Protein Engineering vol.12 no.7 pp.527–534, 1999

Comparing protein sequence-based and predicted secondary structure-based methods for identification of remote homologs

V.Geetha[1,2], Valentina Di Francesco[3], Jean Garnier[1,4] and Peter J.Munson[1]

[1]ABS/MSCL/CIT, National Institutes of Health, Bethesda, MD 20892, [3]The Institute for Genomic Research, Rockville, MD 20850, USA and [4]Laboratoire de Biologie Cellulaire et Moleculaire, Biotechnologies, INRA, 78 352 JOUY-en-JOSAS, Cedex, France

[2]To whom correspondence should be addressed

We have compared a novel sequence–structure matching technique, FORESST, for detecting remote homologs to three existing sequence based methods, including local amino acid sequence similarity by BLASTP, hidden Markov models (HMMs) of sequences of protein families using SAM, HMMs based on sequence motifs identified using meta-MEME. FORESST compares predicted secondary structures to a library of structural families of proteins, using HMMs. Altogether 45 proteins from nine structural families in the database CATH were used in a cross-validated test of the fold assignment accuracy of each method. Local sequence similarity of a query sequence to a protein family is measured by the highest segment pair (HSP) score. Each of the HMM-based approaches (FORESST, MEME, amino acid sequence-based HMM) yielded log-odds score for the query sequence. In order to make a fair comparison among these methods, the scores for each method were converted to Z-scores in a uniform way by comparing the raw scores of a query protein with the corresponding scores for a set of unrelated proteins. Z-Scores were analyzed as a function of the maximum pairwise sequence identity (MPSID) of the query sequence to sequences used in training the model. For MPSID above 20%, the Z-scores increase linearly with MPSID for the sequence-based methods but remain roughly constant for FORESST. Below 15%, average Z-scores are close to zero for the sequence-based methods, whereas the FORESST method yielded average Z-scores of 1.8 and 1.1, using observed and predicted secondary structures, respectively. This demonstrates the advantage of the sequence–structure method for detecting remote homologs.

Keywords: hidden Markov models/motifs/remote homologs/secondary structures

Introduction

Protein sequences emerging from genome sequencing projects are of greatest value to medicine and biology if their structure and function can be identified. With the growing number of unannotated sequences, association of a new sequence to a protein of known structure can be a significant step towards the identification of its biological role. Simple sequence search methods such as FASTA (Pearson and Lipman, 1988) or BLASTP (Altschul *et al*., 1990) readily identify close homologs of protein sequences, whereas sequences of remote homologs have diverged so much, it may be difficult to detect their relationship.

Hidden Markov models (HMM) are sequence-family based techniques which are extensively used in modeling protein families (Hughey and Krogh, 1995). HMMs were found to perform better compared with sequence-based methods like WU-BLASTP for detecting remote homologs (Karplus *et al*., 1998). The performance of different sequence search techniques for identifying homologs has been assessed recently (Pearson, 1995; Agarwal and States, 1998; Brenner *et al*., 1998; Grundy, 1998; Levitt and Gerstein, 1998). In a comprehensive study based on the protein domains of the SCOP database, Levitt and Gerstein (1998) found that structure comparison methods are able to detect twice as many distantly related proteins as sequence comparison methods, at the same error rate. Pairwise sequence comparison methods such as SSEARCH and FASTA detected almost all relationships between proteins whose sequence identities were above 30%, but detected only half of the relationships when pairwise sequence identity is between 20 and 30% (Brenner *et al*., 1998).

Significant pairwise sequence similarity is not a necessary condition for pairs of proteins to adopt a common fold. Indeed it is widely accepted that three-dimensional structure is better conserved than sequence, as is evident from the current classifications of proteins into similar topology and architecture (Murzin *et al*., 1995; Orengo *et al*., 1997). Proteins with sequence identity down to about 30% generally share the same fold (Chothia and Lesk, 1986; Sander and Schneider, 1991; Flores *et al*., 1993), and surprisingly, even proteins with as low as 5% sequence identity have the same fold (Orengo *et al*., 1993). The number of different protein folds was identified to be 327 from the June 1996 release of the Protein Data Bank (Chothia *et al*., 1997). It now seems feasible to choose a fold compatible with a novel protein sequence from the growing repertoire of known folds, rather than attempting to predict the structure *de novo*. Fold assignment is thus becoming a practical approach to protein structure prediction and recent reviews (Lemer *et al*., 1995; Levitt, 1997) suggests that progress in this area is continuing.

We are interested in comparing methods intended to relate new sequences to known structures. The methods studied here include local sequence similarity search by BLASTP (Altschul *et al*., 1990), HMMs constructed from sequences of protein families using the publically available software SAM (Hughey and Krogh, 1995), HMMs based on identified motifs from sequences using meta-MEME (Grundy *et al*., 1997) and FORESST (which stands for FOld REcognition from Secondary Structure), a method based on HMMs built from secondary structure sequences of proteins (Di Francesco *et al*., 1997a). The HMM based methods require many parameters to be estimated. It is therefore essential to measure their cross-validated performance so that simple ‘memorization’ of the training data is excluded as a possibility.

