BIOINFORMATICS

Vol. 19 no. 7 2003, pages 874–881
DOI: 10.1093/bioinformatics/btg097

Improvement of the GenTHREADER method for genomic fold recognition

Liam J. McGuffin and David T. Jones*

Bioinformatics Group, Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK

Received on October 1, 2002; revised on November 29, 2002; accepted on December 14, 2002

ABSTRACT

Motivation: In order to enhance genome annotation, the fully automatic fold recognition method GenTHREADER has been improved and benchmarked. The previous version of GenTHREADER consisted of a simple neural network which was trained to combine sequence alignment score, length information and energy potentials derived from threading into a single score representing the relationship between two proteins, as designated by CATH. The improved version incorporates PSI-BLAST searches, which have been jumpstarted with structural alignment profiles from FSSP, and now also makes use of PSIPRED predicted secondary structure and bi-directional scoring in order to calculate the final alignment score. Pairwise potentials and solvation potentials are calculated from the given sequence alignment which are then used as inputs to a multi-layer, feed-forward neural network, along with the alignment score, alignment length and sequence length. The neural network has also been expanded to accommodate the secondary structure element alignment (SSEA) score as an extra input and it is now trained to learn the FSSP Z-score as a measurement of similarity between two proteins.

Results: The improvements made to GenTHREADER increase the number of remote homologues that can be detected with a low error rate, implying higher reliability of score, whilst also increasing the quality of the models produced. We find that up to five times as many true positives can be detected with low error rate per query. Total MaxSub score is doubled at low false positive rates using the improved method.

Availability: http://www.psipred.net

Contact: l.mcguffin@cs.ucl.ac.uk

INTRODUCTION

The development of rapid and reliable, automatic protein fold recognition methods is important for a more comprehensive annotation of genomic sequences. Traditional pairwise sequence alignment methods can be used to assign folds to sequences with obvious evolutionary relationships to a known structure. Generally, for sequences with identities >30%, fast sequence searching methods such as FASTA (Pearson and Lipman, 1988) and WU-BLAST (Altschul and Gish, 1996) are fairly capable at detecting related proteins by scoring pairwise comparisons and compare in accuracy to the slower, Smith and Waterman (1981) based method SSEARCH (Pearson and Lipman, 1988). However when sequence identities fall below 30%, conventional pairwise sequence comparison methods fail to detect relationships (Brenner *et al.*, 1998), therefore, accurately annotating genes that encode proteins with low sequence identity to any known protein structure remains problematic.

Sequence searching has been improved beyond pairwise comparisons with the introduction of methods such as PSI-BLAST (Altschul *et al.*, 1997), ISS (Park *et al.*, 1997), SAM-T98 (Park *et al.*, 1999) and FFAS (Rychlewski *et al.*, 2000). These methods use information from profiles of related sequences in order to detect more distant relationships, however they often perform poorly at recognising non-homologous proteins with similar folds (Rychlewski *et al.*, 2000).

A number of other automatic methods have been developed that are designed to enhance sequence based searching by incorporating structural information, for example, INBGU (Fischer, 2000), 3D-PSSM (Kelley *et al.*, 2000), FUGUE (Shi *et al.*, 2001) and GenTHREADER (Jones, 1999a). For a comparison of structure prediction servers implementing the methods described above see LiveBench (Bujnicki *et al.*, 2001a,b, http://bioinfo.pl/LiveBench/), EVA (Valencia *et al.*, 2001, http://maple.bioc.columbia.edu/eva/) and CAFASP2 results (Fischer *et al.*, 2001, http://www.cs.bgu.ac.il/~dfischer/CAFASP2/).

In this paper we benchmark improvements to GenTHREADER including the incorporation of PSI-BLAST alignment profiles, which have been 'jump-started' or 'seeded' using FSSP structural alignments, bi-directional alignment scoring, PSIPRED predicted secondary structure and secondary structure element alignments (SSEAs).

*To whom correspondence should be addressed.

Structure is better conserved than sequence between distantly related proteins and it follows that rapid fold recognition methods which incorporate such structural information should therefore benefit from increased accuracy.

