WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre

Riccardo M. Bennett-Lovsey, Alex D. Herbert, Michael J. E. Sternberg, and Lawrence A. Kelley*

Structural Bioinformatics Group, Division of Molecular Biosciences, Imperial College London, London SW7 2AY, United Kingdom

ABSTRACT

Structural and functional annotation of the large and growing database of genomic sequences is a major problem in modern biology. Protein structure prediction by detecting remote homology to known structures is a well-established and successful annotation technique. However, the broad spectrum of evolutionary change that accompanies the divergence of close homologues to become remote homologues cannot easily be captured with a single algorithm. Recent advances to tackle this problem have involved the use of multiple predictive algorithms available on the Internet. Here we demonstrate how such ensembles of predictors can be designed in-house under controlled conditions and permit significant improvements in recognition by using a concept taken from protein loop energetics and applying it to the general problem of 3D clustering. We have developed a stringent test that simulates the situation where a protein sequence of interest is submitted to multiple different algorithms and not one of these algorithms can make a confident (95%) correct assignment. A method of meta-server prediction (Phyre) that exploits the benefits of a controlled environment for the component methods was implemented. At 95% precision or higher, Phyre identified 64.0% of all correct homologous query–template relationships, and 84.0% of the individual test query proteins could be accurately annotated. In comparison to the improvement that the single best fold recognition algorithm (according to training) has over PSI-Blast, this represents a 29.6% increase in the number of correct homologous query–template relationships, and a 46.2% increase in the number of accurately annotated queries. It has been well recognised in fold prediction, other bioinformatics applications, and in many other areas, that ensemble predictions generally are superior in accuracy to any of the component individual methods. However there is a paucity of information as to why the ensemble methods are superior and indeed this has never been systematically addressed in fold recognition. Here we show that the source of ensemble power stems from noise reduction in filtering out false positive matches. The results indicate greater coverage of sequence space and improved model quality, which can consequently lead to a reduction in the experimental workload of structural genomics initiatives.

Proteins 2008; 70:611–625.

Key words: meta-server; remote homology modelling; fold recognition; protein structure prediction; Phyre; ensemble; profile–profile alignment.

INTRODUCTION

The prediction of the three-dimensional (3D) structure of a protein from its amino acid sequence is a long-standing problem, for which many powerful techniques have been developed. The most successful general approach for predicting the structure of proteins involves the detection of homologues of known 3D structure—generally called template-based or fold-recognition methods.

These methods rely on the observation that the number of folds in nature appears to be limited and that many different remotely homologous protein sequences adopt remarkably similar structures. Thus, given a protein sequence of interest, one may compare this sequence to the sequences of proteins with experimentally determined

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-–3585/suppmat/.

Abbreviations: BASIC, bilaterally amplified sequence information comparison; B-DHIP, bi-directional heterogeneous inner product; BLAST, basic local alignment search tool; CASP, critical assessment of structure prediction; FR-systems, fold recognition systems; PCorr, Pearson correlation coefficient (as applied to profile comparison); PDP, probability dot product; PSG, position specific gaps; PSI-Blast, position specific iterated BLAST; QT-pair, query/template pair; SCOP, structural classification of proteins; SVM, support vector machine.

*Correspondence to: Lawrence A. Kelley, Structural Bioinformatics Group, Biochemistry Building, Department of Molecular Biosciences, Imperial College London, London SW7 2AY, U.K. E-mail: l.a.kelley@imperial.ac.uk

Received 23 November 2006; Revised 12 April 2007; Accepted 6 June 2007
Published online 17 September 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21688

structures. If a homologue can be found, an alignment of the two sequences can be generated and used to build a three-dimensional model of the sequence of interest.

Many advances in remote homology and analogy detection have been achieved in the past decade, beginning with sequence–structure threading[1–3] and structural profiles,[4,5] including the use of predicted secondary structure,[6] tertiary structure profiles,[7,8] hidden Markov models,[9,10] and most recently profile–profile matching algorithms.[11–14] All these methods have their strengths and weaknesses. There are a vast number of ways in which two close homologues may diverge over evolutionary time to become remote homologues. This broad spectrum of evolutionary change cannot easily be captured by a single methodology for remote homology detection.

Meta-servers

One of the most recent advances has come from combining many such methodologies in a meta-server (e.g., http://genesilico.pl/meta, http://meta.bioinfo.pl). The first fully automated meta-server, Pcons,[15] worked by collecting the outputs of six different publicly available protein fold-recognition servers, and used a set of neural networks to predict the quality and accuracy of the collected models. Even though Pcons was specifically trained to predict the quality of the final models, rather than whether or not they were of the correct fold, it did allocate higher final scores to folds that were predicted by more than one server. Virtually all meta-servers made available since Pcons work on a similar basis: namely selecting their final answer from a set of results using a consensus approach. However, the consensus methods used in different meta-servers can vary substantially: for example, whereas the original Pcons used a simple frequency count of fold representatives, the 3D-Jury server[16] uses a more sophisticated system of structurally clustering the top models from each of its constituent methods, and selecting the final model from the centre of the largest single cluster. While some methods return one of the initial models used in the consensus calculation as their end result, others make modifications to the final model before returning it to the user: for example, the 3D-SHOTGUN server[17] combines fragments from the initial models based upon the clustering of individual residues during structural superposition.

Some approaches have developed the necessary components of a meta-server in-house: the Shotgun-INBGU server[17,18] uses a meta predictor layer on top of the five prediction components of the original INBGU server[18]; Meta-BASIC[19,20] (Bilaterally Amplified Sequence Information Comparison) uses two versions of search algorithms performing gapped alignments of meta-profiles similar to those used by ORFeus.[21]

The power of the meta-, or ensemble, approach has been repeatedly demonstrated by the top ranking performance of such systems in CASP, the international blind trial of protein structure prediction.[22] These ensemble systems apply forms of majority voting and consensus techniques to the results of several powerful individual fold recognition (FR) algorithms available on the Internet. The most successful ensemble systems currently available are dependent on different algorithms located at different remote sites; are trained and tested according to different criteria and different databases; and rely on an ad hoc pooling of certain strongly performing individual methods. We show that this distributed system restricts the potential power of generating ensembles. The evolution of proteins is accompanied by a diverse spectrum of changes in sequence and secondary structure. By systematically combining methodologies under controlled conditions, it is possible to cover more of this diversity with a patchwork of specialised algorithms.

Development of Phyre

To our knowledge, no systematic and controlled analysis has yet been performed to determine why meta-servers are able to combine individual fold-recognition methods in such a way as to improve the recall and precision of the ensemble over and above the single best constituent method. Here we report the development of an ensemble fold-recognition system, Phyre, which combines in-house component algorithms under controlled conditions. We demonstrate striking improvements in precision and recall of remote homologues using a carefully selected test set, and we show that the power of the ensemble stems mainly from noise reduction in filtering out false positive matches rather than increasing the pool of available true positives.

Based on our benchmarking, we find that PSI-Blast successfully identifies 21.1% of all correct homologous query–template (QT) relationships in the test set at high confidence (i.e., >95% precision); these QT relationships provide enough coverage to accurately annotate 46.0% of all queries in the same test set. Similarly, the single best method (according to training) from a pool of optimised recognition algorithms (25 in Table I) can confidently identify 54.2% of all QT relationships, which accurately annotate 72.0% of all queries. The best ensemble in our study can confidently detect up to 64.0% of all correct homologous QT relationships, corresponding to 84.0% of the individual test query proteins being accurately annotated at 95% precision or higher. This top-performing ensemble forms the core of the Protein Homology/analogY Recognition Engine (Phyre), available on the web (http://www.imperial.ac.uk/phyre).

