doi:10.1006/jmbi.2000.4459 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2001) 306, 1191–1199

JMB

# Prospects for *ab initio* Protein Structural Genomics

**Kim T. Simons, Charlie Strauss and David Baker***

*University of Washington*
*Box 357350, Seattle*
*WA 98195, USA*

We present the results of a large-scale testing of the ROSETTA method for *ab initio* protein structure prediction. Models were generated for two independently generated lists of small proteins (up to 150 amino acid residues), and the results were evaluated using traditional rmsd based measures and a novel measure based on the structure-based comparison of the models to the structures in the PDB using DALI. For 111 of 136 all α and α/β proteins 50 to 150 residues in length, the method produced at least one model within 7 Å rmsd of the native structure in 1000 attempts. For 60 of these proteins, the closest structure match in the PDB to at least one of the ten most frequently generated conformations was found to be structurally related (four standard deviations above background) to the native protein. These results suggest that *ab initio* structure prediction approaches may soon be useful for generating low resolution models and identifying distantly related proteins with similar structures and perhaps functions for these classes of proteins on the genome scale.



*Corresponding author

*Keywords: ab initio* structure prediction; protein folding; proteomics

## Introduction

Protein sequences are being determined at a rate faster than the solutions of their three-dimensional structures. As structural information is critical to our understanding of the basis of the biological properties of protein molecules, there is a tremendous incentive to develop computational methods for obtaining such information. Results in the recent CASP III protein structure prediction experiments demonstrated that considerable progress has been made toward predicting protein structure from primary sequence information alone (Moult *et al*., 1999).

The ROSETTA method for *ab initio* protein structure prediction developed in our group was one of the best methods tested in CASP III (Orengo *et al*., 1999; Simons *et al*., 1999a). ROSETTA is based on the assumption 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 three and nine residue segment of the chain are extracted from the protein structure database using a sequence profile-profile comparison method as described by Simons *et al*. (1997). In the calculations described here, proteins homologous to the sequence being folded (Ψ-blast *e*-value <0.01) were rigorously excluded from the fragment libraries since our goal is to characterize the performance of the method for *ab initio* structure prediction rather than homology modeling. 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 1000 independent simulations are carried out (starting from different random number seeds) for each query sequence, and the resulting structures are clustered as described by Shortle *et al*. (1998).

The structures of large portions of several proteins were reasonably accurately predicted using ROSETTA in the CASP III structure experiment. For example, a 99-residue segment of the transcription factor MarA was predicted to 6.4 Å rmsd (root mean square deviation), and a 75-residue fragment of the dnaB helicase was predicted to 4.7 Å rmsd. To test the method more comprehensively to determine what problems can be tackled currently and which require further methods development, we have carried out large-scale predictions using two independently compiled lists of small protein

Present addresses: K. T. Simons, Harvard University, 7 Divinity Avenue, Cambridge, MA 02138, USA; C. Strauss, Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA.

Abbreviations used: PDB, Protein Data Bank; rmsd, root-mean-square deviation.

E-mail address of the corresponding author: dabaker@u.washington.edu

0022-2836/01/051191–9 $35.00/0

structures, a total of 172 proteins in all. The first list was that used in a recent study by Friesner and co-workers (Eyrich *et al*., 1999) and the second list was derived from the PDB-select 25 protein set (see Methods). The use of large independently compiled lists, while still not as perfect a test as that provided by the CASP process, does alleviate the two dangers of training the method to perform well on a small test set and of tailoring the test set to the strengths of the method.

## Results and Discussion

Sets of 1000 structures were generated and clustered for each protein on each of two lists. The results are summarized in Tables 1 and 2; the proteins are divided into separate categories based on size (less than 50 residues (small), between 50 and 100 residues (medium) and between 100 and 150 residues (large)) and secondary structure (all α, all β, and α/β). The results for each individual protein (except the small proteins to save space) are presented in Table 1, and an overall summary of the performance of the method for the different size and secondary structure classes is presented in Table 2.

A conformation within 5 Å rmsd to the native structure was generated for 24 of the 30 small proteins studied, and a structure within 7 Å rmsd, for all 30 proteins. For eight of ten of the α-helical small proteins and five of 12 of the α/β proteins, one of the top five cluster centers was within 5 Å rmsd of the native structure.

A conformation within 5 Å rmsd to the native structure was generated for 56 of the 127 medium-sized proteins studied, and a structure within 7 Å rmsd, for 93 of the 127 proteins. For 46 of these 127 proteins, one of the top five cluster centers was within 7 Å rmsd of the native structure. Results were significantly worse for the all β proteins: conformations within 7 Å rmsd of the native structure were generated for only 11 of 27 β proteins, and for only five of these 27 proteins was one of the top five cluster centers within 7 Å rmsd of the native. In the large protein category, conformations within 7 Å rmsd were generated for three of the nine α proteins and four of the five α/β proteins.

The large number of simulated structures generated for a large number of different sequences allows a more global evaluation of the scoring function and the search strategy. The scoring function contains terms which describe the density of neighbors surrounding a residue, the distances between pairs of residues given their densities, strand pairing geometry, and the grouping of β-strands in β-sheets (Simons *et al*., 1999a,b). For each sequence, we randomly collected ten structures of the 1000 created and reconstructed the distributions used for the generation of the scoring function. We found little difference between the input distributions obtained from experimentally determined structures and the output distributions obtained from the simulated structures (Figure 1). For the most part, all components of the scoring function were minimized very well by the Metropolis Monte Carlo method (Metropolis *et al*., 1953). Proteins of more than 80 residues and mostly β-strand structure were the exception: the minimization procedure failed to produce structures of better score compared to the native fold (Figure 2). At present, we are exploring the use of alternate minimization approaches for longer sequences and additional features of native protein structure not captured by the current scoring function. There is hope for improvement both from utilization of a more detailed full-atom model and from incorporation of more global features.