Materials and methods

Fold families of proteins and cross-validated testing

A sample of nine fold families (Table I) were selected from the database of protein structure classification, CATH release

Table I. Protein families and proteins tested for fold assignment

Fold family	Family members from CATH used for training HMMs	Protein tested	MPSID[a]	Proteins removed from training HMMs
Phospholipase	1poa,1pod,3p2pA,2phiA,	1poc	17	1poc
	1pshA,4p2p,5p2pA,	1pod	55	1pod
	1pp2L,1ppa,1poc	1pp2L	47	1pp2L
		1ppa	47	1ppa
Globin-like	1eca,1mbd,1myt,1mbs,	1colA	16	1colA
	1ymb,1myiA,2mm1,	1eca	17	1eca,1hbg,1hdsA,1mba,
	3sdhA,1hbg,2lhb,1pbxB,			2lhb,1hdsA,1hdsB,1fdhG,
	1hdsB,1fdhG,2mhbB,			2mhbA,1thbA,1mba
	1hbsB,1pbxA,1hdsA,	1lh1	18	1lh1
	2mhbA,1thbA,1mba,1lh1,	2lhb	21	2lhb,1hdsA,1hdsB,1fdhG,
	1ithA,1cpcA,1cpcB,1colA			2mhbA
		3sdhA	20	3sdhA
Cytochrome-C	1ycc,1ccr,5cytR,1yea,	1cc5	18	1cc5,3c2c,155c
	1ctz,1raq,2ycc,1crg,1cri,	1ycc	22	1ycc,1ccr,5cytR,1yea,1ctz,
	1crj,1cty,1rap,1cyc,			1raq,2ycc,1crg,1cri,1crj,
	2pcbB,351c,1cor,155c,			1cty,1rap,1cyc,2pcbB
	1c2rA,2mtaC,3c2c,1cc5	5cytR	30	1ycc,1ccr,5cytR,1yea,1ctz,
				1raq,2ycc,1crg,1cri,1crj,
				1cty,1rap,1cyc,2pcbB,155c
EF Hand	3cln,1ncx,5tnc,1tnx,	1bod	29	1bod
	1rro,4cpv,5pal,1pal,	2scpA	21	2scpA
	1osa,2pas,2sas,2scpA,	3cln	25	3cln,1ncx,5tnc,1tnx,1osa
	1bod	4cpv	23	4cpv,1rro,5pal,1pal,1osa,
				2pas,1bod
Cytochrome B562	2hmzA,2ccyA,256bA,	256bA	22	256bA,1apc
	2tmvP,1apc,2hmqB,	2hmzA	13	2hmzA,2hmqB,2mhr,
	2mhr,1hmdB,1hmoB,			1hmdB,1hmoB
	1lpe,1le4,1le2,1bbhA,	2ccyA	19	2ccyA
	1aep,1vtmP	2tmvP	14	2tmvP,1vtmP
α/β Hydrolase-lipase topology	1thtA,1cvl,1gpl,1tca,	1thtA	10	1thtA
	1oilA,1ethA,1ede,	1cvl	11	1cvl,1oilA
	1hdeE,1ysc	1gpl	12	1gpl
		1tca	10	1tca
OB fold-dihydrolipo-amide acetyl	1lab,1bdo,1bovA,1afp,	1lab	21	1lab
transferase topology	1mjc,1krs,1rip,1vqb,	1bdo	21	1bdo
	1prtD,1ltsD,2sns,1csp,	1bovA	11	1bovA
	1chbD,1snc	1afp	10	1afp
		1mjc	13	1mjc,1csp
		1krs	10	1krs
		1rip	11	1rip
		1vqb	12	1vqb
		1prtD	11	1prtD
		1ltsD	10	1ltsD,1chbD
		2sns	13	2sns,1snc
Serine protease elongation	1sgt,1thsH,5ptp,3rp2A,	1sgt	25	1sgt,1thsH,5ptp,1etsH,
factor Tu-domain 3 topology	1etsH,1ppfE,4chA,1mctA,			1mctA,1tld,4etsE
	1tld,1trmA,1ton,1gctA,	1thsH	27	1thsH,1etsH,1sgt,5ptp,
	4estE,1hneE			1mctA,4chaA,1tld,1trmA,
				1gctA
		5ptp	41	5ptp,1trmA,1tld,1mctA
		3rp2A	32	3rp2A
		1etsH	37	1etsH,1thsH
Interleukin granulocyte colony	2gmfA,1itl,3inkC,1lki,	2gmfA	15	2gmfA
stimulating factor (form II)	3hhrA,1bgc,1rhgA,1bge,	1itl	13	1itl
	1ilk,1rfbA,1hmcA,1higA,	3inkC	9	3inkC
	1rni	1lki	13	1lki
		3hhrA	14	3hhrA

[a]MPSID is the maximum pairwise sequence identity between the query protein and the rest of the fold family.

1997 (Orengo *et al*., 1997). Each family was required to have sufficient size and diversity to be suitable for cross-validated testing of each recognition algorithm.

The pairwise sequence identity, according to CLUSTALW (Thompson *et al*., 1994) between any two members within the topology family was used as a proxy measure of relatedness. This approach allowed us to perform cross-validated testing by eliminating from the training set all proteins having more than a specified degree of relatedness (Table I). Each test protein then has a maximum pairwise sequence identity (MPSID) with the remaining members of the same topological family used in training each of the models. We also include some examples of close homologs, with relatively high values for the MPSID to explore their effect on Z-scores.