Implications for structural genomics

This improvement in remote homology annotation has implications for the ongoing structural genomics initiatives. The primary goal of such initiatives is to provide

DOI 10.1002/prot

Table I
Performance of Each of the 31 Fold Recognition Systems, and 10 Different Ensemble Systems, on the Full Test Set of 50 Query Protein Sequences

Single methods

Method identifier	Primary structure algorithm	Secondary structure algorithm	Parameter set	Average precision	% Recall at 95% precision	% Queries annotated at 95% precision
PSI-blast	n/a	n/a	n/a	n/a	21.1*	46.0*
01	Seq V Seq	n/a	1 (PSG)	26.6	18.6*	56.0*
02	Seq V Seq	n/a	2	28.4	18.6*	56.0*
03	Seq V PSSM	n/a	1 (PSG)	54.4	44.9*	70.0*
04	Seq V PSSM	n/a	2	53.9	43.3*	70.0*
05	PSSM V Seq	n/a	1 (PSG)	49.5	38.5*	66.0*
06	PSSM V Seq	n/a	2	49.8	38.5*	66.0*
07	BASIC	n/a	1	48.8	27.5*	60.0*
08	**BASIC**	**n/a**	**2**	**49.3**	**32.4***	**60.0***
09	**BASIC**	**n/a**	**3**	**49.2**	**28.7***	**60.0***
10	**BASIC**	**n/a**	**4**	**44.3**	**17.8***	**46.0***
11	**B-DHIP**	**n/a**	**1 (PSG)**	**67.9**	**50.2***	**72.0**
12	**B-DHIP**	**n/a**	**2**	**69.2**	**51.4***	**68.0***
13	**B-DHIP**	**n/a**	**3**	**65.5**	**46.2***	**66.0***
14	B-DHIP	n/a	4	62.0	46.2*	68.0*
15	B-DHIP	n/a	5	56.0	34.4*	62.0*
16	B-DHIP	n/a	6	60.0	36.4*	64.0*
17	PDP	n/a	1	40.6	24.7*	58.0*
18	PDP	n/a	2	41.4	25.5*	62.0*
19	PDP	n/a	3	41.2	24.3*	60.0*
20	PDP	n/a	4	40.6	24.7*	58.0*
21	**PCorr**	**n/a**	**1**	**9.8**	**2.4***	**6.0***
22	**B-DHIP**	**Seq V Seq**	**1**	**60.9**	**39.3***	**66.0***
23	B-DHIP	BASIC	1	63.5	48.6*	70.0*
24	B-DHIP	BASIC	2	67.2	52.2*	74.0
25	**B-DHIP**	**B-DHIP**	**1**	**68.9**	**54.2***	**72.0**
26	B-DHIP	B-DHIP	2	67.3	53.4*	72.0
27	B-DHIP	PDP	1	66.3	52.6*	74.0
28	B-DHIP	PDP	2	68.2	53.8*	74.0
29	B-DHIP	PCorr	1	68.5	55.0*	76.0
30	B-DHIP	PCorr	2	68.9	56.7*	76.0
31	**B-DHIP**	**PCorr(LO)**	**1**	**57.1**	**38.5***	**66.0***

Ensemble systems

Ensemble description	% Recall at 95% precision	% Queries annotated at 95% precision
SVM 3D-Jury Top 10	26.3*	54.0*
Weighted Greedy ep_scores >0.7	25.4*	60.0*
SVM ep_score	34.0*	62.0*
3D-Jury Greedy Top 10	23.9*	82.0
3D-Jury Greedy Top 1	*(25.5)**	*88.0*
SVM 3D-Jury Top 10 + ep_score (top 10)	23.5*	48.0*
3D-Colony Greedy Top 1	(24.7)*	86.0
SVM 3D-Jury Top 10 + ep_scores (all)	21.1*	44.0*
3D-Colony Greedy Top 10	55.9*	68.0*
3D-Colony Greedy ep_scores >0.7 (Phyre)	**64.0**	**84.0**

The figures represent: the average precision of each method; the percentage of correct homologous relationships detected; and percentage of query proteins correctly annotated (at 95% precision or above) **Single methods**: PSG refers to Position-Specific Gaps, that is, individual gap opening and extension parameters were optimized for helix, strand, and coil in the template structure. Numbers in the parameter set column refer to boosted variants of the single algorithms as described in the text. The algorithms used are described in "Materials and Methods." Bold indicates methods selected for the final optimal ensemble by the greedy build-up procedure. The single best method according to the training results is shaded in grey. **Ensemble systems**: ensembles are ranked according to the percentage of correct homologous query-template relationships correctly identified detected during testing with the "impossible" test set (see Table III). *Weighted* refers to weighting of methods according to their average error rate (=1 − average precision). *Top n* refers to how many models were taken from each predictor. All SVM methods used a linear kernel and default parameters in SVM-Light. *Greedy* indicates an ensemble constructed using the greedy build up algorithm. "ep_scores >0.7" indicates only matches with ep_scores greater than 0.7 (marginal confidence) were included in the 3D-Jury/3D-Colony. "SVM 3DJury top 10 + ep_score (top 10)" indicates an SVM with features corresponding to 3D-Jury scores for the top 10 models, represented as discussed in "Materials and Methods." 3D-Jury and 3D-Colony ensembles using just the single top model are bracketed, for high confidence homologous relationship recall, because their recall capability is intrinsically limited by the nature of the ensemble. Standard meta-server 3D-Jury performance is italicized, and the best performing 3D-Colony/Phyre approach (based on training; see Table II) is in bold. Any value that is labeled with an asterisk is significantly different from the respective value for the Phyre ensemble (the last row of the table) at the 5% significance level.

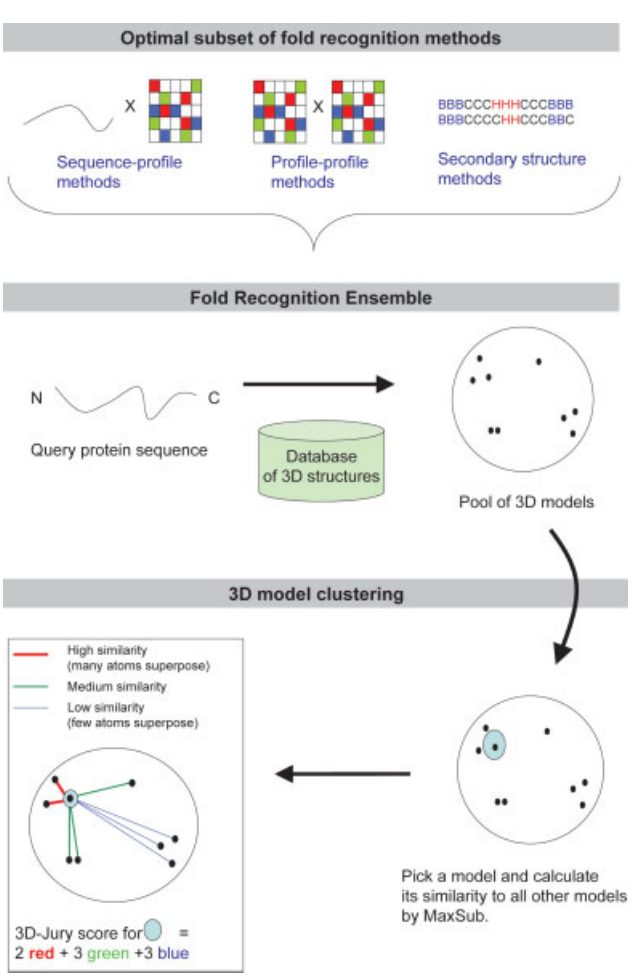

Figure 1

Schematic overview of the 3D-Colony ensemble clustering procedure.

experimentally determined structures for all protein sequences found in nature. However, despite significant improvements in high throughput experimental methods, sequencing projects will continue to outpace structure determination initiatives. To date, only about 1% of known protein sequences have been structurally determined (at the time of writing there were 3.8 million protein sequences in the NCBI protein sequence database, and 37,981 structures in the Protein Data Bank[23]). A major challenge in structural bioinformatics is the development of computational techniques, which can help bridge this gap by reliable prediction of structure from sequence alone.

The minimum number of experimental structures that will be needed to model all proteins using evolutionary relationships depends on the nature of protein sequence space, and the power of remote homology detection algorithms. The further such methods can be pushed to extend the range, at which remote homologies to known structures can be reliably inferred, the more sequences we can reliably model. Consequently, the power of FR has implications for the number of experimental structures required to model a large fraction of the sequenced genomes.[24]

MATERIALS AND METHODS

The construction and assessment of fold-recognition ensembles

Here we provide a description of the methodology used. Figure 1 is a schematic illustrating the flow of events in the final ensemble. A protein sequence is processed by a pool of FR algorithms, which results in a corresponding pool of candidate protein structural models. These models are clustered according to one of the protocols described, and the models ranked. The different clustering protocols are illustrated conceptually in Figure 2.

