Published in final edited form as:

Proteins. 2009 August 1; 76(2): 331–342. doi:10.1002/prot.22348.

Simulating protein folding initiation sites using an alpha-carbon-only knowledge-based force field

Patrick M. Buck and Christopher Bystroff*

Department of Biology, Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, New York

Abstract

Protein folding is a hierarchical process where structure forms locally first, then globally. Some short sequence segments initiate folding through strong structural preferences that are independent of their three-dimensional context in proteins. We have constructed a knowledge-based force field in which the energy functions are conditional on local sequence patterns, as expressed in the hidden Markov model for local structure (HMMSTR). Carbon-alpha force field (CALF) builds sequence specific statistical potentials based on database frequencies for α -carbon virtual bond opening and dihedral angles, pairwise contacts and hydrogen bond donor-acceptor pairs, and simulates folding via Brownian dynamics. We introduce hydrogen bond donor and acceptor potentials as α -carbon probability fields that are conditional on the predicted local sequence. Constant temperature simulations were carried out using 27 peptides selected as putative folding initiation sites, each 12 residues in length, representing several different local structure motifs. Each 0.6 μ s trajectory was clustered based on structure. Simulation convergence or representativeness was assessed by subdividing trajectories and comparing clusters. For 21 of the 27 sequences, the largest cluster made up more than half of the total trajectory. Of these 21 sequences, 14 had cluster centers that were at most 2.6 Å root mean square deviation (RMSD) from their native structure in the corresponding full-length protein. To assess the adequacy of the energy function on nonlocal interactions, 11 full length native structures were relaxed using Brownian dynamics simulations. Equilibrated structures deviated from their native states but retained their overall topology and compactness. A simple potential that folds proteins locally and stabilizes proteins globally may enable a more realistic understanding of hierarchical folding pathways.

Keywords

protein structure prediction; protein folding; knowledge-based potential; simplified model; Brownian dynamics; alpha carbon model; statistical potential; folding pathways; local structure; hydrogen bond potential

*Correspondence to: Christopher Bystroff, 3C07 Jonsson-Rowland Science Center, Department of Biology, 110 Eighth Street, Rensselaer Polytechnic Institute, Troy, NY 12180. bystrc@rpi.edu.

INTRODUCTION

Peptide sequences less than 20 residues in length can have strong structural preferences that are independent of nonlocal interactions, as shown by NMR,^{1–3} and simulation studies.^{4,5} It is thought that sequence patterns for these peptides in the context of a parent sequence become structured early in folding and exist in their native conformation in unfolded proteins. Some of these short sequence patterns, 3–19 residues in length, have been captured in a structural motif library called I-sites (initiation sites)⁶ and the associated hidden Markov model (HMMSTR)⁷ which describes the adjacencies and dependencies of I-sites motifs in the protein structure database.

The existence of strong sequence-structure correlations in the database should enable us to develop and test folding potentials for template-free protein structure prediction.

Knowledge-based potentials based on the statistical occurrences of structural properties in native proteins have proven to be the most successful approach to protein structure prediction.^{8,9} They provide an attractive alternative to molecular mechanics based force fields because they require significantly less detail and do not rely on the assumption that thermodynamic calculations can be extrapolated to model complex macromolecular interactions.¹⁰

In general, there are two types of statistical potentials, one that includes interactions for all residue-specific atom types^{11–13} and others that are coarse-grained, having only one or two residue-specific interaction centers created by “uniting” atoms.^{14–16} Lumping atoms together into interaction centers has been particularly useful in protein structure prediction as it significantly reduces the cost of computing a search for the native structure in hyper-dimensional protein configuration space. However, simplified representations lose some of the detailed geometric dependence on the calculated energy of the system. In particular, hydrogen bond energy depends strongly on the orientation, not just on the distance between interaction centers.

In this study we use a reduced protein representation for folding simulations in an α -carbon-only folding potential. Peptide residues are treated as beads on a string, with backbone atoms for each residue lumped into a single interaction center located at the position of each α -carbon. Additionally, side chain centroids are approximated by the mean relative positions of the side chain centers of mass. Such a model can significantly reduce the cost of computing trajectories to visualize long time-scale dynamics such as in protein folding.¹⁷

In protein structure prediction, effort has been made to discretize the conformational space by fragment insertion Monte Carlo¹⁸ or chain build-up^{19,20} in folding simulations. Although quite successful, these methods may ignore intermediates along the folding pathway by strictly optimizing the global fold energy.²¹ Modeling folding pathways is essential to the understanding of folding kinetics and kinetic stability. Non-native intermediates along the folding pathways may be required in the folding of some knotted proteins.²²

Recently, there has been increased interest in simulating the physical folding process using reduced protein representations and coarse-grained potentials.^{23–25} To our knowledge, no α -carbon-only statistical potential for folding by molecular dynamics simulations has ever

before been tried, as very few of the published statistical potentials act solely on α -carbons,^{26–28} hinting at the difficulty of calculating a realistic energy using a reduced model.

METHODS

Model representation and dynamics

We use a reduced, α -carbon protein representation based on the early work of Levitt.²⁹ When peptide bonds are in the *trans* conformation, the distance between α -carbons is fixed at 3.8 Å. Thus treating the five-heavy atom peptide planar group as a virtual bond is not unreasonable. Adjacent peptide planar groups are related by backbone angles (ϕ , ψ) of connected α -carbons. For convenience and without loss of generality, peptide structure can also be represented by the angles between virtual bonds. CALF simulations use both the bond opening angle δ (), defined by three consecutive α -carbons, and the dihedral angle θ (θ) defined over four adjacent α -carbons (see Fig. 1). Side-chain centers of mass or centroids were calculated at each step in the simulation using the method of Park.³⁰ Centroid positions were assigned to the second through the next-to-last residue of the target sequence, and were updated at each time step.