Here we develop and test several variations of the GenTHREADER protocol. The first method (GT) maintains the essential architecture of the original neural network, however the protocol is modified so that PSI-BLAST is used to generate sequence profiles in the fold library and the back propagation neural network is trained to learn the FSSP *Z*-score (Holm and Sander, 1994) representing the relatedness of proteins. In the second method (GT_FSSP) we use FSSP structural profiles to jump start the PSI-BLAST searches. In the third method (GT_FSSP_SS) we introduce PSIPRED predicted secondary structure to score the alignment, using a similar approach to that of Kelley *et al.* (2000). Finally, in the fourth method (mGT_FSSP_SS) we introduce bi-directional scoring similar to that used, for example, by the original mGenTHREADER (Jones, 1999a; McGuffin *et al.*, 2000), 3DPSSM (Kelley *et al.*, 2000) and FFAS (Rychlewski *et al.*, 2000). For each of these variations the secondary structure element alignment (SSEA) score (McGuffin *et al.*, 2001; McGuffin and Jones, 2002; Marsden *et al.*, 2002) is also incorporated as an extra input into the neural network (GT_SSEA, GT_FSSP_SSEA, GT_FSSP_SS_SSEA and mGT_FSSP_SS_SSEA).

Our findings suggest that incorporating structural information into the GenTHREADER protocol provides increased detection of remote homologues while maintaining a low error rate. We are also able to increase the accuracy of the generated models.

METHODS

Data

A set of 2727 FSSP files were downloaded from the FSSP database (Holm and Sander, 1996) at ftp://ftp.ebi.ac.uk in the directory /pub/databases/fssp. (N.B. This data set will be referred to as FSSP2727 further on in the text).

Observed secondary structures. Secondary structure strings were generated for all proteins using the DSSP secondary structure definition (Kabsch and Sander, 1983). The eight states (H, I, G, E, B, S, T, -) were reduced to three states such that H and G are taken as helix states, E and B are equal to strand states and all other states equal coil.

Predicted secondary structures. Secondary structures were predicted for all proteins using the secondary structure prediction method, PSIPRED (Jones, 1999b). To ensure that predictions were properly cross-validated, PSIPRED was trained using six different training sets producing six different sets of neural network weights. If the target sequence was found to have a homologous sequence in a training set then the corresponding set of weights was excluded.

Similarity scores

All against all comparisons were made for the proteins within the FSSP2727 dataset using the following similarity scoring methods.

Alignment score. A non redundant dataset of 771530 proteins (which will be referred to as NR-UCL) was compiled by concatenating sequence files, in FASTA format, from the following databanks: PDB (Berman *et al.*, 2000), SWISSPROT and TREMBL (Bairoch and Apweiler, 2000), PIR (Barker *et al.*, 2001), ENSEMBL (Birney *et al.*, 2001), WORMPEP (http://www.sanger.ac.uk/Projects/C_elegans/wormpep/) and GENPEPT (ftp://ftp.ncifcrf.gov/pub/genpept/). Each protein sequence in the FSSP2727 dataset was scanned against the NR-UCL plus the FSSP2727 data set using PSI-BLAST (Altschul *et al.*, 1997) with ten iterations. The PSI-BLAST generated alignments were re-scored using the BLOSUM62 matrix; the highest score was taken between the PSI-BLAST score and the BLOSUM62 re-score.

Threading potentials. The PSI-BLAST sequence alignment for each pair was evaluated using threading potentials described by Jones *et al.* (1992); Jones (1999a). The pair energy—pairwise potentials of mean force—was derived using the inverse Boltzman equation originally described by Jones *et al.* (1992). The solvation energy—relating to the degree of residue burial—was also calculated from the equation published by Jones *et al.* (1992).

Secondary Structure Element Alignment (SSEA) score. SSEAs were scored by aligning predicted and observed secondary structure elements using a dynamic programming algorithm based on that of Needleman and Wunsch (1970) and using a scoring scheme similar to Przytycka *et al.* (1999) (Table 1). Secondary structure element alignment using this scheme has been previously described by McGuffin *et al.* (2001); McGuffin and Jones (2002) and Marsden *et al.* (2002).

The modified GenTHREADER protocol (GT)

In this study the original GenTHREADER protocol (Jones, 1999a) was modified. Sequence profiles for the fold library were generated using PSI-BLAST. The neural network architecture was also modifed to give a single output value and the network was trained to learn FSSP *Z*-scores, as a measurement of similarity between two proteins, rather than the binary CATH relationship which had been used previously (Jones, 1999a).

Table 1. Score matrix for SSEA

	C_1	H_1	E_1
C_2	$\min(\text{len}(C_1), \text{len}(C_2))$	$0.5\min(\text{len}(H_1), \text{len}(C_2))$	$0.5\min(\text{len}(E_1), \text{len}(C_2))$
H_2	$0.5\min(\text{len}(C_1), \text{len}(H_2))$	$\min(\text{len}(H_1), \text{len}(H_2))$	0
E_2	$0.5\min(\text{len}(C_1), \text{len}(E_2))$	0	$\min(\text{len}(E_1), \text{len}(E_2))$

C_1 is a coil element in the observed secondary structure, H_2 is a helix element in the predicted secondary structure and $0.5\min(\text{len}(C_1), \text{len}(H_2))$ is half the minimum length of these aligned elements. The overall score was normalised by the mean sequence length of the observed and predicted secondary structures to give a similarity score ranging between 0 and 1.

