Restriction versus guidance in protein structure prediction

Joseph A. Hegler[a,b], Joachim Lätzer[a,b,1], Amarda Shehu[c], Cecilia Clementi[d], and Peter G. Wolynes[a,b,2]

[a]Department of Chemistry and Biochemistry, [b]Center for Theoretical Biological Physics, University of California at San Diego, La Jolla, CA 92093-0365; [c]Department of Computer Science, George Mason University, Fairfax, VA 22030; and [d]Department of Chemistry, Rice University, Houston, TX 77005

Contributed by Peter G. Wolynes, July 17, 2009 (sent for review June 12, 2009)

Conformational restriction by fragment assembly and guidance in molecular dynamics are alternate conformational search strategies in protein structure prediction. We examine both approaches using a version of the associative memory Hamiltonian that incorporates the influence of water-mediated interactions (AMW). For short proteins (<70 residues), fragment assembly, while searching a restricted space, compares well to molecular dynamics and is often sufficient to fold such proteins to near-native conformations (4Å) via simulated annealing. Longer proteins encounter kinetic sampling limitations in fragment assembly not seen in molecular dynamics which generally samples more native-like conformations. We also present a fragment enriched version of the standard AMW energy function, AMW-FME, which incorporates the local sequence alignment derived fragment libraries from fragment assembly directly into the energy function. This energy function, in which fragment information acts as a guide not a restriction, is found by molecular dynamics to improve on both previous approaches.

fragment assembly | associative memory Hamiltonian | protein folding | annealing | molecular dynamics

It is useful to categorize protein structure prediction schemes into two classes: template-based modeling and de novo prediction. Template-based modeling depends on the existence, and identification of, at least one experimentally-solved structure with significant global structural similarity to the target to be predicted, usually a sequence homolog. The identification can be made either by a global sequence–sequence alignment or a global sequence–structure alignment (1). Finding the proper template is a search problem but unlike folding, a search highly restricted to a relatively modest number of possibilities. After finding a template, the homolog structure acts as a global constraint which again severely restricts the remainder of the relevant conformational space to be searched. This leads overall to a much simpler optimization problem to solve. Various energy functions can be used which often lead to successful predictions defined by significant improvement relative to input homolog information (2).

However, modeling a protein structure when no experimentally-determined homologs exist to match the structures globally (or none are recognized to exist) is quite challenging. Such de novo structure prediction can employ all-atom molecular mechanics or hybrid models. Molecular mechanics methods are based on physico-chemical interactions such as van der Waals, electrostatics, hydrogen bonding, solvation energy, and basic backbone steric constraints (covalent bond lengths and angles and torsion angle preferences). Model parameters are generally inferred from experimental measurements and/or quantum chemical calculations on small organic molecules (3, 4). Based on such data, one can generate a transferable energy function (5). The resulting energy function can be used in a variety of search procedures, including template-based modeling. Ultimately, a physically robust energy function alone should be sufficient to carry out molecular dynamics simulations for de novo prediction. However, this intellectually straightforward and satisfying approach comes with a high cost—the complexity of a very detailed energy function leads to very slow computation and hence, great difficulty in searching the full conformation space available to an unconstrained polymer. Except for short peptides, fully atomistic molecular mechanics methods are therefore presently limited in carrying out de novo structure prediction. Hybrid approaches combining bioinformatic information with physical energy functions have been designed to overcome this computational difficulty.

Presently, there are two reasonably successful hybrid approaches: the fragment assembly (FA) methods (6, 7, 8) and knowledge-based energy function methods using specific protein database input (9, 10) such as the associative memory Hamiltonians (AMH) (specifically we will study one with water-mediated interactions (AMW) (11)). Both hybrid approaches use knowledge from the database to either restrict directly the conformational search space (FA) or to design a better guided coarse-grained energy function with most of the physico-chemically relevant features, by using local sequence matching (AMW). The AMW energy function based on both physical chemistry and bioinformatics then guides the molecule toward the native state.

In the FA methods, local sequence homology is used to define allowed local structure observed in naturally-occurring proteins. This approach resembles template-based modeling except, crucially, in FA the restriction on search is strictly local.

A large class of knowledge-based energy functions have been proposed and studied extensively. They are often designed to take advantage of energy landscape theory to optimize their searchability by simulated annealing. One of the earliest hybrid energy functions incorporating energy landscape optimization is the associative memory model (12, 13). The premise of energy landscape design strategy is to learn the parameters by requiring the potential to produce a low energy native state while, according to landscape theory, also creating a gap between the energies of the molten globule states and the native state. Mathematically, the learning procedure involves maximizing over the possible energy parameter values, the energy gap divided by the variance of decoy energies for training proteins (10, 14, 15). The associative memory (AM) terms of the potential are obtained from a sequence–structure threading procedure (1) which, while based on a global alignment, applies only to interactions relatively close in sequence distance, i.e., 12 residues or less. Short and intermediate-range interactions are thereby captured as "memories" from diverse possible global states, much as fragments are assembled in FA. The local information, however, does not act as a strong restriction but merely as a gentle guidance.

Ultimately, every structure prediction approach is characterized by some unique combination of energy function, conformational space, and search procedure. Despite a lack of homologs of experimentally-determined structures, FA has proven in recent years to be successful in de novo structure prediction.