Simulations were done using Over-damped Langevin dynamics or Brownian dynamics. Because protein folding is a slow process relative to chain fluctuations, momentum can effectively be ignored. Random forces ($\mathbf{R}$) are added into the simulation to replace the effect of missing interactions lost by grouping atoms together and ignoring solvent.

$$\frac{d\mathbf{r}_i}{dt} = \frac{\mathbf{F}_i(\mathbf{r}(t))}{\xi_i} + \mathbf{R}_i \quad (1)$$

Here the systematic force ($\mathbf{F}_i$) is determined by taking the derivative of the potential $\mathbf{F}_i = -\nabla_{\mathbf{r}_i} V_{\text{Total}}$ with respect to the current model configuration ($\mathbf{r}(t)$). The friction coefficient (ξ_i) was calculated using Stoke's Law with an atomic radius (α_i) taken as the α -carbon Van der Waals radius (5.0 Å) plus one water molecule (1.4 Å). The viscosity of water (η) was set at 0.001 kg m⁻¹ s⁻¹.

$$\xi_i = 6\alpha\pi_i\eta \quad (2)$$

Random forces ($\mathbf{R}_i$) were taken from a Gaussian distribution of vectors ($\mathbf{R}^G$) with zero mean and variance $2k_B T / \xi_i$ where k_B is the Boltzmann constant, T is temperature in Kelvin and t the time step. No memory or hydrodynamic coupling effects were included. Atomic shifts were updated at each timestep as follows:

$$\mathbf{r}_i(t+\Delta t) = \mathbf{r}_i(t) + \frac{\Delta t}{\xi_i} \mathbf{F}_i(\mathbf{r}(t)) + \sqrt{2k_B T \frac{\Delta t}{\xi_i}} \mathbf{R}^G \quad (3)$$

Virtual bonds between α -carbons were constrained at 3.8 Å using LINCS³¹ during folding simulations.

Peptide folding

Protein segments or peptide sequences predicted to be autonomous folding units were extracted from their parent structures according to the PDB numbering scheme attached to the parent code (Table I). Because the hydrogen bond potential is defined over three-residue segments, one residue was added to each terminus. For example, 14 positions are necessary to model the central 12 residues. In the results presented here, root mean square deviation (RMSD) calculations were performed only on the central 12 residues although 14 α -carbons were used during simulations. Starting from an initial random configuration, folding simulations were run for 0.6 μ s using a time step of 30 fs. The temperature for all simulations was set to 10 K. The units of temperature are relative to the parameters of the energy terms. The temperature, although unphysical, is appropriate for weighting parameters optimized near a value of 1. To ensure adequate sampling, one trajectory for each peptide sequence is run until simulation convergence (discussed in Results section).

Global stability

Proteins are stable under native conditions. Therefore their native structures should also be stable in the presence of an adequate energy function using thermal fluctuations that are appropriate for folding. To test the adequacy of the CALF energy function to stabilize full length proteins, we chose 11 proteins with less than 100 residues that represent a variety of structural classes. Simulations starting from native were run for 60 ns using a time step of 30 fs. The temperature for all simulations was set to 2 K. One relaxation simulation was run for each structure and all structures ceased deviating from native during the trajectory.

Force field

The CALF energy function has three sequence specific energy terms V_{Eij} , V_{Angs} , V_{HB} and the physical interaction energy Van der Waals (V_{VDW}) in its complete form. The derivation of each energetic term and the optimization of weighting parameters w are discussed in the following sections.

$$V_{Total} = W_{Eij} V_{Eij} + W_{Angs} V_{Angs} + W_{HB} V_{HB} + W_{VDW} V_{VDW} \quad (4)$$

Van der Waals potential, V_{VDW}

Alpha-carbons repel each other using the repulsive term of a 12–6 Lennard-Jones potential with a radius of 5 Å ($\sigma_{\alpha\alpha}$). The hard sphere potential prevents any atomic overlap and maintains the peptide in physically realistic configurations. The exaggerated radius size eliminates chain crossing. Van der Waals energy was calculated for all residues pairs at least three sequence positions apart and with distance (r_{ij}) less than 14 Å.

$$V_{VDW} = \sum_{\substack{ij \\ j > i+2}} \left(\frac{\sigma_{\alpha\alpha}}{r_{ij}} \right)^{12} \quad (5)$$

Side-chain centroid to centroid and centroid to α -carbon repulsions were modeled using a the repulsive term of a 12–6 Lennard-Jones potential. Forces experienced by side-chain centroids were summed to the shifts of their corresponding α -carbons. A cutoff of 14 Å was used for calculating centroid Van der Waals energy. The pair-wise radii parameters σ_{ij} are residue specific and were taken from Park.³⁰

Database of known protein structures, B

The sequence specific energy terms V_{Angs} , V_{HB} , and V_{Eij} are statistical potentials derived by inverting the Boltzmann distribution and using conditional probabilities from our dataset of protein structures (B). Structural data used to build statistical potentials were taken from a nonredundant set of 1000 structures with better than 2.0 Å resolution, found on the PISCES Culled Server.³² Any two sequences in this set have less than 25% sequence identity. The database set is treated as a representative sample of sequence-structure correlations at the local level. Statistical potentials were derived by summing conditional probabilities. Initial tests showed that the summed probability distributions, used to derive CALF potentials, did not change significantly when the database was doubled in size, demonstrating that the database is representative for the purposes of this study.

Local structure predictions using HMMSTR

The probabilities of each structural descriptor (virtual bond opening and dihedral angles, hydrogen bond donor-acceptor positions, pair-wise distances) are conditional upon the sequence through HMMSTR, a hidden Markov model (HMM) for local structure. HMMSTR models all common local sequence-structure correlations in proteins. Briefly, a HMM was constructed by linking the sequence-structure motifs within the I-sites library, initially performing all possible pair-wise alignments between motifs, then linking all motifs together to form a HMM. Each Markov state emits backbone angles and a sequence probability distribution (profile) for a single position, as an expression of the local sequence-structure correlation. Standard HMM algorithms are used to make predictions of structure based on sequence.

Single sequences or multiple sequence alignments expressed as profiles were used as input to HMMSTR (in earlier studies profiles were shown to improve the results of local structure predictions over single sequences). Profiles were generated as one of the output options of the database search program Psi-BLAST.³³ To identify more distant homologs, sequences with an e-value of 0.1 or better were kept after two search iterations.

Using the forward/backward algorithm,³⁴ the *a posteriori* probability (γ) was calculated for each Markov state q and each sequence position t .

$$\gamma_{q,t} = P(q|t) \quad (6)$$

Each of the 280 Markov states in HMMSTR defines a sequence-structure context. For example, one HMMSTR Markov state represents the conserved glycine position in a Type-2 diverging β -turn. Another state represents the amphipathic position between the hydrophobic side and the polar side of an amphipathic α -helix. Local structure contexts manifest

preferences in backbone virtual angles, contact distances, and virtual hydrogen bonding geometry. HMMSTR probabilities (γ) were calculated for sequences from the database (γ^B) and from the target (γ^T).