Control dataset for calibration

A dataset of 132 unrelated proteins of known structures, which was obtained by slightly modifying the original database of 125 unrelated proteins (Rost and Sander, 1993), is listed here: 1acx, 2ak3A, 1azu, 1bbpA, 1bds, 1bmv1, 1bmv2, 1cc5, 1cdh, 1cdtA, 1crn, 1cseI, 1eca, 1etu, 1fc2C, 1fdlH, 1fdx, 1fkf, 1fxiA, 1gd1O, 1gp1A, 1hip, 1l58, 1lap, 1mcpL, 1ovoA, 1paz, 1pyp, 1r092, 1rbp, 1rhd, 1s01, 3sdhA, 1sh1, 1tgsI, 1tnfA, 1ubq, 1wsyA, 1wsyB, 256bA, 2aat, 2alp, 2cab, 2ccyA, 2cyp, 1fnd, 2fxb, 2gbp, 2gcr, 2gn5, 2hmzA, 2i1b, 1gdj, 2lhb, 2ltnA, 2ltnB, 2mev4, 2or1L, 2pabA, 2pcy, 2phh, 2rspA, 2sns, 2stv, 2tgpI, 2tmvP, 2tscA, 2utgA, 2wrpR, 3ait, 3b5c, 3blm, 3cla, 3cln, 3gapA, 3hmgA, 3hmgB, 3icb, 3pgm, 3rnt, 3timA, 4bp2, 4cms, 4cpv, 2fox, 4gr1, 4pfk, 4rhv1, 4rhv3, 4rhv4, 4sgbI, 4ts1A, 4×iaA, 5cytR, 5er2E, 5hvpA, 5ldh, 5lyz, 6acn, 6cpa, 6cpp, 6cts, 6dfr, 6tmnE, 7catA, 7icd, 7rsa, 8abp, 8adh, 9apiA, 9apiB, 9pap, 9wgaA, 2ace, 1colA, 3cox, 1f3g, 3gly, 2gmfA, 1hddC, 1hrhA, 1msbA, 1nsbA, 1pi2, 2cpkE, 2hipA, 2ifb, 2pk4, 2sarA, 2scpA, 4fgf, 5p21. These proteins were used throughout our study as negative controls for each recognition method being studied.

Z-Scores

Z-Scores were computed from the raw scores obtained by individual methods compared here. We chose to ignore the significance values given by some of these methods, as they are based upon a variety of statistical assumptions. Therefore, E-values, p-values, rankings, information content, etc. are ignored in favor of the raw scores on which these measures are based. Each method may be assessed objectively by its ability to generate a score for properly matching a query sequence to its own family of proteins which differs markedly from the scores produced by a standard set of unrelated control sequences. The fact that both the test protein and the control proteins had known structure made it possible to eliminate potential control set proteins which belong to the same fold family as the test protein and therefore might bear a distant evolutionary relationship. Knowledge of the structures also allows for proper identification of true positives and true negatives for each query sequence.

The distribution of raw scores for the control database was analyzed for each method. Although the control dataset represented 132 proteins, the empirical distributions of the scores deviated at most mildly from a normal (Gaussian) distribution for each method. To compare different methods, Z-scores were obtained from the raw scores by subtracting the mean and dividing by the standard deviation of the raw scores of the control database (for methods max_LSS and MHMM). When the raw score varied strongly with the length of the control protein (length range between 50 and 600 residues, median around 165), a length dependent correction was applied to the raw scores (for methods SHMM and FORESST), discussed in detail later.

Local sequence similarity (LSS) method

Local sequence similarity was computed using a version of BLASTP with the BLOSUM62 matrix, with the expectation threshold E of 1000, parameters B and V set to 500, T to 1, with the rest of the parameters set to their default values, so as to obtain alignment scores for all members of the control database. Representative proteins of each fold family (between three and 11 sequences per family), were used as query proteins for pairwise sequence comparison against each member of the family and the highest segment pairs (HSP) scores determined. Local sequence comparison was also performed between each protein from the control database to each fold family (Table I). The maximum HSP score (max_LSS) over the members of the family is recorded. This entire procedure was repeated for each fold family.

Motifs, their detection and construction of motif-based HMMs (MHMM)

Motifs of sequences within fold families were automatically detected using MEME and MAST (Bailey and Elkan, 1994) for the same set of sequences of each fold listed in Table I, with default parameter settings available over the Web (http://www.sdsc.edu/MEME). The motifs, ungapped, non-overlapping segments are later combined into a single model using meta-MEME (Grundy *et al*., 1997). The query sequence and the proteins from the control dataset were tested against the resulting motif-based HMM model for each fold using the HMMER software (Eddy, 1996). The 'hmmsw' option of HMMER is a Smith–Waterman based semi-local search program of a sequence database for best matches to a hidden Markov model, resulting in the log-odds score for the query protein.

Sequence-based hidden Markov model (SHMM)

Hidden Markov models were constructed for each fold family using the publically available software SAM version 1.3.1 (Hughey and Krogh, 1995). Default parameters were used for training 15 models with different seed values. The training was done by the expectation maximization method with a maximum of four surgeries, until the relative improvement in the negative log-likelihood (NLL) score dropped to 0.01 (Hughey and Krogh, 1995). A nine-component empirical Dirichlet mixture prior (Brown *et al*., 1993) was used in order to train models with smaller training sets (seven to 25 sequences). In spite of the potential advantage of using larger training sets (Hughey and Krogh, 1996), we did not include more sequences for training a given model because, after removing those proteins which were closely related to the query protein, the number of remaining proteins within the same fold family was limited. Also, we maintained consistency with the training proteins of HMM models based on observed secondary structures of fold families, which were also limited by the number of diverse structures in the same topological group of CATH. The resulting log-odds scores (NLL scores of the model—NLL scores of a 'NULL' model, obtained by average letter frequencies found in match states), were obtained for the true positives in each model and the true negatives of the control dataset.

FORESST

Hidden Markov models for each fold family were trained from unaligned, experimentally derived secondary structure sequences of the proteins listed in Table I by the method FORESST (Di Francesco *et al*., 1999; Di Francesco *et al*., 1997a). During training, the expectation–maximization estimation process was stopped when the relative improvement in the average NLL score of the training sequences with respect to the current model was less than 0.01, and up to four surgeries were allowed (Hughey and Krogh, 1995). The remaining parameters and the transition probability prior distributions were set to default values. The prior distributions of the observation symbols were set to be proportional to the fraction of the residues in the three secondary structure states in the training sequences of each model.

