BIOINFORMATICS

Vol. 17 no. 1 2001
Pages 63–72

What are the baselines for protein fold recognition?

Liam J. McGuffin[1], Kevin Bryson[2] and David T. Jones[1,]*

[1]Bioinformatics Group, Department of Biological Sciences, Brunel University, Uxbridge UB8 3PH, UK and [2]Agent-Based Systems Group, Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK

Received on April 20, 2000; revised on July 19, 2000; accepted on September 23, 2000

ABSTRACT

Motivation: What constitutes a baseline level of success for protein fold recognition methods? As fold recognition benchmarks are often presented without any thought to the results that might be expected from a purely random set of predictions, an analysis of fold recognition baselines is long overdue. Given varying amounts of basic information about a protein—ranging from the length of the sequence to a knowledge of its secondary structure—to what extent can the fold be determined by intelligent guesswork? Can simple methods that make use of secondary structure information assign folds more accurately than purely random methods and could these methods be used to construct viable hierarchical classifications?

Experiments performed: A number of rapid automatic methods which score similarities between protein domains were devised and tested. These methods ranged from those that incorporated no secondary structure information, such as measuring absolute differences in sequence lengths, to more complex alignments of secondary structure elements. Each method was assessed for accuracy by comparison with the Class Architecture Topology Homology (CATH) classification. Methods were rated against both a random baseline fold assignment method as a lower control and FSSP as an upper control. Similarity trees were constructed in order to evaluate the accuracy of optimum methods at producing a classification of structure.

Results: Using a rigorous comparison of methods with CATH, the random fold assignment method set a lower baseline of 11% true positives allowing for 3% false positives and FSSP set an upper benchmark of 47% true positives at 3% false positives. The optimum secondary structure alignment method used here achieved 27% true positives at 3% false positives. Using a less rigorous Critical Assessment of Structure Prediction (CASP)-like sensitivity measurement the random assignment achieved 6%, FSSP—59% and the optimum secondary structure alignment method—32%. Similarity trees produced by the optimum method illustrate that these methods cannot be used alone to produce a viable protein structural classification system.

Conclusions: Simple methods that use perfect secondary structure information to assign folds cannot produce an accurate protein taxonomy, however they do provide useful baselines for fold recognition. In terms of a typical CASP assessment our results suggest that approximately 6% of targets with folds in the databases could be assigned correctly by randomly guessing, and as many as 32% could be recognised by trivial secondary structure comparison methods, given knowledge of their correct secondary structures.

Contact: David.Jones@brunel.ac.uk

INTRODUCTION

As the gap widens between the number of known sequences and the number of experimentally determined protein structures, the pressure has never been greater to develop rapid, fully automated fold prediction methods. Currently fold recognition methods such as THREADER 2 (Jones *et al.*, 1992, 1999) ProCeryon (Domingues *et al.*, 1999) and the method developed by Panchenko *et al.* (1999) for example have limited accuracy, and often interpretation of results is not automated. Methods such as 3D-PSSM (Kelley *et al.*, 2000), SAM-T98 (Karplus *et al.*, 1999), GenTHREADER (Jones, 1999b) and others tested at Critical Assessment of Fully Automated Structure Prediction CAFASP-1 (Fischer *et al.*, 1999) are fast fold prediction methods designed to automatically screen genomic databases. However, given that these methods are intended to be used automatically, it is essential that they are evaluated properly. In order to test the limitations of these methods, benchmarking schemes have been developed (e.g. Fischer *et al.*, 1996; Domingues *et al.*, 2000), however, a proper evaluation of the random baselines for fold recognition has not yet been carried out. This is in contrast with the secondary

*To whom correspondence should be addressed.

© Oxford University Press 2001

structure prediction field, where the random prediction baselines are easily calculated and widely understood.

Given recent improvements in protein secondary structure prediction methods, some groups have attempted to improve the speed of fold prediction by developing methods that incorporate predicted secondary structure information (Russell *et al.*, 1996; Rice and Eisenberg, 1997; Di Francesco *et al.*, 1999). It is thought that these methods may offer an advantage over methods that rely on primary sequence alone (Di Francesco *et al.*, 1999). However, methods that recognise fold directly from secondary structure prediction have not as yet proved to be superior to the best threading methods (Murzin, 1999).

Contrary to the argument that predicted secondary structure offers no real advantage to fold prediction, some groups have put forward the conjecture that protein secondary structure is in fact the major factor determining three-dimensional fold. Based on this conjecture Przytycka *et al.* (1999) have attempted to develop a protein taxonomy by constructing similarity trees based on simple pairwise alignments of secondary structure elements within a set of 183 proteins of known structure. From their results they deduce that pairwise alignments of secondary structure elements may be an effective basis for protein classification.

In this paper we evaluate a number of trivial fold recognition methods ranging from random fold assignment, to methods which consider the order and lengths of secondary structure prediction elements. The first problem tackled in this paper is to establish a set of baselines for fold recognition methods to help in the future to identify automated methods which are capable of producing results well above the random level. The other question addressed here is to ask whether simple methods that make use of secondary structure information can assign folds more reliably than other random methods and further how valuable these methods might be in the rapid construction of useful hierarchical classifications.

METHODS

A number of automated methods ranging in complexity and speed were used to score pairwise similarities between protein domains. Pairs of domains with high similarity scores were taken to indicate proteins of similar fold. Each method was assessed by comparing it with the Class Architecture Topology Homology (CATH) protein structural classification database assignment of fold (Orengo *et al.*, 1997, http://www.biochem.ucl.ac.uk/bsm/cath/). For each method true positives were taken as pairs of domains with high similarity scores found to have the same topology according to CATH. Conversely, false positives were taken as pairs of domains with high similarity scores found to have different topologies according to CATH. The percentage of true positives—taken at a cut-off of 3% false positives (see Section **Results**)—for each method were compared. A less rigorous measurement of the sensitivity of each method—the percentage of correctly assigned top scoring folds—was also made.