Author contributions: J.A.H., C.C., and P.G.W. designed research; J.A.H. and J.L. performed research; J.A.H., J.L., A.S., C.C., and P.G.W. analyzed data; and J.A.H., J.L., and P.G.W. wrote the paper.

The authors declare no conflict of interest.

[1]Present address: BIOMAPS Institute, Rutgers University, Piscataway, NJ 08854.

[2]To whom correspondence should be addressed. E-mail: pwolynes@ucsd.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0907002106/DCSupplemental.

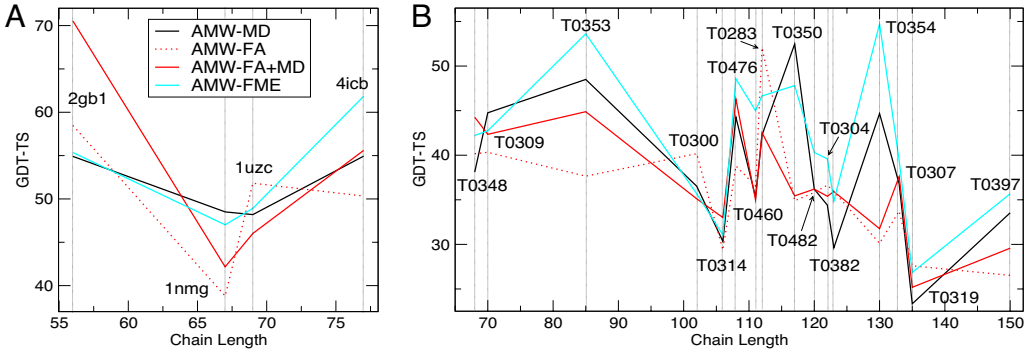

Fig. 1. Best sampled structures taken from 20 simulated annealing runs of the four canonical proteins (*A*) and sixteen CASP proteins (*B*) are compared by the structure similarity measure GDT-TS (larger values are more native-like; see *SI Appendix*). For each simulation approach, 4000 snapshots taken along the trajectory were sampled. *4icb is an AMW training protein.

Recent work has also shown that FA can reliably produce native-like structures of smaller proteins using even a relatively simple energy function (7). Although much progress has been made in improving the FA methods, whether a well-funneled energy function is necessary for its success has been unclear. Generally, the key to FA is a strong restriction of the search. FA makes use of a library of structures constrained uniquely for each sequence region. While certainly important, the quality of the fragment library, defined by the average degree of nativeness for all library members, probably accounts for only part of the overall predictive performance. The highly reduced conformational space of FA may by itself lead to a simpler folding problem to solve, at least up to the best global structure possible from the given library.

Generally, the importance of both funneled local and tertiary forces in protein folding (16–18) leading to mostly minimally frustrated contact energies (11, 19–21) is well captured by the AMW energy function. However, as previously mentioned, the threading procedure for the selection of memories is based on a global sequence–structure alignment. For this reason, it has been recognized that certain peculiar structural motifs, which occur somewhat rarely and are interspersed among helical and β-strand secondary structure elements, would be underrepresented in the associative memory forces. As we shall see, the inclusion of information from fragment library structures, generated from strictly local alignments, better captures such rarer structures and can improve the guidance of the search from the usual AMH procedure. Moreover, proteins with many turns and proline rich proteins with unusual ϕ/ψ backbone dihedral angles reveal significant improvements in local structure when locally chosen memories are included.

In this article, we first explore the two alternative approaches of direct conformational restriction (via FA by Monte Carlo search) and guidance (via local memory choice in molecular dynamics based search). We examine the predictive power of a FA method with an already highly funneled protein folding potential, the AMW energy function (AMW-FA). Then, we compare the resulting structures to those obtained from our standard unrestricted molecular dynamics approach using the same potential (AMW-MD). This comparison reveals the key role of conformational restriction in FA—the energy found by FA is higher than that found by MD, but is structurally more accurate. As it turns out, however, for long proteins FA slows in performance considerably in competition with MD. We show that by incorporating fragment guidance directly in the AMW, we can construct an AMW-fragment memory enriched (AMW-FME) energy function, that while employing molecular dynamics, performs better than the other two strategies, especially for longer sequences (>70 residues).

Results and Discussion

AMW Fragment Assembly (FA) and Molecular Dynamics (MD). To benchmark the performance of the FA method, we chose a set of proteins which includes four "canonical" ones studied previously at various levels of development of the AMH approach as well as sixteen proteins from the CASP7 and CASP8 structure prediction exercises. The test set encompasses α-helical, α/β and all-β proteins ranging in length from 56 to 150 residues. For each sequence of the test proteins, the memory terms required for the AMW energy function were obtained by threading the target sequence onto a set of PDB structures. The energies of the threaded sequences were evaluated with a local contact and burial Hamiltonian (also optimized by energy landscape theory (1)) and the lowest energy structures, with low (less than 25%) sequence homology were used as memory terms. The deliberate omission of homologs in the present study is, of course, useful for testing the procedures but would not be done in practical applications. Next, the libraries of fragments chosen strictly locally were created from a nonredundant set of protein structures (see *Methods* and *SI Appendix* for details). Each library comprised a very diverse set of structures, i.e., for each possible region for a given sequence there were several different conformations possible.