In general, fold recognition systems (henceforth FR-systems) rely on the detection of similarity between a protein sequence of interest and another sequence of known 3D structure. Three major factors contribute to the success of these template-based structure prediction methods:

1. the nature of the algorithms for sequence–sequence comparisons;

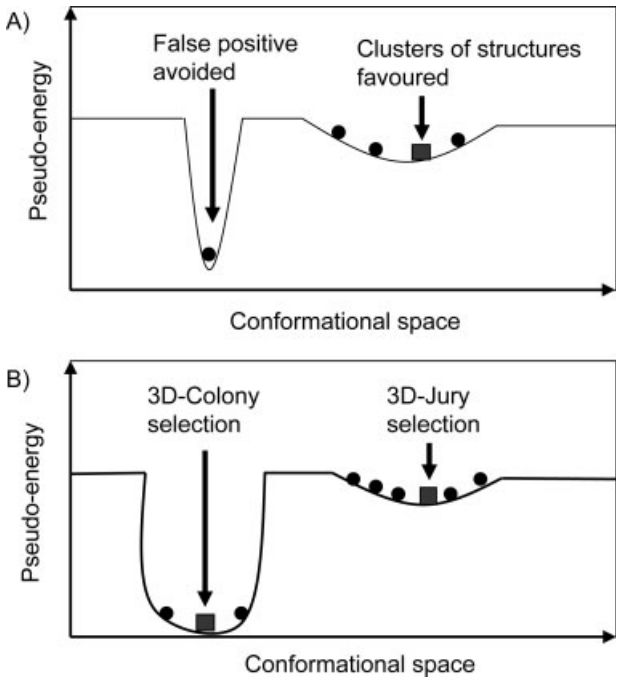

Figure 2

*Schematic illustration of the behavior of the different ensembles. Circles represent models produced by the various algorithms in the ensemble. Squares represent the highest scoring model as judged by the algorithm in question, that is, 3D-Jury or 3D-Colony. (**A**) General principle of structural clustering. (**B**) 3D-Jury only takes into account the population of an area of conformational space, whereas 3D-Colony includes information regarding the "energetics," or confidence, of the fold recognition match.*

DOI 10.1002/prot

2. the databases of known structures and their corresponding sequence profiles;
3. training, tuning, and parameter estimation.

Existing meta-servers are powered by a set of algorithms for which all three of these factors vary. This hinders an understanding of their relative contribution to the overall effectiveness of the ensemble, and limits the feasibility of combining them optimally. In contrast, all FR-systems used in this work have been trained on the same data under the same optimisation strategy and tested under identical conditions. Thus, the performance of the ensemble generated in this work is solely generated by algorithmic and parametric variety.

Combining classifiers or predictive algorithms in ensembles to improve performance is an established research area shared between statistical pattern recognition and machine learning. Unfortunately, even after several decades of research, the theoretical groundwork of ensemble theory does not yet provide us with a recipe for creating optimal ensembles.[25–27] As a result, various heuristics must be explored.

There are several prerequisites to generating an ensemble: (1) a pool of diverse yet accurate methods; (2) a standardised scoring framework; (3) a technique to combine individual methods; and (4) a procedure to select an optimal, or quasi-optimal subset of the methods, which constitute a final ensemble.

Building a pool of methods

To generate a variety of powerful FR-systems we have implemented several successful algorithms for profile–profile and sequence–profile comparison. The generic nature of these algorithms means they can be applied to both primary and secondary structure information. This in turn provides us with a large number of possible algorithm combinations to generate sufficient diversity for a workable ensemble. Each of these algorithmic variants is trained on a set of remote homologies taken from SCOP[28] and its parameters optimised to achieve maximum average precision on a common training set.

Standardised scoring framework

For the purpose of combining the results of different algorithms, it is useful to have a common scoring framework. We introduce the notion of "empirical precision," or *EP-score*: a standardised scoring scheme from zero to one, reflecting the empirically derived probability that a match is correct. Based on measures of performance on a training set of remote homologies, all results from the FR-systems can be converted to this common scoring framework.

Combining methods in an ensemble

Combining the results of different FR-systems can be performed in several ways. However, broadly speaking these fall into two categories: (a) those that work in 3D space on the models produced by such systems, for example, by making structural comparisons between models; and (b) those that work purely on scores and protein identifiers, for example, counting the number of semiconfident matches to a given superfamily.

We introduce a novel method of pooling individual FR results that is a hybrid of these two schemes, which we call the *3D-Colony* protocol (see "Supplementary Materials"). The controlled parallel development of multiple FR-systems enables the use of standardised scoring schemes. 3D-Colony combines confidence measures from a standardised scoring framework with structural similarity clustering. The protocol is analogous to the Colony energy approach used in loop modelling where the confidence measure is analogous to an enthalpy term, and the structural similarity score is analogous to an entropy term.[29]

Selecting methods for an ensemble

Given a set of FR-systems, it is necessary to determine an optimal or quasi-optimal subset for use in an ensemble. A simple pooling of all methods is rarely the best option. Commonly occurring false positive results generated by algorithms, which are not sufficiently diverse can lead to an ensemble, which is in fact worse performing than the single best component.[26] The rigorous selection of a combination of methods, which optimises ensemble performance, leads to a combinatorial explosion. To avoid this problem we use a heuristic that performs a greedy build-up of the ensemble by adding component methods one at a time, searching for the best pair, best triplet, and so forth. This allows one to select a patchwork of algorithms with an aim to maximising ensemble performance on a stringent test set of remote homologues.

As an alternative method of circumventing the problem of component selection we also investigate the use of Support Vector Machines (SVMs)[30] to classify candidate fold matches as correct or incorrect using information from all the methods simultaneously. Alignment scores and structural clustering scores are combined in a high dimensional attribute vector, which can be used as training input to the SVM learning procedure.

Ensemble performance criteria

The training and testing data sets were constructed from SCOP 1.65[28] with no pairs sharing >30% sequence identity from ASTRAL[31] (for greater detail, see "Training and testing data"). There were 105 query training sequences (with a total of 1124 QT relationships) and

50 query test sequences (with a total of 247 QT relationships).

A major problem when assessing the synergistic effect of an ensemble classifier is the issue of trivial answers: for example, a given query protein may be scanned against a template protein database, which contains five correct homologous templates, using a particular FR-system from an ensemble; one homologous template (T1) may be trivial to find at high confidence, whilst the others (T2–T5) may be much harder. In an ensemble that utilises structural clustering as part of its algorithm, templates T2–T5 are now easily identifiable by virtue of their structural similarity to T1; as a result, the harder templates are found simply because there exists an easy template of similar structure in the results of one of the constituent methods of the ensemble. It becomes impossible to distinguish between any improvement in recognition accuracy gained by the process of generating an ensemble and the presence of trivial solutions in one of the constituent methods.

To focus this work on the most remote and undetectable homologies, we have constructed special ("impossible") training and test sets; the training set consists of 105 query proteins (with a total of 1124 correct homologous relationships), and the test set consists of 50 query proteins (with a total of 247 correct homologous relationships). We wish to examine how ensembles perform when none of their constituent methods can make a confident assignment: during ensemble training and testing, specific high confidence (i.e., at 95% confidence or higher) QT relationships found by individual FR-systems are ignored. This is not to say that the given QT relationships are ignored across all FR-systems, it is only ignored in those individual methods, in which it can be confidently identified. It is important to note that, even after all trivial correct homologous QT relationships were removed from across all individual FR-systems, the 247 correct homologous relationships in the testing set were still present within the results pool. Therefore it was not necessary to remove any of the query proteins from the benchmark.

This benchmark was designed to reflect the "real-world" situation of difficult structure prediction targets, where no individual system provides a confident answer. All ensembles were trained on the "impossible" training set, and then tested using the "impossible" test set. This means that any homologous relationships that are confidently detectable by the ensemble are solely related to the use of an ensemble, that is, the combination of multiple weak predictions, and not because of a highly accurate individual method. This procedure of developing the "impossible" set was necessary to provide sufficient QT pairs for training and testing given the present-day size of the protein data bank. All ensembles were also tested using the full test set (i.e., the standard test data still containing all the trivial answers) to assess how well they performed in comparison to the individual FR-systems and PSI-Blast.

