Published in final edited form as:

J Chem Theory Comput. 2018 November 13; 14(11): 6102–6116. doi:10.1021/acs.jctc.8b00683.

Template-Guided Protein Structure Prediction and Refinement Using Optimized Folding Landscape Force Fields

Mingchen Chen^{†,‡,¶}, Xingcheng Lin^{†,§,¶}, Wei Lu^{†,§,¶}, Nicholas P. Schafer^{†,¶}, José N. Onuchic^{†,§,¶,⊥}, Peter G. Wolynes^{*,†,¶,⊥}

[†]Center for Theoretical Biological Physics, Rice University, Houston, Texas 77030, United States

[‡]Department of Bioengineering, Rice University, Houston, Texas 77005, United States

[§]Department of Physics and Astronomy, Rice University, Houston, Texas 77005, United States

[¶]Department of Chemistry, Rice University, Houston, Texas 77005, United States

[⊥]Department of Biosciences, Rice University, Houston, Texas 77005, United States

Abstract

When good structural templates can be identified, template-based modeling is the most reliable way to predict the tertiary structure of proteins. In this study, we combine template-based modeling with a realistic coarse-grained force field, AWSEM, that has been optimized using the principles of energy landscape theory. The Associative memory, Water mediated, Structure and Energy Model (AWSEM) is a coarse-grained force field having both transferable tertiary interactions and knowledge-based local-in-sequence interaction terms. We incorporate template information into AWSEM by introducing soft collective biases to the template structures, resulting in a model that we call AWSEM-Template. Structure prediction tests on eight targets, four of which are in the low sequence identity “twilight zone” of homology modeling, show that AWSEM-Template can achieve high-resolution structure prediction. Our results also confirm that using a combination of AWSEM and a template-guided potential leads to more accurate prediction of protein structures than simply using a template-guided potential alone. Free energy profile analyses demonstrate that the soft collective biases to the template effectively increase funneling toward native-like structures while still allowing significant flexibility so as to allow for correction of discrepancies between the target structure and the template. A further stage of refinement using all-atom molecular dynamics augmented with soft collective biases to the structures predicted by AWSEM-Template leads to a further improvement of both backbone and side-chain accuracy by maintaining sufficient flexibility but at the same time discouraging unproductive unfolding events often seen in unrestrained all-atom refinement simulations. The all-atom refinement simulations also reduce patches of frustration of the initial predictions. Some of the backbones found among

^{*}**Corresponding Author:** P. G. Wolynes. pwolynes@rice.edu. Phone: (713) 348-4101.

[¶]These three authors contributed equally to this work.

Supporting Information

The Supporting Information is available free of charge on the [ACS Publications website](https://doi.org/10.1021/acs.jctc.8b00683) at DOI: 10.1021/acs.jctc.8b00683.

Reduced number of highly frustrated contacts in a protein molecule during molecular dynamics refinement, detailed information on the templates used in this work, and parameters used in our simulations (PDF)

The authors declare no competing financial interest.

the structures produced during the initial coarse-grained prediction step already have CE-RMSD values of less than 3 Å with 90% or more of the residues aligned to the experimentally solved structure for all targets. All-atom structures generated during the following all-atom refinement simulations, which started from coarse-grained structures that were chosen without reference to any knowledge about the native structure, have CE-RMSD values of less than 2.5 Å with 90% or more of the residues aligned for 6 out of 8 targets. Clustering low energy structures generated during the initial coarse-grained annealing picks out reliably structures that are within 1 Å of the best sampled structures in 5 out of 8 cases. After the all-atom refinement, structures that are within 1 Å of the best sampled structures can be selected using a simple algorithm based on energetic features alone in 7 out of 8 cases.

Graphical Abstract

1. INTRODUCTION

Proteins are biopolymer chains made from a set of 20 amino acids, which are covalently linked together by peptide bonds. As the key players in various biological functions, proteins are very diverse in structure. A protein's function is often directly related to its three-dimensional native structure. That protein polymers fold spontaneously to organized structures is a remarkable physical phenomenon. The essential paradoxes of how proteins are able to fold have been successfully resolved by energy landscape theory.¹⁻³ The fact that most globular proteins have a definite structure that is kinetically accessible indicates that the energy landscapes of globular proteins are largely funneled toward their native states, a fact that has come to be known as the Principle of Minimal Frustration.⁴ Distilled to mathematical form, this principle states that the native interactions are on average stronger than the possible non-native interactions, which leads to a funneled folding landscape for the protein that guides the protein's Brownian motions toward native-like configurations without encountering many metastable traps.² The necessary mathematical framework based on an analogy to spin glasses and a machine learning formalism quantifies the concept of minimal frustration and thereby also provide algorithms to learn and to optimize structure prediction force fields.^{2,5}

A twin challenge to understanding intellectually how proteins fold, the problem of practically predicting protein structures from sequence, has attracted much attention over the past three decades.⁶ Describing proteins merely as polymers chained by bonds, purely

physicochemical approaches based on atomistic force fields should eventually work for predicting structure from sequence, but owing to their insatiable demand for computer time these approaches have until recently met with few successes in *de novo* structure prediction. Those few successful cases are largely fast-folding proteins whose landscapes are confirmed (after the fact) to be smoothly funneled toward the native state.^{7,8} The minimal frustration principle is only satisfied by evolved (or well-designed) heteropolymers. Recognizing this point, a recent development in protein structure prediction uses coevolutionary information inferred from entire families of protein sequences which have now become available from large-scale genome sequencing.^{9–14} Strong covariation between pairs of residues in protein sequence families often but not always arises as a result of residue pairs making contact in the three-dimensional native structures of the protein family.^{13,14} Despite the presence of some false-positive signals, however, analysis of residue covariation has proven to be very useful for structure prediction tasks.

The most practical method of protein structure prediction today relies on evolution in a different way:³ the construction of homology models from specific templates that already have experimentally determined structures described in the Protein Data Bank (PDB).¹⁵ The still-growing PDB database already appears to cover almost every possible protein fold, and it is unusual for structural biologists today to reveal new structures with novel folds.¹⁶ For a given target sequence, homology modeling, also termed “comparative modeling”, builds 3D structures based on these already determined 3D structural templates. The canonical procedure of homology modeling consists of five steps: (1) finding specific templates related to the target sequence, (2) aligning the target sequence to the template structures, (3) building the structure framework by copying the aligned regions or by satisfying spatial restraints from templates, (4) constructing the unaligned loop regions, and (5) energy minimization.

MODELLER, which uses simultaneous optimization of an objective function combining energies from an all-atom molecular mechanics force field along with spatial restraints that are represented by a probability distribution function derived from templates, has been successful in efficiently building moderate resolution models.¹⁷ Other models, such as TASSER^{18,19} and I-TASSER,^{20,21} employ distance and contact restraints derived from templates as part of their guidance and have performed well in Critical Assessment of Structure Predictions (CASP). While the methods outlined above yield accurate predictions for many proteins, such template-based modeling methods are sometimes unable to produce accurate predictions even for targets which have more than 30% sequence identity with the template. The “twilight zone”²² for template-based modeling refers to targets whose template has a sequence identity of about 20–30%. For proteins in the “twilight zone”, templates can be hard to identify in the first place and, when found, the alignments to the correctly identified templates can be locally incorrect due to the difficulties that are inherent in aligning two dissimilar but nonetheless structurally related sequences. Mixed approaches that employ template information and coarse-grained force fields that are capable of predicting protein structure even in the absence of template information may help to address these problems that arise when using a template from the “twilight zone” where the target structure can actually be locally quite different from its template or where the alignment between the target and the template is ambiguous and possibly incorrect.