Neural network architecture. Pairwise energy sum, solvation energy sum, alignment score, alignment length, length of the template protein and length of the target protein were used as inputs to a feed forward, multi-layer, back propagation neural network. Each score or input value was scaled to lie in the 0–1 range using the standard logistic function:

$$\text{score} = \frac{1}{1 + e^{-a(x-b)}} \quad (1)$$

where x is the raw input value and a and b are constants such that; for the pair energy, $a = 1$ and $b = 100$; for the solvation energy, $a = 1$ and $b = 10$; for the alignment score, alignment length, template length and target length, $a = 0.01$ and $b = 150$.

The GT network consisted of three neural layers: six neurons in the input layer, six neurons in the hidden layer, and one neuron in the output layer.

Cross validation and training of neural network. Both the training sets and test sets were derived from the FSSP2727 set. The FSSP2727 set of proteins was randomly split into four approximately equal sub test sets. Each sub test set was searched against the whole FSSP2727 set for sequence homologues using FASTA. For each sub test set, FASTA hits with E-values <0.01 or sequence identity $>30\%$ were discarded producing a list of non-homologues that could be used for training. This resulted in four separate jack-knifed training sets.

The network was trained, on each of the four training sets, to recognise relationships between proteins as measured by the FSSP Z-score. Protein pairs in the range $6.0 > Z > 4.0$ were left out in order to minimise ambiguity (see Results section— *Defining a fold*). FSSP Z-scores had also been scaled to lie between 0 and 1 using the standard logistic function (Equation 1, where $a = 0.1$ and $b = 5.0$). The training resulted in four separate sets of neural network weights which were then saved and loaded into the neural network for use with each corresponding sub test set. The neural network output scores from each of the test sets were then pooled to produce a list of all pairwise comparisons.

Further improvements to GenTHREADER

Further improvements were made as follows:

GT_FSSP. The '-B' option in PSI-BLAST allows searches to be jump started from a given multiple alignment (see PSI-BLAST documentation for further instructions). The given alignment is used to construct an initial position specific scoring matrix from which the search is started. FSSP jump started PSI-BLAST profiles were used for the fold library in order to increase the detection of remote homologues.

GT_FSSP_SS. Alignments were scored according to a simple secondary structure matching scheme, similar to that used by (Kelley *et al.*, 2000). For each alignment the predicted secondary structure of the target was compared with the observed secondary structure of the template. Aligned residues with matching secondary structure types were positively weighted and mismatches were negatively weighted.

mGT_FSSP_SS. The alignment score of each target profile searched against the template sequences was taken in addition to the alignment score of each template profile searched against the target sequences. The highest scoring alignment was then chosen. This has previously been referred to as 'bi-directional scoring' (e.g. see Jones, 1999a; Kelley *et al.*, 2000; Rychlewski *et al.*, 2000). Target profiles were constructed using a standard PSI-BLAST search against NR-UCL dataset i.e. no FSSP profiles were used for the targets.

Addition of the SSEA score as an extra input to neural networks (GT_SSEA, GT_FSSP_SSEA, GT_FSSP_SS_SSEA, mGT_FSSP_SS_SSEA). The improved protocols were followed, as outlined above, however an additional neuron was included in the input layer of the neural networks, in order to accommodate the SSEA score. As SSEA scoring produces a score between 0 and 1 no scaling was required for this additional input. An additional neuron was also added to the hidden layer.

Each of these neural networks were trained and cross-validated in the same way as GT, described in the

preceding section. Where secondary structure was used in training, the observed secondary structures of pairs of proteins were aligned. However, during testing, the predicted secondary structures of the target proteins were aligned to the observed secondary structures of the templates.

