Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

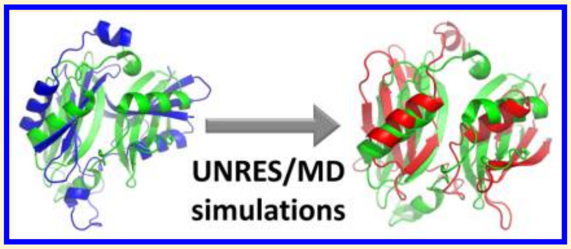
Prediction of Protein Structure by Template-Based Modeling Combined with the UNRES Force Field

Paweł Krupa,[†] Magdalena A. Mozolewska,[†] Keehyoung Joo,[‡] Jooyoung Lee,[§] Cezary Czaplewski,[†] and Adam Liwo*,[†,§]

[†]Faculty of Chemistry, University of Gdańsk, Wita Stwosza 63, 80-308 Gdańsk, Poland

[‡]Center for Advanced Computation and [§]Center for In Silico Protein Structure and School of Computational Sciences, Korea Institute for Advanced Study, 85 Hoegiro, Dongdaemun-gu, Seoul 130-722, Republic of Korea

Supporting Information



ABSTRACT: A new approach to the prediction of protein structures that uses distance and backbone virtual-bond dihedral angle restraints derived from template-based models and simulations with the united residue (UNRES) force field is proposed. The approach combines the accuracy and reliability of template-based methods for the segments of the target sequence with high similarity to those having known structures with the ability of UNRES to pack the domains correctly. Multiplexed replica-exchange molecular dynamics with restraints derived from template-based models of a given target, in which each restraint is weighted according to the accuracy of the prediction of the corresponding section of the molecule, is used to search the conformational space, and the weighted histogram analysis method and cluster analysis are applied to determine the families of the most probable conformations, from which candidate predictions are selected. To test the capability of the method to recover template-based models from restraints, five single-domain proteins with structures that have been well-predicted by template-based methods were used; it was found that the resulting structures were of the same quality as the best of the original models. To assess whether the new approach can improve template-based predictions with incorrectly predicted domain packing, four such targets were selected from the CASP10 targets; for three of them the new approach resulted in significantly better predictions compared with the original template-based models. The new approach can be used to predict the structures of proteins for which good templates can be found for sections of the sequence or an overall good template can be found for the entire sequence but the prediction quality is remarkably weaker in putative domain-linker regions.

INTRODUCTION

Knowledge of protein structure is essential for understanding the mechanisms of the function and dysfunction of living cells (including carcinogenesis and conformational diseases). This knowledge is, in turn, of utmost importance in biological applications, especially drug design. At present, experimental structures are available for only about 7% of known proteins. The gap between the numbers of known sequences and known structures is expected to widen because of exponential growth of the number of known protein sequences. The development of reliable methods for protein structure prediction is therefore necessary not only to acquire new knowledge of cell structure and function but also for practical reasons.

On the basis of the experience from the biannual Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP) exercises, an increasing number of predictors acknowledge that bioinformatics approaches have already reached their upper limit of prediction accuracy and that other routes must be explored to make progress.[1] The physics-based protocol of protein structure prediction,[2,3] which is based on the coarse-grained UNRES force field developed in our laboratory,[4] proved successful in the CASP experiments.[5−10] In the CASP10 experiment, UNRES achieved outstanding predictions of two free-modeling targets: T0663 and T0740. T0663 consists of two $\alpha+\beta$ domains, each of which is highly homologous to proteins from the protein database. However, it exhibits unusual domain packing in which the domains are rotated with respect to each other. UNRES was one of the only two approaches that was able to predict this packing, even though the obtained structure was of low resolution.[9] For T0740, we obtained the best results by using the helix-packing prediction information supplied by the β-sheet topology (BeST) prediction algorithm developed by the Floudas group,[11] which was made available to us as a result of our participation, as the wfCPUNK group, in the WEFOLD collaborative initiative.[10] In earlier CASP exercises, UNRES was able to determine correct folds for proteins that did have templates in the database, whereas all of the bioinformatics methods picked out wrong templates.[7,8,12] On the other hand, UNRES produces only medium-resolution structures, with an average root-mean-square deviation (RMSD) of about 5 Å per 50−60 residue protein segment with the current version of the force field.

Received: March 4, 2015
Published: May 12, 2015

DOI: 10.1021/acs.jcim.5b00117

Recently, a very efficient method of multiple sequence alignment based on seeking the optimal alignment by using the powerful conformational space annealing (CSA) global optimization method[5,13,14] was developed.[15−17] The method achieved top performance in the CASP8−CASP11 experiments for comparative modeling targets.[15,17,18] It should be noted that for target T0663, structures with the highest global distance test (GDT_TS) scores were obtained by the LEEcon group,[19] which used server models as templates. This result was achieved through correct prediction of the structures of individual domains, even though the domain packing was incorrect, as opposed to that obtained with UNRES.[9] Because the UNRES models of this target had the correct domain packing and UNRES supplemented with contact prediction information resulted in the best prediction of T0740,[10] in this work a combined approach is proposed in which knowledge-based modeling is used for those parts of the protein sequence for which the bioinformatics signals are strong and physics-based modeling with UNRES is implemented to obtain the correct domain packing. It is highly probable that the new approach can bridge the gap between the bioinformatics methods, which are accurate when good templates can be found but can fail when all or part of the target sequence exhibits too-weak similarity to those of database structures, and the physics-based methods, which are more robust because of the independence of structural databases but are more expensive and less accurate for template-based modeling (TBM) targets than the bioinformatics approaches.

In this paper, we report an initial implementation of the approach outlined above, in which the knowledge-based information is used in UNRES simulations in the form of $C^\alpha \cdots C^\alpha$ distance and backbone virtual-bond dihedral angle restraint penalties added to the UNRES energy function. In what follows, we first outline the UNRES model of polypeptide chains and force field and the physics-based protocol for protein structure prediction based on UNRES.[3] Subsequently, we describe how the knowledge-based information is implemented with UNRES. Finally, we describe the tests of the method with selected single- and two-domain CASP9 and CASP10 targets. We conclude by sketching the applications of the approach and future directions of its development.

■ METHODS

UNRES Model of Polypeptide Chains. In the physics-based coarse-grained UNRES force field (www.unres.pl),[3,4,8,20−23] a polypeptide chain is represented by a sequence of α-carbon (C^α) atoms, where united peptide groups (p) are positioned halfway between two consecutive C^α atoms and united side chains (SCs) are attached to their C^α atoms; SC and p are the only interaction sites, and C^α atoms serve only as geometric points (Figure 1).

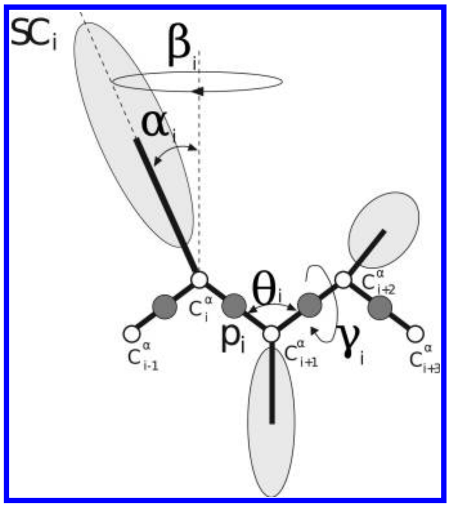


Figure 1. UNRES model of polypeptide chains. The interaction sites are side chain (SC) ellipsoids of different size attached to the corresponding α-carbons (C^α) with different virtual-bond lengths (b_{SC}) and peptide-bond centers (p). The equilibrium length of a peptide bond is 3.8 Å for the trans configuration and 2.8 Å for the cis configuration. The C^α atoms are represented by small open circles. For the *i*th residue, the geometry of the respective chain fragment can be described by using virtual-bond angles θ_i, virtual-bond dihedral angles γ_i, and the polar angles α_i and β_i.

