BIOINFORMATICS

Vol. 15 no. 2 1999
Pages 131–140

FORESST: fold recognition from secondary structure predictions of proteins

V. Di Francesco[1,2], P. J. Munson[1] and J. Garnier[1,3]

[1]Analytical Biostatistics Section, Mathematical and Statistical Computing Laboratory, The Institute, Center for Information Technology, National Institutes of Health, Bethesda, MD 20892-5626, USA, [2]Biolgiza, Room 2041–5626, Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA and [3]Laboratoire de Biologie Cellulaire et Moléculaire, Biotechnologies, INRA, 78352 Jouy-en-Josas, France

Received on June 19, 1998; revised on September 30, 1998; accepted on November 19, 1998

Abstract

Motivation: *A method for recognizing the three-dimensional fold from the protein amino acid sequence based on a combination of hidden Markov models (HMMs) and secondary structure prediction was recently developed for proteins in the Mainly-Alpha structural class. Here, this methodology is extended to Mainly-Beta and Alpha-Beta class proteins. Compared to other fold recognition methods based on HMMs, this approach is novel in that only secondary structure information is used. Each HMM is trained from known secondary structure sequences of proteins having a similar fold. Secondary structure prediction is performed for the amino acid sequence of a query protein. The predicted fold of a query protein is the fold described by the model fitting the predicted sequence the best.*

Results: *After model cross-validation, the success rate on 44 test proteins covering the three structural classes was found to be 59%. On seven fold predictions performed prior to the publication of experimental structure, the success rate was 71%. In conclusion, this approach manages to capture important information about the fold of a protein embedded in the length and arrangement of the predicted helices, strands and coils along the polypeptide chain. When a more extensive library of HMMs representing the universe of known structural families is available (work in progress), the program will allow rapid screening of genomic databases and sequence annotation when fold similarity is not detectable from the amino acid sequence.*

Availability: *FORESST web server at http://absalpha.dcrt.nih.gov:8008/ for the library of HMMs of structural families used in this paper. FORESST web server at http://www.tigr.org/ for a more extensive library of HMMs (work in progress).*

Contact: *valedf@tigr.org; munson@helix.nih.gov; garnier@helix.nih.gov*

Introduction

Both *de novo* design and understanding of the biological function of a protein require the knowledge of the relationship between its amino acid sequence and its three-dimensional (3D) structure. *Ab initio* prediction of protein 3D structure from the amino acid sequence alone remains a challenging problem. A useful approximate solution to this problem consists of modeling the unknown structure of a protein sequence with the structure of an evolutionarily related protein (Browne *et al*., 1969; Greer, 1991). This approach depends on the quality of the sequence alignment to the related protein. Below a sequence identity of ~25%, sequence alignments often become unreliable, although the proteins may indeed be related and may share a common fold and function (Schneider and Sander, 1991; Doolittle, 1992; Jones and Thornton, 1993).

It has been observed that during evolution, the 3D structure of a protein is more conserved than its sequence (Chothia and Lesk, 1986). In order to detect the similarity of fold between sequences of low or undetectable sequence identity, optimal sequence threading methods have been developed [see Bryant (1996) for references]. These methods thread a sequence through known tertiary structures using profile or empirical contact potentials to judge the quality of fit. Some of these methods primarily use secondary structure information to relate protein sequence and its structure. Sheridan *et al*. (1985) generated a set of possible folds for a query protein by aligning its secondary structure sequence to that experimentally determined for proteins of known 3D structure. Unfortunately, they did not test this procedure using the predicted secondary structure sequence of the query protein. Russell *et al*. (1996) mapped predicted secondary structure segments, helices or β strands, to known secondary structure and used various filters to screen out irrelevant topologies. Hubbard and Park (1995) used amino acid sequence-based hidden Markov models (HMMs) and predictions of secondary