RESULTS

Benchmarking the improved methods

FSSP classification as a 'gold standard'. The FSSP database is a fully automatic protein structure classification database which is compiled by carrying out all against all 3D comparisons on structures within the PDB, using the Dali search engine (Holm and Sander, 1993). An FSSP file for a given protein representative contains a list of structurally aligned homologues which are ranked according to the Z-score, which relates to the strength of structural similarity in standard deviations above expected. One of the advantages of using FSSP as a 'gold standard', to train and benchmark methods, is that relationships between multi domain proteins have been classified, whereas in SCOP (Murzin and Bateman, 1997) and CATH (Orengo *et al.*, 1997), domain folds are generally classified as separate units. Another advantage of using FSSP as a benchmark is that it better reflects the upper limit that could be achieved by a fully automatic fold assignment method, i.e. if the 3D structure is actually known how well can an automatic method assign folds? FSSP Z-scores also provide us with a range of 'similarity scores' between protein chains, as opposed to a binary classification system, such as in SCOP or CATH, where proteins either have the same fold or they do not.

Defining a fold. The definition of a fold can sometimes be ambiguous and it has been shown previously that protein classification systems may often disagree on whether two protein domains share the same fold or not (Hadley and Jones, 1999; McGuffin *et al.*, 2001). In this study, there is the added complication of the inclusion of multi-domain proteins in the database, which may contain more than one globular folding unit. Here we take Z-scores $\geqslant$6.0 to indicate similar 'folds' and Z-scores $\leqslant$4.0 to indicate dissimilar 'folds'. The highest Z-score between each pair is taken (see Hadley and Jones, 1999 for a detailed analysis of FSSP Z-scores in comparison to SCOP and CATH fold definitions).

Screening easy hits. The FSSP2727 set was screened using PSI-BLAST E-value cutoff of $E < 0.1$ reducing the size of the data set to 973. Using a Z-score cut-off of $\geqslant$6.0 to indicate protein pairs with similar 'folds', the number of proteins with a matching fold in the data set—or the number of 'known folds'—was 375. Out of 945 756 pairwise comparisons (973^2–973), 6006 pairs have matching folds, leaving 939 750 mismatches.

Confidence estimation and definitions of terms. An important consideration in benchmarking fold recognition is the consistency and reliability of similarity score. In this analysis we carry out a number alternative measures of confidence in similarity score.

The terms used in Figures 1 and 2, and in Table 2 are defined as follows; True positives, the number of matching pairs above a certain score cut-off (i.e. network output or similarity score); False positives, the number of mismatching pairs above a certain score cut-off; (EPQ) Error per query, false positives detected out of the total number of known folds (375); Selectivity, true positives divided by the sum of true positives and false positives; Coverage, the number of correctly assigned top hits divided by the total number of known folds (375).

Figure 1 shows how the error per query increases with increasing detection of true positives. This plot is similar to those produced by Brenner *et al.* (1998). GenTHREADER with incorporated SSEA scores (GT_SSEA) shows an increase in true positive detection over the original method (GT), at similar error per query rates. More strikingly, about twice the number of true positives can be detected when structural profiles are incorporated (GT_FSSP) and this number can be further increased when SSEA scores, structural profiles and secondary structure matching are included (GT_FSSP_SS_SSEA). However, in this plot there appears to be no overall improvement in the reliability of score when bi-directional scoring is taken into account (mGT_FSSP_SS and mGT_FSSP_SS_SSEA).

From the results in Figure 1 we can determine the selectivity values shown in Table 2. These values reflect the confidence in neural network output.

Benchmarking of methods on LiveBench targets

LiveBench (Bujnicki *et al.*, 2001a,b) is a continuous, fully automated, large-scale project to evaluate structure prediction servers. LiveBench targets are selected from newly released PDB entries which show no trivial sequence similarity to any previous PDB entry (BLAST E-value $>$0.1). In order to determine how each method might compare in such a purely automatic blind assessment, 81 targets were downloaded from the ongoing LiveBench-4 evaluation (http://bioinfo.pl/LiveBench/, entries from 2001-11-07 to 2002-03-07 inclusive). 23 of the targets were classified in LiveBench as 'easy' and the remaining 58 were classified as 'hard'. This set of targets was compared against the library of 2727 templates (FSSP2727), using each method.

Analysis of prediction quality using MaxSub. MaxSub (Siew *et al.*, 2001; http://www.cs.bgu.ac.il/~dfischer/MaxSub/) was the official method used to automatically

