ABSTRACT
Motivation: Generation of structural models and recognition of homologous relationships for unannotated protein sequences are fundamental problems in bioinformatics. Improving the sensitivity and selectivity of methods designed for these two tasks therefore has downstream benefits for many other bioinformatics applications.
Results: We describe the latest implementation of the GenTHREADER method for structure prediction on a genomic scale. The method combines profile-profile alignments with secondary-structure specific gap-penalties, classic pair- and solvation potentials using a linear combination optimized with a regression SVM model. We find this combination significantly improves both detection of useful templates and accuracy of sequence-structure alignments relative to other competitive approaches. We further present a second implementation of the protocol designed for the task of discriminating superfamilies from one another. This method, pDomTHREADER, is the first to incorporate both sequence and structural data directly in this task and improves sensitivity and selectivity over the standard version of pGenTHREADER and three other standard methods for remote homology detection.

Contact: d.jones@cs.ucl.ac.uk
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION
Predicting the tertiary structure of a novel protein and determining its relationship to other proteins are two problems of fundamental importance in bioinformatics. Recognizing the correct fold of a protein enables 3D models to be constructed which are essential for structure based drug design, enzyme engineering programmes and general functional analyses. Distinguishing between homologous superfamilies often reveals additional insights into the precise function(s) of the protein through fine sub-classification between similar structures.

At present, the most successful methods for tertiary structure prediction use the structure of template proteins as the basis for their predictions (Moult et al., 2007), based on the scoring of alignments computed between protein sequence profiles (Mittelman et al., 2003; Panchenko, 2003; Rychlewski et al., 2000; Yona and Levitt., 2002). More sophisticated approaches combine sequence profile alignments with structural information and model quality assessment (McGuffin et al., 2006; Zhang, 2007; Zhang et al., 2008; Zhou et al., 2007).

Where sequence conservation is high (i.e. detectable using BLAST; Altschul et al., 1990) structure prediction methods can generate models which are marginally better than any available template, although it remains challenging to reliably select models which achieve this standard (Read and Chavali, 2007). In this region, accessible to standard homology modelling methods such as SWISS-MODEL (Schwede et al., 2003), the observation that sequence conservation following evolutionary divergence always implies conservation of the tertiary structure of the protein suggests structural similarity and homologous relationships are synonymous, and therefore that methods for structure prediction can also be used for inference of distant homologous relationships.

Significant similarities may occur by chance, or represent similar subsequences between architecturally distinct structures (Harrison et al., 2002; Reeves et al., 2006). This observation has resurrected the earlier controversy over whether structural space is continuous or not, and raises questions regarding the reliability of structural conservation as an indicator of a homologous relationship (Abagyan and Batalov, 1997; Cheng et al., 2008; Grishin, 2001; Orengo and Thornton, 2005).

For predicting tertiary structure this is of no consequence since a correct prediction does not necessarily require an evolutionary relationship: the most successful methods for predicting ‘new fold’ targets rely on the re-use of arbitrary structural fragments in the absence of homology between the target sequence and the fragment source (Jones, 2001; Rohlf et al., 2004; Zhang, 2007). Conversely, for superfAMILY discrimination this is clearly a very important issue (Reid et al., 2007).

Since the two problems require similar information a single, generic method is often applied. However, for recognition of distant folds and discrimination between related superfamilies it is reasonable to expect that different information might be more or less important to each task. We therefore developed separate, related methods to address each challenge.

We present pGenTHREADER and pDomTHREADER: two improved versions of the GenTHREADER protocol (Jones, 1999; McGuffin and Jones, 2003) for recognizing and aligning protein sequences and demonstrate their application to structure prediction.
and superfamily discrimination. The two versions use the same core alignment algorithm and in both cases accept features derived from common inputs: protein sequence profiles and structural information. However, the representation and combinations of these features differ between the methods and scoring and confidence values have been tuned to optimize performance in each application domain.

We assessed the performance of these novel implementations using two benchmark sets: a non-redundant consensus domain set for structure prediction and the CATH S35 representative sequences for superfamily identification. For control purposes we use the PSI-BLAST profile–sequence method, the HHpred HMM-HMM comparison tool and the PRC profile–profile comparison tool.

We find that the efficient combination of sequence and structural features using machine learning techniques provides a significant improvement to distant homology recognition and alignment which leads to improved performance in both application areas.