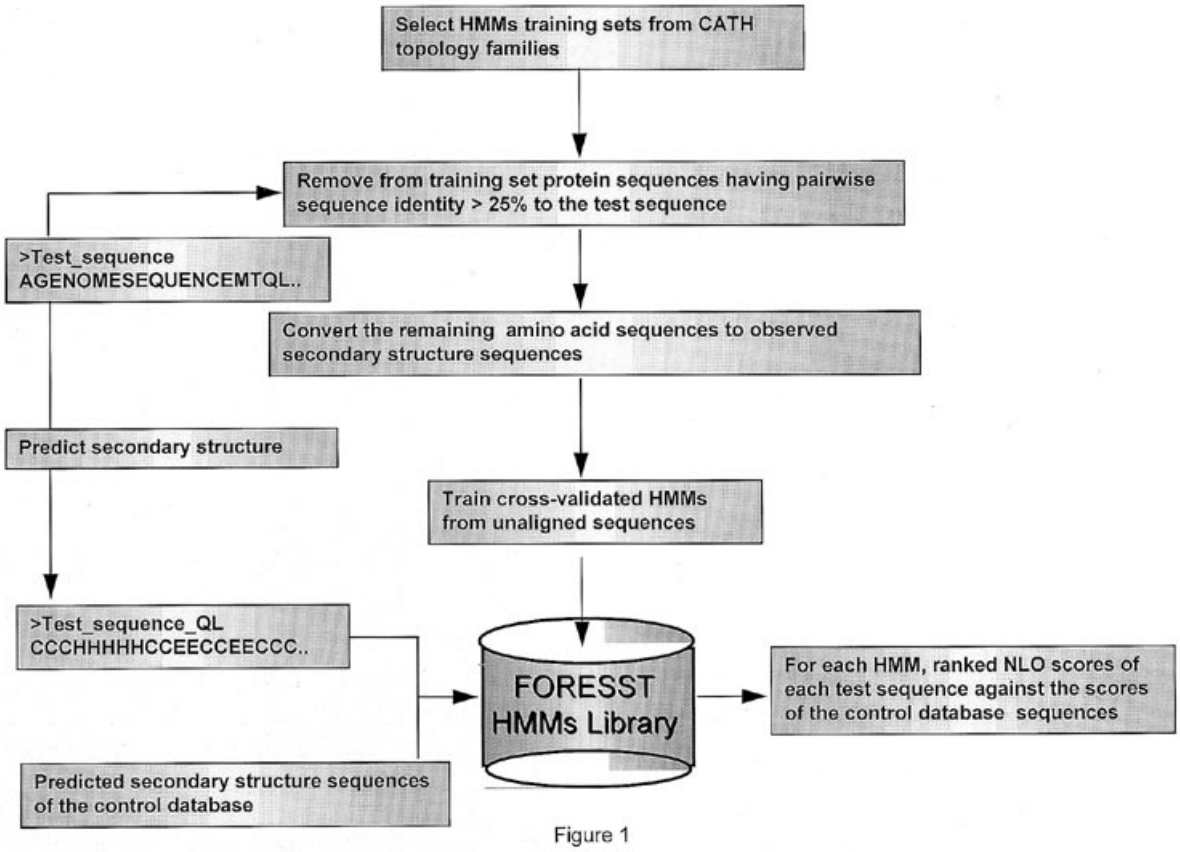

Fig. 1. Schematic representation of the training and testing procedure.

structures, in the case where a high β-strand content is predicted, with a strand-pairing potential used to predict the β-sheet topology. Rost (1995) and Fischer and Eisenberg (1996) combined various characteristics of the protein sequence, e.g. solvent accessibility or amino acid substitution matrices, with the predicted secondary structure in attempting to recognize a fold. Rice and Eisenberg (1997) extended this approach by developing a multi-dimensional substitution matrix which includes amino acid class, solvent accessibility and secondary structure predictions from the PHD server. We recently reported an approach to fold recognition based solely on secondary structure predictions for proteins of Mainly-Alpha structural class (Di Francesco *et al*., 1997a) using HMMs (Rabiner, 1989) of protein folds. Here, we present an extension of this work to the other structural classes, Mainly-Beta and Alpha-Beta, and thereby show the generality of this approach for recognizing a wider variety of existing folds. The accuracy of the method, 59–64% of correctly recognized folds for 44 tested proteins, was prospectively confirmed, achieving five correct fold predictions out of seven (71%) performed prior to knowledge of the experimental structures (Di Francesco *et al*., 1997a,b). This approach, implemented in the program named FORESST (for FOld REcognition from Secondary STructures), offers the advantage of allowing rapid screening of genomic databases and presents a useful alternative to other threading methods.

Materials and methods

The overall procedure involved in developing HMMs of protein structural topologies and validating them in their fold recognition capabilities is presented in Figure 1. A brief description of the methodology follows; further details can be found in Di Francesco *et al*. (1997a).

Hidden Markov models of protein folds

The HMM architecture proposed by Krogh *et al*. (1994) is adopted in this work. Hidden Markov models of protein folds, such as globins or TIM barrels, describe in a probabilistic way the structural similarity and diversity of proteins having similar folds, such as conserved regions and inserted or deleted loops or domains. This is achieved by means of a first-order Markov chain of the hidden states of the models: match, insert and delete, corresponding to each residue in the

protein chain. To implement such models, the software package Sequence Alignment and Modeling (SAM) (Hughey and Krogh, 1995), Versions 1.1 and 1.3.1, was modified as follows. Instead of having each match and insert state produce an observable nucleotide or amino acid, our models produce a secondary structure state: helix (H), strand (E) or coil (C). The reduction in alphabet size for model building (three characters instead of 20 for the different amino acid types) greatly reduces the number of model parameters to estimate (the transition probabilities between the hidden states and the observation symbol probability distributions associated to either the match or the insert states). A model is trained on a set of unaligned observed secondary structure sequences of proteins of a structural family and produces a residue-by-residue alignment of the secondary structure states. Transition probability prior distributions (Hogg and Craig, 1978) are set to the default values in the SAM software package. Observation symbol probability prior distributions for most of the HMMs are set to the frequencies of the three secondary structure states of the sequences in the training set. Proteins with similar folds were selected from the hierarchical database of protein structures CATH, Versions Jan. 1995, 1.0 and 1.1 (Orengo *et al*., 1993, 1994), at the topology level. Proteins in the same topology family are grouped based on both the overall shape and connectivity of the secondary structures. The proteins in each model training set were selected to maximize the diversity of sequences while having a sufficient number of structures to train the models of folds. The lower the percentage amino acid identity in the sequences of the model training set, the more heterogeneity can be found in their secondary structure sequences, and consequently the more generalized is the resulting HMM. The list of proteins in each model training set is available from the authors upon request.