The effective energy function is defined as a potential of mean force (PMF) of a system composed of the protein(s) and the surrounding solvent.[21,22] The PMF is expanded into a (finite) series of Kubo cluster-cumulant functions,[24] enabling us to determine multibody contributions, which are essential to represent regular secondary structures, such as α-helices and β-sheets.[21,25] The resulting UNRES energy function is as follows:

$$\begin{aligned} U_{\mathrm{UNRES}} = {} & w_{\mathrm{SCSC}} \sum_j \sum_{i<j} U_{\mathrm{SC}_i\mathrm{SC}_j} + w_{\mathrm{SCp}} \sum_j \sum_{i \neq j} U_{\mathrm{SC}_i\mathrm{p}_j} \\ & + f_2(T) w_{\mathrm{el}} \sum_j \sum_{i<j-1} U_{\mathrm{p}_i\mathrm{p}_j} + f_2(T) w_{\mathrm{tor}} \sum_i U_{\mathrm{tor}}(\gamma_i) \\ & + f_3(T) w_{\mathrm{tord}} \sum_i U_{\mathrm{tord}}(\gamma_i, \gamma_{i+1}) \\ & + f_2(T) w_{\mathrm{SC\text{-}corr}} U_{\mathrm{SC\text{-}corr}} + w_{\mathrm{b}} \sum_i U_{\mathrm{b}}(\theta_i) \\ & + w_{\mathrm{rot}} \sum_i U_{\mathrm{rot}}(\alpha_{\mathrm{SC}_i}, \beta_{\mathrm{SC}_i}) + \sum_{m=2}^{N_{\mathrm{corr}}} f_m(T) w_{\mathrm{corr}}^{(m)} U_{\mathrm{corr}}^{(m)} \\ & + f_3(T) w_{\mathrm{turn}}^{(3)} U_{\mathrm{turn}}^{(3)} + f_4(T) w_{\mathrm{turn}}^{(4)} U_{\mathrm{turn}}^{(4)} \\ & + w_{\mathrm{bond}} \sum_i U_{\mathrm{bond}}(d_i) + w_{\mathrm{SS}} \sum_{\text{disulfide bonds}} U_{\mathrm{SS}_i} \\ & + n_{\mathrm{SS}} E_{\mathrm{SS}} \end{aligned} \tag{1}$$

in which

$$f_n(T) = \frac{\ln[\exp(1) + \exp(-1)]}{\ln\{\exp[(T/T_0)^{n-1}] + \exp[-(T/T_0)^{n-1}]\}} \tag{2}$$

The terms $U_{\mathrm{SC}_i\mathrm{SC}_j}$ correspond to the mean free energies of solvent-mediated interactions between the side chains.[26] The terms $U_{\mathrm{SC}_i\mathrm{p}_j}$ correspond to the excluded-volume potentials of the side-chain−peptide group interactions.[20] The terms $U_{\mathrm{p}_i\mathrm{p}_j}$ represent the energies of mean-field electrostatic interactions between backbone peptide groups.[20] The terms U_{tor} and U_{tord} are the backbone torsional and double-torsional potentials, respectively, for the rotation about a given virtual bond or two consecutive virtual bonds.[27] The terms $U_{\mathrm{SC\text{-}corr}}$ are torsional potentials for rotation about $C^\alpha \cdots C^\alpha$ virtual bonds involving side-chain centers ($C^\alpha \cdots C^\alpha \cdots C^\alpha \cdots \mathrm{SC}$, $\mathrm{SC} \cdots C^\alpha \cdots C^\alpha \cdots C^\alpha$, and $\mathrm{SC} \cdots C^\alpha \cdots C^\alpha \cdots \mathrm{SC}$). These potentials were introduced for better recognition of loop regions and secondary structures.[28,29] The terms U_b and U_{rot} are the virtual-angle bending and side-chain rotamer potentials, respectively, and the term U_{bond} accounts for backbone and side-chain virtual-bond stretching;[4,30] recently,[31] we extended the backbone virtual-bond stretching term to account for the trans−cis transition of peptide groups. The terms $U_{corr}^{(m)}$ and $U_{turn}^{(m)}$ correspond to the correlations (of order m) between peptide-group electrostatic and backbone local interactions;[21,22] the terms $U_{turn}^{(m)}$ (the "turn" terms) involve consecutive segments of the chain. The terms U_{SS_i} are the energies of distortion of disulfide bonds from their equilibrium configurations. E_{SS} is the energy for the formation of an "unstrained" disulfide bond in the chain (relative to the presence of two free cysteine residues), and n_{SS} is the number of disulfide bonds. The w's are energy-term weights determined by calibration with training proteins. Finally, the multipliers $f_n(T)$ account for the temperature dependence of the respective contributions to the mean-field energy.[3]

Use of Restraints from Templates in UNRES/MREMD Simulations. It is known that attempts to refine the template-based models in high-confidence regions usually results in worsening the prediction quality even when sophisticated all-atom force fields are used. A recent report described the first method that could consistently refine structures by using a series of short MD simulations with restraints;[32] this approach proved to be very successful in the refinement category of the CASP10 and CASP11 experiments. However, even this method was able to achieve an RMSD decrease of only 0.14−0.94 Å, depending on the strength of the restraints and the type of sampling used.

In the approach developed in the present work, the simulations are therefore restrained on the basis of the quality of the alignment in particular regions of the reference structure(s). The penalty function consists of multimodal terms similar to those used in the MODELLER program,[33] as defined by eq 3:

$$U_{restr} = -\sum_{i<j} \ln \sum_{k} \exp\left[-\frac{(d_{ij} - d_{ij;k}^{\circ})^2}{2\sigma_{ij;k}{}^2}\right] - \sum_{i} \ln \sum_{k} \exp\left[-\frac{(\omega_i - \omega_{i;k}^{\circ})^2}{2\sigma_{\omega_i}{}^2}\right] \quad (3)$$

where d_{ij} is the distance between residues i and j (tentatively taken as the distance between the corresponding C^{α} atoms), $d_{ij;k}^{\circ}$ is the corresponding distance in the kth reference structure, ω_i is the ith restrained angle (in the current work, only the virtual-bond dihedral angles γ defined in Figure 1 are considered), $\omega_{i;k}^{\circ}$ is the corresponding angle in the kth reference structure, and the σ's are the "standard deviations" of the restrained quantities, which are assigned on the basis of the reliability of the alignment (see eqs 4 and 5). The logarithms of the sums of Gaussians centered at the values corresponding to given templates enable the system to jump between templates instead of being restrained to follow a given reference structure, which is especially important for ambiguous alignments. In this study, the reference structures were template-based models generated using MODELLER,[33] taken from the Zhang server[34] CASP9 models or the CASP10 LEEcon models; however, the restraints can be generated from any reference structures.

The standard deviations of the residue-pair terms ($\sigma_{ij;k}$) were calculated as products of the prediction-confidence scores of residues i and j taken from the MODELLER output or from the accuracy of the LEEcon models (eq 4):

$$\sigma_{ij;k} = \frac{18}{\sqrt{w_d}\,(\mathrm{conf}_{i;k} + \mathrm{conf}_{j;k})} \quad (4)$$

where w_d is a constant weight factor of the distance terms and $\mathrm{conf}_{i;k}$ and $\mathrm{conf}_{j;k}$ are the confidences of the secondary structure predictions of the ith and jth residues, respectively, calculated using the PSIPRED Protein Sequence Analysis Workbench server[35,36] (the values range from 0 to 9, where 0 is the lowest confidence and 9 the highest). The standard deviations of single-residue terms, $\sigma_{\omega_i;k}$, were calculated similarly (eq 5):

$$\sigma_{\omega_i;k} = \mathrm{conf}_k \frac{36}{\sqrt{w_{\omega}}\,\sum_{l=i}^{i+4} \mathrm{conf}_{l;k}} \quad (5)$$

where w_{ω} is the constant weight factor of the angle terms, conf_k is the confidence of model k, and $\mathrm{conf}_{i;k}$ is the confidence of the secondary structure prediction of the ith residue calculated using PSIPRED.

The final energy U is the sum of the UNRES energy (U_{UNRES}) and the restraint-penalty term (U_{restr}), as defined by eq 6:

$$U = U_{UNRES} + U_{restr} \quad (6)$$

Multiplexed Replica Exchange Molecular Dynamics. The conformational search procedure used in this work is based on coarse-grained molecular dynamics (MD) with UNRES developed in our laboratory.[37−40] The equations of motion for the UNRES chain are Langevin dynamics equations because the solvent is implicit in UNRES. Consequently, it contributes to the conservative forces (through the potential of mean force) and gives rise to nonconservative forces that originate in energy exchange of the polypeptide chain with the solvent (the stochastic and friction forces). The velocity-Verlet algorithm[41] with the adaptive multiple time split (A-MTS) modification[40] was used to integrate the equations of motion.

To sample the conformational space more efficiently than by canonical MD, we extended[42,43] the UNRES/MD approach with the replica-exchange MD (REMD)[44] and multiplexing REMD (MREMD)[45] methods. In the REMD method,[44] M canonical MD simulations are carried out simultaneously, each one at a different temperature. After every $m < M$ steps, an exchange of temperatures between neighboring trajectories ($j = i + 1$) is attempted; the decision about whether the exchange will be made is based on the Metropolis criterion, expressed by eq 7, which reflects the fact that the effective energy function depends on temperature:[3]

$$\Delta = [\beta_j U(\mathbf{X}_j;\, \beta_j) - \beta_i U(\mathbf{X}_j;\, \beta_i)] - [\beta_j U(\mathbf{X}_i;\, \beta_j) - \beta_i U(\mathbf{X}_i;\, \beta_i)] \quad (7)$$

where $\beta_i = 1/RT_i$, in which R is the gas constant and T_i is the absolute temperature corresponding to the ith trajectory, and $\mathbf{X}_i$ denotes the variables of the UNRES conformation of the ith trajectory at the attempted exchange point. If $\Delta \leq 0$, T_i and T_j are exchanged; otherwise the exchange is performed with probability $\exp(-\Delta)$. The multiplexed variant of the REMD method (MREMD)[43,45] differs from the REMD method in that several trajectories are run at a given temperature. Each set of trajectories run at a different temperature constitutes a *layer*. Exchanges are attempted not only within a single layer but also between layers.

