JCTC Journal of Chemical Theory and Computation

Protein Structure Prediction: The Next Generation

Michael C. Prentiss,[†,‡] Corey Hardin,[||] Michael P. Eastwood,[‡] Chenghang Zong,[†,‡] and Peter G. Wolynes*[,†,‡,§]

Center for Theoretical Biological Physics, La Jolla, California 92093, Department of Chemistry and Biochemistry, University of California at San Diego, La Jolla, California 92093, Department of Physics, University of California, La Jolla, California 92093, and Department of Chemistry, University of Illinois, Urbana−Champaign, 600 South Mathews Avenue, Urbana, Illinois 61801

Received January 4, 2006

Abstract: Over the last 10−15 years a general understanding of the chemical reaction of protein folding has emerged from statistical mechanics. The lessons learned from protein folding kinetics based on energy landscape ideas have benefited protein structure prediction, in particular the development of coarse grained models. We survey results from blind structure prediction. We explore how second generation prediction energy functions can be developed by introducing information from an ensemble of previously simulated structures. This procedure relies on the assumption of a funneled energy landscape keeping with the principle of minimal frustration. First generation simulated structures provide an improved input for associative memory energy functions in comparison to the experimental protein structures chosen on the basis of sequence alignment.

Introduction

Every other summer, research groups compare their different protein structure prediction methods via the Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiment. During the CASP experiment, sequences of experimentally determined protein structures that are not publicly available are placed on the Web. This exercise is double blind where neither the organizers nor the participants know the experimentally determined structure. Groups respond with up to 5 ranked predictions, before a predetermined date, such as the publication of the structures. Since the inception of CASP, a three-dimensional structure prediction category has expanded to address related prediction questions such as the sequence to structure alignment quality, amino acid side-chain placement, multidomain domain boundaries, and the ordered or disordered nature of a protein sequence.[1]

* Corresponding author e-mail: pwolynes@chem.ucsd.edu.
[†] Center for Theoretical Biological Physics.
[‡] University of California at San Diego.
[§] University of California, La Jolla.
[||] University of Illinois, Urbana−Champaign.

These different prediction questions can be examined from a common framework: the principle of minimal frustration. The principle of minimal frustration states that native contacts must be more favorable, in a strict statistical sense,[2] than non-native contacts in order for proteins to fold on physiologic time scales.[3] Without a sufficient energetic bias toward the native state, the multidimensional energy surface as a function of native structure possesses too many minima for an efficient stochastic search. Such an energy surface would lead to slow folding kinetics, even if the proteins never found a sufficiently stable native state. This is not true since we know most proteins fold without assistance.[4] The opposite of a rough energy surface is biased toward the native basin without any local minima is an absolute manifestation of the principle of minimal frustration. Funneled energy surfaces have no unfavorable energetic traps (i.e. Gō Models) and have been shown to reproduce most features of experimental folding kinetics.[5−7] These energy landscape concepts can richly be applied in several areas of chemistry and physics.[8] Apparently, evolution's energy function is minimally frustrated.

The correlation between a protein sequence and its three-dimensional structure can be described using similar land-

10.1021/ct0600058 CCC: $33.50
Published on Web 04/26/2006

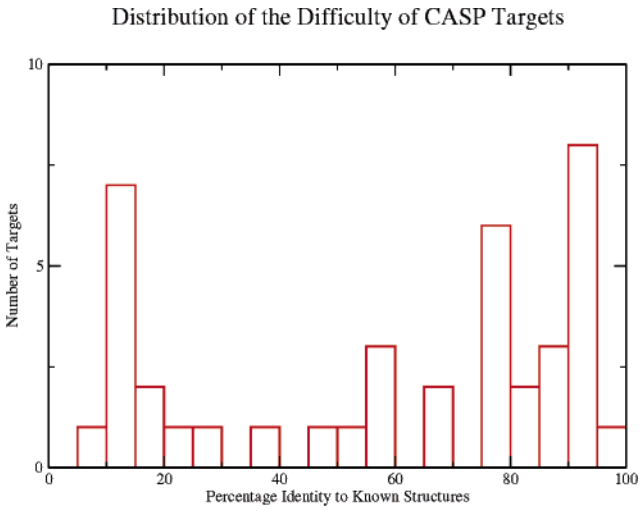

Figure 1. The difficulty of the prediction targets as defined by percent identity. Proteins below 25 sequence identity are usually considered ab initio or fold recognition targets.

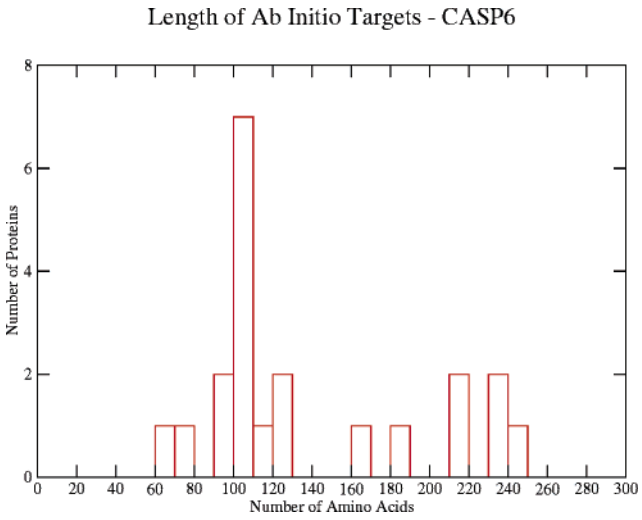

Figure 2. The ab initio prediction targets amino acid lengths for CASP6.

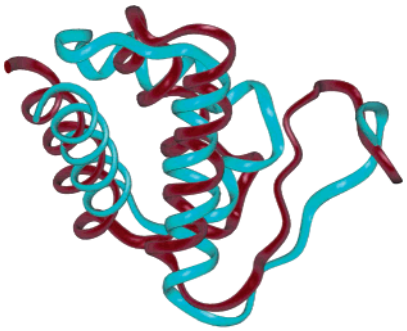

Figure 3. Sequence dependent superpositions of model 1 structure against the native state for CASP5 target T0170 (PDB ID 1UZC). Blue represents the prediction, and the native state is represented with red.

scape language. As a protein sequence diverges away from a consensus wild type sequence, the potential for energetically unfavorable interactions increases. The wild type sequence and its homologues will fold toward the same native basin. Only once enough frustrating contacts are added to wild type sequences will the sequence no longer correspond to the native state ensemble. Sequences with over 25% sequence identity to previously determined protein structures are called comparative modeling targets. The energy landscape underlying such a prediction is a Gō Model based on the structure of the known homologue. This heavily funneled energy surface yields high-resolution structures, with the discrepancies in the turns and residues which have poor sequence to structure alignments. Figure 1 demonstrates the distribution of homology of proteins sequence to known structures included in CASP6. Since proteins below 25% sequence identity are considered new fold recognition targets, 70% of the structures were comparative modeling targets. Recently sequenced genomes such as *E. coli* have the same ratio of ab initio to comparative modeling targets, which suggests the analysis of this ratio over time could be a useful measure of the progress of efforts to experimentally find examples of all of Nature's protein structures.