In addition to predicting protein structures from scratch, efforts have been undertaken to further refine near-native protein structures predicted by moderate resolution algorithms into more fully native states. The rationale for attempting refinement is that, through the work of a large community, the current protein force fields have already been improved to the point that they now capture the funneled nature of protein folding landscapes.^{8,23} Atomistic force fields are particularly useful for correcting any nonphysical clashes found in a predicted coarse-grained structure. Thus, they can move the coarse-grained model structure toward its proper native configuration. Most refinement attempts model the solvation of the initially predicted structures in a water environment, whether implicitly²⁴ or explicitly,^{25,26} followed by MD simulations with a fully atomistically detailed physical force field. Several different refinement schemes have been tested. Previously, a set of long time-scale refinement simulations performed with the Anton machine²⁵ demonstrated that the refinement with restraints based on the initially predicted structure generally improves the quality of that structure prediction, but, at the same time, simulating the protein motion without employing any such restraints often allows the ensemble to drift away from the proper native structure. In a sense, without constraints, we are asking the fully atomistic model not only to predict the native structure but also to predict well the folding temperature. Further studies²⁷ have shown that using a pipeline strategy for structure refinement involving restrained simulations and filtering schemes can achieve modest improvement for most of the structures in CASP refinement sets, while, at the same time, requiring greatly reduced computational effort compared to trying to use fully atomistic level prediction without restraints. Baker and colleagues have recently shown that extensive sampling using a fully atomistic energy function was able to improve the quality of the majority of homology models that had been submitted to the CASP competition.²⁸ Atomistic refinement sensibly increased the number of models that appear to be suitable for use in molecular replacement phase determination.

One might hope that simply increasing the time scale of refinement simulations would result in significantly improved refinements. A recent study,²⁹ however, has shown that only marginally better refinement is achieved even when the simulation time scale is greatly increased. One way to work around this bottleneck may lie in re-examining the type of restraints being implemented in the atomistic refinement. While using harmonic distance restraints based on an initially predicted structure has demonstrated its power,²⁹ making the restraints too strong limits the range of motion of protein segments and slows the exploration of the large conformational space that still must be sampled during refinement simulations. If the initial structure is too far away from the crystal structure, or if there is a kinetic barrier in between the two correct and initial structures arising from backtracking, topological knotting, or hidden allostery, a dramatically more extensive sampling over the configurational space is needed to find the relevant native states. Harmonic restraints thus can add to the search difficulty. Here we explore how the implementation of “soft restraints” based on a collective variable that preserves the folding funnel and does not introduce strong topological constraints allows flexible exploration but still suffices to discourage unproductive unfolding events.

Energy landscape theory not only unravels the puzzle of protein folding²³ but also offers the mathematical framework to learn the nature of the interactions and to optimize transferable potentials for structure prediction.² Based on the energy landscape theory of protein folding,

a series of coarse-grained force fields have been progressively developed and shown to predict accurately protein structures from the ground up without employing any homology information.^{30–32} The physics-based nature of the transferable terms used in these force fields also enables their use to carry out mechanistic studies of biological macromolecular function. The most recent incarnation of these landscape optimized force fields, the Associative memory, Water mediated, Structure and Energy Model (AWSEM)³² has been proven successful not only in moderate resolution structure prediction,^{13,33,34} but also allows the quantitative study of protein–protein association,³⁵ protein allosteric effects, and the early stages of protein aggregation.^{36–40}

In this study, we put forward the use of a harmonic bias not in the Cartesian coordinates of the template but on a collective variable, Q , which measures the similarity between the simulated and template structures. This way of incorporating structural information from homologues of the target sequence into a structure prediction algorithm allows local changes to correct for differences between the target and template without too much energetic penalty. In addition to this bias, the forces are provided by the transferable and optimized AWSEM structure prediction force field to guide structure prediction. We call this combined protocol AWSEM-Template. By using a bias to a collective variable Q and setting the strength of the bias such that structures with moderate Q values are not severely penalized, the protocol allows the sampled structures to deviate from the template where necessary. These deviations are guided by the transferable terms of AWSEM, which by themselves are capable of predicting the structures of proteins to moderate resolution. Unlike using constraints individually on all pair distances, the gentle biasing based on the collective variable Q allows portions of the molecule to differ significantly from the template without straining the rest of the molecule because the penalty of the collective Q bias is proportional to the fraction of unsatisfied constraints, not to the degree to which individual distance constraints are violated.

To demonstrate the prediction capability of a protocol using soft template constraints, we selected eight protein targets having different levels of sequence identity to their templates. Seven of these proteins are previous CASP targets. We also study myoglobin. Myoglobin has been used often before as a benchmark for landscape optimized force fields. No heme is included in the prediction protocol, so trying to predict myoglobin structure should give some idea of what happens when cofactors are either missing or are unknown in making structure predictions. With guidance from templates, the transferrable potentials from AWSEM efficiently and accurately guide the structure predictions to high resolution for all eight test cases, even though four out of eight targets fall into the “twilight zone” of sequence identity with their template, with the lowest sequence identity between target and template being 15%. In most cases, structures with quality comparable to the highest resolution structures that are sampled during the coarse-grained prediction simulations can be selected blindly using a clustering procedure that seeks consensus between the lowest energy structures sampled during multiple independent simulated annealing trajectories.

After the coarse-grained prediction using the AWSEM-Template force field, we also refine the resulting structures using an atomistic force field (CHARMM36), again employing a soft collective coordinate constraint on the C_α atoms that is now derived from the predicted

structures provided by AWSEM-Template. The short refinement simulations consistently improve the accuracies of the initial all-atom models that are reconstructed from the coarse-grained structures using SCWRL4.⁴⁶ Molecular dynamics approaches naturally lead to an ensemble of choices for the “predicted structure”. We show here that picking the best refined structures out of this ensemble can be done using a scoring system based on all-atom and AWSEM energies whose parameters were optimized from the decoy set of MD sampled structures of a single example protein. The structurally most accurate structures from the refinement trajectory of the other test proteins are usually among the highest-scored structures. In this paper, we aim to document that coarse-grained predictions with AWSEM-Template followed by atomistic refinements accurately predicts protein structures with relatively high accuracy and resolution, including instances of proteins where templates are in the “twilight zone” of sequence identity.

2. METHOD

The prediction protocol described in this paper is summarized in Figure 1. First, the templates of the target sequence are identified, and structural information from the templates is incorporated into the AWSEM Hamiltonian (see below for details about how this is done). Next, 20 simulated annealing simulations are performed using the AWSEM-Template Hamiltonian starting from completely extended structures. The resulting 20 predicted structures are then clustered. The central structure from the cluster with the highest mutual structural similarity is then chosen as the final “predicted” coarse-grained structure. An all-atom model is then rebuilt from this final predicted coarse-grained structure and refined through 10 molecular dynamics simulations of 20 ns each in explicit solvent with a soft collective bias of the peptide backbone toward the coarse-grained predicted structure. Finally, the all-atom structures produced by the refinement simulations are evaluated to infer their structural accuracy using a novel machine learning scheme based on the AWSEM and atomistic energies. The top scoring structures are selected as the final predicted structures.

2.1. The AWSEM Force Field.

The Associative memory, Water mediated, Structure and Energy Model (AWSEM) is a predictive, coarse-grained, protein folding force field that represents the conformation of each amino acid residue using the location of three atoms within the residue: C_α , C_β , O. A complete backbone can be reconstructed on the fly from the positions of these three atoms by using an ideal peptide bond geometry assumption. AWSEM’s parameters were optimized by employing the principles of energy landscape theory²³ using as a training set a database of experimentally determined protein structures. Interested readers are encouraged to consult Davtyan et al.³² for detailed information about the AWSEM force field and Schafer et al.² for a pedagogical review of the history of the AWSEM model and the principles used for its optimization. The AWSEM-Template Hamiltonian as summarized in eq 1 is made up of both the optimized transferable tertiary energy terms and the specific knowledge-based terms of ordinary AWSEM, but now also includes the template-based biasing term.

$$V_{\text{AWSKM-Template}} = V_{\text{backbone}} + V_{\text{contact}} + V_{\text{burial}} + V_{\text{FM}} + V_{\text{HB}} + V_{\text{template}} \quad (1)$$

Among the transferable terms in eq 1, the backbone term, V_{backbone} , ensures that the peptide backbone stays connected and does not overlap itself. The many body burial term, V_{burial} , takes into account the instantaneous local density around each residue and attempts to sort each residue into its preferred burial environment: exposed, partially buried, or completely buried. The contact term, V_{contact} , includes a direct contact interaction and a water- or protein-mediated interaction. As a knowledge-based term, the fragment-based associative memory term, V_{FM} , locally biases the formation of secondary and supersecondary structures. This local in sequence bias uses structural information on peptide fragments found in experimentally solved structures from the Protein Data Bank (PDB). The fragments can be chosen either with or without knowledge of global sequence homology.³² This bias can alternatively employ structures sampled in atomistic simulations.^{33,34} The hydrogen bonding term, V_{HB} , favors formation of α -helices or β -sheets and can optionally be informed by the input of a secondary structure prediction.

