Improving the Performance of Rosetta Using Multiple Sequence Alignment Information and Global Measures of Hydrophobic Core Formation

Richard Bonneau,[1] Charlie E.M. Strauss,[2] and David Baker[1]*
[1]*Department of Biochemistry, Box 357350, University of Washington, Seattle, Washington*
[2]*Biosciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico*

ABSTRACT **This study explores the use of multiple sequence alignment (MSA) information and global measures of hydrophobic core formation for improving the Rosetta ab initio protein structure prediction method. The most effective use of the MSA information is achieved by carrying out independent folding simulations for a subset of the homologous sequences in the MSA and then identifying the free energy minima common to all folded sequences via simultaneous clustering of the independent folding runs. Global measures of hydrophobic core formation, using ellipsoidal rather than spherical representations of the hydrophobic core, are found to be useful in removing non-native conformations before cluster analysis. Through this combination of MSA information and global measures of protein core formation, we significantly increase the performance of Rosetta on a challenging test set. Proteins 2001;43:1–11.** © **2001 Wiley-Liss, Inc.**

Key words: ab initio structure prediction; homology; Rosetta

INTRODUCTION

Recent blind tests (CASPIII) have demonstrated significant progress in the field of ab initio protein fold prediction.[1,2] Although encouraging, this progress still leaves the field without methods that can reliably predict protein tertiary structure in the absence of homology to a sequence whose structure is known. One of the major hurdles that must be overcome in the development of consistently reliable ab initio protocols is the difficulty of discriminating near-native models from incorrect models.[3–6] The method our group has developed, Rosetta, is able to generate low root-mean-square deviation (RMSD) structures for most small proteins (good = 3–7.5 Å RMSD, small = <100 residues), but it is not always possible to recognize these structures amidst the larger population of incorrect decoys.[7] This work deals with this problem of recognition via two main approaches: the use of multiple sequence alignment (MSA) information and the use of global measures of hydrophobic core formation.

MSAs contain a great deal of information not available to prediction methods that use only a single sequence. Central to the main method described in this article is the empirical observation that two sequences sharing 25% sequence identity, for more than 60 amino acids, almost always share the same fold.[8–11] Each sequence alignment can therefore be thought of as representing a single fold, with each position in the MSA representing the preference for different amino acids and gaps at the corresponding position in the fold. Badretdinov and Finkelstein and colleagues used the theoretical framework of the random energy model to argue that the decoy discrimination problem is not currently possible unless the information in the MSA is used to smooth the energy landscape.[12,13] Using a three-dimensional (3D) cubic lattice model for a polypeptide, these investigators demonstrated that averaging a scoring function containing random errors over several homologous sequences allowed the correct fold to be separated from the rest of the averaged energy distribution, in spite of the random errors introduced into the energy function. Keasar et al.[14,15] showed that coupling the folding of several homologous sequences on a tetrahedral lattice significantly improved the performance of their Monte Carlo routine for a 36-residue peptide hormone. In this study, we test four methods for using MSA information to assist in ab initio structure prediction. The method we find most effective allows each homologous sequence to fold independently of other aligned sequences, only using the information in the MSA after the independent homologous folding runs are completed.

One of the major assumptions behind many ab initio folding potentials currently used is that the free energy of a conformation can be described as a sum of several pair-additive terms meant to describe specific interactions present in the molecule. There are many cases, however, for which this fundamental assumption is likely to be in error, especially cases involving entropies.[16,17] Solvation free energies are largely dominated by entropic terms and are therefore not well described by pair additive terms. The nearly ubiquitous presence of well-formed single hydrophobic cores in small proteins suggests using measures that explicitly monitor hydrophobic core formation.[4,18,19] We have developed three global features that together screen for non-native core packing and topology. Unlike previous global features designed to recognize well-formed hydrophobic cores, our features use an ellipsoi-

*Correspondence should be addressed to DB; e-mail: dabaker@u.washington.edu

dal rather than a spherical representation, allowing them to recognize a broader range of native core arrangements.

In the present study, we demonstrate that recognition of native-like structures produced by Rosetta can be enhanced using MSA information and our measures of hydrophobic core assembly. Using a combination of these two approaches has allowed us to produce good models for 15 of the 18 query sequences used to test the methods described in this work. Most of this increase in performance is the result of our novel use of MSA information. The global features are used to filter the decoy populations before the analysis of the homologous decoy populations via our clustering procedure, ridding the analysis of all recognizably misfolded structures, and thus offer a smaller but significant improvement in Rosetta's performance.

RESULTS AND DISCUSSION

Test Set Selection

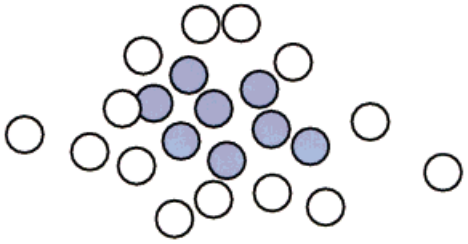
Initial tests of Rosetta were performed on a test set of 70 proteins less than 120 amino acids in length as described elsewhere (Simons et al., J Mol Biol, in press). For testing more computationally intensive methods, 18 query sequences were chosen from the larger set of 70. The smaller test set includes five positive controls, for which good models were generated using Rosetta and selected with our clustering routine. Ten of the 18 structures folded were structures for which Rosetta could produce good models (<7 Å RMSD) that were not identified by clustering (presumably because they occurred too rarely in the overall population). The remaining three were cases for which Rosetta did not produce any good models. Any improvement of the method must rescue some of the 13 cases for which no good models were selected while preserving the performance of the positive controls. The set contains 10 α/β, 5 β, and 3 α class proteins with a mean length of 79 residues. Because the main improvement to the method involved homologous sequence information, the smaller test set includes only sequences for which at least two sequences, not highly correlated to the query or each other, were aligned.