Scoring

For each HMM, the normalized negative log-odds score (NLO) is calculated for the predicted secondary structure sequences of the proteins in a control database and of the query protein. The score is based on a ratio of the probability that the particular sequence arises from the model to the probability that it arises from a null model based on the average secondary structural composition of the training sequences in each model match states. The more negative the NLO score, the better the model fits the sequence. The negative log-odds scores are normalized by the length of each sequence. For each model, the assigned normalized score of the query is ranked against the scores of the control database proteins so that the best low ranks (i.e. rank 1, 2, etc.) were associated with the most negative scores. Ranks were utilized, rather than NLO scores, because scores from different models are not directly comparable, but depend strongly on the model length, the scored sequence length and the prior distributions used for model training. The model assigning the lowest rank to the query sequence becomes the predicted fold, even when the rank is not number one. For example, if the globin-like test protein 1eca is assigned a score by the HMM of the globin-like structural family which ranks sixth, while the scores assigned by the HMMs of the wrong structural families are ranked seventh or worse, the globin fold of 1eca would be declared to be correctly recognized. Whenever a test protein is assigned scores that rank 20th or worse by all the models, no fold prediction is carried out for that sequence.

Cross-validated training of models

To test whether correct fold recognition is attained in the presence of low or undetectable sequence similarity between proteins with similar folds, the models were trained with cross-validation. For each test protein sequence (see below), a model was trained on the other proteins having the same topology that were <25% sequence identical to the test protein. This cross-validation procedure sometimes reduces the number of proteins used in the training set considerably. For the EF-Hand proteins, it was necessary to raise the cut-off to 30% in order to have enough proteins to train the model.

The control database and the test set

The control database consists of a non-redundant set of 137 proteins representing some of the known folds. [PDB names (Bernstein et al., 1977) of the control database polypeptide chains, suffixed with the chain identifiers: 1ace, 1acx, 1ak3A, 1azu, 1bbpA, 1bds, 1bmv1, 1bmv2, 1cc5, 1cd4, 1cdtA, 1colA, 1cox, 1crn, 1cseI, 1eca, 1etu, 1f3g, 1fc2C, 1fdlH, 1fdx, 1fkf, 1fxiA, 1gd1O, 1gly, 1gmfA, 1gp1A, 1hddC, 1hip, 1hrhA, 1l58, 1lap, 1lib, 1mcpL, 1msbA, 1nsbA, 1ovoA, 1paz, 1pi2, 1pyp, 1r092, 1rbp, 1rhd, 1s01, 1sdhA, 1sh1, 1tgsI, 1tnfA, 1ubq, 1wsyA, 1wsyB, 256bA, 2aat, 2alp, 2cab, 2ccyA, 2cpkE, 2cyp, 2fnr, 2fxb, 2gbp, 2gcr, 2gn5, 2hipA, 2hmzA, 2i1b, 2ifb, 2lh4, 2lhb, 2ltnA, 2ltnB, 2mev4, 2or1L, 2pabA, 2pcy, 2phh, 2pk4, 2rspA, 2sarA, 2scpA, 2sns, 2stv, 2tgpI, 2tmvP, 2tscA, 2utgA, 2wrpR, 3ait, 3b5c, 3blm, 3cla, 3cln, 3fgf, 3gapA, 3hmgA, 3hmgB, 3icb, 3pgm, 3rnt, 3timA, 4bp2, 4cms, 4cpv, 4fxn, 4gr1, 4pfk, 4rhv1, 4rhv3, 4rhv4, 4sgbI, 4ts1A, 4xiaA, 5cytR, 5er2E, 5hvpA, 5ldh, 5lyz, 5p21, 6acn, 6cpa, 6cpp, 6cts, 6dfr, 6tmnE, 7catA, 7icd, 7rsa, 8abp, 8adh, 9apiA, 9apiB, 9pap, 9wgaA, 1ak1, 1exf, 1sro, 3dhq. The last 4 proteins in the control database list were target proteins for fold recognition for CASP2. The structure of 3dhq is not yet available in PDB.] The majority of them (81%) are a subset of a non-redundant database of 125 proteins collected by Rost and Sander (1993). Additional protein sequences have been added to it such that the pairwise sequence identity is <25%.

Forty-four proteins from this data set were used for fold recognition belonging to 14 different CATH topologies. Sixteen

proteins belonged to the Mainly-Alpha class, 15 to the Mainly-Beta class and 13 to the Alpha-Beta class. For each test protein, cross-validated training was performed, thereby obtaining 44 models. Each of the 44 proteins was then scored by each of the 44 cross-validated HMMs and ranked appropriately.

Jack-knife cross-validated secondary structure predictions for the proteins in the control database, including the test proteins, were obtained using the algorithm QL (Munson *et al*., 1994) and homologous sequences with an average prediction accuracy Q_3 = 68% (Q_3 is the percentage of correctly predicted residues in the three conformations).

Results

Fold recognition

Fold recognition of a test protein was considered successful when its predicted secondary structure sequence attained lower rank with the cross-validated HMM of its true fold than with other HMMs. For each model, the ranking of the test protein is in comparison to the scores achieved by the predicted sequences of proteins in the control database having different topologies. The summary of the fold recognition results is presented in Table 1. The fold topology was correctly identified for 26 of 44 proteins (59%). Another two (2ccyA and 5cytR) of the 44 obtained the same rank by the HMM of their true topology and by the HMM of another fold (± in Table 1). Including these two, the overall success rate becomes 64%. The sequence 2ccyA obtained rank 1 by the cross-validated models of its own fold, but also by the five cross-validated models of the globins in the test set (data not shown). All the proteins whose fold was considered to be correctly recognized obtained scores having ranks lower than 10, and 24 of them, including 2ccyA and 5cytR, obtained rank = 5.

