ^{min} of eq 16], and panel D shows RMSD averaged over the sets of native-like conformations [$\langle\rho\rangle^{nat}_{clust}(T_a)$ of eq 17]; each set is defined as a sum of clusters with ensemble-average RMSD per residue less than 0.1 Å. See Table 2 for the populations of conformations of panels (B) and (D) and for the temperatures T_a at which the analysis was performed. (E) Bar diagram of the Global Distance Test (GDT_TS)[71] measure at the 6.0 Å distance cutoff for the clusters with the lowest RMSD.

position of the computed structure on the experimental structure. An example of such a superposition for 1GAB is shown in Figure 9. Sample plots for 1GAB and 1POU are shown in Figures 10 and 11, respectively, for all three values of $w_{SC-corr}$. It can be seen from Figures 10 and 11 that the most significant improvement has been achieved in the loop regions (those between the α-helices). For 1GAB, the quality of the calculated structure is most improved in the loop regions, that is, at residues 19–26 and 33–42 (Figures 9 and 10). This is where the sequence-specific local correlation terms were expected to make the greatest difference. However, for 1POU (Figure 11), not only are the loop regions (residues 18–28 and 33–41) of the computed structure closer to those of the experimental structure with the new potential but the α-helices of the computed structure (residues 29–38, 44–51, and 62–75) also move closer to their counterparts in the experimental structure. This finding demonstrates that the enhancement of the representation of local interactions not only improves the quality of the calculated loop structures but also influences the packing of regular secondary-structure elements and can, therefore, be critical for the correct prediction of the topology of the whole fold. The importance of chain reversals and loop regions was demonstrated in our earlier experimental[72−75] and theoretical work.[76] In the last study,[76] we demonstrated that the loop regions between helices in 1BDD persist as chain reversals beyond the melting-transition temperature. For 1CLB, 1E0G, and 1KOY the same improvement in the quality of the calculated loop regions and packing of secondary structure elements is observed as for 1GAB and 1POU.

To appraise the improvement in the loop regions of the test proteins quantitatively, resulting from the introduction of the side-chain backbone correlation potentials, we computed the mean distances between the C^{α} atoms of the loop regions of the calculated structures and those of the experimental

dx.doi.org/10.1021/ct4004977 | J. Chem. Theory Comput. 2013, 9, 4620−4632

Figure 9. Superposition of the average structure of 1GAB obtained in MREMD simulations with $w_{SC-corr}$ = 0 (yellow ribbons and lines), 0.57 (pink ribbons and lines), onto the experimental structure (green ribbons and lines). The ribbons are thicker for fragments for which the most significant improvement of the quality of the calculated structure was observed after introducing the new potentials.

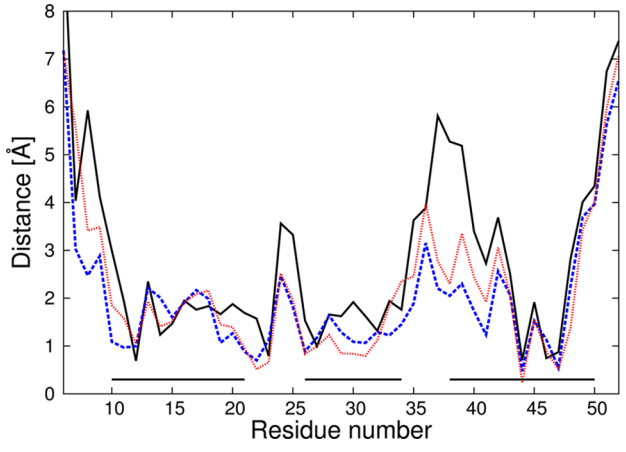


Figure 10. Plot of distances between C^{α} atoms of the average structures of 1GAB after superposition of the structure simulated with $w_{SC-corr}$ = 0 (solid black line), 0.57 (dashed blue line), and 1.0 (dotted red line) onto the experimental structure. Horizontal lines close to the abscissa mark α-helices in the experimental structure.