In addition to the above knowledge-based terms, which are also used when no template is available, here we introduce an additional template-based modeling term, V_{template} , which guides the global folding with tertiary information from the identified templates. V_{template} is a harmonic potential based on the collective coordinate, Q , which measures structural similarity by comparing pairwise distances between two structures (eq 3). Q values range between 0 and 1 with 1 indicating near perfect identity up to an angstrom or so for pair distances.

$$V_{\text{template}} = k_{\text{template}}(Q_{\text{template}} - 1)^2 \quad (2)$$

$$Q_{\text{template}} = \frac{\sum_{i,j} \theta(r_{ij}^N - 2\text{\AA}) e^{-(r_{ij} - r_{ij}^N)^2 / 2\sigma_{ij}^2}}{\sum_{i,j} \theta(r_{ij}^N - 2\text{\AA})} \quad (3)$$

In eq 3, $\theta(r_{ij}^N - 2\text{\AA})$ is 1 for $r_{ij}^N \geq 2.0\text{\AA}$ and 0 otherwise, $\sigma_{ij} = |j-i|^{0.15}$, and r_{ij}^N is the pairwise residue-residue distance in the templates after alignment, while r_{ij} is the pairwise distance in the structure whose Q value is being measured. k_{template} can be used to scale the strength of the template term relative to the other terms used in AWSEM-Template. A detailed description of how to obtain the set of r_{ij}^N for a given target will be given in the following sections. Additionally, detailed parameters for structure prediction using AWSEM-Template are shown in the Supporting Information.

2.2. Predicting Secondary Structural Information for the AWSEM-Template Hamiltonian.

The backbone term in AWSEM-Template, V_{backbone} , can optionally take advantage of a secondary structure prediction. When present, the secondary structure prediction is used to apply additional biases to the backbone Φ and Ψ angles toward either α -helical or β -strand conformations. In this study, we adopted PSSpred developed by Zhang and his co-workers to provide the predicted secondary structural information on the target sequence. PSSpred is a neural network algorithm, which has been proven to be effective in secondary structure prediction.⁴¹

2.3. Obtaining Fragment Memories for Structure Prediction Targets.

The fragment memory term for AWSEM, V_{FM} , applies local-in-sequence structural biases by aligning short fragments from experimentally resolved structures or atomistic simulations to the sequence of the structure prediction target. In the past, when testing the ability of AWSEM to perform template-free *de novo* structure prediction, we excluded information from fragments derived from structures that have sequences that are globally homologous to the target sequence. In some of our previous studies of *de novo* structure prediction, we used BLAST⁴² to find the short fragments that locally match in sequence but are not from globally homologous sequences. When performing template-based modeling, a sensible strategy would seem to be to include only fragment memories from structures that have homologous sequences. In practice, however, limiting fragment information to that from homologues only inevitably leads to unfilled gaps in the fragment memory alignments and results in reduced performance due to uneven coverage of the alignments across the target sequence. Therefore, in this work, fragments taken from homologous sequences are supplemented by fragments taken from nonhomologous sequences and the weights of the two sources of fragment memories are balanced based on the number of aligned memories in each fragment from the target sequence. In this way, those parts of the protein target that do not match the template are being driven by the same forces as in a strict *de novo* AWSEM prediction.

2.4. Obtaining Tertiary Structural Information from Templates.

We used HHpred⁴³ to find templates for our target sequences using a minimum threshold confidence score of 95. In order better to simulate a scenario in which template-based modeling would be employed in practice, we excluded all templates that were deposited in the PDB after the target structure. The aligned regions from the templates were selected and renumbered according to their alignment to the target sequence from HHpred. Then, a pairwise distance matrix was created and the entries in the matrix were used to define the r_{ij}^N values in eq 3.

2.5. Simulation Details for Structure Predictions.

Structure prediction simulations were performed using the Large-Scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) software package,⁴⁴ in which the AWSEM-Template force field is available as an open source add-on.³² Annealing simulations were carried out for the eight targets annealing from either 600 to 200 K or 500 to 200 K in 8 million steps. The initial temperature of the annealing was 600 K for T0766, T0833 and

T0815; the initial temperature was 500 K for all other targets. All the predictions start from a fully extended configuration that was built using PyMol (www.pymol.org). For the proteins in this study, an eight million step annealing simulation using the AWSEM-Template model takes 10–24 h to complete when running on a single CPU.

To compute the free energy landscapes of a protein, we used umbrella sampling along an order parameter, Q with respect to the crystal structure of this protein. The harmonic biasing potential used for constant temperature (400 K) umbrella sampling simulations for 8 million steps is shown in eq 4.

$$V_{Q\text{-bias}} = k_{Q\text{-bias}}(Q - Q_0)^2 \quad (4)$$

In eq 4, $k_{Q\text{-bias}} = 200 \text{ kcal/mol}$. The biasing center values, Q_0 , were chosen to be equally spaced from 0 to 0.98 with an increment of 0.02. The unbiased free energy landscapes were then reconstructed from the umbrella sampling data using the weighted histogram analysis method (WHAM).⁴⁵ The free energy landscapes can also be extrapolated to temperatures other than the temperature at which the simulation was performed.

2.6. Rebuilding All-Atom Models Based on Coarse-Grained Predictions from AWSEM.

The first round of selection for coarse-grained structures was done on the basis of the AWSEM-Template energy (eq 1). The structures with the lowest AWSEM energy from each of the 20 simulated annealing trajectories were clustered based on their mutual- Q values. High average mutual- Q values within a cluster indicates strong mutual structural similarity. A similar clustering method was used in a previous paper.³⁴ The centroid structure of the cluster with the highest average mutual- Q was selected as the final candidate for atomistic refinement.

Since the structures produced by the coarse-grained AWSEM simulation have only backbone and C_β atoms, rebuilding of the side-chains on these coarse-grained structures was performed using SCWRL4.⁴⁶ We also tried using MODELLER¹⁷ to rebuild the side-chains but found that, in general, SCWRL4 performed somewhat better in terms of choosing native-like rotamer states for side chains (data not shown). After rebuilding side-chains onto the selected backbone structure, one is left with a complete protein structure, with coordinates for all of the atoms, which is ready for refinement using all-atom molecular dynamics simulations.

2.7. Simulation Details for Structural Refinement.

Prior to atomistic refinement, the selected candidate structure was electrostatically neutralized by adding an appropriate number of either Na or Cl ions, solvated in explicit water with 0.15 M NaCl, and simulated for 20 ns with a time step of 2 fs. A total of 10 independent runs with different random seeds as input to the velocity initializations were performed for each protein. A soft harmonic C_α Q -bias with a spring constant of 200.0 kJ/mol toward the coarse-grained structure predicted by AWSEM-Template was added. This

soft bias allows for significant backbone movement but disfavors unproductive unfolding. The CHARMM36 force field⁴⁷ was used for our refinement simulations, and the simulations were performed with GROMACS 5.0.4.^{48,49} For the proteins in this study, each all-atom refinement simulation takes about 24 h to complete when running in parallel on 16 CPUs.

