PROTEINS: Structure, Function, and Genetics, Suppl. 1:129–133 (1997)

Protein Folds From Pair Interactions: A Blind Test in Fold Recognition

Hannes Flöckner, Francisco S. Domingues, and Manfred J. Sippl*
Center for Applied Molecular Engineering, Institute for Chemistry and Biochemistry, University of Salzburg, Salzburg, Austria

ABSTRACT **We submitted nine predictions to CASP2 using our fold recognition program ProFIT. Two of these structures were still unsolved by the end of the experiment, six had a recognizable fold, and one fold was new. Four predictions of the six recognizable folds were correct. Two models were excellent in terms of alignment quality (T0031, T0004): in one the alignment was partially correct (T0014), and one fold was correctly identified (T0038). We discuss improvements of the program and analyze the prediction results. Proteins, Suppl. 1:129–133, 1997.** © 1998 Wiley-Liss, Inc.

Key words: knowledge-based potentials; energy functions; molecular modeling; prediction of protein structure; prediction evaluation

INTRODUCTION

In the first Critical Assessment of techniques for protein Structure Prediction (CASP1) fold recognition correctly predicted the folds of several targets.[1] In particular, we identified the histon H1 fold to be closely related to the structure of replication terminator protein.[2] Another successful prediction was the structure of ferredoxin, which is similar to the fold of the subtilisin propeptide.[2] In these and other cases the correct result scored highest among all possible folds, and in this sense they were clear successes.

However, the quality of the sequence–structure alignments, when compared to the optimum geometric superimposition, was poor. Therefore, the most important conclusion from CASP1 was perhaps the need to improve alignment quality.

For CASP2 we again employed ProFIT[2,3] for fold recognition. In this program potentials of mean force derived from a database of known structures are used in combination with dynamic programming techniques. We based our predictions exclusively on these potentials, deliberately neglecting multiple sequence alignments, secondary structure prediction, or any other information available on the target sequences. As demonstrated before[4] use of multiple sequence and secondary structure information definitely improves prediction results. However, using this information would not reveal the predictive power of mean force potentials in a clear manner. Since we were interested in the performance of methods based exclusively on pair interactions, we deliberately did not use additional information.

In our view, the goal of fold recognition is to provide a good model for a given target sequence. Obviously, the best model that can be obtained is the most similar structure in the fold database with the target sequence correctly aligned. A clear goal for critical assessment of techniques is to evaluate how far different methods are able to identify a related fold and align the target sequence correctly. This is straightforward when one model is submitted for each target. Consequently, we submitted one model or in some cases a group of related folds to CASP2.

Below we discuss the major changes and improvements of ProFIT since CASP1, followed by an analysis of the predicted models (6 targets). Then we discuss successes and failures as well as problems concerning evaluation of predictions.

METHODS

In summary, our goals for CASP2 were as follows:

1. Improve alignments.
2. Evaluate the predictive value of mean force potentials.
3. Submit a single model.

The version of ProFIT used for CASP2 is similar to previous releases.[2,3,5] New features are restrictions on insertions and deletions and an updated set of potentials.

A key to better alignments turned out to be proper gap control. It is well known that insertions and deletions in secondary structure elements and core regions of proteins are rare and often not tolerated. We incorporated these rules in our automated alignment procedure:

Contract grant sponsor: Fonds zur Förderung der wissenschafltichen Forschung, Austria; Contract grant numbers: P11205-MOB and P11601-GEN; Contract grant sponsor: Junta Nacional de Investigação Científica e Tecnológica (F.D.); Contract grant number: PRAXIS XXI/BD/4528/94.
*Correspondence to: Dr. Manfred J. Sippl, Center for Applied Molecular Engineering, Institute for Chemistry and Biochemistry University of Salzburg, Jakob-Haringerstr. 3, A-5020 Salzburg, Austria.
E-mail: sippl@came.sbg.ac.at
Received 6 May 1997; Accepted 26 August 1997

1. Gaps are prohibited inside secondary structure elements (but whole elements can be skipped or shortened).
2. A gap between adjacent residues i and $i + 1$ in the sequence is only accepted if the spatial distance is less than a threshold (7 Å).

The current state of our potentials has been reported recently.[6,7] Since CASP1 the database of folds increased considerably with a corresponding increase in the statistical reliability of the potentials. In additions this increases the chances of finding a related fold in CASP2.