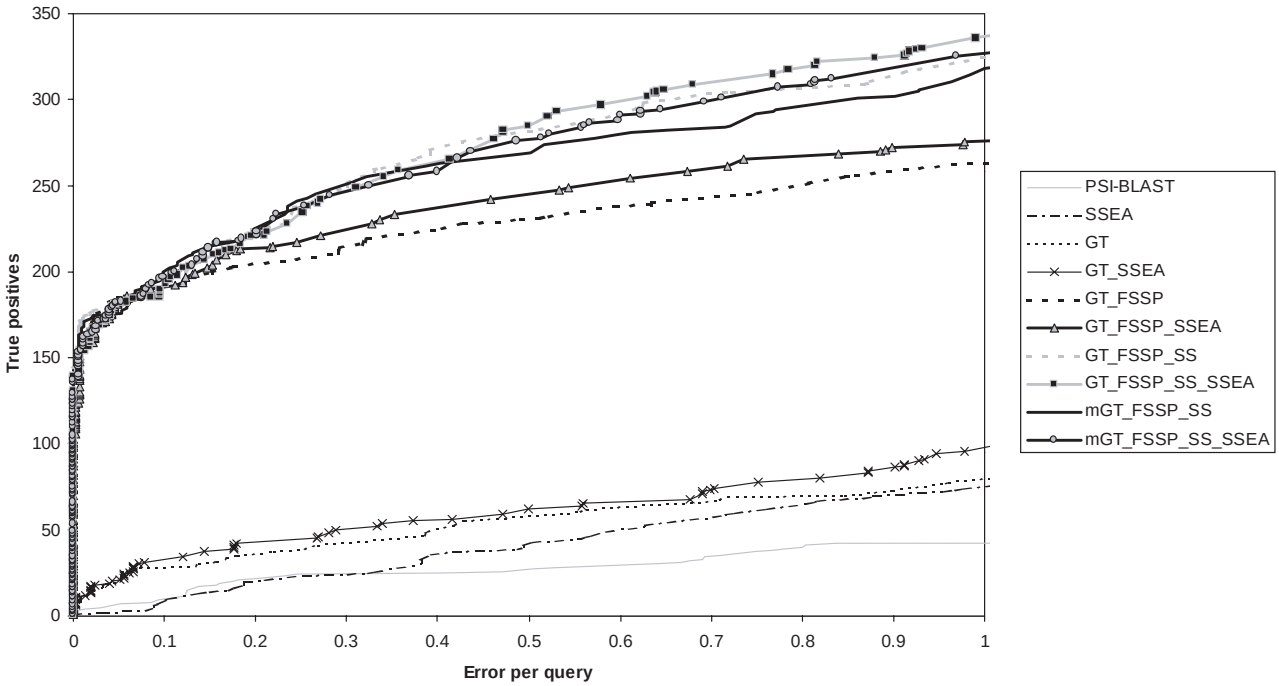

Fig. 1. Confidence estimation—coverage versus error plot. True positives are plotted against error per query (see Results section for definition of terms). PSI-BLAST, position specific iterative basic local alignment search tool (Altschul *et al.*, 1997), SSEA, secondary structure element alignment (predicted versus known secondary structure); GT, GenTHREADER; GT_FSSP, GenTHREADER using FSSP structural profiles; GT_FSSP_SS, GenTHREADER with structural profiles and PSIPRED predicted secondary structure incorporated into alignment score; mGT_FSSP_SS, as the previous method but with bi-directional scoring. The extension _SSEA to methods indicates the incorporation of secondary structure element alignment score (predicted versus known secondary structure) as an added input to the underlying neural network. All pairs of proteins in the analysis have PSI-BLAST *E*-values $\geqslant$0.1.

Table 2. Selectivity versus neural network output for the improved version of GenTHREADER incorporating structural alignments profiles, predicted secondary structure, secondary structure element alignments and bi-directional scoring (mGT_FSSP_SS_SSEA). See Results section for definition of terms. All pairs of proteins in the analysis have PSI-BLAST E-values $\geqslant$ 0.1

Network score cut-off	True positives	False positives	Selectivity	Coverage
0.615	137	0	1.0	0.253
0.548	183	19	0.9	0.341
0.517	214	53	0.8	0.387
0.502	244	101	0.7	0.416

assess prediction quality at CAFASP-2 (Fischer *et al.*, 2001) and it is also one of the methods used to evaluate the performance of prediction servers participating in LiveBench. The MaxSub method produces a single score, between 0 and 10, representing the quality of the models (correct models are those with scores $>$ 0). The score is normalised such that many scores may simply be added across a number of targets to produce total scores for each structure prediction method (Siew *et al.*, 2001).

The 'model quality versus error' plot in Figure 2 allows us to assess the influence of each improvement on model quality. The cumulative MaxSub score is plotted against false positives. From this plot we can see a gradual increase in MaxSub total as each improvement is made. It can be seen that mGT_FSSP_SS_SSEA clearly outperforms the other methods in terms of model quality with similar numbers of false positives.