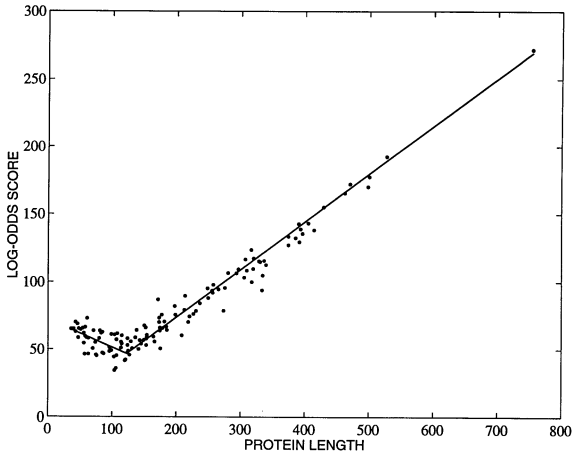

Fig. 1. Log-odds scores plotted as a function of the length of the proteins in the control dataset for a cross-validated sequence-based hidden Markov model for the protein 1poc in the phospholipase fold family. Predicted log-odds scores are shown as a solid line. The model length was 132, roughly corresponding to the 'knot' position (127) of the two linear segments.

To evaluate the capabilities of each model at recognizing its family members, Z-scores were calculated from log-odds scores as described above. The log-odds score is a measure of how a given HMM of a specific fold family fits a query sequence better than a generic null model related to some underlying background distribution. For each HMM, the null model was defined with default transition probability distributions and observation symbol probability distributions equal to the relative frequencies of helical, extended and coil residues of the training set sequences in the model match states.

When evaluating the recognition capabilities of the models using predicted secondary structure sequences (FORESST), both the query sequences and the sequences in the control database were predicted with the quadratic–logistic (QL) method (Munson *et al*., 1994). When evaluating the models using observed secondary structure sequences (denoted FORESST-obs), the query sequences and the sequences in the control database were obtained from the output of DSSP (Kabsch and Sander, 1983).

Length dependence of the log-odds scores

The log-odds scores by the HMM methods for the negatives of the control database were found to be generally related to the length of the protein sequence, often in a non-linear fashion. Accordingly, for each model, a simple linear regression and a two-segment linear regression model were fitted to the scores of the control proteins (Figure 1) to determine the expected value of the score for a sequence of given length. For the two-segment model, two lines were fit, with the break-point or 'knot' adjusted to fit the data. The linear fit was accepted unless the two-segment fit displayed a residual sum of squares values less than 90% of that for the linear fit. In either case, the Z-scores were determined by subtracting the regression value (expected value given the sequence length) from the raw score and dividing the result by the root mean square (r.m.s.) error from the regression. In most cases, the expected score was adequately fit by two linear segments with the knot placed near the effective length of the hidden Markov model.

Effect of size of the database for computing Z-scores

To resolve any question concerning the modest size of our control database, and the possible effect on the Z-scores for the local alignment method, a much larger control database was developed by randomly sampling a very large database of sequences from GENPEPT 103 (ftp://ftp.ncifcrf.gov/pub/genpept) to produce a collection of 1000 protein sequences. Since these sequences were not all associated with known structures, we were less confident about their status as negative controls. But sequences bearing an obvious relationship to the query were removed from the set when the BLAST E-values for those sequences were <0.00001. The distribution of HSP scores of the true negatives for the small database of 132 proteins gave a mean of 28.37 and a standard deviation of 5.17. The same distribution with respect to the larger database resulted in a mean of 28.14 and a standard deviation of 5.20 and we concluded that the smaller database was adequate for the current study.

The effect of size of the database was also closely examined for the HMM based methods. The standard deviation of log-odds scores after subtracting the length-dependent expected log-odds scores for the small database of 132 proteins was 9.74, whereas the value was 9.39 for the larger database. The change in the standard deviation is not greater than would be expected from the sampling error for samples of size 132 and 1000 and we again concluded that the smaller dataset was adequate here.

Scores and their evaluation

The pairwise sequence identity determined by the CLUSTALW alignment (Thompson *et al*., 1994) between proteins is often used as a measure of how likely the two proteins are to adopt similar folds. This serves as rough measure of the difficulty of the recognition problem, since protein pairs with higher pairwise sequence identity are more easily recognized. We define the MPSID as the maximum pairwise sequence identity (MPSID) of the sequences used in the various models.

Sensitivity and specificity for fold assignment

Z-Scores were computed for members and non-members of each fold family to examine the discriminatory power of each method. If the Z-score for a query against a particular protein family is greater than 2.0, the query is considered 'positive' for that family. Sensitivity is defined as the ratio of true positives to all true family members and specificity is defined as the ratio of true positives to all positives for that family. The percentage sensitivity and specificity therefore vary anywhere between 0 and 100. A good method should have both high sensitivity and high specificity values.