2.8. Evaluating the All-Atom Structures with a Machine Learning Method Based on Linear Regression with Energies As Features.

To provide a way to select blindly the most native-like structures from among the structures generated by the all-atom refinement simulations, we formulated a classification problem and trained a supervised learning algorithm using a set of “energetic” features. Training data were generated by assigning the labels of “high” and “low” quality to all-atom structures corresponding, respectively, to high GDT-TS structures (among the top 2.5%) and to low GDT-TS structures (the rest). Training this postselection was done using only a single one of the prediction targets. We chose three energy terms as our feature set. These energies were the statistical energy calculated by RWplus,⁵⁰ which is sensitive to the side chain orientations, the AWSEM template energy (V_{template}), and the total AWSEM energy excluding the template energy. All energy terms for each protein were renormalized by subtracting the mean and dividing by the standard deviation over the training ensemble.

For the specific training set to learn the weights of the three features, we used 80 percent of the structures generated during the all-atom refinement simulations of one protein, T0792. We then optimized the coefficients of a logistic regression model with features x and weights θ . The cost function that was minimized during training is given in eq 5, where

$$\hat{p} = \frac{1}{1 + e^{-\theta^T \cdot x}}. \text{ The weights obtained during training and used for evaluation of the}$$

structures generated during the all-atom refinement simulations are -0.2 , -1.6 , and -0.2 for RWplus, V_{template} , and $V_{\text{AWSEM}} = V_{\text{backbone}} + V_{\text{contact}} + V_{\text{burial}} + V_{\text{FM}} + V_{\text{HB}}$, respectively.

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m \left[y^{(i)} \log(\hat{p}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{p}^{(i)}) \right] \quad (5)$$

The trained model yields a probability that a structure is of high quality. From each set of refinement simulations, we then select those structures that have the highest predicted probability of being of high quality as the final predicted structures. Aside from GDT-TS score, we have also tried carrying out the training using other metrics like Q and RMSD. The quality of prediction when using these other metrics is similar to what was found when using the GDT-TS score as the metric (data not shown).

2.9. Structure Prediction Accuracy Metrics.

To evaluate the accuracy of the structure predictions provided by the AWSEM-Template protocol, we use multiple metrics: C_{α} -RMSD, CE-RMSD, Q, and GDT-TS. These quantities all measure the structural similarity between two structures, but in different ways. In the case of the structure prediction, these metrics were used to measure the accuracy of each predicted model based on the corresponding native structure. C_{α} -RMSD captures the root-

mean-square-deviation of C_{α} atoms between the predicted and native structures. CE (combinatorial extension) RMSD calculates the RMSD of the aligned regions of proteins after CE alignment.⁵¹ All CE-RMSDs are given along with the percentage of aligned residues. The RMSD is particularly relevant when considering the potential of using predicted models for molecular replacement to phase X-ray diffraction data. Q is a metric with the form given by

$$Q = \frac{2}{(N-2)(N-3)} \sum_{j-i > 2} e^{-\left(r_{ij} - r_{ij}^N\right)^2 / 2\sigma_{ij}^2} \quad (6)$$

where N is the total number of residues. r_{ij} and r_{ij}^N are the distances between two residues i and j in the model and native structures. $\sigma_{ij} = |j-i|^{0.15}$ is a tolerance that depends weakly on the sequence separation of residues i and j . GDT-TS (Global Distance Test Total Score) is another measure frequently used in the CASP competition⁵² to assess the quality of structure prediction. The GDT-TS is defined as the percentage of C_{α} atoms falling within a specified tolerance of the corresponding distance in the native structure. Typically, one uses the cutoff values of 1, 2, 4, and 8 Å to compute the percentage, and GDT-TS is the average of these percentages, and this is what we have done in this study as well.

3. RESULTS

3.1. Summary of Template-Guided Structure Prediction and Refinement of Eight Proteins with Varying Degrees of Sequence Similarity to Their Templates.

The overall scheme for producing all-atom structures of prediction targets that have identifiable template structures is sketched in Figure 1. The scheme can be divided into two stages: first, coarse-grained structures are produced using AWSEM-Template and then atomistic refinement of selected coarse-grained structures that have been chosen without reference to any knowledge of the native structure is carried out. The results of these stages of structure prediction are summarized using four structure quality metrics (RMSD, CE-RMSD, GDT-TS, and Q) in Table 1. To understand how the structures obtained directly from the AWSEM-Template coarse-grained annealing and those specific coarse-grained structures that were subsequently chosen for refinement using all-atom simulations compare to structures obtained when using other state-of-the-art homology modeling methods, we have included a summary of the corresponding prediction results obtained using MODELLER and I-TASSER on all eight of the prediction targets in Table 1. MODELLER and I-TASSER were run with default settings and utilized the same template structures that were employed for the AWSEM-Template predictions. Overall, the coarse-grained AWSEM-Template structures that were blindly selected are comparable in quality to those produced by I-TASSER but have significantly lower RMSD values than those produced by MODELLER when using the same template to guide structure prediction. In 4 out of 8 cases, the selected AWSEM-Template coarse-grained annealing in fact yields the lowest RMSD structure of all the prediction tools. In 3 out of the 4 remaining cases, AWSEM-Template yields the second lowest RMSD structure. Besides RMSD, the other structure quality metrics also indicate that

both the coarse-grained AWSEM-Template structures and the all-atom structures obtained by refining AWSEM-Template structures are comparable or better in quality than those produced by other state-of-the-art methods.

Table 1 also documents that atomistic refinement, as described below, except in two cases, improves the RMSD of the selected coarse-grained structure. The final refined structures always have better RMSDs than MODELLER results and, in 5 out of 8 cases, considerably better RMSDs than I-TASSER. Interestingly, the best coarse-grained annealing results are sometimes substantially better than the results obtained by atomistic refinement of the selected templates. This argues that a more sophisticated selection scheme of the coarse-grained annealings would be a valuable addition to this technology.

3.2. Coarse-Grained Structure Prediction for Eight Proteins Using AWSEM-Template.

The coarse-grained structure prediction results using AWSEM-Template are summarized in Figure 2 and Table 1. In Figure 2, we have plotted the Q values of the structures that were selected using the clustering analysis described in the Methods section versus the sequence identity of the template that was used as input to the V_{template} energy term. The results for the AWSEM predictions without template information are shown as blue diamonds, while those from simulations with template guidance are shown as red diamonds. As expected, the results from AWSEM-Template predictions are consistently better than pure AWSEM predictions that use no template information. As a control, we also carried out predictions with a Hamiltonian that included only the template bias and backbone geometry (green diamonds) but that contained no contributions from AWSEM forces. The results indicate that predictions obtained using a combination of AWSEM potential with template-based bias usually are better than those found using templates alone.

Template-based modeling can be challenging for targets in the “twilight zone” of low sequence identity. Among our eight targets, at least four have a sequence identity below 30% compared to their templates. AWSEM-Template predictions for the twilight zone targets are nonetheless of high quality, which suggests that soft constraints from templates work well in template-based modeling which is also informed by a transferable and optimized structure prediction model like AWSEM. Although the predictions for the four targets that are not in the “twilight zone” are better on average than those that are in the “twilight zone”, the best structures that are sampled for targets with low sequence identity to their templates still tend to be of quite high quality ($Q > 0.7$, RMSD of 2–3 Å). In Figure 2, the results for the blindly selected structures are shown in red squares. The blindly selected structures are obviously not as good as the best ones sampled but nevertheless are usually quite close to the best sampled structures (within ≈ 0.1 Q units). These blindly selected structures are used both as the starting structures for all-atom refinement and also provide the biasing information in the all-atom refinement simulations as described below. The GDT-TS, Q, CE-RMSD, and RMSD values of the best sampled and blindly selected structures can be found in Table 1. Again, the AWSEM-Template and refined AWSEM-Template predictions fare very well in comparison to the existing methods (MODELLER and I-TASSER).