SIMILARITY SCORING BETWEEN PROTEIN DOMAINS

A representative set of 1087 protein domains with resolutions $\leqslant$2.5 Å was selected from the CATH list of sequence family representatives (S-reps, v1.6) (ftp://ftp.biochem.ucl.ac.uk/pub/cathdata/v1.6/Sreps).

Secondary structure was assigned for each domain using both the DSSP method of Kabsch and Sander (1983) and a backbone dihedral angle method similar to Przytycka *et al.* (1999). Helical residues were taken as those with backbone dihedral angles ($-80° \leqslant \phi \leqslant -40°$, $-65° \leqslant \varphi \leqslant -5°$) or ($-110° \leqslant \phi \leqslant -40°$, $-74° \leqslant \varphi \leqslant -0°$) (N.B. Przytycka *et al.* (1999) allow only isolated residues in the range ($-110° \leqslant \phi \leqslant -40°$, $-74° \leqslant \varphi \leqslant -0°$), however for our data set we have found that it makes no significant changes to our results when we allow all residues in this range). Strand residues were taken as those with backbone dihedral angles ($-180° \leqslant \phi \leqslant -60°$, $60° \leqslant \varphi \leqslant 180°$) or ($-180° \leqslant \phi \leqslant -60°$, $-180° \leqslant \varphi \leqslant -140°$).

For each domain a file was generated containing 4 strings: (1) CATH domain name (four-character PDB code followed by chain identifier and domain number); (2) DSSP amino acid sequence; (3) DSSP assigned secondary structure (Kabsch and Sander, 1983); (4) Backbone dihedral angle assigned secondary structure (Przytycka *et al.*, 1999). A clarifying example follows:

```
>1atx00
GAAaLbKSDGPNTRGNSMSGTIWVFGcPSGWNNbEGRAIIGYacKQ
  EEE TTS S  TTSSEEEEEESS   TT EEE  SSSSSEEEE
CEEEEEHHECEEEECCCECEEEECCCEECCEECEEECCEECEEEEC
```

These files are available as zipped archives from http://insulin.brunel.ac.uk/~mcguffin/baseline.html.

The strings were interpreted so that primary sequence lowercase letters were taken to be cysteine residues, a strand would equal three or more consecutive *E*s and a helix would equal five or more consecutive *H*s. All other secondary structure elements were taken as coil.

Similarity scores were calculated between pairs of domains by the 11 methods listed in Table 2 (see Section **Results**). (N.B. Secondary structure has been interpreted from the 4th string unless otherwise stated. Methods are roughly numbered by increasing complexity and decreasing speed. A detailed explanation of each method used is available at http://insulin.brunel.ac.uk/~mcguffin/baseline.html.)

Table 1. Cut-offs used to assign class to protein domains from secondary structure composition

Assigned class	Class number (CATH)	Percentage helix residues	Percentage strand residues
Mainly alpha	1	$\geqslant 24$	<20
Mainly beta	2	$\leqslant 15$	>20
Alpha/beta	3	>15	$\geqslant 20$

Optimisation of similarity scoring methods using class prediction as a pre-filter

Class was used as a filter to improve the true positive rate of the methods. Two domains were only considered to be similar if they had identical predicted class.

For each domain class was assigned purely from secondary structure composition using a similar method to that of Michie *et al.* (1996). Percentages of residues constituting helices and strands were calculated for each domain within the representative set. The CATH class assignment for each domain was then taken and plots (Figures 1a–c) were produced. The scatter plot of secondary structure composition (Figure 1a) was used to set minimum percentage alpha composition required for alpha class assignment and minimum percentage beta composition required for beta class assignment.

Alpha/beta domains were seen to overlap the composition regions of both alpha and beta domains. Alpha and alpha/beta domain regions were separated by calculating the percentage beta cut-off that would allow the majority of domains to lie in their correct regions. Similarly, beta and alpha/beta domain regions were separated by calculating the percentage alpha cut-off that would allow the majority of domains to lie in their correct regions. The points on the graphs where the lines cross in Figures 1b and c indicate the optimum alpha and beta cut-offs. The cut-offs used to assign class are tabulated in Table 1.

Percentage primary identity filter

Percentage primary sequence identity was calculated between all pairs of domains using method 11 (alignment of primary sequence). The representative set was screened for redundant pairs with sequence identities $>25\%$ prior to comparison of methods with CATH.

