PROTEINS: Structure, Function, and Genetics 23:376–386 (1995)

Progress in Fold Recognition

Hannes Flöckner, Michael Braxenthaler, Peter Lackner, Markus Jaritz, Maria Ortner, and Manfred J. Sippl

Center for Applied Molecular Engineering, Institute for Chemistry and Biochemistry, University of Salzburg, A-5020 Salzburg, Austria

ABSTRACT **The prediction experiment reveals that fold recognition has become a powerful tool in structural biology. We applied our fold recognition technique to 13 target sequences. In two cases, replication terminating protein and prosequence of subtilisin, the predicted structures are very similar to the experimentally determined folds. For the first time, in a public blind test, the unknown structures of proteins have been predicted ahead of experiment to an accuracy approaching molecular detail. In two other cases the approximate folds have been predicted correctly. According to the assessors there were 12 recognizable folds among the target proteins. In our postprediction analysis we find that in 7 cases our fold recognition technique is successful. In several of the remaining cases the predicted folds have interesting features in common with the experimental results. We present our procedure, discuss the results, and comment on several fundamental and technical problems encountered in fold recognition.** © 1995 Wiley-Liss, Inc.

Key words: knowledge based potentials, molecular modeling, prediction of protein structure, protein function, genome projects

INTRODUCTION

A long standing goal in protein structure theory is the computation of native protein structures solely from the information contained in amino acid sequences. The last decades have seen a variety of different approaches and many difficulties have been encountered. But the field has made progress along several directions. In particular, since the pioneering work by David Eisenberg's group[1] fold recognition is regarded as a promising new technology having the potential to reveal the approximate structures of many proteins by computational means.

The idea that fold recognition can be used to predict structures is based on the observation that proteins often have similar three-dimensional folds even if there is no detectable homology on the sequence level. Given an amino acid sequence there is a good chance that its unknown native fold is already represented in the data base of known structures. By combining sequences with structures it should be possible to identify such coincidences.

The idea is convincing and was followed or independently pursued by several groups[2–11]. The main goal is to demonstrate that fold recognition is indeed capable of recognizing relationships undetectable on the sequence level. Some case studies reported success for such distant relationships but a true blind test where a structure has been predicted ahead of experiment and has later been verified by experimental data has been lacking. This prediction experiment provided the proper basis for an objective assessment of the current state of the art in fold recognition and it was an exciting challenge for the fold recognition methods currently developed in our laboratory.

Fold recognition is an elegant approach to protein structure prediction but its realisation as a computer program meets several obstacles. Many of these are technical problems but some are fundamental. In the following section we summarize the main problems we encountered in our implementation followed by a presentation of our fold recognition algorithm. We submitted results for 13 targets. We demonstrate that in at least four cases the predictions are to a significant extent in agreement with the experimental data. Our goal was to submit results for every target, but we were unable to achieve this within the available time. For three additional targets the technique would have been successful. We close with a discussion of the results and implications for future developments and applications. The fold recognition program ProFIT used in this experiment is available from gundi.came.sbg.ac.at by anonymous ftp.

STRATEGIES AND PROBLEMS IN FOLD RECOGNITION

In fold recognition the goal is to identify one or more structures in a data base of known protein folds which are related to the unknown structure of a given target sequence. The sequence is combined

Received March 21, 1995; revision accepted June 9, 1995.

Address reprint requests to Manfred J. Sippl, Center for Appliced Molecular Engineering, Institute for Chemistry and Biochemistry, University of Salzburg, Jakob Haringer Str. 1, A-5020 Salzburg, Austria.

with each structure in some optimal way. In general reasonable alignments can be obtained only if gaps are allowed in the sequence and/or structure. The alignment obtained is equivalent to an approximate model of some conformational state of the target sequence. In most cases this model is incomplete, since parts of the sequence have been removed in the alignment production. Some of the models generated may be related to the unknown fold of the target sequence (but most of them will be unrelated). A crucial final step in fold recognition is to identify these native like models.

On the technical side there are two crucial steps in fold recognition (1) alignment or model production and (2) model assessment. The task in alignment production is to find an optimal or at least useful arrangement of the amino acid sequence on the structural template. This requires some kind of scoring function expressing the quality of fit of alternative arrangements as well as an efficient algorithm to find a good solution among the huge number of possible variants.

What are the features of a good scoring function? Most importantly the function has to be able to discriminate native like conformations from nonnative alternatives for the given sequence. The lack of such functions is one of the major obstacles in the classic protein folding problem but advances have been made. As described below we use a knowledge-based energy function for alignment production.

Efficient algorithms are available for the alignment of two sequences, but these techniques cannot be applied to the alignment of sequences and structures, at least not without approximations. Sequence alignment techniques are efficient because comparison of two sequence positions a_i and b_j depends only on the two amino acids involved so that the elements m_{ij} of the comparison matrix are strictly local quantities. The situation is quite different in sequence structure alignment. To determine the fitness or energy m_{ij} of amino acid a_i at structure position b_j the arrangement of all other amino acids a_k has to be known, since they define the environment of a_i with respect to the structural template.

In a recent analysis Rick Lathrop[11] has shown that there is little hope of finding an efficient algorithm for this problem and there are only two options: Spend a large amount of computing time on a single alignment if you want to find an optimum or almost optimum solution or use heuristics to produce reasonable alignments.

The problem with the first option is that the computing requirements are prohibitive for data bank searches. The problem with heuristics is that we have no guarantee of finding a good solution even if it exists. Nevertheless, our fold recognition is based on heuristics as shown below. Using heuristic or exact algorithms, the result still depends on the scoring function employed. The optimum alignment obtained with an unsuitable scoring function will be quite useless.

An additional complication is the treatment of gaps. What is the score or energy penalty for opening a gap? Gaps are necessary to achieve reasonable alignments, but there is no sound physical basis for choosing a gap penalty. Heuristic arguments and trial and error are the choices. In fact this points to a fundamental problem of the fold recognition approach. Alignment of sequences and structures and hence fold recognition does not correspond to a naturally occurring physical process. Rules and features applicable to the folding of protein chains are not necessarily valid in fold recognition.