The Rosetta Method

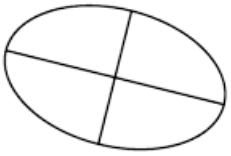
The basic method used to generate and select models has been previously described but will be reviewed briefly because of its importance as a starting point for the following discussion.[7,20] One of the fundamental assumptions underlying Rosetta is that the distribution of conformations sampled for a given nine residue segment of the chain is reasonably well approximated by the distribution of structures adopted by the sequence and closely related sequences in known protein structures. Fragment libraries for each 3- and 9-residue segment of the chain are extracted from the protein structure database using a sequence profile–profile comparison method as described previously. At no point is knowledge of the native structure used to select fragments or fix segments of the structure. The conformational space defined by these fragments is then searched using a Monte Carlo procedure with an energy function that favors compact structures with paired β strands and buried hydrophobic residues. A total of 1,000 independent simulations are carried out (starting from different random number seeds) for each query sequence, and the resulting structures are clustered as described in Materials and Methods and as previously reported.[21] The most reliable selection method, before this study, was simply to choose the centers of the largest clusters as the highest confidence models.[22] These cluster centers are then rank-ordered according to the size of the clusters they represent, with the cluster centers representing the largest clusters representing the highest confidence models. This protocol produces good models with rank 25th or better for 5 of the 18 queries folded in this study. In the larger set, the performance is somewhat better (as the 18-protein set was chosen to be a challenging set) and ~40% of small proteins (<100 residues) produce good models with the protocol described above. Before clustering, most structures produced by Rosetta are incorrect (i.e., good structures account for less than 10% of the conformations produced); for this reason, we refer to conformations generated by Rosetta as decoys. The problem of discriminating between good and bad decoys in Rosetta populations is the primary problem addressed in this work.

The Use of Global Features for Filtering Large Sets of Decoys

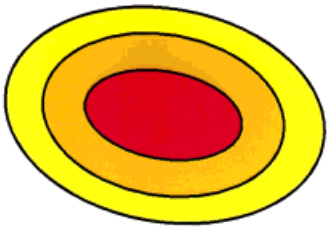
One of the problems we encountered during CASPIII, and during the tests of Rosetta immediately after CASPIII, was that the sequence dependent terms in Rosetta's energy function (the environment score and the pair score) allowed topologies that formed multiple small hydrophobic cores, as opposed to a single unified hydrophobic core. For proteins of the size attempted, the probability of having multiple domains is low,[23] and most hydrophobic cores can be roughly described by a single ellipsoid. In an effort to address this problem and other commonly encountered errors in decoys generated by Rosetta, we have developed three measures of hydrophobic core formation loosely based on a micelle model of proteins, which requires that hydrophobic residues partition to the interior of a protein and that the core is uniformly surrounded by backbone and hydrophilic residues. These features are evaluated using a method that partitions conformations into inner, middle, and outer ellipsoidal spaces for the purpose of recognizing single well-formed hydrophobic cores, as illustrated in Figure 1. The first feature, the core score, measures the extent to which hydrophobic side-chains partition to the central region of the molecule, to the exclusion of polar and charged side-chains. The second feature, the Cα partition score, measures the partitioning of backbone elements to the outer region of the molecule. Finally, the angle distribution score measures the degree to which the backbone uniformly surrounds the central core. Thus, the second and third score together check for the uniform enclosing of the core cavity by the backbone elements, while the first feature measures the extent to which the central cavity is filled by hydrophobic side-chains.



1. Select the most buried hydrophobic residues (blue).



2. Define inertial ellipsoid for this sub-set of residues.



3. Grow partitions in this elipsoidal basis to encompass 25 (red), 50 (orange) and 75% (yellow) of all residue centroids.

4. The core score is calculated by summing the log-odds table entries for all centroids in the red and orange region. The C-alpha score is the number of C-alpha's in the red region divided by the number in the yellow region.

Fig. 1. Schematic diagram of the ellipsoidal partitioning performed in order to calculate the core score and Cα-partitioning score.

Figure 2 presents a scatter plot of RMSD versus the core score for three proteins. A standard correlation coefficient is not useful for such a distribution because of the high number of false-positive results and the fact that RMSD of ≥8 Å are roughly equivalent for our purposes. Lower RMSD models, however, generally have higher core scores than the overall population as shown in the histograms in Figure 2. This suggests that the score should be used to filter out decoys falling in the worst part of the distribution (i.e., that the feature should be used to determine whether a decoy is bad, but not to determine whether the decoy is good, due partly to the high rate of false-positive results). To quantitate the value of each score as a filter, we calculate enrichment values. The decoys are sorted with respect to a given score (in this case, the core score), and all but the top percentage of the population (according to the score) is eliminated. In the case of Table I, all but 15% or 50% of the decoys are eliminated, so that only the decoys with the best core scores remain. The degree to which the ratio of good structures to incorrect structures is increased in the best scoring population as compared with the total population is calculated and reported as the enrichment value in Table I. An average enrichment value of 1.61 for the core score is reported in Table I. This shows that for the 18 homologous decoy sets the core score increased the ratio of good structures to bad structures by a factor of 1.61 when used to reduce the populations to 15% of their original size.

The Cα occupancy score was used to filter all 18 homologous decoy sets to eliminate topologies that violate the simple micelle model, and the enrichment values are reported in Table I. This score is useful for most proteins of <100 amino acids but, as the length of the query increases, the value of this score decreases, as the probability of strands traversing the inner core gets larger. The score is used here to rid the decoy populations of the worst scoring topologies (worst 25%) before clustering.

The angle distribution score was found to be most effective when used as a fixed cutoff, such that sets of decoys that do not minimize this score are filtered more heavily than sets of decoys that all fall below the cutoff. The angle distribution score enrichment values for all-β proteins was 1.7-fold with the angle-bin cutoff set such that fewer than two large gaps were allowed in the angular distribution of backbone elements around the common core.

The effectiveness of the features in the larger framework of the protocol for clustering multiple homologues is shown in column 12 of Table II. Before clustering, the decoys were filtered with the three global scores, described above and in Materials and Methods, so that the overall size of the population was reduced and the proportion of good decoys increased. The overall quality of the models produced by the filtered clustering did not improve drastically as compared with column 11 (unfiltered clustering). The rank of the first occurrence of a good model produced by the filtered clustering, however, was closer to first in many cases. In two cases (1kde, 2ife), models were selected when the unfiltered method failed to select good models.