A detailed explanation of the profile methods used and the methods used to generate and score the ensembles is available in the Supplementary Material.[32–39]

Training and testing data

A common training set was constructed for use by all algorithms based on SCOP 1.65[28] with no pairs sharing >30% sequence identity from ASTRAL.[31] One query sequence for each superfamily was chosen, ensuring at least five other members of that superfamily are present in the fold library. The set of "correct" homologues (proteins in the same SCOP superfamily) was filtered on structural similarity grounds using Mammoth.[40] Many pairs of proteins in the same SCOP superfamily are significantly structurally dissimilar. Such dissimilar proteins were removed from the list of correct homologues if they were not superposable, in a sequence independent manner, by Mammoth with a Z-score >5. These relationships, although possibly valid from an evolutionary perspective, are not useful in protein structure prediction and add noise to the system. This is because any model built based on an alignment to such a homologue, assuming perfect alignment accuracy, would not be an accurate model of the query sequence. Any query sequences with less than five "correct" homologues following the Mammoth filtering step were removed from the query list. For testing, a set of query sequences all from SCOP folds different to any used in training were selected and filtered in the same way as the training set. However, to enable a sufficiently large test set, the criteria for the number of homologues available in the fold library after Mammoth filtering was reduced from 5 to 2. This process provided us with 105 training query sequences (with a total of 1124 QT relationships) and 50 query test sequences (with a total of 247 QT relationships).

Each of the various algorithms require several parameters, such as gap opening and extension, relative weights of sequence versus secondary structure, normalising zero-shifts, and so forth, to function optimally. To determine the optimum value of these parameters one needs a measure of performance, or fitness, to use standard optimisation protocols. The performance of a particular system is normally represented by a ROC curve or precision/recall graph.

For training, one may fix, say, the precision of the system at some threshold and optimise the recall. However, deciding where to choose this threshold can be problematic. In addition, if one only looks at the recall at a fixed precision, any information regarding improvements in recall at lower precision will not be available to the optimisation protocol. Thus the optimisation can be incomplete and become more easily trapped in local minima.

To avoid these problems we use the average precision or AP value of the system as a single fitness figure. The AP can be thought of as an approximation to the area

 DOI 10.1002/prot

under the precision–recall curve. This thus rewards improvements at all scales of precision and recall. Assuming the target databank contains N correct entries for a given query, a search algorithm then returns M entries of which K are true positives. Following on from this, recall $= K/N$ and precision $= K/M$. Therefore

$$\mathrm{AP} = \frac{1}{N} \sum_{i}^{K} \frac{1}{p_i}$$

where p_i is the rank of the i-th true positive. Note that $1/p_i$ is just the precision value of the i-th true positive in this iterative process, and the above equation is an approximate integral to calculate the area under the resulting precision–recall curve.

In this work we have used the simplex method[41,42] for parameter optimisation. Simplex is a simple optimisation algorithm seeking the vector of parameters corresponding to the global extreme (maximum or minimum) of any n-dimensional function searching through the parameter space. An $n + 1$ geometric figure in an n-dimensional space is called a simplex. The simplex moves through the parameter space by various transformations, searching the surface in a relatively efficient and intuitive sense. Although not guaranteed to find a global minimum, its performance is expected to be adequate for the purposes of this study.

Each method was optimised on the training data using the simplex algorithm, and the resulting parameters used in testing. The results of these tests are shown in Table I.

This procedure resulted in 31 independently optimised algorithmic variants for FR. The next step was to generate ensembles of these methods by both the empirical greedy algorithm and SVMs. To mimic the real-world scenario of detecting extremely remote homologous relationships, the standard training set was modified as follows.

The "impossible" training and test sets

For our ensemble training and test set we use the same training and testing data described earlier (i.e., 105 training queries and 50 testing queries), but exclude (from every individual method listed in Table I) any homologous relationship detectable by that method at >95% confidence (as defined by 0.95 EP score). It should be noted that all non-homologous (i.e., false positive) relationships detected by any method at >95% confidence are allowed to remain. This leaves a pool of methods which independently cannot discover a single correct relationship in the test set at >95% confidence (see Fig. 6). Similarly, when using the SVM ensembles, any structural comparisons that used models built from homologous relationships detectable above 95% confidence were excluded. This mimics a common annotation problem where, regardless of which state-of-the-art FR-systems on the Internet one uses, no confident matches can be found. Thus, any recall greater than 0% using an ensemble directly illustrates the power of the ensemble rather than any particularly strong individual method.

RESULTS AND DISCUSSION

Generating a pool of algorithms

Any ensemble system requires a diverse pool of component predictors with reasonably high accuracy. Of course, an exhaustive analysis of the techniques currently used to predict protein structure from sequence is beyond current resources. As a result, we must restrict our analysis to a subset of state-of-the-art methods shown to perform well in blind trials. Many of the widely-used programs for the detection of remote homology can be classified according to their use of primary and secondary structure information in the form of sequences or profiles. To generate a pool of such methods, we performed independent optimisation of 31 algorithmic variants for remote homology detection. A breakdown of the methods and their features is shown in Table I. We aimed to derive a wide variety of methods without an exhaustive enumeration of the possibilities. We included standard sequence–sequence, sequence–profile and profile–sequence methods. In addition, we have implemented four generic methods for profile–profile comparison (*PDP, BASIC, B-DHIP,* and *PCorr*—described in "Supplementary Materials") that vary in the manner in which they handle the comparison of profiles. Each of these methods may be applied to either primary or secondary structure information, in various combinations.

For several systems, we experimented with using secondary structure-specific gap penalties. In these systems, gap penalties for helix, strand and coil in the template structure are individually optimised (indicated by PSG in Table I). These systems did not perform significantly differently from their simpler counterparts. The ineffectiveness of this approach is surprising as it is commonly observed that gaps are more prevalent in coil regions than in regions of secondary structure.[43] When attempting to recognise extremely remote homologues, it may be that such subtleties are outweighed by other factors, such as profile composition and accuracy of secondary structure prediction.

Individual method performance

Throughout, the terms precision and recall will be used. Precision and recall are defined as:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

where TP, FP, and FN represents the number of true positives, false positives and false negatives, respectively.

First we present the performance results of each of the individual methods in comparison to a standard PSI-Blast search, Table I. The figures represent the recall and queries annotated at 95% precision, and the average precision (defined in "Materials and Methods") on the full test set of 50 query protein sequences. As expected, each method (excluding a simple BLOSUM62 search and the PCorr method—the latter is discussed in "Supplementary Materials") demonstrates superiority to standard PSI-Blast searching. This work is not concerned with demonstrating superiority of one method over another. Indeed, different groups using small variations in algorithm or different training criteria may achieve a different ranking of methods than that displayed in Table I. What is of primary importance is to illustrate that all of the methods are reasonably accurate recognition methods under the same controlled test.

Boosting, and simplified boosting

In the machine learning field, the process known as *boosting* has been repeatedly and successfully used to improve the performance of learning algorithms in an ensemble setting.[44,45] The principle of boosting is to increase progressively the importance of training examples that are difficult to classify. For several FR systems we have used a simplified form of boosting to generate a family of algorithms, which differ only in parameter tuning. To do this we remove all query sequences from the training set that are annotated with a precision above some threshold. The new, reduced set of query sequences constitute the next training set, on which the algorithm is retrained, and this process is repeated either a fixed number of times or until no further improvement is found.

Unfortunately, this approach did not produce any real improvement in the accuracy of the algorithms. We believe this is due to two factors: (1) There are not very many tuneable parameters involved with FR systems in general. In this case the parameters are: insert opening penalty and extension, deletion opening penalty and extension, Z-shift, and relative weight of primary versus secondary structure. (2) The system is not very sensitive to the values of these parameters. It is not in general possible to detect very difficult remote homologies by altering parameters alone. FR methods are not generic learning systems, and this suggests that standard boosting methodologies applied to existing FR methods may not be beneficial in general.