A profound feature of protein chains is that amino acid neighbors along the sequence are constrained to remain close in space. This constraint dictates much of the features of protein folds. Within protein chains, segments can be forced to adopt strained conformations as long as these structures are stabilized by the remainder of the chain. What is the result when a cut is introduced at some position along the chain? The structure may remain stable, but it also can unfold, especially if the cut hits a strained section. There are numerous experimental examples where proteins unfold or rearrange when the polypeptide backbone is cleaved.

In sequence structure alignment an amino acid sequence can slide along a solid framework and split at unfavorable regions to find an arrangement which is favorable under the applied scoring function and the structural constraints. Even if a sequence is aligned with its known native fold there is no guarantee that the native arrangement is obtained. In fact, the native arrangement is never found with our algorithm, if gaps are not penalized. On the other hand large gap penalties always yield the native arrangements—but this is a rather trivial result.

Gap penalties are artificial constructs, but they are necessary to suppress strong fragmentation in the alignments. They are perhaps the greatest obstacles in the application of fold recognition techniques since a proper universal gap penalty is hard to find. Native protein structures are frustrated systems. Although the whole system is in equilibrium, parts of the system are relaxed or stressed as a function of the local environment. Depending on the location a large or small gap penalty may be appropriate, but there is no obvious way to define a location dependent penalty.

A given sequence structure combination may correspond to a native like model for the unknown fold, but in a data base search most if not all of the models will be incorrect and we have to identify those which have a good chance to be similar to the unknown native fold. What are the features of a good model? A first guess would be that the total energy of a model can be used to judge its quality. When alignments are calculated sequences are trimmed and frag-

mented to fit optimally into the structural framework. As a result the individual models obtained by combining a single sequence with the structures in the data base generally differ substantially in their sequences. The problem is that energies of different sequences cannot be compared, at least not for the energy function employed here.[12] Instead the energy of each model has to be compared to the energy distribution of its associated sequence in conformation space and the quality of each model is expressed in terms of a score obtained from this distribution.

Trivially, a native-like model can be obtained only if the unknown fold is represented in the data base. In other words, the data base must contain one or more folds which are similar in geometric terms to the unknown fold of the target sequence. But what is the meaning of similar? To what extent must a model match the native fold in order to be useful for further studies?

There is no general answer to this question. Obviously, the result is ideal if the model is very similar to the native fold. In other cases the fold class may be correct, but the model may differ considerably from the native fold, or parts of the model may be right and other parts may differ. Many of the similarities among protein structures described in the literature can be recognized only with sophisticated geometric alignment techniques and often only a fraction of atoms can be superimposed. It is not at all clear whether such distant similarities have any significance in terms of molecular biology or whether such similarities are useful in modeling one structure on the basis of the other. And there is certainly a point where even the most reliable fold recognition techniques conceivable will fail.

In a sense fold recognition is even harder than the classic folding problem. In the latter the goal is to predict the genuine native fold of a protein. If we had proper energy functions whose global minima correspond to native structures and a search strategy to locate these minima the protein folding problem could be solved. But in fold recognition we try to find structures differing from native folds to an a priori unknown extent and this could be even more difficult, especially if the energy functions are very sensitive to deviations from the native fold. Of course, the advantage of fold recognition is that the search problem is considerably reduced, provided we have an efficient alignment technique at hand.

Let's assume for the moment that we have an ideal fold recognition technique: an optimal scoring function, an efficient algorithm finding the optimum alignment in each case, proper gap penalties, and reliable quality assessment. Would this solve the fold recognition problem in all cases with reliable results? Unfortunately the answer is no. There is still a fundamental problem. In an alignment the sequence is forced to fit a structural framework. In most cases this is possible only by removing parts of the sequence and by splitting the sequence at several locations. The result is a new sequence which may differ considerably from the original. This new sequence may indeed have a structure closely related to the predicted fold. But the native structure of the original sequence could be very different.

In fact, fold recognition predicts the structure of the transformed sequence, not the fold of the original sequence. These two structures may be similar but we do not know unless we build complete models for the original sequence from the results obtained in fold recognition. Here we arrive at the classic protein folding problem. Given the tools currently available for protein prediction and modeling chances are that we succeed in building a complete structure sufficiently close to the native fold. But success or failure will depend on the model obtained from fold recognition. In any case this second step, model completion and refinement, generally will be more demanding and time consuming than fold recognition.

After all, in fold recognition we are confronted with a set of interdependent problems and a successful fold recognition technique must incorporate reasonable solutions to all of them. What follows is a description of our fold recognition implementation used in the prediction experiment.

FOLD RECOGNITION IMPLEMENTATION USED FOR PREDICTIONS

The main components of any fold recognition technique are

- An energy function or scoring scheme for the evaluation of sequence/structure compatibility or fitness.
- A sequence/structure alignment technique for optimal or at least useful alignment production.
- Proper quality assessment to determine whether or not the models obtained from sequence/structure alignment have native like features.
- A structure data base containing as many folds as possible.

Energy Function

For the evaluation of sequence structure compatibility we use a knowledge-based energy function consisting of mean force potentials.[9,12–17] The function is composed of pairwise atom–atom interactions and a surface term which accounts for protein solvent interactions. Although our present force field has terms for all backbone atoms, in the prediction experiment we used only C^{β}–C^{β}-based potentials for pair and surface interactions.