In contrast to comparative modeling, ab initio structure predictions do not have the advantage of creating Gō-like energy surfaces. While many ab initio targets contain less than 150 residues, and thus are candidates for standard techniques, there are several that are longer as shown in Figure 2. Most longer sequences will be multidomain proteins. This causes new problems. Folding a protein with two hydrophobic cores allows for new sources of frustration, beyond those present in single domain proteins. To obtain predictions for such problematic sequences, they usually must be divided into their constituent domains. Current methods for dividing the sequence into domains range from purely sequence based algorithms, which look for sequence patterns in multiple sequence alignments, to simulation techniques that look for hydrophobic core formation among multiple independent simulations.[9−11]

The case studies we highlight of difficult structure predictions were chosen from our participation in the CASP5 and CASP6 experiments. In CASP5, we utilized several improved techniques, such as a backbone hydrogen bond term for the proper formation of beta sheets and a liquid crystal-like term to ensure parallel or antiparallel sheet formation.[12] We also performed target sequence averaging which enhances the funneling of the prediction landscape[13] and assessed our ensemble of sampled structures with a 20 letter contact for submission.[14] Our most striking result from this round of blind prediction was a prediction for target T0170 protein databank[15] code (PDB ID IUZC). Figure 3 presents the sequence dependent overlay of our model 1 structure with the experimentally determined structure. The sequence dependent alignment quality of this structure is high as measured by a Q score of 0.38. Q is an order parameter defined in eq 1 that measures the sequence dependent structural complementarity of two structures, where Q is defined as a normalized summation of C-alpha pairwise contact differences.

$$Q = \frac{2}{(N-1)(N-2)} \sum_{i<j} \exp\left[-\frac{(r_{ij} - r_{ij}^{N})^2}{\sigma_{ij}^{2}} \right] \qquad (1)$$

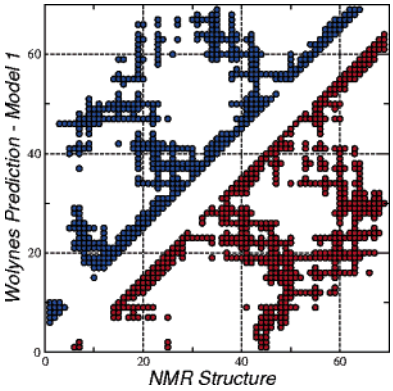

Figure 4. Contact map of target T0170 (PDB ID 1UZC) model 1 structure against the NMR structure.

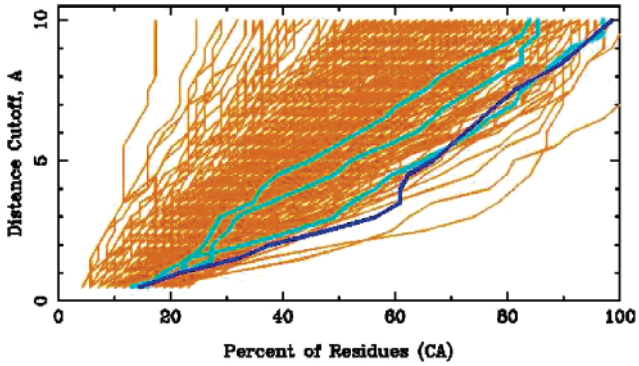

Figure 5. Percentage of residues under a RMSD limit (dark blue − model 1, light blue − model 2-5, orange − other groups prediction).

The resulting order parameter, Q, ranges from 0, when there is no similarity between structures at a pair level, to 1 which is an exact match. Q has been shown to be more sensitive in determining the quality of intermediate quality protein structure predictions.[13] Q scores of 0.4 for single domain proteins equals an RMSD of 5 Å. In most cases the reference state for the Q score is the native state, but often one wants to compare structural similarity between structures in a simulation. A sequence independent measure CE[16] also scores well (CE Z-score = 4.1). The CE Z-score measures structural complementarity without regard to sequence information and is parametrized such that structure between a Z-score greater than 4 belong to the same protein structure family. The contact map of the prediction, Figure 4, which identifies all of the C-alpha intermolecular interactions within 9 Å where the axes are the index of the protein, shows the correct packing of the helices. Figure 5 shows the size of partially correct continuous in sequence segments under an RMSD cutoff. When compared against the other predictions, our model 1 prediction (dark blue) was among the best of all submitted structures. Also the relative success of the prediction classifies this target as being of moderate difficulty. In this example CASP demonstrates that small (70 residues) all-alpha proteins are beginning to be successfully predicted by a variety of ab initio techniques.

Methods

Energy Functions and Sampling. We used an Associative Memory Hamiltonian (AMH) with optimized parameters to sample and predict structures.[17−19] The AMH uses a reduced description of the amino acid chain in order to gain the orders of magnitude computational acceleration over all atom models needed to fold moderate length proteins with ordinary computational resources and has been described in great detail before.[13] This is possible due to reducing the number of atoms per residue from over 10 to only three backbone atoms: the C_α, C_β, and O. The remaining backbone heavy atoms (N, C′) can be reconstituted using the ideal geometry of the peptide bond as a template. Also we reduced the complexity of the amino acid code from 20 letters to four. We chose the four letter code, which has the advantage of preserving a diversity of contacts, because it is still simple enough that the number of coefficients that need to be optimized does not create problems of inaccurate statistics due to limits of interactions encountered in the molten globule state. Specifically four amino acid classes are defined: hydrophilic (A, G, P, S, T), hydrophobic (C, I, L, M, F, W, Y, V), acidic (N,D,Q,E), and basic (R,H,K).[20] The optimization procedure produces an energy landscape that discriminates the native state from misfolded states, while avoiding kinetic traps reasonably well.[2,21] The AMH is an analogue to the neural networks designed by Hopfield to synthesize information from multiple previous experiences.[22] This energy function recalls structural patterns in a set of known protein structures. The Hamiltonian produces an energetically favorable minimum when there is sufficient coherence between a set of three-dimensional protein structures.

The AMH energy function, in its most general sense, contains a backbone term, E_{back}, and interaction term, E_{int}, defined by

$$E_{total} = E_{back} + E_{int} \quad (2)$$

The backbone energy term consists of several terms that reproduce the self-avoiding behavior of the polypeptide chain given by

$$E_{back} = -(E_{SHAKE} + E_{rama} + E_{ev} + E_{chain} + E_{chi}) \quad (3)$$

As in many molecular mechanics energy functions, covalent bonds are preserved by using the SHAKE algorithm[23] E_{SHAKE}, which enables an increase of the time step size and eliminates the need for a traditional harmonic calculation. The SHAKE algorithm preserves the distances between neighboring $C_\alpha-C_\beta$ and $C_\alpha-O$ atoms. The neighboring residues limit the variety of angles the backbone atoms can occupy, producing a Ramachadran plot.[24] This distribution of angles is reinforced by a potential, E_{rama},with low barriers to encourage rapid local backbone movements. Another term, E_{ev}, maintains a sequence specific excluded volume constraint between $C_\alpha-C_\alpha$, $C_\beta-C_\beta$, $O-O$, $C_\alpha-C_\beta$ atoms. The chain connectivity and planarity of the peptide bond due to resonance is ensured by means of a harmonic potential, E_{chain}. Also the chirality of the C_α, due to its four different bonding partners, is maintained using scalar product of neighboring unit vectors of carbon and nitrogen bonds, E_{chi}.

While E_{back} creates peptide-like stereochemistry, it does not introduce the majority of the attractive interactions that result in folding. Such interactions are supplied by the rest

of the potential E_{int}. The interactions described by E_{int} depend on the sequence separation $|i - j|$. Specifically, they are divided into three proximity classes $x(|i - j|)$: $x =$ short ($|i - j| < 5$), $x =$ medium ($5 \leq |i - j| \leq 12$) and $x =$ long ($|i - j| > 12$) as defined by eq 4

$$E_{int} = E_{short} + E_{med} + E_{long} \tag{4}$$

Also these distance classes are also referred to as local, supersecondary, and tertiary, respectively.

The AMH interaction potential E_{int} is based on correlations between a target's sequence signified by i, j, and the sequence-structure patterns in a set of memory proteins μ represented as i', j', and a pairwise contact potential. The pairs in the target and in the memory are first associated using a sequence-structure threading algorithm.[14] The database is assumed to contain a subset of pair distances, which may match the associated pair distances in the target structure. The general form of the associative memory interaction is

$$E_{int} = -\frac{\epsilon}{a}\sum_{\mu}^{N_{mem}} \sum_{j\text{-}12\leq i\leq j-3} \gamma(P_i, P_j, P^{\mu}_{i'}, P^{\mu}_{j'}) \exp\left[-\frac{(r_{ij} - r^{u}_{i'j'})}{2\sigma^2_{ij}}\right] + -\frac{\epsilon}{a}\sum_{k=1}^{3} C_k(N)\gamma(P_i, P_j, k)U_k(r_{ij}) \tag{5}$$

where the similarity between target pair distances r_{ij}, with aligned memory pair distances $r^{u}_{i'j'}$, is measured by Gaussian functions whose widths are given by $\sigma_{ij} = |i - j|^{0.15}$ Å. The set of parameters, γ, encode the similarity between residues i and j and the memories residues i' and j'. Favorable interactions occur during coherence in the distances achieved in the sequence to structure alignments. The encoding of the alignment information in eq 5 is only an example of what is used for the all-alpha energy functions. Other encodings have been used in the alpha−beta energy function[12] to improve the discrimination between helices and strands. While the first term in eq 5 is the superposition of interactions over a set of experimentally determined structures, it also shares a dependence on the sequence separation between the interacting residues. For residues separated by greater than 12 residues, a contact potential E_{long}, as described by the second term in eq 5, which does not depend on interaction information from the structures, is used to define local in sequence interactions. In this term $C_k(N)$ represents a sequence length dependence scaling to account for the variation in probability distributions based on sequence length. Five wells instead of the three defined here by $U_k(r_{ij})$ determine interactions in the alpha−beta energy function.[12] Energy units ϵ are defined excluding backbone contributions in terms of a native state energy in eq 6

$$\epsilon = \frac{|E^{N}_{amh}|}{4N} \tag{6}$$

where N is the number of residues. A distance class scaling a is constant in each of the energy classes because they are designed to be equal during the optimization.