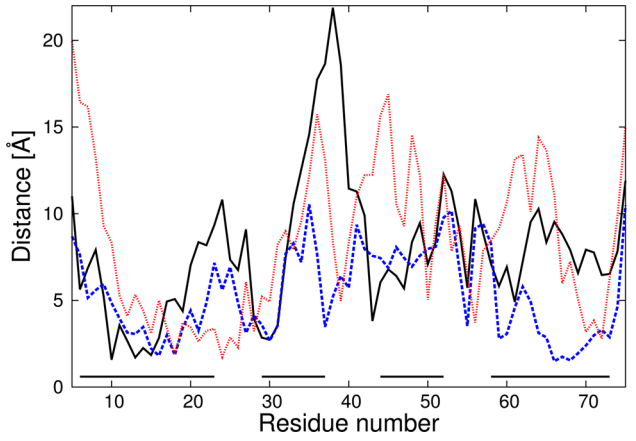


Figure 11. Same as Figure 10 but for 1POU.

structures, after optimal superposition of the computed structures onto the respective experimental structures. These average distances for each of the test proteins, as well as their averages over all test proteins are shown, for $w_{SC-corr}$ = 0 and $w_{SC-corr}$ = 0.57, in Table 3.

Table 3. Average Improvement of Distances between Corresponding C^{α} Atoms from Loop Regions with the Mean Value for All Test Proteins, Which Folded during Simulation into Native-like Structures

	distances [Å]			
protein	$w_{SC-corr}$ = 0.0	$w_{SC-corr}$ = 0.57	improvement [Å]	improvement [%]
1BDD	2.78	3.06	−0.28	−10.07
1CLB	5.14	4.67	0.47	9.14
1GAB	2.98	1.69	1.29	43.29
1POU	9.86	6.62	3.24	32.86
average			1.18	18.81

Another illustrative example of the influence of the new $U_{SC-corr}$ terms on structure is presented in Figure 12 with the example of 1E0G (an $\alpha + \beta$ protein). Without the new terms (Figure 12A), the calculated structure is all-α-helical. With $w_{SC-corr}$ = 0.57 (Figure 12B), the C-terminal segment becomes a β-strand, as in the experimental structure. Finally, with $w_{SC-corr}$ = 1.0 (Figure 12C), two antiparallel β-strands are present in the calculated structure which has the native

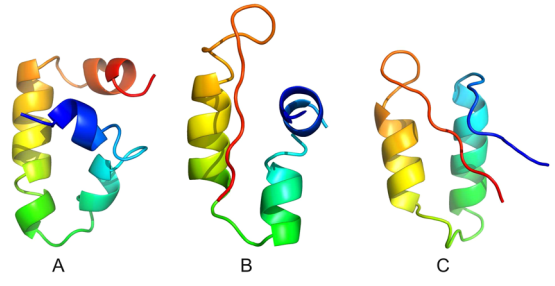


Figure 12. Representatives of the most probable clusters of 1E0G obtained in MREMD simulations with (A) $w_{SC-corr}$ = 0, (B) $w_{SC-corr}$ = 0.57, and (C) $w_{SC-corr}$ = 1.0. The C^{α} RMSDs from the experimental 1E0G structures are 8.5 Å, 5.7 Å, and 5.7 Å, respectively. The chains are colored blue to red from the N- to the C-terminus.

dx.doi.org/10.1021/ct4004977 | J. Chem. Theory Comput. 2013, 9, 4620−4632

topology. However, the β-strands are not tightly packed into a β-sheet, and the RMSD is equal to 5.7 Å, which is about 1 Å too high to consider the calculated structure native-like by the criterion adapted in this work (0.1 Å per amino-acid residue). It should be noted that the force field used in these calculations was parametrized with only 1GAB, which is an α-helical protein[11] and further parametrization is in progress.