Figure 3 shows a couple of examples of the improvements to model quality. In Figure 3a the same top hit is selected by GT as is selected by mGT_FSSP_SS_SSEA for LiveBench target 1hqz1, however the alignment produced by the improved method is more accurate. Figure 3b

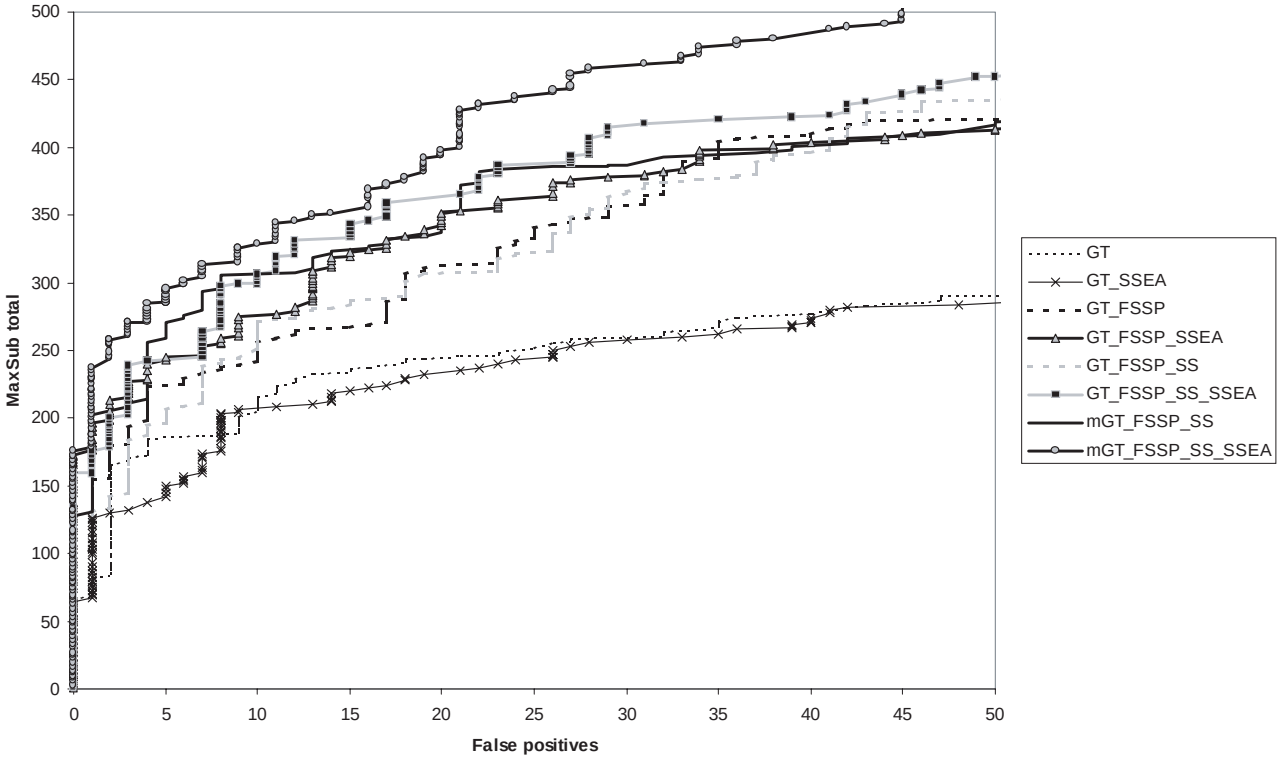

Fig. 2. Benchmarking methods on LiveBench targets (Bujnicki *et al.*, 2001b)—MaxSub cumulative total is plotted against false positives as an assessment of model quality and reliability. For definitions of methods see legend of Figure 1.

shows an example where mGT_FSSP_SS_SSEA selects a different top hit and produces a better model than GT for LiveBench target 1ktzB.

DISCUSSION

The results of this study indicate that the incorporation of structural information, such as FSSP structural alignment profiles and alignments of PSIPRED predicted secondary structures to observed secondary structures, increases both the reliability of the GenTHREADER method and the accuracy of the alignments produced.

In a previous study the alignment of predicted secondary structure elements of a target to a library of DSSP assigned secondary structures (SSEA) was shown to detect more remote homologues than could be detected using purely sequence based methods such as the original GenTHREADER method (McGuffin and Jones, 2002). However, for automated fold recognition methods to be useful it is important that the scoring scheme used for fold recognition is consistent and reliable so that scores are comparable. One of the disadvantages of relying on SSEA scoring for genome annotation is the low consistency of the scoring scheme

Coverage versus error plots, such as those produced by Brenner *et al.* (1998) allow us to compare the usefulness of methods for genome annotation by placing a premium on the consistency of the score. The plot in Figure 1 clearly illustrates that SSEA is a poor method to use alone if a low error rate per query is required. However, combining the SSEA score with GenTHREADER (GT) increases the number of true positives whilst maintaining a low error rate (GT_SSEA).