The pair interactions are defined by

$$E(a,b,k,r) = -ln\left[\frac{f(a,b,k,r)}{f(k,r)}\right] \quad (1)$$

where a and b represent the C^{β} atoms of amino acids

(e.g., a = VAL-C^{β}, b = ALA-C^{β}), k is the separation of a and b along the amino acid sequence, and r is the spatial distance between a and b. The potentials are asymetric, i.e., $E(a,b,k,r) \neq E(b,a,k,r)$. Distinction of potentials by k is important for small sequence separations but not for larger values. Individual potential types were used for sequence separations k = 1,2,3,4,5,6, and two additional potential types for k = 7,8,9 and k = 10, . . . ,∞. Potentials were compiled to a distance cut off of 15 Å, and set to zero for larger values of r. For data sampling the distance range was divided into intervals of 0.25 Å.

The frequencies f were obtained from a data base of known structures by

$$f(a,b,k,r) = n(a,b,k,r)/n(a,b) \tag{2}$$

and

$$f(k,r) = n(k,r)/n \tag{3}$$

where $n(a,b,k,r)$ etc. are absolute frequencies obtained from the data base and the energies were derived by applying a sparse data treatment.[13]

The surface term is of the form[12,17]

$$E(a,s) = -kT \ln\left[\frac{f(a,s)}{f(s)}\right] \tag{4}$$

with s the number of C^{β} atoms in a sphere of 10 Å around a and

$$f(a,s) = n(a,s)/n(a), \qquad f(s) = n(s)/n. \tag{5}$$

Alignment Technique

The alignment technique is based on the Needleman–Wunsch algorithm.[18] The algorithm determines a path through a matrix of elements $m(i,j)$ corresponding to a globally optimal alignment between two sequences A and B. In sequence comparisons $m(i,j)$ is a measure of similarity between amino acid i of sequence A and j of sequence B. To apply this concept to sequence–structure alignment a suitable definition for $m(i,j)$ must be found. $m(i,j)$ should express the fitness or compatibility of amino acid a_i of sequence A at location j of conformation C. To evaluate this quantitity the position of most residues of A on C has to be known, since $m(i,j)$ depends on the environment of i

$$m(i,j) = \sum_{q} E(a_{i_j}, b_{p_q}, k, r) \tag{6}$$

where a_{i_j} represents the amino acid at sequence position i placed at position j of conformation C with a corresponding meaning for the symbol b_{p_q}. For a given combination (i,j) there are, of course, many possibilities to arrange the sequence on the structural template. Jones et al.[3] determine the optimum arrangement (p,q) for each combination (i,j) and define this to be the value of $m(i,j)$. Hence, the calculation of $m(i,j)$ involving optimization is rather time consuming.

To keep the computational requirements low, we use the energy field generated by the original sequence of conformation C. We simply mutate the residue at position c_j to the residue type found at position a_i and calculate the energy of this residue in the original environment of C

$$m(i,j) = \sum_{q \neq j} E(a_{i_j}, c_q, k, r) \tag{7}$$

and the optimum arrangement is obtained from a standard Needleman–Wunsch alignment applied to the matrix $m(i,j)$. This is a crude approximation, which may or may not work and whose value can be judged only by the results obtained.

Gap Penalty

The alignments obtained depend quite sensitively on the actual gap penalty used. A single gap penalty works in some situations but fails in others as discussed above and position-dependent penalties are also problematic. In our strategy we determined the range of reasonable gap penalties and defined one default value. Alignments were generated using the default value and, if time permitted, several alternative alignments were generated by changing the gap penalty. Judgment of alternative solutions was deferred to the model assessment stage.

Model Assessment

For a given target sequence alignments (and in some cases alternative alignments) with every structure in the data base were generated and the quality of each alignment was expressed and ranked in terms of a z-score. The score was calculated by the polyprotein technique as described by Sippl and Jaritz.[15] For a given sequence the method estimates the average energy $\bar{E}$ and standard deviation σ in conformation space. Then the energy of the corresponding model E is calculated and transformed to a z-score by

$$z = (E - \bar{E})/\sigma. \tag{8}$$

Note that the sequences are generally changed in the alignment process. Even for a single target sequence the individual model sequences will be distinct and $\bar{E}$ and σ have to be determined for each individual model. The program to calculate scores and to analyze structures,[16] called PROSA-II, is available via anonymous ftp from gundi.came.sbg.ac.at.

Structure Data Base

Data bases of structures were derived from the Brookhaven Protein Data Bank.[19] Sequence–structure alignment is often very sensitive to the particular circumstances. The algorithm may produce a

TABLE I. Prediction Targets

Target	Recognizable fold	rms/res	Fold class	Submitted
rtp	1hst-a	1.6 39	α + β	Yes
prosub	2nck-1	2.0 32	α + β	Yes
pcna	2pol-a	2.1 130	α/β	Yes
ppdk 4	5tim-a	2.3 88	TIM-barrel	No
ppdk 3	1aco	2.4 34	TIM-barrel	No
staufen3	1pda	2.6 36	α + β	Yes
kauB	2bpa-1	2.7 29	IG-like domain	No
xylanase	1btc	2.8 122	TIM-barrel	Yes
synapto	1cob-1	2.8 29	IG-like	Yes
kauA	5tim-a	2.9 71	TIM-barrel	No
pbdg	1nar	3.0 108	TIM-barrel	Yes
ce-1	Unique	— —	—	Yes
bhted	Unique	— —	—	Yes
smanucecs	Unique	— —	—	Yes
chmut	Unique	— —	All-α	Yes
ppdk 1	Unique	— —	—	Yes
mystery	?	— —	?	Yes

good solution in one case, but may fail on a closely related structure. Therefore, the structure data base should contain as many structures as possible and it is advantageous to include closely related structures.

Although, our fold recognition algorithm is rather efficient in terms of CPU time we had to use data bases of varying size to meet the submission deadlines. For the calculations on the catalytic core of xylanase, for example, we used a data base of approximately 1,000 structures but only 300 structures for most of the other targets.