Determining the Most Probable Conformational Ensembles from MREMD Simulations. To determine ensemble-average quantities (specific heat, energy, RMSD from the experimental structure, etc.) as well as the conformational ensembles, the weighted histogram analysis method (WHAM)[46] has been implemented.[3] In particular, WHAM together with clustering of conformations is used to determine candidate predictions. First, the melting temperature (T_m) is determined, and subsequently, a cluster analysis of conformations is carried out at temperature $T = T_\mathrm{m} - 10$ K. To save time, only such conformations whose contributions together make a fraction of 0.99 of the partition function are considered in clustering. After clustering is accomplished, the fractions (or probabilities) of families in the conformational ensemble at the selected temperature, P_i, where i ranges from 1 to the number of families, are computed from eq 8:

$$P_i = \sum_{k\in\{i\}} p_k \tag{8}$$

in which

$$p_k = \frac{\exp[\omega_k - U_k/RT]}{\sum_{k=1}^{N} \exp[\omega_k - U_k/RT]} \tag{9}$$

where $\{i\}$ denotes the set of conformations that belong to family i, U_k denotes the UNRES energy of conformation k, and ω_k is the logarithm of the weight factor of conformation k determined by WHAM from MREMD simulation (see eq 15 of ref 3). The families are then ranked according to decreasing values of P_i. For each family, the weighted-average structure is calculated, the weights being the components of the sum on the right-hand side of eq 8. Then the structure of the family with the lowest RMSD from the average structure is selected as the representative of the family. The reader is referred to our earlier work[3] for more details of the approach.

In this work, we used Ward's minimum-variance clustering.[47,48] The conformational ensemble was dissected into five clusters, and a representative structure was selected from each cluster to give a total of five models. This procedure followed the CASP rules, because only five models per target can be submitted by a given group.

Simulation Details. All of the simulations were carried out with the version of the UNRES force field described in ref 49, which was calibrated with tryptophan cage and tryptophan zipper, augmented with the newly developed $U_{\mathrm{SC\text{-}corr}}$ potentials described in ref 29 with the weight of 0.25.

For each target, a 1,000,000-step MREMD simulation was run initially with the weight of the distance-restraint part of the penalty function (eqs 3 and 4) set at $w_d = 0.001$. The purpose of this simulation with weak distance restraints was to avoid getting stuck in local minima. Subsequently, four separate 5,000,000-step (for single-domain proteins) or 6,000,000-step (for two-domain proteins) production runs were carried out with distance-penalty weights of $w_d = 0.001$, 0.01, 0.1, and 1.0 (eqs 3 and 4). Thus, the length of each simulation was 24.45 or 29.34 ns of UNRES time for single- or two-domain proteins, respectively, which is equivalent to 24.45 or 29.34 μs of effective time, respectively. The weight of the angle-penalty term was $w_\omega = 1.0$ (eqs 3 and 5). As a reference, plain UNRES/MREMD simulations, with restraints only on secondary structure based on PSIPRED predictions, as in the CASP experiments,[7,9] were run. For single-domain proteins, the simulations were run for 20,000,000 steps (97.8 ns of UNRES time, which is equivalent to 97.8 μs of effective time). For two-domain proteins, the simulations were run for 25,000,000 steps (122.25 ns of UNRES time, which is equivalent to 122.25 μs of effective time).

For two-domain proteins, the plain UNRES runs were carried out only for the targets that were not treated by the Cornell−Gdansk group in CASP10; otherwise, the results of the simulations performed during CASP10 were utilized. The simulations were run at the following 32 temperatures: 250, 255, 260, 265, 270, 275, 280, 285, 290, 295, 300, 305, 310, 315, 320, 325, 330, 335, 340, 345, 350, 360, 380, 390, 400, 410, 420, 430, 440, 460, 480, and 500 K. It should be noted that the purpose of using a wide range of temperatures in MREMD simulations is to explore the conformational space efficiently. The UNRES-simulated folding transition usually occurs at 310−350 K; therefore, the conformational ensembles from which candidate predictions were selected were usually collected at 280−320 K. Two trajectories were run per temperature, constituting a total of 64 trajectories. Simulations were run using the A-MTS algorithm[40] with a major time step of 4.89 fs, and replicas were exchanged every 20,000 MD steps. For single-domain proteins (see Single-Domain Proteins), all of the simulations were started from the fully extended structures. For two-domain proteins (see Multidomain Proteins), the simulations for a given target were started from either the extended structure or the LEEcon model 1 structure.[17] The conformations from the final 4,000,000 steps of the simulations were processed by WHAM[3,46] and clustered[47,48] as described above in Determining the Most Probable Conformational Ensembles from MREMD Simulations".

For single-domain proteins, the RMSD over the C$^\alpha$ atoms was calculated to measure the structural similarity between the model and its experimental structure. For two-domain proteins, the global distance test (GDT)[50,51] was used as another measure of similarity along with the RMSD. GDT measures the percentage of the C$^\alpha$ atoms of a model that are within a given RMSD cutoff value from the optimal superposition with the experimental structure. Three kinds of C$^\alpha$ RMSDs were used: the minimum RMSD over a given simulation for a given target ($\rho^{\min}$), the average of the 100 lowest RMSD values over a simulation for a given target ($\rho^{\min}_{100}$), and the lowest of the average RMSDs over the five clusters of conformations obtained from the WHAM and cluster analysis (see Determining the Most Probable Conformational Ensembles from MREMD Simulations); the last quantity, $\langle\rho\rangle^{\min}_{\mathrm{clust}}$, can be considered as the average RMSD over the most nativelike cluster. These quantities are defined as follows:

$$\rho^{\min} = \min_i \rho_i \tag{10}$$

$$\rho^{\min}_{100} = \frac{1}{100} \sum_{\substack{k=1 \\ \rho_{i_k} < \rho_{i_{k+1}}}}^{100} \rho_{i_k} \tag{11}$$

$$\langle\rho\rangle^{\min}_{\mathrm{clust}}(T_\mathrm{a}) = \min_I \sum_{i\in I} \rho_i p_i(T_\mathrm{a}) \tag{12}$$

where ρ_i is the C$^\alpha$ RMSD of the ith conformation, i_k (k = 1, 2, ..., 100) are the indices of conformations sorted according to increasing C$^\alpha$ RMSD, and $p_i(T_\mathrm{a})$ is the weight of the ith conformation (eq 9) calculated at the temperature of analysis of the conformational ensembles (T_a).

RESULTS AND DISCUSSION

Single-Domain Proteins. Before the extent to which the new method can improve template-based predictions could be assessed, it had to be determined whether the method recovers template-based models from the distance and virtual-bond dihedral angle restraints incorporated into the penalty function defined by eq 3. For this purpose, five single-domain targets from the CASP9 experiment with sizes ranging from 54 to 105 amino acid residues whose complete structures were accurately predicted by template-based methods (T0538, T0539, T0559, T0560, and T0580) were selected. The PDB codes of the respective experimental structures (revealed after CASP9) are listed in Table 1. Two independent series of calculations with different sets of restraints were carried out. One set was calculated from template-based models generated with MODELLER,[33] which provides models for complete sequences. The second set was calculated from the models submitted to CASP9 by the Zhang server,[34] which scored best in that experiment. However, these models corresponded to truncated sequences (Table 2).

Table 1. Templates Used To Create the Models of the Selected Targets from the CASP9 Experiment

	PDB codes	
target number	experimental structure	templates
T0538	2L09	2KRU
T0539	2L0B	2JMD, 1X4J, 1IYM, 3HCU, 1V87
T0559	2L01	1QBJ, 1LVA, 1XMK
T0560	2L02	1QBJ, 2V9V, 1DPU
T0580	3NBM	1E2B, 1H9C, 2WY2

Table 2. Lengths of the Sequences of the Selected CASP Targets

	numbers of residues			
target ID	CASP sequence	experimental structure	MODELLER model	Zhang server model
T0538	54	62	54	53 (2−54)
T0539	81	81	81	68 (14−81)
T0559	69	69	69	67 (3−69)
T0560	74	74	74	64 (3−66)
T0580	105	104 (2−105)	105	104 (2−105)

Five models of each protein were generated by MODELLER. The experimental structures of the target proteins (which became available after CASP9) were removed from the database from which the templates were selected. The templates selected by MODELLER to generate the final models are listed in Table 1.