For each test protein, Monte Carlo-based (MC) fragment assembly (AMW-FA) was performed as well as molecular dynamics simulated annealing (AMW-MD) with the same potential energy function. Since an accepted MC step does not relate directly to a time step in MD, the temperature annealing schedule and radius of gyration (Rg) constraint were tuned such that the MC temperature range, relative to the approximate glass transition regime, and ease of motion (controlled partly by the Rg) in MC, were similar to MD. Consistent with previous studies (7), we find that structures that result directly from FA exhibit poor β strand–strand alignments and poor associated hydrogen bond formation. The alignment of the strands is usually improved by short MD runs of the structure predicted by FA. Therefore, we refer to structures obtained after such MD runs as postrefinement structures (AMW-FA+MD) in our analysis.

To properly quantify how the restricted conformational space via FA changes the energy landscape, one is required to calculate free energies which can be done with the reversible FA approach used here (see *Methods*) (7). However, we take a cruder approach by simply comparing structures sampled during simulated annealing runs (or low energy structures). This methodology is common practice in the Critical Assessment of Structure Prediction (CASP). The results obtained in this way are summarized in Table S1. To illustrate the performance, we show sampled structures deemed "best" by the order parameter GDT-TS (see *SI Appendix* for details) of the canonical protein set (Fig. 1*A*) and the CASP proteins (Fig. 1*B*) during 20 simulated annealing runs. Briefly,

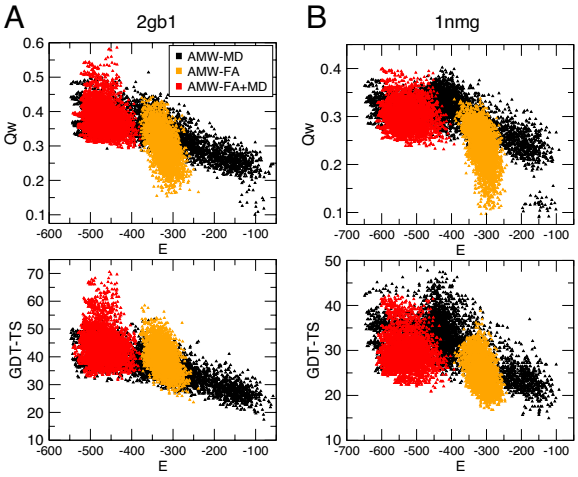

Fig. 2. Fragment assembly before (AMW-FA) (orange) and after MD-refinement (AMW-FA+MD) (red) are compared with standard AMW-MD (black) for the α/β protein 2gb1 (*A*) and the all-β protein 1nmg (*B*). The AMW energy and degree of nativeness are compared by both Q_w (*Upper*) and GDT-TS (*Lower*). In both measures, larger values are more native-like. Each point is one of 4000 snapshots from either 20 runs (AMW-MD and AMW-FA) or 100 runs (AMW-FA+MD; the final structure from each AMW-FA run (T = 0.5) was used as a starting structure for 5 refinement MD runs).

GDT-TS is the percentage of total residues which can be superimposed (averaged over distances of 1, 2, 4 and 8 Å). For the short proteins with fewer than 70 residues (including three canonical and one CASP protein, T0348), AMW-FA performs quite well in predicting structures with low root mean square deviation (rmsd) from their respective native PDB structure and high GDT-TS scores. While the all-β protein 1nmg is poorly predicted by AMW-FA, the other short proteins are equally well or better predicted by AMW-FA compared with AMW-MD.

The structures we obtained with the AMW-FA method and minimized with short MD annealing runs (AMW-FA+MD) compare favorably with those based on (what we believe are) simpler energy functions by other groups. Generally, structures resulting from AMW-FA are significantly higher in energy before (Fig. 2, orange dots) MD-refinement than after (Fig. 2, red dots). However, compared with AMW-MD, their energies remain high even after refinement, illustrating the effect of restricting the conformational search. Importantly, for the α/β protein 2gb1, we find a shift to more native-like structures after refinement (Fig. 2*A*, red dots) compared with standard MD (Fig. 2*A*, black dots). Since the shift is quite significant and occurs over a small range of energies, it seems the AMW-FA ensemble occupies a region of conformational space not frequently occupied in AMW-MD.

In a few cases, such as the α-helical protein 1uzc, the MD refinement step results in significantly less native-like structures than pure FA. Consistently, we find refinement of α-helical proteins leads to structures of similar quality to those from standard AMW-MD without FA. This finding is not too surprising as α-helical structures are able to undergo significant rearrangement during low temperature refinement more easily than β-sheet containing proteins, where significant energetic barriers to strand reorientation are hard to overcome at low temperature.

Advantages and Disadvantages of FA-Based Methods. We examine the roles of the energy function, extent of the conformational space, and search algorithm in the various methods. The strength of the FA method lies in the fact that a diminished conformational space needs to be searched. For short α-helical and α/β proteins, AMW-FA performs very well (predicting with high fidelity native-like structures) and often outperforms the plain AMW-MD simulations which do not restrict the local search space but merely guide the molecule to presumed better local conformations (Fig. 1; Table S1). Good predictions are made by all methods for specific proteins, suggesting the AMW is a sufficiently funneled energy function. So, the remaining differences in performance must stem from the search in a reduced conformational space. However, β-strand containing proteins are often more poorly predicted with FA.