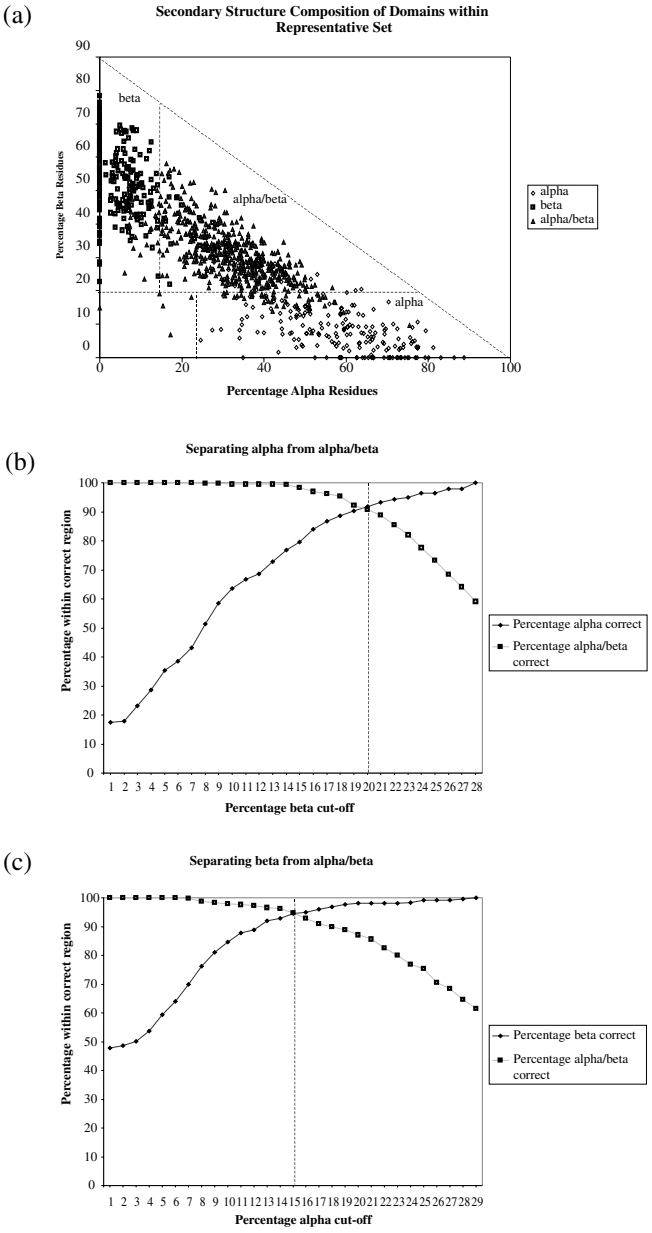

Fig. 1. (a) Secondary structure composition of domains within representative set. Class assignment for each domain has been taken from CATH. Separation lines indicate cut-offs used to assign class as shown in (b) and (c). (b) Calculation of percentage beta cut-off separating alpha domains from alpha/beta domains. (c) Calculation of percentage alpha cut-off separating beta domains from alpha/beta domains.

A RIGOROUS COMPARISON OF SIMILARITY SCORING METHODS WITH CATH

In this first comparison we were concerned with measuring the percentage of true positives at a fixed low false positive percentage.

Each similarity scoring method produced a list of 590 241 ($\frac{1}{2}n(n-1)$, where $n = 1087$) pairs of domains with similarity scores. These lists were sorted by descending similarity score and were read line by line. The CATH topology codes (CAT codes) for each domain within a pair were compared. If the CAT codes were seen to be dissimilar then the number of false positives was

Table 2. Results of assessment of similarity scoring methods by rigorous comparison with CATH. Column (a) shows percentage true positives at a false positive rate of 3% and similarity score for methods 1–11 without using predicted class as a pre-filter. Column (b) shows percentage true positives at a false positive rate of 3% and similarity score for methods 1–11 with predicted class as a pre-filter

Method title	Method number		Similarity score at 3% false positives		Percentage true positives at 3% false positives	
	(a)	(b)	(a)	(b)	(a)	(b)
Alignment of primary sequence	11	11	0.14	0.12	3.56	8.52
Absolute difference in length	1	1	0.97	0.93	6.25	13.28
Absolute difference in number of secondary structure elements	2	2	0.93	0.92	14.49	16.43
Simple alignment of secondary structure elements	3	3	0.74	0.7	12.25	17.70
Alignment of secondary structure elements with absolute difference in length as scoring scheme	9	9	0.47	0.45	15.68	19.89
Alignment of full length secondary structure strings	10	10	0.64	0.63	20.01	21.06
Alignment of secondary structure elements with gap penalty	7	7	0.59	0.58	20.53	21.95
Alignment of secondary structure elements with gap penalty for long elements	8	8	0.66	0.66	25.33	25.91
Alignment of secondary structure elements (Przytycka *et al.*, 1999)	4	4	0.73	0.72	25.91	26.39
Alignment of secondary structure elements using DSSP as secondary structure assignment	6	6	0.72	0.71	25.86	26.45
Alignment of secondary structure elements without additional scoring	5	5	0.68	0.67	26.92	27.18

incremented. Conversely, if a pair of domains was seen to have equal CAT codes the number of true positives was incremented. Thus, the percentage of false positives was taken as the number of pairs of dissimilar CAT codes at the top of the list divided by total number of pairs with dissimilar CAT codes within the whole list. The percentage of true positives was taken as the number of matching CAT codes at the top of the list divided by the total number of matches within the whole list. The accuracy of methods was measured by comparing the percentage of true positives when the percentage of false positives reached 3% (see Section **Results**).

Lower control—random assignment of fold compared to CATH

CAT codes were randomly assigned to each domain without replacement, according to the frequency of each CAT code within the representative set. The randomly assigned folds were compared against the real CATH assignments and the percentage true positives and false positives were calculated. The random simulation was carried out 100 times and the average true positive and false positive percentages were recorded.

In order to test the validity of the random simulation, formulae for the theoretical true positive and false positive values were derived. True positives were calculated by y/x and false positives by $(x - y)/(1 - x)$, where x equals the sum of squares of relative fold frequencies and y equals the sum of cubes of relative fold frequencies.

Upper control—comparison of FSSP with SCOP and CATH

Protein structural classification schemes are not in 100% agreement as shown by a systematic comparison of SCOP (Structural Classification of Proteins Murzin *et al.*, 1995), CATH and FSSP (Families of Structurally Similar Proteins Holm and Sander, 1994) carried out by Hadley and Jones (1999). In order to determine the percentage of true positives that could be achieved by an automated fold assignment method, given knowledge of 3D structure, FSSP was compared against both CATH and SCOP individually. This set a target or upper level of accuracy for fold assignment methods.

