BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 12 2006, pages 1456–1463
doi:10.1093/bioinformatics/btl102

Structural bioinformatics

A machine learning information retrieval approach to protein fold recognition

Jianlin Cheng and Pierre Baldi*

Institute for Genomics and Bioinformatics, School of Information and Computer Sciences, University of California, Irvine, CA, USA

Received on January 17, 2006; revised on March 4, 2006; accepted on March 15, 2006
Advance Access publication March 17, 2006
Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Recognizing proteins that have similar tertiary structure is the key step of template-based protein structure prediction methods. Traditionally, a variety of alignment methods are used to identify similar folds, based on sequence similarity and sequence-structure compatibility. Although these methods are complementary, their integration has not been thoroughly exploited. Statistical machine learning methods provide tools for integrating multiple features, but so far these methods have been used primarily for protein and fold classification, rather than addressing the retrieval problem of fold recognition-finding a proper template for a given query protein.

Results: Here we present a two-stage machine learning, information retrieval, approach to fold recognition. First, we use alignment methods to derive pairwise similarity features for query-template protein pairs. We also use global profile–profile alignments in combination with predicted secondary structure, relative solvent accessibility, contact map and beta-strand pairing to extract pairwise structural compatibility features. Second, we apply support vector machines to these features to predict the structural relevance (i.e. in the same fold or not) of the query-template pairs. For each query, the continuous relevance scores are used to rank the templates. The FOLDpro approach is modular, scalable and effective. Compared with 11 other fold recognition methods, FOLDpro yields the best results in almost all standard categories on a comprehensive benchmark dataset. Using predictions of the top-ranked template, the sensitivity is ~85, 56, and 27% at the family, superfamily and fold levels respectively. Using the 5 top-ranked templates, the sensitivity increases to 90, 70, and 48%.

Availability: The FOLDpro server is available with the SCRATCH suite through http://www.igb.uci.edu/servers/psss.html.

Contact: pfbaldi@ics.uci.edu

Supplementary information: Supplementary data are available at http://mine5.ics.uci.edu:1026/gain.html

1 INTRODUCTION

The key step of template-based protein structure prediction approaches (comparative modeling and fold recognition) is to recognize proteins that have similar tertiary structures. This task becomes very challenging when there is little sequence similarity between the query and the template protein. Several alignment methods have been used to try to identify fold similarity, using sequence information, structural information or both. Instead of developing a new specialized alignment method for fold recognition (Shi *et al*., 2001; Xu *et al*., 2003; Zhou and Zhou, 2004), or integrating existing fold recognition servers (Lundstrm *et al*., 2001; Fischer, 2003; Ginalski *et al*., 2003*a*), here we propose a machine learning information retrieval approach that leverages features extracted using existing, general-purpose, alignment tools as well as protein structure prediction program and combines them using support vector machines (SVMs) to rank all the templates.

1.1 Classical approaches to fold recognition

Alignment methods for fold recognition include sequence–sequence, sequence–profile (or profile–sequence), profile–profile and sequence–structure methods.

Sequence–sequence alignment methods (Needleman and Wunsch, 1970; Smith and Waterman, 1981; Dayhoff *et al*., 1983; Pearson and Lipman, 1988; Altschul *et al*., 1990; Henikoff and Henikoff, 1992; Vingron and Waterman, 1994) are effective at detecting homologs with significant sequence identity (>40%).

Sequence–profile (or profile–sequence) alignment methods (Baldi *et al*., 1994; Krogh *et al*., 1994; Hughey and Krogh, 1996; Altschul *et al*., 1997; Bailey and Gribskov, 1997; Karplus *et al*., 1998; Eddy, 1998; Park *et al*., 1998; Koretke *et al*., 2001; Gough *et al*., 2001) are more sensitive at detecting distant homologs with lower sequence identity (>20%). Profiles can correspond to simple multiple alignments, to position specific scoring matrices (PSSMs), or to hidden Markov models (HMMs).

Profile–profile alignment approaches (Thompson *et al*., 1994; Rychlewski *et al*., 2000; Notredame *et al*., 2000; Yona and Levitt, 2002; Madera and Gough, 2002; Mitelman *et al*., 2003; Ginalski *et al*., 2003*b*; Sadreyev and Grishin, 2003; Edgar and Sjolander, 2003, 2004; Ohlson *et al*., 2004; Wallner *et al*., 2004; Wang and Dunbrack, 2004; Marti-Renom *et al*., 2004; Söding, 2005) are even more sensitive at detecting distant homologs and compatible structures, and often achieve even better performance than sequence–structure alignment methods that leverage template structural information (Rychlewski *et al*., 2000).

Sequence–structure alignment methods (or threading) (Bowie *et al*., 1991; Jones *et al*., 1992; Godzik and Skolnick, 1992; Bryant and Lawrence, 1993; Abagyan *et al*., 1994; Murzin and Bateman, 1997; Xu *et al*., 1998; Jones, 1999; Panchenko *et al*., 2000; David *et al*., 2000; Shi *et al*., 2001; Skolnick and Kihara, 2001; Xu *et al*., 2003; Kim *et al*., 2003) align query sequences with template structures and compute compatibility scores according to

*To whom correspondence should be addressed.

structural environment fitness and contact potentials. These methods are particularly useful for detecting proteins with similar folds but no recognizable evolutionary relationship.

The separation between sequence-based and structure-based methods, however, is becoming blurred as new methods are developed that combine both kinds of information together. Combining both sequence and structure information has been shown to improve both fold recognition (Elofsson *et al*., 1996; Jaroszewski *et al*., 1998; Al-Lazikani *et al*., 1998; Fischer, 2000; Kelley *et al*., 2000; Panchenko *et al*., 2000; Shan *et al*., 2001; Tang *et al*., 2003; Pettitt *et al*., 2005) and alignment quality (Thompson *et al*., 1994; Al-Lazikani *et al*., 1998; Domingues *et al*., 2000; Notredame *et al*., 2000; Griffiths-Jones and Bateman, 2002; Tang *et al*., 2003; O'Sullivan *et al*., 2004). Even the sequence-derived predicted secondary structure can be used to increase the sensitivity of fold recognition (Rost and Sander, 1997; Jones, 1999; Ginalski *et al*., 2003*b*; Xu *et al*., 2003; Zhou and Zhou, 2004).