One example where prediction with FA is poor, in our hands, is the all-β protein 1nmg (Fig. 2*B*), mentioned previously. While AMW-MD simulations for 1nmg lead to quite native-like structures at the lowest energy sampled (see Table S1), ensembles obtained with AMW-FA are, on average, shifted toward less native structures. Inadequate funneling of the energy function as well as overly slow search and incompleteness of the fragment library derived conformational space are all potential reasons why specific protein structures might not be well predicted. In the case of 1nmg, we conclude that the poor quality predictions are not a consequence of a poor energy function since plain AMW-MD simulations of 1nmg produce reasonably native-like structures. In addition, the quality of the predictions is significantly improved after MD minimization of the structures obtained with FA with the same Hamiltonian—primarily resulting in better hydrogen bonded networks of β-strands. But this is not the whole story. Even after MD refinement of 1nmg fragment assembled structures, those generated by standard AMW-MD are still superior. Apparently the conformational space is overly restricted and seems to be somewhat inconsistent with the native ensemble structures.

With limited computational resources, success in predicting low energy, native structures with MC-based FA depends on both chain topology and length. FA move steps lead to kinetic slowing because of the increased likelihood of steric clashes when a protein adopts more compact molten globule conformations, as was encountered in MC studies decades ago in lattice models (22). There is significant difficulty in carrying out the subtle rearrangements, in compact protein conformations, necessary for proper β-strand formation and corresponding hydrogen bonding. In the cases of 2gb1, 1nmg, and T0348, only by relaxing the structures by FA+MD are strands efficiently rearranged to reach a reasonable β-sheet topology. Such rearrangement is especially important for the central regions of the protein chain. With the standard FA algorithm, accepted MC moves are strongly biased to occur near the chain terminal regions. Clearly, there will be greater sampling difficulty with increasing chain length as proportionately more of the chain becomes buried during collapse. The R_g-bias forces can be adjusted to control the collapse of structures and may need further development.

Since there is favorable rearrangement of β-strands (for most proteins in our set) upon MD refinement, the search procedure alone must be partly responsible for the poor performance of the pure AMW-FA method for 1nmg. However, two other proteins with similar β-strand content, namely 2gb1 and T0348, show significant performance improvement of AMW-FA over plain AMW-MD, even without the final MD refinement (Fig. 1 *A* and *B*), suggesting the poor quality of the prediction could also be related to the quality of the fragment library for 1nmg. Since the conformational space searched by FA is entirely determined by the structures present in the original fragment library, to examine this point we first defined the quality of the fragment library as the average fragment nativeness (see *SI Appendix*) and then computed the library quality for each protein in our study (summarized in Table S2). Protein 1nmg has the lowest quality fragment library, while protein 4icb has the highest. While native-like structures of 1nmg cannot be produced with the AMW-FA method (best rmsd = 7.39), structures obtained for 4icb are consistently very similar to the native state (best rmsd = 4.73), confirming a direct relationship between the nativeness of the local structures in the original fragment library and the global performance. 1nmg thus illustrates both possible contributing factors to poor FA perfor-

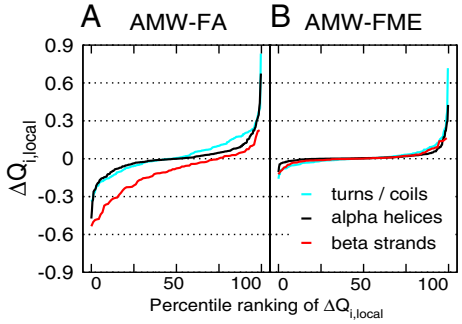

Fig. 3. Local prediction performance by secondary structure type of AMW-FA (*A*) and AMW-FME (*B*) are compared with standard AMW-MD. A measure of structural similarity of local regions (of 9 residue length) to experimentally-observed native structures, $Q_{i,\mathrm{local}}$ (defined in *SI Appendix*), is used to capture the difference in local nativeness, $\Delta Q_{i,\mathrm{local}}$. For all studied proteins, all fragments of the 20 final annealed structures ($T = 0$) were compared with native. For each simulation method, the average $Q_{i,\mathrm{local}}$ was calculated for each fragment region. Fragments are categorized into three structural classes (α-helical, β-strands, or turns/coils) based on the DSSP annotation (23) for the central residue, *i*, and sorted by $\Delta Q_{i,\mathrm{local}}$. In the case that $\Delta Q_{i,\mathrm{local}} < 0$, AMW-MD generated local structures are more native-like than the corresponding alternative sampling strategy. For illustration purposes, in *A*, a value of $\Delta Q_{i,\mathrm{local}} \geq 0.15$ occurs in $\approx$20% of the fragments around turns and coils. That is, 20% of the local structures generated by AMW-FA are at least 15% more native-like (by $Q_{i,\mathrm{local}}$) than those generated by standard AMW-MD.