FSSP files were downloaded from ftp://ftp.ebi.ac.uk/pub/databases/fssp, the SCOP list from http://scop.mrc-lmb.cam.ac.uk/scop/parse/dir.dom.scop.txt_1.48 and the CATH list from ftp://ftp.biochem.ucl.ac.uk/pub/cathdata/v1.6/cath.list.

The FSSP classification scheme was treated essentially as a similarity scoring method. A list of pairwise comparisons and FSSP Z-scores was compiled for proteins with sequence identity $\leqslant$25%. The FSSP, SCOP and CATH lists were then screened for shared single domains existing in all three databases.

FSSP fold assignments of shared single domains were then compared against SCOP and CATH assignments individually, in the same way as the similarity scoring methods. In each case, the percentage of true positives at 3% false positives was taken as a measurement of accuracy of FSSP (see Section **Results**).

A LESS RIGOROUS CRITICAL ASSESSMENT OF STRUCTURE PREDICTION-LIKE COMPARISON OF SIMILARITY SCORING METHODS WITH CATH

In this second comparison we were interested in assessing the sensitivity of each method. In terms of a typical Critical Assessment of Structure Prediction (CASP) assessment, we may be simply concerned with measuring the probability of a method correctly guessing each fold. Sensitivity values—percentages of correctly assigned top hits—were calculated as follows.

The data set was initially screened for domains with no matching folds so that 'novel folds' were not included in the sensitivity calculations. Each domain was assigned the fold with the highest similarity score or top hit, however, a homologous superfamily filter was imposed. Top hits were only valid if the target domain and the top hit had dissimilar CATH codes to the *H*-level. If two or more hits were found to have the highest score a fold was randomly chosen from them. In order to account for this randomisation, the sensitivity calculation for each method was carried out 100 times and the average value was taken.

Upper and lower controls

Sensitivity scores were calculated for the FSSP upper control—with the homologous superfamily filter—as above. A random proportional fold assignment—with the homologous superfamily filter and with replacement—was carried out for the lower control.

CALCULATION OF DISTANCE MATRICES AND SIMILARITY TREE DRAWING

Distances matrices were calculated for domain pairs. The distance between a pair was defined as one minus their similarity score. Clustering was carried out by inputting distance matrices into the program QCLUST (by John Brzustowski). QCLUST is free to download from ftp://www.biology.ualberta.ca/pub/jbrzusto/trees/. NJPLOT (Perrière and Gouy, 1996) was used in order to produce tree diagrams. NJPLOT is freely available from ftp://pbil.univ-lyon1.fr/pub/mol_phylogeny/njplot/.

RESULTS

Comparison of the percentage true positives attained by similarity scoring methods

The experimental comparison of the random assignment of fold with CATH, or lower control, set a threshold level of 10.79% true positives and 2.88% false positives, hence a cut-off of 3% false positives was set for all other methods. These values are near the theoretical values of 10.99% true positives and 2.86% false positives.

The upper control comparison of FSSP with SCOP set a benchmark level of 61.14% true positives at a *Z*-score of 6.1 and at a rate of 3% false positives. The upper control comparison of FSSP with CATH set a benchmark level of 46.71% true positives at a *Z*-score of 5.8 and at a rate of 3% false positives.

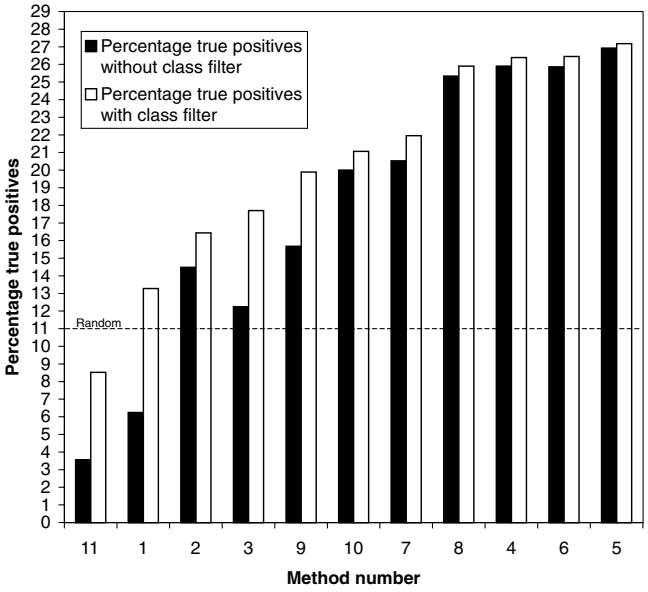

Fig. 2. Percentage true positives at a false positive rate of 3% for methods 1–11 with and without using class as a pre-filter. The dashed horizontal line at 11% represents the threshold of accuracy set by the lower control at 3% false positives.

In Figures 2 and 3 the solid and dashed horizontal lines indicate the approximate true positive levels at 3% false positives set by the upper controls and lower control respectively.

Using percentage true positives at 3% false positives as a measure of accuracy, the most accurate method applied to our data set appears to be number 5—the alignment of secondary structure elements without additional scoring. In most cases, at a 3% false positive percentage, methods that score the similarity of domains by the alignment of secondary structure elements are in better agreement with CATH than other methods. In method 6, using DSSP to assign secondary structure as opposed to using backbone dihedral angles does not appear to significantly affect the overall true positive percentage (Table 2a and Figure 2).

Methods in Table 2b are sorted by increasing accuracy. The order of accuracy of methods does not appear to correspond to their relative speed or complexity as indicated by the non-sequential ordering of the method numbers.