Local Structure Calculations

In addition to fold recognition we calculated the local structural preferences (helix, strand, coil) for the submitted target sequences.[20] Conformational ensembles for all possible pentapeptides were computed and assembled to contiguous backbone conformations. In these calculations only pair interactions were used, since the surface terms are unsuitable for small peptides. Assembly of ensembles is a pure geometric process. Since pentapeptides contain only interactions whose sequence separation is less than 5 residues the backbone conformations obtained are functions of the short range interactions only. In general the results are indicative of the local structural preferences of target sequences but not their overall fold.

The local structures calculated by this procedure were compared to the models obtained as an additional criterion to judge the fold recognition results and to get an estimate for the structural preferences for those parts of a target sequence which remained unaligned in the models.

RESULTS

We submitted results for 13 targets in accordance with the rules set by the organizers. A list of the 20 highest scoring models was submitted for each target, together with the alignment obtained for the highest scoring model. Results obtained from the assembly of fragments were submitted in terms of secondary structure assignments.

A fair judgment of the predictions turned out to be a complex and difficult task. Detection of similarities among distantly related structures in general requires sophisticated algorithms. Current methods agree when the structures are closely related, but they produce different and sometimes complementary results in less obvious cases. The different criteria to represent structural similarities provide additional obstacles. Similarities are reported in terms of the percentage of atoms superimposable with an error less than some prescribed threshold, but the results depend on the respective method, its parameters, and the specific values chosen.

Table I presents all targets and indicates whether or not models were submitted. The targets are sorted with respect to their similarity to a known fold as determined by the assessors, indicating whether or not the structure of a particular target could have been recognized by fold recognition. In Table II we assemble our results. In some cases, e.g., for target ce-1, our result indicated that the model obtained was to some extent similar to the experimentally determined structure. Subsequent structure–structure comparison using alternative techniques indeed revealed structural similarities (32 residues of ce-1 and 4sbg-i are superimposable to an rms error of 3.3 Å—L. Holm, personal communication). This and similar examples highlight the difficulties in evaluating the quality of predictions.

In presenting our results we try to keep things as simple as possible. Perhaps, the following scheme provides a simple reference frame which helps to qualify the predictions:

TABLE II. Predicted Structures

Target	Prediction	rms/res
	Class 1 predictions	
rtp	1hst-a	2.4 65
prosub	2fxb	3.5 —
	Class 2 predictions	
ce-1	4sgb-i	3.3 32
xylanase	1tim-b	3.9 158
	Class 3 predictions	
staufen3	1ctf	— —
chmut	1cdp	7.2 53
synapto	2hpr	4.5 51
	Class 4 predictions	
smanucecs	5cpv/1dra	
pbdg	1mat	
bhted	1lz6	
pcna	1ypi-a	
ppdk 1	1gky	
mystery	1pgd	

1. The predicted structure is very similar to the experimental result. The model obtained can be used to investigate specific structural problems and it is a useful starting point for model refinement and biochemical studies.

2. The predicted structure corresponds to the correct folding class, but the molecular details differ from the experimental results. The result may indicate some structural features, but it is insufficient to initiate detailed structural studies.

3. The predicted structure has some features in common with the experimental result, but the overall fold is different.

4. There is no obvious relationship between predicted structure and experimental result that could be used for further studies.

Table II organizes our predictions in terms of this scheme. The predictions for replication terminator protein (rtp) and the propiece of subtilisin (prosub) are very similar to the respective experimental results. We put the results for the catalytic core of xylanase and the protease inhibitor ce-1 in category 2. Although the fold class is correctly predicted in both cases the models and actual structures differ in several details. The four predictions in category 3, although dissimilar in overall structure, still have interesting similarities to the experimentally determined folds.

Replication Terminator Protein

Perhaps the most interesting target for fold recognition was the replication terminator protein from *B. subtilis* (rtp). The sequence is unique. There are no homologous sequences in the current sequence data bases, so that methods based on multiple alignments were not applicable. Our fold recognition algorithm predicted that rtp adopts a conformation related to histone H5 (1hst-a).

Indeed the rtp fold turned out to be very similar to 1hst-a and the degree of similarity is in itself surprising if not spectacular (Fig. 1). There is no homology between the two sequences, but nevertheless the C^{α} atoms of 64 residues can be superimposed to an rms error of 2.4 Å (or 39/1.6 Å). The sequence–structure alignment (Fig. 2) obtained from fold recognition is in good, but not perfect agreement with the structure–structure alignment. Gaps are set at appropriate positions and in the alignment 91% of the secondary structure assignments (helix, strand, or random) match. rtp has a C-terminal extension which has no counterpart in 1hst-a. The local structure calculation predicted a helix for this region, which also agrees well with the experimental result.

The rtp/1hst-a case is important for several reasons. It shows that structures can be quite similar even if their sequences are unrelated and the successful blind prediction demonstrates in a completely objective way that computational techniques are capabale of recognizing this structural similarity. Moreover, the structural model obtained for rtp provides a sound basis for further biochemical and structural studies even if its genuine native fold is unknown.

Propeptide of Subtilisin

The 77-residue propeptide of subtilisin BPN′ from *B. subtilis* (prosub) was found to be compatible with the structure of ferredoxin 2fxb. The structure of prosub is similar to 2fxb (Fig. 3) and superposition yields an rms error of 3.5 Å (32/2.0 Å). The result for the prosub structure is again similar enough to its experimentally determined fold so that the model could be used as a basis for further investigations.

Our data base contained another structure, carboxypeptidase A (1pca), even more similar to prosub than 2fxb, but this fold was not recognized. A subsequent analysis revealed that our algorithm failed to generate an alignment of sufficient accuracy resulting in a model of comparatively low score. In fact, it was impossible to find suitable alignment parameters, but a useful model was obtained when the 18 C-terminal residues, corresponding to a contiguous α-helix, were removed from the structure.

Chymotrypsin/Elastase Inhibitor

The predicted model for chymotrypsin/elastase inhibitor-1 from *Ascaris lumbricoides* (ce-1) was an alignment with 4sgb-i, chymotrypsin inhibitor from *P. tuber*. The two structures were not considered to be similar by the assessors. However, an alignment generated by L. Holm reveals significant structural similarities corresponding to an exposed loop in both molecules (Fig. 4).