In fold recognition, different alignment tools are often used independently to search protein databases for similar structures. Previous research (Jaroszewski *et al*., 1998; Lindahl and Elofsson, 2000; Shan *et al*., 2001; Ohsen *et al*., 2003; Wallner *et al*., 2004) has shown that these alignment methods are complementary and can find different correct templates. But combining these methods is difficult (Lindahl and Elofsson, 2000). Meta or jury approaches (Lundstrm *et al*., 2001; Fischer, 2003; Ginalski *et al*., 2003*a*; Juan *et al*., 2003; Wallner *et al*., 2004) collect the predicted models from external fold recognition predictors and derive predictions based on a small set of returned candidates. This popular, hierarchical approach increases the reach of fold recognition. However, it relies on the availability of external predictors and cannot recover true positive templates discarded prematurely by individual predictors.

1.2 A machine learning information retrieval approach to fold recognition

Statistical machine learning methods provide powerful means for integrating disparate features in pattern recognition. So far, however, machine learning integration of features has been used in this area primarily for coarse homology detection, such as protein structure/fold classification (Jaakkola *et al*., 2000; Leslie *et al*., 2002). Classifying proteins into a few categories or even dozens of families, superfamilies and folds, however, does not provide the specific templates required for template-based structure modeling. Furthermore, current classification methods are not likely to scale up to the thousands of families, superfamilies and folds already present in current protein classification databases, such as SCOP (Murzin *et al*., 1995). Fold recognition is different from protein classification—it is fundamentally a retrieval problem, very much like finding a document or a web page (Rocchio, 1966; Page *et al*., 1998). Given a query protein, the objective of fold recognition is rather to rank all possible templates according to their structural relevance, like Google and other search engines rank web pages associated with a user's query.

Machine learning methods (such as binary classifiers) have been used also in threading approaches (Jones, 1999; Xu *et al*., 2003) to combine multiple scores produced by threading into a single scores to rank the templates. Here we generalize this idea and derive a broad machine learning framework for the fold recognition/retrieval problem. The framework integrates a variety of similarity features and feature extraction tools, including standard alignment tools. However, unlike meta approaches, it does not require any pre-existing fold recognition programs or servers.

Consistently with the major trend in machine learning towards kernel methods (Schölkopf and Smola, 2002), we first focus on the computation of a variety of similarity measures between query-template pairs. Instead of extracting features and analyzing individual sequences, we focus exclusively on pairs of sequences and use a variety of complementary alignment tools to align the query protein with the template proteins, rather than to search the database of templates. The alignment scores for query-template pairs are used as similarity measures. Furthermore, based on alignments (e.g. profile–profile) between query and template, we further extract pairwise structural compatibility features by checking the predicted secondary structure, solvent accessibility, contact map and beta-strand pairings of the query protein against the tertiary structure of the template protein. Second, these alignment and structural similarity scores as well as other sequence and structural features derived using three standard similarity measures (cosine, correlation and Gaussian kernel) are fed into SVMs (Vapnik, 1998) to learn a relevance function to evaluate whether the query and template belong to the same fold. Finally, the continuous output scores produced by the SVMs are used to rank the templates with respect to the query. The top-ranked templates can be used to model the structure of the query.

2 METHODS

2.1 Feature extraction

We extract five categories of pairwise features (similarity scores) for each query–template pair associated with sequence or family information, sequence alignment, sequence–profile (or profile–sequence) alignment, profile–profile alignment and structure (Table 1).

Sequence/family information features. To compare the sequences of query and template proteins, we compute their single amino acid (monomer) and ordered pair of amino acids (dimer) compositions. The composition vectors x and y of the query and template are compared and transformed into six similarity scores using the cosine $(x \cdot y/|x||y|)$, correlation $((\sum_i (x_i - \bar{x})(y_i - \bar{y})/(\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}))$, and Gaussian kernel $(e^{-||x-y||^2})$ respectively. We apply the same techniques to the monomer and dimer residue composition vectors of the family of sequences associated with the query and the template to extract another set of six similarity measures, to measure the family composition similarity. The sequences for both query and template families are derived from multiple sequence alignments generated by searching the NCBI non-redundant sequence database (NR release 1.21, 28-Apr-2003) using PSI-BLAST(Altschul *et al*., 1997). The e-value ($-e$ option) threshold for inclusion in the profile is set to 0.001; the cut-off threshold ($-h$ option) for building iterative profiles is set to $e - 10$; and the number of iteration ($-j$ option) is set to 3. Thus the sequence/family information feature subset includes 12 (6 + 6) pairwise features in total.

Sequence–sequence alignment features. Two sequence alignment tools, PALIGN (Ohlson *et al*., 2004) and CLUSTALW (Thompson *et al*., 1994), are used to extract pairwise features associated with sequence alignment scores. PALIGN uses local alignment methods and produces a score and an e-value. The score is divided by the length of the query to remove any length bias. CLUSTALW generates a global sequence alignment score between the query and the template. This score is also normalized by the length of the query sequence. Thus the sequence alignment feature subset includes three pairwise features.