Using structural profiles from FSSP to seed the initial sequence searches noticeably enhances fold recognition (GT_FSSP). The number of distant homologues correctly assigned is effectively quadrupled at low error rates, suggesting more consistent scores. Adding SSEA in combination with structural profiles increases the number of correct assignments slightly further (GT_FSSP_SSEA).

Incorporating secondary structure matching directly into the alignment scoring scheme increases the reliability of the score further still (GT_FSSP_SS). Although there remains some advantage in combining this alternative secondary structure scoring method with SSEA as an extra input to the neural network (GT_FSSP_SS_SSEA).

In addition to the score reliability, the alignment accuracy is also an important consideration when assessing fold recognition methods, as accurate sequence-to-

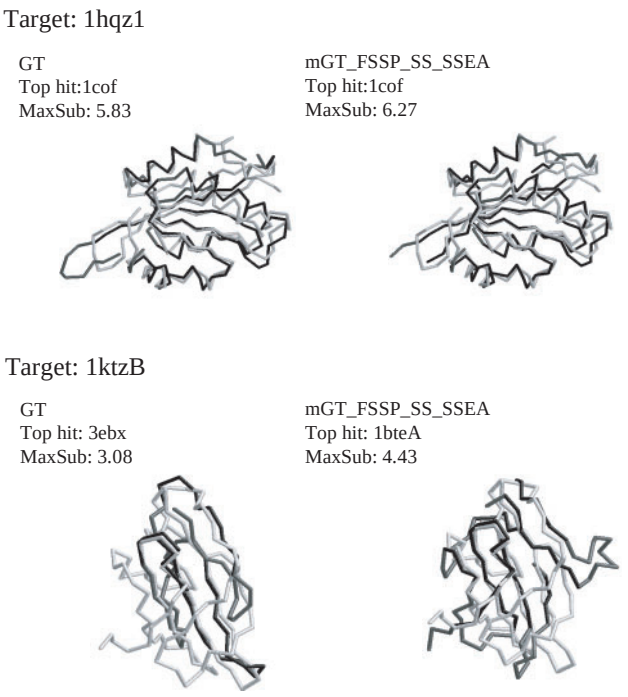

Fig. 3. Superposition of backbones of targets (grey) with backbones of predicted models (white). Residues of model within 3.5 Å of target are shown in black. MaxSub, MaxSub score between target and model. For definitions of methods see legend of Figure 1. a, LiveBench easy target 1hqz1—an example of improvement in alignment quality. The same top hit is selected by both GT and mGT_FSSP_SS_SSEA but the MaxSub score is improved by the latter method (right model). b, LiveBench hard target 1ktzB—a more appropriate model is selected by the improved method (right model) compared to the GT model (left model).

structure alignments imply accurate models. If a fast method such as SSEA were to be used alone for genome annotation, alignment accuracy would be an issue, as it is difficult to relate the alignment of intact secondary structure elements to an alignment on a residue-by-residue basis. However, by incorporating SSEA scores into GenTHREADER(GT_SSEA) we benefit from the detection of more remote homologues whilst also generating a usable sequence to structure alignment.

The advantage of carrying out bi-directional scoring becomes clear when using MaxSub to score model quality. In the cumulative MaxSub score versus false positives plot (Figure 2), mGT_FSSP_SS_SSEA is shown to achieve the highest cumulative MaxSub score with low numbers of false positives.

The effect of the addition of the FSSP profiles and the secondary structure matching is to directly increase the alignment accuracy. An example of this can be seen in Figure 3a where the same top hit is chosen by both GT and mGT_FSSP_SS_SSEA, however the MaxSub score is increased. Conversely, the addition of the bi-directional score and the SSEA score only alters the neural network output score and there is no change in the alignment. However, as a consequence of the change in neural network output, hits may be ranked differently so that a different top hit for a particular target may be selected. Figure 3b shows an example where a different top hit is selected by the method mGT_FSSP_SS_SSEA which is shown to have a higher MaxSub score than the top hit selected by GT.

It must be said that here we have focused primarily on improving the quality of the input data to the underlying neural network in order to improve GenTHREADER. We have purposefully kept the neural network architectures simple in order to assess directly the effect of the improvements and no rigorous attempt has yet been made at optimising the machine learning. However, we are currently investigating the possibility of using a support vector machine in order to enhance the machine learning aspects of GenTHREADER.