As can be seen from Figure 2, the lowest average RMSD of the most nativelike cluster [$\langle\rho\rangle^{\min}_{\text{clust}}(T_{\text{a}})$ of eq 12] and the average values of the 100 lowest RMSDs ($\rho^{\min}_{100}$ of eq 11) are achieved using the highest distance-penalty weight of w_d = 1.0 (eqs 3 and 4). Because cluster representatives are selected as candidate predictions, the use of stronger restraints seems to be a better choice, at least for single-domain proteins. The RMSD values and fractions of each cluster for test proteins from CASP9 are shown in Table S1 in the Supporting Information.

It can also be seen from Figure 2 that $\rho^{\min}$ decreases with weaker restraints. This is understandable because with weaker restraints the conformational space can be explored more extensively, including the low-RMSD regions. However, these better-quality structures cannot be selected without a prior knowledge of the experimental structure unless a model quality assessment system could be developed, which remains as a challenging problem to solve in the future.

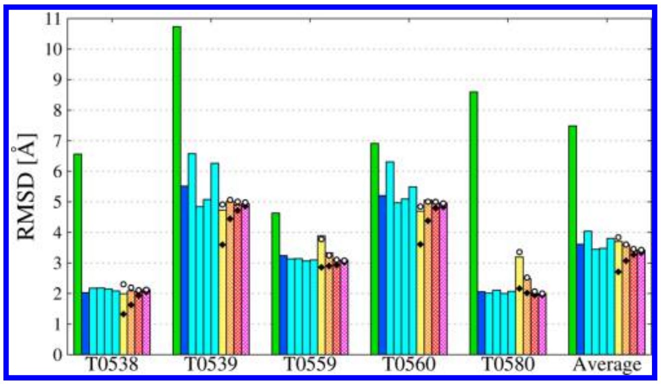


Figure 2. Bar diagrams of the RMSDs of the most nativelike clusters [$\langle\rho\rangle^{\min}_{\text{clust}}(T_{\text{a}})$ of eq 12] for the five targets from the CASP9 exercise. Green bars: plain UNRES simulations with restraints only on virtual-bond dihedral angles obtained by secondary-structure prediction. Blue bars: model 1 from the MODELLER program.[33] Cyan bars: models 2−5 from the MODELLER program. Bars with dashed-line pattern: combined method with w_d = 0.001 (yellow), w_d = 0.01 (orange), w_d = 0.1 (red), or w_d = 1.0 (purple). Black diamonds: $\rho^{\min}$. Small circles: $\rho^{\min}_{100}$. The numbers and populations of the clusters corresponding to the models generated in this work are given in Table S1 in the Supporting Information. The numbers of the models from the MODELLER program are given according to the order in which they were created. The RMSD values were calculated for complete structures.

It can also be seen from Figure 2 that the simulations with the strongest distance restraints (w_d = 0.1 and 1.0) produce structures with resolution comparable to that of the best of the models used to create the restraints in terms of $\rho^{\min}$, $\rho^{\min}_{100}$, and $\langle\rho\rangle^{\min}_{\text{clust}}(T_{\text{a}})$. This resolution is better than that of the models obtained by simple averaging of the structures of the five models used to compute the restraints. Thus, for the TBM targets, UNRES does not produce models resulting from averaging the restraints from the input models but rather the best consistent models, because of its ability to pack the elements of secondary and supersecondary structure correctly.

The results of the calculations carried out with the restraints calculated from the Zhang server models[34] submitted to the CASP9 experiment (prior to the release of the respective experimental structures) are shown in Figure 3. Because the Zhang server models were not assigned confidence scores (even though, qualitatively, they were ranked from the most to the least reliable), the restraints from these models were weighted equally in eq 3. Moreover, most of the Zhang server models corresponded to truncated sequences (Table 2); therefore, restraints were imposed only on the sections of the structures modeled by the Zhang server, while the structure of the rest of the protein was unrestricted.

It can be seen in Figure 3 that the Zhang server models from which the restraints were calculated are of diverse quality. For models 1−5 of T0559, the RMSD values are 1.669, 1.703, 1.906, 2.832, and 10.242 Å, respectively. However, even though the restraints from these models were weighted equally in eq 3, the models of this target obtained by using UNRES simulations with restraints have RMSDs of 1.767 and 1.722 Å from the

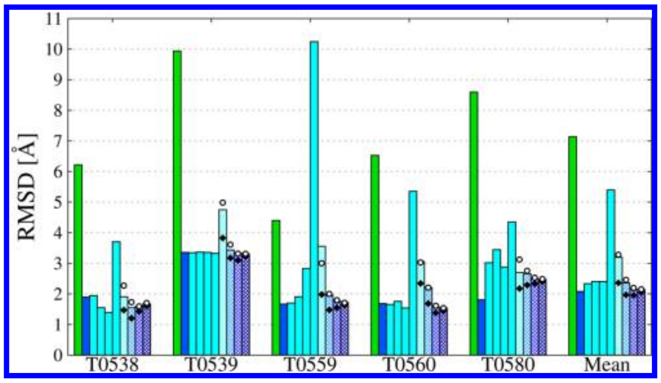


Figure 3. Bar diagrams of the RMSDs of the most nativelike clusters [$\langle\rho\rangle_{\text{clust}}^{\text{min}}(T_{\text{a}})$ of eq 12] for the five targets from the CASP9 exercise. Green bars: plain UNRES simulations with restraints imposed only on virtual-bond dihedral angles obtained by secondary-structure prediction. Blue bars: model 1 submitted by the Zhang server group during CASP9.[34] Cyan bars: models 2−5 submitted by the Zhang server group during CASP9. Bars with dashed-line pattern: the combined method with w_d = 0.001 (cyan), w_d = 0.01 (light blue), w_d = 0.1 (blue), or w_d = 1.0 (dark blue). Black diamonds: ρ^{min}. Small circles: ρ_{100}^{min}. The numbers and populations of the clusters corresponding to the models generated in this work are given in Table S2 in the Supporting Information. The RMSD values were calculated for truncated sequences corresponding to the Zhang server models (Table 2).

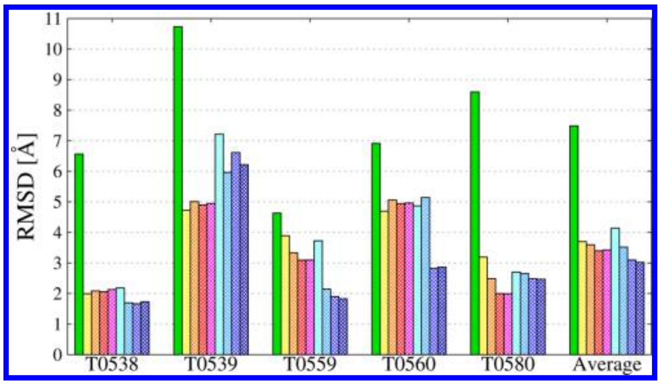


Figure 4. Bar diagrams of the RMSDs of the most nativelike clusters [$\langle\rho\rangle_{\text{clust}}^{\text{min}}(T_{\text{a}})$ of eq 12] for the five targets from the CASP9 exercise. Green bars: plain UNRES simulations with restraints imposed only on virtual-bond dihedral angles obtained by secondary-structure prediction. Bars with dashed-line pattern: the combined method with restraints calculated from the MODELLER[33] models with w_d = 0.001 (yellow), w_d = 0.01 (orange), w_d = 0.1 (red), or w_d = 1.0 (purple). Bars with dashed-line pattern: the combined method with restraints calculated from the Zhang server models[34] with w_d = 0.001 (cyan), w_d = 0.01 (light blue), w_d = 0.1 (blue), or w_d = 1.0 (dark blue). The RMSD values were calculated for complete structures.

experimental structure (with the weights of the distance restraints equal to 0.1 and 1.0, respectively). Similar results were obtained for all of the other proteins except T0580, for which Zhang server models 2−5 had significantly higher RMSDs from the experimental structure than Zhang server model 1 (by up to 2.54 Å). Because the restraints from models 2−5 had a greater summary contribution than those from model 1, the structures resulting from the combined method also had a higher RMSD (by about 0.7 Å) compared with Zhang server model 1.

The results obtained with the restraints calculated from the MODELLER and Zhang server models, respectively, are compared in Figure 4. Here the RMSDs were computed over the complete structures and not only over those sections of the structures that were predicted by the Zhang server. As shown in Figure 4, the structures obtained with the restraints calculated from the Zhang server models have RMSDs lower by about 0.4 Å compared with those from the MODELLER-generated restraints. However, for T0539, for which residues 1−13 are absent in the Zhang server models (Table 2), significantly better results were obtained with the MODELLER-generated restraints.

In summary, it can be concluded that for good TBM targets, the proposed method is able to retrieve structures close to the best models from which the restraints were derived.