Using MSA Information to Enhance Rosetta's Performance Energy Averaging

One possible way to use MSA information is to generate a population of decoys using the single query sequence, and then compute for each resultant model the average of the scores of each of the homologous sequences when mapped onto that structure.[13] We initially tested this score averaging method using the larger test set, which contained 1,000 decoys for each of 70 different proteins. The enrichment value for the core score increased from 1.84 to 2.26 when MSA information was summed using sequence weights to account for redundancies in the MSA.[24] Smaller improvements were seen when the pair

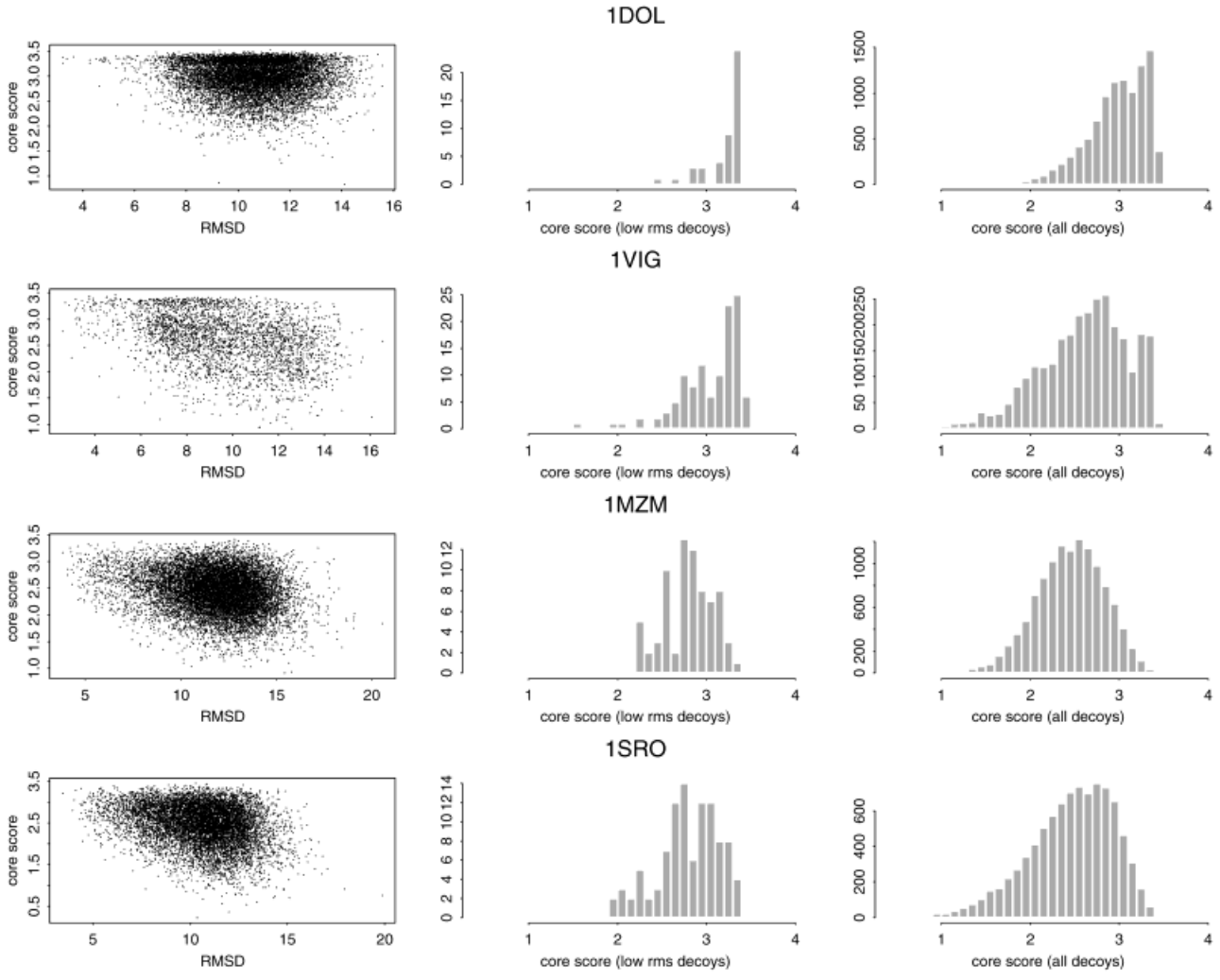


Fig. 2. Performance of the core score demonstrated for four proteins. Column 1, scatter plots of root-mean-square deviation (RMSD) versus the core score for populations of homologous decoys for 1dol, 1vig, 1mzm, and 1sro;column 2, histograms of the core score for decoys closer than 6.5 Å RMSD to native; column 3, core score histogram for the entire population.

TABLE I. Enrichment Values for Global Features[†]

Feature	Enrichment in top 15%	Enrichment in top 50%
Core score	1.61-fold	1.40-fold
C_α score	1.73-fold	1.28-fold

[†]Enrichment values averaged over the 18 sets of homologous decoys. The geometric mean of the enrichment values when homologous populations are filtered until only 15% or 50% of the original set are shown.

score and environment scores from Rosetta's scoring function were summed over the MSA in this way. Although there was some enhancement of the predictive power of these features, the enhancement was not large enough to improve the performance of Rosetta significantly.

Clustering Using Common-Residue RMSD

The regions that are unimportant to the structure and function of a protein may be less conserved, or absent, from many homologues in the alignment. One simple use of this information is to pay attention only to positions that are present in most of the sequences aligned to the query when attempting to select good models from a decoy population. When the positions that are gaps in many of the aligned sequences (often the edges of the alignment) are ignored, the average length of the targets in our set of 18 proteins drops from 79 to 68 residues. The yellow bar in Figure 3 shows the positions that remain after considering the MSA for 2acy. For each target, the positions common to all sequences folded were determined, and the number of these common positions is shown in Table II. Based on the assumption that the positions present in all aligned homologues are more important than positions often absent, we repeated the clustering for the full-length query sequence folding runs using RMSD over just these common residues as a distance metric (instead of global RMSD). Column 8 of Table II shows the results from clustering with global RMSD while column 9 shows the results obtained using

TABLE II. Summary of Clustering Results[†]