Results

Comparison of methods for fold assignment

The Z-scores computed for the local pairwise sequence comparison (max_LSS), sequence-based HMMs (SHMM), motif-based HMMs (MHMM), FORESST and FORESST-obs will be discussed here. There were in total 45 proteins from different fold families tested for fold assignment (Table I). The local pairwise sequence comparison method gave rise to highest scoring segment pairs (HSP) for each protein tested against each family member. To obtain a single score for the entire family, the maximum HSP score (max_LSS) was taken on the theory that a family is recognized if the query recognizes at least a single member. We analyzed the entire data where proteins with MPSID were less than 55% (Figure 2). The mean Z-score (standard error) for the max_LSS method is 7.7(2.1). The SHMM method has a mean Z-score of 4.4(0.9)

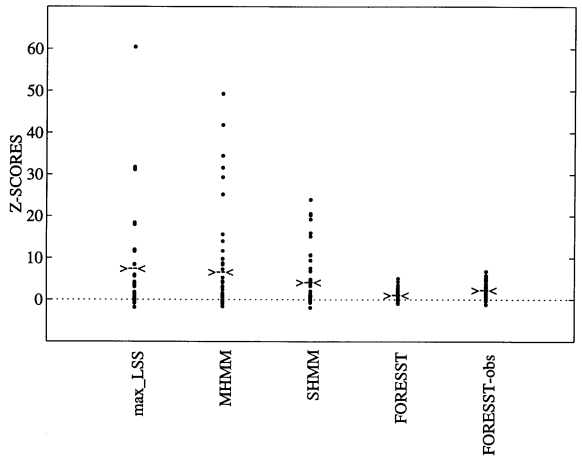

Fig. 2. Z-Scores obtained for various methods for all the test proteins. The average Z-score for each method is also indicated (as >--<).

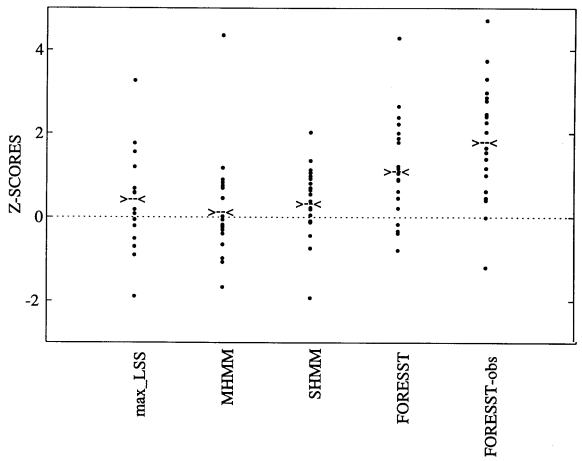

Fig. 3. Z-scores obtained for various methods for proteins with MPSID <15%, namely, the remote homologs.

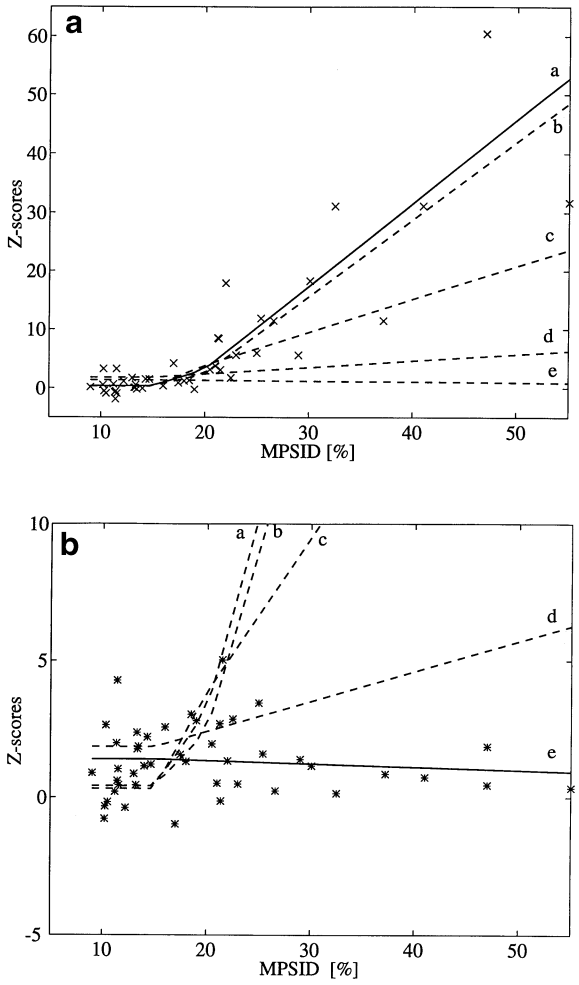

Fig. 4. (**a**) Z-Scores plotted as a function of MPSID. a, max_LSS; b, MHMM; c, SHMM; d, FORESST-obs; and e, FORESST. Data points (x) shown only for max_LSS. Data for all the methods are modeled as a segmented linear fit. (**b**) Z-Scores plotted as a function of MPSID. a, max_LSS; b, MHMM; c, SHMM; d, FORESST-obs; e, FORESST. Data points (*) shown only for FORESST. Data for all the methods are modeled as a segmented linear fit.

and MHMM method has a mean Z-score of 6.9(1.8). The FORESST and FORESST-obs methods have means 1.31(0.19) and 2.58(0.27) with maximum Z-scores of 5.1 and 6.8, respectively.

Sequence-based methods are generally successful in finding homologous proteins with MPSID greater than 20% but are less successful for lower values of MPSID. The Z-scores for the 20 proteins with MPSID less than 15% are plotted as a function of the different methods (Figure 3). When MPSID is less than 15%, average Z-score for max_LSS drops to 0.5(0.3), making recognition unreliable for the remote homologous pairs. For this range of MPSID, MHMM and SHMM methods also have low average Z-scores, 0.1(0.2) and 0.3(0.1), respectively. In contrast, FORESST has a mean of 1.1(0.2) (Figure 3). When observed rather than predicted secondary structure is used, the mean Z-score rises to 1.8(0.3), suggesting that improvements in prediction accuracy would materially improve the overall performance of the FORESST method. Moreover, the average Z-scores for FORESST (using prediction or observation) is higher than for any of the sequence-based methods analyzed in this paper.