Table 1. Features used in fold recognition. cos/corr/Gauss denote cosine, correlation, and Gaussian kernel functions

Category	Feature	Method	Num
Seq and Family Info.	Seq monomer compo	cos/corr/Gauss	3
	Seq dimer compo	cos/corr/Gauss	3
	Fam monomer compo	cos/corr/Gauss	3
	Fam dimer compo	cos/corr/Gauss	3
Seq–Seq Align.	Local alignment	PALIGN	2
	Global alignment	CLUSTALW	1
Seq–Prof Align.	Prof versus seq	PSI-BLAST	3
	Prof versus seq	IMPALA	3
	Prof versus seq	HHMER	2
	Seq versus prof	RPS-BLAST	3
	Seq versus prof	HMMER	1
Prof–Prof Align.	Multiple alignment	CLUSTALW	1
	PSSM	COMPASS	2
	HMM prof	PRC	6
	HMM prof	HHSearch	1
Structural Info.	SS and RSA match	ratio	2
	SS and RSA compo	cos/corr/Gauss	4
	Contact probability	average	2
	Residue contact order	cos/corr	4
	Residue contact num	cos/corr	4
	Beta-sheet pair prob.	average	1
Total	-	-	54

SS and RSA represent secondary structure and relative solvent accessibility respectively.

Sequence–profile (or profile–sequence) alignment features. We use three different profile–sequence alignment tools [PSI-BLAST, HMMER-hhmsearch (Eddy, 1998) and IMPALA (Schaffer *et al*., 1999)] to extract profile–sequence alignment features between the query profile and the template sequence. The profiles (or multiple alignments) for queries are generated by searching the NR database using PSI-BLAST, as described above. Identical sequences in the multiple alignments are removed. No sophisticated weighting scheme is used. The multiple alignments are used by all profile alignment tools directly, or as the basis for building customized profiles. For instance, the HMM models in HMMER are built from the multiple alignments using the hmmbuild and hmmcalibrate tools of HMMER. Note that, instead of using these tools to search sequence databases, we use them to align individual query profiles against individual template sequences to extract pairwise features. The alignment score normalized by the query length, the logarithm of the *e*-value and the alignment length normalized by the query length from the most significant PSI-BLAST and IMPALA local alignment are used as features. The alignment scores, normalized by the length of the query sequence, and the logarithm of the *e*-value produced by hmmsearch alignments are used as features too. Thus the profile-sequence alignment tools generate eight pairwise features.

For sequence–profile alignments, we use RPS-BLAST in the PSI-BLAST package and hmmpfam in the HMMER package to align the query sequence with the template profiles. The template profiles are generated in the same way as the query profiles. In this way, RPS-BLAST generates three features similar to PSI-BLAST. The logarithm of the *e*-value produced by hmmfam is also used as one feature. Thus the subset of profile–sequence (or sequence–profile) alignment features includes 12 (8 + 4) pairwise features in total.

Profile–profile alignment features. We use five profile–profile alignment tools including CLUSTALW, COACH of LOBSTER (Edgar and Sjolander, 2004), COMPASS (Sadreyev and Grishin, 2003), HHSearch (Söding, 2005) and PRC (Profile Compiled by M. Madera, http://supfam.org/PRC) to align query and template profiles. The global alignments produced by CLUSTALW and LOBSTER and the most significant local alignments produced by COMPASS, PRC and HHSearch are used to extract the pairwise features. Specifically, CLUSTALW aligns query multiple alignments with template multiple alignments. COACH aligns query HMMs with template HMMs built from the multiple alignments produced by LOBSTER. HHSearch also aligns query HMMs with template HMMs generated from the multiple alignments using the hhmake function of HHSearch. The alignment scores produced by CLUSTALW and HHSearch are normalized by query length and used as pairwise features. The alignment scores produced by LOBSTER are not used directly as features because their dependence on template length would introduce a bias toward long templates.

PRC, an HMM profile–profile alignment tool, is used with two different kinds of profiles: HMM models built by HMMER and chk profiles built by PSI-BLAST. In each case, PRC produces three scores (co-emission, simple and reverse), which are normalized by query length. COMPASS, which uses internally a log-odds ratio score and a sophisticated sequence weighting scheme, is used to align query multiple alignments with template multiple alignments. The Smith–Waterman local alignment score normalized by query length and the logarithm of the *e*-value from the COMPASS alignments are used also as pairwise features. Thus the subset of profile–profile alignment features includes 10 pairwise features in total.

Structural features. Based on the global profile–profile alignment between query and template obtained with LOBSTER, we use predicted 1D and 2D structural features including secondary structure (3-class: alpha, beta, loop), relative solvent accessibility (2-class: exposed or buried at 25% threshold), contact probability map at 8 and 12 Å, and beta-sheet residue pairing probability map to evaluate the compatibility between query and template structures. These structural features for query proteins are predicted using the SCRATCH suite (Pollastri *et al*., 2001*a*, *b*; Pollastri and Baldi, 2002; Cheng *et al*., 2005; Cheng and Baldi, 2005; http://www.igb.uci.edu/servers/psss.html).

The predicted secondary structure (SS) and relative solvent accessibility (RSA) of the query residues are compared with the nearly exact SS and RSA of the aligned residues in the template structure. The fractions of correct matches for both SS [as in Jones (1999), Xu *et al*. (2003)] and RSA are used as two pairwise features. The SS and RSA composition (helix, strand, coil, exposed and buried) are transformed into four similarity scores by cosine, correlation, Gaussian kernel and dot product. So this 1D structural feature subset has six features in total.

For the aligned residues of the template which have sequence separation >5 and are in contact at 8 Å threshold (resp. 12 Å), we compute the average contact probability of their counterparts in the predicted 8 Å (resp. 12 Å) contact probability map of the query. The underlying assumption is that the counterparts of the contact residues in the template should have high contact probability in the query contact map if the query and template share similar structure. Similarly, for each paired beta-strand residues in the template structures, we compute the average pairing probability of their beta-strand counterparts in the predicted beta-strand paring probability map of the query, assuming that two proteins will share similar beta-sheet topology if they belong to the same fold.