2 METHODS

2.1 pGenTHREADER : parametric profile–profile based fold recognition

The new fold recognition method employed here (parametric-GenTHREADER - pGenTHREADER) is a significant development of our earlier algorithm (nGenTHREADER). Similarly to nGenTHREADER profile–profile comparisons from PSI-BLAST position-specific scoring matrices (PSSM matrices) were built using eight iterations of PSI-BLAST (-j 8) against Uniref90 (Birney et al., 2007) sequences with low complexity regions, coiled coil regions and transmembrane segments filtered out. The profile–profile scoring scheme is based on the weighted sum of the dot products of the two PSSM vectors \( X \) (from the target sequence at position \( x \)) and \( Y \) (from the template at position \( y \)), and their respective target frequency transformations, \( X^{TF} \) and \( Y^{TF} \):

\[
g_{p}^{prof}(X,Y) = \sum_{i=1}^{20} \sum_{j=1}^{20} X^{TF}_{x_{i}} Y^{TF}_{y_{j}} \cdot w_{ij}^{prof} + 20 \cdot \sum_{i=1}^{20} X^{TF}_{x_{i}} \cdot w_{i}^{prof} + 20 \cdot \sum_{j=1}^{20} Y^{TF}_{y_{j}} \cdot w_{j}^{prof} + 20 \cdot \sum_{i=1}^{20} \sum_{j=1}^{20} X^{TF}_{x_{i}} \cdot Y^{TF}_{y_{j}}
\]

Two additional scores for sequence–profile and profile–sequence are also combined into the final score. The sequence–profile and profile–sequence scores controlled for profile drift ensuring high scoring profile–profile matches that differed from high scoring sequence–profile or profile–sequence matches to be down-weighted in the final score:

\[
g_{p}^{seq prof}(X,Y) = \sum_{i=1}^{20} \sum_{j=1}^{20} X_{x_{i}} Y_{y_{j}} \cdot w_{ij}^{seq prof} + 20 \cdot \sum_{i=1}^{20} X_{x_{i}} \cdot w_{i}^{seq prof} + 20 \cdot \sum_{j=1}^{20} Y_{y_{j}} \cdot w_{j}^{seq prof} + 20 \cdot \sum_{i=1}^{20} \sum_{j=1}^{20} X_{x_{i}} \cdot Y_{y_{j}}
\]

Where \( a_{ix} \) is the amino-acid type in the target sequence at position \( x \) and \( a_{iy} \) is the amino-acid type observed in the template sequence at position \( y \). Weights \( w_{1} \ldots w_{20} \) are adjustable weights (see optimization procedure below).

Additional terms were added (again with adjustable weights) to account for agreement between predicted and observed secondary structure and hydrophobic burial. For secondary structure, a 3 x 3 matrix of similarity scores was used for the each predicted secondary structure state (helix, strand and coil) compared against the same three secondary structure states in the template structure. For example, a residue predicted to be a helix by PSIPRED (Jones, 1999b) would accrue a large negative score if aligned with a template position observed to be in a b-strand. The exact values used in this 3 x 3 score matrix were treated as adjustable parameters in the optimization procedure.

To bias alignments towards correct positioning of hydrophobic groups in the target sequence, a hydrophobic burial term was added to the final score based on the solvation potential as already used in the GenTHREADER algorithm. The final change to the alignment algorithm involved the implementation of secondary structure dependent opening and extension gap penalties. For the three secondary structure states observed in the template (helix, strand and coil) adjustable affine gap penalties can be specified. Again, the final values of these gap penalties were determined by optimization.

In total, the behaviour of the fold recognition alignment algorithm is specified by 20 adjustable parameters: four for the profile–profile scoring function, nine for the secondary-structure scoring matrix, six for the gap penalties and finally a weighting for the burial term.

Weight parameters were initially optimized using a coarse grid search and refined using a generic algorithm to maximize the sum of TM-scores (Zhang and Skolnick, 2004) for each top hit across a benchmark set of 158 fold recognition targets taken from LiveBench-8 (Rychlewski and Fischer, 2005). Optimized parameter values appear as Supplementary Table S1.

Both pGenTHREADER and pDomTHREADER employ a similar approach to deriving a single measure of confidence from the profile–profile score, the pair and solvation potential terms. For both methods, a linear regression model using logistic functions were used to rescale several of the input features. A comparison of the parameter choices and features is shown in Supplementary Table S2.

In the case of pGenTHREADER, the feature weights were optimized using linear SVM regression and using the complete set of chain pairs from LiveBench-8. The target values in this case were 3D-scores. Match \( P \)-values were determined by fitting the density of predicted 3D-Scores for false matches (3D-Score < 30) to an extreme-value distribution.

pDomTHREADER was trained in classification mode to provide a clearer distinction between separate homologous superfamilies that could be aligned with high scores. The classification target comprised pairs of CATH S35 representative sequences and 5-fold cross validation experiments were carried out to establish the best parameters for the linear SVM model. The bias parameter \( j \) was set to equal the ratio of number of training negative examples to number of training positive examples to simulate training on a balanced class dataset. The cost parameter \( C \) was selected by optimizing the precision-recall break-even point over coarse and subsequently fine grid searches ranging between 1e-3 and 1e6.