1	2	3	4	5	6	7	8	9	10	11	12	13	14
PDB code	2′ class	Best of 1,000 single seq. runs	No. of seq.	No. of res. in query seq.	No. of res. in MSA core	No. of res. common to all in MSA	Cluster centers from clustering 1,000 decoys (query seq.)	Cluster centers from 1,000 decoys (RMSD over just common residues)	Cluster centers from clustering MSA cores	Homologous clustering (no filtering)	Homologous clustering (after filtering)	Best MSA core match to cluster center	Best full-length match to cluster center
1a68	α/β	6.3	6	87	86	83	—	—	—	—	13th 7.5	—	—
1aca	α	5.0	9	86	79	79	—	—	7th 6.2 (6.2)	3rd 5.2	3rd 5.2	5.2 (5.2)	5.6/5.7
1ah9	β	5.2	9	71	52	49	—	—	—	—	—	—	—
1aoy	α/β	5.4	3	78	62	60	4th 5.9/7.1	5th 5.7/7.2	—	4th 6.2	3rd 6.2	6.2 (6.3)	6.4/7.0
1coo	α	5.3	13	81	55	52	5th 5.5/13.5	10th 3.7/10.4	—	2nd 2.8	2nd 2.8	3.2 (4.0)	4.3/6.3
1ctf	α/β	4.1	13	68	58	51	2nd 4.1/6.0 5th 3.2/4.9	2nd 4.0/5.5	3rd 4.3 (5.2)	1st 6.9 2nd 3.3	2nd 4.0	4.3 (6.3)	6.8/7.8 3.7/5.1
1dol	α/β	5.5	11	71	56	52	—	—	—	4th 6.9	4th 7.5	—	—
1hqi	α/β	6.8	8	90	70	70	—	—	—	10th 7.4	—	7.7 (7.7)	6.7/8.4
1kde	β	7.6	6	65	60	60	—	—	—	—	10th 7.3	7.6 (7.7)	—
1mzm	α	4.9	12	93	90	85	1st 4.9/7.0	1st 4.9/7.0	1st 4.7 (5.4)	1st 4.6	1st 4.4	5.2 (5.3)	6.8/6.9
1pse	β	6.9	11	69	63	53	—	—	14th 5.6 (6.2)	9th 5.9	4th 6.2	7.0 (7.3)	5.9/7.2
1sro	β	4.6	13	76	62	59	1st 6.0/6.7	1st 4.9/5.9	1st 5.0 (5.2)	1st 5.3	1st 4.8	4.8 (4.9)	5.2/7.0
1stu	α/β	6.3	2	68	64	64	—	—	—	2nd 6.4	1st 6.3	5.9 (5.9)	7.1/7.2
1tnt	α/β	5.2	3	76	65	46	6th 5.2/10.2	2nd 5.0/8.7	—	1st 6.1	1st 5.1	5.7 (5.8)	5.0/9.7
1vig	α/β	6.6	3	71	59	59	—	—	11th 6.5 (6.8)	1st 6.3	1st 6.6	5.1 (5.1)	6.9/10.3
1wkt	β	9.3	2	88	86	86	—	—	—	—	—	—	—
2acy	α/β	8.2	8	98	66	38	1st 5.9/14.2	1st 5.6/15.0	2nd 5.1 (5.6)	1st 2.8	1st 2.8	2.2 (4.7)	3.2/11.6
2ife	α/β	4.7	12	91	86	82	1st 5.4/6.0	1st 5.4/6.0	12th 6.5 (7.0)	—	14th 4.4	6.6 (6.7)	7.7/8.7

[†]PDB, Protein Data Bank; RMSD, root-mean-square deviation; MSA, multiple sequence alignment. *Column 1,* PDB code for each protein folded. *Column 3,* RMSD of the best structure generated in 2,000 Rosetta runs using only the native sequence. *Column 4,* number of sequences folded for each query. *Columns 5, 6, 7,* length of the full sequence, the length of the MSA core, and the number of residues common to all positions in MSA, respectively. *Column 8,* results from clustering the full-length sequence runs using full-length RMSD as a distance metric; *Column 9,* results for the same decoy population clustered with common residue RMSD. *Columns 8, 9, 14,* rank is given followed by the RMSD over the common residues and the global RMSD (common/global). *Columns 10, 13,* common residue RMSD followed by the RMSD over the MSA core. *Column 10,* rank of the best clusters when just the MSA cores are folded with the common residue RMSD followed by the RMSD over the length of the MSA core in parentheses. *Columns 11, 12,* clustering results for the unfiltered and global feature filtered, respectively, simultaneous clustering of multiple homologues, with only the rank and common residue rmsd given. *Columns 13, 14,* common residue and full-length RMSD for models selected from the MSA core and full-length single sequence runs using RMSD to the homologous models as a selection criterion. In all columns, the clusters reported are the best clusters in the top 15 clusters produced.

common position RMSD. Positions were considered common if they were present in all homologues selected for folding as described in the methods section and above. The decoy sets clustered for columns 8 and 9 are identical. Some improvement in the quality of models selected using common position RMSD is obtained but in no case are good models selected for targets that failed to produce good models using global RMSD as the distance metric.

Restricting Folding to Common Cores

An alternate approach is to fold only the cores of the MSA (the green bar in Fig. 3). In tests of ab initio folding algorithms, short unstructured N- and C-terminal regions have been eliminated to bolster the performance of the method being tested.[3,7] In these cases, the investigators used knowledge of the native structure to decide what positions to exclude from the folding simulations. We have made these decisions based only on the MSA. One might expect that by removing disordered tail segments, the quality of the models would increase over the remaining core region in cases where the tails interfered with Rosetta's folding of the more important core region.

We did this manually during CASPIII while predicting the structure of MarA, which had a large and diverse alignment to all but the N- and C-terminal residues. Ten terminal positions were excluded from the folding runs based on the alignment. These termini ended up largely unstructured in the experimental structure, and the runs with the unaligned tail positions excluded converged to good models (which were submitted) while the full-length runs failed to converge to good models.

After refolding just the MSA cores, and clustering using RMSD over all residues remaining in the core of the MSA, we find that 9 of the 18 targets have good models in the top 15 cluster centers. Using only this simplest form of information present in the MSA, we achieve a moderate increase in performance without increasing the time required for the calculation.

Simultaneous Clustering of Multiple Independent Homologous Folding Runs

The methods above are limited in that they are all reliant on populations of decoy conformations constructed using only the single query sequence and therefore do not use MSA information to influence the buildup procedure. To overcome this limitation, we have developed a method whereby a subset of the aligned sequences are allowed to fold independently; the resultant homologous populations