Moreover, we compute the contact order (sum of sequence separation of contacts) and contact number (number of contacts) for each aligned residue in both query and templates. This information is easy to derive for the template sequences since their tertiary structure is known. For the query sequence, we let the contact order for residue i to be $\sum_{|i-j|>5} C_{ij}|i-j|$, where C_{ij} is the predicted contact probability for residues i and j. The contact number for residue i in the query is defined as the sum of the contact probabilities $\sum_{|i-j|>5} C_{ij}$. The contact order and contact number vectors of the aligned residues are not used directly as features. Instead, they are compared and transformed into pairwise similarity scores using the cosine and correlation functions. For both the 8 and 12 Å contact maps, eight pairwise features of contact order and contact number are extracted. So the 2D structural feature subset has 11 features in total. Thus the entire 1D and 2D structural feature subset has 17 features in total.

The entire feature set contains 54 pairwise features measuring query-template similarity (Table 1). Initially we used a larger set of 74 features (data not shown) that also included non-pairwise features, such as the proportion of helices and strands in each chain. Information gain analysis (see Results) and experiments led us to remove the 20 least informative or most biased features to optimize performance. All the alignment tools for extracting pairwise features are run with default parameters, except the *e*-value thresholds (-*e* option) of PSI-BLAST, RPS-BLAST and IMPALA which are set to larger values (100, 50 and 20 respectively), to ensure that alignments between sequences with very little similarity are generated in most cases. If no features are generated by these tools, the corresponding similarity features are set to 0.

2.2 Fold recognition with support vector machines

Each feature vector associated with a pair of proteins in a given training set correspond to a positive or negative example, depending on whether the two proteins are in the same fold or not. These feature vectors in turn can be used to train a binary classifier. Here we train SVMs and learn an optimal decision function $f(x)$ to classify an input feature vector x into two categories ($f(x) > 0$: same fold; $f(x) < 0$: different fold). The decision function $f(x) = \sum_{x_i \in S} \alpha_i y_i K(x, x_i) + b$ is a weighted linear combination of the similarities $K(x_i, x)$ between the input feature vector x and the feature vectors x_i in the training dataset S. Here K is a user-defined kernel function that measures the similarity between the feature vectors x_i and x corresponding in general to four proteins. α_i is the weight assigned to the training feature vector x_i and y_i is the corresponding label (+1:positive, −1:negative). All protein pairs in the same fold are labeled as positive examples, and the remaining ones as negative examples. We use SVM-light (Joachims, 1999) to learn the SVM parameters. The continuous value $f(x)$ is indicative of how likely the corresponding sequences are in the same fold, and therefore it is used to evaluate the structural relevance and rank all the templates for a given query. We tested polynomial, tanh and Gaussian radial basis kernels (RBF: $e^{-\gamma||x-y||^2}$). We report the results obtained with the RBF kernel which worked best for this task, with $\gamma = 0.015$. Preliminary tests indicated that the results are robust with respect to γ. All other SVM parameters are set to their default values. A thorough parameter optimization may help further improve the accuracy.

2.3 Training and benchmarking

To compare the performance of our method with other well-established methods, we use the large benchmark dataset (Lindahl and Elofsson, 2000) derived from the SCOP (Murzin *et al*., 1995) database. The Lindahl's dataset includes 976 proteins. The pairwise sequence identity is ≤40%. We extract a feature vector for all 976 × 975 distinct pairs. In this dataset, 555 sequences have at least one match at the family level, 434 sequences have at least one match at the superfamily level and 321 sequences have at least one match at the fold level.

We split all protein pairs evenly into 10 subsets for 10-fold cross validation purposes. All the query–template pairs associated with the same query protein are put into the same subset. Nine subsets are used for training and the remaining subset is used for validation. The pairs in the training dataset that use queries in the test dataset as templates are removed. The procedure is repeated 10 times and the sensitivity/specificity results are computed across the 10 experiments. Training takes about 3 days for a single data-split on a single node with dual Pentium processors, hence 3 days for the entire 10-fold cross-validation experiment using 10 nodes in a cluster. Using the same evaluation procedure as in Lindahl and Elofsson (2000), Shi *et al*. (2001) and Zhou and Zhou (2004), we evaluate the sensitivity by taking the top 1 or the top 5 templates in the ranking associated with each test query. Furthermore, as in Lindahl and Elofsson (2000) and Shi *et al*. (2001), we also evaluate the performance of our method for all positive matches using specificity-sensitivity plots.

Table 2. The 20 top-ranked features using information gain

Feature	Information gain
HHSearch score	0.0375
COMPASS *e*-value	0.0370
PRC reverse score on chk profile	0.0354
PRC reverse score on HMM profile	0.0341
HMMer pfam *e*-value	0.0287
Dot product of SS and RSA vectors	0.0266
HMMer search *e*-value	0.0264
SS match ratio	0.0263
Correlation of SS and RSA vectors	0.0263
PRC simple score on HMM profile	0.0248
Cosine of SS and RSA vectors	0.0246
Gaussian kernel on SS and RSA vectors	0.0237
COMPASS score	0.0235
PRC coemis score on HMM profile	0.022
PSI-BLAST *e*-value	0.0205
IMPALA *e*-value	0.0181
RPS-BLAST *e*-value	0.0180
SA match ratio	0.0154
Cosine of residue contact num (8 Å)	0.0150
HMMer search score	0.0142

3 RESULTS

Table 2 lists the 20 top features ranked using the information gain measure (Yang and Pedersen, 1997) (complete table for the 54 features is available as Supplementary Materials). The table shows that profile–profile alignment features are the most informative. For instance, the alignment features of HHSearch, COMPASS and PRC are ranked first, second and third respectively. Thus profile–profile alignment methods have the strongest discriminative power in fold recognition, consistently with previous studies (Rychlewski *et al*., 2000; Wallner *et al*., 2004; Ohlson *et al*., 2004). Profile–sequence (or sequence–profile) alignment features and some structural features based on the LOBSTER alignment between queries and templates have also strong discriminative power according to the information gain measure. For instance, the *e*-values of HMMer pfam and HMMer search are ranked fifth and seventh respectively. Our results, confirm also the importance of predicted structural features. The dot product of secondary structure and solvent accessibility composition vectors, and the secondary structure match ratio, rank sixth and eighth respectively.