Performance of full ensembles on the "impossible" training set

Table II reports the training results of 10 different protocols for building an ensemble, ranked by their recall on the "impossible" training set. Additionally, the percentage of unique queries successfully annotated is shown. These measures differ of course as one query may have several possible correct annotations—one for each of its homologues in the fold library. Only one relationship is required for detection, but the total coverage gives a more useful assessment when considering the general quality of the ensemble.

All SVM values are listed as "n/a" because the nature of SVM learning would mean that the final classifiers would be intrinsically biased when classifying the training data. In an ideal scenario there would be three sets of data available for benchmarking SVM models: a training set, an "evaluation" set, and a test set. The "evaluation"

Table II
Comparison of 10 Different Ensembles on the Stringent, "Impossible" Training Set

Ensemble description	% Recall at 95% precision	% Queries annotated at 95% precision
Any single method in Table 1	**0**	**0**
3D-Jury Greedy Top 10	24.5*	42.9*
Weighted Greedy ep_scores >0.7	41.5	29.5*
SVM ep_score	n/a	n/a
SVM 3D-Jury Top 10	n/a	n/a
3D-Jury Greedy Top 1	(20.5)*	76.2
SVM 3D-Jury Top 10 + ep_score (top 10)	n/a	n/a
3D-Colony Greedy Top 1	(21.0)*	75.2
SVM 3D-Jury Top 10 + ep_scores (all)	n/a	n/a
3D-Colony Greedy Top 10	29.8*	50.5*
3D-Colony Greedy ep_scores >0.7 (Phyre)	**41.0**	**81.0**

By definition of the "impossible" set (see "Materials and Methods"), none of the individual component methods (see "Single Methods" in Table 1) used in any of the ensembles can detect any confident matches in the training set. Any recall greater than 0% using an ensemble occurs as a result of the ensemble rather than any particularly strong individual method. Figures represent the percentage of correct homologous query–template relationships detected, and percentage of query proteins correctly annotated (at 95% precision or above are ranked according to the percentage of correct homologous query–template relationships correctly identified detected during testing with the "impossible" test set (see Table 3). *Weighted* refers to weighting of methods according to their average error rate (=1 − average precision). *Top n* refers to how many models were taken from each predictor. All SVM methods used a linear kernel and default parameters in SVM-Light. *Greedy* indicates an ensemble constructed using the greedy build up algorithm. "ep_scores>0.7" indicates only matches with ep_scores greater than 0.7 (marginal confidence) were included in the 3D-Jury/3D-Colony. "SVM 3DJury top 10 + ep_score (top 10)" indicates an SVM with features corresponding to 3D-Jury scores for the top 10 models, represented as discussed in "Materials and Methods." All SVM values are listed as "n/a" because the nature of SVM learning would mean that the final classifiers would be intrinsically biased towards the training data. 3D-Jury and 3D-Colony ensembles using just the single top model are listed as "n/a," for high confidence homologous relationship recall, because their recall capability is intrinsically limited by the nature of the ensemble. Given the total number of correct relationships in the training set (1124), any comparison with other ensembles would be meaningless. Standard meta-server 3D-Jury performance is italicized, and the best performing 3D-Colony/Phyre approach is in bold. Any value that is labeled with an asterisk is significantly different from the respective value for the Phyre ensemble (the last row of the table) at the 5% significance level.

DOI 10.1002/prot

set would be constructed in the same way as the training and test sets (see "Materials and Methods"), but would be used to predict, which SVM model would be the best classifier before they were all finally tested using the test set. Unfortunately, due to the limited amount of data available (and given our extremely stringent selection criteria) it was only possible to construct the training and test sets and still maintain a reasonable number of queries remaining in each. Instead, it was decided to take all (four) of the SVM models that were built, and test them on the "impossible" and full test sets, under the assumption that the SVM models with larger feature vector inputs would perform better than those with smaller feature vector inputs.

The percentage of high confidence correct homologous relationship detected by the 3D-Jury and 3D-Colony ensembles that use just the single top model from each of their constituent methods are bracketed because the recall capability of these methods is intrinsically limited by the nature of the ensemble: whilst it would still be theoretically possible to find all correct homologous relationships in the training and test sets, in practice the results would be much more restricted when compared with an ensemble that used the top 10 models from each of their constituent methods (they are included here for reference purposes). This limitation does not apply to the successful annotation of individual query proteins.

From the results of the ensemble process, in Table II, it is clear that all the methods explored (that have available data) are capable of confidently annotating remote homologues despite the fact that their individual components can make no confident assignments at all. The "impossible" training set used for the ensembles produces 0% recall at 95% precision for each FR method in isolation; therefore any improvement above 0% recall is as a result of the ensemble process alone. In Figure 3, an example is given of the 3D-Colony approach with the greedy build-up algorithm as a combination method. The graph shows how progressively, greedily adding methods to the system during training results in an initial increase in coverage at high confidence followed by a subsequent fall. The vertical line in this graph corresponds to the peak performance point found on the training set—itself corresponding to a final trained ensemble containing 10 FR-systems.

A similar pattern of an initial rise in performance and subsequent fall with the addition of further methods is found across the majority of techniques to form ensembles reported here. Of course the peak in testing is not always as predicted by the training regime, and this will be due to differences of how the individual methods behave with different query sequences. Nevertheless, the peaks are generally in agreement between training and testing.

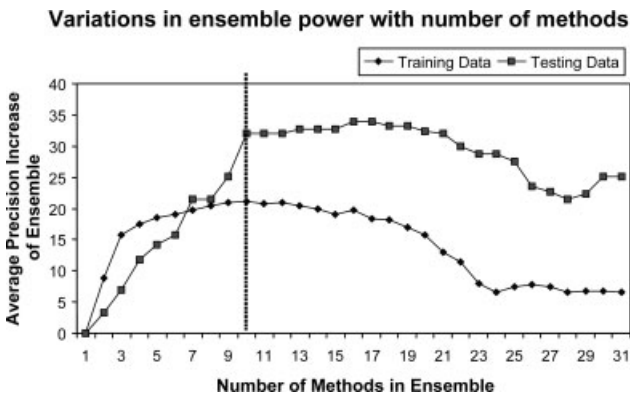

Figure 3

Graph illustrating the progression of the greedy algorithm in training (diamonds). As methods are added to the ensemble, the average precision rises and falls. The peak in this training process is highlighted by a dashed vertical line at the point where the ensemble contains 10 methods. In testing (squares), a similar pattern of rise and fall is seen, with a peak in performance roughly in agreement with the training regime.

Performance on testing data

Table III shows the results of the 10 protocols for forming ensembles when tested using the "impossible" test set. Analogous to the "impossible" training set, the "impossible" test set used for the ensembles produces 0% recall at 95% precision for each FR method in isolation. The lower section of Table I shows the results of the same 10 ensembles when tested using the full test set. In both tables all results are made available: results for the SVMs are valid as the test sets are unseen data. The percentage of high confidence correct homologous relationship, detected by 3D-Jury and 3D-Colony Top 1 ensembles, are still bracketed even though the number of test set relationships (247) is significantly smaller than the number of training set relationships (1124). This is merely to indicate that one should still be aware of the intrinsic limitations of these particular ensembles.

SVMs versus greedy build-up algorithm

Combining individual methods with an SVM is generally superior in performance to the greedy build-up algorithm on corresponding data. For example, on the "impossible" test set, the performance of an SVM applied to EP-scores (*SVM ep_score* in Table III; 58% queries annotated) is clearly superior to a technique using the same data but combined using the greedy build up procedure (Weighted Greedy ep_scores >0.7; 12% queries annotated).