While rmsd is the standard measure for evaluating structure models, it is somewhat removed from the ultimate interest of a user of a structure prediction method. If functional insights can be gained from a 7 Å model, it is a better model in a very real sense than a 5 Å model which does not allow such insights. In particular, in the context of genome level structure predictions, there is a very real value to models which allow for correct annotation of previously unannotated sequences. With the goals both of providing an alternative to rmsd for evaluating structure predictions and for assessing the current prospects of an ''*ab initio* structural genomics'' strategy, we used the structure-structure comparison method DALI (Holm & Sander, 1995) to identify the closest structure in the PDB for each of the top ten clusters for each of the proteins in the two lists. Since there are many very closely related structures in the PDB, a positive result is not only a match to the native structure, but also a match to a structure similar to the native structure. Thus, each of the matches to a cluster center in the PDB was in turn compared using DALI to the native structure. A *Z*-score of >4 was somewhat arbitrarily chosen as a structural similarity threshold; i.e. a model was considered a successful prediction if the closest DALI match in the PDB had a *Z*-score of >4 when compared with the true structure.

For roughly half of the all α and α/β proteins in the medium size class, one of the top ten cluster centers was a success according to this DALI-based criteria. More remarkable is that for five of the nine large α-helical proteins and four of the five large α/β proteins, one of the top ten cluster centers was a success according to this measure. This success with the large α/β proteins should be contrasted with the complete failure to generate structures within 5 Å rmsd for any of these proteins: although the models are not very accurate, they retain sufficient features characteristic of the true structure to specifically recognize proteins structurally related to the true structure. It must be emphasized, however, that the majority of the DALI matches are false positives (i.e. matches to proteins not in the same superfamily), and thus the *ab initio* structure prediction followed by the DALI

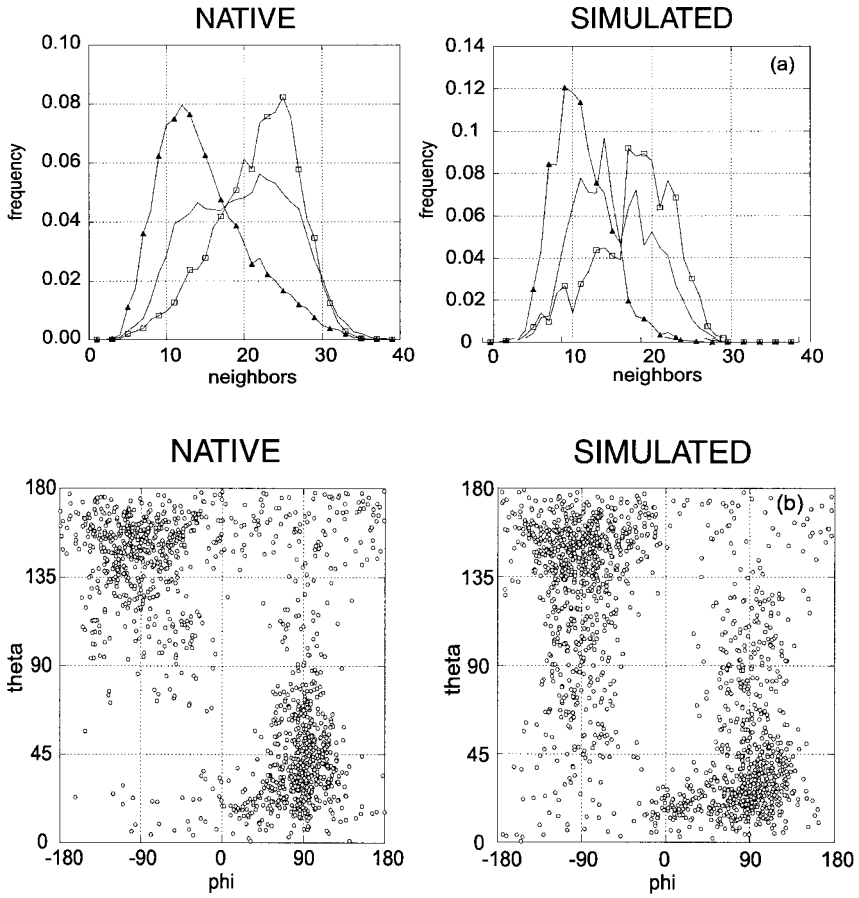


**Figure 1.** Comparison of residue-environment and strand pairing distributions in native and simulated structures. (a) The frequencies of different numbers of residue centroids within 10 Å for isoleucine (squares), alanine and aspartate (triangles) in native (left) and simulated structures (right). The *x*-axis is the neighbor density (the number of residues within 10 Å), and the *y*-axis, the frequency with which this neighbor density is observed for the particular residue type. The neighbor densities extend to larger numbers for native proteins because the simulated structures are shorter than the majority of proteins contributing to the native plots. (b) The relative orientation of dipeptides in paired β-strands in native structures is similar to that in simulated structures. The angles are defined by Simons *et al.* (1999b).

approach is currently better viewed as narrowing down the set of possible structures than uniquely identifying the correct family.

Cartoon diagrams illustrating some of the more interesting models are presented in Figure 3. The predicted structure (Figure 3, right) is the cluster center indicated in the DALI > 2 column of Table 1. The models provide an indication of how DALI is able to detect the similarities to the native structures. In some cases, although there is not global