The solvent in these energy functions is treated in a mean field manner, where the implicitly solvated native states of the proteins define the energy gap to the molten globule state. Solvent effects are also present in the sequence to structure alignment energy functions, but they are not explicitly represented in the molecular dynamics energy function. Water mediated contacts with an expanded 20 letter code in the contact potential were introduced,[25] based upon previous work which examined protein recognition.[26,27] The water mediated contacts along with a new one-dimensional burial term has shown promising results especially for long proteins.

Once the energy function is optimized, the minima of the energy function are probed via simulated annealing with molecular dynamics simulations. This minimization technique integrates Newton's equations of motions to determine the energy of the next time step. Simulated annealing slowly reduces the temperature from a high value as in the tempering of steel in metallurgy. This minimization algorithm allows for local searches, while allowing modest energy barriers to be overcome.

Energy landscape ideas have generated an optimization scheme for creating funneled energy surfaces. While funneled, the parametrization does not eliminate all non-native minima. The superposition of several energy surfaces reduces the likelihood of such trapping in local minima.[28,29] The flexibility of the AMH framework provides several ways of incorporating multiple sequence alignment information. Some of the options include creating a consensus sequence,[13] simulating different homologue sequences concurrently, and averaging the resulting forces and energies.[12] The averaged AMH energy function we used averages the forces and the energies of these simulation over a set of sequences, because it allows for more generalizable results than may occur with other techniques, and is described as in eqs 7 and 8

$$E_{short+med} = -\frac{1}{N_{seq}}\frac{\epsilon}{a}\sum_{1}^{seq}\sum_{\mu}^{N_{mem}} \sum_{j\text{-}12\leq i\leq j-3} \gamma(P_i, P_j, P^{\mu}_{i'}, P^{\mu}_{j'} \exp\left[-\frac{(r_{ij} - r^{u}_{i'j'})^2}{2\sigma^2_{ij}}\right] \tag{7}$$

$$E_{long} = -1/N_{seq}\frac{\epsilon}{a}\sum_{1}^{seq}\sum_{k=1}^{3} C_k(N)\gamma(P_i, P_j, k)U_k(r_{ij}) \tag{8}$$

To superimpose multiple energy landscapes, we need a multiple sequence alignment to a set of sequence homologues. Sequences homologous to the target sequence are first identified by using PSI-Blast with default parameters.[30] Each sequence above and below a certain sequence identity thresholds (70% and 30% in this work) is then aligned against each other, and proteins that have greater than 90% sequence identity to other identified sequence homologues are removed. The culling of the sequence homologues via open source bioinformatic libraries is necessary for two reasons.[31] Some classes of proteins have a large number of sequence homologues, and performing a multiple sequence alignment can be impractical. Also removing sequence homologues attempts to remove biases introduced when there are few homologues. The remaining sequences were aligned using

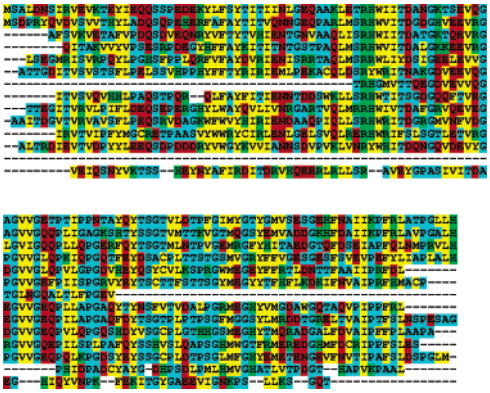

Figure 6. Multiple sequence alignment for target T0212 (PDB 1TZA) colored with respect to a four letter code, where red represents acidic residues, blue represents polar residues, yellow represents nonpolar residues, and green represents basic residues.

a multiple sequence alignment algorithm.[32] Within the AMH energy function, gaps occurring in a sequence alignment could be addressed in a variety of ways, in this work gaps in the target sequence are ignored, while gaps within homologues are completed with residues from the target protein. This strategy may introduce biases toward the target sequence, but this approach is preferred to ignoring interactions. Figure 6 shows a representative multiple sequence alignment for a target, colored with respect to the four letter code of the AMH. If one focuses on the hydrophobic yellow residues, the alternating hydrophobic hydrophilic patterns for beta strands formation are apparent.

Another way of introducing the characteristics of multiple funneled energy landscapes is using information derived from neural networks trained on multiple sequence alignments. Even with different architectures, neural networks typically achieve 75% accuracy when predicting secondary structure. Recently it has been shown that artful combinations of two different predictions can slightly improve the results.[33] This secondary structure information was added by a biasing energy function to either a helix or a strand via, $E_{Q_{ss}} = 10^5\epsilon(Q-Q_{ss})^4$,[13] where Q_{ss} is defined by eq 9

$$Q_{ss} = \sum_{k}^{n} \frac{2}{(N_k - 1)(N_k - 2)} \sum_{i<j-1} \exp\left[-\frac{(r_{ij} - r_{ij}^{ss})^2}{\sigma_{ij}^2}\right] \quad (9)$$

Q_{ss} takes the same form of the Q defined before in eq 1 except that potential acts over n independent secondary structures units derived from secondary structure prediction. The distances that define energy minimum, r_{ij}^{ss}, are determined from experimentally determined Cartesian distances. Previously in an effort to incorporate this secondary structure information, the Ramachandran potential has been altered to bias the backbone.[34] The local in sequence potential $E_{Q_{ss}}$ is preferred to the Ramachandran potential biasing because it avoids SHAKE violations when the strength of the bias is increased.

For most selected CASP6 targets, we followed the same protocol. We averaged the AMH potential over multiple sequence homologues when they were available. In most cases, information from secondary structure prediction was used to bias secondary structure units to their predicted structures. Molecular dynamics with simulated annealing sampled low-energy structures. Also constant temperatures slightly above the predicted glass temperature were used to generate candidate structures. We collected structures above T_K, which usually gives the fastest folding thereby compromising between the funneled and glassy behavior of the energy function. Once the kinetics of the structure slows, the diversity of structures encountered disappears. The slow kinetics regime typically predominates around a temperature of 0.9. While using a linear annealing schedule up to T_K, about 25 different collapsed structures were collected during each simulation. The amount of sampling performed for each structure varied from about 500 to 20 000 different structures. While this was roughly 50 times more sampling than we had previously performed in the CASP setting, it is dwarfed by the efforts of others who can sample in the millions of structures by using more powerful computational resources.[35] Subsequently, a smaller subset of structures was selected for submission by evaluating the size of the hydrophobic core and the hydrophilic surface area. Further selection criteria included visual inspection, agreement with the preliminary secondary structure prediction, and low energies predicted from a second optimized contact energy function.