Virtual backbone angle potential, V_{Angs}

Position specific virtual bond opening angle (Δ) and dihedral (θ) potentials for a target sequence (T), position i , were generated by summing joint conditional probabilities $\gamma_{q,i}^T \gamma_{q,j}^B$ over all HMMSTR states q and over all database positions j , whose angle set falls within the bin (Δ, θ). Bins were created by dividing the angle space into 10° intervals over the range of $0^\circ < \Delta < 360^\circ$ and $0^\circ < \theta < 180^\circ$, giving $n_{\text{bins}} = 36 \times 18 = 648$. Summing joint probabilities assures that only database positions with local sequence similarity were counted.

$$P_i(\Delta_j, \theta_j) = \sum_{\Delta_j, \theta_j} \sum_q \gamma_{q,i}^T \cdot \gamma_{q,j}^B \quad (7)$$

After summing, P_i was normalized to one. The virtual angle energy was then expressed as a log-likelihood ratio, where the expected value for each bin is $1/n_{\text{bins}}$. Adding one to the likelihood ratio sets the maximum energy to zero. Virtual angle energy was smoothed using Gaussian filtering³⁵ (see Fig. 2).

$$E_i(\Delta, \theta) = -k_B T \cdot \ln \left(\frac{N_i(\Delta, \theta)}{n_{\text{bins}}^{-1}} + 1 \right) \quad (8)$$

The total backbone angle energy is summed over positions two through $n-2$ where n is the total number of positions,

$$V_{\text{Angs}} = \sum_{i=2}^{n-2} E_i(\Delta_i, \theta_i) \quad (9)$$

and (Δ_i, θ_i) are the current angles at position i . Derivatives of the virtual angle energy were taken numerically using a five-point stencil³⁵ and gradient derivatives were taken analytically.³⁶ Because it is extremely rare to observe a delta angle outside of the range $70-150^\circ$ we set $V_{\text{Angs}}/\Delta_i = 1$ for $\Delta_i > 150^\circ$ and $V_{\text{Angs}}/\Delta_i = -1$ for $\Delta_i < 70^\circ$, to keep all target delta angles physically realistic (not shown in Fig. 2).

Hydrogen bond potential, V_{HB}

Hydrogen bonds were identified in the database by searching $2.8 \text{ \AA}$ away from all backbone nitrogen atoms. A hydrogen bond was declared when a backbone carbonyl was met and its members (α -carbons) were designated as donor and acceptor, respectively (see Fig. 3). Each residue was allowed one donor and one acceptor hydrogen bond; Bifurcated hydrogen bonds were ignored. Donor-acceptor pairs were collected for all proteins in the data set. The coordinates of the donor residue α -carbon were transformed into the frame of reference of the acceptor residue using adjacent α -carbons and vice versa.

Position specific donor potentials (**D**) for a target sequence (T) position i were generated by summing the joint conditional probabilities $\gamma_{q,i}^T \gamma_{q,j}^B$ over all HMMSTR states q and over all acceptor database (B) positions j whose donor coordinates (x_d, y_d, z_d) after transformation into the acceptor frame of reference fall within the three-dimensional bin (x, y, z) . Bins were created by dividing the $20^3 \text{ \AA}$ Cartesian space centered at the acceptor j into $0.5^3 \text{ \AA}$ bins, giving a total of $n_{\text{bins}} = 64,000$ bins (see Fig. 4). Similarly, position specific acceptor potentials **A_i** were generated by binning acceptors positions around an α -carbon donor position. After summing, both **A_i** and **D_i** were normalized to one. Summing joint probabilities ensures that only database positions with local sequence similarity to the target were counted.

$$\mathbf{D}_i(x_d, y_d, z_d) = \sum_j \sum_q \gamma_{q,i}^T \cdot \gamma_{q,j}^B \quad (10)$$

The hydrogen bond energy V_{HB} was calculated using both log-likelihood ratios for geometry-specific donor and acceptor probabilities, summed over the current hydrogen bond network (hbn). Adding one to the likelihood ratio sets the maximum energy to zero. All position specific acceptor and donor energies were smoothed with Gaussian filtering and derivatives in the transformed space were taken numerically using a five point stencil.

$$V_{\text{HB}} = \sum_{\text{hbn}} -k_B T \left[\ln \left(\frac{\mathbf{A}_i(\mathbf{r}_j)}{n_{\text{bins}}^{-1}} + 1 \right) + \ln \left(\frac{\mathbf{D}_i(\mathbf{r}_j)}{n_{\text{bins}}^{-1}} + 1 \right) \right] \quad (11)$$

The hbn was updated periodically (~ 1000 update steps) to allow hydrogen bond partnering to change. To establish new donor-acceptor pairings, hydrogen bond energies $\mathbf{H}_{ij} = -k_B T [\ln(\mathbf{D}_i(\mathbf{r}_j) + 1) + \ln(\mathbf{A}_j(\mathbf{r}_i) + 1)]$ were calculated for all ij pairs at a sequence separation of $|i-j| > 3$, assuming that i could potentially donate to any position j . The lowest energy pairing $\mathbf{H}_{ij}$ was assigned as the first hydrogen bond and all other possible partnering for those members were set to zero, because bifurcated hydrogen bonds were not allowed. The process was continued until all nonzero $\mathbf{H}_{ij}$ were found, resulting in the new hbn.