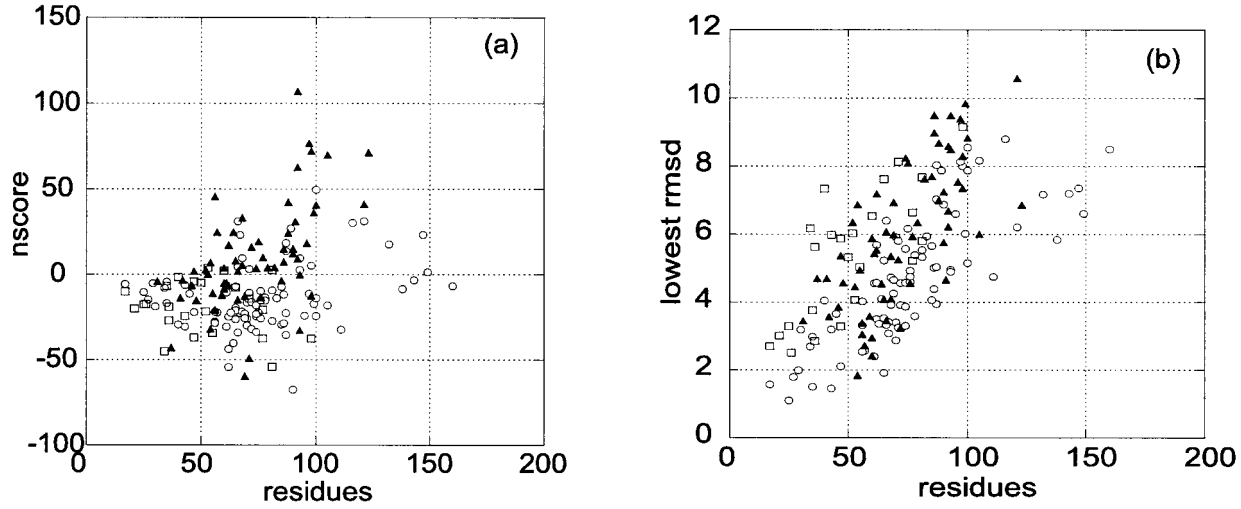


**Figure 2.** Effect of protein length on the folding simulations. (a) Normalized score *versus* protein length. For each of the 172 sequences, the average score of the 1000 structures minus the native score (*y*-axis) *versus* the length of the protein (*x*-axis). (b) The lowest rmsd structure generated in the 1000 simulations *versus* the length of the protein. Symbols indicate the secondary structure of the native protein: >25% β-strand (triangles), >25% α-helical (circles), and other (squares).