Protein domain class assignment was predicted with an accuracy of 85.2% (926/1087 domains predicted correctly) using the secondary structure composition method.

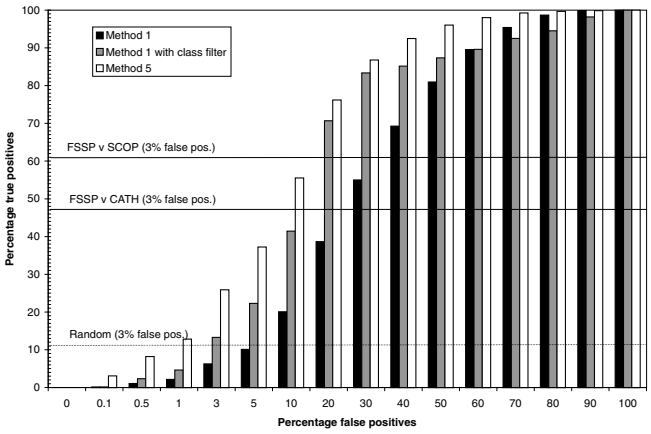

Fig. 3. Percentage true positives versus percentage false positives for method 1, method 1 with class as a pre-filter and method 5. The dashed horizontal line at 11% true positives represents the threshold set by the lower control at 3% false positives. The solid horizontal lines at 61 and 47% true positives represent the benchmarks or targets set by the upper controls at 3% false positives.

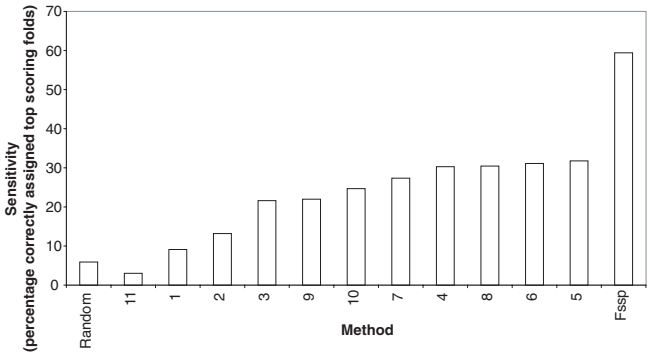

Fig. 4. Sensitivity values (percentage correctly assigned top scoring folds) for random fold assignment, methods 1–11 and FSSP.

The effect of using prediction of class as a pre-filtering stage to all similarity scoring methods is illustrated in Table 2b and Figure 2. There is a significant increase in the mean population percentage of true positives at 3% false positives across all similarity scoring methods at the 5% significance level according to a paired samples *t*-test. The increase in true positives is marked for domain comparison methods that use little or no secondary structure information. Methods that are more complex and incorporate more secondary structure information do not appear to benefit notably from the class pre-filter.

Figure 3 clearly illustrates the effect of using class as a filter to method 1, the 'absolute difference in length' method. The relative accuracy of method 1—with and without the class pre-filter—is compared to method 5—the most accurate method—at increasing percentages of false positives. At $<10\%$ false positives method 5 is considerably more accurate than both method 1 alone and method 1 with the class pre-filter. However when the false positive rate is increased to $\geqslant 10\%$, the effect of the filter becomes more distinct and the gap between method 1 with the class pre-filter and method 5 becomes less extensive. At $\geqslant 20\%$ false positives method 1 with the class pre-filter and method 5 exceed the targets or benchmarks set by the upper controls at 3% false positives. At a rate of $\geqslant 60\%$ false positives the differences in accuracy between methods start to fluctuate and become less apparent.

Assessment of the sensitivity of similarity scoring methods

The random lower control set a threshold sensitivity value of 5.91% and the FSSP upper control set a benchmark sensitivity value of 59.42%. The optimum alignment of secondary structure elements—method 5—achieved a sensitivity value of 31.78%, however this method does not perform significantly better than other secondary structure alignment methods. The order of sensitivity of similarity scoring methods does not correlate with the speed or simplicity of each method (Figure 4).

Similarity trees

Figure 5 shows similarity trees, which compare some of the most common folds within each class. All similarity trees were calculated from distance matrices produced by method 5. Figure 5a shows a complete similarity tree featuring all globin-like domains versus all casein kinase domains. Due to the high number of entries of folds featured in Figures 5b and c, for clarity only sub-trees that are representative of the complete trees are shown.

In Figure 5a there is a distinct separation of globin-like and casein kinase domain folds. Conversely, there are many regions of the sub-tree for immunoglobulin-like folds versus folds represented by thrombin, subunit *H* which show no clear separation of folds (Figure 5b). Figure 5c shows that there are isolated regions where alpha–beta plaits are separated from Rossmann folds although in some cases these folds are not clearly differentiated.

A full similarity tree comparing all 1087 folds within the representative set has been calculated and is available to download as a PostScript file from http://insulin.brunel.ac.uk/~mcguffin/baseline.html. Although there are small areas of this tree where similar folds are clustered the tree is generally disordered.