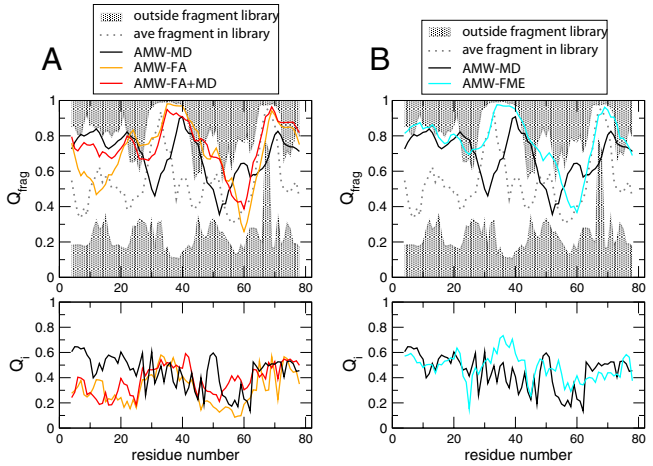

Fig. 4. Local structure prediction for protein T0353 by FA methods (*A*) and AMW-FME (*B*). The solid gray area represents a local structure similarity to native that cannot be reached by the fragment library itself—except by hybrid fragments (7). The gray dotted line through the middle represents the average value of Q_{frag} for the 20 library fragments at the given position. The solid lines depict the Q_{frag} (*Upper*) and Q_i (*Lower*) values for each 9 residue region along the chain from the best sampled structure (highest GDT-TS) by each of the four sampling methods (AMW-MD, AMW-FA, AMW-FA+MD, and AMD-FME). For a fragment centered at residue *i*, Q_{frag} includes all interactions within the fragment while Q_i includes interactions from residue *i* to the remainder of the protein.

mance—a conformational space overly constrained by a poor quality finite fragment library and a kinetically slow search algorithm hindered by unphysical FA move steps.

To further explore the advantages and disadvantages of the FA method, we can compare the nativeness of secondary structure resulting from AMW-FA and standard AMW-MD simulations. We plot the difference in nativeness, $\Delta Q_{i,local}$ (see *SI Appendix*), of local regions (9 residue length) for all proteins used in this study. Fig. 3*A* (red line) clearly shows that the β-strands for the AMW-FA algorithm are significantly less native-like, with most $\Delta Q_{i,local}$ values being negative. AMW-MD produced more native-like structures for the given fragment than AMW-FA. MD appears to be the method of choice for predicting β-strands. Not surprisingly, relaxation of MC-generated strands with MD leads to some improvement of β-strands. In contrast, pure FA does show advantages in turn regions which are not so well predicted with standard AMW-MD as shown in Fig. 3*A* (cyan line). This result is consistent with the fact that local turn regions do not align well during the initial global sequence-structure alignment used for the AMW. Additionally, the α-helical regions (Fig. 3, black lines) are, on average, slightly more native compared with AMW-MD simulations.

We further probed how the input local fragment library quality affects local and global structure prediction. For each local fragment region of 9 residue length from all proteins in the study set, centered at residue *i*, we computed the average degree of nativeness of each of the library structures to the corresponding native structure, $Q^{lib}_{i,frag}$ (see *SI Appendix*), and recorded the minimum and maximum values of nativeness. Plotting these values against sequence gives an idea of the average quality of the library as well as the possible range of quality of the input fragments. For each of the different prediction algorithms we also plot for the same fragment the degree of nativeness $Q_{i,frag}$ for the best sampled structure. We find the $Q_{i,frag}$ of the predicted structure is quite close to, or, on account of hybrid fragments [composed of parts of multiple library fragments (7)], slightly better than the best fragment library structure. This means we are sampling global structures whose local regions have structures quite similar to the best predictions that could have been obtained by any of the local sequence alignments used in the assembly. The results are presented in Fig. 4 for protein T0353. In Fig. 4*A*, the results for AMW-FA (orange line) and AMW-FA+MD (red line) are shown—for comparison the AMW-MD result (black line) is also plotted. Other than the fragments located in the N terminus and around residue 60, in all three methods the predicted local structures are significantly closer to the native structure than the average fragment library structure, and often are very near to the best library structure locally.

However, improvement in local structure does not necessarily mean improvement in global structure, as illustrated by the global degree of nativeness for each residue *i*, Q_i, shown in the *Lower* of Fig. 4 (same color coding). The tertiary nativeness Q_i is calculated over the full length of the protein and measures how native-like are the interactions of residue *i* with the remainder of the protein—a local measure of correct tertiary structure. While the prediction of local fragments is better for FA, the tertiary structure as a whole (as measured by Q_i) is often worse for FA than it is for AMW-MD. For example, in protein T0353 the fragments around residue 50 show improvement compared with AMW-MD in local structure prediction, but the tertiary structure at that area becomes less native. Ultimately, the final global prediction performance seems to be reflected more in Q_i, the global nativeness parameter, which captures tertiary interactions. The significant improvement of Q_i from AMW-MD over AMW-FA around the first 30 residues and residue 50 captures the superior AMW-MD global prediction performance (Fig. 1*B*). Since the protein T0353 does not have a particularly poor fragment library (unlike the case of 1nmg), the poor performance likely arises from the kinetic sampling limitations of FA observed with increasing chain length.