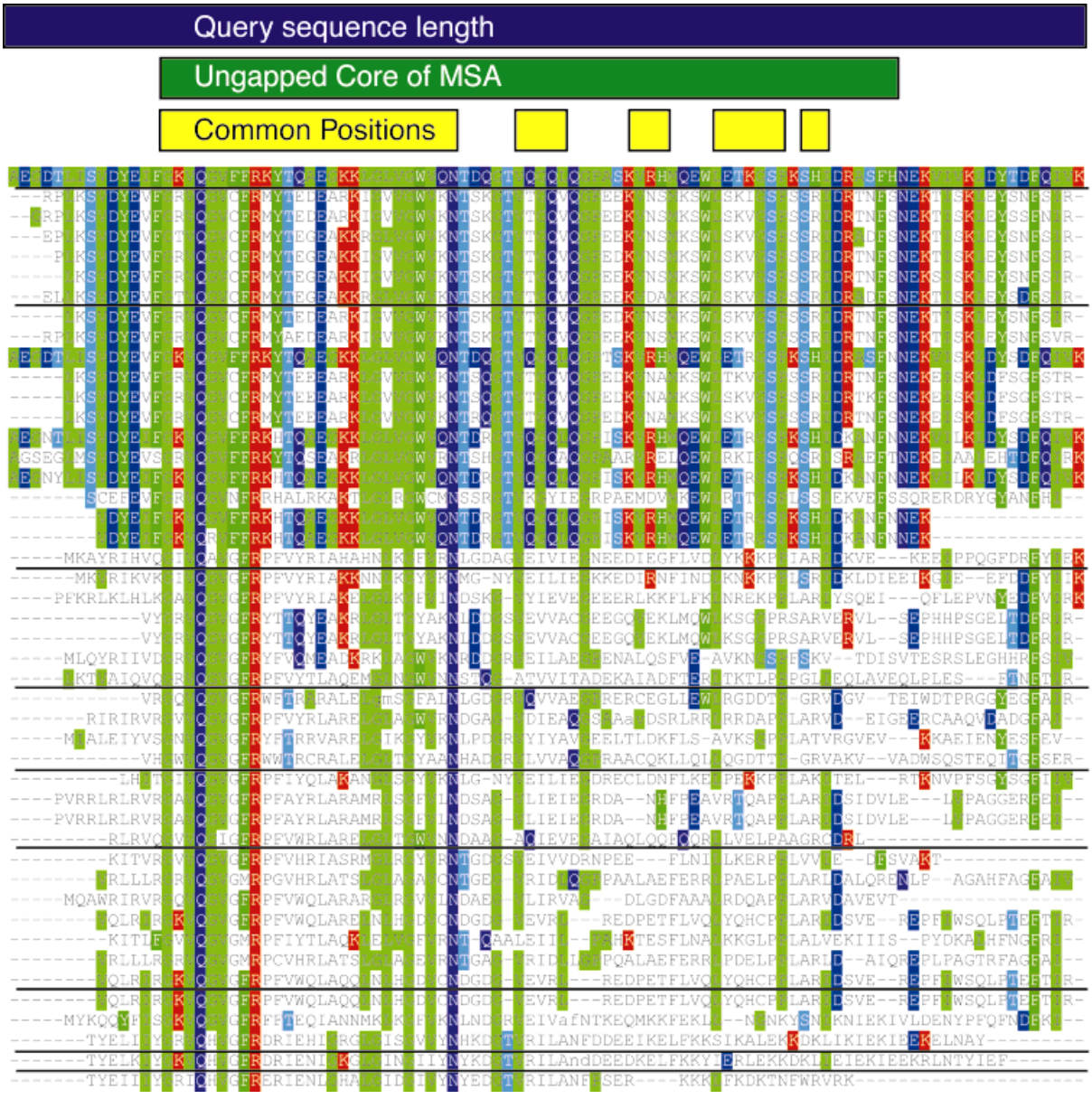


Fig. 3. MSA for 2acy. Sequences are ordered by PSI-BLAST *E*-value. The sequences folded using Rosetta are underlined. Bars at top represent the full length of the query sequence (blue), the ungapped core of the MSA (green), and the positions common to sequences folded (yellow).

are clustered simultaneously to produce the final models. Each homologous sequence thus finds its own energy minima independent of other homologous sequences. We then compare all minima, each corresponding to a cluster of decoy conformations generated with a single sequence with each other via simultaneous clustering.

The ideal result is that only the correct free energy minima will be present for all the homologues folded, and that the false minima (incorrect conformations) will be sufficiently different for different aligned sequences so as to be diluted out as the homologous decoy populations are combined. One advantage of this procedure comes from the way the method handles gaps and insertions. If a gap or insertion is present in one sequence and absent in another sequence in the alignment, the gap/insertion will effect a distance constraint when the structures generated for these sequences are compared using common residue RMSD clustering. The degree to which the constraint will aid in the selection of the correct cluster/decoy is obviously low if the gap is short; nevertheless, the information is automatically considered by the simultaneous clustering procedure.

Of the four methods discussed, the most consistent improvement in our ability to generate good models comes from folding multiple homologues for each query. In Table II, column 4 shows the number of sequences selected, as

described in the methods section, for each query. Once a representative subset of the entire MSA was selected, an independent Rosetta run was carried out for each of the selected sequences, producing 1,000 decoys for each sequence. Once generated, the multiple decoys were clustered using RMSD over positions common to all sequences folded as a distance metric. This common position RMSD was chosen so that pairwise RMSD would be comparable for homologues of variable length and composition.

The results for simultaneously clustering homologues for the 18 queries in the test set are shown in column 11 of Table II. In 13 of the 18 cases, good models are produced in the top 10 clusters when the clusters are sorted by size, and in all but 2 of these 13 successes, the good model is in the top 5. This method selects better models with ranks closer to 1 than does the original global RMSD clustering of the single-query sequence folding runs. Generated models are compared with the corresponding native structure for an α, a β, and an α/β class protein in Figure 2.

Once homologous models are obtained, the remaining problem is how to use these homologous models to build models completely consistent with the full-length query sequence (the homologous models may contain gaps or insertions that obviously complicate this process). Using the homologous models to fish through full-length query and MSA-core folding runs (using common position RMSD to the homologous model as a selection criterion) allowed us to select good models for all queries for which there were good query-sequence models. The last two columns in Table II show that good models over the common residues were almost always selected, but globally correct models were sometimes not present in the full-length single-query sequence runs and were therefore not selected. This is the case for 2acy, where the core 68 residues converged to a very low RMSD model, but the C-terminal strands were never folded correctly in the query single sequence runs and were absent in most of the homologues. Consequently, no good models were produced for this region, although 68 residues (the MSA core) were predicted to 4.7 Å RMSD.

One slight drawback to this method is that it increases the length of the folding run by, on average, 6.5-fold for each query. Currently this is not an issue; the 6.5-fold increase still leaves the calculation short enough so that it can be completed in less than 1 week on a 500-MHz Intel Celeron processor. If this increase in computer time were to become an issue, the length of the calculation could then be reduced by reducing the number of homologous sequences folded or resorting to the simpler methods described above. For five of the queries, only 2–3 sequences were folded due to the shallowness of their MSA. For these sequences, some improvement is still gained by simultaneous clustering even with this small number of homologues, suggesting that folding fewer homologues when the length of the calculation becomes problematic will not completely destroy the increase in performance.

CONCLUSIONS

Recognition of native-like structures produced by Rosetta is considerably improved using MSA information and global measures of hydrophobic core formation. Low RMSD (<7.5 Å) models were ranked 15th or better for 15 of the 18 queries attempted and 5th or better for 13 of the 18 queries (Table II). To indicate the quality of models with RMSD within this range, the models generated for three of the targets are shown in Figure 4.