The example shows the difficulties encountered in the evaluation of predictions, but it also demon-

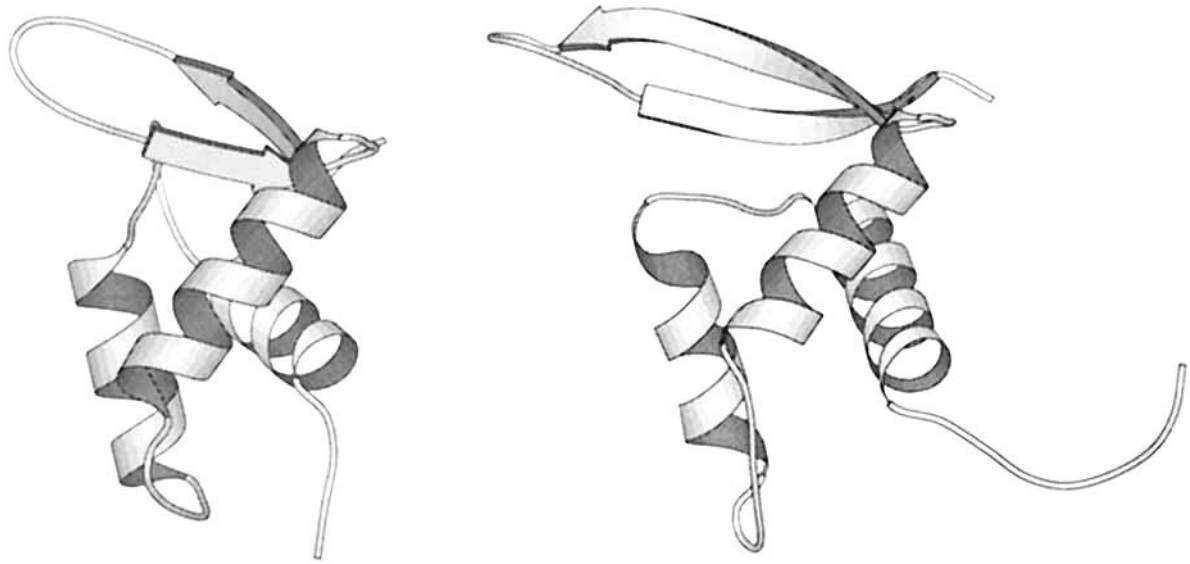

Fig. 1. Histon H5 from chicken, 1hst-1 (left) and replication terminator protein (rtp) from *B. subtilis* (right). The C-terminal helix of rtp is not shown. The C^{α} atoms of 65 residues in 1hst-a and rtp can be superimposed to an rms error of 2.4 Å.

```
                        HHHHHHHHHHHHHHHHH     HHHHHHHHHHHH          HHHHHHHH
rtp      MKEEKRSSTGFLVKQRAFLKLYMITMTEQERLYGLKLLEVLRSEFKEIGFKPNHTEVYRSL
                        =========             ===========           ========
1hst-a   ....SHPT-----YSEMIAAAIRAE--KSRGGSSRQSIQKYIKSHYKVG-HNADLQIKLSI
                       HHHHHHHHHH           EEHHHHHHHHHHH         HHHHHHHHH

         HHHHH    EEEEEEE         EEEEEEE  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
rtp      HELLDDGILKQIKVKKEGAKLQEVVLYQFKDYEAAKLYKKQLKVELDRCKKLIEKALSDNF
         =====       ====          ====
1hst-a   RRLLAAG---VLKQTKGVGA-SGSFRLA-K.............................
         HHHHH       EEEE          EEEE
```

Fig. 2. Sequence/structure alignment of rtp and 1hst-a. The secondary structure type is indicated by H for helices and E for strands. '=' marks paired residues of equivalent secondary structure.

strates the considerable power and usefulness of fold recognition techniques. The method is able to reveal structural coincidences which are not obvious even when both structures are known, and it can reveal important structural and functional information which may be inaccessible by other methods. For example, if the function of ce-1 would be unknown (which is the case for a large fraction of sequenced genes) the result obtained here would clearly point to its biochemical role.

Catalytic Core of Xylanase

The catalytic core of xylanase was predicted to fold into a TIM-barrel, which was confirmed by the experimental structure. However, the result is less significant as compared to the previous examples. The method correctly identified the overall fold by matching the target with 1tim-b, triosephosphate isomerase, a classic TIM-barrel, but the sequence–structure alignment is quite distinct from the optimal geometric superposition and the two proteins differ in length. There is considerable structural variability among TIM-barrels and in this case it is less clear to what extent the predicted model would support further studies. Moreover, it should be noted that the TIM-barrel motif occurs quite frequently. In fact, five of the target sequences turned out to belong to this family, and consequently a consistent bet on TIM-barrels would have been rewarded by five successful fold recognition results.

Comments on Other Targets

For the remaining submitted targets no recognizable structure is known or our algorithm failed to