Since EP scores and 3D-Jury methods by themselves appear to produce significant performance gains when

Table III
Comparison of 10 Different Ensembles on the Stringent, "Impossible" Test Set

Ensemble description	% Recall at 95% precision	% Queries annotated at 95% precision
Any single method in Table 1	**0**	**0**
3D-Jury Greedy Top 10	27.9*	22.0*
Weighted Greedy ep_scores > 0.7	28.3*	12.0*
SVM ep_score	28.7*	58.0*
SVM 3D-Jury Top 10	34.8*	62.0*
3D-Jury Greedy Top 1	(38.5)*	84.0
SVM 3D-Jury Top 10 + ep_score (top 10)	40.0*	66.0
3D-Colony Greedy Top 1	(40.1)*	82.0
SVM 3D-Jury Top 10 + ep_scores (all)	45.3*	72.0
3D-Colony Greedy Top 10	47.8*	48.0*
3D-Colony Greedy ep_scores > 0.7 (Phyre)	**58.3**	**82.0**

By definition of the "impossible" set (see "Materials and Methods"), none of the individual component methods (see "Single methods" in Table 1) used in any of the ensembles can detect any confident matches in the test set. Any recall greater than 0% using an ensemble occurs as a result of the ensemble rather than any particularly strong individual method. Figures represent the percentage of correct homologous relationships detected, and percentage of query proteins correctly annotated (at 95% precision or above). Ensembles are ranked according to the percentage of correct homologous query–template relationships correctly identified. *Weighted* refers to weighting of methods according to their average error rate (=1 − average precision). *Top n* refers to how many models were taken from each predictor. All SVM methods used a linear kernel and default parameters in SVM-Light. *Greedy* indicates an ensemble constructed using the greedy build up algorithm. "ep_scores>0.7" indicates only matches with ep_scores greater than 0.7 (marginal confidence) were included in the 3D-Jury/3D-Colony. "SVM 3DJury top 10 + ep_score (top 10)" indicates an SVM with features corresponding to 3D-Jury scores for the top 10 models, represented as discussed in "Materials and Methods". 3D-Jury and 3D-Colony ensembles using just the single top model are bracketed, for high confidence homologous relationship recall, because their recall capability is intrinsically limited by the nature of the ensemble. Standard meta-server 3D-Jury performance is italicized, and the best performing 3D-Colony/Phyre approach (based on training; see Table 2) is in bold. Any value that is labeled with an asterisk is significantly different from the respective value for the Phyre ensemble (the last row of the table) at the 5% significance level.

used on the "impossible" data sets, we explored the idea of combining the two methodologies in the form of a feature vector for an SVM (described in "Supplementary Materials"). The application of an SVM to the 3D-Jury Top 10 structural clustering data increases the number of correctly annotated queries to 62%, compared with 22% using the same 3D-Jury Top 10 data combined using the greedy algorithm. This performance difference is unsurprising because of the far from optimal heuristic used in the greedy build-up method. It is worth noting that an SVM using 3D-Jury Top 1 data was not trained separately because the equivalent information was already contained as a subset within the 3D-Jury Top 10 SVM data; as a result, the 3D-Jury Top 1 data would have been implicitly used by the SVM during training because of the nature of the learning algorithm.

Interestingly, when the SVMs were tested using the full test set, there was a change in the accuracy of the SVMs compared with the results on the "impossible" test set (see Tables I and II): as the size of the input feature vector increases, the accuracy of the ensemble not only falls, but falls by a larger amount. As a result, the SVMs that use smaller input vectors perform better than those with larger input vectors. These results were unexpected given the relative success of the SVMs on the "impossible" test data. They show that the SVMs are capable of efficiently learning relationships for specific data, however they were not able to build a generalised model that could distinguish easy examples from hard examples. Therefore, training an SVM on an "impossible" training set will build a model that can reliably classify similar "impossible" testing examples; however, if one wishes to classify easy testing examples, a better classifier would be built using an easy training set.

Fortunately a simpler and more intuitive approach, avoiding the use of black-box learning with an SVM proved most successful of all. We call this approach 3D-Colony.

3D-Colony

A simple and computationally tractable approach is to use EP scores both as a filter and as weighting terms as input to a 3D-Jury protocol. Thus, extremely low confidence matches are excluded from the ensemble to avoid pollution by noise, and those matches that are permitted into the system contribute to the structural clustering in accordance with the predicted confidence. This approach was the most successful ensemble tested and forms the basis of the Phyre web server.

The Phyre ensemble is capable of striking performance enhancement: when tested on the "impossible" test set, 82% of queries and 58% of all relationships can be detected at high confidence by this ensemble. This highlights the power of the ensemble process: none of the 58% of homologous relationships detected by Phyre (at >95% confidence) were detectable by any of the component methods that were the input into the ensemble. Only 10 of the 31 tested methods are required for this strong performance. Unlike the SVM-based approaches, this system requires 1/3 of the computational resources, yet performs the best in this benchmark, and is robust enough to classify easy as well as "impossible" examples. When tested using the full test set, the Phyre ensemble can confidently detect up to 64.0% of all correct homologous QT relationships, and 84.0% of the individual test query proteins could be accurately annotated at 95% precision or higher. In comparison to the improvement that the single best FR-system based on the training data (number 25) has over PSI-Blast, this represents a 29.6% increase in the number of correct homologous QT relationships, and a 46.2% increase in the number of accurately annotated queries, for the Phyre ensemble (see Tables I and IV).

DOI 10.1002/prot

Table IV
An Overview of Full Test Results for PSI-Blast, the Single Best Fold Recognition Method From the Training Benchmark (see 25 Under "Single methods" in Table 1), and the Phyre Ensemble

Description	% Recall at 95% precision	% Queries annotated at 95% precision
PSI-Blast	21.1*	46.0*
Single Best FR-system (25)	54.2*	72.0
Phyre Ensemble (best ensemble)	64.0	84.0

Figures represent the percentage of correct homologous query–template relationships detected, and percentage of query proteins correctly annotated (at 95% precision or above). Any value that is labeled with an asterisk is significantly different from the respective value for the Phyre ensemble (the last row of the table) at the 5% significance level.

The above is the correct approach for comparison of algorithms as one selects the methods and the parameters on unseen training data. However, even if one had the benefit of hindsight and chose the best single FR algorithm from the testing data (30 in Table I); this best single method would only confidently identify 56.7% of all QT relationships and annotate 76.0% of all queries. Thus, in comparison to the improvement that this method has over PSI-Blast, the ensemble provides a 20.5% increase in the number of correct homologous QT relationships, and a 26.7% increase in the number of accurately annotated queries (see Tables I and IV).

Despite the number of queries annotated by the Phyre ensemble on the full test set, it was still not as high as the 3D-Colony Top 1 and the 3D-Jury Top 1 ensembles (although in real terms these were only able to annotate one and two more queries, respectively, out of a possible 50). The reason why this particular ensemble was chosen over the others was because of its high percentage of correct homologous QT relationships detected at high confidence. As mentioned previously, the clustering ensembles that use the single top result from each of their constituent methods are intrinsically limited in their coverage of homologous relationships; therefore, even though they may be able to identify one or two correct homologies to annotate a given query, they cannot provide the breadth of coverage that is required for other bioinformatics tasks such as function prediction. Therefore, it was decided that the very small drop in performance of the Phyre ensemble, when measuring the number of queries annotated, was a sacrifice worth making given the huge increase in the percentage of correct homologous QT relationships it could detect.

Statistical analysis

For every set of results, a two-tailed McNemar's test[46] was used to test for any significant difference between the Phyre ensemble and any other comparable method. The McNemar's test is a standard approach for finding the statistical significance by evaluating the probability of χ^2, where:

$$\chi^2 = \frac{(b-c)^2}{(b+c)}$$

where b is the number of times that the prediction of the first method is wrong and the prediction of the second method is correct, and c is the number of times that the prediction of the first method is correct and the prediction of the second method is wrong.

The results are shown in last two columns of Tables I–IV: any value that is labelled with an asterisk is significantly different from the respective value for the Phyre ensemble (the last row of every table) at the 5% significance level. The test was performed for all results: training, testing, high confidence recall of correct homologous QT relationships, and high confidence annotation of individual queries. A Yates continuity correction[47] was used for small data values.

Precision, model quality, and the role of structural information

The 10 FR-systems used in the Phyre ensemble are methods 8, 9, 10, 11, 12, 13, 21, 22, 25, and 31. Many of the highest performing individual systems appear in this list (highlighted in Table I). However, several relatively poorly performing systems (including the worst performing) are also included. This is interesting as it highlights the presumption that false positive predictions made by one system in isolation can be overpowered by other more precise systems, yet the predictions made by noisy, low precision systems are adding value to the ensemble as a whole.

Equally interesting is that the boosted variants, which vary only in parameterisation, are included in the ensemble; even though individually these systems do not perform very differently from one another (in terms of percentage recall at high confidence), in combination they clearly boost performance.