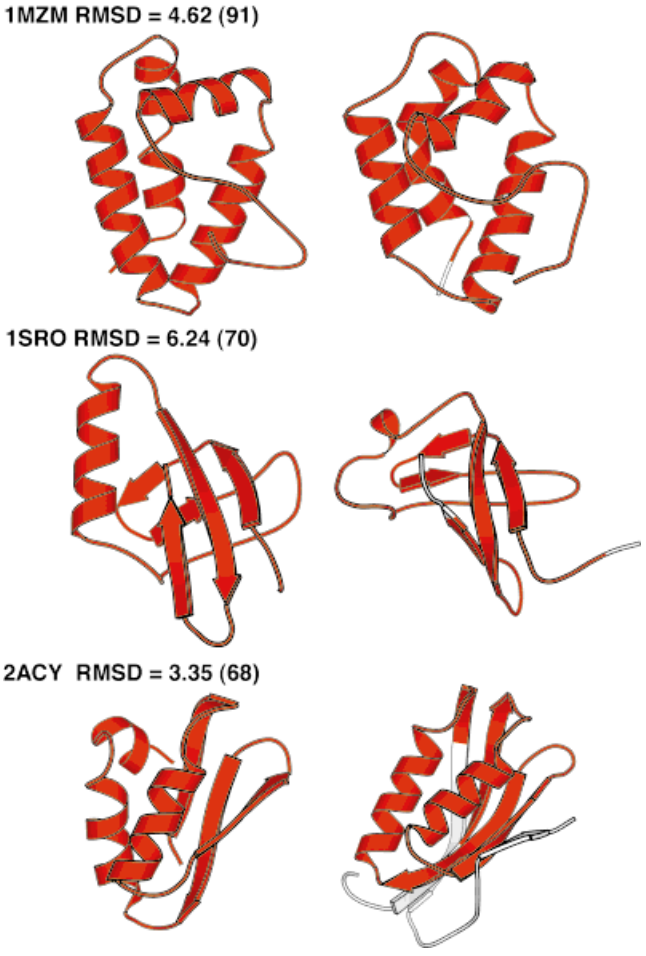


Fig. 4. Models for three of the 18 queries for which multiple homologs were clustered. Models selected by the clustering routine are shown on the left. Native structures are shown on the right with residues matching the homologous model shown in red. RMSD values are give followed by the number of matching residues in parenthesis.

The methods described for assessing hydrophobic core formation and using multiple sequence information improve on previously developed methods. The ellipsoid model of protein cores is considerably more general than the spherical models implicit in previous methods—the ellipsoidal models of the three native proteins shown in Figure 6 are clearly better representations than the best spherical model. The hydrophobic core score also avoids some of the problems inherent in pair-additive measures of hydrophobic core association. Our use of multiple sequence information—the clustering of structures generated in independent simulations with different homologue sequences—outperforms methods based on simple score averaging in facilitating recognition of low RMSD structures. The searches of conformational space for each

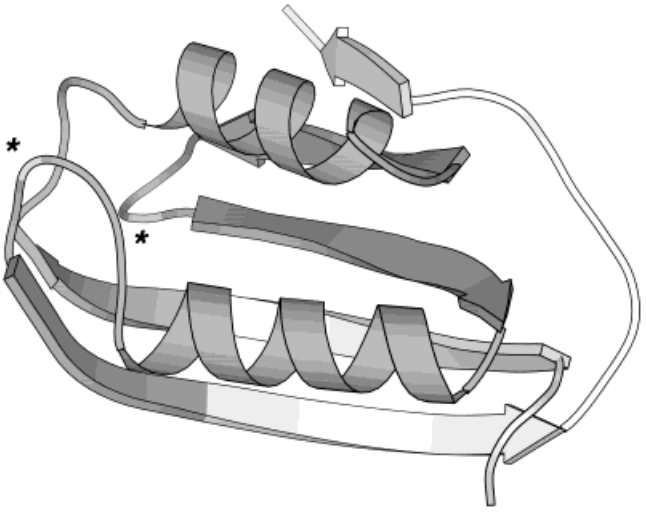


Fig. 5. Native structure of 2ACY shown with the positions common to all homologues folded, shown in gray. Two internal gaps, evident upon inspection of Fig. 3, are indicated with asterisks.

homologous sequence yield the lowest free energy structures consistent with the local sequence biases for that sequence, and hence the clustering of the structures generated for a number of different homologues identifies the conformations that are most consistent with the (different) local sequence biases and nonlocal interactions represented within the protein family.

MATERIALS AND METHODS

MSA Generation and Homologue Selection

Each query sequence was taken directly from the corresponding PDB file. The MSA was generated using PSI-BLAST[26] and then inspected using MView.[27] Because of time constraints, we folded only a representative subset of the MSA. The sequences were chosen to be a representative, diverse, subset of the MSA. Sequences shorter than 60% of the query sequence length were discarded, as were sequences greater than 60% or less than 20% sequence identical to the query. Pearson correlation coefficient matrices based on sequence identity were then generated for each MSA. A total of 12 sequences were chosen such that their summed row in the correlation matrix was lowest (in cases in which fewer sequences were present all sequences were selected). Once a subset was chosen, sequences with greater than 60% sequence identity to other sequences in the subset were removed. If the shortest sequence, over 60% of the length of the query, was not in the set selected for folding, it was added to ensure that the shortest allowable sequence was included in the final set of homologues to be folded (regardless of its sequence identity to sequences in the subset before adding the shortest sequence). Once the final set of homologues to be folded was selected, the positions common to all sequences selected were determined. At this stage, the common core of the MSA is the region between the first and last common residues in the subset; positions common to all homologues are also in reference to this subset (see Figs. 3 and 5).

General Clustering Routine

Our general procedure for clustering populations of decoys has been described previously but will be reviewed here due to its importance to the methods that follow.[7] Two structures are considered neighbors if they are closer in Cα RMSD than an empirically derived cutoff. The clustering procedure is iterative and begins by calculating a list of neighbors for each structure. The structure with the largest number of neighbors according to this list is then the center of the first, largest, cluster. The cutoff for considering two structures neighbors is started at 8.0 Å and iteratively reduced until the first cluster contains 50–100 decoys or until the cutoff has reached 3.0 Å. Once one of these conditions is met, the cutoff is fixed for the remaining iterations. The first cluster center is then written out and its neighbors are removed from the population. The process is then repeated until the clusters produced contain fewer than 5 neighbors. For populations of $<$3,000 decoys, the first cluster was set to contain 50 members, for populations of $>$3,000, the first cluster was set to contain 100 decoys.