The fold recognition success rates for each class, Mainly-Alpha, Mainly-Beta and Alpha-Beta, were 63% (75% if one also includes the ± cases), 40% and 77%, respectively. Mainly-Beta protein folds are not as easily placed into the correct fold families as the folds in the two other classes. The average Q_3 accuracy measure for secondary structure prediction of proteins in the Mainly-Alpha and Alpha-Beta classes is 76.6 and 70.7%, respectively, while the average Q_3 for Mainly-Beta proteins is only 66.1%, ~4.5 standard deviations lower than for Mainly-Alpha (Table 1). There are five proteins whose incorrect predicted topology belongs to the wrong structural class: the Mainly-Alpha proteins 5cytR and 2tmvP, and the proteins in the Alpha-Beta class (4xiaA, 8adh, 2fxb). Perhaps the most surprising cases are the helical proteins 5cytR and 2tmvP, which were both predicted to have a topology in the Mainly-Beta class. Contrary to what one would expect, this is not due to a high predicted content of β-strand residues for those two sequences. In fact, for example, the fraction of predicted helix and strand residues of 5cytR is 31 and 11%, respectively. The wrong class prediction of 5cytR is mainly due to long predicted coil and short predicted β-strand regions that match coil and β-strand regions in the consensus structure of the predicted immunoglobulin-like topology. Insert regions are found in the consensus structure of immunoglobulin-like proteins where helices are predicted in the 5cytR sequence.

Table 1. Results of fold recognition of the test proteins

Class and protein topology[a]	Protein name	Q_3 (%)[b]	Topology recognition[c]	Correct topology model rank[d]	Wrong topology prediction[e]
Mainly-Alpha					
Globin-like	1eca	73.5	+	6	
	2lhb	71.8	–	3	Granulocyte colony-stimulating factor
	1sdh	71.2	+	2	
	2lh4	69.9	–	3	Granulocyte colony-stimulating factor
	1colA	71.6	+	3	
Cytochrome C550	1cc5	84.3	+	1	
	5cytR	68.0	±	5	Immunoglobulin-like
Four-helix bundle [Hemerythrin (Met), subunit A]	2hmzA	81.6	+	2	
	256bA	87.7	+	1	
	2ccyA	85.8	±	1	Globin-like OB folds
	2tmvP	66.9	–	>20	
EF-Hand	4cpv	75.9	+	6	
	3icb	93.3	+	1	
	3cln	86.7	+	1	
	2scpA	56.9	–	>20	None

Table 1. *continued*

Class and protein topology[a]	Protein name	Q_3 (%)[b]	Topology recognition[c]	Correct topology model rank[d]	Wrong topology prediction[e]
Granulocyte colony-stimulating factor (Form II RCG-CSFII) subunit A	1gmfA	80.7	+	1	
Mainly-Beta					
Lipocalin (Streptavidin, subunit A)	1bbpA	70.5	+	1	
	1rbp	62.1	–	>20	Immunoglobulin-like
	1lib	52.7	–	>20	None
	2ifb	55	–	>20	Immunoglobulin-like
Elongation factor TU, domain 3	2alp	63.6	–	4	Immunoglobulin-like
Immunoglobulin-like, 1 domain[f]	2pcy	76.8	–	>20	OB folds
	1paz	78.3	–	15	Lipocalin (Streptavidin, subunit A)
	1azu	67.7	–	>20	OB folds
	2pabA	54.4	–	>20	None
	1acx	69.2	+	5	
	3ait	68.9	+	5	
Immunoglobulin-like, 2 domain[f]	3cd4	65.9	+	4	
	1fdlH	70.6	+	1	
	1mcpL	63.6	+	3	
OB folds	2sns	73.0	–	4	Immunoglobulin-like
Alpha-Beta					
Flavoprotein 390, subunit A	3timA	79.5	+	5	
	4xiaA	73.8	–	8	Globin-like
	1wsyA	83.2	+	2	
Nitrogenase molybdenum-iron protein, subunit A, domain 3. 1 domain[f]	1etu	64.4	+	1	
	4fxn	71.7	+	8	
	5p21	69.2	+	2	
Nitrogenase molybdenum-iron protein, subunit A, domain 3, 2 domain[f]	8abp	64.3	+	3	
	2gbp	70.2	+	3	
	5ldh	63.1	+	10	
	8adh	64.4	–	18	Elongation factor TU, domain 3
	4pfk	74.0	+	2	
	1gd1O	65.9	+	5	
Crambin	2fxb	75.3	–	6	OB folds
	Average Q_3	71.3			

[a]Topology names correspond to version 1.1 of CATH (Orengo *et al*., 1993, 1994).
[b]Correctly predicted residues in three states, helix, strand and coil, after cross-validation with QL method and evolutionary information when available. Q_3 is calculated with respect to the DSSP assignments of observed secondary structures (Kabsch and Sander, 1983).
[c]+ indicates correct fold recognition, – indicates wrong fold recognition. ± indicates ambiguous cases in which the test protein was assigned identical, low ranks by the model of the correct topology and by a model of the wrong topology. Of a total of 44 test proteins, the fold topology was correctly identified for 26 (59%). If one also considers the two ambiguous cases as correct, the percentage of correctly identified folds is 64%.
[d]These are ranks of the NLO score assigned by the cross-validated model of the correct topology of the test protein. For each HMM, the rank of a test protein is calculated after removal of the scores assigned to the true positives for that specific topology model in the control database. For example, rank 5 means that four proteins of the control database were ranked higher and they all were false positives.
[e]This column contains the wrongly predicted topology name for the test protein in those cases in which the fold is not correctly recognized (–) or when the fold prediction is ambiguous (±). 'None' indicates that the test protein was not predicted to have any fold since it did not achieve a rank >20 with any of the models.
[f]The CATH topology group was divided into two parts, one containing one-domain protein sequences and the other containing two-domains, which may or may not belong to the same topology group.