Ensemble power by improved precision

When trying to decipher the source of the performance gain by the ensemble process, a clear pattern emerges when one analyses the frequency with which the 10 methods find the same QT pair (see Fig. 4). It can be seen that for true positive matches (i.e., correct homologous relationships) a large number are found by all methods, whereas it is far rarer for a correct match to only be found by a handful of methods. In contrast, when one observes false positive matches (i.e., erroneous hits) the vast majority is uniquely found by just a single method. This illustrates the source of most of the power of the ensemble. It is not so much different methods

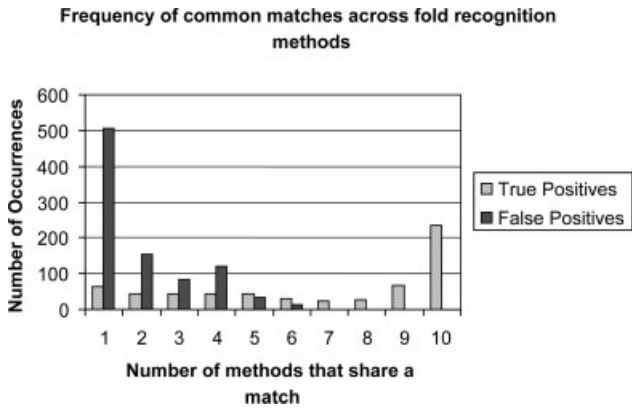

Figure 4

Histogram illustrating the difference in frequency distributions for all true positives and false positives, found above an empirical precision (EP) score of 0.7, across the 10 members of the best performing ensemble during testing. Query–template pairs with an EP score >0.7 were included in the analysis, that is, reasonably confident matches.

finding different homologies, but more the fact that the mistakes made by one method are rarely made by the rest of the ensemble. The mistakes made by the methods are often unique, or rare, and as a result the ensemble can effectively filter out false positive results.

Model quality improvement in ensembles

None of the ensembles used in this analysis perform any further modification of their final models; they simply select a structure from the pool of models produced by the constituent FR-systems.

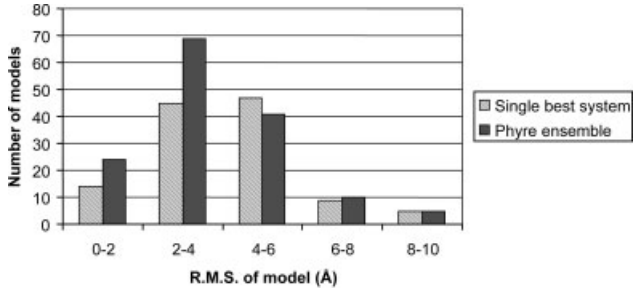

Figure 5

Histogram showing the RMSD of the models produced at 95% confidence (according to the 95% confidence point achieve in training) by the ensemble compared with those produced by the single best method. It can be seen that overall more models are produced by the ensemble, and that these tend to be at the higher accuracy ranges of 0–4 Å. Using bin sizes of 1 Å, gives a student t*-test probability of 0.02 (i.e., 98% probability the distributions are different).*

It is worth examining the quality of the models built using the ensemble compared with those built using the single best method. To make meaningful comparisons, we applied the single best performing method from the testing (30) and the final ensemble to the full test set of 50 proteins. For each of the two methods we gathered those test set models produced at 95% confidence (or above) according to where the 95% confidence point appeared during training. This point was chosen to make model selection as close to a "real world" scenario as possible. Each of these test models was then compared with the true structure of the protein in question. Figure 5 illustrates how the ensemble produces more correct matches (155 compared with 124 by the single best

	Correct or Incorrect	M_1	M_2	M_3	M_4
$Q_1 T_{11}$	Correct	✕	●	✕	✕
$Q_1 T_{12}$	Correct	●	✕	●	✕
$Q_1 T_{13}$	Incorrect	○	○	○	○
$Q_2 T_{21}$	Correct	✕	✕	✕	●
$Q_2 T_{22}$	Correct	●	●	●	●
$Q_2 T_{23}$	Correct	✕	●	✕	✕
$Q_3 T_{31}$	Correct	●	✕	✕	✕
$Q_3 T_{32}$	Incorrect	○	○	○	○
$Q_3 T_{33}$	Correct	✕	●	✕	●

Figure 6

An illustration of "impossible" data. Several queries (Q_1, Q_2, and Q_3) are listed against four different recognition methods (M_1, M_2, M_3, and M_4), along with some of the templates to which the queries are matched. Table cells filled with black crosses represent correct homologous query–template (QT) relationships that have been identified at empirical precision (EP) values of 0.95 (or above) by the respective recognition methods; these results are ignored (i.e., their EP values are reset to 0). Table cells filled with black circles represent correct homologous QT relationships (correct examples) that have been identified at EP values below 0.95 by the respective recognition methods; these results remain unchanged. Note that the same homologous QT relationships may be ignored in one recognition method but left unchanged in another. Table cells filled with open circles represent non-homologous QT relationships (incorrect examples); these remain unchanged regardless of whether they were identified above or below an EP value of 0.95.

DOI 10.1002/prot

method), and how most of these extra correct assignments are in the 0–2 Å and 2–4 Å RMSD bins. Thus the ensemble is adding considerably to our ability to model remote homologies at a resolution that is high enough to be of use to biology.

Sequence and structural features of remote homologues

Ensemble systems can only function when there is variety in their constituent methods. A major source of variety in the ensembles reported here is the presence or absence of predicted secondary structure information. It may be expected that those members of the ensemble that use predicted secondary structure as part of the scoring function for a match would have a greater tendency to detect matches with greater secondary structure similarity than sequence-only methods. We analysed the performance of each of the 10 members of the final ensemble by examining all QT pairs found by the methods at reasonable confidence (expected precision >70%; this lower expected precision score was chosen to increase the amount of data available for analysis). In testing, 64 QT pairs were uniquely found by methods using secondary structure information. Each of these QT pairs was structurally superimposed using Mammoth[40] and we measured the percentage identity of secondary structure elements in these alignments, that is, the fraction of the time a helical residue was matched with a helical residue and a strand residue was matched with a strand residue. The average secondary structure percentage identity of these 64 pairs was 64% (SD 11) and the average sequence identity was 12% (SD 4). Surprisingly, for the QT pairs uniquely found by the members of the ensemble that use sequence alone (20 pairs), the secondary structure percentage identity of query template pairs was remarkably similar at 66% (SD 12), whereas the sequence identity was far higher at 21% (SD 13). A *t*-test measuring the probability that these means differ due to chance alone returns 0.48 for the secondary structure differences and 0.005 for the sequence identity difference. Thus there is 99.5% confidence that the difference in sequence identity distributions is not due to chance. Therefore it appears that including predicted secondary structure information in FR methods does not permit one to detect homologues with more similar secondary structure distributions as expected. Instead, its power appears to lie in its indirect ability to detect homologues that are more remote in sequence space. Since the percentage coverage of QT relationships is not significantly different between methods that do include predicted secondary structure and those that do not, it is reasonable to conclude that methods that do include such information are capable of finding more distantly related homologies, but are also more likely to miss closer homologies, when compared with purely sequence-based methods. While such methods may not offer an increase in recall on an individual basis, in an ensemble such variety could potentially be a powerful addition.

Secondary structure is represented in these systems as a 3-state vector of probabilities, and only predicted secondary structure is used. Thus, one possible explanation for this phenomenon is that the 3-state vector is capturing non-local properties of the surrounding sequence (PsiPred uses a window of 15 residues as input to its neural network), and that these properties may be conserved even when a particular amino acid position has very different mutational propensities to its aligned partner.

Finally, to compare performance of the ensemble versus the single best method in an international blind trial, both techniques were entered into the CASP7 experiment (http://predictioncenter.org/casp7/). Although an experimental version of the ensemble was entered into CASP to test some novel techniques, across multiple prediction categories in CASP the ensemble system (designated Phyre-2 on the CASP website) consistently outperformed the single best method (Phyre-1). In an overall comparison of automated servers based on the GDT scores of their models to the native, Phyre-2 ranked 28th (13% worse than rank 1) whilst Phyre-1 ranked 46th (10% worse than Phyre-2 and 22% worse than rank 1). We believe this illustrates the potential advantages of ensemble systems. Future work will attempt to increase absolute performance by including further alignment techniques and more sophisticated profile generation strategies.