Simultaneous Clustering of Multiple Homologues

For each sequence, secondary structure predictions are made and profiles generated as described above. Independently generating a sequence profile for each homologue is necessary because of the presence of gaps and insertions in the homologous sequences. Once the profile is generated, secondary structure predictions are made using PSI-PRED, DSC, and PHD.[15,28–30] Fragment libraries are then generated for each sequence. A total of 1,000 decoys are generated for each homologous sequence and the original query sequence using Rosetta. A total of 1,000 decoys were also generated for each query's core MSA positions using the fragments generated with the full length query. The sequences are then clustered as above, using RMSD over the positions common to all homologous sequences folded and the MSA as the key for mapping the different homologous decoys onto each other. Clusters reported in Table II are ranked by size (i.e. the number of neighbors in each cluster). This practice of ranking by cluster size is supported by our results from CASP3[30] and the work of Shortle et al.,[22] who found that, in populations of decoys generated by Rosetta, native-like decoys were surrounded by a larger number of similar conformations than non-native decoys.

Global Measures of Core Formation

Ellipsoidal demarcations of protein cores were calculated based on the most buried 75% of the hydrophobic residues. Burial is approx. by the number of centroid–centroid contacts of less than 10 Å, the 25% of residues with the fewest contacts are removed from the demarcating ellipsoid calculation. The principle axes of the remaining buried hydrophobic residues are calculated based on the coordinates of their Cβ, Cα, and centroid atoms (the centroids are weighted according to the number of hydrophobic heavy atoms they represent). Once the axes of the best-fit ellipsoids (the principal axes) are determined the

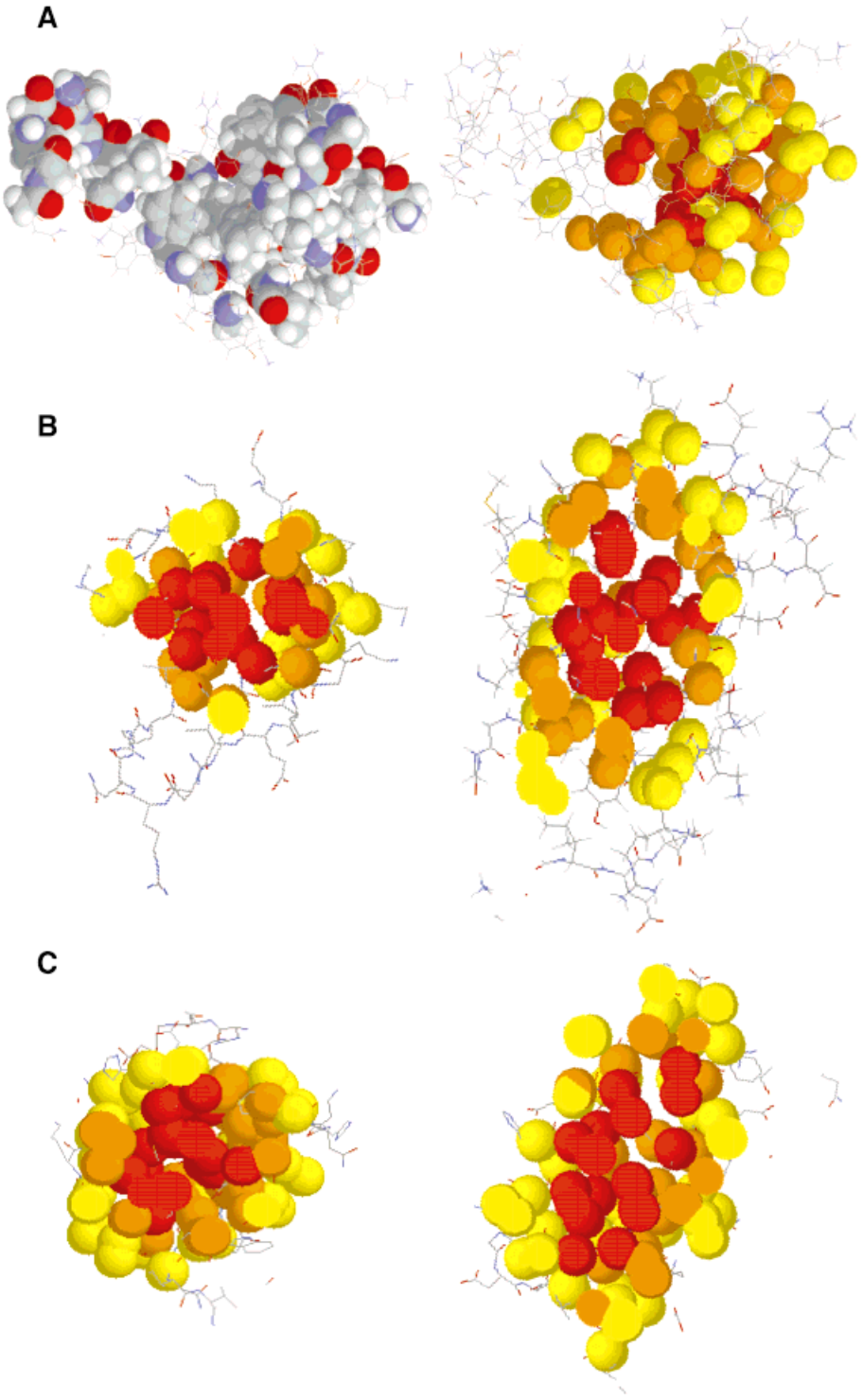


Fig. 6. **A:** 1PSE with all hydrophobics shown as CPK space-filling (left) and with all hydrophobics partitioned into inner (red), middle (orange), and outer (yellow) cores (right). Note the exclusion of a large number of unstructured surface hydrophobics. **B:** Two orthogonal cutaway views of the native structure for 1SRO with inner, middle, and outer cores labeled red, orange, and yellow space filling, respectively. Surface residues are shown in wireframe. **C:** Same view for 2ACY. Core boundaries for A–C were generated as described in Materials and Methods.

conformation is divided into four regions by scaling the principal axis such that the resultant three ellipsoids contain 25%, 50%, and 75% of all centroids. These three ellipsoids now partition the decoy into four regions as shown in two dimensions in Figure 4. The case for using only the most buried 75% of the hydrophobic residues is

important for proteins such as 1PSE, which have considerable numbers of exposed hydrophobic residues (Fig. 4).