**Table 1.** Prediction results

| Name | $N_{res}$ | $N_{\alpha}$ | $N_{\beta}$ | <5 Å | <6 Å | <7 Å | Low rmsd | Cluster cutoff | DALI rank $z > 2$ | DALI rank $z > 4$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Medium α | | | | | | | | | | |
| 1a1z* | 83 | 62 | 0 | - | - | 87 | 5.9 | 6.0 | Low | Low |
| 1a32* | 65 | 55 | 0 | 1 | 1 | 1 | 1.9 | 2.3 | 1 | 1 |
| 1a6s* | 87 | 54 | 0 | - | 96 | 96 | 5.1 | 7.0 | Low | Low |
| 1aab* | 74 | 40 | 0 | - | - | 43 | 5.6 | 4.0 | 1 | 2 |
| 1aca* | 86 | 58 | 0 | - | 75 | 40 | 5.0 | 6.6 | 2 | - |
| 1acp | 73 | 39 | 0 | - | 16 | 16 | 4.4 | 6.2 | - | - |
| 1adr* | 76 | 44 | 0 | - | 18 | 18 | 4.9 | 6.0 | 4 | 4 |
| 2af8* | 86 | 43 | 0 | - | 1 | 1 | 4.4 | 7.0 | Low | - |
| 1ail | 67 | 57 | 0 | - | 27 | 23 | 3.1 | 6.5 | 4 | Low |
| 1ail* | 70 | 59 | 0 | - | - | 58 | 5.9 | 7.0 | 2 | 2 |
| 1aj3 | 95 | 84 | 0 | - | - | - | 6.6 | 6.0 | 1 | 2 |
| 1am3 | 57 | 41 | 0 | 1 | 1 | 1 | 2.6 | 3.0 | Low | Low |
| 1bw6* | 56 | 32 | 0 | 1 | 1 | 1 | 2.5 | 3.0 | 1 | 8 |
| 1c5a | 62 | 45 | 0 | 2 | 1 | 1 | 3.5 | 4.1 | 2 | - |
| 1cc5 | 76 | 39 | 0 | - | - | 2 | 4.6 | 7.0 | 10 | 10 |
| 1cei* | 85 | 49 | 0 | - | - | - | 5.7 | 7.0 | 5 | Low |
| 1coo* | 81 | 35 | 0 | - | - | 32 | 5.3 | 7.0 | - | - |
| 1ddf | 87 | 60 | 0 | 420 | 420 | 270 | 4.0 | 4.5 | Low | Low |
| 2ezh | 65 | 45 | 0 | 1 | 1 | 1 | 3.4 | 5.0 | 4 | 9 |
| 2ezh* | 65 | 45 | 0 | 17 | 14 | 14 | 3.5 | 4.0 | 2 | - |
| 2ezk | 93 | 63 | 0 | 210 | 210 | 210 | 5.0 | 7.0 | 3 | Low |
| 2ezl* | 99 | 58 | 0 | - | - | - | 6.0 | 7.0 | 6 | Low |
| 1hp8* | 68 | 43 | 0 | - | 3 | 1 | 4.7 | 5.0 | 1 | - |
| 2hp8 | 56 | 38 | 0 | 6 | 2 | 2 | 4.0 | 4.5 | 2 | - |
| 1hsn | 62 | 43 | 0 | - | 78 | 1 | 5.7 | 3.0 | 5 | 5 |
| 1hyp* | 75 | 43 | 0 | - | - | 25 | 6.2 | 6.1 | 4 | - |
| 1jvr | 74 | 38 | 0 | - | 20 | 7 | 4.6 | 6.5 | - | - |
| 1kjs* | 74 | 45 | 0 | 1 | 1 | 1 | 3.3 | 4.5 | 1 | 1 |
| 1lfb | 69 | 39 | 0 | - | 138 | 7 | 4.7 | 6.5 | 3 | 7 |
| 1mzm | 71 | 46 | 0 | 1 | 1 | 1 | 3.3 | 5.3 | 1 | 2 |
| 1mzm* | 93 | 57 | 0 | - | - | 1 | 4.9 | 7.0 | 1 | 3 |
| 1ner* | 74 | 36 | 0 | - | 60 | 25 | 3.9 | 4.0 | Low | Low |
| 1ngr* | 85 | 52 | 0 | - | 3 | 3 | 4.1 | 7.0 | 3 | 3 |
| 1nkl | 70 | 55 | 0 | 1 | 1 | 1 | 3.4 | 3.1 | 1 | Low |
| 1nkl* | 78 | 57 | 0 | 15 | 15 | 15 | 3.6 | 5.0 | 2 | 7 |
| 1nre | 66 | 53 | 0 | 69 | 5 | 3 | 3.5 | 5.7 | 5 | 5 |
| 1nre* | 81 | 55 | 0 | - | 16 | 5 | 5.5 | 7.0 | 6 | 8 |
| 2pac | 77 | 18 | 0 | - | - | 1 | 5.2 | 7.0 | Low | Low |
| 1pou | 70 | 48 | 0 | 28 | 28 | 28 | 2.9 | 6.0 | 1 | Low |
| 1r69 | 61 | 39 | 0 | 1 | 1 | 1 | 2.4 | 2.9 | 1 | 1 |
| 1rpo* | 61 | 55 | 0 | 92 | 40 | 27 | 4.6 | 4.0 | 4 | 4 |
| 1utg | 62 | 49 | 0 | 33 | 10 | 10 | 4.6 | 3.6 | 2 | - |
| Medium β | | | | | | | | | | |
| 3ait* | 74 | 0 | 35 | - | - | - | 8.3 | 6.7 | 9 | - |
| 1aiw* | 62 | 0 | 19 | - | - | - | 7.2 | 7.0 | - | - |
| 1ark | 55 | 0 | 12 | - | 49 | 7 | 5.0 | 4.6 | 7 | - |
| 1bdo | 75 | 0 | 32 | - | - | - | 8.1 | 7.0 | - | - |
| 1bq9* | 53 | 0 | 10 | - | 72 | 2 | 4.1 | 5.7 | Low | Low |
| 2cdx | 54 | 0 | 15 | - | - | - | 6.9 | 6.5 | - | - |
| 1csp | 64 | 0 | 34 | 8 | 8 | 2 | 4.6 | 6.5 | 8 | 8 |
| 1fbr | 93 | 0 | 26 | - | - | - | 9.5 | 7.0 | - | - |
| 1gvp | 82 | 0 | 37 | - | - | - | 7.6 | 7.0 | - | - |
| 1iyv* | 79 | 0 | 32 | - | - | - | 6.4 | 7.0 | 1 | 1 |
| 1kde* | 65 | 0 | 8 | - | - | - | 7.6 | 6.2 | - | - |
| 1ksr | 92 | 0 | 35 | - | - | - | 8.6 | 7.0 | 1 | 1 |
| 1msi | 60 | 4 | 8 | - | - | 28 | 6.5 | 6.2 | - | - |
| 2ncm | 96 | 0 | 53 | - | - | - | 7.5 | 7.0 | 2 | 2 |
| 2ncm* | 99 | 0 | 55 | - | - | - | 9.8 | 7.0 | - | - |
| 1nxb | 53 | 0 | 22 | - | - | 4 | 4.5 | 6.2 | - | - |
| 1pse* | 69 | 0 | 21 | - | - | - | 6.9 | 7.0 | - | - |
| 1rip* | 81 | 0 | 4 | - | - | - | 7.7 | 7.0 | - | - |
| 1sro | 66 | 4 | 25 | 1 | 1 | 1 | 3.5 | 4.1 | 1 | 4 |
| 1tit | 85 | 0 | 32 | - | - | - | 7.7 | 7.0 | 7 | 7 |
| 1tit* | 89 | 0 | 32 | - | - | - | 5.4 | 7.0 | 2 | 2 |
| 1tul | 97 | 0 | 49 | - | - | - | 9.4 | 7.0 | - | - |
| 1vif* | 60 | 0 | 26 | - | - | 3 | 5.9 | 7.0 | - | - |
| 1who | 88 | 0 | 40 | - | - | - | 7.0 | 7.0 | 4 | 4 |
| 1wiu | 90 | 0 | 44 | - | - | - | 7.3 | 7.0 | 2 | 2 |
| 1wiu* | 93 | 0 | 48 | - | - | - | 8.5 | 7.0 | 3 | - |
| 1wkt* | 88 | 0 | 30 | - | - | - | 8.7 | 7.0 | - | - |
| Medium α/β | | | | | | | | | | |
| 1a68* | 87 | 31 | 18 | - | - | - | 7.0 | 7.0 | - | - |
| 1aa3 | 56 | 18 | 4 | 8 | 8 | 7 | 3.3 | 4.0 | 3 | 3 |
| 1aba* | 87 | 30 | 16 | - | - | - | 8.0 | 7.0 | - | - |
| 1ap0* | 52 | 7 | 22 | - | - | - | 6.4 | 5.2 | Low | - |
| 2acy | 92 | 24 | 38 | - | - | - | 6.7 | 7.0 | 1 | 1 |