In general, cross-validated training of the HMM increases the ranking of the scores of the test proteins by one or two positions. However, there are cases, such as the lipocalin, immunoglobulin-like (1 domain), and flavoprotein 390 HMMs, for which cross-validated training considerably increases the ranks of the scores of certain proteins. For instance, the lipocalin proteins 1rbp, 1lib and 2ifb had a rank of 1, 6 and 1, respectively (data not shown), while, with cross-validated training, the three proteins have rank >20 (Table 1). The immunoglobulin-like proteins 2pcy and 1paz had ranks 2 and 1, respectively, when the members of their homology family were not removed from the training set of immunoglobilin-like HMM (compare with Table 1). For two other test proteins, 4cpv and 2sns, their ranks went up from 1 to 6 and 4, respectively.

The FORESST technique is here used to identify proteins in the control database that are structurally similar, but have low sequence identity with the test protein. We wished to assess whether sequence homology detection tools used to search the control database would also associate the same proteins from the structural point of view. Using each of the 44 test proteins, both BLAST (Altschul *et al*., 1990) and SSEARCH [Smith and Waterman algorithm from the FASTA package (Pearson, 1996)] searches of the control database were performed, seeking significant homologies between the query sequence and the proteins in the control database. If a significant similarity is found, one can immediately associate a protein fold to the test sequence. Note that, according to the CATH database, only the 44 test proteins in the control database belong to the 14 structural topologies represented by the HMMs. The remaining proteins in the control database belong to topologies that are different from those represented by the HMMs used here. For 33 test proteins, no significant similarity (cut-off value: $P < 0.01$) with any other protein in the control database could be found with either method. The pairwise sequence alignment scores were significant for only 25% of the test proteins. For four of these proteins, 1lib, 2fib, 2pcy and 1paz, the fold was also not correctly recognized with the HMMs. However, sequence homology detection tools could not identify topologically similar proteins in the control database as well as the HMMs.

Results of blind experiments of fold recognition

Fold recognition predictions prior to the knowledge of the structural data for a test protein are of particular interest since they correspond to real experimental conditions. Here we briefly discuss the results of fold recognition predictions for seven additional proteins whose structure was unknown at the time of the prediction. Two target sequences were those of interleukin-6 (hIL6) and leptin, the mouse obese gene product; five others were prediction targets for the second Critical Assessment of techniques for protein Structure Prediction (CASP2), held at Asilomar, CA, in December 1996. For those proteins, the predicted secondary structures were obtained using several programs: QL (Munson *et al*., 1994), GOR-IV (Garnier *et al*., 1996), SIMPA (Levin and Garnier, 1988) (Version 96), PREDATOR (Frishman and Argos, 1996) and PHD (Rost and Sander, 1993) with evolutionary information when available. The use of several prediction algorithms increases the chance of having a predicted sequence of secondary structure elements with a smaller number of mistakes. Moreover, since these proteins did not have detectable sequence similarity to any protein of known fold, this search for the correct topology used models that were not cross-validated, in order to increase the set of possible structures represented by each model.

Both leptin and hIL6 were predicted to have the helical cytokine topology (Di Francesco *et al*., 1997a). The recently published NMR structure of leptin (Kline *et al*., 1997) reveals that its topology is indeed similar to that of four-helix-bundle short-chain cytokines. For hIL6, an NMR structure (Xu *et al*., 1996) and an X-ray structure at 1.9 Å resolution (Somers *et al*., 1997) have now been published. Both experimental structures indicate the four-helix bundle of the long-chain cytokine topology, as was anticipated by our studies.

The CASP2 target proteins were polyribonucleotide nucleotidyltransferase (T0004, 1sro), 3-dehydroquinase (T0014, 3dhq), ferrochelatase (T0020, 1ak1), elongation factor TU, domain 3 fold for exfoliative toxin A (T0031, 1exf) and β-cryptogein (T0032, 1beo). Three fold predictions were correct: two Mainly-Beta fold types—an OB fold for polyribonucleotide nucleotidyltransferase and an elongation factor TU, domain 3 fold for exfoliative toxin A; and an Alpha-Beta fold—a TIM Barrel for 3-dehydroquinase. A fourth prediction for ferrochelatase would have been correct if we had an HMM for the topology of the two-domain nitrogenase molybdenum-iron proteins (NMIP), subunit A, domain 3. At that time, we only had a model for one-domain proteins of that topology. The fifth prediction for β-cryptogein was wrongly attributed to the Mainly-Alpha fold phospholipase A2, whereas the experimental X-ray structure revealed it to be a unique previously unobserved fold.

For the three targets whose fold was correctly predicted, a comparison was made by the assessors of CASP2 between the HMM alignments and the structure-based alignments of the experimental structures of the target proteins with the PDB structure. This comparison showed that the percentage of correctly aligned residues (ASns) ranged between 16 and 29%, and the mean alignment shift error (Shft) was between 1.4 and 6.2 residues. Further details of these predictions are reported elsewhere (Di Francesco *et al*., 1997b) and a direct comparison of the performance of this method with that of other fold recognition approaches can be found in the proceedings of CASP2 (Levitt, 1997; Marchler-Bauer *et al*., 1997).