To assess the consistency of the results from many annealing runs, we plotted the best Q values for each of the 20 annealing runs (sorted in ascending order) and the results are

shown in Figure 3. Three models were tested, namely the combined AWSEM-Template model, the standard AWSEM model without any template information, and a model with only template information guiding tertiary structure formation. The AWSEM-Template predictions are seen to be consistently of higher quality than the predictions using AWSEM without template information. For the majority of targets (7 out of 8), the combined AWSEM-Template model produces the highest quality structures, indicating the utility of combining a transferable structure prediction model with template information. The structural alignments of the best sampled structures to the X-ray crystal structures, shown in Figure 3, indicate that the best sampled structures are highly similar to the X-ray crystal structures. In some cases, such as T0766 and T0784, AWSEM-Template sampled structures are close to 1 Å RMSD from the crystal structure after Combinatorial Extension (CE) alignment.⁵¹ The RMSDs after CE alignment are below 3 Å for all the other proteins. The above results taken together suggest that AWSEM-Template can generate accurate models so long as there are acceptable templates available (sequence identity >15%). Also, consistent with what is shown in Figure 2, the predictions from the Hamiltonian that includes only template information (green) but no transferable tertiary interactions or fragment information are lower in quality than those from AWSEM-Template Hamiltonian.

3.3. All-Atom Molecular Dynamics Refinement of Coarse-Grained Structures Predicted from AWSEM-Template.

We refined a subset of blindly selected coarse-grained model structures for each of the eight target proteins by performing 10 replicates of 20 ns all-atom molecular dynamics simulation starting from the same blindly selected structure. A soft collective bias based on the C_{α} Q value of the simulated all-atom structure compared with the blindly selected predicted coarse-grained AWSEM-Template model structure was added to the full atomistic energy function (see the Methods section for details). The soft and collective nature of the bias toward the template allows for significant backbone movement as guided by side-chain rearrangements but also disfavors unproductive unfolding. As shown in Table 1, in terms of GDTTS scores and Q values, all eight structures were improved by carrying out the atomistic refinement. Among the improved structures, 1MBA, T0784, and T0792 show the largest gains in GDT-TS.

In order to visualize for which regions refinement makes significant structural changes, the contact maps of the initial (reconstructed from coarse-grained predictions using AWSEM-Template), best refined (structure with the highest Q value), and native structure are plotted with a cutoff distance of 9.5 Å for C_{α} atoms. Also, the corresponding three-dimensional structures were aligned using the CE alignment algorithm⁵³ (Figure 4). We observe that the largest improvements during refinement occur in helical regions of the proteins that we studied, which is consistent with observations made by others.²⁷

In addition, we calculated the change of side-chain accuracy during our atomistic refinement simulations. We first categorize the residues according to their normalized solvent accessible surface area in the X-ray crystal structure.⁵⁴ The side-chain orientations of the buried residues are presumably better defined by tertiary interactions and thus fluctuate less than the orientations of the floppy side-chains protruding from the protein surface into the

solvent. Therefore, we calculate and highlighted the change of the side-chain accuracy of the buried side-chains alone (buried accuracy or “bAccuracy”) during refinement. The side-chain accuracy was calculated based on the two major side-chain dihedrals (χ_1 and χ_2) for each residue and this accuracy is plotted as a function of simulation time. The average and standard deviation of the bAccuracies as a function of time were calculated based on the total of 10 refinement simulations for each protein. As shown in Figure 5, the atomistic refinement typically increases the average χ_1 accuracy for most proteins. The accuracy of the exposed side-chains is also shown in Figure 5 for comparison. Unlike the bAccuracy, the accuracy of the surface-exposed side-chains, or “sAccuracy”, shows no systematic trend in the refinement simulations, which is consistent with our presumption that the χ_1 and χ_2 angles for surface-exposed side-chains are relatively poorly defined due to the lack of tertiary interactions.

3.4. Reduction of Frustration during All-Atom Refinement Simulations.

While overall the frustration in native proteins is kept to a minimal level, patches of frustration do occur in accurate protein structures. These patches in accurate structures signal regions of importance to binding⁵⁵ and to allosteric function.⁵⁶ Obviously, in the context of prediction we must be aware also that additional frustration also can be introduced by errors in these predicted structures. These patches may signal either errors in the force fields or insufficiency of the sampling.

The Frustratometer⁵⁵ can be used to localize energetic frustration in proteins. Frustration analysis allows us to pinpoint where frustration due to errors in coarse-grained predictions can be alleviated by all-atom refinement simulations, as measured by the residue-level AWSEM energy function. As shown in Figure S1, S2, S4, S5, the four proteins that show the largest improvement in terms of GDT-TS score, 1MBA, T0251, T0784 and T0792 also show significant reduction in localized patches of frustration after atomistic structure refinement. For example, the frustration in T0784 is reduced primarily in the region where the atomistic simulation converted a loop structure in the coarse-grained model into a better formed β -sheet.

In order to better understand the overall changes in frustration during the all-atom refinement simulations, we used the Frustratometer to calculate the number of minimally frustrated and highly frustrated interactions during the refinement simulations (Figure 6). Most targets do not show large changes in the overall number of minimally or highly frustrated interactions during the refinement simulations. For 1MBA, the average numbers of both the minimally frustrated interactions and of the highly frustrated interactions increase significantly during the refinement simulation, with about a 2-fold larger increase in minimally frustrated interactions compared to highly frustrated interactions. For 1MBA, the final average number of both minimally frustrated interactions and highly frustrated interactions match those found in the crystal structure, consistent with our finding that its structure is significantly refined during the all-atom simulations. We note, of course, that 1MBA in our simulations lacks the crucial heme cofactor which would explain the presence of highly frustrated interactions in the apo form.

4. DISCUSSION

4.1. Template Information Can Be Naturally Incorporated into the AWSEM Framework, Yielding a Template-Based Modeling Method That Is Supported by a Realistic and Transferable Coarse-Grained Model.

The transferrable AWSEM code has proved successful in moderate-resolution structure prediction when no structural templates are available.³² Nevertheless, for practical situations when homologues are often available, it is useful to incorporate structural information that can reliably be inferred from templates. In the AWSEM-Template protocol, the structural information from templates is encoded using a collective but soft bias, which, by collaborating with AWSEM's transferable components, generally yields very high-quality coarse-grained structures.

To better understand the connection between the transferable and template-based components of AWSEM-Template potential, we computed free energy profiles for T0792 protein with and without the template-based component. The free energy profiles of T0792 were computed using umbrella sampling simulations and are plotted as a function of Q . As shown in Figure 7, at this temperature the free energy profile for AWSEM with no template-based bias has a free energy minimum at $Q \approx 0.3$, while the free-energy minimum was shifted to $Q \approx 0.58$ by including the bias (with $k = 200$ kcal/mol). Thus, the template-based term does indeed guide the structure prediction to a more native-like structure.

To understand the influence of the strength of the constraint on the collective variable Q , we compared the free energy profiles of T0792 with two different strengths of bias toward the template structure (Figure 7). The two profiles both have free energy minima near $Q = 0.6$. The softer constraint allows the model to sample more structures with lower Q values. For targets with low sequence identity to their templates, or for the purpose of modeling proteins that undergo significant conformational changes, the presence of this flexibility may be important. Finding the optimal balance between the template-based and the transferable terms for studying mechanistic aspects of protein function that go beyond simple structure prediction is an interesting challenge that we will pursue in the future.