CONCLUSIONS

In this study we have shown that the incorporation of secondary structure element alignments (SSEAs), FSSP structural alignment profiles, secondary structure matching and bi-directional scoring into the GenTHREADER protocol increases the number of folds that can be correctly assigned, improves the consistency of score and increases the quality of models. The reliability for recognising the folds of evolutionarily distant proteins is approximately 4 to 5 times that of the original GenTHREADER method when using these modifications. The model quality is approximately twice that of the original method according to automated assessment method MaxSub at low false positive rates. We expect that improved versions of GenTHREADER, using these modifications, will provide a more reliable, accurate and comprehensive automated annotation of completed genomes.

ACKNOWLEDGEMENTS

This work was supported by the Biotechnology and Biological Sciences Research Council and Inpharmatica Ltd., London. We would like to thank Alejandro Schaffer at the National Institutes for Health, USA for his useful comments concerning PSI-BLAST.

REFERENCES

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

Altschul,S.F. and Gish,W. (1996) Local alignment statistics Methods. *Enzymol.*, **266**, 460–480.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein data base search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL. *Nucleic Acids Res.*, **28**, 45–48.

Barker,W.C., Garavelli,J.S., Hou,Z., Huang,H., Ledley,R.S., McGarvey,P.B., Mewes,H.W., Orcutt,B.C., Pfeiffer,F., Tsugita,A., Vinayaka,C.R., Xiao,C., Yeh,L.L. and Wu,C. (2001) Protein Information Resource: a community resource for expert annotation of protein data. *Nucleic Acids Res.*, **29**, 29–32.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Birney,E., Clamp,M., Kraspcyk,A., Slater,G., Hubbard,T., Curwen,V., Stabenau,A., Stupka,E., Huminiecki,L. and Potter,S. (2001) Ensembl: a multi-genome computational platform. *Am. J. Hum. Genet.*, **69**, 219.

Brenner,S.E., Chothia,C. and Hubbard,T.J.P. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.

Bujnicki,J.M., Elofsson,A., Fischer,D, and Rychlewski,L. (2001a) LiveBench-1: Continuous benchmarking of protein structure prediction servers. *Protein Sci.*, **10**, 352–361.

Bujnicki,J.M., Elofsson,A., Fischer,D. and Rychlewski,L. (2001b) LiveBench-2: large-scale automated evaluation of protein structure prediction servers. *Proteins*, **45**, 184–191.

Fischer,D., Elofsson,A., Rychlewski,L., Pazos,F., Valencia,A., Rost,B., Ortiz,A.R. and Dunbrack,R.L. (2001) CAFASP2: the second critical assessment of fully automated structure prediction methods. *Proteins*, **45**, 171–183.

Fischer,D. (2000) Hybrid fold recognition: combining sequence derived properties with evolutionary information. In *Proceedings of the Pacific Symposium on Biocomputing*. World Scientific Press, Hawaii, pp. 119–130.

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. (1996) Mapping the protein universe. *Science*, **273**, 595–602.

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

Holm,L and Sander,C. (1993) Protein sructure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–128.

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

Jones,D.T. (1999b) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

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

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.

McGuffin,L.J., Bryson,K. and Jones,D.T. (2001) What are the baselines for protein fold recognition? *Bioinformatics*, **17**, 63–72.

McGuffin,L.J., Bryson,K. and Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.

McGuffin,L.J. and Jones,D.T. (2002) Targeting novel folds for structural genomics. *Proteins*, **48**, 44–52.

Marsden,R.L., McGuffin,L.J. and Jones,D.T. (2002) Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Sci.*, **11**, 2814–2824.

Murzin,A.G. and Bateman,A. (1997) Distant homology recognition using structural classification of proteins. *Proteins*, **29**, 105–112.

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

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

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.

Park,J., Karplus,K., Barret,C., Hughey,R., Haussler,D., Hubbard,T. and Chothia,C. (1999) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.*, **284**, 1201–1210.

Park,J., Teichmann,S.A., Hubbard,T. and Chothia,C. (1997) Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol.*, **273**, 349–354.

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

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

Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.*, **9,**, 232–241.

Siew,N., Elofsson,A., Rychlewski,L. and Fischer,D. (2001) MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, **16**, 776–785.

Shi,J., Blundell,T.L. and Mizuguchi,K. (2001) FUGUE: Sequnece-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.*, **310**, 243–257.

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

Valencia,A., Sali,A. and Rost,B. (2001) EVA: continuous automatic evaluation of protein structure prediction servers. *Bionformatics*, **17**, 1242–1243.