| Name | $N_{res}$ | $N_{\alpha}$ | $N_{\beta}$ | <5 Å | <6 Å | <7 Å | Low rmsd | Cluster cutoff | DALI rank $z > 2$ | DALI rank $z > 4$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 2acy* | 98 | 24 | 41 | - | - | - | 8.3 | 7.0 | 1 | 2 |
| 1afi* | 72 | 21 | 21 | 9 | 9 | 9 | 3.3 | 4.5 | 2 | 2 |
| 1ag2 | 97 | 54 | 4 | - | - | - | 8.1 | 7.0 | - | - |
| 1ah9* | 71 | 3 | 28 | - | - | - | 5.3 | 7.0 | Low | Low |
| 1aoy* | 78 | 32 | 8 | - | - | 1 | 5.4 | 6.0 | 1 | 6 |
| 1bb8* | 71 | 11 | 14 | - | - | - | 8.1 | 7.0 | - | - |
| 2bby* | 69 | 35 | 4 | - | 1 | 1 | 4.3 | 5.0 | 2 | 4 |
| 1beg* | 98 | 59 | 4 | - | - | - | 8.0 | 7.0 | 2 | - |
| 1bor | 52 | 0 | 0 | - | - | 32 | 6.0 | 4.6 | - | - |
| 1btb | 89 | 37 | 16 | - | - | - | 7.9 | 7.0 | - | - |
| 1ctf | 67 | 35 | 13 | 2 | 1 | 1 | 3.4 | 3.7 | 1 | 2 |
| 1ctf* | 68 | 38 | 18 | 4 | 4 | 4 | 4.1 | 3.5 | 4 | 8 |
| 1dol* | 62 | 13 | 17 | - | - | - | 5.6 | 5.0 | 1 | 1 |
| 1dvc* | 98 | 13 | 26 | - | - | - | 7.4 | 7.0 | 1 | 1 |
| 2fdn | 55 | 4 | 10 | - | - | 1 | 4.6 | 6.5 | - | - |
| 2fmr* | 65 | 18 | 18 | 2 | 2 | 2 | 4.1 | 5.5 | 2 | 2 |
| 2fow | 66 | 30 | 6 | 6 | 2 | 1 | 3.3 | 4.3 | 4 | Low |
| 2fow* | 68 | 31 | 8 | 2 | 1 | 1 | 3.9 | 4.3 | 1 | 2 |
| 1fwp | 66 | 21 | 14 | - | - | - | 6.4 | 5.8 | Low | - |
| 2fxb* | 81 | 16 | 14 | - | - | - | 5.8 | 7.0 | 4 | 4 |
| 1gb1 | 54 | 13 | 16 | 2 | 1 | 1 | 1.9 | 3.6 | 2 | 7 |
| 1hqi* | 90 | 30 | 16 | - | - | - | 6.9 | 7.0 | Low | - |
| 5icb | 72 | 41 | 4 | 32 | 1 | 1 | 3.9 | 4.7 | 6 | - |
| 2ife* | 91 | 28 | 23 | - | 5 | 2 | 4.7 | 7.0 | 1 | 1 |
| 1lea* | 72 | 39 | 6 | - | 89 | 89 | 4.6 | 4.0 | 10 | Low |
| 1leb | 63 | 37 | 4 | 41 | 4 | 4 | 3.4 | 4.1 | 4 | 4 |
| 1orc | 56 | 24 | 16 | 15 | 3 | 1 | 3.4 | 3.2 | - | - |
| 2orc* | 64 | 23 | 8 | 50 | 3 | 3 | 4.1 | 3.8 | Low | Low |
| 1pgx | 57 | 14 | 26 | 3 | 2 | 1 | 2.7 | 2.6 | 2 | 2 |
| 1pgx* | 56 | 14 | 28 | 13 | 1 | 1 | 3.1 | 2.8 | 1 | 1 |
| 2pni* | 86 | 3 | 24 | - | - | - | 9.0 | 7.0 | 1 | - |
| 5pti | 55 | 8 | 14 | - | 66 | 2 | 5.0 | 6.0 | - | - |
| 2ptl | 60 | 12 | 22 | 1 | 1 | 1 | 3.0 | 3.5 | 1 | 1 |
| 2ptl* | 60 | 12 | 22 | 9 | 9 | 9 | 2.4 | 4.0 | 3 | 10 |
| 1ris | 92 | 28 | 45 | - | - | 5 | 6.2 | 7.0 | 3 | 7 |
| 1sap* | 66 | 20 | 27 | - | - | 11 | 6.1 | 4.5 | - | - |
| 1sro* | 76 | 4 | 27 | - | 30 | 2 | 4.6 | 6.5 | 2 | 2 |
| 1stu* | 68 | 26 | 20 | - | 42 | 42 | 5.4 | 6.7 | - | - |
| 1svq | 90 | 25 | 24 | - | 19 | 19 | 5.8 | 7.0 | - | - |
| 1tif* | 59 | 13 | 22 | 1 | 1 | 1 | 3.6 | 3.2 | Low | Low |
| 1tnt* | 65 | 24 | 4 | - | - | 51 | 5.2 | 5.7 | - | - |
| 1tsg* | 98 | 10 | 4 | - | - | - | 9.2 | 7.0 | - | - |
| 1tuc* | 61 | 3 | 25 | - | - | 19 | 5.4 | 5.0 | 8 | 8 |
| 2u1a* | 76 | 22 | 18 | - | - | - | 4.8 | 5.0 | 2 | 2 |
| 4ull* | 69 | 11 | 21 | - | - | - | 6.0 | 6.5 | - | - |
| 1vcc* | 77 | 11 | 25 | - | - | 19 | 5.9 | 7.0 | - | - |
| 1vig* | 71 | 25 | 17 | - | - | - | 5.8 | 6.0 | 10 | 10 |
| 1vqh* | 86 | 6 | 40 | - | - | - | 9.5 | 7.0 | - | - |
| Large α | | | | | | | | | | |
| 1aa2 | 105 | 57 | 0 | - | - | - | 8.2 | 7.0 | 2 | - |
| 1eca | 132 | 94 | 0 | - | - | - | 7.2 | 7.0 | 1 | - |
| 2fha | 160 | 123 | 0 | - | - | - | 8.5 | 7.0 | 1 | 1 |
| 2gdm | 149 | 106 | 0 | - | - | - | 6.6 | 7.0 | 1 | 2 |
| 1hlb | 138 | 101 | 0 | - | - | 7 | 5.9 | 7.0 | 7 | 7 |
| 2lfb* | 100 | 60 | 0 | - | - | - | 8.6 | 7.0 | 3 | 6 |
| 1lis | 111 | 86 | 0 | - | - | 44 | 4.8 | 7.0 | 1 | 4 |
| 1mbd | 147 | 112 | 0 | - | - | - | 7.4 | 7.0 | 1 | Low |
| 1pal | 100 | 52 | 4 | - | - | - | 7.9 | 7.0 | - | - |
| 1vls | 143 | 112 | 0 | - | - | - | 7.2 | 7.0 | 1 | 2 |
| Large β | | | | | | | | | | |
| 4fgf | 121 | 0 | 47 | - | - | - | 10.6 | 10.0 | - | - |
| 1ksr* | 100 | 0 | 39 | - | - | - | 8.8 | 7.0 | - | - |
| Large α/β | | | | | | | | | | |
| 1acf | 123 | 40 | 41 | - | - | - | 6.9 | 7.0 | 1 | 10 |
| 1erv | 105 | 43 | 28 | - | - | 5 | 6.0 | 7.0 | 1 | 5 |
| 1kte | 100 | 48 | 18 | - | 171 | 171 | 5.2 | 7.0 | Low | Low |
| 1lz1 | 116 | 39 | 10 | - | - | - | 8.5 | 3.0 | 3 | 3 |
| 1pdo | 121 | 59 | 18 | - | - | 3 | 6.2 | 7.0 | 1 | 1 |