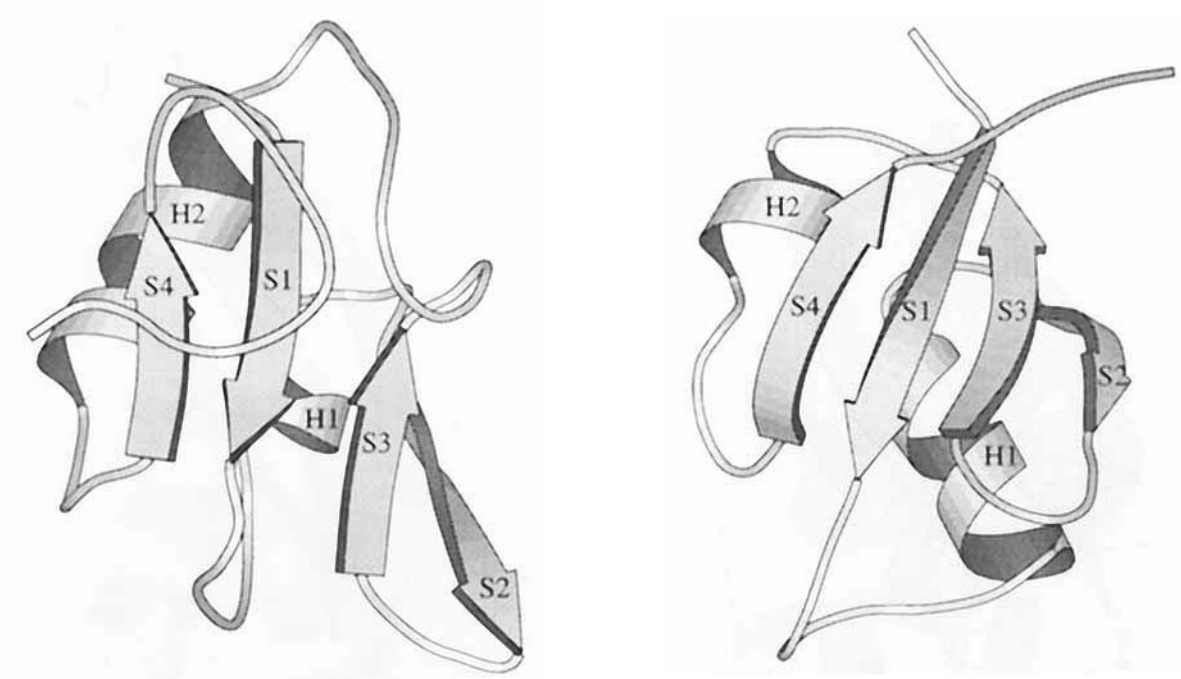

Fig. 3. Ferredoxin (2fxb) from *B. thermoproteolyticus* (left) and propeptide of subtilisin BPN (prosub) from *B. subtilis* (right). Prosub and 2fxb can be superimposed to an rms error of 3.5 Å.

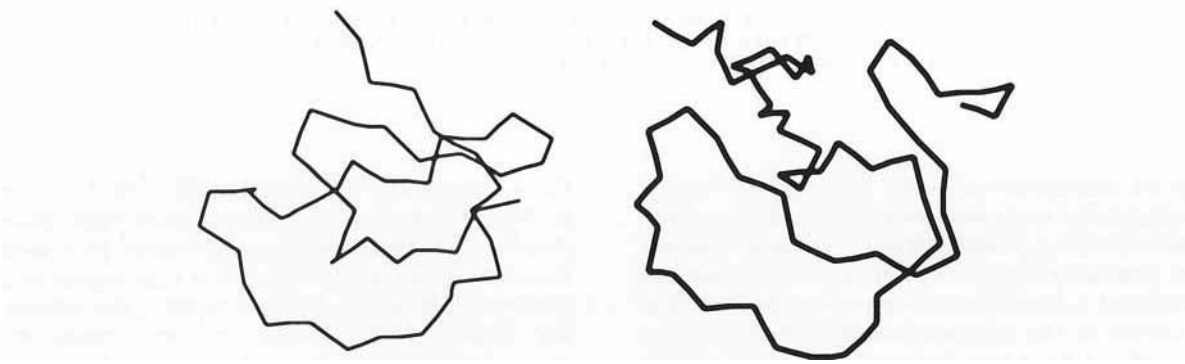

Fig. 4. Plot of chymotrypsin/elastase inhibitor-1 (ce-1) from *Ascari lumbricoides* (bold line) and chymotrypsin inhibitor (4sgb-i) from potato tuber (thin line). Structural similarities are significant in the exposed loop of both folds.

detect a recognizable fold (Tables I and II). One example is staufen3, domain 3 of staufen from *Drosophila*. The Brookhaven data base contains at least one related fold, 1pda, porphobilinogen deaminase from *E. coli*, with 36 C^{α} atoms superimposable to an rms error of 2.6 Å. Our data base did not contain 1pda and our subsequent analysis revealed that the algorithm would have failed to detect 1pda anyway. Instead, we predicted 1ctf, the C-terminal fragment of ribosomal protein L7/12 from *E. coli*. In fact, staufen3 and 1ctf have quite similar structures but different topologies (Fig. 5) with distinct connections of secondary structure elements. Nevertheless, the similarity between the two molecules is striking.

No recognizable structure was available for chmut, chorismate mutase from *E. coli*. The structure of chmut is composed of three long α-helices. The top scoring folds obtained for this target are all α-helix rich proteins, like the calcium binding protein 1cdp or myohemerythrin, 2mhr, and one may speculate to what extent these predictions would be suitable starting points for subsequent modeling and refinement.

The first C2 domain of synaptotagmin (synapto)