The results (Figure 4a) for all the sequence-based methods show a striking dependency on the difficulty of the recognition problem as measured by the maximum pairwise sequence identity (MPSID). The Z-scores appear to increase approximately linearly for MPSID above 20% but are evidently flat for MPSID below 15%. Accordingly, we fit the individual points with a segmented linear model, with a break point set at 15%. Although the exact position of the break could not be determined precisely, it lies within the range of 15–20%. Of the four sequence-based methods, the max_LSS generally performs better than other sequence-based methods for MPSID over 20%. The FORESST-obs method performs better compared with any of the methods for MPSID below 15%.

The Z-scores appear to be largely independent of MPSID for FORESST (Figure 4b). The average (standard error) Z-score over the entire range of MPSID for FORESST is 1.31(0.19). The average Z-score for FORESST using observed secondary structures is 2.58(0.27) over the entire range, although the Z-scores increase gradually as the MPSID increases, reaching a maximum Z-score around 7.0 for MPSID of 55%.

For each family, methods can recognize not only the proteins of their own members but also proteins from other families suggesting that recognition specificity is an issue. The overall sensitivity and specificity rates for all the methods are found to be strongly dependent on the difficulty of the recognition problem. Table II summarizes the sensitivity and specificity

Table II. Overall sensitivity and specificity by various methods

METHODS	All test proteins		MPSID ⩽15%		16 ⩽ MPSID ⩽ 30%	
	Sensitivity[a] (%)	Specificity[b] (%)	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
max_LSS	47	58	7	10	77	63
MHMM	44	71	6	7	66	63
SHMM	40	35	6	7	58	34
FORESST	24	41	25	66	39	37
FORESST-obs	64	51	50	58	66	48

[a]Sensitivity is defined as the ratio of true positives to all true family members.
[b]Specificity is defined as the ratio of true positives to all positives for that family. Detection is based on Z-scores greater than 2.0.

for all the test proteins at varying levels of the MPSID. The max_LSS method has greater sensitivity and specificity for proteins with MPSID between 16 and 30% (Table II), compared with proteins with MPSID less than 15%. This sequence similarity search method is not capable of reliably identifying the very remote homologous pairs.

The MHMM method seems to perform well in cases where it is possible to identify fingerprints or sequence motifs of the fold families as is evident from the overall specificity rates for all the query proteins and for proteins with MPSID ranging from 16 to 30%. The moderate sensitivity and higher overall specificty rate (Table II, columns 2 and 3) by MHMM for all the test proteins when compared with other methods, seems to roughly corroborate a similar observation which found the motif-based method MEME to have moderate power and higher confidence, while comparing multiple protein sequence alignment servers (Briffeuil *et al*., 1998). Both the sensitivity and specificity rates decrease for MPSID below 15%.

The SHMM method has an overall sensitivity of 40% and the overall specificity is 35%. The overall sensitivity rate increases for MPSID between 16 and 30%. In essence, all the sequence-based methods have a very low sensitivity and specificity rate for MPSID below 15%, thus making recognition of these very remote homologs very unlikely.

The overall sensitivity and specificity rates by the FORESST method (Table II) is lower compared with other sequence-based methods for all the test proteins. The specificity rate is slightly better than the SHMM method but still less than other sequence-based methods. However, the sensitivity and the specificty rates are much higher than sequence-based techniques when proteins have MPSID less than 15% (Table II, columns 4 and 5). When the FORESST method is based on observed secondary structures for the query protein, the sensitivity, specificity rates are even better compared with all the other methods analyzed.

Discussion

We have assessed various methods for detection of remote homologs using a uniform, objective, statistically valid comparison. Each HMM-based method was trained on the identical database of proteins and tested with carefully controlled cross-validation. The protein families we have discussed in this paper include both close and remote homologs. Even though the remote homologs are not easily detectable by sequence search methods alone, such methods may at times recognize distant homologs. The results also show that the local similarity search method performs relatively well compared with other sequence or structure based approaches, mainly for an MPSID greater than 20%. This corroborates an earlier study, where BLAST was found to outperform other sequence based approaches, namely, HMMER and MEME (Grundy, 1998). However, that study did not examine their performance as a function of the pairwise sequence identity which we show to be critical to the understanding of the relative performance of these methods. Use of a gapped BLAST or PSI-BLAST (Altschul *et al*., 1997) might also improve the results obtained by the LSS method here for some of the remote homologs. Although an earlier study comparing the PSI-BLAST and HMM methods found the HMM method to detect 35% of the true homologs while PSI-BLAST detected only 30% (Park *et al*., 1998).

Close homologs are readily recognized by sequence based methods. The recognition power of each method improves roughly linearly with sequence identity above 15–20%, except for FORESST which maintains a nearly constant Z-score between 1.0 and 1.5, regardless of the sequence identity of the target. That the Z-scores for FORESST do not increase with MPSID may be attributed to the errors in the secondary structure prediction. The method based on experimental secondary structures (FORESST-obs) is included to explore the upper limit of the performance of any of the secondary structure prediction methods. It has been noted that the greater the sequence identity between a pair of related proteins, the greater the agreement of secondary structures (Chothia and Lesk, 1986; Russell and Barton, 1994). This would explain the increase of Z-scores with MPSID seen here for FORESST based on observed secondary structures as opposed to predicted ones. The maximum attainable Z-score for the FORESST method is smaller than for any of the sequence-based methods. This is likely due to the reduced size of the alphabet used for representing secondary structures (namely, H, E and C, see Materials and methods), which in turn implies a higher chance of falsely matching a given position of the secondary structure sequence in the control database.