Other profile–sequence (or sequence–profile) alignment features such as PSI-BLAST, IMPALA, BLAST and structural features such as the cosine of the residue contact number lead also to significant information gains. On the other hand, compared with other local profile–profile alignment scores, the CLUSTALW global profile–profile alignment score carries a lesser weight. This suggest that CLUSTALW is optimized for alignment, but not for direct fold recognition, which is consistent with previous results (Marti-Renom *et al*., 2004). Since the pairwise sequence identity in the dataset is <40%, sequence alignment and sequence/family information features have a lesser, albeit still noticeable, impact.

We evaluate the performance of our FOLDpro method against 11 other fold recognition methods. The 11 other methods are PSI-BLAST, HMMER, SAM-T98 (Karplus *et al*., 1998), BLASTLINK,

Table 3. The sensitivity of 12 methods on the Lindahl's benchmark dataset at the family, superfamily, and fold levels

Method	Family (%)		Superfamily (%)		Fold (%)	
	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5
PSI-BLAST	71.2	72.3	27.4	27.9	4.0	4.7
HMMER	67.7	73.5	20.7	31.3	4.4	14.6
SAM-T98	70.1	75.4	28.3	38.9	3.4	18.7
BLASTLINK	74.6	78.9	29.3	40.6	6.9	16.5
SSEARCH	68.6	75.5	20.7	32.5	5.6	15.6
SSHMM	63.1	71.7	18.4	31.6	6.9	24.0
THREADER	49.2	58.9	10.8	24.7	14.6	37.7
FUGUE	82.2	85.8	41.9	53.2	12.5	26.8
RAPTOR	75.2	77.8	39.3	50.0	25.4	45.1
SPARKS	81.6	88.1	52.5	69.1	24.3	47.7
SP^3	81.6	86.8	55.3	67.7	28.7*	47.4
FOLDpro	85.0*	89.9*	55.5*	70.0*	26.5	48.3*

*denotes the best results.

SSEARCH, SSHMM (Hargbo and Elofsson, 1999), THREADER (Jones *et al*., 1992), FUGUE (Shi *et al*., 2001), RAPTOR (Xu *et al*., 2003), SPARKS (Zhou and Zhou, 2004) and SP^3 (Zhou and Zhou, 2005). SPARKS, for instance, was one of the top predictors during the sixth edition of the CASP evaluation (Moult *et al*., 2005). The results for PSI-BLAST, HMMER, SAM-T98, BLASTLINK, SSEARCH, SSHMM and THREADER are taken from Lindahl and Elofsson (2000). The results for the other methods are taken from the corresponding articles. One caveat is that the sequence databases used to generate the profiles are being updated continuously, and so are some of the methods. Thus the comparative analysis is only meant to provide a broad, rough assessment of performance rather than a precise and stable ranking.

Table 3 shows the sensitivity of FOLDpro and the other methods at the family, superfamily and fold levels, for the top 1 and top 5 predictions respectively. Here sensitivity is defined by the percentage of query proteins (with at least one possible hit) having at least one correct template ranked first, or within the top 5 (Lindahl and Elofsson, 2000). It shows that in almost all situations the performance of FOLDpro is better than that of other well-established methods such as SPARKS, SP^3, FUGUE and RAPTOR.

Specifically, at the family level, the sensitivity of FOLDpro for the top 1 or 5 predictions is 85.0 and 89.9%, about 2–4% higher than FUGUE, SPARKS and SP^3, and significantly higher than all other methods. At the superfamily level, the sensitivity of FOLDpro for the top 1 or 5 predictions is 55.5 and 70.0%, slighlty higher than SPARKS and SP^3 and significantly higher than all other methods. At the fold level, the sensitivity of FOLDpro for the top 1 predictions is 26.5%, about 2% lower than SP^3, 1–3% higher than RAPTOR and SPARKS, and significantly higher than all other methods. For the top 5 predictions, at the fold level, the sensitivity of FOLDpro is 48.3%, about 0.6–3% higher than RAPTOR, SPARKS and SP^3, and significantly higher than all other methods.

The performance of FOLDpro is significantly better than pure sequence- or profile-based approaches, such as PSI-BLAST, HMMER, SAM-T98 and BLASTLINK. It is also significantly better than threading approaches, such as THREADER, in all three categories. For example, compared with PSI-BLAST, FOLDpro is

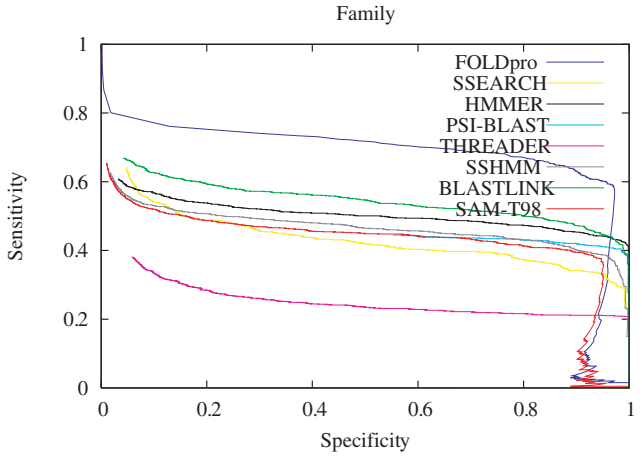

Fig. 1. Specificity–sensitivity plot at the family level.