Multidomain Proteins. The following four two-domain proteins from the CASP10 experiment, for which single domains were predicted with high accuracy by knowledge-based methods but none of these methods were able to predict correct domain packing, were selected to test the combined method proposed in this work: T0651 (PDB code 4F67), T0663 (PDB code 4EXR), T0717 (PDB code 4H0A), and T0724 (PDB code 4FMR). We used a similar procedure as for the single-domain proteins, but the simulations were started from both extended structures and the model 1 structures submitted by LEEcon during CASP10.[17] The reasons for using the LEEcon group models were that these were consensus models derived from models produced by various servers and that they corresponded to complete structures. Overall, LEEcon scored second after Zhang (see Table 2 and Figure 4B of ref 19); however, for T0663, for example, it scored the best, and for the other two-domain proteins considered in this work, its scores were comparable to those of Zhang.

In the first step, 1 million MREMD steps were run to relax the structure, and then longer (6 million steps, equal to 29.34 ns of UNRES time, which is equivalent to 29.34 μs of effective time) production MREMD simulations were run. Restraints were imposed only on the domain level; no interdomain distance restraints were applied. For each protein, domains were identified by visual inspection of model 1 from the LEEcon group. The domain compositions are listed in Table 3.

Table 3. Domain Compositions of the Four Two-Domain CASP10 Targets

target number	domain 1	domain 2
T0651	1−92	115−254
T0663	1−123	147−205
T0717	1−193	225−345
T0724	1−126	144−265

As can be seen from Figure 5, the application of UNRES along with intradomain distance restraints from the respective LEEcon models (but with interdomain distances unrestricted) results in a significant improvement of $\langle\rho\rangle_{\text{clust}}^{\text{min}}(T_{\text{a}})$ compared with plain UNRES simulations. The RMSD values of the models calculated using the combined approach are also lower than those corresponding to the parent models (Figure 5). The RMSD values and fractions of each cluster for test proteins from CASP10 are shown in Table S2 in the Supporting Information. An analysis of the $\langle\rho\rangle_{\text{clust}}^{\text{min}}(T_{\text{a}})$ values for individual domains of the proteins considered in this section shows that these values are comparable to those corresponding to the parent models (Figure 6). This result demonstrates that the improvement in the models obtained using our UNRES-based

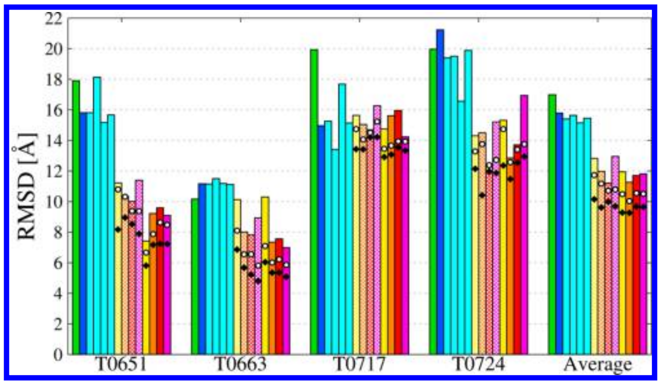


Figure 5. Bar diagrams of the RMSDs of the most nativelike clusters [$\langle\rho\rangle_{\text{clust}}^{\text{min}}(T_a)$ of eq 12] for the four two-domain proteins from the CASP10 exercise. Green bars: plain UNRES simulations with restraints imposed only on secondary structure elements (for T0663 and T0717, the results were taken directly from the models submitted by the Cornell−Gdansk group[9]). Blue bars: model 1 of the LEEcon group submitted during the CASP10 experiment.[17] Cyan bars: models 2−5 of the LEEcon group submitted during the CASP10 experiment.[17] Bars with dashed-line pattern: the combined method with w_d = 0.001 (yellow), w_d = 0.01 (orange), w_d = 0.1 (red), or w_d = 1.0 (purple), in which the simulations were started from the extended structure of the corresponding protein. Solid bars in the rightmost section of each target: the combined method with w_d = 0.001 (yellow), w_d = 0.01 (orange), w_d = 0.1 (red), or w_d = 1.0 (purple), in which the simulations were started from the LEEcon model 1 of a given protein. Black diamonds: ρ^{min}. Small circles: ρ_{100}^{min}. The numbers and populations of the clusters corresponding to the models generated in this work are given in Table S3 in the Supporting Information.

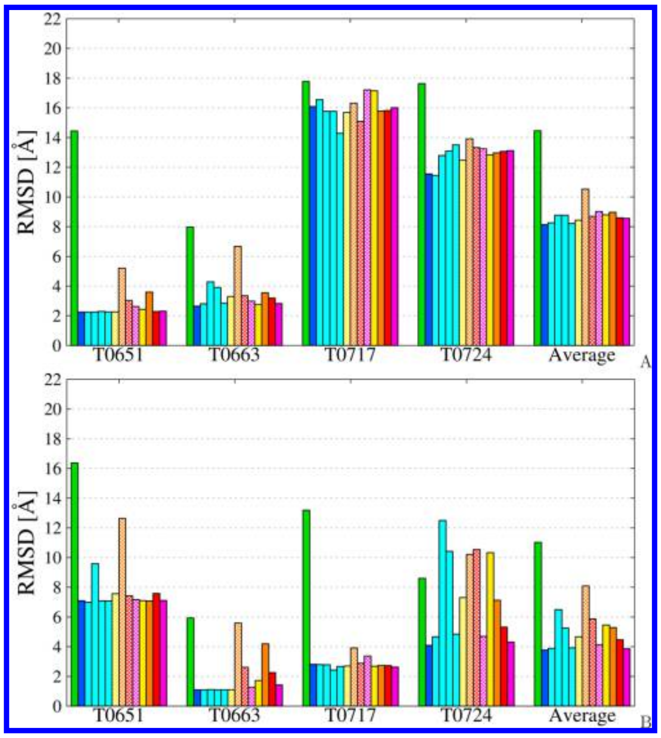


Figure 6. Bar diagrams of the RMSDs of the most nativelike clusters [$\langle\rho\rangle_{\text{clust}}^{\text{min}}(T_a)$ of eq 12] for individual domains of the four two-domain proteins from CASP10: (A) the N-terminal domain (domain 1); (B) the C-terminal domain (domain 2). The meanings of bar colors, patterns, and symbols are the same as in Figure 5.

methodology results from the ability of the UNRES force field to find correct domain packing, which was the assumption of the method proposed in this work.

It can be observed from Figure 5 that lower $\langle\rho\rangle_{\text{clust}}^{\text{min}}(T_a)$ values were obtained in the simulations started from the structures of the LEEcon group models compared with those started from the extended structures. This difference can be explained by the fact that the simulations started from the LEEcon group models already have premodeled domains (the structures of which are well-predicted by knowledge-based methods), and therefore, simulation time is not spent on domain formation but only on domain packing. By inspection of the values of ρ^{min} and ρ_{100}^{min}, it can be concluded that the simulations started from the structure of model 1 reached convergence faster. In contrast to the results obtained for the single-domain proteins, rather small differences between ρ^{min} and ρ_{100}^{min} were observed for the results of the simulations started from extended structures compared with those started from LEEcon group model 1 for the distance-restraint weights (eqs 3 and 4) of w_d = 0.1 and w_d = 0.01, which produce the most nativelike clusters of structures. The convergence is a much more significant issue for the proteins considered in this section compared with the single-domain proteins considered in Single-Domain Proteins, which are smaller. It should be noted that for the strongest distance restraints (w_d = 1.0), many of the trajectories started from the extended structures got stuck and did not produce any nativelike structures. The best average results were obtained with w_d = 0.1 and w_d = 0.01 for both the simulations started from the extended structure and those started from LEEcon group model 1.

The GDT plots corresponding to the results obtained in this work and those obtained in CASP10 for the two-domain proteins considered here are shown in Figure 7. As can be seen from the figure, for three of the four proteins considered in this study (T0651, T0663, and T0724), the GDT plots show significantly increased percentages of residues within a given RMSD cutoff value from the corresponding experimental structure compared with those of the models submitted in CASP10. It should be noted that no information on the experimental structures of the four proteins was utilized in our simulations; all of the knowledge-based information was taken directly from the LEEcon group models, which were submitted to CASP10 before the experimental structures were disclosed.

Figures 8 and 9 show the best models of targets T0651 and T0663 obtained in this work, along with their corresponding experimental structures and the best LEEcon-group models. The domain orientation of the best model from this work for T0651 (Figure 8c) is almost identical to that of the experimental structure (Figure 8b). The largest structural discrepancy arises from the C-terminal domain, as it was not as accurately predicted by the knowledge-based method.

In the CASP10 exercise, using our approach based on the UNRES force field, we predicted correct domain packing of T0663.[9] Two of our models for this target were featured by the CASP assessor because of their correct domain-packing topology. However, our models had poor resolution of individual domains and therefore were not so well distinguishable from the other predictions on the basis of the GDT plot alone. As can be seen from Figure 7, the models obtained with UNRES assisted by knowledge-based information are distinguishable in the GDT plots. The method proposed in this work can not only produce the global fold properly (Figure 9) but also predict individual domains with high accuracy (Figure 6).