Contact potential, V_{Eij}

The contact energy term V_{Eij} is based on the previous work of Shao.³⁸ The pair-wise contact potential between any two HMMSTR states p and q ($\mathbf{G}(p, q, s)$) was calculated as the log of the joint conditional probability of these two states in contacting residues where $r_{i,(i+s)}$ less than $8 \text{ \AA}$ at a sequence separation s and over all proteins in the database.

$$\mathbf{G}(p, q, s) = -\ln \frac{\sum_B \sum_{\{i: r_{i,(i+s)} < 8 \text{ \AA}\}} \gamma_{(i,p)} \gamma_{(i+s,q)}}{\sum_B \sum_i \gamma_{(i,p)} \gamma_{(i+s,q)}} \quad (12)$$

The contact potential $\mathbf{E}(i,j)$ between residues i and j in the target was calculated as the probability weighted sum of the pair-wise potential function $\mathbf{G}$.

$$\mathbf{E}_{ij}(i,j) = \sum_p \sum_q \gamma_{(i,p)} \gamma_{(j,q)} \mathbf{G}(p,q,s) \quad (13)$$

For use in these simulations, a smooth function was formulated that limits to $\mathbf{E}_{ij}$ as r_{ij} approaches zero, and to zero as the r_{ij} approaches 13 Å. Contact energies are summed over all positions with a sequence separation of at least three.

$$V_{\mathbf{E}_{ij}} = \sum_{\substack{ij \\ j > i+2}} -k_B T \cdot \mathbf{E}_{ij}(i,j) \cdot \left[\frac{\arctan(0.7 \cdot (r_{ij}-8))}{\pi} - 0.5 \right] \quad (14)$$

Parameter optimization

The parameters w Eq. (4) were derived empirically based on global structure relaxation simulations for 1pgb. With a temperature and time step set before parameterizing, the weight for Van der Waals w_{vdw} was chosen to prevent chain crossing and atomic overlap. By fixing w_{vdw} , the contact energy parameter w_{Eij} was optimized to maintain the radius of gyration for the native structure. This is important because over collapsing structures can significantly improve RMSD calculations although unjustifiably. Keeping w_{vdw} and w_{Eij} fixed, the parameters w_{Angs} and w_{HB} were optimized to increase the energetic gap between the native structure and all non-native structures encountered during relaxation simulations. Empirically deriving weighting parameters by relaxing one native structure has been done before,³⁹ and the process was automated by relaxing a large training set.⁴⁰ The same set of optimized weights for 1pgb were used for all native structure relaxation simulations and peptide folding simulations. Their values are listed here for convenience. ($w_{\text{vdw}} = 0.01$, $w_{\text{Angs}} = 1.0$, $w_{\text{HB}} = 0.1$, $w_{\text{Eij}} = 0.35$)

RESULTS AND DISCUSSION

The new knowledge-based energy function includes potentials for virtual bond opening and dihedral angles, hydrogen bond donor and acceptor probability fields, and a local-structure dependent pair-wise potential. The potentials are fit to probabilities summed from a non-redundant set of known proteins. All are position-specific and conditional on their unique amino acid sequences as opposed to being residue-specific. In this study, we fold 27 short protein segments of length 12 that were predicted to be autonomous folding units. As a test of the ability of the energy function to stabilize nonlocal interactions, 11 full length native structures were relaxed using Brownian dynamics. All simulations converged within reasonable CPU time. Most of the native structural preference was accounted for by local virtual bond angle preferences and predicted contacts, but the inclusion of a hydrogen bond probability field significantly decreased the deviation of native structures during relaxation simulations.

There are two problems recognized in the protein structure prediction community that should be addressed before making a template-free prediction (1) Search/Simulation Convergence: Does a particular method have adequate sampling such that the native structure can actually be reached during a simulation? And how should it be determined that further sampling will not lead to additional conformations? (2) Discrimination/Prediction: If the native structure can be reached, does it have a lower energy than alternative conformations such that it can be identified as native? Or how should the predicted protein structure be chosen from a large set of candidate structures generated over the course of a simulation? Although it is probably safe to assume that the native ensemble is narrowly distributed around a mean structure, selecting only the global minimum in energy ignores the larger ensemble of near-native structures⁴¹ that are more likely to be encountered during simulations. Many successfully template-free prediction strategies use clustering⁴² to approximate this kinetically accessible free energy minimum.

Peptide predictions/clustering

Peptide predictions were made by clustering 10,000 structures sampled 6 ns apart from simulation trajectories and choosing the largest cluster center as the predicted structure. There have been a variety of clustering methods previously reported to determine the highest population conformation and thus lowest free energy from simulations.^{5,41,43,44} The clustering method for this study was previously described by Daura⁴⁵ and begins by building a pair-wise distance matrix containing the RMSD between all structures submitted for clustering. The first cluster center was chosen as the structure with the most neighbors below a distance cut-off of 2.0 Å. All structures clustered were removed from the distance matrix and the process was repeated until all structures were clustered.

In Table I, RMSD values for the most representative conformation during simulations based on clustering are shown. Additionally, the confidence of our predictions was assessed by determining how much of the simulation was spent in the largest cluster center compared to all other clusters. If more than half of those structures submitted for clustering fell into the largest cluster then those protein segments were regarded as having a structural preference. Of the 27 protein segments predicted, 21 were found to have trajectories where more than half of the total simulation could be clustered into one conformation.

Peptide simulation convergence

One outstanding problem in template-free protein structure prediction has been adequate sampling of conformational space.⁹ The simplest way to improve sampling is to allow simulations to continue longer. Still, there needs to be a way to know when simulations have converged such that no more new structures will be discovered by allowing simulations to proceed indefinitely. There have been a variety of equilibrium sampling or simulation convergence measures previously reported^{46–48} and most check for sampling consistency by dividing trajectories in half and checking for structural diversity. Measures of simulation convergence have been used to validate the use of single trajectories for dynamic processes that occur on long time-scales. In this study, a similar strategy was used by separating structures from each cluster into two subclusters according to whether they seem in the first or second half of the trajectory (see Fig. 5). If the subcluster of cluster one from the first half

of the trajectory was the largest subcluster of all first half subclusters and the same was true for the second half subcluster of cluster one compared to all other second half subclusters, then Stage-1 equilibrium was achieved. Stage-2 equilibrium was achieved when this was also true for the second largest cluster. The method forces the relative ordering of cluster sizes in trajectory halves to be preserved before simulations have converged. A more stringent test of convergence could preserve both the relative size and ordering of clusters in trajectory halves.

Of the 27 protein segments studied, 17 reached stage-2 equilibrium and nine reached stage-1 equilibrium (Table I). The remaining peptide, that did not pass either test, both showed little or no structural preference, making stage-1 simulation convergence as it is defined here difficult to achieve. There were six cases where clustering indicated a structure preference but the trajectory did not reach stage-2 equilibrium, either because the second and third largest clusters were nearly equal in size or because the second largest cluster was much smaller than the first. Although these results cannot rule out the possibility that new structures would be encountered by running longer simulations, the possibility that one of the new structures would be the structural preferred one is remote.