We submitted predictions to CASP2 only when ProFIT results were conclusive. Scores[8,9] were used for ranking, but we did not regard the resulting list as an absolute quality measure for the models. The submitted model was selected by human decision from the top ranks based on a detailed analysis regarding alignment quality, specifically fragmentation of alignment and location and size of gaps. ProSUP[10] was used to search for common motifs among the top scoring models. The confidence in a prediction was considered high when two or more similar folds had high ranks.

We refrained from submission when this procedure gave no clear answers. This either indicates that no related fold is contained in the library or that the program was unable to detect such a fold.

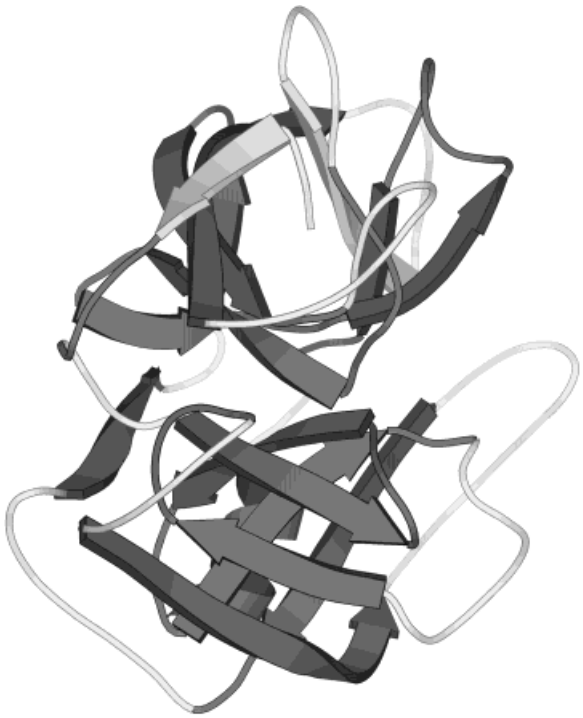

Fig. 1. Model of T0031. Those regions where sequence–structure alignment matches the geometrical equivalencies of model 1ppf-E and native fold of T0031 are dark gray.

RESULTS

We submitted predictions for nine targets. For two of these, the structures were not solved in time for assessment. One adopts a newly observed fold and the remaining six were reported to have known folds.

T0031: Exfoliative Toxin A

The predicted fold for T0031 (241 residues) was leukocyte elastase 1ppf-E (218 residues). The structure of T0031 is indeed very similar to 1ppf-E, and superposition yields a root-mean-square (RMS) of 1.6 Å for 166 geometrical equivalent residues (Fig. 1). The sequence–structure alignment obtained from fold recognition matches to a large extent the structure alignment of 1ppf-E and the structure of T0031; 135 residue pairs are aligned identically (Fig. 2).

Recently a relationship of exfoliative toxin A and the serine proteinases was suggested,[11] and a model based on multiple sequence alignment was proposed.[12] Surprisingly, this was a very good model based solely on sequence information where 105 residues match the topologically equivalent residues. Nevertheless, the ProFIT result based on single-sequence information and pair interactions only, is even more accurate (135 correct aligned residues). According to the assessment, this is also the best alignment among all the predictions in CASP2. In addition, from all the serine proteinases contained in the fold library used, 1ppf-E is the most similar to T0031.

T0004: Polyribonucleotide Nucleotidyltransferase—S1 Motif

For the S1 motif of polyribonucleotide nucleotidyltransferase (75 residues), we submitted three very similar models, two based on the major cold shock proteins 1mjc (69 residues) and 1csp (67), and another one based on verotoxin-1 1bov-A (69). All turned out to be very similar to the experimental fold of T004 (Fig. 3). Deviations are restricted to loop regions and the orientation of the C-terminal helix. The sequence–structure alignments of these folds differ slightly, but nevertheless they are in good agreement with the geometric equivalencies, for example, in the case of 1mjc 41 of a total of 52 geometrical equivalent residues were placed correctly (Fig. 4).

The target sequence and the major cold shock proteins do have sequence similarity, but only at a very limited level[13] and T004 may be regarded as a rather simple test case for fold recognition.