Selection of Structures

To select candidate structures from independent simulated annealing or constant temperature trajectories, we calculated both the buried hydrophobic surface area and the exposed hydrophilic surface area along the trajectory. In an effort to calculate the buried or exposed surface area, we assigned residues which have greater than the mean total surface area as solvent exposed, and the converse as solvent buried. We scaled each surface area by a weight to represent the likelihood of amino acid burial. It was modeled to the free energy cost of transferring each amino acid from octanol to water[36] in an effort to introduce a sequence specificity as shown in eq 10

$$E_{\text{Burial}} = \sum_{i}^{N} \begin{cases} \gamma_i * \text{SA}_i, \text{ if } \text{SA}_i > \text{total surface}/N \\ 0, \text{ if } \text{SA}_i < \text{total surface}/N \end{cases} \quad (10)$$

This normalization is desirable because the surface accessibility is calculated from our minimal C_α, C_β, and O atoms, which produces amino acids of the same volume. Such an energy term would be more valuable if nonadditive interactions and a larger number of hydration layers were added. The unavoidable inaccuracies in atomistic force fields and the slow glassy kinetics of side-chain rearrangements prevented any completion of the backbone and side chains with all-atoms or minimization of putative structures.[37]

Another parameter we used after sampling to select and examine structures was based on sequence specific backbone probabilities. The specificity of local interactions has been fruitful for improving collapsed proteins structure predictions.[38] In a similar spirit sequence specific nearest neighbor probabilities were also used.[39] Local signals have also been

Table 1: Linear Regression of Hydrophobic Burial Energy

proteins	fold class	correlation coefficient
1R69	α	0.22
1BG8	α	0.33
1UTG	α	0.63
1MBA	α	0.40
2MHR	α	0.46
1IGD	α/β	−0.70
3IL8	α/β	−0.06
1TIG	α/β	0.02
1BFG	β	0.16
1CKA	β	−0.14
1JV5	β	0.11
1K0S	β	0.27

Table 2: Linear Regression of Mscore

proteins	fold class	correlation coefficient
1R69	α	0.29
1BG8	α	0.04
1UTG	α	0.26
1MBA	α	0.26
2MHR	α	0.10
1IGD	α/β	0.37
3IL8	α/β	0.13
1TIG	α/β	0.19
1BFG	β	0.08
1CKA	β	0.03
1JV5	β	−0.07
1K0S	β	−0.10

theoretically shown to contribute roughly a third of the total folding gap for α helical proteins.[40] Similarly we started looking at such probabilities to further improve the backbone potential of the AMH but without needing secondary structure prediction.

$$E_{\text{trimer}} = \sum_{i=2}^{N-1} \text{LogP}(i-1, i, i+1, \phi, \psi) \qquad (11)$$

Somewhat surprisingly, the summation of the resulting log probabilities from 4012 highly resolved protein structures could be used as an additional measure as part of a strategy for the selection of structures out of an ensemble. Table 1 shows the linear correlation coefficients between structures of varying Q-scores, sampled above T_K which is where the best predictions usually occur before glassy dynamics dominates the kinetics. For both proteins with all α and α/β compositions, the summed log probabilities provide discrimination but not within the all β folds. These results shown in Table 2 echo the previous findings in terms of the ϕ, ψ probability maps and also that all beta structures are less well predicted when a dihedral angle energy function is minimized. The weakness of nearest neighbor excluded volume effects to determine local structure is also demonstrated in the consistent weakness of secondary structure prediction with respect to beta strands. Alpha helices are correctly predicted to roughly 80% accuracy, while beta strands average 60% accuracy by such pure sequence based algorithms. The difficulty of predicting some circular dichroism spectroscopy results for beta to coil transitions can also be attributed to the weakness of the local backbone excluded volume interactions.

Table 3: CASP6 Results: Best Submitted and Sampled Structures

target	length	fold	sub Q	samp Q	temp	traj	CASP
T0281	70	α/β	0.34	0.48	0.85	986	NF
T0201	94	α/β	0.36	0.44	1.39	199	NF
T0212	123	β	0.26	0.42	1.30	97	FR/A
T0230	102	α/β	0.31	0.42	1.05	395	FR/A
T0207	76	α/β	–	0.39	0.98	297	–
T0224	87	α/β	0.30	0.38	1.20	501	FR/H
T0263	97	α/β	0.34	0.38	0.94	404	FR/H
T0272-a	85	α/β	0.30	0.37	0.94	30	FR/A
T0265	102	α/β	0.29	0.34	0.83	374	CM/H
T0213	103	α/β	0.26	0.32	0.98	448	FR/H
T0243	88	α/β	0.31	0.32	0.95	418	FR/H
T0239	98	α/β	0.25	0.32	0.99	424	FR/A
T0214	110	α/β	0.24	0.30	0.41	348	FR/H
T0242	115	α/β	0.27	0.30	0.89	358	NF
T0270-b	125	α/β	0.27	0.28	0.99	32	–
T0270-a	122	α/β	0.25	0.27	0.80	47	–
T0272-b	124	α/β	–	0.26	0.81	34	FR/A
T0273	186	α/β	0.22	0.24	0.98	189	NF

Results

Blind Simulations. For ab initio blind predictions in CASP6, we selected sequences if there were no experimentally determined homologous structures found by automated comparative modeling servers. The overall results for the ab initio structure prediction simulation are summarized in Table 3, where the abbreviations are length = the number of amino acids, temp = temperature where the best structure was encountered, sub Q or samp Q = the best sampled and submitted structures, respectively, as judged by a function of Q, and traj = number of independent trajectories simulated. The CASP6 targets are classified under the following categories (NF=new fold, FR/A=fold recognition analogue, FR/H=fold recognition homologue, CM/H=comparative modeling hard). Targets T0207 and T0270 were removed from the experiment so their CASP class are undefined. Structures for T0207 and T0272-b were not submitted. There are a few main points from these data. Using a Q of 0.4 as a measure of successful prediction, we were able to encounter high quality structures for 4 targets and nearly so for 4 others. The temperature at which the best structures were sampled was between the 1.2 and 0.8, which is the annealing regime we investigated most throughly. This suggests our annealing schedules were close to the behavior we sought a priori. The longer the length of the target sequence clearly reduced the quality of our predictions. Also the proteins where we had a greater number of trajectories naturally showed better structures. A final observation identifies the difference between the best submitted structure and the best sampled structure as disappointingly large for some of the targets. This can be attributed to our strategy of maximizing the number of simulations performed rather than more carefully studying our trajectories. This

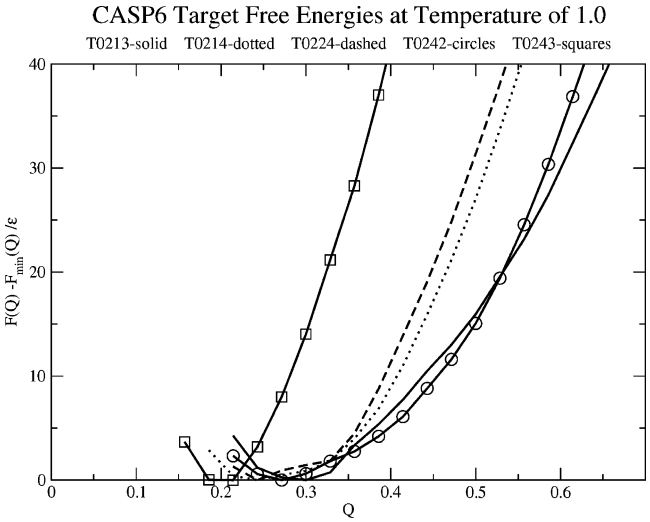

Figure 7. Free energy calculations for CASP6 targets T0213, T0214, T0224, T0242, and T0243.

Table 4: Likely Quality of Structure Seen at a Free Energy of 10 CASP6

target	PDB	length	probable *Q*	sampled *Q*
T0213	1TE7	103	0.43	0.32
T0214	1S04	110	0.40	0.30
T0224	1RHX	87	0.39	0.38
T0242	2BLK	123	0.45	0.30
T0243	–	88	0.28	0.32

difference would be smaller if greater care was taken in the selection of the structures, but the number of high quality structures would have been less.

Calculating the free energy of several randomly chosen CASP6 targets in Figure 7 provides us with probabilities of what we would have expected to see if more simulations had been performed during the CASP season. We can estimate how many independent structures need to be seen at this temperature to sample the region 10 k_BT greater than the minimum of the free energy. We see roughly $e^{10} \approx 2 \times 10^4$ independent sampled structures would be needed at a temperature of 1.0. Target T0242 (PDB ID 2BLK) illustrates why the best structure we encountered had a *Q* score of 0.3. For this target, we sampled roughly 7000 different structures. To achieve a *Q* of 0.45, according to the free energy analysis we would need to increase our sampling by a factor of 3.