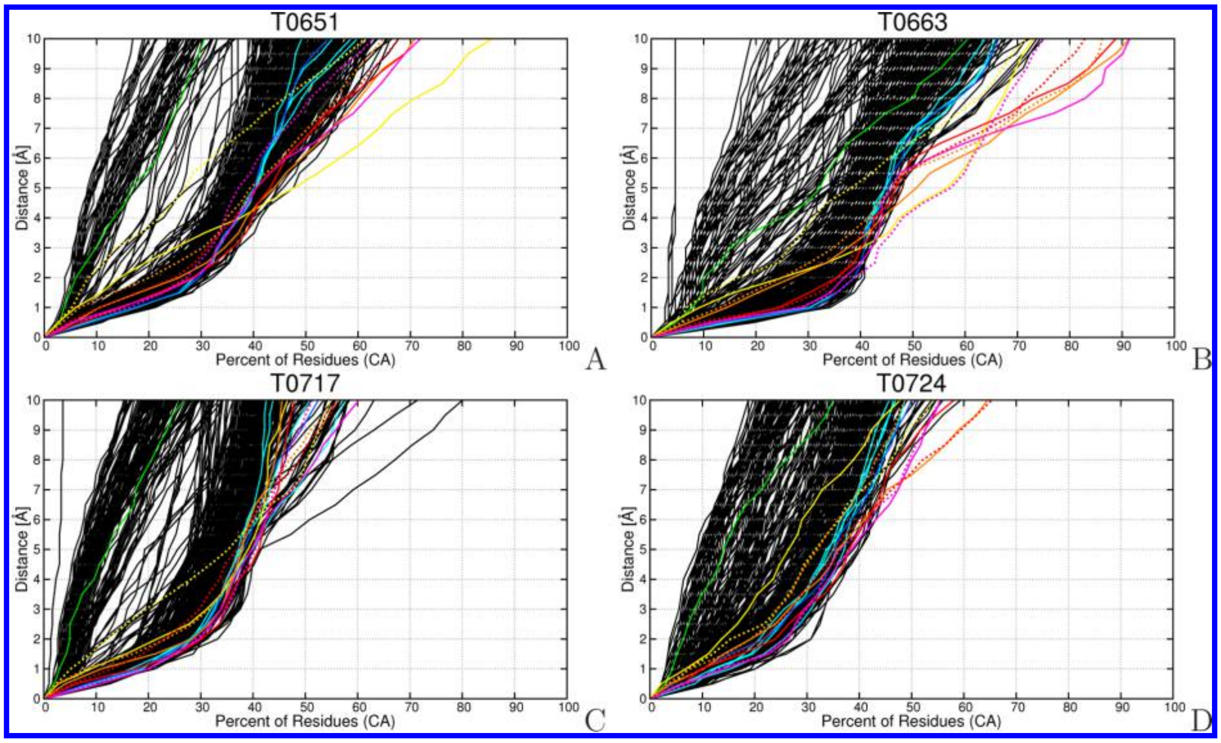


Figure 7. Global Distance Test (GDT) plots for the models of the CASP10 targets (A) T0651, (B) T0663, (C) T0717, and (D) T0724. Black lines: models submitted by all groups except for Cornell−Gdansk and LEEcon. Green lines: models obtained by use of plain UNRES with secondary-structure restraints; for T0663 and T0717, the models from the Cornell−Gdansk group[9] submitted during the CASP exercise were used. Blue lines: model 1 from the LEEcon group submitted during CASP10.[17] Cyan lines: models 2−5 from the LEEcon group submitted during CASP10.[17] Dashed yellow, orange, red, and purple lines: models obtained by the combined approach used in this work with UNRES simulations started from the extended structures, with w_d = 0.001, 0.01, 0.1, and 1.0, respectively. Solid yellow, orange, red, and purple lines: as above but for simulations of a given protein started from model 1 of the LEEcon group. The "all group" parts of the plots (black lines) were taken from the CASP10 site (http://www.predictioncenter.org/casp10/results.cgi), while the plots for the LEEcon group models, plain UNRES models, and the models resulting from the combined approach proposed in this work were calculated with the LGA server (http://proteinmodel.org/AS2TS/LGA/lga.html).

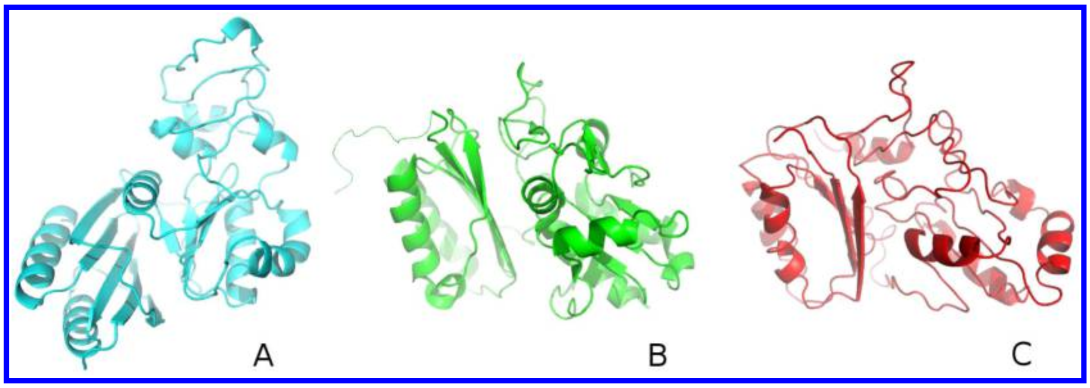


Figure 8. Cartoon representations of structures of T0651: (A) model 1 submitted by the LEEcon group during CASP10;[17] (B) the experimental structure (PDB code 4F67); (C) the best model from the combined method (with the weight of the distance-restraint part of the penalty function equal to 0.001, cluster 1).

It should also be noted that some discrepancy in the domain packing between the calculated and the experimental structure of T0663 arises from the interference of the C-terminal α-helix, which is absent in the experimental structure.

■ CONCLUSIONS

We have proposed a combined approach to protein structure prediction in which knowledge-based information is used in MREMD simulations with the UNRES force field in the form of restraints calculated from models obtained by the use of bioinformatics tools, mainly template-based modeling. The tests with single-domain TBM targets demonstrated that proper usage of the template information in the physics-based UNRES force field results in structures comparable in quality to the best of the parent models. This result can be attributed to the ability of UNRES to pack elements of secondary and supersecondary structures correctly and not to simple averaging over the input models.

The added value of UNRES in the combined approach proposed in this work was demonstrated in tests with two-domain proteins. For three of the four two-domain targets from the CASP10 exercise (T0651, T0663, and T0724), use of the

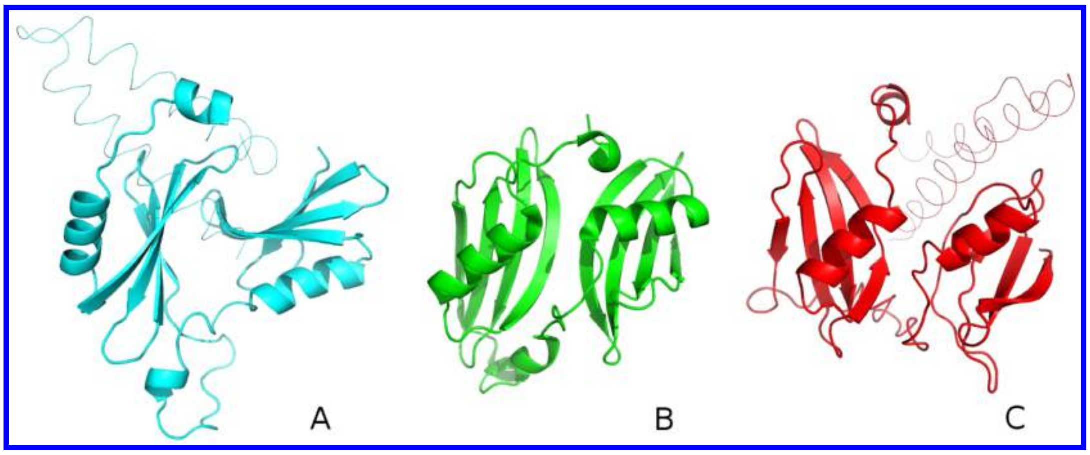


Figure 9. Cartoon representations of structures of T0663: (A) model 1 submitted by the LEEcon group during CASP10;[17] (B) the experimental structure (PDB code 4EXR); (C) the best structure after clustering from the combined method (with the weight of the distance-restraint part of the penalty function equal to 1.0, cluster 1).

combined approach resulted in models with correct domain packing and, consequently, significantly lower RMSDs from the experimental structures (Figure 5) and GDT plots extending more to the right side (Figure 7) than for the models submitted in the CASP10 exercise. Consistent with the results obtained for single-domain proteins, the accuracy of the modeling of individual domains is high because the method takes advantage of knowledge-based information (Figure 6). This is a major improvement over using plain UNRES, which at the present stage of development has the ability to pack domains correctly but generates individual domains at only medium resolution.[9]

On the basis of the results obtained in this work, the combined approach can offer a considerable advantage for modeling of the structures of those proteins whose compact sections (domains) but not their complete structures are well-predicted by TBM approaches. However, even when the overall similarity score of a sequence is high, and consequently, pure TBM models are assessed as very reliable, there still might be linker regions predicted with significantly lower accuracy, which can result in wrong domain packing (e.g., target T0663 from CASP10). In such situations, the combined approach can improve the result significantly over using TBM alone.