Discussion

Hidden Markov models provide a stochastic description of the consensus structure and the structural variations, such as extra secondary structure elements and differences in chain length, of proteins that have the same overall topology. The first-order Markov chains of the hidden states imply that insertions, deletions or matches depend only on the preceding hidden state. This is certainly not adequate to describe real protein structures fully, but these models serve only as a first approximation of protein folds which may be improved by, for example, encoding a minimum length for regions with match states and secondary structure elements of some specified type. The fold recognition accuracy of the models and the alignment quality may be enhanced by taking into account the secondary structure prediction probabilities and other biochemical aspects of the query sequence, such as the hydrophobicity profile. In Di Francesco *et al*. (1997c), another use of the HMMs of protein folds was shown which consists of refining the secondary structure prediction of a query sequence with the HMM of its predicted structural topology. Since the HMMs describe protein folds, they effectively incorporate global structural information that can be used in the secondary structure prediction scheme. The use of the HMMs of the predicted folds to modify QL predicted sequences provided an improvement in prediction accuracy Q_3 of 3%.

The motivation for training HMMs of protein folds at the topology level of the CATH structural classification, rather than at the homology superfamily level at which the proteins are evolutionarily related, followed from the need for a sufficient number of proteins for the HMM training sets and the need to perform model validation. Cross-validated training required the elimination of proteins with significant similarity to the test protein from the model training set. Often, an entire homology family for the test protein was removed, with the consequence that the reduced training set did not contain the protein structures most similar to the query protein. This cross-validation procedure is, therefore, a more severe test of the method than in actual applications where the HMM is trained with a set of sequences representing all the known structures of proteins belonging to a certain topology level of CATH.

In prospective blind predictions discussed here, those for leptin, interleukin-6 and the five CASP2 proteins, several predicted secondary structure sequences were obtained with various algorithms for each query sequence, thereby increasing the likelihood of having a better quality prediction. It was shown previously (Di Francesco *et al*., 1997a), and additional evidence is available (data not shown), that the recognition capabilities of the models would increase considerably if one could use experimentally derived secondary structure sequences, such as those obtained at the early stages of structure determination NMR protocols. It was out of the scope of the present work to investigate the effect of the various secondary structure prediction algorithms on the performance of this fold recognition approach. It should be observed that the fold recognition success rate obtained here by this approach is also likely to be affected as a more extensive library of HMMs of structural families becomes available, since the likelihood of having false positives and negatives will be higher. A more extensive library of HMMs, representative of a larger set of the known protein structural families, is currently being built. It will be used to calculate the effect of a larger number of HMMs on the method's performance and will soon be available (work in progress) to the scientific community on the Internet at http://www.tigr.org/.

The choice of the proteins for model training proved to be crucial in those cases in which proteins have multiple structural domains. An illustrative example is shown in Figure 2, where it can be observed that some proteins having the subunit A topology of NMIP received poor scores by the two-domain model of that topology, even using observed sequences. Indeed the best scores corresponding to the lowest three ranks are assigned to two-domain proteins, while the worst two scores (Figure 2a, filled bars, rank 29 and 59) are in fact of the one-domain proteins 5p21 and 4fxn. The remaining one-domain protein having this topology in the test set, 1etu, obtained a score that was ranked in the lowest 10. On the other hand, when training the model for one-domain proteins in this topology (data not shown), 5p21, 1etu and 4fxn were assigned ranks 1, 1 and 6, respectively, with predicted sequences. Multi-domain proteins had rank 14 or higher with the one-domain model of NMIP. There are also two-domain proteins receiving poor scores by the two-domain model. The domains of those proteins typically consist of non-adjacent segments of the protein chain, such as when the N and the C termini together form the structural domain of the topology of interest, e.g. proteins 8adh or 1gd1O. It has been observed (Hughey and Krogh, 1996; Barrett *et al*., 1997) that there is a linear dependence of the HMM scores on the protein sequence length, whose effect is more evident when scoring short sequences with long models (such is the case of one-domain proteins scored by two-domain models), where model length is defined as the number of match positions in the model. In order to skip match positions, short sequences use several delete states that do not emit secondary structure states and thereby give rise to poor scores, and therefore false negatives.

The results of Table 1 are obtained with cross-validated predicted secondary structures. It is noticeable that Mainly-Alpha and Alpha-Beta proteins have a higher fold recognition success rate, 63% (or 75%, including the ± cases) and 77%, than Mainly-Beta proteins at 40%. One might expect Alpha-Beta proteins to carry more information compared to Mainly-Alpha and Mainly-Beta proteins, because the sec-