When extrapolating to lower temperatures, we see lower barriers to the folded state, and thus if sampling were more complete one would see better structures at these temperatures. This further cooling would be a favorable strategy except that dynamic slowing due to the approach of the glass transition interferes, which occurs at a temperature of 0.9. Naturally, it is best to sample just above the glass transition temperature, which can be approximately found from $Q-Q$ correlation ($< Q(t)Q(t+\tau) >$),[41] and by using the Kolmogorov-Smirnov test to assess the independence of samples.[42] Table 4 indicates what was the best structure we would be likely to see under such sampling conditions. The differences between thermodynamically accessible structures and those that were sampled suggests that increased simulations would have improved the best structures sampled considerably. The free energy of target T0243 (PDB ID not available) is significantly different due to its unusual architecture that contains a buried helix.

As in Figure 4, we compare contact maps between the predictions and the experimentally resolved structure. Often contact maps give more insight than superimposed structures especially when viewing in 2 dimensions. We compare the submitted structures with the best structure encountered during our sampling to determine what aspects of folding are being captured by our energy functions. For a short target T0201 (PDB ID 1S12), we see that sometimes a small difference in the contact maps in Figure 8 can greatly improve the quality of the prediction even though a large number of contacts are already correct. There was a larger fraction of incorrect contacts in our best submitted structure for target T0230 (PDB ID 1WCJ) than we would have seen in the best generated structure as shown in Figure 9. The incorrect parallel docking of the first two helices is largely resolved in the best sampled structure, and the *Q* score improves considerably. Similar analysis for target T0281 (PDB ID 1WHZ) shows incorrect long range contacts between the two otherwise properly oriented helices and disordered intermediate interactions as in Figure 10. Again the best sampled structure has these problems largely resolved.

One amusing way to analyze predicted structures is to view the results of different structure prediction schemes as intermediates along a kinetic folding coordinate. How far did the simulated annealing get in the folding pathway? By mapping the likelihood of folding[43] against its location on a folding free energy surface, we can assess how close the model structure is to the folded state in a kinetic sense. The energy function for the kinetic modeling is a Gō model i.e. ideally nonfrustrated energy function. The difference between the Gō model and the structure prediction energy functions is a measure of the quality of those structure prediction schemes. A pairwise additive Gō model was created based on the native structure of the experimentally determined protein. As it has been discussed previously,[13] this Gō model has both polypeptide backbone energy terms that are the same as in the structure prediction energy function as described by eq 3 and an interaction potential where the Gaussian interaction potential distances r_{ij}^N are determined by the native state formally described in eq 12.

$$E_{G\bar{o}} = -\frac{\epsilon}{a}\sum_{i<j} - 3\gamma_{G\bar{o}}[x_{(|i-j|)}]\exp\left[-\frac{(r_{ij}-r_{ij}^N)^2}{\sigma_{ij}^2}\right] \quad (12)$$

The interactions are defined in this minimal model as residues with greater than three residues in sequence separation between $C^\alpha-C^\alpha$, $C^\alpha-C^\beta$, $C^\beta-C^\alpha$, $C^\beta-C^\beta$ atom pairs. The weights $\gamma_{G\bar{o}}$ or the depth of the Gaussian wells are set to (0.177,0.048,0.430) in order to approximately divided the interaction energy equally between the different distance classes as defined in the original structure prediction energy function. The width of the gaussians is defined by the sequence separation as before. Notice that the Gō Hamilto-

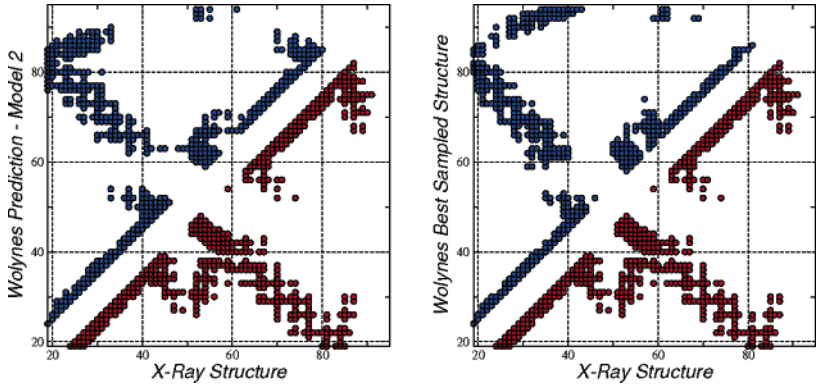

Figure 8. Contact maps for the best submitted (Q=0.36) and the best sampled (Q=0.44) structures for target T0201.

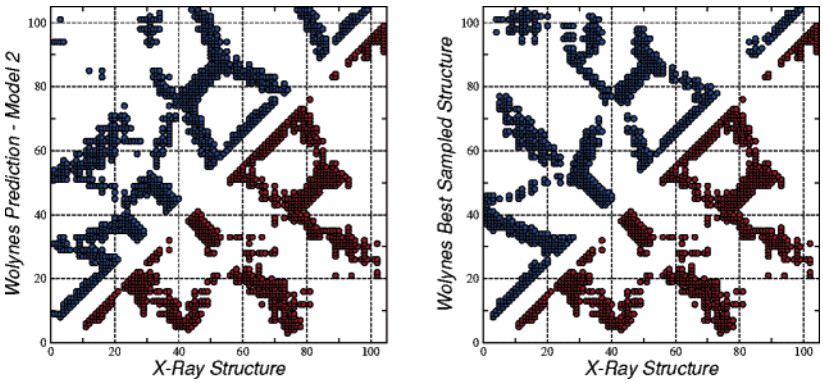

Figure 9. Contact maps for the best submitted (Q=0.31) and the best sampled (Q=0.42) structures for target T0230.

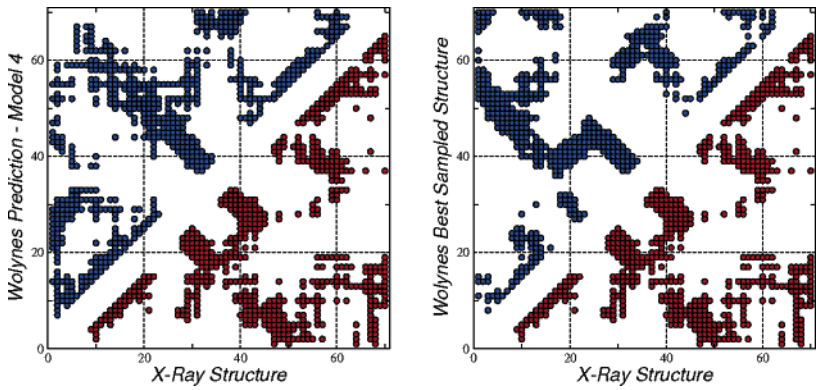

Figure 10. Contact maps for the best submitted (Q=0.34) and the best sampled (Q=0.48) structures for target T0281.

nian does not contain a summation over a set of memory structures as in the AMH; this is because all of the contacts in this definition of a Gō model use only the native state. One hundred independent simulations of this Gō energy function are performed starting with the best structure of three different structure prediction groups. Pfold is then calculated by simply determining whether the simulation started from the model structure folds to the native structure. The results in Figure 11 compare three minimalist models, one of which (the Baker Group) has undergone a further atomistic refinement. The minimalist models are only a few k_BT from the barrier's peak; they only infrequently cross it. It also suggests that a detailed less coarse grain sampling procedure may be necessary for correctly assigning hydrophobic packing and hydrogen bonding patterns.

The Next Generation in Structure Prediction

Examining the contact maps of structures encountered during the CASP experiment, we observed that contacts between residues with a large separation in sequence can be inaccurate, even when most of the contacts within a 12 residues sequence separation are nativelike. A different way of expressing this idea is that the amount of funneling is different within the different distance classes. While this was not used in the recent CASP exercise, we thought it would be interesting and straightforward to improve the prediction energy function by using these first generation results as better memory structures in the AMH. Sequence to structure alignments yield gapless identity alignments thereby eliminating any possibility of secondary structure registry shift irregularities.