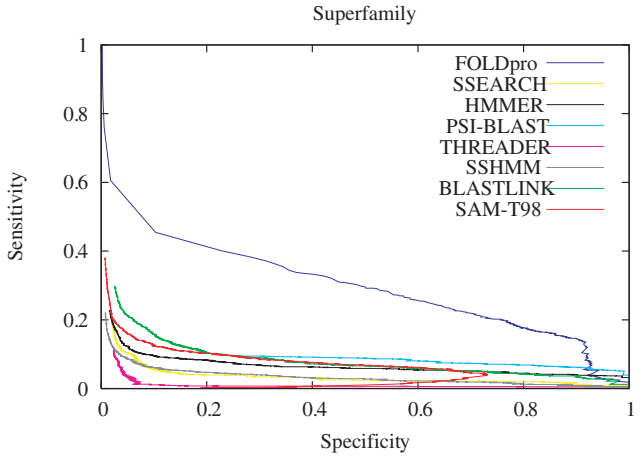

Fig. 2. Specificity–sensitivity plot at the superfamily level.

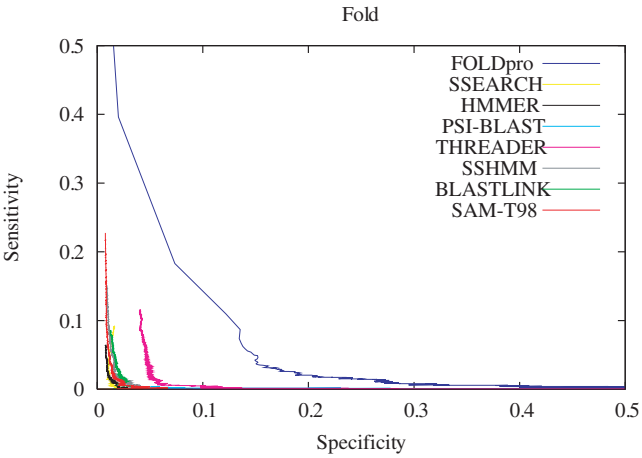

Fig. 3. Specificity–sensitivity plot at the fold level.

about 14, 28 and 23% more sensitive at recognizing members of the same family, superfamily and fold, respectively, using the top 1 predictions; using the top 5 predictions, these improvements are 18, 42 and 44% respectively.

As in Lindahl and Elofsson (2000), we also compare the performance of FOLDpro using specificity–sensitivity plots (Fig. 1–3), to better assess the trade-offs between specificity and sensitivity, using the Lindahl's dataset. We compute the sensitivity and specificity of FOLDpro for different thresholds applied to the SVM scores. Specificity is defined as the percentage of predicted positives (above threshold) that are true positives (in the same family, superfamily or fold). Sensitivity is defined as the percentage of true positives that are predicted as positives (above threshold). The advantage of the specificity–sensitivity plots is that they measure the ability of a method to reliably identify all positive matches in the dataset beyond the top hits. Sensitivity–specificity results for 7 of the 11 methods above were kindly provided by Dr Elofsson (http://www.sbc.su.se/~arne/protein-id/).

In the family category (Fig. 1), FOLDpro consistently outperforms all other methods by >10% for almost all specificity values. However, like SAM-T98, the sensitivity of FOLDpro drops rapidly when the specificity is close to 1. This suggests that some false positives may be receiving very high scores. However, after manually inspecting the dozen of 'false positives' with high scores, some of them turn out to be true positives that were misclassified in the original dataset. For instance, the pair (1XZL,3TGL) belongs to the same superfamily and fold (alpha/beta-Hydroplases) in the latest SCOP 1.69 release, while 1XZL was wrongly classified into another fold (Flavodoxin-like) in the Lindahl's dataset based on the old SCOP 1.37 release. This shows that FOLDpro is capable of correcting some human annotation errors and that 'false positives' with high scores must be verified carefully. Although these wrongly classified pairs with high scores lead one to slightly under-estimate the performance of FOLDpro, we did not attempt to correct them in the evaluation, because of their small effect and to maintain consistency with previous evaluations.

At the superfamily level (Fig. 2), FOLDpro has more than twice the sensitivity of the second best method for almost all specificity levels. For instance, at 50% specificity, the sensitivity of FOLDpro is 30%, ∼20% higher than the second best method, PSI-BLAST. At the fold level (Fig. 3), fold recognition remains challenging for all methods. However, FOLDpro's performance is significantly better than all other methods, including the second best method THREADER, a threading method specifically designed for this purpose. For instance, at 5% specificity, FOLDpro achieves sensitivity of 28%, ∼23% higher than THREADER, while the sensitivity of all other methods is close to 0.

The specificity–sensitivity plots show that FOLDpro significantly outperforms a variety of different methods in all categories, indicating that the integration of complementary alignment tools and sequence and structural information can improve fold recognition across the board.

4 DISCUSSION

We have presented a general information retrieval framework for the fold recognition problem that leverages similarity methods at two fundamental levels. Rather than directly classifying individual proteins, we first consider pairs of proteins and derive a set of pairwise features (feature vector) consisting of many different similarity scores (e.g. profile–profile alignment scores). We then apply supervised classification methods (e.g. SVM) to these feature vectors to learn a relevance function to measure whether or not the query–template pairs are structurally relevant (same versus different fold). For a given query, the continuous relevance values are used to rank the templates.

The learning process involves measuring the similarity between pairs of feature vectors associated with four proteins, which differs from the two-protein comparison of traditional classification approaches. From the standpoint of using structural information in fold recognition, our approach differs also from traditional threading approaches, which use structural information to produce alignments and compute statistical contact potentials to evaluate sequence–structure fitness. In contrast, our approach employs sequence-based profile–profile alignment tools to align a query against the possible templates, without using structural information. Then, based on these alignments, it checks the predicted secondary structure, solvent accessibility, contact probability map and beta-sheet pairings of the query against the template structures to evaluate fitness.