In summary, 14 of the 21 protein segments found to have structural preferences had cluster centers that were at most 2.6 Å RMSD from native. Some improvements were seen in the second largest cluster for some protein segments (not shown) although generally the largest cluster center was closest to native. Scatter plots of energy versus RMSD to native for all 27 protein segment trajectories are shown in Figure 6. In many cases, the densest sampling was both closer to native and lower in energy. For well predicted protein segments, trajectories sampled configuration space that is primarily higher in energy than the native. This feature almost certainly ensures that the native structure will have the lowest free energy during simulations. For these peptides, a correlation can be observed between the energy and distance from native, which suggests a funnel-like energy landscape that draws simulations towards their native state.

Peptide case studies

It is instructive to look at those cases where the preferred structure did not match the native. Each of the 27 protein segments studied here were chosen as putative folding initiation sites, blindly predicted using HMMSTR. The failure to match the lowest free energy state to the native state in these simulations could be a fault of HMMSTR, or of imperfections in the CALF energy function, or it could be an artifact of separating the peptide from the rest of the protein when performing simulations. HMMSTR measures the database correlation between sequence and structure, which should be high when the sequence in question, and all of its nearest neighbor sequences, are always found in the same structure. A high correlation is suggestive of an energetically stable folding motif, but it is not proof. CALF calculates the energy based on database probabilities of backbone virtual angles, contacts and virtual hydrogen bonds. A low energy reflects a high database probability when the conformation is viewed as the sum of its parts. In the majority of cases, folding simulations of putative initiation sites were in close agreement with the native segment structure. Given

the correctness of this majority, we may analyze the exceptions as artifacts of taking peptides out of their structural context.

Of the protein segments that showed a structural preference in simulations but deviated ($>2.6 \text{ \AA}$) from native, four were found to have important nonlocal interactions in the parent structure, which were of course missing in the peptide isolated simulations. For example, in the segment 1cfr_135-148, a strand + helix motif, the strand is prevented from contacting the helix in the parent structure because a nonlocal strand is contacting the helix instead. The predicted structure forms a second helical turn instead of preserving the eight residue β -strand. In the segment 1aoqA_176-189, a strand + loop + strand motif, nonlocal hydrogen bonds prevent strands from pairing with one another in the native structure. The largest cluster center prediction shows a N-capped helix, where the second strand hydrogen bonds back against the turn forming a single helical turn. Protein segments that have hydrogen bonding otherwise satisfied through nonlocal interactions or through solvent are difficult to predict. It is not surprising that predicted structures satisfy hydrogen bonds locally. These structures are easy to identify in Figure 6 as their energy is often much lower than their corresponding native structures.

In the segment 1mwpA_22-35, a diverging turn motif, strands on either side of the β -turn do not pair to form a hairpin but instead wrap around the protein core. A diverging turn has been shown to be stable in isolation using NMR,⁴⁹ despite being less compact than a similar length β -hairpin. Its stability is a result of interactions between three inwardly-pointing side-chains. Without detailed side-chains, the CALF energy function cannot capture specific side-chain interactions well. However, the presence of the diverging turn sequence pattern enforces virtual backbone angles that favor the diverging turn. For the protein segment 1ahjB_46-59, a strand + helix N-cap, no contacts are made between the strand and the helix in the native structure, a fact that was reproduced in the prediction. In the simulation, the strand hydrogen bonds back against the second helical turn, forming a three turn helix. The second largest cluster is indeed much closer to native ($2.2 \text{ \AA}$ RMSD) although with higher energy than the first, this cluster was thus less represented during the simulation.

In all, six simulations showed no structural preference by our metric, but it is interesting to note that two of the six contained a near native largest cluster center, with RMSD less than $3.5 \text{ \AA}$. For 1pgs_192-205, a rare turn motif, the scatter plot indicates sampling of mostly an extended structure. On inspection of the contact energy, attractive interactions that would stabilize the closed native form are missing (see Fig. 7). The root of each potential is the assignment of Markov states according to HMMSTR, which in this case fails. It is not surprising that our knowledge-based potential fails to predict a structure that has poor database statistics, and this presents a challenge for the future of coarse-grained knowledge-based potentials.

Forcefield versus sequence similarity

Because the database is a representative set of sequences and structures, it is reasonable to think that simply searching the database using the peptide sequence might result in a good local structure prediction. In fact a similar strategy was used to build the I-sites library. To test the likelihood of finding structures with RMSD $> 2.6 \text{ \AA}$ using sequence similarity alone,

we performed all pairwise ungapped sequence alignments against the database using the BLOSUM40 substitution matrix. Structures for those sequences were then compared against the native for each of the peptides in this study. We found that an average 9.6% of the top 100 sequence-similar peptides had RMSD ≤ 2.6 Å, whereas at least 44.7% of the peptide structures from CALF simulations passed this test. For sequences of length 12, the force field captures local structure preferences better than simple sequence similarity alone.

Native structure relaxation

The ability of an energy function to stabilize native protein structures is adequate if structures do not unfold during relaxation simulations starting from the experimental native state. Such a test is not trivial to pass, as even all-atom, molecular mechanic force fields can cause structures to drift far away from native during relaxation simulations.⁵⁰ It is therefore a challenge upon structure predictors to create potential functions such that native structures have the lowest energy within their near-native energy basin. In an attempt to meet this criterion, the CALF energy function was parameterized to minimize relaxation drift for one native structure (1 pgb), and tested on a set of 10 other proteins. To our knowledge, this test has not previously been applied to α -carbon-only models.

Table II shows the results of relaxing a diverse set of small proteins less than 100 residues. Three metrics are provided to evaluate deviations from native, including RMSD, Q (the fraction of native contacts⁵¹), and TM-score (a measure of structural similarity that ranges between 0 and 1.0 for identical structures and a value of 0.4 or greater for a pair of significantly related structures⁵²). Of the 11 native structures relaxed, six have RMSD to native less than 6.5 Å, indicating topological similarity. Although 2a3d deviates from native greater than 6.5 Å, the converged structure conserves 87.8% of its native contacts (Q value). All structures, including those that deviate far from native, maintain good β -sheet geometry and α -helix periodicity, primarily due to the three-dimensional hydrogen bond potential. In general we find that the energy function is sufficient to maintain local structure, topology, and overall compactness.