The higher rates of sensitivity and specificity for the FORESST method compared with all the sequence-based methods in the range of MPSID less than 15%, suggests the advantage of the secondary structure-based method over the sequence-based methods for identification of these very remote homologs. However, the moderate sensitivities and specificities with predicted secondary structures emphasizes the need to improve the accuracy of prediction and encourages the use of NMR secondary structure assignments for fold recognition.

The relatively poor performance of the SHMM method for recognition of some of the remote homologs may be attributed in part to the fact that the hidden Markov models were not

trained with an adequate number of sequences. The training set of the SHMM model includes few sequences, which resulted in the artificially low specificity and sensitivity of some of these models. Recognition might improve for the SHMM or even the MHMM method if more sequences were used for training, as has been shown by Karplus and co-workers (Karplus *et al*., 1997). In this study, we used a limited number of sequences for training the SHMM method, in order to allow for the rigorous cross-validation of the models and to confine our study only to proteins of known structure, a prerequisite for determining membership in a structural fold family. In any case, it should be observed that with a smaller number of sequences in the training set, the HMMs based on secondary structures are more successful in identifying remote homologs than the HMMs based on amino acid sequences.

The protein sequences whose structures were not yet published, served as prediction targets for various fold recognition methods (Lemer *et al*., 1995; Levitt, 1997) during the second Critical Assessment of Structure Prediction experiment, namely CASP2 (Moult *et al*., 1997). Some of the prediction targets at the CASP2 had functional relationship with the existing protein folds (Russell *et al*., 1998) and those proteins which had sequence motifs were identified as 'easy targets' (Marchler-Bauer and Bryant, 1997). The method responsible for the accurate predictions for almost all of the submitted targets in the fold recognition category had combined knowledge from various sources, namely, sequence, function, predicted secondary structure and information from the literature (Murzin and Bateman, 1997). In the absence of function or other information, as is the case in high throughput genome sequencing projects, a fold recognition method based on structure–sequence information alone might help to infer the function for the novel protein.

As an illustration of our Z-score methodology, we applied these five methods to the two CASP2 prediction targets (T0004 and T0031) which are among the fold families studied here. Post hoc analysis of the T0004 sequence by the current approach resulted in a Z-score of 2.20 by the max_LSS method, identifying the protein 1mjc, a member of the OB fold family. The Z-scores obtained for the OB fold family by the SHMM and FORESST methods were 2.46 and 2.86, respectively, whereas MHMM yielded a Z-score of 0.70. Three out of five methods studied here seem to be able to recognize this remote homolog of the OB fold family with Z-scores greater than 2.0. In the CASP2 contest, we correctly predicted T0004 to be a member of the OB fold family using the FORESST method (Di Francesco *et al*., 1997a,b). The success of sequence similarity search methods like BLASTP in identifying T0004 as an OB fold, even though the sequence identity is in the twilight zone, is partly due to the conservation of the key functional residues. The sequence motif Phe22, Val33, His34, Ser36, Ileu38 in T0004 is also found in the RNP-1 motif of cold-shock proteins (Schindelin *et al*., 1994; Marchler-Bauer and Bryant, 1997; Murzin and Bateman, 1997).

The Z-score for the prediction target T0031 turned out to be 4.28 against a serine protease SHMM, 1.58 by the max_LSS method and 2.44 by the FORESST method. Here again, the MHMM method gave the Z-score of 0.15, and was unable to recognize this protein as a member of the serine protease fold family. Our CASP2 FORESST prediction for T0031 as a member of the serine protease fold is confirmed by the post hoc analysis of the Z-scores on the target by this method. Furthermore, T0031 has conserved functional residues similar to the catalytic triad of serine proteases (Nienaber *et al*., 1993; Marchler-Bauer and Bryant, 1997; Murzin and Bateman, 1997).

We were also able to identify a novel pleckstrin homology (PH) domain using FORESST and SHMM in the mammalian phospholipase D (PLD) proteins PLD1 and PLD2 (Holbrook *et al*., 1999). Previously, these PLD protein sequences were reported to lack signaling domains. The presence of these PH domains in PLDs has also been demonstrated with independent biochemical and sequence homology searches (Steed *et al*., 1998). This finding would therefore resolve the contradictory observations about PLD regulation (Steed *et al*., 1998).

After fold recognition, the query sequence must then be correctly aligned to a known fold in order to build a successful three-dimensional model for the novel sequence. Unfortunately, the sequence to structure alignment quality of current fold recognition methods is not yet adequate and stands in the way of better protein structure prediction. Fold recognition currently helps in the annotation of sequences for which no function is known. Future progress in fold recognition is also hampered by the fact that there are only a limited number of known folds. Even attempts to build novel folds by understanding the topological rules of existing folds (Reva and Finkelstein, 1996) have not advanced the current repertoire of fold libraries. It remains to be seen if existing or new and improved fold recognition methods are capable of identifying a greater number of remote homologs successfully at the CASP3 contest.

Conclusion

We conclude that the sequence-based methods are successful in recognizing close homologs, but structure–sequence based methods, such as FORESST, are more appropriate for fold recognition of remote homologs (MPSID less than 15%). The performance of FORESST-obs, which used the experimentally-derived secondary structure of the query sequence, suggests that improving secondary structure prediction can improve automated recognition in some cases. None of the sequence based methods studied here can provide evidence for relatedness of these distant homologs, in general. In choosing between sequence-based or sequence–structure-based methods for fold recognition, one should be guided by the degree of relatedness of the homolog being sought. Our results suggest that a hybrid method utilizing both sequence (for close homolog searches) and secondary structure prediction methods (for remote homolog searches), would be an even better approach for automatic recognition of folds of novel protein sequences.