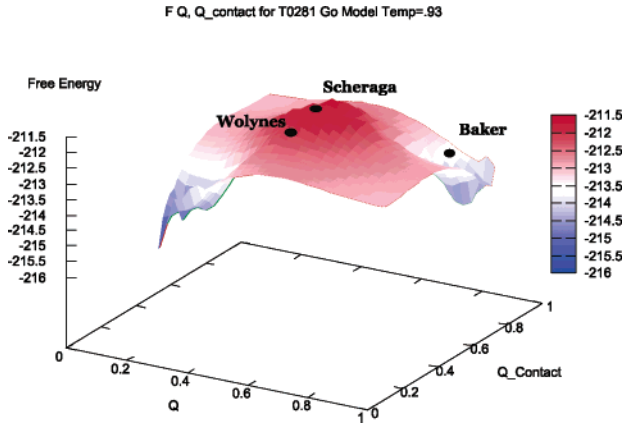

Figure 11. Gō Model free energy surface with final prediction structures shown. The Pfold values for the three proteins are the Wolynes Group 0.07, the Scheraga Group 0.02, and the Baker Group 0.97 with an error of ±0.1.

Different energy functions have been used to identify nativelike proteins from an ensemble of simulated structures. Alternatively, one can rely on energy landscape ideas and assume a mean field contact potential derived from the energy minima of the simulated energy function. This approach has the additional advantage that it does not rely on using a distinct energy function: one is simply seeing how close simulated annealing was to completely accessing the global minimum of the prediction energy function. To select structures a pairwise Q denoted by a lower case q is calculated between all of the ground-state structures encountered in 200 independent simulations.

By dividing the interchain interactions under the same definitions as used in the energy function, the potential for improvements from such second generation structures over the original memories is considerable for protein 256B. As seen in Figure 12, the low-temperature structure as identified by q has an increased amount of nativelike contacts in all distance classes. This style of analysis also suggests potential changes in the energy function. The long distance in sequence interactions are also improved over that original memory used in the energy function. To utilize this improvement the energy function in the distant interaction class was modified. The original function used a multiwell contact potential,

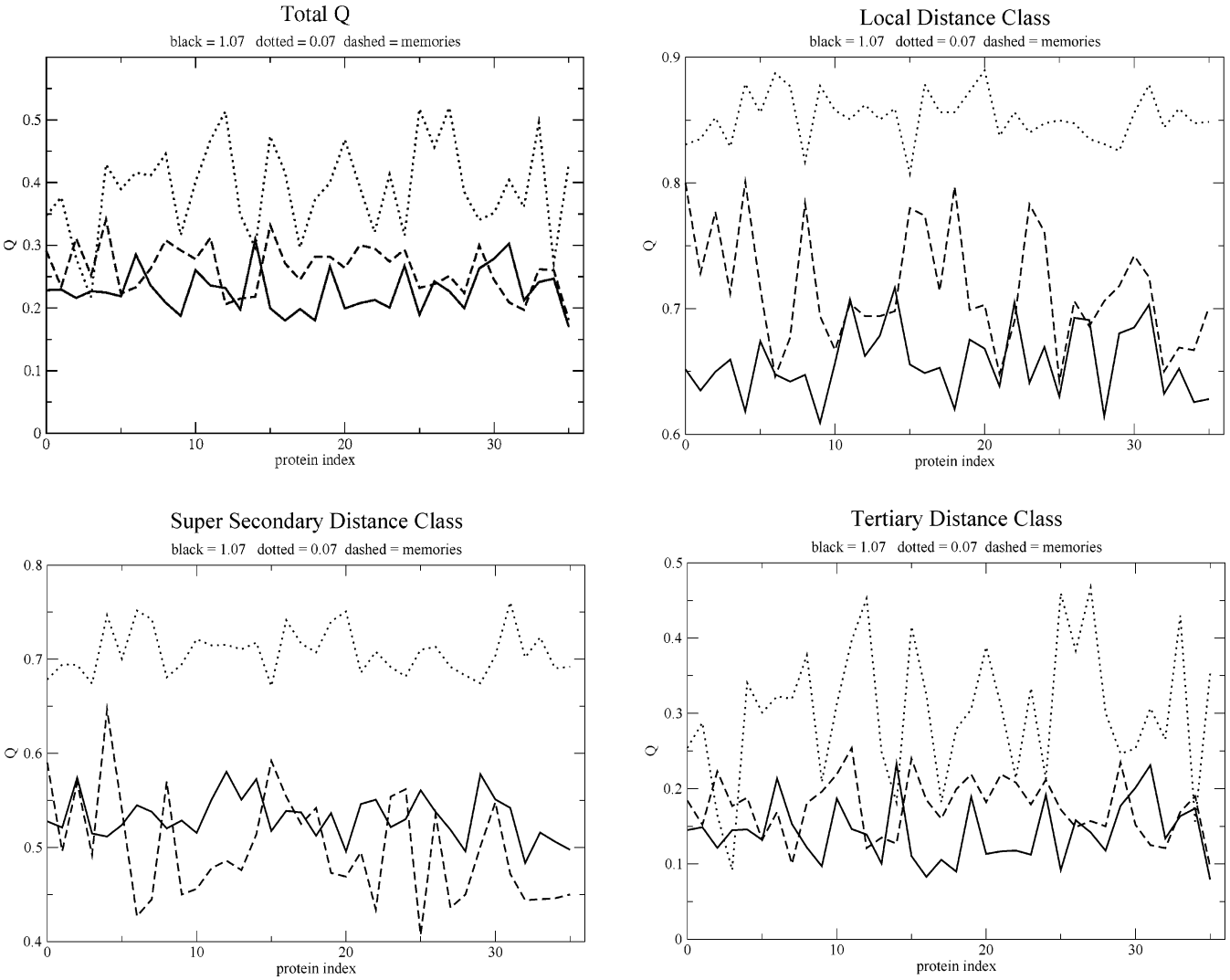

Figure 12. These figures show the total Q and the Q in the different distance classes between PDB structures, structures from a temperature of 1, and a temperature near zero for structures used as inputs to AMH simulations. The lowest temperature show the largest improvement because they are fully collapsed.

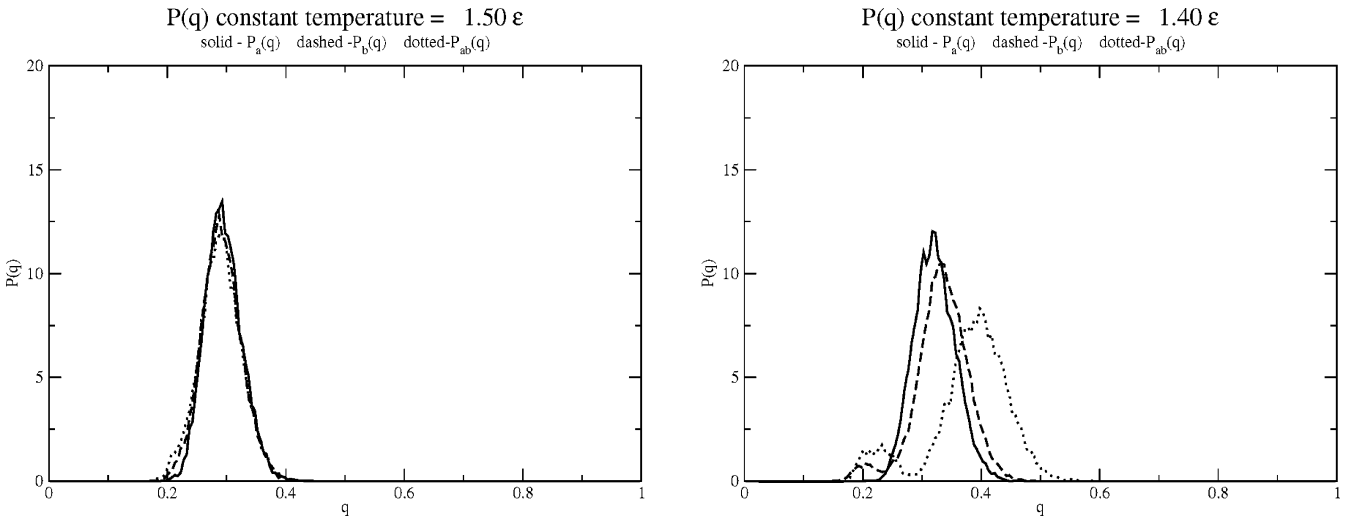

Figure 13. Kolmogorov-Smirnov test shows the constant temperature simulation falling out of equilibrium at a lower temperature of 1.4. The different probability distributions of structures between two independent simulations is no longer the same.