For the four structures that deviated from native, imperfections in the contact potential was the primary cause. For example, two-sheet beta sandwiches 1fna and 1who are structurally very similar. However during simulations, 1fna over-collapses while 1who maintains a native-like radius of gyration. The relaxed structure for 1fna shows new hydrogen bond partnering that is comparable in energy to native, although over-collapsing causes some strand rearrangement in the sheets. The case for ubiquitin (1ubq) is similar. In the relaxed molecule, the four-stranded β -sheet is broken up by the attraction of the hairpin to the helix. Some structures over-collapse while others, such as 1shg, expand. In this case, the contact potential is insufficient to stabilize native contacts as evidenced by the low Q value of the relaxed structure. Deviations from native for these structures suggest the need for improvements in the sequence dependence, and perhaps orientation dependence, of the contact potential.

Hydrogen bond energy

Initial studies, using an α -carbon-only potential based on backbone virtual angles, Van der Waals repulsion and contact energy terms, but without an orientation dependent hydrogen bond term, showed errors in strand alignment of β -sheets and other irregularities that could be traced to poor hydrogen bonding geometry. For example, three β -strands, all rich in nonpolar side-chains, would arrange themselves in a collagen-like triple helix rather than in a sheet. The backbone angles permitted this and it increased the total number of contacts made between strands. To capture the directional nature of hydrogen bonds, three-dimensional energy fields were imposed for both donor and acceptor residues.

Others have shown that the inclusion of an α -carbon-only hydrogen bond potential significantly improves the cooperativity and kinetics of the folding reaction for proteins G and L.⁵³ However in their model, hydrogen bonds are preferentially perpendicular to the α -carbon chain. This results in β -sheets that lack the characteristic right-handed twist, and α -helices with a helical pitch of 4.0 rather than 3.6 residues per turn. Our hydrogen bond potential correctly captures the right-handed twist of β -sheets and the pitch of α -helices. Using the sequence-structure database to build hydrogen bond potentials more accurately represents the secondary structure dependence of hydrogen bond orientation.

CONCLUSIONS

Folding a diverse set of short protein segments is prerequisite to developing a hierarchical folding model for larger proteins. It has long been thought that proteins fold locally first, forming secondary structures which are then able to nucleate tertiary contacting.^{54,55} Recently, it has been reported that this type of folding mechanism could be implemented in a procedure called zipping and assembly.⁵⁶ Briefly, peptide chains are broken up into short segments and folded independently. Segments that show structure preferences are zipped together by setting distance restraints on hydrophobic residues. Simulations are then restarted by growing those segments out by adding more residues to either side of the segment or assembling previously zipped segments together. The success of folding a diverse set of protein segments in the current study indicates that the zipping and assembly technique could also be implemented with the CALF energy function. In finding the native conformation in simulations of several different local structure motifs that are expected to fold autonomously, and that a diverse set of global native structures remains folded after energy relaxation, the force field promises generality and provides hope that a simple representation can be used to fold full-length proteins.

Acknowledgments

Grant sponsor: NSF; Grant number: DBI-0448072.

The authors thank Ryan Day for helpful comments on this manuscript.

Abbreviations

CALF	carbon-alpha based force field
-------------	--------------------------------

HMMSTR	hidden Markov model for local structure
RMSD	root mean square deviation

References

1. Munoz V, Blanco FJ, Serrano L. The hydrophobic-staple motif and a role for loop-residues in alpha-helix stability and protein folding. *Nat Struct Biol.* 1995; 2:380–385. [PubMed: 7664095]
2. Blanco FJ, Rivas G, Serrano L. A short linear peptide that folds into a native stable beta-hairpin in aqueous solution. *Nat Struct Biol.* 1994; 1:584–590. [PubMed: 7634098]
3. Viguera AR, Serrano L. Experimental analysis of the Schellman motif. *J Mol Biol.* 1995; 251:150–160. [PubMed: 7643384]
4. Bystroff C, Garde S. Helix propensities of short peptides: molecular dynamics versus bioinformatics. *Proteins.* 2003; 50:552–562. [PubMed: 12577261]
5. Ho BK, Dill KA. Folding very short peptides using molecular dynamics. *PLoS Comput Biol.* 2006; 2:e27. [PubMed: 16617376]
6. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol.* 1998; 281:565–577. [PubMed: 9698570]
7. Bystroff C, Thorsson V, Baker D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J Mol Biol.* 2000; 301:173–190. [PubMed: 10926500]
8. Skolnick J. In quest of an empirical potential for protein structure prediction. *Curr Opin Struct Biol.* 2006; 16:166–171. [PubMed: 16524716]
9. Bradley P, Misura KM, Baker D. Toward high-resolution de novo structure prediction for small proteins. *Science.* 2005; 309:1868–1871. [PubMed: 16166519]
10. Sippl MJ. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comput Aided Mol Des.* 1993; 7:473–501. [PubMed: 8229096]
11. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* 2002; 11:2714–2726. [PubMed: 12381853]
12. Qiu J, Elber R. Atomically detailed potentials to recognize native and approximate protein structures. *Proteins.* 2005; 61:44–55. [PubMed: 16080157]
13. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol.* 1998; 275:895–916. [PubMed: 9480776]
14. Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol.* 1996; 256:623–644. [PubMed: 8604144]
15. Liwo A, Oldziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. A united-residue force field for off-lattice protein-structure simulations. 1. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J Comput Chem.* 1997; 18:849–873.
16. Tobi D, Elber R. Distance-dependent, pair potential for protein folding: results from linear optimization. *Proteins: Struct Funct Genet.* 2000; 41:40–46. [PubMed: 10944392]
17. Tozzini V. Coarse-grained models for proteins. *Curr Opin Struct Biol.* 2005; 15:144–150. [PubMed: 15837171]
18. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol.* 1997; 268:209–225. [PubMed: 9149153]
19. Zhang J, Lin M, Chen R, Liang J, Liu JS. Monte Carlo sampling of near-native structures of proteins with applications. *Proteins.* 2007; 66:61–68. [PubMed: 17039507]
20. Hamelryck T, Kent JT, Krogh A. Sampling realistic protein conformations using local structural bias. *PLoS Comput Biol.* 2006; 2:e131. [PubMed: 17002495]