We also investigated how adding the template bias influences the funneling to the native state by calculating the expectation value of the individual energy terms (as well as the total energy) as a function of Q (Figure 8), using T0792 as an example. A well funneled energy function will lead to monotonically decreasing energy expectation values with increasing Q . The standard AWSEM model is funneled to a Q value of ≈ 0.4 , and structures with high Q values have approximately equal energy as structures at $Q = 0.4$. However, as can be seen in panel E of Figure 3, the state at $Q \approx 0.4$ "traps" all of the simulated annealing simulations in the standard AWSEM model, effectively preventing the high Q structures from being sampled. In contrast, when guidance from the template structure is added, V_{total} decreases monotonically to very high Q values (Figure 8), and this increased funneling results in improved structure prediction performance when using simulated annealing simulations to predict structure (panel E, Figure 3).

4.2. Combination of Template-Based Modeling with Coevolutionary Information under the AWSEM Framework Is Probably Optimal for Structural Prediction.

As the most practical approach for structure prediction, template-based modeling has enjoyed success during past CASP contests. A general analysis of CASP targets shows that templates can be found in existing structure databases for more than 60% of the targets. Template-free targets nonetheless make up a substantial fraction of all targets encountered today. Therefore, for the purposes of both improving the modeling of targets that are only partially covered by templates and for predicting the structure of targets for which no template information is available or identifiable, it is important to continue to develop approaches to structure prediction that complement pure template-based modeling.

Coevolutionary constraints inferred from large sequence data sets also can help guide accurate structure prediction of proteins.^{9–11,13,14} In contrast to template-based modeling, coevolutionary constraints do not directly provide three-dimensional coordinates but only provide inferred contact information, which is not generally as complete as the contact information inherent in template-based approaches. In practice, we postulate that an algorithm that includes both template 3D information and coevolutionary pair information into a framework based on an optimized and transferable coarse-grained model should be quite useful to efficiently and accurately predict protein folds most generally. We are currently preparing another paper on such a hybrid protocol.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was supported by Grant R01 GM44557 from the National Institute of General Medical Sciences. Additional support was also provided by the D.R. Bullard-Welch Chair at Rice University, Grant C-0016. We thank the Data Analysis and Visualization Cyberinfrastructure funded by National Science Foundation Grant OCI-0959097. Additional support was provided by the Center for Theoretical Biological Physics sponsored by the NSF (Grant PHY-1427654) and by NSFCHE 1614101.

REFERENCES

- (1). Dill KA; MacCallum JL The Protein-Folding Problem, 50 Years On. *Science* 2012, 338, 1042–1046. [PubMed: 23180855]
- (2). Schafer NP; Kim BL; Zheng W; Wolynes PG Learning To Fold Proteins Using Energy Landscape Theory. *Isr. J. Chem* 2014, 54, 1311–1337. [PubMed: 25308991]
- (3). Wolynes PG Evolution, energy landscapes and the paradoxes of protein folding. *Biochimie* 2015, 119, 218–230. [PubMed: 25530262]
- (4). Ferreiro DU; Komives EA; Wolynes PG Frustration in biomolecules. *Q. Rev. Biophys* 2014, 47, 285–363. [PubMed: 25225856]
- (5). Goldstein RA; Luthey-Schulten ZA; Wolynes PG Optimal protein-folding codes from spin-glass theory. *Proc. Natl. Acad. Sci. U.S. A* 1992, 89, 4918–4922. [PubMed: 1594594]
- (6). Service RF Problem Solved* (*sort of). *Science* 2008, 321, 784–786. [PubMed: 18687949]
- (7). Lindorff-Larsen K; Piana S; Dror RO; Shaw DE How fast-folding proteins fold. *Science* (Washington, DC, U. S.) 2011, 334, 517–520.

- (8). Best RB; Hummer G; Eaton WA Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Natl. Acad. Sci. U. S. A* 2013, 110, 17874–17879. [PubMed: 24128758]
- (9). Morcos F; Pagnani A; Lunt B; Bertolino A; Marks DS; Sander C; Zecchina R; Onuchic JN; Hwa T; Weigt M Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U. S. A* 2011, 108, E1293–E1301. [PubMed: 22106262]
- (10). Marks DS; Hopf TA; Sander C Protein structure prediction from sequence variation. *Nat. Biotechnol* 2012, 30, 1072. [PubMed: 23138306]
- (11). Sułkowska JI; Morcos F; Weigt M; Hwa T; Onuchic JN Genomics-aided structure prediction. *Proc. Natl. Acad. Sci. U. S. A* 2012, 109, 10340–10345. [PubMed: 22691493]
- (12). Morcos F; Schafer NP; Cheng RR; Onuchic JN; Wolynes PG Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proc. Natl. Acad. Sci. U. S. A* 2014, 111, 12408–12413. [PubMed: 25114242]
- (13). Sirovetz BJ; Schafer NP; Wolynes PG Protein structure prediction: making AWSEM AWSEMER by adding evolutionary restraints. *Proteins: Struct., Funct., Genet* 2017, 85, 2127–2142. [PubMed: 28799172]
- (14). Ovchinnikov S; Park H; Varghese N; Huang P-S; Pavlopoulos GA; Kim DE; Kamisetty H; Kyrpides NC; Baker D Protein structure determination using metagenome sequence data. *Science (Washington, DC, U. S.)* 2017, 355, 294–298.
- (15). Berman HM; Westbrook J; Feng Z; Gilliland G; Bhat TN; Weissig H; Shindyalov IN; Bourne PE The Protein Data Bank. *Nucleic Acids Res* 2000, 28, 235–242. [PubMed: 10592235]
- (16). Khor BY; Tye GJ; Lim TS; Choong YS General overview on structure prediction of twilight-zone proteins. *Theor. Biol. Med. Modell* 2015, 12, 10.1186/s12976-015-0014-1.
- (17). Sali A; Blundell TL Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol* 1993, 234, 779–815. [PubMed: 8254673]
- (18). Zhang Y; Arakaki AK; Skolnick J TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins: Struct., Funct., Genet* 2005, 61 (Suppl 7), 91–98. [PubMed: 16187349]
- (19). Zhou H; Skolnick J Template-based protein structure modeling using TASSER(VMT). *Proteins: Struct., Funct., Genet* 2012, 80, 352–361. [PubMed: 22105797]
- (20). Roy A; Kucukural A; Zhang Y I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc* 2010, 5, 725–738. [PubMed: 20360767]
- (21). Yang J; Yan R; Roy A; Xu D; Poisson J; Zhang Y The ITASSER Suite: protein structure and function prediction. *Nat. Methods* 2015, 12, 7–8. [PubMed: 25549265]
- (22). Rost B Twilight zone of protein sequence alignments. *Protein Eng., Des. Sel* 1999, 12, 85–94.
- (23). Bryngelson JD; Onuchic JN; Socci ND; Wolynes PG Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Struct., Funct., Genet* 1995, 21, 167–195. [PubMed: 7784423]
- (24). Chopra G; Summa CM; Levitt M Solvent dramatically affects protein structure refinement. *Proc. Natl. Acad. Sci. U. S. A* 2008, 105, 20239–20244. [PubMed: 19073921]
- (25). Raval A; Piana S; Eastwood MP; Dror RO; Shaw DE Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins: Struct., Funct., Genet* 2012, 2071–2079. [PubMed: 22513870]
- (26). Raval A; Piana S; Eastwood MP; Shaw DE Assessment of the utility of contact-based restraints in accelerating the prediction of protein structure using molecular dynamics simulations: Contact Restraints in Protein Simulations. *Protein Sci* 2016, 25, 19–29. [PubMed: 26266489]
- (27). Feig M; Mirjalili V Protein structure refinement via molecular-dynamics simulations: What works and what does not?: MD-Based Protein Structure Refinement. *Proteins: Struct., Funct., Genet* 2016, 84, 282–292. [PubMed: 26234208]
- (28). Park H; Ovchinnikov S; Kim DE; DiMaio F; Baker D Protein homology model refinement by large-scale energy optimization. *Proc. Natl. Acad. Sci. U. S. A* 2018, 115, 3054–3059. [PubMed: 29507254]
- (29). Heo L; Feig M What makes it difficult to refine protein models further via molecular dynamics simulations. *Proteins: Struct., Funct., Genet* 2018, 86, 177–188.