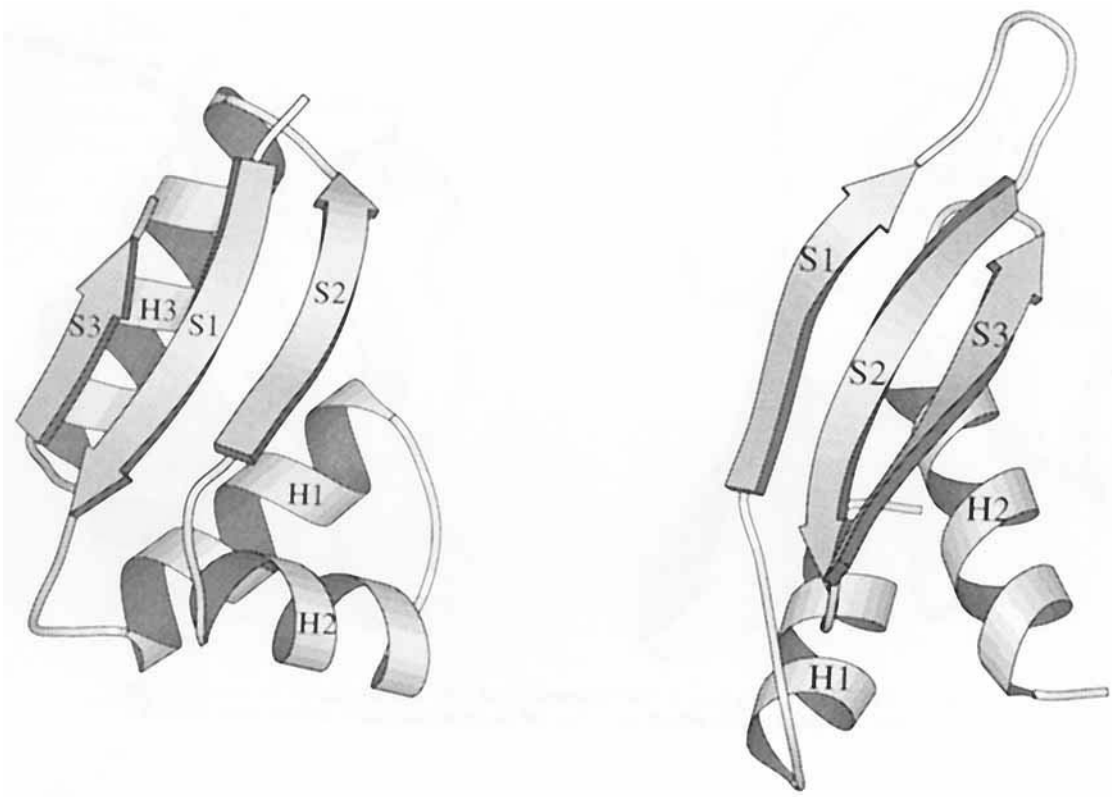

Fig. 5. C-terminal fragment of ribosomal protein L7/12 (1ctf) from *E. coli* (left) and domain 3 of staufen from *Drosophila*. At first glance the fold of staufen seems to be very similar to 1ctf, but the connections of secondary structure elements are distinct.

has an immunoglobulin-like fold but our highest scoring model was a combination with 2hpr, a histidine-containing phosphocarrier. The latter has several β-strands similar to immunoglobulins but two additional helices. There is considerable structural variation in the immunoglobulin-like family. Our data base did not contain those folds which provide good models for synaptotagmin. If these structures are supplied the algorithm identifies 1ten, whose fold is an immunoglobulin-like β-sandwich. Hence, for synapto, the algorithm was not successful because proper structures were missing from the data base.

For the remaining targets (Table II) the models obtained have little overall resemblance to the experimental structures, although the local structures and folding types often coincide. According to the assessors, folds were recognizable for pbdg and pcna (Table I). pbdg belongs to the TIM-barrel family. Our top scoring model, methionine aminopeptidase (1mat), is an α + β fold and for pcna, an α/β fold we obtained triosephosphate isomerase (1ypi-a), a classic TIM-barrel.

For the artificial sequence mystery, designed to fold into a TIM-barrel, the actual structure is unknown and there are experimental indications that the sequence does not adopt a stable fold (L. Holm, personal communication). We predicted 1pgd, phosphogluconate dehydrogenase, composed of β–α–β-units. The topology of this fold is reminiscent of a distorted TIM-barrel. The structures of the remaining targets, bhted, ppdk1, and smanucecs, are unique and no similarities to known structures have been reported.

Our goal was to submit predictions for all targets, but in several cases the deadline was too close to finish the calculations. Among these targets are four cases, ppdk3, ppdk4, kauA, and kauB, whose structures were recognizable by fold recognition (Table I). Our algorithm correctly identifies the structure of kauA as a TIM-barrel and kauB as an immunoglobulin-like fold, but not the structures of ppdk3 and ppdk4 which are TIM-barrels. Interestingly, for ppdk4 we obtained 1bnh, a large ribonuclease inhibitor composed of a series of β–α–β– units, quite similar to the basic architecture of TIM-barrels. Again this example shows that it is quite difficult to assess the quality of predictions and the potential value of the results for further modeling studies. A detailed analysis in most cases reveals some relationship between predicted and actual structures, but detailed

TABLE III. Prediction Results on Recognizable Folds

Target	Correct	Correct/after	Total
	Submited targets		
rtp	*		*
prosub	*		*
ce-1	*		*
xylanase	*		*
synapto		*	*
staufen3			
pcna			
pbdg			
	Not submitted		
kauA		*	*
kauB		*	*
ppdk 3			
ppdk 4			
	Fold unknown		
mystery			

documentation and discussion of such coincidences are beyond our intentions.

DISCUSSION

Table III summarizes the results. The total number of targets having recognizable folds is 12. We submitted predictions for 8 of these targets. In four cases the predictions were correct and the models obtained provide significant structural and functional information. For synaptotagmin the recognizable fold was not contained in the data base but the algorithm is successful if the 1ten structure is supplied. The postprediction analysis shows that in two cases, where no results were submitted, the algorithm detects a suitable fold. If our data base had contained all available folds and if there had been enough time the procedure would have been successful in 7 out of 12 cases.

The results indicate that there have been significant advances in the development of energy functions and fold recognition techniques. For the first time the unknown structures of proteins have been predicted ahead of experiment to an accuracy approaching molecular details. This has been achieved in a public blind test.

The organizers ensured that no information on the structure of targets was available to the predictors so that the results can be evaluated in a completely objective fashion. In at least two cases, rtp and prosub, the predictions come close to atomic resolution and provide significant information on the structure and function of these molecules. In addition the study demonstrates that two proteins whose sequences are unrelated can have surprisingly similar folds. This is an important requirement in fold recognition, since the quality of the predictions depends on the available structures and their similarity to the genuine native fold.