4. SUMMARY

In this work, statistical potentials for the virtual-bond-dihedral angles have been determined and implemented in the coarse-grained UNRES force field[11,25,26,29,36−44] developed in our laboratory. The potentials have the functional forms of one-dimensional Fourier series in the virtual-bond-dihedral angles $\tau^{(1)}$, $\tau^{(2)}$, and $\tau^{(3)}$ involving side-chain centers (Figure 2). The new potentials not only improve the quality of the calculated structures but also result in an increase of the population of native-like clusters on average. Moreover, improvement of the resolution of the calculated structure occurs not only in the loop regions but in better packing of secondary-structure elements following the improvement of the quality of the loop regions (Figure 11). The UNRES force field with new side-chain backbone correlation potentials might prove to be a powerful tool in refining loop conformations even when the structure of the nonloop regions is well described.

Although the newly introduced $U_{SC-corr}$ potentials definitely improve the accuracy of UNRES in reproducing loop structures, even with first-guess values of the weights of these energy terms, all energy-term weights must be redetermined by calibration on structural and thermodynamic data for protein folding, to complete the implementation of $U_{SC-corr}$ in the force field, as in our earlier work.[11] This work is currently being carried out in our laboratory. We are also determining physics-based $U_{SC-corr}$ potentials from AM1 energy surfaces of terminally blocked amino-acid residues.

■ ASSOCIATED CONTENT

Supporting Information

Table S1: Coefficients of Fourier-series fit (eq 5 of the main text) to the $U_{SC-corr;XY}(\tau^{(1)})$ torsional potentials of mean force for all pairs of extended amino-acid types (Table 1 of the main text). Table S2: Coefficients of Fourier-series fit (eq 5 of the main text) to the $U_{SC-corr;XY}(\tau^{(2)})$ torsional potentials of mean force for all pairs of extended amino-acid types (Table 1 of the main text). Table S3: Coefficients of Fourier-series fit (eq 5 of the main text) to the $U_{SC-corr;XY}(\tau^{(3)})$ torsional potentials of mean force for all pairs of extended amino-acid types (Table 1 of the main text). Table S4: Parameters of the least-squares fitting of $U_{SC-corr;XY}(\tau^{(m)})$ potentials to the statistical backbone potentials, $U_{stor;XY}(\gamma)$ (eq 6 of the main text). Figure S1: Plots of the 35 side-chain backbone correlation potentials $U_{XY}(\tau^{(m)})$,m = 1,2,3, where X and Y each denote one residue type of the reduced amino-acid alphabet (see Table 1), calculated from the statistical analysis of loop structures. The values of X and Y are stated above each plot. Black circles represent the values of the dimensionless PMFs calculated from histograms (eq 3) and red lines represent one-dimension Fourier series fits (eq 5) to the PMF values. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +48585235430. Fax: +48585235472. E-mail: adasko@sun1.chem.univ.gda.pl.

Author Contributions

⊥These authors contributed equally to the paper.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by Grant No. MPD/2010/5 from the Foundation for Polish Science (FNP), Grant No. DEC-2011/01/N/ST4/01772 from the National Science Center of Poland, and by grants from the National Institutes of Health (GM-14312) and the National Science Foundation (MCB10-19767). This research was supported by an allocation of advanced computing resources provided by the National Science Foundation (http://www.nics.tennessee.edu/), and by the National Science Foundation through TeraGrid resources provided by the Pittsburgh Supercomputing Center. Computational resources were also provided by (a) the supercomputer resources at the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdańsk, (b) the 624-processor Beowulf cluster at the Baker Laboratory of Chemistry, Cornell University, and (c) our 184-processor Beowulf cluster at the Faculty of Chemistry, University of Gdańsk.