CONCLUSION

The Phyre ensemble is capable of striking performance enhancement: when tested on the "impossible" test set, 82% of queries and 58% of all relationships can be detected at high confidence by this ensemble. When tested using the full test set, the Phyre ensemble can confidently detect up to 64.0% of all correct homologous QT relationships, and 84.0% of the individual test query proteins could be accurately annotated at 95% precision or higher. In comparison to the improvement that the single best FR-system (number 25) has over PSI-Blast, this represents a 29.6% increase in the number of correct homologous QT relationships, and a 46.2% increase in the number of accurately annotated queries, for the Phyre ensemble (see Tables I and IV).

The application of ensembles of FR classifiers to remote homology detection can result in dramatic improvements in precision and recall. It is well known that simply pooling all the results from a set of classifiers is usually far from optimal and often inferior to a single system. One of the most difficult aspects of ensemble construction is the optimal selection of some subset of available methods to achieve maximum performance. This is because different methods are correlated to vary-

ing degrees in their output and have individual baseline accuracies. This complex trade-off between accuracy and diversity in an ensemble has not been satisfactorily dealt with to date. However, we demonstrate here that the use of either greedy algorithms or SVMs can ameliorate this problem. SVMs are capable of capturing higher order features generated by combinations of methods and can automatically derive appropriate weighting terms. However, they do not so readily produce models capable of distinguishing harder examples from easier ones. Greedy algorithms can give rough and quick solutions to the combinatorial search of all possible subsets of methods so that accuracy and diversity are appropriately balanced.

Structural clustering in the form of 3D-Jury and 3D-Colony is clearly a powerful tool. It harnesses simultaneously the similarity of templates with the consistency of alignments. The analogy to entropy in the 3D-Colony energy view of structural clusters helps explain, albeit rather abstractly, why this should be so. A more intuitive and direct explanation would be that there are more ways of being wrong than there are of being right. If two relatively diverse algorithms detect similar templates and generate similar alignments, then this strengthens the prediction. The fact that the component algorithms are selected in an empirical greedy way avoids some of the problem of reinforcing false negative predictions. This is also confirmed by the analysis of the degree to which true positive matches are shared across methods, whereas false positives are either uniquely found by one method, or shared by a small handful of methods.

The controlled development of meta-predictors demonstrates marked improvement over ad hoc combinations of systems. This is illustrated by the rise and fall of performance as methods are progressively added to an ensemble (Fig. 3). Clearly it is vital to choose methods wisely, or it is likely that one may actually reduce performance rather than enhance it. The controlled development in this work permits the use of a standardised scoring framework, which in turn permits the 3D-Colony approach. This approach is significantly superior to a more conventional 3D-Jury approach as illustrated in Tables I and III. The 3D-Colony approach improves upon 3D-Jury by weighting models by their confidence. Thus a small number of highly confident, self-similar models are more heavily weighted than a large number of weakly predicted self-similar models.

We suggest that the system could be further improved by using a more diverse pool of (already existing) algorithms with a more sophisticated search of the space of possible ensemble components—such as simulated annealing, and so forth. However, even at this stage the improvement is dramatic. The top performing ensemble system has been made available to the community as part of the Phyre protein structure prediction web server (http://www.imperial.ac.uk/phyre).

ACKNOWLEDGMENTS

We wish to thank Dr. Keiran Fleming (funded by the BBSRC) for his help constructing and maintaining our fold recognition database. Riccardo Bennett-Lovsey is sponsored by the MRC, and Alex Herbert is sponsored by the BBSRC. Lawrence Kelley is supported by the APRIL II project (PE0701).

REFERENCES

1. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213:859–883.
2. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
3. Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through the folding motif. Proteins 1993; 16:92–112.
4. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
5. Godzik A, Skolnick J. Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination. Proc Natl Acad Sci USA 1992;89:12098–12102.
6. Fischer D, Eisenberg D. Protein fold recognition using sequence-derived predictions. Protein Sci 1996;5:947–955.
7. Kelley LA, MacCallum RM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299:499–520.
8. Tang CL, Xie L, Koh IY, Posy S, Alexov E, Honig B. On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. J Mol Biol 2003;334:1043–1062.
9. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;14:846–856.
10. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14: 755–763.
11. von Ohsen N, Sommer I, Zimmer R. Profile–profile alignment: a powerful tool for protein structure prediction. Pac Symp Biocomput 2003:252–263.
12. Ohlson T, Wallner B, Elofsson A. Profile–profile methods provide improved fold-recognition: a study of different profile–profile alignment methods. Proteins 2004;57:188–197.
13. Panchenko AR. Finding weak similarities between proteins by sequence profile comparison. Nucleic Acids Res 2003;31:683–689.
14. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A. FFAS03: a server for profile–profile sequence alignments. Nucleic Acids Res 2005; 33:W284–W288.
15. Lundstrom J, Rychlewski L, Bujnicki J, Elofsson A. Pcons: a neural-network-based consensus predictor that improves fold recognition. Protein Sci 2001;10:2354–2362.
16. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018.
17. Fischer D. 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. Proteins 2003;51:434–441.
18. Fischer D. Hybrid fold recognition: combining sequence derived properties with evolutionary information. Pac Symp Biocomput 2000:119–130.
19. Rychlewski L, Zhang B, Godzik A. Fold and function predictions for Mycoplasma genitalium proteins. Fold Des 1998;3:229–238.
20. Ginalski K, von Grotthuss M, Grishin NV, Rychlewski L. Detecting distant homology with Meta-BASIC. Nucleic Acids Res 2004;32: W576–W581.

 DOI 10.1002/prot

21. Ginalski K, Pas J, Wyrwicz LS, von Grotthuss M, Bujnicki JM, Rychlewski L. ORFeus: Detection of distant homology using sequence profiles and predicted secondary structure. Nucleic Acids Res 2003;31:3804–3807.
22. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 2005;15:285–289.
23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
24. Yan Y, Moult J. Protein family clustering for structural genomics. J Mol Biol 2005;353:744–759.
25. Jain AK, Duin RPW, Mao JC. Statistical pattern recognition: A review. IEEE Trans Pattern Anal 2000;22:4–37.
26. Kuncheva LI, Whitaker CJ. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach Learn 2003;51:181–207.
27. Zhou ZH, Jiang Y, Yang YB, Chen SF. Lung cancer cell identification based on artificial neural network ensembles. Artif Intell Med 2002;24:25–36.
28. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
29. Xiang Z, Soto CS, Honig B. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. Proc Natl Acad Sci USA 2002;99:7432–7437.
30. Vapnik V, Chapelle O. Bounds on error expectation for support vector machines. Neural Comput 2000;12:2013–2036.
31. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res 2000;28: 254–256.
32. Yona G, Levitt M. Within the twilight zone: a sensitive profile–profile comparison tool based on information theory. J Mol Biol 2002; 315:1257–1275.
33. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
34. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics 2000;16:404–405.
35. Przybylski D, Rost B. Improving fold recognition without folds. J Mol Biol 2004;341:255–269.
36. Siew N, Elofsson A, Rychlewski L, Fischer D. MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 2000;16:776–785.
37. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.
38. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710.
39. Joachims T. Learning to classify text using support vector machines. Boston: Kluwer Academic Publishers; 2002. xvi, 205 pp.
40. Ortiz AR, Strauss CE, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 2002;11:2606–2621.
41. Nelder JA, Mead R. A simplex method for function minimization. Comput J 1965;7:308–313.
42. Lagarias JC, Reeds JA, Wright MH, Wright PE. Convergence properties of the Nelder-Mead Simplex method in low dimensions. SIAM J Optim 1998;9:112–147.
43. Goonesekere NC, Lee B. Frequency of gaps observed in a structurally aligned protein pair database suggests a simple gap penalty function. Nucleic Acids Res 2004;32:2838–2843.
44. Breiman L. Arcing classifiers. Ann Stat 1998;26:801–824.
45. Breiman L. Random forests. Mach Learn 2001;45:5–32.
46. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12:153–157.
47. Yates F. Contingency tables involving small numbers and the chi-squared test. J R Stat Soc 1934;Suppl 1:217–235.