For rtp/1hst-a the sequence–structure alignment is in good agreement with the alignment obtained from the geometric superimposition of these structures (Figs. 1 and 2). In other cases there are substantial differences. This could be due to problems in the alignment technique and/or unsuitable gap penalties, but the interesting point is that such models produce significant scores even if the sequence–structure alignment is distinct from the optimal geometric solution. In any case a detailed study of these effects is a complex task which may be necessary to improve the performance of the technique.

In fact there are many problem areas where improvements are conceivable. It seems that the technique performs less reliably for large proteins, or in cases where sequences and structures have distinct lengths. This could be due to the alignment technique used. The Needleman–Wunsch algorithm searches for a global solution and this may be a suboptimal strategy in cases where sequences and structures differ in length. Systematic variation of parameters and use of alternative alignment algorithms are promising extensions increasing the probability to generate one or several suitable sequence–structure combinations whose scores are significant.

From this study it is clear that fold recognition has become a powerful tool in structural biology. Structural and functional information can be obtained for sequences in cases where no biological significance can be assigned by sequence comparison or other established techniques. In a recent large scale study[21] we applied fold recognition to approximately 500 putative genes found in the central gene cluster of *C. elegans* chromosome III. The biological role for 60% of these sequences is unknown. Around 10% of these sequences can be combined with some known structure yielding significant scores. Genome projects produce a huge number of sequences whose biological role cannot be interpreted. For a substantial number of these sequences fold recognition will yield otherwise inaccessible structural and functional information.

As already noted there is no guarantee that fold recognition results are correct even if the scores indicate native-like quality. Additional studies are required to corroborate the results. Given the success rate obtained here, roughly half of the predictions can be expected to be useful starting points for further studies and there are obvious possibilities to increase the number of successful predictions: use of larger fold data bases, refinement of energy functions, and improvements in alignment techniques.

ACKNOWLEDGMENTS

We thank Liisa Holm, John Moult, Chris Lemer, Marianne Rooman, and Shoshana Wodak for several structure–structure alignments which were invaluable in assessing the predictions. Figures for molec-

ular structures were prepared using the program MOLSCRIPT.[22] We are most grateful for the continuous support and encouragement by Edgar Morscher, Rektor of the University of Salzburg. This work was supported by grant P09661-MOB (Fonds zur Förderung der wissenschaftlichen Forschung, Austria) and grant 5158 (Österreichische Nationalbank, Austria).

REFERENCES

1. Bowie, J.U., Luethy, R., Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170, 1991.
2. Sippl, M.J., Weitckus, S. Detection of native like models for amino acid sequences of unknown three dimensional structure in a data base of known protein conformations. Proteins 13:258–271, 1992.
3. Jones, D.T., Taylor, W.R., Thornton, J.M. A new approach to protein fold recognition. Nature (London) 358:86–89, 1992.
4. Godzik, A., Kolinski, A., Skolnick, J. Topology fingerprint approach to the inverse folding problem. J. Mol. Biol. 227: 227–238, 1992.
5. Ouzounis, C., Sander, C., Scharf, M., Schneider, R. Prediction of protein structure by evaluation of sequence-structure fitness. Aligning sequences to contact profiles derived from 3D structures. J. Mol. Biol. 232:805–825, 1993.
6. Wilmans, M., Eisenberg, D. Three-dimensional profiles from residue-pair preferences: identification of sequences with beta/alpha-barrel fold. Proc. Natl. Acad. Sci. U.S.A. 90:1379–1382, 1993.
7. Bryant, S.H., Lawrence, C.E. An empirical energy function for threading protein sequence through folding motif. Proteins 16:92–112, 1993.
8. Johnson, M.S., Overington, J.P., Blundell, T.L. Alignment and searching for common protein folds using a data bank of structural templates. J. Mol. Biol. 231:735–752, 1993.
9. Sippl, M.J., Weitckus, S., Floeckner, H. In search of protein folds. In: "The Protein Folding Problem and Tertiary Structure Prediction." Merz, K.H., LeGrand, S. (eds). Boston: Birkhäuser, 1994: 353–407.
10. Sippl, M.J., Weitckus, S., Floeckner, H. Fold recognition. In: "Modelling of Biomolecular Structures and Mechanisms." Pullman, A., Jortner, J., Pullman, B. (eds.) Kluwer, 1994, in press.
11. Lathrop, R.H. The protein threading problem with sequence amino acid interaction preferences is NP-complete. Protein Eng. 7:1059–1068, 1994.
12. Sippl, M.J. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. J. Comput. Aided Mol. Des. 7:473–501, 1993.
13. Sippl, M.J. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J. Mol. Biol. 859–883, 1990.
14. Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., Sippl, M.J. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. J. Mol. Biol. 216: 167–180, 1990.
15. Sippl, M.J., Jaritz, M. Predictive power of mean force pair potentials. In: "Protein Structure by Distance Analysis." Bohr, H. Brunak, S. (eds.). Amsterdam: IOS Press, 1994: 113–134.
16. Sippl, M.J. Recognition of errors in three-dimensional structures of proteins. Proteins 17:355–362, 1993.
17. Sippl, M.J., Jaritz, M., Hendlich, M., Ortner, M., Lackner, P. Applications of knowledge based mean fields in the determination of protein structures. In: "Statistical Mechanics, Protein Structure and Protein-Substrate Interactions." Doniach, S. (ed). New York: Plenum, 1994: 297–315.
18. Needleman, S.B., Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48:443–453, 1970.
19. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Data Bank: A computer-assisted archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
20. Sippl, M.J., Hendlich, M., Lackner, P. Assembly of polypeptide an protein backbone conformations from low energy ensembles of short fragments. Development of strategies and construction of models for myoglobin, lysozyme and thymosin β_4. Protein Sci. 1:625–640, 1992.
21. Braxenthaler, M., Sippl, M.J. Screening genome sequences for known folds. In press.
22. Kraulis, P.J. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. J. Appl. Crystal. 24:946–950, 1991.