■ REFERENCES

(1) Whitford, D. *Proteins Structure and Function*; Wiley: Chichester, 2007.
(2) The Uniprot Consortium. *Nucleic Acid Res.* **2011**, *39*, D214−D219.
(3) Berman, H. M. *Acta Cryst. A* **2008**, *64*, 88−95.
(4) Moult, J.; Hubbard, T.; Fidelis, K.; Pedersen, J. T. *Proteins* **1999**, *3*, 2−6.
(5) Moult, J. *Curr. Opin. Struct. Biol.* **2005**, *15*, 285−289.
(6) Moult, J.; Fidelis, K.; Kryshtafovych, A.; Rost, B.; Hubbard, T.; Tramontano, A. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 3−9.
(7) Moult, J.; Fidelis, K.; Kryshtafovych, A.; Rost, B.; Tramontano, A. *Proteins: Struct. Func. Bioinf.* **2009**, *77*, 1−4.
(8) Zhang, Y. *Proteins: Struct., Funct., Bioinf.* **2009**, *77*, 100−113.
(9) Marti-Renom, M.; Stuart, A.; Fiser, A.; Sanchez, R.; Melo, F.; Sali, A. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291−325.
(10) Chen, Y.; Ding, F.; Nie, H.; Serohijos, A. W.; Sharma, S.; Wilcox, K. C.; Yin, S.; Dokholyan, N. V. *Archiv. Biochem. Biophys.* **2008**, *469*, 4−19.
(11) Liwo, A.; Khalili, M.; Czaplewski, C.; Kalinowski, S.; Ołdziej, S.; Wachucik, K.; Scheraga, H. *J. Phys. Chem. B* **2007**, *111*, 260−285.
(12) Yeh, I.-C.; Lee, M. S.; Olson, M. A. *J. Phys. Chem. B* **2008**, *112*, 15064−15073.
(13) Durrant, J. D.; McCammon, J. A. *BMC Biol.* **2011**, 71.
(14) Okimoto, N.; Futatsugi, N.; Fuji, H.; Suenaga, A.; Morimoto, G.; Yanai, R.; Ohno, Y.; Narumi, T.; Taiji, M. *PLoS Comput. Biol.* **2009**, *5*, e1000528.
(15) Terstappen, G.; Reggiani, A. *Trends Pharmacol. Sci.* **2001**, *22*, 23−6.
(16) Rao, V. S.; Srinivas, K. *J. Bioinf. Sequence Anal.* **2011**, *3*, 89−94.
(17) Lee, J.; Wu, S.; Zhang, Y. In *From Protein Structure to Function with Bioinformatics*; Rigden, D., Ed.; Springer: The Netherlands, 2009; pp 3−25.
(18) Juraszek, J.; Bolhuis, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 15859−15864.
(19) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740−744.