21. Skolnick J. Putting the pathway back into protein folding. *Proc Natl Acad Sci USA*. 2005; 102:2265–2266. [PubMed: 15703287]
22. Taylor WR. A deeply knotted protein structure and how it might fold. *Nature*. 2000; 406:916–919. [PubMed: 10972297]
23. Kmiecik S, Kolinski A. Characterization of protein-folding pathways by reduced-space modeling. *Proc Natl Acad Sci USA*. 2007; 104:12330–12335. [PubMed: 17636132]
24. Fujitsuka Y, Takada S, Luthey-Schulten ZA, Wolynes PG. Optimizing physical energy functions for protein folding. *Proteins: Struct Funct Genet*. 2004; 54:88–103. [PubMed: 14705026]
25. Liwo A, Khalili M, Scheraga HA. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc Natl Acad Sci USA*. 2005; 102:2362–2367. [PubMed: 15677316]
26. Loose C, Klepeis JL, Floudas CA. A new pairwise folding potential based on improved decoy generation and side-chain packing. *Proteins*. 2004; 54:303–314. [PubMed: 14696192]
27. Zhang J, Chen R, Liang J. Empirical potential function for simplified protein models: combining contact and local sequence-structure descriptors. *Proteins*. 2006; 63:949–960. [PubMed: 16477624]
28. Wu Y, Lu M, Chen M, Li J, Ma J. OPUS-Ca: a knowledge-based potential function requiring only C-alpha positions. *Protein Sci*. 2007; 16:1449–1463. [PubMed: 17586777]
29. Levitt M. A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol*. 1976; 104:59–107. [PubMed: 957439]
30. Park B, Levitt M. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J Mol Biol*. 1996; 258:367–392. [PubMed: 8627632]
31. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: a linear constraint solver for molecular simulations. *J Comput Chem*. 1997; 18:1463–1472.
32. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. *Bioinformatics*. 2003; 19:1589–1591. [PubMed: 12912846]
33. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997; 25:3389–3402. [PubMed: 9254694]
34. Rabiner LR. A tutorial on hidden markov-models and selected applications in speech recognition. *Proc IEEE*. 1989; 77:257–286.
35. Press, WH. Numerical recipes in C: the art of scientific computing. Vol. 26. Cambridge, New York: Cambridge University Press; 1997. p. 994
36. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr*. 2004; 60:2126–2132. [PubMed: 15572765]
37. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph*. 1996; 14:33–38. [PubMed: 8744570]
38. Shao Y, Bystroff C. Predicting interresidue contacts using templates and pathways. *Proteins*. 2003; 53 (Suppl 6):497–502. [PubMed: 14579339]
39. Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Oldziej S, Scheraga HA. A united-residue force field for off-lattice protein-structure simulations. 2. Parameterization of short-range interactions and determination of weights of energy terms by Z-score optimization. *J Comput Chem*. 1997; 18:874–887.
40. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys J*. 2003; 85:1145–1164. [PubMed: 12885659]
41. Shortle D, Simons KT, Baker D. Clustering of low-energy conformations near the native structures of small proteins. *Proc Natl Acad Sci USA*. 1998; 95:11158–11162. [PubMed: 9736706]
42. Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)–round 6. *Proteins*. 2005; 61 (Suppl 7):3–7. [PubMed: 16187341]
43. Betancourt MR, Skolnick J. Finding the needle in a haystack: educing native folds from ambiguous ab initio protein structure predictions. *J Comput Chem*. 2001; 22:339–353.
44. Zhang Y, Skolnick J. SPICKER: a clustering approach to identify near-native protein folds. *J Comput Chem*. 2004; 25:865–871. [PubMed: 15011258]

45. Daura X, van Gunsteren WF, Mark AE. Folding-unfolding thermodynamics of a beta-heptapeptide from equilibrium simulations. *Proteins: Struct Funct Genet.* 1999; 34:269–280. [PubMed: 10024015]
46. Lyman E, Zuckerman DM. Ensemble-based convergence analysis of biomolecular trajectories. *Biophys J.* 2006; 91:164–172. [PubMed: 16617086]
47. Luchko T, Huzil JT, Stepanova M, Tuszynski J. Conformational analysis of the carboxy-terminal tails of human beta-tubulin isotypes. *Biophys J.* 2008; 94:1971–1982. [PubMed: 17993481]
48. Hess B. Convergence of sampling in protein simulations. *Phys Rev E Stat Nonlin Soft Matter Phys.* 2002; 65(3 Part 1):031910. [PubMed: 11909112]
49. Yi Q, Bystroff C, Rajagopal P, Klevit RE, Baker D. Prediction and structural characterization of an independently folding substructure in the src SH3 domain. *J Mol Biol.* 1998; 283:293–300. [PubMed: 9761691]
50. Wroblewska L, Skolnick J. Can a physics-based, all-atom potential find a protein's native structure among misfolded structures? I. Large scale AMBER benchmarking. *J Comput Chem.* 2007; 28:2059–2066. [PubMed: 17407093]
51. Nymeyer H, Garcia AE, Onuchic JN. Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proc Natl Acad Sci USA.* 1998; 95:5921–5928. [PubMed: 9600893]
52. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins.* 2004; 57:702–710. [PubMed: 15476259]
53. Yap EH, Fawzi NL, Head-Gordon T. A coarse-grained alpha-carbon protein model with anisotropic hydrogen-bonding. *Proteins: Struct Funct Bioinformatics.* 2008; 70:626–638.
54. Kim PS, Baldwin RL. Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annu Rev Biochem.* 1982; 51:459–489. [PubMed: 6287919]
55. Rose GD. Hierarchic organization of domains in globular proteins. *J Mol Biol.* 1979; 134:447–470. [PubMed: 537072]
56. Ozkan SB, Wu GA, Chodera JD, Dill KA. Protein folding by zipping and assembly. *Proc Natl Acad Sci USA.* 2007; 104:11987–11992. [PubMed: 17620603]

Figure 1.

The reduced protein representation shows four α -carbons connected by virtual bonds. Side-chain centroids (SC) are placed on each α -carbon except residues at the termini. The α -carbon angle delta (Δ) acts as a bond opening angle. Torsions around the α -carbon bond are modeled by angle theta (θ).

Figure 2.