- (30). Friedrichs MS; Wolynes PG Toward protein tertiary structure recognition by means of associative memory hamiltonians. *Science* (Washington, DC, U. S.) 1989, 246, 371–373.
- (31). Papoian GA; Ulander J; Eastwood MP; Luthey-Schulten Z; Wolynes PG Water in protein structure prediction. *Proc. Natl. Acad. Sci. U. S. A* 2004, 101, 3352–3357. [PubMed: 14988499]
- (32). Davtyan A; Schafer NP; Zheng W; Clementi C; Wolynes PG; Papoian GA AWSEM-MD: Protein Structure Prediction Using Coarse-Grained Physical Potentials and Bioinformatically Based Local Structure Biasing. *J. Phys. Chem. B* 2012, 116, 8494–8503. [PubMed: 22545654]
- (33). Chen M; Lin X; Zheng W; Onuchic JN; Wolynes PG Protein Folding and Structure Prediction from the Ground Up: The Atomistic Associative Memory, Water Mediated, Structure and Energy Model. *J. Phys. Chem. B* 2016, 120, 8557–8565. [PubMed: 27148634]
- (34). Chen M; Lin X; Lu W; Onuchic JN; Wolynes PG Protein Folding and Structure Prediction from the Ground Up II: AAWSEM for Proteins. *J. Phys. Chem. B* 2017, 121, 3473–3482. [PubMed: 27797194]
- (35). Zheng W; Schafer NP; Davtyan A; Papoian GA; Wolynes PG Predictive energy landscapes for protein-protein association. *Proc. Natl. Acad. Sci. U. S. A* 2012, 109, 19244–19249. [PubMed: 23129648]
- (36). Zheng W; Schafer NP; Wolynes PG Free energy landscapes for initiation and branching of protein aggregation. *Proc. Natl. Acad. Sci. U. S. A* 2013, 110, 20515–20520. [PubMed: 24284165]
- (37). Zheng W; Tsai M-Y; Chen M; Wolynes PG Exploring the aggregation free energy landscape of the amyloid- β protein (1–40). *Proc. Natl. Acad. Sci. U. S. A* 2016, 113, 11835–11840. [PubMed: 27698130]
- (38). Chen M; Zheng W; Wolynes PG Energy landscapes of a mechanical prion and their implications for the molecular mechanism of long-term memory. *Proc. Natl. Acad. Sci. U. S. A* 2016, 113, 5006–5011. [PubMed: 27091989]
- (39). Chen M; Tsai M; Zheng W; Wolynes PG The Aggregation Free Energy Landscapes of Polyglutamine Repeats. *J. Am. Chem. Soc* 2016, 138, 15197–15203. [PubMed: 27786478]
- (40). Chen M; Wolynes PG Aggregation landscapes of Huntingtin exon 1 protein fragments and the critical repeat length for the onset of Huntington's disease. *Proc. Natl. Acad. Sci. U. S. A* 2017, 114, 4406. [PubMed: 28400517]
- (41). Yan R; Xu D; Yang J; Walker S; Zhang Y A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. *Sci. Rep* 2013, 3, 10.1038/srep02619.
- (42). Altschul SF; Gish W; Miller W; Myers EW; Lipman DJ Basic local alignment search tool. *J. Mol. Biol* 1990, 215, 403–410. [PubMed: 2231712]
- (43). Soding J; Biegert A; Lupas AN The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* 2005, 33, W244–W248. [PubMed: 15980461]
- (44). Plimpton S Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys* 1995, 117, 1–19.
- (45). Kumar S; Rosenberg JM; Bouzida D; Swendsen RH; Kollman PA The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem* 1992, 13, 1011–1021.
- (46). Krivov GG; Shapovalov MV; Dunbrack RL Improved prediction of protein side-chain conformations with SCWRL4. *Proteins: Struct., Funct., Genet* 2009, 77, 778–795. [PubMed: 19603484]
- (47). Best RB; Zhu X; Shim J; Lopes PEM; Mittal J; Feig M; Mackerell AD Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone, and side-chain (1) and (2) dihedral angles. *J. Chem. Theory Comput* 2012, 8, 3257–3273. [PubMed: 23341755]
- (48). Abraham MJ; Murtola T; Schulz R; Pall S; Smith JC; Hess B; Lindahl E GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 2015, 1–2, 19–25.
- (49). Tribello GA; Bonomi M; Branduardi D; Camilloni C; Bussi G PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun* 2014, 185, 604–613.

- (50). Zhang J; Zhang Y A Novel Side-Chain Orientation Dependent Potential Derived from Random-Walk Reference State for Protein Fold Selection and Structure Prediction. PLoS One 2010, 5, e15386. [PubMed: 21060880]
- (51). Shindyalov IN; Bourne PE Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng., Des. Sel 1998, 11, 739–747.
- (52). Zemla A; Venclovas C; Moult J; Fidelis K Processing and analysis of CASP3 protein structure predictions. Proteins: Struct., Funct., Genet 1999, 37 (Suppl 3), 22–29.
- (53). Shindyalov IN; Bourne PE Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng., Des. Sel 1998, 11, 739–747.
- (54). Fraczekiewicz R; Braun W Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. J. Comput. Chem 1998, 19, 319–333.
- (55). Ferreira DU; Hegler JA; Komives EA; Wolynes PG Localizing frustration in native proteins and protein assemblies. Proc. Natl. Acad. Sci. U. S. A 2007, 104, 19819–19824. [PubMed: 18077414]
- (56). Ferreira DU; Hegler JA; Komives EA; Wolynes PG On the role of frustration in the energy landscapes of allosteric proteins. Proc. Natl. Acad. Sci. U. S. A 2011, 108, 3499–3503. [PubMed: 21273505]

Figure 1.
Protocol for structure prediction using AWSEM-Template and all-atom refinement.

Figure 2.

Summary of the coarse-grained structure prediction results for eight proteins with different sequence identities to their templates. From left to right, the prediction targets are T0251, 1MBA, T0815, T0833, T0792, T0784, T0803, and T0766. Q , a measure of structural similarity, is plotted on the y -axis. The maximum Q values obtained during the AWSEM annealing simulations without template information (blue diamonds) and the AWSEM-Template annealing simulations (red diamonds) are plotted along with the Q values of the blindly selected AWSEM-Template structures (red squares). Blind selection of structures typically yields structures that are within ≈ 0.1 Q units of the best sampled structure. For comparison, the maximum Q values obtained from annealing simulations using a Hamiltonian having only the template bias and backbone geometry constraints are drawn in green diamonds.

Figure 3.

Prediction quality and consistency for each of the eight structure prediction targets during the coarse-grained prediction step. Red diamonds correspond to AWSEM-Template predictions and blue diamonds correspond to AWSEM predictions without template information. Green diamonds correspond to predictions using a model that has only template information to guide tertiary structure prediction and lacks the transferable energy terms in the standard AWSEM model (V_{contact} , V_{burial} , and V_{HB}). In each case, the highest Q values from each of the 20 annealing simulations are plotted in ascending order. The alignments of the best sampled structures (those sampled structures that have the highest Q value) from the AWSEM-Template annealing simulations with the corresponding X-ray crystal structures are shown on the right side of each panel. The predicted structures are shown in red and the X-ray crystal structures are shown in white. Below the structures, several accuracy metrics are given including the Q value, the RMSD value, and the CE-RMSD value along with the percentage of aligned residues in the CE alignment.⁵¹

Figure 4.

Comparison of initial and refined structures of each protein. Left: Three dimensional alignment of initial, refined, and native structures. The initial structure (red) and the refined structure with the maximum Q value (blue) are aligned to the native structure (white). Middle: The contact maps comparing the initial and native structures (upper left) and comparing the refined and native structures (lower right). The contact maps are calculated based on the C_{α} atoms with the cutoff distance of 9.5 Å. Hollow black squares are native contacts. Filled red squares are false-positive contacts. Filled green squares are false-negative contacts. Right: Changes in the incorrectly predicted contacts during refinement. Here, hollow symbols indicate a decrease and filled symbols indicate an increase. The upper left corner shows the change of false-positive contacts, with red hollow squares indicating the loss of a false-positive contact and red filled squares indicating the gain of a false-positive contact. The lower right corner shows the changes in false-negative contacts, with green hollow squares indicating the loss of a false-negative contact and green filled squares indicating the gain of a false-negative contact.