(20) Ripoll, D. R.; Vila, J. A.; Scheraga, H. A. *J. Mol. Biol.* **2004**, *339*, 915−925.
(21) Freddolino, P.; Schulten, K. *Biophys. J.* **2009**, *97*, 2338−2347.
(22) Shaw, D. E.; Deneroff, M. M; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Ierardi, D. J.; Kolossvary, I.; Klepeis, J. L.; Layman, T.; Mcleavey, C.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C. *Commun. ACM* **2008**, *51*, 91−97.
(23) Skolnick, J.; Zhang, Y.; Arakaki, A. K.; Koliński, A.; Boniecki, M.; Szilagyi, A.; Kihara, D. *Proteins: Struct. Func. Genet.* **2003**, *53*, 469−479.
(24) Czaplewski, C.; Liwo, A.; Makowski, M.; Ołdziej, S.; Scheraga, H. A. In *Multiscale Approaches to Protein Modeling*; Kolinski, A., Ed.; Springer: New York, 2010; Chapter 3.
(25) Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A. *J. Chem. Phys.* **2001**, *115*, 2323−2347.
(26) Liwo, A.; Czaplewski, C.; Ołdziej, S.; Rojas, A. V.; Kazmierkiewicz, R.; Makowski, M.; Murarka, R. K.; Scheraga, H. A. In *Coarse-Graining of Condensed Phase and Biomolecular Systems*; Voth, G., Ed.; CRC Press: Boca Raton, FL, 2008; Chapter 8, pp 1391−1411.
(27) Khalili, M.; Liwo, A.; Jagielska, A.; Scheraga, H. *J. Phys. Chem. B* **2005**, *109*, 13798−13810.
(28) Liwo, A.; Khalili, M.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2362−2367.
(29) Sieradzan, A. K.; Scheraga, H. A.; Liwo, A. *J. Chem. Theory Comput.* **2012**, *8*, 1334−1343.
(30) Sieradzan, A. K.; Hansmann, U. H. E.; Scheraga, H. A.; Liwo, A. *J. Chem. Theory Comput.* **2012**, *8*, 4746−4757.
(31) Sieradzan, A. K.; Liwo, A.; Hansmann, U. H. E. *J. Chem. Theory Comput.* **2012**, *8*, 3416−3422.
(32) Liwo, A.; Lee, J.; Ripoll, D. R.; Pillardy, J.; Scheraga, H. A. *Proc. Natl. Acad. Sci., U.S.A.* **1999**, *96*, 5482−5485.
(33) Lee, J.; Scheraga, H. A. *Int. J. Quantum Chem.* **1999**, *75*, 255−265.
(34) Ołdziej, S.; Czaplewski, C.; Liwo, A.; Chinchio, M.; Nanias, M.; Vila, J. A.; Khalili, M.; Arnautova, Y. A.; Jagielska, A.; Makowski, M.; Schafroth, H. D.; Kaźmierkiewicz, R.; Ripoll, D. R.; Pillardy, J.; Saunders, J. A.; Kang, Y. K.; Gibson, K. D.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7547−7552.
(35) Liwo, A.; He, Y.; Scheraga, H. A. *Phys. Chem. Chem. Phys.* **2011**, *13*, 16890−16901.
(36) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *Protein Sci.* **1993**, *2*, 1715−1731.
(37) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Ołdziej, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 874−887.
(38) Liwo, A.; Ołdziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 849−873.
(39) Liwo, A.; Kazmierkiewicz, R.; Czaplewski, C.; Groth, M.; Ołdziej, S.; Wawak, R. J.; Rackovsky, S.; Pincus, M. R.; Scheraga, H. A. *J. Comput. Chem.* **1998**, *19*, 259−276.
(40) Liwo, A.; Ołdziej, S.; Czaplewski, C.; Kozłowska, U.; Scheraga, H. A. *J. Phys. Chem. B* **2004**, *108*, 9421−9438.
(41) Kozłowska, U.; Liwo, A.; Scheraga, H. A. *J. Phys.: Cond. Matter* **2007**, *19*, 285203.
(42) Kozłowska, U.; Maisuradze, G. G.; Liwo, A.; Scheraga, H. A. *J. Comput. Chem.* **2010**, *31*, 1154−1167.
(43) Makowski, M.; Liwo, A.; Sobolewski, E.; Scheraga, H. A. *J. Phys. Chem. B* **2011**, *115*, 6119−6129.
(44) Makowski, M.; Liwo, A.; Scheraga, H. A. *J. Phys. Chem. B* **2011**, *115*, 6130−6137.