T0014: 3-Dehydroquinase

The result for T0014, 3-dehydroquinase, is a TIM barrel fold. We submitted two models of this type, pyruvate kinase 1pky-C and triosephophate isomerase 1mss-A. The structure of T0014 is very similar to

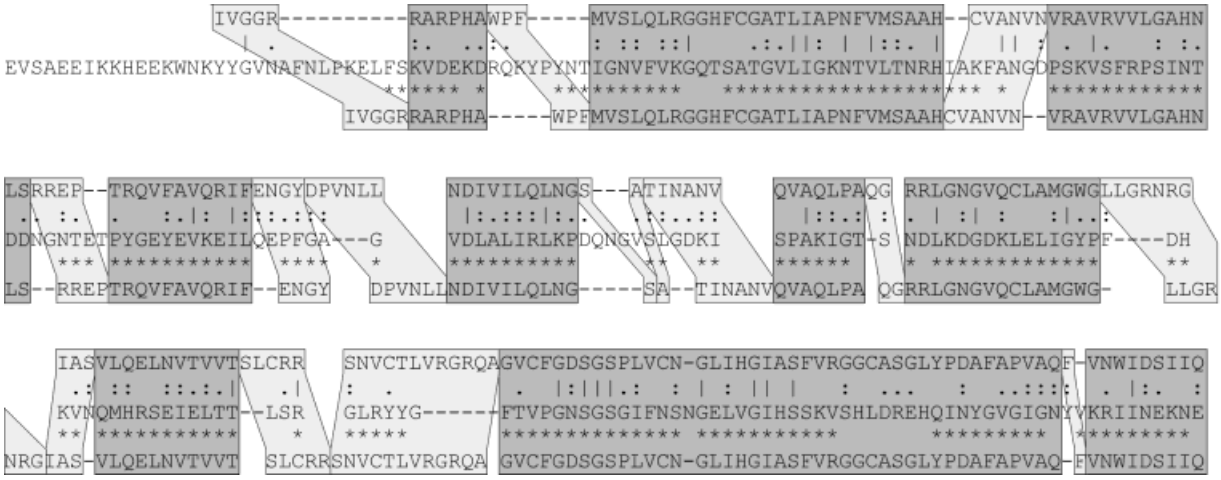

Fig. 2. Predicted versus observed alignment for T0031 and 1ppf-E. The target sequence is positioned in the middle. Dark gray rectangles indicate residues aligned identically in the sequence–structure (line above) and in the ProSUP alignment (line below). Geometrical equivalent residues are marked with an asterisk.

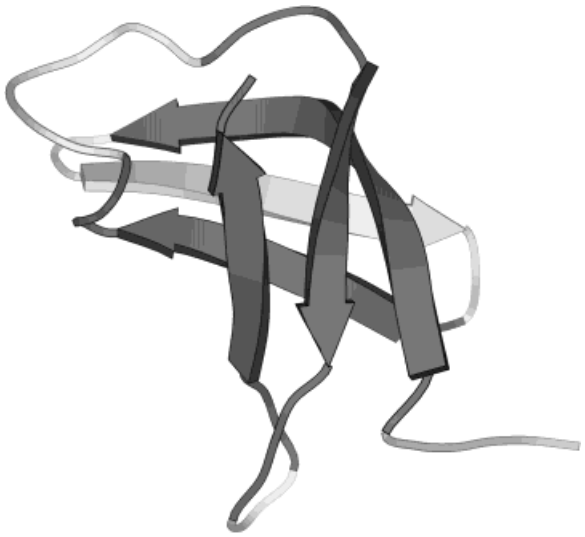

Fig. 3. Native fold of target T0004. In the sequence structure alignment four out of five strands are aligned correctly (dark gray).

these proteins where structure comparison yields 143 equivalent residues superimposable to 1,9 Å in the case of 1pky-C and 123 equivalent residues at an RMS of 2.0 Å for 1mss-A.

In both models the alignment was far from optimum, but still some regions coincide with the geometric equivalencies (22 residues of 142 are correctly aligned). TIM barrel folds consist of repetitive alpha/beta units with several loops of variable length. Periodic motifs are difficult to align correctly which may explain the rather poor alignment quality. The automatic alignment procedure accurately recognized two large insertions in 1pky-C (Fig. 5).

T0038: Fructose-1,6-Bisphosphatase—CBDN1

CBDN1 was predicted to have a structure similar to transthyretin (1roy-A) and azurin (1arn). Both are beta sandwich proteins having similarities to the experimental structure of T0038. For example 1roy-A superimposes on T0038 with an RMS error of 2.4Å for 52 equivalent residues. The equivalent regions correspond to five strands (out of nine) of T0038 (Fig. 6). The remaining four strands do have topological equivalencies in 1roy-A but of different connectivity. Similar results are obtained for 1arn.

However, the fold database contains structures that are more similar to T0038. For these structures, a proper alignment requires many large gaps. ProFIT failed to produce proper alignments mainly because our gap restrictions were too stringent for this situation. Nevertheless, the submitted models have a clear structural relationship, but the assessors of CASP2 considered this to be a wrong prediction.