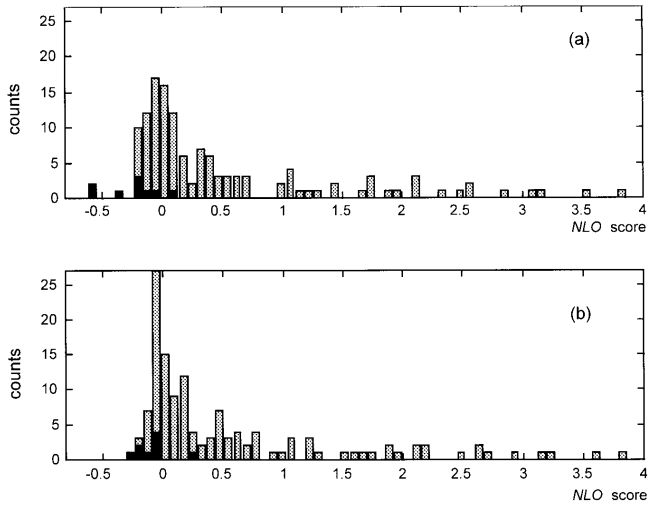

Fig. 2. Histograms of NLO scores assigned to the 137 proteins in the control database by the un-cross-validated HMM trained on proteins with the two-domain nitrogenase molybdenum-iron proteins (NMIP). Filled bars show the scores of proteins with the NMIP topology (true positives), including one-domain and multi-domain proteins. The scores of the proteins in the control database not having any structural domain in the NMIP topology (true negatives) are indicated by hashed bars. (**a**) The score distributions for observed secondary structures. (**b**) The score distributions for predicted secondary structures. The distribution of scores obtained with predicted secondary structure sequences is narrower than that with observed sequences and the true positives (in filled bars) are not as easily distinguished from other proteins as when observed sequences are used. The distribution of scores assigned to the control database of observed secondary structure sequences provides an indication of how well the model of two-domain NMIP topology distinguishes true positives from true negatives. The distribution of scores for the predicted sequences, which corresponds to more realistic experimental conditions, shows the degrading effect of errors made during secondary structure prediction. Even with observed structure sequences, though, the NMIP model does not fit some NMIP protein sequences well, as indicated by less negative, and in one case positive, scores. (A positive score indicates that the model fits less well than the null model.) This deficiency is due mainly to the fact that one-domain proteins having the NMIP topology were considered as true positives for the two-domain NMIP model.

ondary structure sequence of proteins in the latter two groups mostly consists of only two letters ({H,C} or {E,C}, respectively) instead of a sequence of three letters. This expectation is realized at least with respect to the Mainly-Beta class. However, the fold recognition prediction accuracy for each class appears to be correlated to the secondary structure prediction accuracy. It is well known that secondary structure prediction algorithms do not predict β-strand residues as well as the residues having the helical or coil conformation, which is probably due to the non-locality of residue contacts and hydrogen bonds formed in β sheets.

In general, it is difficult to compare success rates for fold recognition between various methodologies: for example, some methods use gapped while others use ungapped threading, there are often differences in the test sets used and in the reported accuracy measures. Moreover, the threading analysis presents some inherent difficulties, such as, first, the existence of many acceptable structural models for a query sequence and, second, many possible sequence-to-structure alignments (Lemer *et al*., 1995; Levitt, 1997). Prediction experiments such as CASP2 are well suited to comparison of the performances of different methods, since the various approaches are tested on the same set of proteins and the same accuracy measures are applied for evaluation by independent assessors. In that experimental context, this approach was found to have sustained an overall good performance in both fold recognition and in threading accuracy (Levitt, 1997; Marchler-Bauer *et al*., 1997). This result is a substantial indication of the potential of this approach relative to other approaches. However, it should be noticed that in CASP2 this approach has been used together with additional information about the target sequences deriving from other sources, such as from the literature or from biological insights of the prediction team members (Di Francesco *et al*., 1997b). Therefore, the fold recognition success rate obtained in CASP2 does not entirely reflect the inherent capabilities of this approach. Since the set of target proteins was too small (only seven) to constitute a statistically representative sample, more blind prediction experiments of the CASP type will allow for more definitive comparisons of various approaches to fold recognition.

To conclude, helices and strands have been commonly recognized as important protein structural features. Their arrangement in space (Richardson, 1981) and their connectivity (Robson and Garnier, 1986) allow us to describe and classify their 3D fold (Orengo *et al*., 1993; Murzin *et al*., 1995). As some secondary structure prediction schemes currently achieve an average Q_3 of 65–75%, it has been shown here, for a variety of folds from three different structural classes, that indeed enough information is embedded into the length and nature of the predicted secondary structures along the amino acid sequence to characterize a fold topology with an accuracy of 60–70%.

Because of its limited requirement for CPU time, FORESST is suitable for screening large genomic databases. It may be used as a complement to sequence-based tools when analyzing genomic sequences and attempting to assign a biological role to hypothetical proteins. In fact, for 59% of the proteins in the test set, this approach was capable of detecting fold similarity among protein sequences in the control database, while only 25% of those proteins could be found to be structurally similar using sensitive algorithms such as

BLASTP and SSEARCH. This approach should also benefit from future improvements in secondary structure predictions and possibly of higher order Markov chains.

Acknowledgements

We would like to thank Dr Geetha Vasudevan for enlightening discussions and assistance with the SSEARCH and BLAST searches of the query proteins against the control database. We also thank Mr Will Cushing, who developed the NIH Web server for FORESST.

References

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Barrett,C., Hughey,R. and Karplus,K. (1997) Scoring hidden Markov models. *Comput. Applic. Biosci*., **13**, 191–199.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanovichi,T. and Tasumi,M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol*., **112**, 535–542.

Browne,W.J., North,A.C.T., Phillips,D.C., Brew,K., Vanaman,T.C. and Hill,R.L. (1969) A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme. *J. Mol. Biol*., **42**, 65–86.

Bryant,S.H. (1996) Evaluation of threading specificity and accuracy. *Proteins: Struct. Funct. Genet*., **26**, 172–185.

Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J*., **5**, 823–826.