(45) Shen, H.; Liwo, A.; Scheraga, H. A. *J. Phys. Chem. B* **2009**, *113*, 8738−8744.
(46) Kolinski, A.; Skolnick, J. *J. Chem. Phys.* **1992**, *97*, 9412−9426.
(47) Johansson, M. U.; de Chateau, M.; Wikström, M.; Forsön, S.; Drakenberg, T.; Björck, L. *J. Mol. Biol.* **1997**, *266*, 859−865.
(48) Solis, A. D.; Rackovsky, S. *Proteins: Struct., Funct., Bioinf.* **2000**, *38*, 149−164.
(49) Gouda, H.; Torigoe, H.; Saito, A.; Sato, M.; Arata, Y.; Shimada, I. *Biochemistry* **1992**, *31*, 9665−9672.
(50) Skelton, N. J.; Krdel, J.; Chazin, W. J. *J. Mol. Biol.* **1995**, *249*, 441−462.
(51) Bateman, A.; Bycroft, M. *J. Mol. Biol.* **2000**, *299*, 1113−1119.
(52) Macias, M. J.; Gervais, V.; Civera, C.; Oschkinat, H. *Nat. Struct. Biol.* **2000**, *7*, 375−379.
(53) Fukushima, K.; Kikuchi, J.; Koshiba, S.; Kigawa, T.; Kuroda, Y.; Yokoyama, S. *J. Mol. Biol.* **2002**, *321*, 317−327.
(54) Assa-Munt, N.; Mortishire-Smith, R. J.; Aurora, R.; Herr, W.; Wright, P. E. *Cell* **1993**, *73*, 193−205.
(55) Nagadoi, A.; Morikawa, S.; Nakamura, H.; Enari, M.; Kobayashi, K.; Yamamoto, H.; Sampei, G.; Mizobuchi, K.; Schumacher, M. A.; Brennan, R. G. *Structure* **1995**, *3*, 1217−24.
(56) Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R.; Thompson, J. D.; Gibson, T. J.; Higgins, D. G. *Bioinformatics* **2011**, *23*, 2947−2948.
(57) Rhee, Y. M.; Pande, V. S. *Biophys. J.* **2003**, *84*, 775−786.
(58) Czaplewski, C.; Kalinowski, S.; Liwo, A.; Scheraga, H. A. *J. Chem. Theory Comput.* **2009**, *5*, 627−640.
(59) Hansmann, U. H. E.; Okamoto, Y. *J. Comput. Chem.* **1993**, *14*, 1333−1338.
(60) Hansmann, U. H. *Chem. Phys. Lett.* **1997**, *281*, 140−150.
(61) Sugita, Y.; Okamoto, Y. *Phys. Rev. Lett.* **2000**, *329*, 261−270.
(62) Khalili, M.; Liwo, A.; Rakowski, F.; Grochowski, P.; Scheraga, H. *J. Phys. Chem. B* **2005**, *109*, 13785−13797.
(63) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(64) Rosta, E.; Buchete, N. V.; Hummer, G. *J. Chem. Theory Comput.* **2009**, *5*, 1393−1399.
(65) Zimmerman, S. S.; Pottle, M. S.; Némethy, G.; Scheraga, H. A. *Macromolecules* **1977**, *10*, 1−9.
(66) Levitt, M.; Chothia, C. *Nature* **1976**, *261*, 552−558.
(67) Ołdziej, S.; Kozłowska, U.; Liwo, A.; Scheraga, H. A. *J. Phys. Chem. A* **2003**, *107*, 8035−8046.
(68) Nanias, M.; Czaplewski, C.; Scheraga, H. A. *J. Chem. Theory Comput.* **2006**, *2*, 513−528.
(69) Späth, H. *Cluster Analysis Algorithms*; Halsted Press: New York, 1980.
(70) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011−1021.
(71) Zemla, A. *Nucleic Acids Res.* **2003**, *31*, 3370−3374.
(72) Skwierawska, A.; Makowska, J.; Oldziej, S.; Liwo, A.; Scheraga, H. A. *Proteins: Struct., Funct., Bioinf.* **2009**, *75*, 931−953.
(73) Skwierawska, A.; Żmudzińska, W.; Ołdziej, S.; Liwo, A.; Scheraga, H. *Proteins: Strut., Funct., Bioinf.* **2009**, *76*, 637−654.
(74) Lewandowska, A.; Ołdziej, S.; Liwo, A.; Scheraga, H. A. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 723−737.
(75) Lewandowska, A.; Ołdziej, S.; Liwo, A.; Scheraga, H. A. *Biophys. Chem.* **2010**, *151*, 1−9.
(76) Maisuradze, G. G.; Liwo, A.; Ołdziej, S.; Scheraga, H. A. *J. Am. Chem. Soc.* **2010**, *132*, 9444−9452.