AMW-FME – Fragment Memory Enriched. The results for longer proteins from the CASP test set are shown in Fig. 1*B* and Table S1. Not surprisingly, longer proteins are harder to predict with FA methods. Indeed, most of the best structures found by FA do show lower GDT-TS values from the native PDB structure relative to those found by direct MD. As described previously, with MC torsion-angle rotation based schemes, cooperative

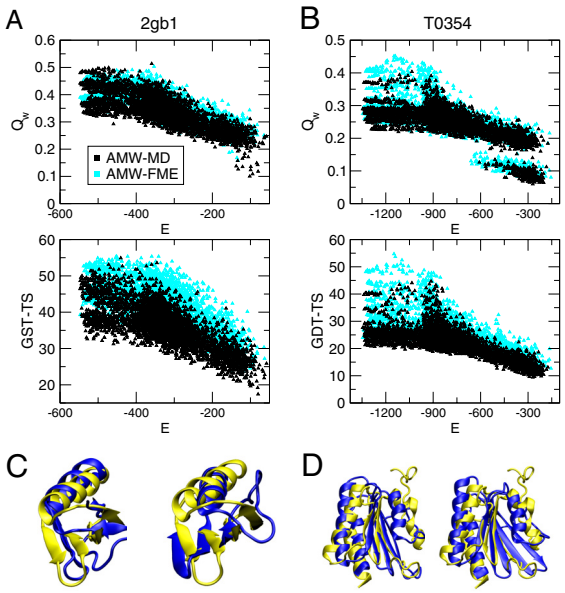

Fig. 5. AMW-FME (cyan) is compared with standard AMW-MD (black) for the α/β proteins 2gb1 (*A*) and T0354 (*B*). The AMW energy (or AMW-FME energy scaled to AMW-MD) and degree of nativeness are compared by both Q_w (*Upper*) and GDT-TS (*Lower*). Each point is one of 4000 snapshots from 20 runs for each method. For both 2gb1 (*C*) and T0354 (*D*), the best structures (by GDT-TS; $T = 0$) by AMW (*Left*) and AMW-FME sampling (*Right*) are superimposed against the corresponding native structures.

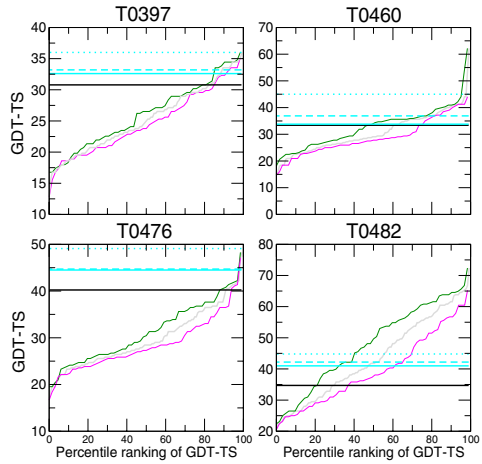

Fig. 6. AMW-MD and AMW-FME performance for the four CASP8 Free Model targets with 150 or fewer residues. For each target, the GDT-TS values of the best submitted structures from each participating group are ordered. Performance for all groups (gray line) is distiguished from human-only (green) and server-only groups (magenta). The best submitted structure by the Wolynes group (id 093) is indicated (black line). The postexercise structures generated by AMW-FME are indicated in cyan. From 100 AMW-FME simulations, the best of five are indicated after filtering by energy alone (solid line) or by a simple filtering procedure (dashed line). From these, the best sampled structure is shown for reference (dotted line).

long-range rearrangements cannot simultaneously compensate for each other, as in the parallel-search MD procedures. Also, the number of states of the system increases dramatically with chain length, leading to even greater difficulties besides those inherent in the MC motions.

Besides the kinetic problems with the FA sampling procedure, the restriction of conformational space in the FA method can either increase or decrease the overall performance depending on the quality of the underlying fragment library (low for 1nmg and high for 4icb). This analysis suggests seeking a way to avoid the limitations of the standard AMW energy function while keeping what is good about FA. The guidance in AMW must deal with the more unusual structural motifs such as regions near prolines, turns and coils which are well captured by the FA implementation. Overall, while using global sequence–structure alignment to choose memories in the AMW is quite successful in achieving highly native-like tertiary interactions (as in the long-range β-strands in 1nmg), the dramatic improvement of AMW-FA+MD over AMW-MD in the case of 2gb1 supports the notion that some local motifs would have been better predicted by restricting search to structures constructed by the purely local sequence alignment procedure used for the fragment libraries.

To get the best of both worlds, we developed a modified version of the AMW code, which we call AMW-fragment memory enriched (AMW-FME). This approach simply supplements the standard AM forces (a component of the full AMW energy function) based on global alignments with additional forces based on local structures taken directly from the fragment library (see *Materials and Methods* for details of the AMW-FME energy function).

The results for AMW-FME on all proteins in our set are shown in Fig. 1, cyan lines, and Table S1. For most proteins, AMW-FME shows measurable improvement in prediction results over the standard AMW. Moreover, the proteins showing the most significant improvement are those with the highest local library quality, including the α-helical protein 4icb and the three α/β proteins T0353, T0354, and 2gb1. We compare the AMW and AMW-FME simulated annealing snapshots for proteins 2gb1 and T0354 in Fig. 5. In the case of 2gb1, only slightly more native-like structures are sampled with the Q_w order parameter, while structures with significantly higher values of GDT-TS are frequently populated. The comparison for the protein T0354 (Fig. 5*B*) shows a dramatic shift toward more native-like structures when ordered by either Q_w or GDT-TS.

That the improvement in local memories is key is shown by the fact that when a poor fragment library is used, as in 1nmg, with AMW-FME, there is no improvement. Pleasantly, while longer proteins with good fragment library quality, such as T0354 (Fig. 5*B*), show poor global performance on account of sampling deficiencies using FA, the AMW-FME method successfully takes advantage of the local fragment library information and shows significant performance improvement.