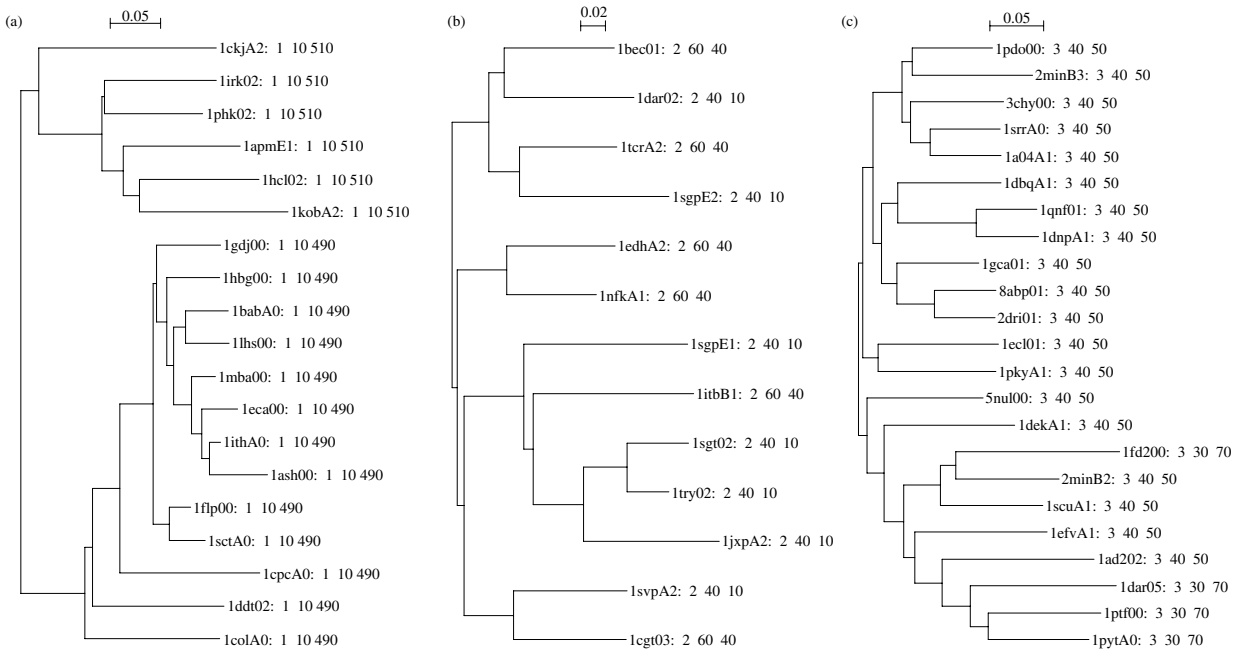

Fig. 5. Similarity trees for selected domains within the representative set. Folds of domains are indicated by CATH topology codes (CAT codes). Branch lengths are proportional to the mean dissimilarity between clusters. (a) Globin-like versus casein kinase (CAT codes '1 10 490' and '1 10 510' respectively), complete tree shown. (b) Immunoglobulin-like versus Thrombin, subunit *H* (CAT codes '2 60 40' and '2 40 10' respectively), sub-tree shown. (c) Alpha–beta plaits versus Rossmann folds (CAT codes '3 30 70' and '3 40 50' respectively), sub-tree shown.

DISCUSSION

In this paper we intended to answer the question 'What are the baselines for fold recognition?' by evaluating a number of simple similarity scoring methods that use varying amounts of structural information. One additional difficulty turned out to be that the question can be answered in different ways, depending on the way the methods are evaluated. We assessed the validity of each method firstly by stringently measuring of the percentage of true positives at a low percentage of false positives, and secondly by a less stringent measurement of the percentage of correctly assigned folds as top hits (sensitivity), which is perhaps the most frequently used evaluation method for fold recognition methods. In addition we assess the suitability of simple secondary structure alignments in classifying folds by comparing them with random and simple fold assignments and by the construction of similarity trees.

The reliability of benchmarks based on SCOP, CATH or FSSP

For routine application of fold recognition methods for, say, genome annotation, it is important that folds be assigned with a low rate of false positives. For this reason it is now common to benchmark methods on the entire set of similarities found in a structure classification resource such as CATH or SCOP. As we wished to compare results to simple random fold selection (according to the observed distribution of folds) the precise false positive rate had to be fixed at 3% for each method to ensure the results were comparable across the board.

In this paper, our 'gold standard' was taken as being the CATH classification scheme. However, it is quite apparent that fold assignment even based on known 3D structures is an inexact science. One surprising result from the analysis here, and which has also been investigated previously in a different way (Hadley and Jones, 1999), is that the agreement between the competing structure classification systems is actually quite low. When evaluated at the same 3% false positive rate as the 'prediction' methods, the highest agreement was found to be between FSSP and SCOP where a true positive rate of only 61% was achieved. As we have discussed before (Hadley and Jones, 1999) this is mainly due to the differences of opinion between the curators of these classification systems: the SCOP team favouring a relatively subjective evolutionary view of fold similarity and the FSSP team favouring an entirely automated approach. In particular the very common doubly-wound alpha/beta folds are hardest hit in this comparison. In SCOP these 'Rossmann-like fold' proteins are split into different fold groups, but FSSP (and

CATH) tends to group them together as being sufficiently similar. Clearly there is no right answer, but the effect of this uncertainty causes severe problems when fold recognition methods are benchmarked. In our opinion, it may be better to benchmark automatic fold recognition methods on a consensus of structural classifications (i.e. just on cases where SCOP, CATH and FSSP are in complete or at least partial agreement). Such a consensus can be obtained from the following web site: http://globin.bio.warwick.ac.uk/~hadley/db. It is fair to say that it makes no sense to penalise a fold recognition method for incorrectly assigning folds which cannot be unambiguously assigned even given the two sets of 3D co-ordinates.

Percentages of true positives at fixed percentages of false positives

Despite the above concerns about the maximum achievable success rates, these concerns are not so important when a single benchmark is used simply to rank different methods. In view of this we picked a single 'gold standard' (CATH) and stuck to it through the evaluations.