The proteins are divided into categories based on size and secondary structure content; proteins with an asterisk are from the PDB select list of proteins, those without, from the Friesner set (see Methods). $N_{res}$ is the number of residues in the sequence. $N_{\alpha}$ and $N_{\beta}$ are the number of α-helical and β-strand residues as assigned in the native structure by DSSP (Kabsch & Sander, 1983). The ranks of the cluster centers are shown for three cutoffs, <5, <6 and <7 Å rmsd to native. The low rmsd is the lowest rmsd structure in the set of structures. The clustering cutoff is described in Methods. The rank of the top cluster found by DALI related to the native structure is indicated in the last two columns (using a DALI Z-score cutoff of 2 and 4 as indicated). Low, indicates that none of the top ten clusters were similar to the native structure, but the lowest rmsd structure in the set had a DALI Z-score to the native structure greater than the cutoff.

**Table 2.** Summary of predictions

| Protein class | Total | <5 Å top five | <5 Å set | <7 Å top five | <7 Å set | DALI hit top five | DALI hit top ten |
|---|---|---|---|---|---|---|---|
| All | 172 | 32 | 71 | 73 | 130 | 52 | 69 |
| α | 65 | 18 | 40 | 30 | 56 | 25 | 32 |
| β | 36 | 2 | 11 | 13 | 19 | 7 | 9 |
| α/β | 71 | 12 | 30 | 30 | 55 | 20 | 28 |
| Small all | 30 | 14 | 24 | 25 | 30 | 4 | 4 |
| Small α | 10 | 8 | 10 | 10 | 10 | 4 | 4 |
| Small β | 8 | 1 | 7 | 8 | 8 | 0 | 0 |
| Small α/β | 12 | 5 | 7 | 7 | 12 | 0 | 0 |
| Medium all | 127 | 18 | 56 | 46 | 93 | 41 | 56 |
| Medium α | 46 | 10 | 29 | 20 | 43 | 17 | 23 |
| Medium β | 27 | 1 | 4 | 5 | 11 | 7 | 9 |
| Medium α/β | 54 | 7 | 23 | 21 | 39 | 17 | 24 |
| Large all | 15 | 0 | 1 | 2 | 7 | 7 | 9 |
| Large α | 9 | 0 | 1 | 0 | 3 | 4 | 5 |
| Large β | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Large α/β | 5 | 0 | 0 | 2 | 4 | 3 | 4 |

The number of proteins for which one of the top five clusters was within 5 Å of the native structure is reported in the <5 Å top five column, while the number of proteins for which one of the proteins in the set of generated structures was within 5 Å rmsd of the native structure is reported in the <5 Å set column. The following two columns report the same statistics for a 7 Å cutoff. The last two columns indicate whether one of the top five or ten cluster centers was found by DALI to be similar to the native structure.

similarity, a significant piece of the structure is related to the true structure, more so than to the vast majority of unrelated proteins in the structural database. The DALI Z-score was found to be a modest indicator of the prediction quality: the more similar a cluster center is to a structure in the PDB, the more likely the structure is similar to the native structure (data not shown).

To assess whether the tertiary structure predictions produced by ROSETTA provide information beyond that contained in the predicted secondary structure used in the fragment library generation, we sent the secondary structure predictions for ten large proteins and ten medium proteins to the FORESST server (Di Francesco *et al.*, 1999). FORESST uses the predicted secondary structure string to predict the fold. For six of the ten large proteins FORESST correctly assigns the native fold but is not as successful on the smaller proteins (three of ten medium proteins are correctly assigned). The low resolution 3D information in the ROSETTA predictions may be complementary to the 1D information utilized by many current fold recognition methods, as FORESST and ROSETTA succeed on different proteins in both size ranges.

Beyond structural similarity identified by DALI, we were interested if the matches between the predicted structures and the most related structure in the PDB could be used to predict the function of the sequence. To this end, for the proteins where the top DALI match correctly identified the fold of the protein (34 cases), the function of the protein being folded and the top DALI match without significant sequence similarity were compared. In a number of cases there are interesting similarities in function in the absence of significant sequence similarity (BLAST *e*-value <10). These functional similarities cover a gamut of protein function. Some proteins are nucleotide binding: 1r69 (434 repressor) found structurally similar to 1qpz (purine nucleotide synthesis repressor), 1bw6 (centromere DNA binding) to 1mbe (*c*-myb DNA binding), and 1sro (polyribonucleotide nucleotidyl transferase) to 1a0i (DNA ligase). Similarities were found both in enzymatic activity (1iyv (dihydrolipamide acetyltransferase) matched with 1htp (decarboxylase with a lipoamide arm moeity)) and protein-protein interactions: 1dvc (proteinase inhibitor) to 1ugi (glycosylase inhibitor), the IgG binding proteins 1pgx (protein G) matched to 2ptl (protein L), 2ptl (protein L) to 2igd (protein G), and 2pni (PI3 kinase, SH3 domain) to 1aoj (Eps8, SH3 domain). These results suggest that *ab initio* structure prediction followed by a global structure comparison based search of the PDB can give insights into protein function; this approach is complementary to methods which focus on matching active site templates (Skolnick & Fetrow, 2000; Wallace *et al.*, 1996).