To better understand these effects, we return to the protein T0353. Comparing the best sampled structures, we see the AMW-FME method (Fig. 4*B*) clearly improves local as well as global structure over standard AMW-MD. The local structure performance of AMW-FME (Fig. 4*B Upper*, cyan line) is quite similar to that obtained with AMW-FA+MD (Fig. 4*A Upper*, red line). However, across the first half of the protein chain, the local structure obtained with AMW-FME is more native-like. For T0353, better local fragment structures are sampled with AMW-FME than with the pure FA even followed by MD refinement. Also, there is secondary structure improvement over the full study set (Fig. 3*B*). However, the tertiary structure, described by Q_i (as shown in the *Lower* of Fig. 4*B*), is generally better for AMW-FME than pure AMW-MD. Apparently, the kinetic limitations of sampling with AMW-FME are small.

The "chimeric" AMW-FME MD approach can take advantage of the existence of high quality fragment libraries for larger proteins but does not degrade in performance even when the fragment library is of low quality.

CASP8. The most recent CASP (CASP8) took place in the summer of 2008. Of 27 targets of length 150 residues or less, only four were categorized as free model (FM). FM targets are those for which no

structurally similar template was identified or submitted. For these proteins, in Table S3, we present the results we submitted using the standard AMW-MD procedure. We also present postexercise results based on the AMW-FME method developed here. Typically, human judgement based on visual inspection, and other filtering tools, are used in selecting targets for submission. To avoid bias in this post hoc assessment, we cannot include selection based on visual inspection. Instead, we selected the five best candidate structures according to lowest energy structures and a simple automated filtering strategy based on quantifying the frustration level of a sampled structure (see *SI Appendix*). A comparison of the AMW-MD and AMW-FME performance with all CASP8 participating groups is presented in Fig. 6.

Conclusions

The complementary advantages of AMW and FA methods for structure prediction can be combined to design a better performing method. In particular, we find that the combination of FA with the AMW potential (AMW-FA) often performs better than MD simulated annealing using the same potential energy function (AMW) in relatively short proteins with fewer than 70 residues, when there is a fragment library of reasonable quality. However, a distinct disruption of β-strand local structure is clearly observed in FA. Despite this, in shorter proteins, improvements in local regions around turns and α-helices usually drive the system toward better structures. A short MD relaxation run of the final structures obtained from FA is often sufficient to produce remarkable improvement in structure prediction, associated with fine-tuning in β-strand arrangement. For instance, the proteins 2gb1 and T0348 exhibit this behavior in our study. However, the all-β protein 1nmg, which has the poorest fragment library quality of the proteins studied, fails to sample the native structure well. MD relaxation proves insufficient to lead to structures of similar quality to standard AMW-MD simulations. In longer proteins, FA suffers from a number of deficiencies intrinsic to the awkward nonparallel move steps (MD search procedures are naturally parallel) of the standard MC-based FA procedures. Alternative procedures might avoid this. To combine the benefits of these approaches, we propose the fragment enriched version of the standard AMW energy function, AMW-FME, that naturally takes advantage of local alignment derived fragment libraries without paying the price of the accompanying kinetic difficulties created by the FA method. The AMW-FME algorithm produces significantly improved prediction performance (with no advanced postprocessing techniques) of de novo structures of longer proteins.

Materials and Methods

Fragment Assembly with AMW. From library fragments based on sequence alignments of 9 residue length (see *SI Appendix* for library details), the torsion angles between the central three (from $i-1$ to $i+1$) and six (from $i-3$ to $i+2$) residues were extracted. For the AMW-FA method, two fragment assembly procedures were used, differing by the length of the fragments substituted (3 or 6 residues). We took this approach since, during development with a set of separate proteins, we found each of the methods to produce better results in different cases. Simulations with 9 residue fragments rarely outperformed those with shorter fragments and were therefore not studied further. For each protein, ten simulations were performed each with 3 and 6 residue fragments. At each MC move step, a random position along the sequence was selected. Next, the torsion angles were substituted with those from a randomly selected fragment from the library associated with the region (composed of 20 conformations). The reversible version of the fragment assembly method, as described by Takada (7), was generally found to outperform the irreversible version. As such, for the 20 proteins included in this study, we ran simulations only with the reversible algorithm.

AMW-Fragment Memory Enriched (AMW-FME) Energy Function. The enriched version of our AMW energy function takes the form: $H_{\mathrm{AMW-FME}} = H_{\mathrm{AMW}} + H_{\mathrm{AM-frag}}$ (see *SI Appendix* for details of H_{AMW}). $H_{AM-frag}$ is equivalent to H_{AM} except that we remove the sequence dependence from $\gamma_{frag}[|i-j|]$:

$$H_{AM-frag} = \varepsilon_{frag} \sum_f \sum_{i<j-2} \gamma_{frag}[|i-j|] e^{-\frac{(r_{ij}-r_{ij}^f)^2}{(2\sigma_{ij}^2)}},$$

where f is an index over all fragments and the sum over i and j includes all pairs of atoms of type (C^α-C^α, C^α-C^β, C^β-C^α, C^β-C^β) given $i<j-2$. The distances r_{ij} and r_{ij}^f are between atoms i and j in the current and fragment conformations, respectively. The Gaussian well widths are given by $\sigma_{ij} = (i-j)^{0.15}$Å. The weights $\gamma_{frag}[|i-j|]$ depend only on the sequence distance class (short, medium, and long range) but are different in α/β and α-only simulations. They were chosen such that the balance of energy between the short, medium, and long range is maintained, to be consistent with the balance of energy in the standard H_{AM} (10). We chose ε_{frag} such that the balance of the total energy between the standard and fragment enriched AM terms is approximately in the ratio 2:1. We found that a ratio of 1:1 leads to poor global prediction performance as the local structure becomes overly constrained.