Di Francesco,V., Garnier,J. and Munson,P.J. (1997a) Protein topology recognition from secondary structure sequences—Application of the hidden Markov models to the alpha class proteins. *J. Mol. Biol*., **267**, 446–463.

Di Francesco,V., Geetha,V., Garnier,J. and Munson,P.J. (1997b) Fold recognition using predicted secondary structure sequences and hidden Markov models of protein folds. *Proteins: Struct. Funct. Genet. Suppl*., **1**, 123–128.

Di Francesco,V., McQueen,P., Garnier,J. and Munson,P.J. (1997c) Incorporating global information into secondary structure prediction with hidden Markov models of protein folds. In *Proceedings of the 5th International Conference on Intelligent Systems in Molecular Biology*, pp. 100–103. AAAI Press, Menlo Park, CA.

Doolittle,R.F. (1992) Stein and Moore Award address. Reconstructing history with amino acid sequences. *Protein Sci*., **1**, 191–200.

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

Frishman,D. and Argos,P. (1996) Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng*., **9**, 133–142.

Garnier,J., Gibrat,J.-F. and Robson,B. (1996) GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol*., **266**, 540–553.

Greer,J. (1991) Comparative modeling in homologous proteins. *Methods Enzymol*., **202**, 239–252.

Hogg,R.V. and Craig,A.T. (1978) *Introduction to Mathematical Statistics*. Macmillan, New York.

Hubbard,T.J. and Park,J. (1995) Fold recognition and ab initio structure predictions using hidden Markov models and β-strand pair potential. *Proteins: Struct. Funct. Genet*., **23**, 398–402.

Hughey,R. and Krogh,A. (1995) SAM: Sequence alignment and modeling software system. Technical Report, UCSC-CRL-95-7. University of California, Santa Cruz, CA.

Jones,D. and Thornton,J. (1993) Protein fold recognition. *J. Comput. Aided Mol. Des*., **7**, 439–456.

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

Kline,A.D. *et al*. (1997) Leptin is a four-helix bundle: secondary structure by NMR. *FEBS Lett*., **407**, 239–242.

Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D. (1994) Hidden Markov models in computational biology: Applications to protein modeling. *J. Mol. Biol*., **235**, 1501–1531.

Lemer,C.M.-R., Rooman,M.J. and Wodak,S.J. (1995) Protein structure prediction by threading methods: evaluation of current techniques. *Proteins: Struct. Funct. Genet*., **23**, 337–355.

Levin,J.M. and Garnier,J. (1988) Improvements in secondary structure prediction method based on search for local sequence homologies and its use as a model building tool. *Biochim. Biophys. Acta*, **955**, 283–295.

Levitt,M. (1997) Competitive assessment of protein fold recognition and alignment accuracy. *Proteins: Struct. Funct. Genet. Suppl*., **1**, 92–104.

Marchler-Bauer,A., Levitt,M. and Bryant,S.H. (1997) A retrospective analysis of CASP2 threading predictions. *Proteins: Struct. Funct. Genet. Suppl*., **1**, 83–91.

Munson,P.J., Di Francesco,V. and Porrelli,R.N. (1994) Protein secondary structure prediction using periodic-quadratic-logistic models: statistical and technical issues. In *Proceedings of the 27th Hawaii Int. Conf. Sys. Sci*. IEEE Press, Los Alamitos, CA, Vol. V, pp. 375–384.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*., **247**, 536–540.

Orengo,C.A., Flores,T.P., Taylor,W.R. and Thornton,J.M. (1993) Identification and classification of protein fold families. *Protein Eng*., **6**, 485–500.

Orengo,C.A., Jones,D.T. and Thornton,J.M. (1994) Protein superfamilies and domain superfolds. *Nature*, **372**, 631–634.

Pearson,W.R. (1996) Effective protein sequence comparison. *Methods Enzymol*., **266**, 227–258.

Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.

Rice,D.W. and Eisenberg,D. (1997) A 3D-1D substitution matrix for protein fold recognition that included predicted secondary structure of the sequence. *J. Mol. Biol*., **267**, 1026–1038.

Richardson,J.S. (1981) The anatomy and taxonomy of protein structure. *Adv. Protein Chem*., **34**, 167–339.

Robson,B. and Garnier,J. (1986) *Introduction to Proteins and Protein Engineering*. Elsevier Science, Amsterdam.

Rost,B. (1995) TOPITS: Threading One-dimensional Predictions Into Three-dimensional Structures. In *Proceedings of the 3rd International Conference on Intelligent Systems in Molecular Biology*, pp. 314–321. AAAI Press, Menlo Park, CA.

Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol*., **232**, 584–599.

Russell,R.B., Copley,R.R. and Barton,G.J. (1996) Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol*., **259**, 349–365.

Schneider,R. and Sander,C. (1991) Database of homology derived structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet*., **9**, 56–68.

Sheridan,R.P., Dixon,J.S. and Venkataraghavan,R. (1985) Generating plausible folds by secondary structure similarity. *Int. J. Peptide Protein Res*., **25**, 132–143.

Somers,W., Stahl,M. and Seehra,J.S. (1997) 1.9 Å crystal structure of interleukin 6: implications for a novel mode of receptor dimerization and signaling. *EMBO J*., **16**, 981–997.

Xu,G.Y., Hong,J., McDonagh,T., Sathl,M., Kay,L.E., Seehra,J. and Cumming,D.A. (1996) Complete 1H, 15N and 13C assignments, secondary structure, and topology of recombinant human interleukin-6. *J. Biomol. NMR*, **8**, 123–135.