The results presented here are a dramatic improvement over our results of several years ago in which good models were produced only for small all helical proteins (Simons *et al.*, 1997). With continued improvements, the method should be able to make an important contribution to the interpretation of genome sequence information. We envision generating low resolution models for all globular protein domains of less than 150 amino acid residues, and using the DALI-based approach described here to identify structurally related known proteins where they exist. Such an approach has the advantage over more traditional fold recognition methods that models are produced even in cases where there is not a related structure already in the PDB, and in cases where there is a

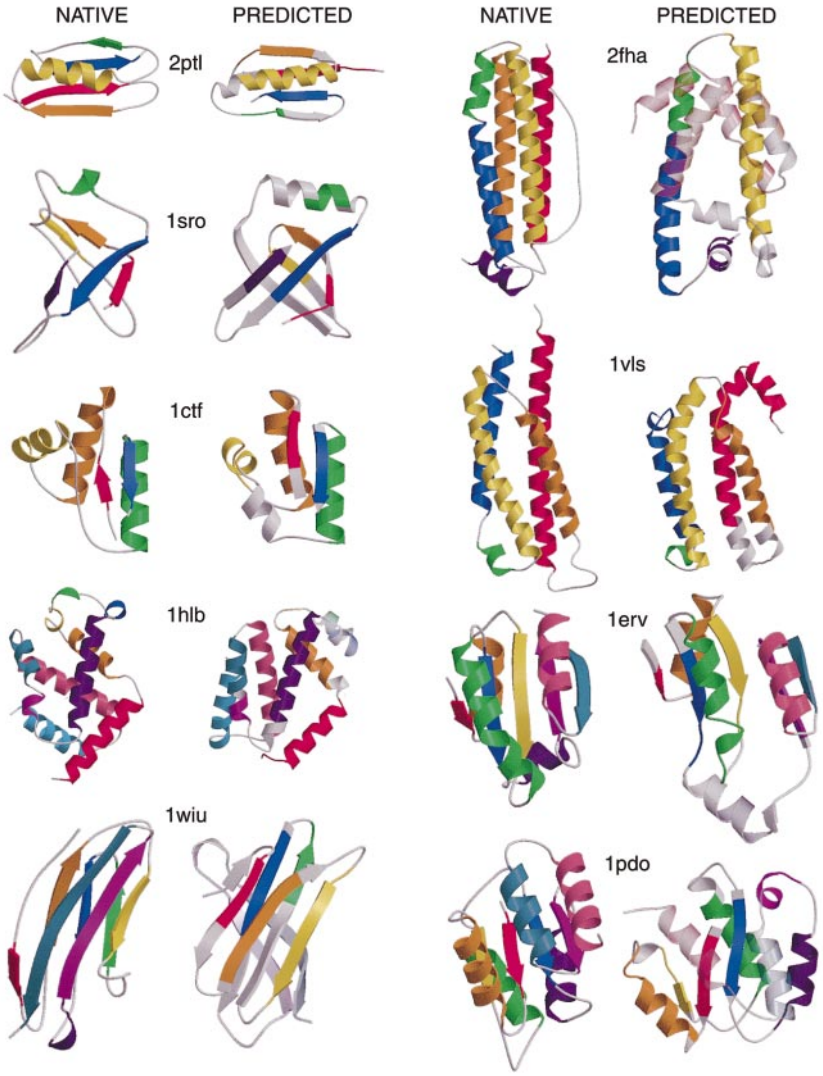


**Figure 3.** Comparison of native and predicted structures. The left column depicts the native structure and the right column is the best cluster center as identified in Table 1. The coloring of the secondary structural elements is demarcated by the native secondary structure assignment. Beginning with the N terminus, the coloring scheme is red, orange, yellow, green, blue, indigo, violet, turquoise and cyan. Images were prepared using Molscript (Kraulis, 1991) and Raster3d (Merritt & Bacon, 1997).

distantly related structure in the PDB, the identification of the structure does not depend on a low energy threading of the query sequence through the structure (which may not exist because of differences in helix orientation, β-strand/sheet twist, etc., between the true structure of the query sequence and the structurally related protein in the PDB). The results presented here suggest that this method is almost ready to be used on the genome scale. However, there is also quite clearly considerable room for improvement, particularly on the β-sheet containing proteins, and the large-scale tests reported here should provide a useful benchmark/standard for evaluating future improvements in *ab initio* structure prediction methodology.

## Methods

A total of 1000 structures were generated for each sequence in two large sets of sequences of proteins of known structure. The first list of proteins was compiled by selecting proteins greater than 50 and less than 100 amino acid residues with no chain identifier and less than 30% sequence identity from the PDB culled list of proteins, resulting in a list of 77 proteins (http://chaos.fccc.edu/research/labs/dunbrack/culledpdb.html; Hobohm *et al.*, 1992). The set from the Friesner laboratory was chosen to cover proteins from all secondary structure classes and sizes up to 150 residues (Eyrich *et al.*, 1999). Each sequence was folded in a pseudo-blind manner: the native and homologous sequences of known structure were eliminated from the nearest neighbor sets. The contribution of the native structure to the scoring function is very minimal; in test cases where the native structure and sequence homologs are removed from all components of the scoring function, the results were unchanged (Simons *et al.*, 1999a,b; results not shown). The simulation method was similar to that used in our earlier work (Simons *et al.*, 1997, 1999a,b) except for two improvements.

First, information from three different secondary structure prediction methods: PHD (Rost *et al.*, 1994), DSC (King *et al.*, 1997), and PSI-PRED (Jones *et al.*, 1999), is used in the fragment picking process to reduce the sensitivity to errors in any one of the methods. Six neighbors

were chosen with each of the three secondary structure predictions and seven nearest neighbors were chosen with sequence profiling alone using the following equation:

$$DISTANCE = \sum_{i}^{9}\sum_{aa}^{20} |S(aa,i) - X(aa,i)| + w^{*}\sum_{i}^{9}\sum_{s}^{3} SSconf(i,s)^{*}\delta[SS(i,s), XX(i)]$$

where $S(aa,i)$ and $X(aa,i)$ are the frequencies of amino acid $aa$ at position $i$ in nine residue segments of multipe sequence alignments for either the sequence being folded ($S$) or of one of the proteins in the pdb_select_25 set($X$). The second part of the equation uses the secondary structure prediction $SS$ ($s$ is helical, strand or other) and the confidence of the prediction ($SSconf$). The delta function is one if the secondary structure of the known fragment, $XX(i)$, is the same as the prediction, $SS(i,s)$, or zero if otherwise. The weight, $w$, normalizes the sequence profiling score with the secondary structure matching score and was optimized for local structure prediction ($w = 0.05$). The combination of sequence profiling and secondary structure matching has been used in prior work for protein fold recognition (Fischer & Eisenberg, 1996).