T0022: L-Fucose Isomerase

T0022 is a multidomain protein of 591 residues. Only one domain of 109 residues is similar to a fold in the database. Since our alignment technique is global, that is, it tries to align a complete sequence with a complete structure, chances to be successful in such cases are very small.

We predicted a TIM barrel, and when we analyzed the results we were surprised to find 61 structural equivalent residues superimposable between the two structures (RMS of 1.9 Å). The best match would have been 2liv, which has 109 equivalent residues and an RMS of 2.0 Å.

T0002: Threonine Deaminase

The result we submitted for T0002 is embarrassing. By human failure a wrong prediction was sent

Fig. 4. Comparison of predicted versus ProSUP alignment of T0004 and the predicted fold 1mjc. In the sequence–structure alignment 41 residues match the 52 geometrically equivalent residues (dark gray) of model and native fold.

Fig. 5. 1pky-C, the proposed model for T0014. In the predicted alignments two large gaps (~60 and 130 residues) were correctly placed to exclude two additional domains (displayed in light gray), which are missing in the native fold T0014.

for the wrong target domain. As we did not submit for the fold recognition domain the prediction was not considered by the assessors.

DISCUSSION

We summarize the results: Alignments of high quality were obtained for T0031 and T004. A correct fold was assigned to T0014 with a partially correct alignment. A similar fold to T0038 was predicted, but there are more closely related folds in the database. To find the closest possible match was too difficult for T0022, but the fold assigned by ProFIT has some structural similarity. T0002 was a (human) failure. We did not submit for T0020 (recognizable, but not detected) and submitted a model for T0030 (not recognizable, hence the model is wrong). As a corollary, in all cases but one, results were obtained only when there was a recognizable fold in the database. When we refrained from submission, the fold was in most cases not recognizable ("none prediction"), although we did not state this explicitly.

In our predictions human intervention is restricted to the selection of the resulting models. The threading procedure itself is solely based on mean force potentials neglecting all available additional information. It is now interesting how combination with multiple sequence and secondary structure information will affect the quality of the predictions.

It is clear that the alignments have improved drastically compared to the CASP1 results, approaching the quality of superimposition of structures where the full geometric information of both folds is required.

Three targets (T0002, T0004, T0031) were rather easy, two perhaps intermediate (T0014, T0038) and two very difficult (T0020, T0022). Seemingly a large number of groups got the easy ones but with varying degree of alignment quality. We do not attempt to compare our results with those of other groups, but there are several comments we have on the evaluation problem.

As already mentioned in the introduction ideally a prediction should consist of one model. In this case evaluation is easy and straightforward. On the other hand in CASP, results are usually submitted as a ranked list of possible models, but it is obvious that a list of models is not a clear prediction. Taken to the extreme one could submit a range of folds each with a non vanishing weight. In the evaluation such a prediction gets at least some positive score (where all competitors get nothing), but such a result is of little practical value.

In our submission we adopted the very stringent rule to submit only one fold or a group of related structures. Hence the result is either correct or incorrect. If a related fold would have been found for example on third position it would have been disregarded.