References

Agarwal,P. and States,D.J. (1998) *Bioinformatics*, **14**, 40–47.

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

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., and Miller,W.D.J.L. (1997) *Nucleic Acids Res*., **25**, 3389–3402.

Bailey,T.L. and Elkan,C. (1994), *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology.* Stanford University, Palo Alto, pp. 28–36.

Brenner,S.E., Chothia,C. and Hubbard,T.J.P. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.

Briffeuil,P., Baudoux,G., Lambert,C., De Bolle,X., Vinals,C., Feytmans,E. Depiereux,E. (1998) *Bioinformatics*, **14**, 357–366.

Brown,M.P., Hughey,R., Krogh,A., Mian,I.S., Sjolander,K. and Haussler,D. (1993), *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology.* National Library of Medicine, Bethesda, pp. 47–55.

Chothia,C., Hubbard,T., Brenner,S., Barns,H. and Murzin,A. (1997) *Annu. Rev. Biomol. Struct*., **26**, 597–627.

Chothia,C. and Lesk,A.M. (1986) *EMBO J.*, **5**, 823.

Di Francesco,V., Garnier,J. and Munson,P.J. (1997a) *J. Mol. Biol.*, **267**, 446–463.
Di Francesco,V., Geetha,V., Garnier,J. and Munson,P.J. (1997b) *Proteins Struct. Funct. Genet.*, **1**, 123–128.
Di Francesco,V., Munson,P.J. and Garnier,J. (1999) *Bioinformatics*, **15**, 131–140.
Eddy,S.R. (1996) *Curr. Opin. Str. Biol.*, **6**, 361–365.
Flores,T.P., Orengo,C.A., Moss,D.M. and Thornton,J.M. (1993) *Protein Sci.*, **2**, 1811–1826.
Grundy,W.N. (1998), *Proceedings of the Second Annual International Conference on Computational Molecular Biology.* New York City, 94–100.
Grundy,W.N., Bailey,T.L., Elkan,C.P. and Baker,M.E. (1997) *CABIOS*, **13**, 397–406.
Holbrook,P., Geetha,V., Beaven,M.A. and Munson,P.J. (1999) *FEBS Lett.*, **448**, 269–272.
Hughey,R. and Krogh,A. (1996) *CABIOS*, **12**, 95–107.
Hughey,R. and Krogh,A. (1995) *SAM: Sequence Alignment and Modeling Software System*. University of California, Santa Cruz. Report no. UCSC-CRL-95-7.
Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.
Karplus,K., Barrett,C. and Hughey,R. (1998) *J. Mol. Biol.*, in press.
Karplus,K., Sjolander,K., Barrett,C., Cline,M., Haussler,D., Hughey,R., Hol,L. and Sander,C. (1997) *Proteins Struct. Funct. Genet.*, Suppl., **1**, 134–139.
Lemer,M.-R., Rooman,M.J. and Wodak,S.J. (1995) *Proteins Struct. Funct. Genet.*, **23**, 337–355.
Levitt,M. (1997) *Proteins Struct. Funct. Genet.*, **92**, 92–104.
Levitt,M. and Gerstein,M. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 5913–5920.
Marchler-Bauer,A. and Bryant,S.H. (1997) *Trends Biochem. Sci.*, **22**, 236–240.
Moult,J., Hubbard,T., Bryant,S.H., Fidelis,K. and Pedersen,J.T. (1997) *Proteins Struct. Funct. Genet.*, Suppl., **1**, 2–6.
Munson,P.J., Di Francesco,V. and Porrelli,R.N. (1994), *27th Hawaii International Conference System Sciences V,* Maui, Hawaii, pp. 375–384.
Murzin,A.G. and Bateman,A. (1997) *Proteins Struct. Funct. Genet.*, Suppl., **1**, 105–112.
Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol.*, **247**, 536–540.
Nienaber,V.L., Breddam,K. and Birtoft,J.J. (1993) *Biochemistry*, **32**, 11469–11475.
Orengo,C.A., Flores,T.P., Jones,D.T., Taylor,W.R. and Thornton,J.M. (1993) *Curr. Biol.*, **3**, 131–139.
Orengo,C.A., Michie,A.D., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) *Structure*, **5**, 1093–1108.
Park,J., Karplus,K., Barrett,C., Hughey,R., Haussler,D., Hubbard,T. and Chothia,C. (1998) *J. Mol. Biol.*, **284**, 1201–1210.
Pearson,W.R. (1995) *Protein Sci.*, **4**, 1145–1160.
Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444.
Reva,B.A. and Finkelstein,A.V. (1996) *Protein Engng*, **9**, 399–411.
Rost,B. and Sander,C. (1993) *J. Mol. Biol.*, **232**, 584–599.
Russell,R.B. and Barton,G.J. (1994) *J. Mol. Biol.*, **244**, 332–350.
Russell,R.B., Saqi,M.A.S., Bates,P.A., Sayle,R.A. and Sternberg,M.J.E. (1998) *Protein Engng*, **11**, 1–9.
Sander,C. and Schneider,R. (1991) *Proteins Struct. Funct. Genet.*, **9**, 56–68.
Schindelin,H., Jiang,W., Inouye,M. and Heinemann,U. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 5119–5123.
Steed,P.M., Clark,K.L., Boyar,W.C. and Lasala,D.J. (1998) FASEB, **12**, 1309–1317.
Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.

Received November 15, 1998; revised March 19, 1999; acccepted March 22, 1999