Further development of the approach will include imposing restraints not only on the $C^{\alpha}\cdots C^{\alpha}$ distances and backbone virtual-bond dihedral angles γ but also including the backbone virtual-bond valence angles θ and the side-chain center locations with respect to the backbone (Figure 1), which should improve the quality of the local structure. The currently used "multiharmonic" restraint function (eq 3) can be replaced with a more ergodic one (e.g., Lorenzian[17]). Finally, a more sophisticated method to determine the weights of the distance restraints on the basis of the alignment confidence of the corresponding residue pairs, which will take into account the alignment of the whole sequence fragment between the interacting residues, has to be designed. This work is currently being carried out in our laboratory.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

RMSD values and fractions of clusters for the nine selected CASP9 targets obtained by using the method developed in this work with restraints derived from TBM models obtained by using MODELLER (Table S1); RMSD values and fractions of clusters for the nine selected CASP9 targets obtained by using the method developed in this work with restraints derived from the Zhang server models submitted to CASP9 (Table S2); and RMSD values and fractions of clusters for the four selected CASP10 targets obtained by using the method developed in this work with restraints derived from the models submitted by the LEEcon group to CASP10 (Table S3). The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00117.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +48 58 523 5124. Fax: +48 58 523 5012. E-mail: adam@sun1.chem.univ.gda.pl.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Dr. Adam K. Sieradzan for helpful assistance with the UNRES code. This work was supported by Grant DEC-2013/10/M/ST4/00640 from the National Science Center of Poland (Narodowe Centrum Nauki), and from the Foundation for Polish Science [Grant Mistrz 7./2013 and Grant MPD/2010/5, which is a project operated within the Foundation for Polish Science International Ph.D. Projects (MPD) Programme cofinanced by the EU European Regional Development Fund, Operational Program Innovative Economy 2007–2013]. K.J. and J.L. were supported by a National Research Foundation of Korea (NRF) grant funded by the Government of Korea (MEST) (2008-0061987). Computational resources were provided by the supercomputer resources at the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdańsk, the supercomputer resources at the Interdisciplinary Center of Mathematical and Computer Modeling (ICM) at the University of Warsaw, the 952-processor Beowulf cluster of the Baker Laboratory of Chemistry at Cornell University, our 528-processor Beowulf cluster at the Faculty of Chemistry of the University of Gdańsk, and computing resources at the KIAS Center of Advanced Computations.

■ REFERENCES

(1) Grishin, N. V. Predictive Landscape of CASP. Presented at the CASP10 meeting. http://www.predictioncenter.org/casp10/doc/presentations/CASP10_Keynote_NG.pdf (accessed April 28, 2015) .

(2) Lee, J.; Liwo, A.; Ripoll, D. R.; Pillardy, J.; Scheraga, H. A. Calculation of Protein Conformation by Global Optimization of a Potential Energy Function. *Proteins: Struct., Funct., Genet.* **1999**, *37* (Suppl. 3), 204−208.

(3) Liwo, A.; Khalili, M.; Czaplewski, C.; Kalinowski, S.; Ołdziej, S.; Wachucik, K.; Scheraga, H. A. Modification and Optimization of the United-Residue (UNRES) Potential Energy Function for Canonical Simulations. I. Temperature Dependence of the Effective Energy Function and Tests of the Optimization Method with Single Training Proteins. *J. Phys. Chem. B* **2007**, *111*, 260−285.

(4) Liwo, A.; Czaplewski, C.; Ołdziej, S.; Rojas, A. V.; Kaźmierkiewicz, 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.

(5) Lee, J.; Liwo, A.; Scheraga, H. A. Energy-Based *de Novo* Protein Folding by Conformational Space Annealing and an Off-Lattice United-Residue Force Field: Application to the 10−55 Fragment of Staphylococcal Protein A and to Apo Calbindin D9K. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 2025−2030.

(6) Pillardy, J.; Czaplewski, C.; Liwo, A.; Lee, J.; Ripoll, D. R.; Kaźmierkiewicz, R.; Oldziej, S.; Wedemeyer, W. J.; Gibson, K. D.; Arnautova, Y. A.; Saunders, J.; Ye, Y.-J.; Scheraga, H. A. Recent Improvements in Prediction of Protein Structure by Global Optimization of a Potential Energy Function. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 2329−2333.

(7) 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. Physics-Based Protein-Structure Prediction Using a Hierarchical Protocol Based on the UNRES Force Field: Assessment in Two Blind Tests. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7547−7552.

(8) Liwo, A.; He, Y.; Scheraga, H. A. Coarse-Grained Force Field: General Folding Theory. *Phys. Chem. Chem. Phys.* **2011**, *13*, 16890−16901.

(9) He, Y.; Mozolewska, M. A.; Krupa, P.; Sieradzan, A. K.; Wirecki, T. K.; Liwo, A.; Kachlishvili, K.; Rackovsky, S.; Jagieła, D.; Ślusarz, R.; Czaplewski, C. R.; Ołdziej, S.; Scheraga, H. A. Lessons from Application of the UNRES Force Field to Predictions of Structures of CASP10 Targets. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 14936−14941.

(10) Khoury, G. A.; Liwo, A.; Khatib, F.; Zhou, H.; Chopra, G.; Bacardit, J.; Bortot, L. O.; Faccioli, R. A.; Deng, X.; He, Y.; Krupa, P.; Li, J.; Mozolewska, M. A.; Sieradzan, A. K.; Smadbeck, J.; Wirecki, T.; Cooper, S.; Flatten, J.; Xu, K.; Baker, D.; Cheng, J.; Delbem, A. C. B.; Floudas, C. A.; Kesar, C.; Levitt, M.; Popović, Z.; Scheraga, H. A.; Skolnick, J.; Crivelli, S. N.; Foldit Players. WeFold: Large-Scale Coopetition for Protein Structure Prediction. *Proteins: Struct., Funct., Bioinf.* **2014**, *82*, 1850−1868.

(11) Subramani, A.; Floudas, C. A. β-Sheet Topology Prediction with High Precision and Recall for β and Mixed α/β Proteins. *PLoS One* **2012**, *7*, No. e32461.

(12) Liwo, A.; Lee, J.; Ripoll, D. R.; Pillardy, J.; Scheraga, H. A. Protein Structure Prediction by Global Optimization of a Potential Energy Function. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 5482−5485.

(13) Lee, J.; Scheraga, H. A.; Rackovsky, S. Conformational Analysis of the 20-Residue Membrane-Bound Portion of Melittin by Conformational Space Annealing. *Biopolymers* **1998**, *46*, 103−115.

(14) Lee, J.; Scheraga, H. A. Conformational Space Annealing by Parallel Computations: Extensive Conformational Search of Met-Enkephalin and of the 20-Residue Membrane-Bound Portion of Melittin. *Int. J. Quantum Chem.* **1999**, *75*, 255−265.

(15) Joo, K.; Lee, J.; Lee, S.; Seo, J.-H.; Lee, S. J.; Lee, J. High Accuracy Template Based Modeling by Global Optimization. *Proteins: Struct., Funct., Bioinf.* **2007**, *69* (Suppl. 8), 83−89.

(16) Joo, K.; Lee, J.; Kim, I.; Lee, S. J.; Lee, J. Multiple Sequence Alignment by Conformational Space Annealing. *Biophys. J.* **2008**, *95*, 4813−4819.

(17) Joo, K.; Lee, J.; Sim, S.; Lee, S. Y.; Lee, K.; Heo, S.; Lee, I.-H.; Lee, S. J.; Lee, J. Protein Structure Modeling for CASP10 by Multiple Layers of Global Optimization. *Proteins: Struct., Funct., Bioinf.* **2014**, *82*, 188−195.

(18) Krieger, E.; Joo, K.; Lee, J.; Lee, J.; Raman, S.; Thompson, J.; Tyka, M.; Baker, D.; Karplus, K. Improving Physical Realism, Stereochemistry, and Side-Chain Accuracy in Homology Modeling: Four Approaches That Performed Well in CASP8. *Proteins: Struct., Funct., Bioinf.* **2009**, *77* (Suppl. 9), 114−122.

(19) Huang, Y. J.; Mao, B.; Aramini, J. M.; Montelione, G. T. Assessment of Template-Based Protein Structure Predictions in CASP10. *Proteins: Struct., Funct., Bioinf.* **2014**, *82*, 43−56.