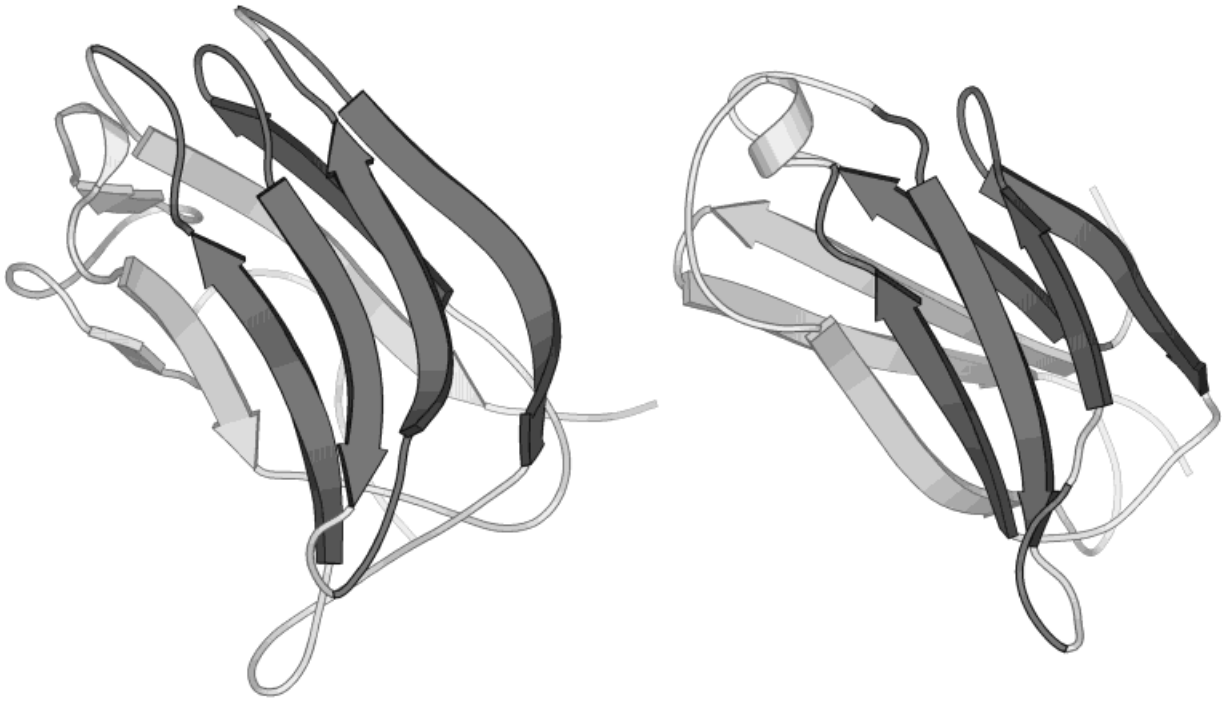

Fig. 6. **Left:** Target T0038. **Right:** Predicted model 1roy-A. The five superimposable strands are shown in dark gray. The remaining strands have some topological equivalents, but the connectivity differs.

Evaluation and assessment of threading predictions is a difficult and controversial procedure. Noticeably CASP2, the rules for evaluation were not defined while submissions were accepted. However, it would be desirable for clear rules to be known by all participants before submission. In spite of these difficulties, the assessors and organizers managed to make CASP2 a successful experiment.

ACKNOWLEDGMENT

Figures for molecular structures were prepared using the program MOLSCRIPT.[14] ProFIT, ProSUP, and other software are available from http://www.came.sbg.ac.at/

REFERENCES

1. Lemer, C.M-R., Rooman, M.J., Wodak, S.J. Protein structure prediction by threading methods: Evaluation of current techniques. Proteins 23:337–355, 1995.
2. Flöckner, H., Braxenthaler, M., Lackner, P., Jaritz, M., Ortner, M., Sippl, M.J. Progress in fold recognition. Proteins 23:376–386, 1995.
3. Sippl, M.J., Flöckner, H. Threading thrills and threats. Structure 4:15–19, 1996.
4. Fischer, D., Eisenberg D. Protein fold recognition using sequence-derived predictions. Protein Sci. 5:947–955, 1996.
5. Sippl, M.J. Boltzman's principle, knowledge based mean force potentials and protein folding: An approach to the computational determination of protein structure. J. Comput. Aided Mol. Design 7:473–501, 1993.
6. Sippl, M.J., Ortner, M., Jaritz, M., Lackner, P., Flöckner, H. Helmholtz free energies of atom pair interactions in proteins. Folding Design. 1:289–298, 1996.
7. Sippl, M.J. Helmholtz free energy of peptide hydrogen bonds in proteins. J. Mol. Biol. 260:644–648, 1996.
8. Sippl, M.J., Jaritz, M. The predictive power of mean force pair potentials. In: "Protein Structure by Distance Analysis." Bohr, H., Brunak, S. (eds.). Amsterdam: IOS Press, 113–134, 1994.
9. Sippl, M.J. Recognition of errors in three-dimensional structures of proteins. Proteins 17:355–362, 1993.
10. Feng, Z.-K., Sippl, M.J. Optimum superimposition of protein structures, ambiguities and implications. Folding Design 1:123–132, 1996.
11. Dancer, S.J., Garrat, R., Saldanha, J., Jhoti, H., Evans, R. The epidermolytic toxins are serine proteases. FEBS Lett. 268:129–132, 1990.
12. R.G.Barbosa, J.A., Saldanham J.W., Garrat, R.C. Novel features of serine protease active site and specificity pockets: Sequence analysis and modelling studies of glutamate-specific endopeptidases and epidermolytic toxins. Protein Eng. 9:591–601, 1996.
13. Sander, C., Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–68, 1991.
14. Kraulis, P.J. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. 24:946–950, 1991.