The core score was parameterized using 560 small, single core, native structures. For each native structure the ellipsoidal demarcations were calculated as described above and the frequencies of all amino acids occurring in each of the four regions were tabulated. Our initial set of hydrophobics was as follows: A, F, G, I, L, M, P, V, W, and Y. After looking at the initial frequency table for the inner and middle cores we revised our hydrophobic set to: A, C, F, H, I, L, M, V, W, Y (−G, −P, +C, +H). The inclusion of C is not surprising and H is probably present due to its participation in metal binding sites. Once the new set of hydrophobics was determined a frequency table (describing the probability of finding different amino acids in the four possible shells defined individually for each native structure) was regenerated and normalized to account for the overall frequency of each amino acid in the 560 sequences. From this normalized frequency table, a log-odds table was generated. The core score is determined by calculating the demarcating ellipsoids for a given decoy and then summing the log odds table entries for each residue in each of the four regions; this sum now includes hydrophilic residues as well as the hydrophobics used to calculate the demarcating ellipsoids. We found that the most significant signal was obtained by combining the log odds sums for the inner and middle core. For the prefiltered clustering the 25% of the decoy population with the worst core score was removed and the remaining decoys were then filtered with the remaining features and clustered.

The Cα score is simply the number of Cα atoms in the innermost core divided by the number of atoms in the outer core. This quantity is low for most proteins and high for ~50% of the decoys generated by Rosetta. In the filtered tests of the simultaneous clustering, the Cα partition score was used to filter rare topologies (topologies with many backbone elements penetrating the core) from homologous decoy sets by removing the 25% of decoys with the highest Cα partition score.

The third feature conceptually derived from the micelle model is the angle distribution score. This score attempts to exclude topologies where the backbone atoms do not surround the core. Using spherical polar coordinates, we divide the surface of a sphere centered at the conformation's center of mass into bins with equal surface area (ϕ boundaries at $-\pi$, -2.356, -1.571, -0.7853, 0, .785, 1.571, 2.356, and π; θ boundaries at: 0, 0.72, 1.04, 1.57, 2.09, 2.42, and π). Each bin has eight neighboring bins. The score is calculated by first binning the Cα atoms and tabulating the number of Cα in each bin. Using this table, the number of "holes" in the angle distribution are counted, where a hole is an empty bin with five or more empty neighboring bins. The angle distribution score is then the number of holes averaged over three orthogonal choices for the *z*-axis, to minimize effects due to the limitations of the differently shaped longitude/latitude bins. Decoys having an angle distribution score of >6.0 (>6 holes) were removed from the population. The fixed angle distribution score of 6.0 allows for roughly two holes in the distribution of backbone atoms around the single unified core.

Each filtering is performed on the original decoy population; thus, decoys removed by multiple methods are equivalent to decoys removed by just one feature. Once filtering is complete the remaining decoys are passed to the simultaneous clustering routine.

ACKNOWLEDGMENTS

We thank Carol Rohl, Brian Kuhlman, Jerry Tsai, and Peter Bowers for their careful reading of the manuscript and for the discussion that followed. Keith E. Laidig was instrumental to the work through effective administration of the resources necessary for completing the calculations. We also thank Ingo Ruczinski for several useful suggestions. R.B. acknowledges support from a HHMI predoctoral fellowship.

REFERENCES

1. Moult J, Hubbard T, Fidelis K, Pedersen JT. Critical assessment of methods of protein structure prediction (CASP): round III. Proteins 1999;Suppl 3:2–6.
2. Venclovas C, Zemla A, Fidelis K, Moult J. Some measures of comparative performance in the three CASPs. Proteins 1999; Suppl 3:231–237.
3. Eyrich VA, Standley DM, Friesner RA. Prediction of protein tertiary structure to low resolution: performance for a large and structurally diverse test set. J Mol Biol 1999;288:725–742.
4. Park BH, Huang ES, Levitt M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J Mol Biol 1997;266:831–834
5. Samudrala R, Xia Y, Levitt M, Huang ES. A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. Pacific Symp Biocomput 1999;505–516.
6. Sippl MJ. Knowledge-based potentials for proteins. Curr Opin Struct Biol 1995;5:229–235.
7. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34:82–95.
8. Abagyan RA, Batalov S. Do aligned sequences share the same fold? J Mol Biol 1997;273:355–368.
9. Brenner SE, Chothia C, Hubbard TJ. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci USA 1998;95:6073–6078.
10. Chung SY, Subbiah S. A structural explanation for the twilight zone of protein sequence homology. Structure 1996;4:1123–1127.
11. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.
12. Bryngelson JD, Wolynes PG. Intermediates and barrier crossing in a random energy model. J Phys Chem 1989;93:6902–6915.
13. Badretdinov A, Finkelstein AV. How homologs can help to predict protein folds even though they cannot be predicted for individual sequences. J Comput Biol 1998;5:369–376.
14. Reva BA, Skolnick J, Finkelstein AV. Averaging interaction energies over homologs improves protein fold recognition in gapless threading. Proteins 1999;35:353–359.
15. Keasar C, Tobi D, Elber R, Skolnick J. Coupling the folding of homologous proteins. Proc Natl Acad Sci USA 1998;95:5880–5883.
16. Dill KA. Additivity principles in biochemistry. J Biol Chem 1997;272:701–704.
17. Mark AE, van Gunsteren WF. Decomposition of the free energy of a system in terms of specific interactions. J Mol Biol 1994;240:167–176.
18. Bowie JU, Eisenberg D. An evolutionary approach to folding small α-helical proteins that uses sequence information and an empirical guiding fitness function. Proc Natl Acad Sci USA 1994;91:4436–4440.
19. Huang ES, Subbiah S, Levitt M. Recognizing native folds by the

arrangement of hydrophobic and polar residues. J Mol Biol 1995;252:709–720.
20. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
21. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 1999;37:171–176.
22. Shortle D, Simons KT, Baker D. Clustering of low-energy conformations near the native structures of small proteins. Proc Natl Acad Sci USA 1998;95:11158–11162.
23. Orengo CA, Pearl FM, Bray JE, Todd AE, Martin AC, Lo Conte L, Thornton JM. The CATH Database provides insights into protein structure/function relationships. Nucleic Acids Res 1999;27:275–279.
24. Sander C, Schneider R. The HSSP data base of protein structure–sequence alignments. Nucleic Acids Res 1993;21:3105–3109.
25. Vingron M, Sibbald PR. Weighting in sequence space: a comparison of methods in terms of generalized sequences. Proc Natl Acad Sci USA 1993;90:8777–8781.
26. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
27. Brown NP, Leroy C, Sander C. MView: a web-compatible database search or multiple alignment viewer. Bioinformatics 1998;14:380–381.
28. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
29. King RD, Sternberg MJ. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Protein Sci 1996;5:2298–2310.
30. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.