The approach used in FOLDpro has several advantages in terms of integration, scalability, simplicity, reliability and performance. First the approach readily integrates complementary streams of information, from alignment to structure, and additional features can easily be added. It is worth pointing out this integrative approach is slower than some individual alignment methods such as PSI-BLAST. However, it can scan a fold library with about 10 000 templates in a few hours, for an average-size query protein, on a server with two Pentium processors. Second, most features can readily be derived using publicly available tools. This is simpler than trying to develop a new, specialized, alignment tool for fold recognition as in SPARKS, SP^3, FUGUE and RAPTOR, which usually requires a lot of expertise. Third, our approach can be included in a meta server but, unlike a meta-server, it is self-contained and does not rely on external fold-recognition servers. Unlike meta servers, this approach produces a full ranking of all the templates and does not discard any templates early on during the recognition process. Finally, the approach delivers state-of-the-art performance on current benchmarking datasets. And while fold recognition remains a challenging problem, the approach provides clear avenues of exploration to improve the performance, such as adding new features to the feature vector, enlarging the training set, using different machine learning tools to learn the relevance function and leveraging ensembles.

Conflict of Interest: none declared.

REFERENCES

Abagyan,R. *et al.* (1994) Recognition of distantly related proteins through energy calculations. *Proteins*, **19**, 132–140.

Al-Lazikani,B. *et al.* (1998) Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to sh2 domains of janus kinase. *Proc. Natl Acad. Sci. USA*, **98**, 14796–14801.

Altschul,S. *et al.* (1990) Basic local aligment tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bailey,T. and Gribskov,M. (1997) Score distributions for simultaneous matching to multiple motifs. *J. Comput. Biol.*, **4**, 45–59.

Baldi,P. *et al.* (1994) Hidden markov models of biological primary sequence information. *Proc. Natl Acad. Sci. USA*, **91**, 1059–1063.

Bowie,J. *et al*. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.

Bryant,S. and Lawrence,C. (1993) An empirical energy function for threading protein sequence through the folding motif. *Proteins*, **16**, 92–112.

Cheng,J. and Baldi,P. (2005) Three-stage prediction of protein beta-sheets by neural networks, alignments, and graph algorithms. *Bioinformatics*, **21** (suppl. 1), i75–i84.

Cheng,J. *et al*. (2005) SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res*., **33**, w72–w76.

David,R. *et al*. (2000) 3D–1D threading methods for protein fold recognition. *Pharmacogenomics*, **1**, 445–455.

Dayhoff,M. *et al*. (1983) Establishing homologies in protein sequences. *Methods Enzymol*, **91**, 524–545.

Domingues,F. *et al*. (2000) Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J. Mol. Biol*., **297**, 1003–1013.

Eddy,S. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Edgar,R. and Sjolander,K. (2003) Simultaneous sequence alignment and tree construction using hidden Markov models. *Bioinformatics*, **19**, 1404–1411.

Edgar,R. and Sjolander,K. (2004) COACH: profile–profile alignment of protein families using hidden markov models. *Bioinformatics*, **20**, 1309–1318.

Elofsson,A. *et al*. (1996) A study of combined structure/sequence profiles. *Fold Des*., **1**, 451–461.

Fischer,D. (2000) Hybrid fold recognition: combining sequence derived properties with evolutionary information. In Altman,R., Dunker,A., Hunter, Lauderdale,K. and Klein,T. (eds), *Pacific Symposium Biocomputing*. World Scientific, New York, pp. 119–130.

Fischer,D. (2003) 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. *Proteins*, **51**, 434–441.

Ginalski,K. *et al*. (2003a) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, **19**, 1015–1018.

Ginalski,K. *et al*. (2003b) ORFeus: detection of distant homology using sequence profiles and predicted secondary structure. *Nucleic Acids Res*., **31**, 3804–3807.

Godzik,A. and Skolnick,J. (1992) Sequence–structure matching in globular proteins: application to supersecondary and tertiary structure determination. *Proc. Natl Acod. Sci. USA*, **89**, 12098–12102.

Gough,J. *et al*. (2001) Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. *J. Mol. Biol*., **313**, 903–919.

Griffiths-Jones,S. and Bateman,A. (2002) The use of structure information to increase alignment accuracy does not aid homologue detection with profile hmms. *Bioinformatics*, **18**, 1243–1249.

Hargbo,J. and Elofsson,A. (1999) A study of hidden markov models that use predicted secondary structures for fold recognition. *Proteins*, **36**, 68–87.

Henikoff,S. and Henikoff,J. (1992) Amino acid substitutes matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Hughey,R. and Krogh,A. (1996) Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Comput. Appl. Biosci*., **12**, 95–107.

Jaakkola,T. *et al*. (2000) A discriminative framework for detecting remote protein homologies. *J. Comput. Biol*., **7**, 95–114.

Jaroszewski,L. *et al*. (1998) Fold prediction by a hierarchy of sequence, threading, and modeling methods. *Protein Sci*., **7**, 1431–1440.

Joachims,T. (1999) In Schölkopf,B., Burges,C. and Smola,A. (eds), *Making large-scale SVM Learning Practical. Advances in Kernel Methods—Support Vector Learning*. MIT Press, Cambridge, MA.

Jones,D. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol*., **287**, 797–815.

Jones,D. *et al*. (1992) A new approach to protein fold recognition. *Nature*, **358**, 86–98.

Juan,D. *et al*. (2003) A neural network approach to evaluate fold recognition results. *Proteins*, **50**, 600–608.

Karplus,K. *et al*. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–846.

Kelley,L. *et al*. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol*., **299**, 499–520.

Kim,D. *et al*. (2003) PROSPECT ii: protein structure prediction method for genome-scale applications. *Protein Eng*., **16**, 641–650.

Koretke,K.K. *et al*. (2001) Fold recognition from sequence comparisions. *Proteins*. (Suppl. 5), 68–75.

Krogh,A. *et al*. (1994) Hidden Markov models in computational biology. Appllications to protein modeling. *J. Mol. Biol*., **235**, 1501–1531.

Leslie,C. *et al*. (2002) The spectrum kernl: a string kernel for SVM protein classification. *Pac Symp Biocomput*., 564–575.