The optimum 'baseline' similarity scoring method under these conditions is method 5, the alignment of secondary structure elements without additional scoring. This method achieved 26.92% true positives without the class pre-filter and 27.18% true positives with the class pre-filter. True positives were measured under stringent conditions—false positives were kept below 3% and percentage primary identity between domains was less than 25%. The optimum methods are found to be primarily adaptations of the method originally put forward by Przytycka *et al.* (1999) although slight improvements can be made by the addition of class pre-filters and by disallowing division of helix and strand elements to align with coil.

The addition of the class pre-filtering stage increases the mean population percentage true positives at the 5% significance level using a paired samples *t*-test, implying that this filter is beneficial. The increase in true positives is most striking for methods that do not rely on secondary structure information such as the 'absolute difference in length' (method 1) and the 'alignment of primary sequence' (method 11). The more complex alignment methods that use the most secondary structure information may not benefit so much from the measurement of the secondary structure composition by the class pre-filtering stage, as a similar type of filter may already be inherent in the methods (Figure 2 and Table 2).

When the percentage of true positives is measured less stringently, differences between the more complex methods and the more simplistic methods become less considerable. That is to say that when false positives rates are fixed at 30%, the differences between the 'alignment of secondary structure elements without additional scoring' (method 5) and the 'absolute difference in length' (method 1) with the class pre-filter is negligible. In order to reach the 'FSSP versus CATH' target of 46.71% true positives (at 3% false positives) and 'FSSP versus SCOP' target of 61.14% true positives (at 3% false positives), false positives rates for these methods must increase to beyond 10%. At these higher levels of false positives faster simpler methods such as method 1 with the class pre-filter are seemingly no less accurate than slower more complex methods such as method 5 (Figure 3).

Sensitivity of similarity scoring methods—average percentage correctly assigned folds

Despite the obvious stringency of benchmarking across all pairs of structures in a benchmark set, it is much more common to evaluate fold recognition methods based on a smaller benchmark of structurally similar pairs. A widely known example of this is of course the international CASP experiment (Moult *et al.*, 1999) where truly blind predictions are evaluated. The usual way in which CASP prediction success is stated for fold recognition is simply the percentage of folds correctly recognised. Generally speaking, little attention in this case is paid to false positive rates. When the false positive rates are ignored, it is reasonable to simply take the top scoring fold as the prediction, but in this case of course, every target protein will be predicted, even when no correct answer exists in the data bank of known folds. Despite this, however, such a crude estimate of a methods sensitivity (percentage of folds correctly recognised) is easy for non-specialists to understand, and so it does remain a popular benchmarking metric.

Evaluating the methods on the basis of how many correct folds are found as the best match also puts method 5 in first place with a success rate of 31.78% using our data set, however this value clearly falls very short of the benchmark of 59.42% set by FSSP. These results clearly imply that simple alignment of secondary structure elements cannot be considered a sensitive method for classifying non-homologous folds. These results do suggest, however, that given the perfect secondary structure of a fold target these simple methods can achieve limited success at recognising folds of distantly related protein domains (Figure 4). Certainly an accurate knowledge of the secondary structure of a protein (e.g. from NMR chemical shift analysis) would provide a certain advantage in the fold recognition process.

Differences between similarity trees

Although up until now we have only considered the benchmarking of fold recognition methods, one issue that has been raised in previous work (Przytycka *et al.*, 1999) is that simple similarity methods might be sufficient to

produce a reasonable structural classification of proteins in its own right. This idea, whilst attractive, must be considered rather contrary to the popular belief that useful fold 'taxonomies' can only be produced by detailed analysis of the 3D structures.

The example similarity trees in Figure 5 illustrate that folds with distinctly different secondary structures such as globin-like folds and casein kinase folds can be cleanly separated by simple methods (Figure 5a). The popular belief seems to be wrong in this case. However, when obvious differences in secondary structure become less clear, such as the difference between immunoglobulin-like folds and folds represented by thrombin, subunit *H*, the methods do not distinguish between them as effectively (Figure 5b). It would thus appear that the popular belief is correct for this case.

To explain the difference of opinion, we can look more closely at the two cases. According to CATH, globin-like folds have on average 8 helices and 153 residues and casein kinase folds have on average 12 helices, 4 strands, and 204 residues. At a glance it is apparent that based on differences in number and type of secondary structure elements and sequence length these folds can be easily separated. Conversely, immunoglobulin-like folds have on average 1 helix, 8 strands and 109 residues, similar to folds represented by thrombin, subunit *H* which have on average 2 helices, 7 strands and 101 residues. Clearly in this case, we might expect that a tree produced by a simplistic comparison of secondary structure would be disordered, and indeed this is the case (Figure 5b). Similar problems occur with the separation of alpha–beta plaits from Rossmann folds as shown in Figure 5c. Repeating 'helix–strand–helix-strand' motifs common to both folds and similarity in relative composition of secondary structure may be responsible for disorder in this case. The isolated areas that show separation may be accounted for by differences in average sequence length and distinctive strand–helix–strand–strand–helix–strand' motifs that are common in alpha-beta plaits.

Clearly from these examples it is apparent that simple secondary structure alignments alone can not be relied upon to construct a viable taxonomy of protein folds. Although isolated sub-trees can show separation of folds, the similarity tree for all folds is generally disordered. It must be re-emphasised, however, that as we have shown, surprisingly, even established automated and semi-automated fold classifications systems (FSSP and CATH respectively) can disagree considerably when making pairwise comparisons on single domains of known structure. Again, these results generally agree with the findings of Hadley and Jones (1999).