(20) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. Prediction of Protein Conformation on the Basis of a Search for Compact Structures; Test on Avian Pancreatic Polypeptide. *Protein Sci.* **1993**, *2*, 1715−1731.

(21) Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A. Cumulant-Based Expressions for the Multibody Terms for the Correlation between Local and Electrostatic Interactions in the United-Residue Force Field. *J. Chem. Phys.* **2001**, *115*, 2323−2347.

(22) Liwo, A.; Ołdziej, S.; Czaplewski, C.; Kozłowska, U.; Scheraga, H. A. Parameterization of Backbone-Electrostatic and Multibody Contributions to the UNRES Force Field for Protein-Structure Prediction from Ab Initio Energy Surfaces of Model Systems. *J. Phys. Chem. B* **2004**, *108*, 9421−9438.

(23) Liwo, A.; Baranowski, M.; Czaplewski, C.; Gołas, E.; He, Y.; Jagieła, D.; Krupa, P.; Maciejczyk, M.; Makowski, M.; Mozolewska, M. A.; Niadzvedtski, A.; Ołdziej, S.; Scheraga, H. A.; Sieradzan, A. K.; Ślusarz, R.; Wirecki, T.; Yin, Y.; Zaborowsk, B. A Unified Coarse-Grained Model of Biological Macromolecules Based on Mean-Field Multipole−Multipole Interactions. *J. Mol. Model.* **2014**, *20*, 2306.

(24) Kubo, R. Generalized Cumulant Expansion Method. *J. Phys. Soc. Jpn.* **1962**, *17*, 1100−1120.

(25) Kolinski, A.; Godzik, A.; Skolnick, J. A General Method for the Prediction of the Three-Dimensional Structure and Folding Pathway of Globular Proteins: Application to Designed Helical Proteins. *J. Chem. Phys.* **1993**, *98*, 7420−7433.

(26) Liwo, A.; Ołdziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. A United-Residue Force Field for Off-Lattice Protein-Structure Simulations. I. Functional Forms and Parameters of Long-Range Side-Chain Interaction Potentials from Protein Crystal Data. *J. Comput. Chem.* **1997**, *18*, 849−873.

(27) Ołdziej, S.; Kozłowska, U.; Liwo, A.; Scheraga, H. A. Determination of the Potentials of Mean Force for Rotation about $C^\alpha \cdots C^\alpha$ Virtual Bonds in Polypeptides from the ab Initio Energy Surfaces of Terminally-Blocked Glycine, Alanine, and Proline. *J. Phys. Chem. A* **2003**, *107*, 8035−8046.

(28) Krupa, P.; Sieradzan, A. K.; Rackovsky, S.; Baranowski, M.; Ołdziej, S.; Scheraga, H. A.; Liwo, A.; Czaplewski, C. Improvement of the Treatment of Loop Structures in the UNRES Force Field by Inclusion of Coupling between Backbone- and Side-Chain-Local Conformational States. *J. Chem. Theory Comput.* **2013**, *9*, 4620−4632.

(29) Sieradzan, A. K.; Krupa, P.; Scheraga, H. A.; Liwo, A.; Czaplewski, C. Physics-Based Potentials for the Coupling between Backbone- and Side-Chain-Local Conformational States in the United Residue (UNRES) Force Field for Protein Simulations. *J. Chem. Theory Comput.* **2015**, *11*, 817−831.

(30) Kozłowska, U.; Maisuradze, G. G.; Liwo, A.; Scheraga, H. A. Determination of Side-Chain-Rotamer and Side-Chain and Backbone Virtual-Bond-Stretching Potentials of Mean Force from AM1 Energy Surfaces of Terminally-Blocked Amino-Acid Residues, for Coarse-Grained Simulations of Protein Structure and Folding. 2. Results, Comparison with Statistical Potentials, and Implementation in the UNRES Force Field. *J. Comput. Chem.* **2010**, *31*, 1154−1167.

(31) Sieradzan, A. K.; Scheraga, H. A.; Liwo, A. Determination of Effective Potentials for the Stretching of $C^\alpha \cdots C^\alpha$ Virtual Bonds in

Polypeptide Chains for Coarse-Grained Simulations of Proteins from *ab Initio* Energy Surfaces of *N*-Methylacetamide and *N*-Acetylpyrrolidine. *J. Chem. Theory Comput.* **2012**, *8*, 1334−1343.

(32) Mirjalili, V.; Noyes, K.; Feig, M. Physics-Based Protein Structure Refinement through Multiple Molecular Dynamics Trajectories and Structure Averaging. *Proteins: Struct., Funct., Bioinf.* **2014**, *82*, 196−207.

(33) Fiser, A.; Šali, A. MODELLER: Generation and Refinement of Homology-Based Protein Structure Models. *Methods Enzymol.* **2003**, *374*, 463−493.

(34) Dong, X.; Jian, Z.; Ambrish, R.; Zhang, Y. Automated Protein Structure Modeling in CASP9 by I-TASSER Pipeline Combined with QUARK-Based Ab Initio Folding and FG-MD-Based Structure Refinement. *Proteins: Struct., Funct., Bioinf.* **2011**, *79* (Suppl. 10), 147−160.

(35) Jones, D. T. Protein Secondary Structure Prediction Based on Position-Specific Scoring Matrices. *J. Mol. Biol.* **1999**, *292*, 195−202.

(36) 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*, W340−W348.

(37) Khalili, M.; Liwo, A.; Rakowski, F.; Grochowski, P.; Scheraga, H. A. Molecular Dynamics with the United-Residue Model of Polypeptide Chains. I. Lagrange Equations of Motion and Tests of Numerical Stability in the Microcanonical Mode. *J. Phys. Chem. B* **2005**, *109*, 13785−13797.

(38) Khalili, M.; Liwo, A.; Jagielska, A.; Scheraga, H. A. Molecular Dynamics with the United-Residue Model of Polypeptide Chains. II. Langevin and Berendsen-Bath Dynamics and Tests on Model *α*-Helical Systems. *J. Phys. Chem. B* **2005**, *109*, 13798−13810.

(39) Liwo, A.; Khalili, M.; Scheraga, H. A. Ab Initio Simulations of Protein-Folding Pathways by Molecular Dynamics with the United-Residue Model of Polypeptide Chains. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2362−2367.

(40) Rakowski, F.; Grochowski, P.; Lesyng, B.; Liwo, A.; Scheraga, H. A. Implementation of a Symplectic Multiple-Time-Step Molecular Dynamics Algorithm, Based on the United-Residue Mesoscopic Potential Energy Function. *J. Chem. Phys.* **2006**, *125*, No. 204107.

(41) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. A Computer Simulation Method for the Calculation of Equilibrium Constants for the Formation of Physical Clusters of Molecules: Application to Small Water Clusters. *J. Chem. Phys.* **1982**, *76*, 637−649.

(42) Nanias, M.; Czaplewski, C.; Scheraga, H. A. Replica Exchange and Multicanonical Algorithms with the Coarse-Grained United-Residue (UNRES) Force Field. *J. Chem. Theory Comput.* **2006**, *2*, 513−528.

(43) Czaplewski, C.; Kalinowski, S.; Liwo, A.; Scheraga, H. A. Application of Multiplexing Replica Exchange Molecular Dynamics Method to the UNRES Force Field: Tests with *α* and *α*+*β* Proteins. *J. Chem. Theory Comput.* **2009**, *5*, 627−640.

(44) Hansmann, U. H. E.; Okamoto, Y. Comparative Study of Multicanonical and Simulated Annealing Algorithms in the Protein Folding Problem. *Physica A* **1994**, *212*, 415−437.

(45) Rhee, Y. M.; Pande, V. S. Multiplexed-Replica Exchange Molecular Dynamics Method for Protein Folding Simulation. *Biophys. J.* **2003**, *84*, 775−786.

(46) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* **1992**, *13*, 1011−1021.

(47) Murtagh, F. *Multidimensional Clustering Algorithms*; Physica-Verlag: Vienna, 1985.

(48) Murtagh, F.; Heck, A. *Multivariate Data Analysis*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1987.

(49) He, Y.; Xiao, Y.; Liwo, A.; Scheraga, H. A. Exploring the Parameter Space of the Coarse-Grained UNRES Force Field by Random Search: Selecting a Transferable Medium-Resolution Force Field. *J. Comput. Chem.* **2009**, *30*, 2127−2135.

(50) Zemla, A. LGA: A Method for Finding 3D Similarities in Protein Structures. *Nucleic Acids Res.* **2003**, *13*, 3370−3374.

(51) Moult, J.; Fidelis, K.; Kryshtafovych, A.; Schwede, T.; Tramontano, A. Critical Assessment of Methods of Protein Structure Prediction (CASP) Round X. *Proteins: Struct., Funct., Bioinf.* **2013**, *82* (Suppl. 2), 1−6.