Second, ''smooth'' moves which reduce the overall perturbation to the structure caused by a fragment insertion are used late in the simulation to increase the acceptance rate and thus improve minimization of the scoring function. Following the approach of Gunn (1997), the relative orientation of coordinate systems embedded in the first and last residues of each fragment was computed and stored at the beginning of the simulation, and fragments producing relatively small perturbations to the overall structure were selected based on the similarity of the relative orientation of these coordinate systems to those in the segment of the current conformation being replaced.

To generate 1000 structures for a 100-residue protein required 12 hours of computer time on a pentium III 450 MHz Personal Computer. Each simulation consisted of 20,000 attempted insertions of nine-residue fragments, 4000 attempted insertions of three-residue fragments, and 8000 attempted insertions of ''smooth'' three-residue fragments. The score being minimized includes terms for hydrophobic burial ($P_{density}$, $P_{env}$), polar side-chain interactions ($P_{pair}$), hydrogen bonding between β-strands ($P_{HS\text{-}dist}$, $P_{HS\text{-}\phi\theta}$, $P_{SS\text{-}dist}$, $P_{SS\text{-}\phi\theta}$, $P_{hb}$, $P_{sheet}$) and hard sphere repulsion ($VdW$) as described by Simons *et al*. (1999a,b). The bins of the $P_{XX\text{-}\phi\theta}$ functions were every 10° for φ and every 5° for θ totaling 1296 bins. We increased the number of bins from earlier work (Simons *et al*., 1999b) because the structure generation procedure could minimize this function very well and there is enough data for finer binning. The set of structures were clustered on global $C^{\alpha}$ rmsd (Shortle *et al*., 1998) with an rmsd cutoff chosen such that 80 to 100 structures were in the largest cluster. The cluster ''center'' was taken to be the structure in the cluster with the best score.

The lists of PDB codes of the lists of sequences folded and additional details on the models are available upon request. All structures and cluster center identities are available in tarred and compressed format and can be requested from the authors.

## Acknowledgments

We are indebted to the system administration skills of Keith Ladig, Qian Yi, and Jerry Tsai and helpful discussion with Tanya Kortemme, Brian Kuhlman, Jerry Tsai, and Ingo Ruczinski. We thank Liisa Holm for the DALI program and Phillip G. McQueen for assistance with the FORREST server. This work was supported by a Molecular and Cellular Biology training grant from the Public Health Service, NSRA T32 GM07270 (K.T.S.) and by young investigator awards (D.B.) from the NSF and the Packard foundation.

## References

Di Francesco, V., Munson, P. J. & Garnier, J. (1999). FORESST: fold recognition from secondary structure predictions of proteins. *Bioinformatics,* **2**, 131-40.

Eyrich, V. A., Standley, D. M. & Friesner, R. A. (1999). Prediction of protein tertiary structure to low resolution: performance for a large and structurally diverse test set. *J. Mol. Biol.* **288**, 725-742.

Fischer, D. & Eisenberg, D. (1996). Protein fold recognition using sequence-derived predictions. *Protein Sci.* **5**, 947-955.

Gunn, J. R. (1997). Sampling protein conformations using segment libraries and a genetic algorithm. *J. Chem. Phys.* **106**, 4270-4281.

Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409-417.

Holm, L. & Sander, C. (1995). DALI: a network tool for protein structure comparison. *Trends Biochem. Sci.* **20**, 478-480.

Jones, D. T., Tress, M., Bryson, K. & Hadley, C. (1999). Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. *Proteins: Struct. Funct. Genet.* **37**, 104-111.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* **22**, 2577-2637.

King, R. D., Saqi, M., Sayle, R. & Sternberg, M. J. (1997). DSC: public domain protein secondary structure prediction. *Comput. Appl. Biosci.* **13**, 473-474.

Kraulis, P. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946-950.

Merritt, E. A. & Bacon, D. J. (1997). Raster3D: photorealistic molecular graphics. *Methods Enzymol.* **277**, 505-524.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machine. *J. Chem. Phys.* **21**, 1087-1092.

Moult, J., Hubbard, T., Fidelis, K. & Pedersen, J. T. (1999). Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins: Struct. Funct. Genet.* **37 Suppl. 3**, 2-6.

Orengo, C. A., Bray, J. E., Hubbard, T., LoConte, L. & Sillitoe, I. (1999). Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. *Proteins: Struct. Funct. Genet.* **37 Suppl. 3**, 149-170.

Rost, B., Sander, C. & Schneider, R. (1994). PHD-an automatic mail server for protein secondary structure prediction. *Comput. Appl. Biosci.* **10**, 53-60.

Shortle, D., Simons, K. T. & Baker, D. (1998). Clustering of low-energy conformations near the native structures of small proteins. *Proc. Natl Acad. Sci. USA,* **95**, 11158-11162.

Skolnick, J. & Fetrow, J. S. (2000). From genes to protein structure and function: novel applications of computational approaches in the genomic era. *Tibtech,* **18**, 34-39.

Simons, K. T., 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.

Simons, K. T., Bonneau, R., Ruczinski, I. & Baker, D. (1999a). *ab initio* protein structure prediction of CASP III targets using ROSETTA. *Proteins: Struct. Funct. Genet.* **37 Suppl. 3**, 171-176.

Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C. & Baker, D. (1999b). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins: Struct. Funct. Genet.* **34**, 82-95.

Wallace, A. C., Laskowski, R. A. & Thornton, J. M. (1996). Deviation of 3D coordinate templates for searching structural databases: application of Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.* **5**, 1001-1013.

***Edited by B. Honig***

*(Received 3 October 2000; received in revised form 12 December 2000; accepted 12 December 2000)*