Lindahl,E. and Elofsson,A. (2000) Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol*., **295**, 613–625.

Lundstrm,J. *et al*. (2001) Pcons: a neural network based consensus predictor that improves fold recognition. *Protein Sci*., **10**, 2354–2362.

Madera,M. and Gough,J. (2002) A comparison of profile hidden markov model procedures for remote homology detection. *Nucleic Acids Res*., **30**, 4321–4328.

Marti-Renom,M. *et al*. (2004) Alignment of protein sequences by their profiles. *Protein Sci*., **13**, 1071–1087.

Mitelman,D. *et al*. (2003) Probabilistic scoring measures for profile–profile comparison yield more accurate short seed alignments. *Bioinformatics*, **19**, 1531–1539.

Moult,J. *et al*. (2005) Critical assessment of methods of protein structure prediction (CASP)—round VI. *Proteins*, **61** (Suppl. 7), 3–7.

Murzin,A. and Bateman,A. (1997) Distance homology recognition using structural classification of proteins. *Proteins*, (Suppl. 1), 105–112.

Murzin,A. *et al*. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*., **247**, 536–540.

Needleman,S. and Wunsch,C. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol*., **48**, 443–453.

Notredame,C. *et al*. (2000) T-Coffee: a novel method for multiple sequence alignment. *J. Mol. Biol*., **302**, 205–217.

Ohlson,T. *et al*. (2004) Profile–profile methods provide improved fold—recognition. a study of different profile-profile alignment methods. *Proteins*, **57**, 188–197.

O'Sullivan,O. *et al*. (2004) 3DCoffee: combing protein sequences and structures within multiple sequence alignment. *J. Mol. Biol*., **340**, 385–395.

Page,L., Brin,S., Motwani,R. and Winograd,T. (1998) The PageRank citation ranking: Bringing order to the web. *Technical report*, Stanford University, Standford, CA.

Panchenko,A. *et al*. (2000) Combination of threading potentials and sequence profiles improves fold recognition. *J. Mol. Biol*., **296**, 1319–1331.

Park,J. *et al*. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol*., **284**, 1201–1210.

Pearson,W. and Lipman,D. (1988) Improved tools for biological sequences analysis. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Pettitt,C. *et al*. (2005) Improving sequence-based fold recognition by using 3d model quality assessment. *Bioinformatics*, **21**, 3509–3515.

Pollastri,G. and Baldi,P. (2002) Predition of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics*, **18** (Suppl. 1), S62–S70.

Pollastri,G. *et al*. (2001a) Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, **47**, 142–153.

Pollastri,G. *et al*. (2001b) Improving the prediction of protein secondary strucure in three and eight classes using recurrent neural networks and profiles. *Proteins*, **47***, 228–235*.

Rocchio,J.J. (1966) Document retrieval systems—optimization and evaluation. Ph.D Thesis, Harvard University, Cambridge, MA.

Rost,B. and Sander,C. (1997) Protein fold recognition by prediction-based threading. *J. Mol. Biol*., **270**, 471–480.

Rychlewski,L. *et al*. (2000) Comparison of sequence profiles. strategies for structural predictions using sequence information. *Protein Sci*., **9**, 232–241.

Sadreyev,R. and Grishin,N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol*., **326**, 317–336.

Schaffer,A. *et al*. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.

Schölkopf,B. and Smola,A.J. (2002) *Learning with Kernels, Support Vector Machines, Regularization, Optimization and Beyond*. MIT University Press, Cambridge, MA.

Shan,Y. *et al*. (2001) Fold recognition and accurate query-template alignment by a combination of PSI-BLAST and threading. *Proteins*, **42**, 23–37.

Shi,J. *et al*. (2001) FUGUE: sequence–structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol*., **310**, 243–257.

Skolnick,J. and Kihara,D. (2001) Defrosting the frozen approximation: PROSPECTOR-a new approach to threading. *Proteins*, **42**, 319–331.

Smith,T. and Waterman,M. (1981) Identification of common molecular subsequences. *J. Mol. Biol*., **147**, 195–197.

Söding,J. (2005) Protein homology detection by HMM–HMM comparison. *Bioinformatics*, **21**, 951–960.

Tang,C. *et al*. (2003) On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. *J. Mol. Biol*., **334**, 1043–1062.

Thompson,J. *et al*. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.

Vapnik,V. (1998) *Statistical Learning Theory*. Wiley, New York, NY.

Vingron,M. and Waterman,M. (1994) Sequence alignment and penalty choice. review of concepts, case studies and implecations. *J. Mol. Biol*., **235**, 1–12.

Ohsen,N. *et al*. (2003) Profile–profile alignment: a powerful tool for protein structure prediction. *Pac. Symp. Biocomput*., 252–263.

Wallner,B. *et al*. (2004) Using evolutionary information for the query and target improves fold recognition. *Proteins*, **54**, 342–350.

Wang,G. and Dunbrack,R.J. (2004) Scoring profile–profile sequence alignments. *Protein Sci*., **13**, 1612–1626.

Xu,J. *et al*. (2003) Protein structure prediction by linear programming. *Pac Symp Biocomput*., 264–275.

Xu,Y. *et al*. (1998) An efficient computational method for globally optimal threadings. *J. Comput. Biol*., **5**, 597–614.

Yang,Y. and Pedersen,J. (1997) A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML97),* Morgan Kaufmann, San Francisco, CA, pp. 412–420.

Yona,G. and Levitt,M. (2002) Within teh twilight zone: a sensitive profile–profile comparison tool based on information theory. *J. Mol. Biol*., **315**, 1257–1275.

Zhou,H. and Zhou,Y. (2004) Single-body residue-level knowledge-based energy score combined with sequence–profile and secondary structure information for fold recognition. *Proteins*, **55**, 1005–1013.

Zhou,H. and Zhou,Y. (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins*, **58**, 321–328.