The basis on which the taxonomy produced by Przytycka *et al.* (1999) is constructed begins to break down when obvious differences in secondary structure between folds become subtler. The main differences that account for the successful cases seems to be simple such as differences in secondary structure composition, sequence length, and the lengths, type and number of secondary structure elements. These may be useful distinguishing characteristics at some level but are not sufficient alone to identify 3D structure in a majority of cases. Przytycka *et al.* (1999) anticipate that their method will improve with the incorporation of information on super secondary structural motifs, and this is likely to be true, but this moves the simple method much closer to a more traditional structure comparison approach.

WHAT ARE THE BASELINES FOR FOLD RECOGNITION?

So what conclusions can be reached as to the baselines for fold recognition? Clearly the simplest answer to this question is to consider the case where folds are randomly assigned in proportion to fold frequencies within current databases. Our results indicate that by randomly assigning folds in proportion to their frequency within our data set, a level of $\sim$11% true positives and $\sim$3% false positives is attained. The upper limit of simple secondary structure alignment methods tested here is $\sim$27% allowing 3% false positives. These true positive limits for simple methods can be treated as baseline levels over which fold recognition techniques must exceed. In terms of a typical CASP assessment (Moult *et al.*, 1999)—where the sensitivity of a method is the prime consideration—our results suggest that on average 6% of targets with folds in the databases would be correctly assigned if folds were assigned randomly and that as many as 32% could be assigned correctly given perfect secondary structures of the targets. Perhaps this rather quantifies the so-called 'Jones rule' discussed by Murzin (1999). Assuming that the proteins considered at CASP are representative, it is reasonable to say that groups making 'bets' according to their knowledge of fold distributions, and the secondary structure and lengths of the proteins concerned do have a reasonable chance of success in CASP-like evaluations. However, this does ignore the thorny problem of generating an accurate sequence to structure alignment at the end of the day.

Although the simple similarity scoring methods tested here are obviously limited in reliability, it is still possible that as they are comparatively fast, they might still have some value as an automatic pre-filtering stage to enhance threading methods. For example they could be used either in combination with or as a replacement for the secondary structure 'masking' stage in THREADER 2 (Jones *et al.*, 1999). In addition they may help to increase the sensitivity of automatic genome annotation methods. One final aspect which we are currently evaluating is the sensitivity of

these methods to the use of predicted secondary structure. Methods for randomly simulating prediction errors are currently being developed in order to assess these effects.

ACKNOWLEDGEMENTS

This work was supported by the BBSRC (LJM).

REFERENCES

Di Francesco,V., Munson,P.J. and Garnier,J. (1999) FORREST: fold recognition from secondary structure predictions of proteins. *Bioinformatics*, **15**, 131–140.

Domingues,F.S., Koppensteiner,W.A., Jaritz,M., Prlic,A., Weichenberger,C., Wiederstein,M., Floeckner,H., Lackner,P. and Sippl,M.J. (1999) Sustained performance of knowledge based potentials in fold recognition. *Proteins Struct. Funct. Genet.*, **3** (Suppl.), 112–120.

Domingues,F.S., Lackner,P., Andreeva,A. and Sippl,M.J. (2000) Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J. Mol. Biol.*, **297**, 1003–1013.

Fischer,D., Elofsson,A., Rice,D.W., LeGrand,S. and Eisenberg,D. (1996) Assessing the performance of fold recognition by means of a comprehensive benchmark. In *Proceedings of the Pacific Symposium on Biocomputing* World Scientific Press, Hawaii, pp. 300–318.

Fischer,D., Christian,B., Bryson,K., Elofsonn,A., Godzik,A., Jones,D., Karplus,K.J., Kelley,L.A., MacCallum,R.M., Pawowski,K., Rost,B., Rychlewski,L. and Sternberg,M. (1999) CAFASP-1: Critical Assessment of Fully Automated Structure Prediction methods. *Proteins Struct. Funct. Genet.*, **3** (Suppl.), 209–217.

Hadley,C. and Jones,D.T. (1999) A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure*, **7**, 1099–1112.

Holm,L. and Sander,C. (1994) The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res.*, **22**, 3600–3609.

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

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) A new approach to protein fold recognition. *Nature*, **358**, 86–89.

Jones,D.T., Tress,M., Bryson,K. and Hadley,C. (1999) Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. *Proteins Struct. Funct. Genet.*, **3** (Suppl.), 104–111.

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

Karplus,K., Barrett,C., Cline,M., Diekhans,M., Grate,L. and Hughley,R. (1999) Predicting protein structure using only sequence information. *Proteins Struct. Funct. Genet.*, **3** (Suppl.), 121–125.

Kelley,L.A., MacCallum,R.M. and Sternberg,M.J.E. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.*, **299**, 499–520.

Michie,A.D., Orengo,C.A. and Thornton,J.M. (1996) Analysis of domain structural class using an automated class assignment protocol. *J. Mol. Biol.*, **262**, 169–185.

Moult,J., Hubbard,T., Fidelis,K. and Pedersen,J.T. (1999) Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins Struct. Funct. Genet.*, **3** (Suppl.), 2–6.

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

Murzin,A.G. (1999) Structure classification-based assessment of CASP3 predictions for the fold recognition targets. *Proteins Struct. Funct. Genet.*, **3** (Suppl.), 88–103.

Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Panchenko,A., Marchler-Bauer,A. and Bryant,S.H. (1999) Threading with explicit models for evolutionary conservation of structure and sequence. *Proteins Struct. Funct. Genet.*, **3** (Suppl.), 133–140.

Perrière,G. and Gouy,M. (1996) WWW-query: an on-line retrieval system for biological sequence banks. *Biochimie*, **78**, 364–369.

Przytycka,T., Aurora,R. and Rose,G. (1999) A protein taxonomy based on secondary structure. *Nat. Struct. Biol.*, **6**, 672–682.

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

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