which does not use any information from the memory proteins. For this third distance class the next generation energy function uses associative memory contacts much as was done before for modeling with homologues.[44] The energy function now takes the form

$$E_{\text{int}} = -\sum_{3}^{c}\frac{\epsilon}{a_c}\sum_{\mu}^{n}\sum_{i<j}^{N}\gamma(P_i, P_j, P_{i'}^{\mu}, P_{j'}^{\mu})\Theta(r_{ij} - r_{i'j'}^{\mu}) \quad (13)$$

The parameters for this new distance class are taken from the second distance class. The total energy is defined over the set of memory structures as defined by eq 14

$$\epsilon = \frac{1}{36}\sum_{1}^{\mu}\frac{E_{\text{amh}}^{\text{model}}}{4N} \quad (14)$$

instead of using the values taken from the optimization. Some next generation memory structures are more collapsed than the memory structures used in the initial round of simulation. Furthermore the scaling is changed from the initial round of simulation's 1:1:1 scaling among the three different (local, super-secondary, tertiary) distance classes to 1.5:0.5:1 in an effort to approximate the equal division of energy in each distance class. To examine the equilibrium properties of this energy function, we need to estimate the glass transition temperature. As previously explored,[42] we use the Kolmogorov-Smirnov test to determine if two independent simulations have been sampled from the same equilibrium distribution. This test ensures that simulations are equilibrated.

Once the glass transition temperature (T_K) is estimated using the Kolmogorov-Smirnov test as seen in Figure 13, we can use now standard techniques to quantify the equilibrium proprieties of different energy functions. The proteins we used for study of the next generation AMH strategy are cytochrome B562 (PDB ID 256b) and HDEA (PDB ID 1BG8), because they are both of moderate size and one of them (1BG8) was not in the training set of proteins that optimized the original energy function. An additional advantage of this choice is these proteins have different fold types. According to CATH[45] HDEA belongs to the orthogonal bundle architecture, while cytochrome B562 represents an up−down bundle. Using umbrella sampling combined with the weighted histogramming method, we are able to sample parts of phase space that would rarely be encountered during a simulation.[46] When using memories with a larger number of native contacts, we see improved free energy and energy profiles as shown in Figure 14. This is even more impressive when we consider this energy function has not yet been properly optimized for this new Hamiltonian. For the other target, the results are also not surprising. In this case the next generation memories used to simulate this protein were not of greater structural quality than the initial set. Thus a very similar free energy profile was generated as seen in Figure 15. Our use of q as an order parameter successfully identified the high Q structures for the 256B test case. This is due to the highly funneled characteristic of the first generation energy function. The original energy function for 1BG8 is not as funneled, so therefore there is poorer enrichment by scanning with q. This limitation could be overcome by increasing the amount of sampling of structures in the first generation simulations. More simulations would guarantee better structure as was demonstrated during the CASP5 exercise summarized in Table 3. This difference in the enrichment could be anticipated by using the Kolmogorov-Smirnov measure to differentiate the distribution of the little q values encountered between the memories derived from simulation and the protein databank.

Conclusion

These case studies from our participation in the CASP experiment only provide a snapshot of our group's prediction schemes. It produces a series of lessons for us, and we hope for others. In the future, more balanced efforts between the sampling and selection of structures from that ensemble would appear to be desirable. More efforts in selection would have clearly improved the results submitted in CASP6. While it was computationally impractical to quench all of the structures simulated during the prediction season, the com-

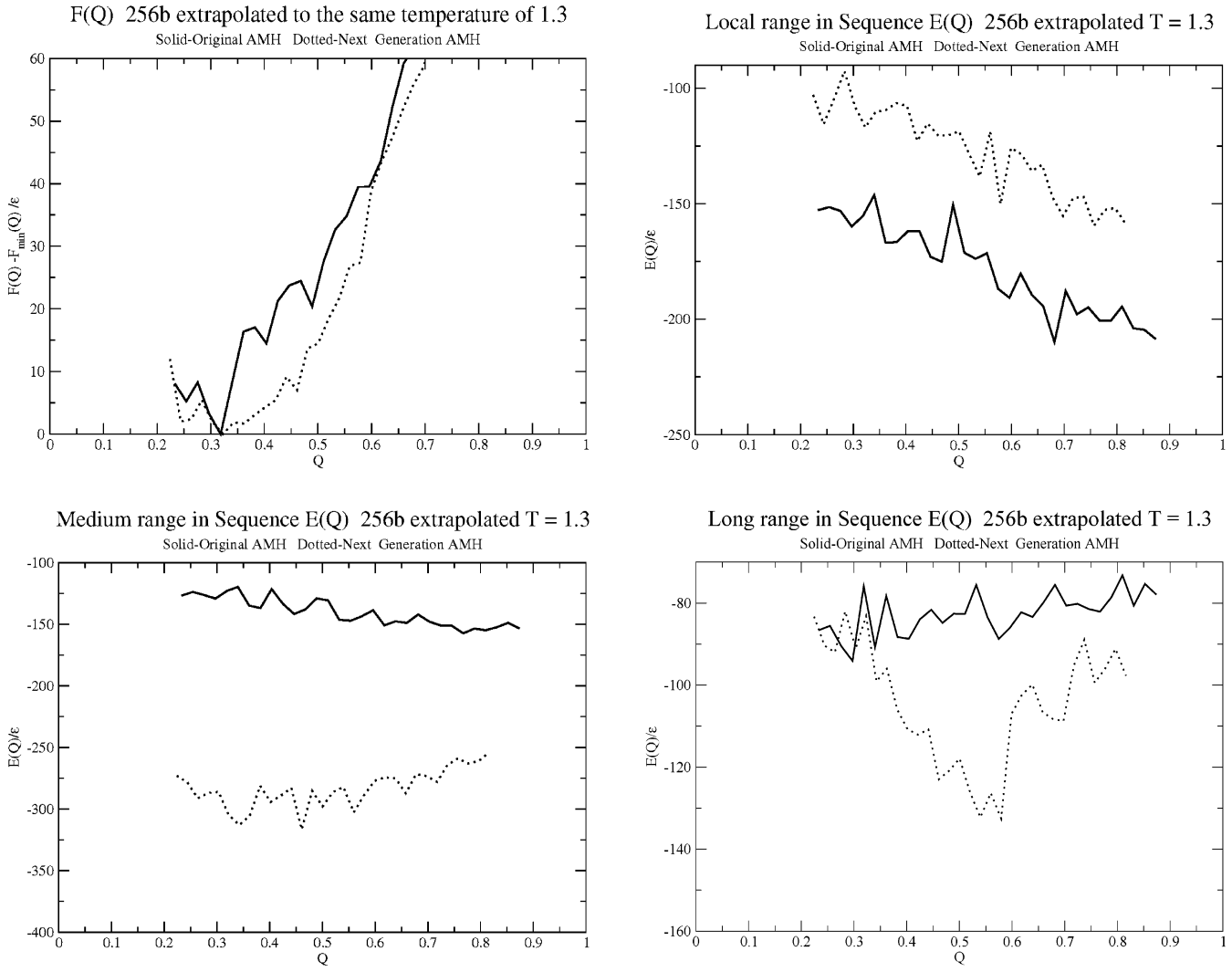

Figure 14. The free energy the two different energy functions for the protein 256B, shows roughly a 5−10 $k_B T$ improvement for this protein. The primary improvements are in the medium- and long-range nonlocal distance classes.