A. bAccuracy

B. sAccuracy:

Figure 5.

Change of accuracy of side-chain packing during atomistic refinement. The accuracy was calculated for (A, “bAccuracy”) the residues whose normalized solvent accessible surface areas are less than 0.2 in the X-ray crystal structure and (B, “sAccuracy”) the residues whose normalized solvent accessible surface areas are greater than 0.5 in the X-ray crystal structure. The accuracy was evaluated based on the side-chain dihedrals χ_1 (blue) and χ_2 (orange) of each residue. The average (solid line) and standard deviation (shadow) were calculated over a total of 10 independent refinement simulations.

Figure 6.

Number of minimally frustrated and highly frustrated interactions as a function of time during the all-atom refinement simulations. The average number of minimally frustrated interactions during the 10 all-atom refinement simulations is shown as a function of time in green for all prediction targets. The average number of highly frustrated interactions is shown in red. The standard deviations of these quantities are indicated with errorbars. For comparison, the number of minimally frustrated and highly frustrated interactions present in the corresponding crystal structures are shown as solid horizontal lines.

Figure 7.

Soft constraints from templates guide the folding of T0792 toward a native-like state. Free energy profiles of T0792 as a function of Q are shown using different models: AWSEM without template information (blue circles), AWSEM-Template with soft constraints (red circles), and AWSEM-Template with strong constraints (yellow circles).

Figure 8.

Expectation values of energy terms in the AWSEM-Template model as a function of Q for the protein T0792. The expectation values are shown for several different models: AWSEM without template information (blue circles), AWSEM-Template with soft constraints (red circles), and AWSEM-Template with strong constraints (yellow circles).

Table 1.

Summary of Structure Prediction Results for Coarse-Grained AWSEM-Template Structures and for AWSEM-Template Structures That Have Been Refined with All-Atom Simulations^a

Target Information					RMSD (Å)			
ID	Length	Solved Structure	Template Structure	Template Identity (%)	AWSEM-Template		Refined AWSEM-Template	
					MODELLER	I-TASSER	Best	Selected
T0251	111	1XG8	1J0F	16.1	7.64	6.62	2.44	3.44
IMBA	145	IMBA	1X3K	26.0	2.70	2.28	2.16	2.86
T0815	106	4U13	3FH1	28.7	3.56	3.07	2.70	2.73
T0833	136	4R03	4R8O	29.2	5.11	3.31	3.37	4.71
T0792	80	5A49	3RCO	32.8	7.70	4.49	2.59	5.05
T0784	155	4QEY	2LRG	57.0	1.81	1.96	1.21	1.50
T0803	139	4OGM	4PE2	63.1	6.94	3.95	2.81	5.74
T0766	130	4Q53	4ORL	65.5	2.18	2.34	1.16	1.37
Target Information					CE-RMSD (Å/°)			
ID	Length	Solved Structure	Template Structure	Template Identity (%)	AWSEM-Template		Refined AWSEM-Template	
					MODELLER	I-TASSER	Best	Selected
T0251	111	1XG8	1J0F	16.1	4.83/72	4.58/86	2.36/96	2.20/96
IMBA	145	IMBA	1X3K	26.0	2.20/94	1.82/99	2.01/93	2.78/93
T0815	106	4U13	3FH1	28.7	2.46/98	2.02/98	2.46/97	2.16/91
T0833	136	4R03	4R8O	29.2	3.46/76	2.14/76	2.93/96	4.22/89
T0792	80	5A49	3RCO	32.8	2.36/90	2.26/90	1.63/90	2.15/90
T0784	155	4QEY	2LRG	57.0	0.85/77	0.79/77	1.08/96	1.42/96
T0803	139	4OGM	4PE2	63.1	3.44/86	3.36/92	2.61/96	3.40/90
T0766	130	4Q53	4ORL	65.5	0.81/80	1.08/80	0.97/96	1.08/96
Target Information					GDT-TS			
ID	Length	Solved Structure	Template Structure	Template Identity (%)	AWSEM-Template		Refined AWSEM-Template	
					MODELLER	I-TASSER	Best	Selected
T0251	111	1XG8	1J0F	16.1	45.72	45.05	75.22	66.21
IMBA	145	IMBA	1X3K	26.0	71.23	77.57	72.43	65.24
T0815	106	4U13	3FH1	28.7	76.18	79.01	74.29	72.17

Target Information										RMSD (Å)			
ID	Length	Solved Structure	Template Structure	Template Identity (%)	MODELLER	I-TASSER	AWSEM-Template		Refined AWSEM-Template		Best	Selected	Selected
							Best	Selected	Best	Selected			
T0833	136	4R03	4R80	29.2	64.58	79.86	61.58	47.92	53.71	47.92	53.71	47.92	47.92
T0792	80	5A49	3RCO	32.8	64.69	73.12	80.31	74.06	84.38	81.25	84.38	81.25	81.25
T0784	155	4QEY	2LRG	57.0	95.20	95.80	91.80	86.00	92.60	92.20	92.60	92.20	92.20
T0803	139	4OGM	4PE2	63.1	78.92	71.08	79.66	74.07	78.17	74.18	78.17	74.18	74.18
T0766	130	4Q53	4ORL	65.5	94.44	90.28	93.06	90.97	95.60	90.97	95.60	90.97	90.97

Target Information										Q			
ID	Length	Solved Structure	Template Structure	Template Identity (%)	MODELLER	I-TASSER	AWSEM-Template		Refined AWSEM-Template		Best	Selected	Selected
							Best	Selected	Best	Selected			
T0251	111	1XG8	1J0F	16.1	0.40	0.44	0.71	0.61	0.69	0.61	0.69	0.61	0.61
IMBA	145	1MBA	1X3K	26.0	0.70	0.75	0.72	0.64	0.78	0.64	0.78	0.73	0.73
T0815	106	4U13	3FH1	28.7	0.71	0.73	0.77	0.69	0.79	0.69	0.79	0.65	0.65
T0833	136	4R03	4R80	29.2	0.84	0.94	0.72	0.54	0.56	0.54	0.56	0.52	0.52
T0792	80	5A49	3RCO	32.8	0.58	0.65	0.75	0.65	0.77	0.65	0.77	0.76	0.76
T0784	155	4QEY	2LRG	57.0	0.92	0.92	0.90	0.87	0.90	0.87	0.90	0.88	0.88
T0803	139	4OGM	4PE2	63.1	0.69	0.68	0.77	0.66	0.70	0.66	0.70	0.65	0.65
T0766	130	4Q53	4ORL	65.5	0.93	0.89	0.94	0.91	0.93	0.91	0.93	0.90	0.90

^aStructure prediction results from MODELLER and I-TASSER are included for the sake of comparison. MODELLER and I-TASSER were run with default settings and utilized the same template that was employed for the AWSEM-Template predictions ("Template Structure"). All structures for each given target were compared to the same experimentally solved structure ("Solved Structure"). Four structure quality measures are shown: RMSD, CE-RMSD, GDT-TS, and Q. Structure quality metrics were computed for the best sampled AWSEM-Template structures ("Best") and the AWSEM-Template structures that were chosen without reference to the native structure ("Selected"). Best structures were chosen from the coarse-grained ("AWSEM-Template") and all-atom refinement ("Refined AWSEM-Template") stages of prediction by identifying the structure having the highest Q value. Selected structures were chosen by clustering of low energy structures for the coarse-grained prediction step or by a machine learning algorithm that uses energies as features for the all-atom refinement step.