ACKNOWLEDGMENTS. We thank the Center for Theoretical Biological Physics (CTBP) for computational resources. This work supported by National Science Foundation Grant PHY-0822283 (Center for Theoretical Biological Physics) and National Institutes of Health Grant R01GM44557. In addition, work was supported by the National Science Foundation Grants NSF-Career CHE-0349303, NSF-CCF-0523908, NSF-CDI CHE-0835824, and the Welch Foundation Grant C-1570. Simulations were performed in the Rice Computational Research Cluster funded by NSF Grants CNS-0421109, CNS-0454333, and EIA-0216467, a partnership between Rice University, AMD and Cray, and a partnership between Rice University, Sun Microsystems, and Sigma Solutions, Inc.

1. Koretke KK, Luthey-Schulten Z, Wolynes PG (1996) Self-consistently optimized statistical mechanical energy functions for sequence structure alignment. *Protein Sci* 5:1043–1059.
2. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T (2007) Assessment of casp7 predictions for template-based modeling targets. *Proteins: Struct, Funct, Bioinf* 69 Suppl 8:38–56.
3. MacWood GE, Urey HC (1935) Raman spectrum of methyl deuteride. *J Chem Phys* 3:650–651.
4. Westheimer FH, Mayer JE (1946) The theory of the racemization of optically active derivatives of diphenyl. *J Chem Phys* 14:733–738.
5. Ponder JW, Case DA (2003) Force fields for protein simulations. *Adv Protein Chem* 66:27–85.
6. Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J Mol Biol* 268:209–225.
7. Chikenji G, Fujitsuka Y, Takada S (2003) A reversible fragment assembly method for *de novo* protein structure prediction. *J Chem Phys* 119:6895–6903.
8. Shehu A, Kavraki LE, Clementi C (2009) Multiscale characterization of protein conformational ensembles. *Proteins: Struct, Funct, Bioinf Online* 4:837–851.
9. Lee J, Liwo A, Scheraga HA (1999) Energy-based de novo protein folding by conformational space annealing and an off-lattice united-residue force field: Application to the 10–55 fragment of staphylococcal protein A and to apo calbindin D9K. *Proc Natl Acad Sci USA* 96:2025–2030.
10. Eastwood MP, Hardin C, Luthey-Schulten Z, Wolynes PG (2001) Evaluating protein structure-prediction schemes using energy landscape theory. *IBM J Res Dev* 45:475–497.
11. Papoian GA, Ulander J, Eastwood MP, Luthey-Schulten Z, Wolynes PG (2004) Water in protein structure prediction. *Proc Natl Acad Sci USA* 101:3352–3357.
12. Friedrichs MS, Wolynes PG (1989) Toward protein tertiary structure recognition by means of associative memory Hamiltonians. *Science* 246:371–373.
13. Goldstein RA, Luthey-Schulten ZA, Wolynes PG (1992) Optimal protein folding codes from spin glass theory. *Proc Natl Acad Sci USA* 89:4918–4922.
14. Hardin C, Eastwood MP, Prentiss M, Luthey-Schulten Z, Wolynes PG (2002) Folding funnels: The key to robust protein structure prediction. *J Comput Chem* 23:138–146.
15. Hardin C, Eastwood M, Prentiss M, Luthey-Schulten Z, Wolynes PG (2003) Associative memory Hamiltonians for structure prediction without homology: α/β proteins. *Proc Natl Acad Sci USA* 100:1679–1684.
16. Ueda Y, Taketomi H, Go N (1978) Studies of protein folding, unfolding, and fluctuations by computer simulation, II. 3-dimensional lattice model for lysozyme. *Biopolymers* 7:1531–1548.
17. Onuchic JN, Luthey-Schulten Z, Wolynes PG (1997) Theory of protein folding: The energy landscape perspective. *Annu Rev Phys Chem* 48:545–600.
18. Onuchic JN, Wolynes PG (2004) Theory of protein folding. *Curr Opin Struct Biol* 14:70–75.
19. Zong C, Papoian GA, Ulander J, Wolynes PG (2006) Role of topology, nonadditivity, and water-mediated interactions in predicting the structures of α/β proteins. *J Am Chem Soc* 128:5168–5176.
20. Sutto L, Laetzer J, Hegler JA, Ferreiro DU, Wolynes PG (2007) Consequences of localized frustration for the folding mechanism of the IM7 protein. *Proc Natl Acad Sci USA* 104:19825–19830.
21. Ferreiro DU, Hegler JA, Komives EA, Wolynes PG (2007) Localizing frustration in native proteins and protein assemblies. *Proc Natl Acad Sci USA* 104:19819–19824.
22. Hilhorst HJ, Deutch JM (1975) Analysis of Monte Carlo results on the kinetics of lattice polymer chains with excluded volume. *J Chem Phys* 63:5153–5161.
23. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637.
24. Wang G, Dunbrack RL (2003) PISCES: A protein sequence culling server. *Bioinformatics* 19:1589–1591.