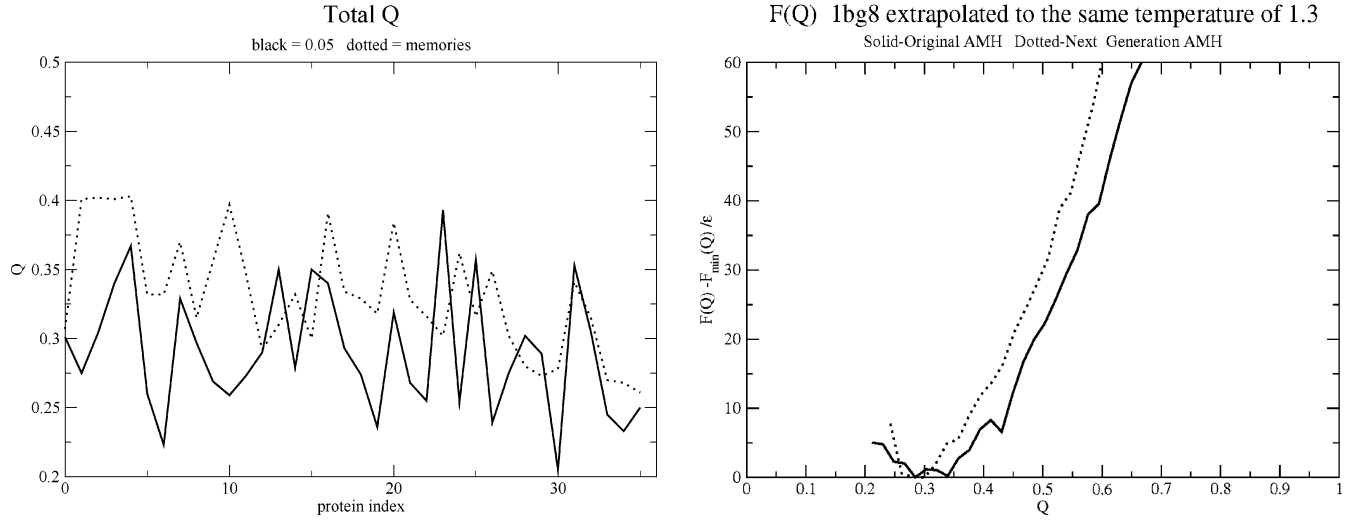

Figure 15. The free energy the two different energy functions for the protein 1BG8 show little improvement. The memories though show no enrichment in native contacts.

parison of the contact maps demonstrated further that tempering of the structure would have improved intermediate range ordering. Using preliminary structures as input to a next generation of AMH modeling improves the quality of the prediction results. While these results may initially appear to be a model or energy function specific, we feel that any algorithm that uses structures as an input would benefit from similar next generation approaches.

Acknowledgment. The authors thank Joe Hegler, Zaida Luthey-Schulten, Garegin Papoian, and Marcio Von Muhlen for their key roles in developing codes used in this study and for many helpful discussions over the years. The efforts of P.G.W. are supported through the National Institutes of Health Grant 5RO1GM44557. Computing resources were supplied by the Center for Theoretical Biological Physics through National Science Foundation Grants PHY0216576 and PHY0225630.

References

(1) Moult, J.; Fidelis, K.; Zemla, A.; Hubbard, T. *Proteins* **2003**, *53 Suppl 6*, 334−339.

(2) Goldstein, R. A.; Luthey-Schulten, Z. A.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 4918−4922.

(3) Bryngelson, J. D.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 7524−7528.

(4) Anfinsen, C. B. *Science* **1973**, *181*, 223−230.

(5) Gō, N. *Annu. Rev. Biophys. Bioeng.* **1983**, *12*, 183−210.

(6) Koga, N.; Takada, S. *J. Mol. Biol.* **2001**, *313*, 171−180.

(7) Portman, J. J.; Takada, S.; Wolynes, P. G. *Phys. Rev. Lett.* **1998**, *81*, 5237−5240.

(8) Wales, D. *Energy Landscapes;* Cambridge University Press: Cambridge, U.K., 2003.

(9) Wheelan, S. J.; Marchler-Bauer, A.; Bryant, S. H. *Bioinformatics* **2000**, *16*, 613−618.

(10) George, R. A.; Heringa, J. *J. Mol. Biol.* **2002**, *316*, 839−851.

(11) Rigden, D. J. *Protein Eng.* **2002**, *15*, 65−77.

(12) Hardin, C.; Eastwood, M.; Prentiss, M.; Luthey-Schulten, Z.; Wolynes, P. G. *Proc. Nat. Acad. Sci. U.S.A.* **2002**, *100*, 1679−1684.

(13) Eastwood, M. P.; Hardin, C.; Luthey-Schulten, Z.; Wolynes, P. G. *IBM Syst. Res.* **2001**, *45*, 475−497.

(14) Koretke, K. K.; Luthey-Schulten, Z.; Wolynes, P. G. *Protein Sci.* **1996**, *5*, 1043−1059.

(15) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucl. Acids Res.* **2000**, *28*, 235−242.

(16) Shindyalov, I.; Bourne, P. *Protein Eng.* **1998**, *11*, 739−747.

(17) Friedrichs, M. S.; Wolynes, P. G. *Science* **1989**, *246*, 371−373.

(18) Friedrichs, M.; Wolynes, P. G. *Tet. Comp. Meth.* **1990**, *3*, 175.

(19) Friedrichs, M. S.; Goldstein, R. A.; Wolynes, P. G. *J. Mol. Biol.* **1991**, *222*, 1013−1034.

(20) Hardin, C.; Eastwood, M.; Luthey-Schulten, Z.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 14235−14240.

(21) Goldstein, R.; Luthey-Schulten, Z. A.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 9029−9033.

(22) Hopfield, J. J. *Proc. Natl. Acad. Sci. U.S.A.* **1982**, *79*, 2554−2558.

(23) Ryckaert, J.; Ciccotti, G.; Berendsen, H. *J. Comput. Phys.* **1977**, *23*, 327−341.

(24) Ramachandran, G.; Sasisekharan, V. *Adv. Protein Chem.* **1968**, *23*, 283−438.

(25) Papoian, G. A.; Ulander, J.; Eastwood, M. P.; Luthey-Schulten, Z.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 3352−3357.

(26) Papoian, G. A.; Wolynes, P. G. *Biopolymers* **2003**, *68*, 333−349.

(27) Papoian, G. A.; Ulander, J.; Wolynes, P. G. *J. Am. Chem. Soc.* **2003**, *125*, 9170−9178.

(28) Maxfield, F. R.; Scheraga, H. A. *Biochemistry* **1979**, *18*, 697−704.

(29) Finkelstein, A. V. *Phys. Rev. Lett.* **1998**, *80*, 4823−4825.

(30) Altschul, S.; Madden, T.; Schaffer, A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. *Nucl. Acids Res.* **1997**, *25*, 3389−3402.

(31) Stajich, J. E. et al. *Genome Res.* **2002**, *12*, 1611−1618.

(32) Thompson, J.; Higgins, D.; Gibson, T. *Nucl. Acids Res.* **1994**, *22*, 4673−4680.

(33) Zhang, Y.; Kolinski, A.; Skolnick, J. *Biophys. J.* **2003**, *85*, 1145−1164.

(34) Hardin, C.; Eastwood, M.; Prentiss, M.; Luthey-Schulten, Z.; Wolynes, P. G. *J. Comput. Chem.* **2002**, *23*, 138−146.

(35) Bonneau, R.; Tsai, J.; Ruczinski, I.; Chivian, D.; Rohl, C.; Strauss, C. E. M.; Baker, D. *Proteins* **2001**, *Suppl 5*, 119−126.

(36) Zhou, H.; Zhou, Y. *Proteins* **2004**, *54*, 315−322.

(37) Kussell, E.; Shakhnovich, E. I. *Phys. Rev. Lett.* **2002**, *89*, 168101.

(38) Simons, K.; Kooperberg, C.; Huang, E.; Baker, D. *J. Mol. Biol.* **1997**, *268*, 209−225.

(39) Betancourt, M.; Skolnick, J. *J. Mol. Biol.* **2004**, *2*, 635−649.

(40) Saven, J. G.; Wolynes, P. G. *J. Mol. Biol.* **1996**, *257*, 199−216.

(41) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids;* Clarendon Press: New York, U.S.A., 1987.

(42) Eastwood, M.; Hardin, C.; Luthey-Schulten, Z.; Wolynes, P. G. *J. Chem. Phys.* **2003**, *118*, 8500−8512.

(43) Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. I. *J. Chem. Phys.* **1997**, *108*, 334−350.

(44) Koretke, K. K.; Luthey-Schulten, Z.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 2932−2937.

(45) Pearl, F. et al. *Nucl. Acids Res.* **2005**, *33*, D247−251.

(46) Kong, X.; Brooks, C. L., III *J. Chem. Phys.* **1996**, *105*, 2414−2423.

CT0600058