A sample virtual angle space for an α -helical sequence position. The z -axis is the negative log probability ratio Eq. (8) in $k_B T$ units. There is a very low probability for delta angles less than 70° or greater than 150° . Two distinct minima are seen for theta in this case. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Figure 3.

After identifying hydrogen bonds in the database of protein structure, each donor residue α-carbon position (Cα₅) was transformed into their acceptor residue α-carbon (Cα₂) frame of reference. The *x*-axis of the frame of reference lies along the Cα₂-Cα₁ virtual bond, the *z*-axis in the plane of the three α-carbons, and the *y*-axis is the cross-product of *z* and *x*. Similarly, the acceptor residue α-carbon positions (Cα₂) were transformed into their donor residue frame of reference.

Figure 4.

(a) A set of donor residue α -carbon positions for one acceptor residue α -carbon located at the center of the tri-peptide for one α -helical sequence position created in VMD.³⁷ (b) A contoured probability density map for the same weighted data points, created in MOE (Chemical Computing Group, Montreal, Canada). Data points were weighted based on sequence similarity as described in the methods section. Donor position probability density exists within a spherical shell surrounding the acceptor with multiple preferred orientations (blue) in this case. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Figure 5.

The entire trajectory is clustered and clusters are ranked based on size. To test for simulation convergence, structures from each cluster are separated according to whether they seemed in the first or second half of the trajectory. The example shows stage-1 equilibrium because the relative ordering of subclusters (2'' and 3'') is reversed in the second half of the trajectory.

Figure 6. Scatter plots of energy versus RMSD to native for each 0.6 μ s trajectory of all 27 peptides. The dotted line indicates the energy of the native structure.

Figure 7. Contact maps for (a) the β -hairpin 1qks_169-182 and (b) the turn motif 1pgs_189-202. Strongly predicted contacts in the potential V_{Eij} are shown in red, predicted noncontacts in blue. The black outline indicates contacts ($\text{distance}(i,j) < 8 \text{ \AA}$) found in the native structure. The sequence pattern is illustrated on the diagonal, glycine residues colored orange, hydrophobic residues black, polar residues grey, and charged residues with no bar.

Table 1

Results of Folding All 27 Putative Initiation Sites Using the CALF Potential Function

Code ^a	Motif ^b	Sequence	EQ-Stage ^c	% of Structures in cluster 1 ^d	RMSD of cluster 1 center ^e
1ahjB_46-59	Strand + helix N-cap	LGAFSVDEVRYV	2	81	2.9
1amma_94-107	Diverging turn	RDDFRGQMSEIT	1	98	4.7
1aoqA_176-189	Strand + loop + strand	LDTGYAVHISRL	2	31	4.4
1c0pA_8-21	Strand + helix N-cap	VVVLGSGVIGLS	1	68	1.5
1cfrA_135-148	Strand + helix N-Cap	SIIDIRGRREEL	1	51	3.6
1cpoA_2-15	Loop	EPGSGIGYPYDN	0	2	4.4
1ctjA_8-21	Frayed helix	AVFDGNCAACHA	1	74	1.6
1dxyA_153-166	Helix	GHIGQVVAIKLKF	2	100	0.5
1galA_23-36	Helix N-cap	GGGLTGLTTAAR	2	100	0.5
1guqB_8-21	Type I beta hairpin	PHRRYNPLTGQW	2	55	1.3
1hrdA_235-248	Helix N-cap	AGFGNVAWGAAK	2	98	0.8
1idlA_9-22	Helix N-cap	GHSILAINLILQ	1	96	2.6
1k00A_138-151	Strand + turn	RLIALDAATGKE	1	84	1.2
1kitA_307-320	DG Hairpin	YSVYDVASGNWQ	1	99	0.7
1mkaA_64-77	Proline helix C-cap	WFFGCHFIGDPV	2	4	6.7
1mwpa_22-35	Strand + diverging turn + strand	GKWDSDPSGTKT	2	33	3.3
1pexA_104-117	Type I beta hairpin	AAVHFEDTGKTL	2	81	5.4
1pgsA_189-202	Turn	HAKPYDAGSRGC	2	49	6.0
1pgsA_60-73	Type I beta hairpin	NVYVKNKTTGEW	2	83	0.9
1qksA_169-182	Type I beta hairpin	QIALIDGSTYEI	2	50	1.0
1tdeA_151-164	Helix N-cap	GGGNTAVEEALY	2	94	2.4
1tupA_109-122	Type I beta hairpin	VEYLDNRNTPRH	1	98	4.8
1vdcA_161-174	Helix N-cap	GGGDSAMEEANF	2	87	2.3
1vpsB_233-246	Turn + strand + turn	DENGVGPLCKGE	1	18	2.9
2nacA_198-211	Helix N-cap	AGRIGLAVLRRL	2	100	1.2
3grsA_160-173	Strand + loop + helix	DGFFQLLELPGR	2	94	4.2
3grsA_277-290	Turn + strand + turn	TKDLSLNKLGIQ	2	90	4.1

^a Peptides were 12 residues in length although numbering indicates 14 residues. One residue was added to both termini to allow all 12 residues to experience the complete CALF potential.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

b_i The type of secondary structure motif for each protein segment.

c_i Equilibrium sampling stage reached as measured by simulation convergence.

d_i The number of structures in the largest cluster (1) divided by the total number of structures clustered (10,000).

e_i The RMSD of the largest cluster center to native by alpha-carbon coordinates.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table II

Results for Native Structure Relaxation Simulations

Code	Structure	Length	RMSD ^a	Q ^b	TM ^c
1dxg	β	36	4.6	0.540	0.40
1fna	β	91	6.7	0.412	0.36
1opd	α/β	85	6.0	0.612	0.36
1pgb	α/β	56	3.3	0.770	0.52
1prb	α	53	5.8	0.778	0.42
1shg	β	57	7.3	0.362	0.30
1srl	β	56	4.5	0.683	0.51
1ubq	α/β	76	6.9	0.521	0.38
1who	β	94	3.5	0.707	0.59
2a3d	α	73	7.0	0.878	0.38
2c12l	α/β	65	7.9	0.531	0.36

^a RMSD of maximum deviated structure.

^b Percentage of native contacts remaining in maximum deviated structure.

^c TM-Score of maximum deviated structure. (>0.4 are significant structures).