2.2 Structure prediction benchmark

Structure prediction performance of pGenTHREADER was tested using a set of 2873 consensus domains derived from the overlap of the ASTRAL 1.73 database (Chandonia et al., 2004) and the CATH 1.35 representative sequences. Sequences were submitted for prediction both as single domains and full PDB chains in order to determine the sensitivity of the methods to identification of domain boundaries (Supplementary Methods).

PSI-BLAST (Altschul et al., 1997) profiles (six iterations, profile inclusion score 0.001) and PSIPRED (Jones, 1999) predictions were generated for the sequences using the Uniref90 database (Suzek et al., 2007). Both query and database sequences were filtered for low-complexity regions using pIthD (Jones and Swindells, 2002). Predictions were run using a full-chain fold library filtered for non-redundancy using ASTRAL 1.73. The fold library contained 6694 chain sequences. pGenTHREADER models were generated using its own procedure, which generates models by transferring coordinates from aligned regions only with no loop modelling.

For comparison we ran PRC (Madura, 2006) and HHpred (Soding, 2005) using the sixth iteration PSI-BLAST profiles. In each case the default options supplied with the software were used to generate profiles for all 6694 chain representatives and these were then searched, again with default parameters. Additionally we used PSI-BLAST to search the sequence database corresponding to the 6694 sequences in the fold library. Models were generated for all alignments with \( e \)-value \( \leq 10 \) for each method using MODELLER 7v7 (Sali and Blundell, 1993). Models were assessed using an in-house implementation of the MAXSUB score (Siew et al., 2000) with an equivalence threshold of 2.0 Å.
To assess domain detection power of pGenTHREADER we compared it with PRC, HHpred and PSI-BLAST using 2873 domain sequences which overlap between the CATH3.1 S35 representative set and the ASTRAL 1.73 40% non-redundant set and have identical structural definitions. A further set of 717 full chains containing multiple domains were also assessed separately (Supplementary Material). Sequences were scanned against a fold library containing 6694 sequences (30% non-redundant at the chain level). We compared models built from alignments generated by the four methods and assessed template selection performance and alignment accuracy since these are the main determinants of template-based structure prediction success. Methods are compared on a top-hit basis and the difficulty of the structure prediction task is varied by including or excluding members of the same SCOP superfamily.

### 3 RESULTS

#### 2.3 Domain detection benchmark

pDomTHREADER domain recognition was benchmarked using 4008 full-chain queries derived from the entire CATH 3.1 S35 representative set. Full-chain queries were used to make the analysis more realistic since in the real case the domain boundaries would not be known. Alignments were generated to a domain-based fold library derived from the CATH S60 representative set using the standard pGenTHREADER protocol. Confidence scores were assigned to matches using the Platt algorithm (Platt, 1999) to model posterior probabilities from the SVM scores.

For HHpred, PRC, PSI-BLAST and pDomTHREADER algorithms, the third iteration PSI-BLAST profiles (appropriately converted to checkpoint, a3m or matrix files using default parameters) for each query sequence were scanned against the domain based S35 superfamily library. For pGenTHREADER, the threading library was constructed from whole chain PDB entries as opposed to domain delineated entries. The maximum e-value threshold for HHpred, PRC and PSI-BLAST results was 10, and for pDomTHREADER and pGenTHREADER algorithms, score thresholds of 0 were used. Positive matches were assigned to the true class if the CATH classifications were identical to the H level; otherwise they were considered false. Unclassified chain regions and curated SAS exceptions (those that obtained a high Structural Alignment Score, Reid et al., 2007) were considered ambiguous and omitted from the benchmark.

#### 2.4 Domain boundary benchmark

To assess domain boundary predictions, the from-to residues for each CATH S35 superfamily numbered from residue 1 in the corresponding PDB chain sequence were used. For discontinuous domains, the longest fragment only was considered. For each method, the predicted boundaries of correct hits only were compared to actual boundaries and the absolute residue deviations recorded.

Fig. 1 charts the mean proportion of equivalent residues predicted within length ranges of 50 amino acids. The mean percentage of equivalent residues at 2 Å for the top selected template in each length bin is shown for each method. Error bars indicate standard errors. Methods are annotated as best (highest scoring amongst all methods), pGT (pGenTHREADER), HHP (HHpred), PRC and PH (PSI-BLAST). The X-axis shows length ranges for each bin with the number of sequences in that bin in brackets.

Fig. 2 shows ROC-style plots of the top hits reported for the four methods. Template quality is defined by the highest score achieved for any method with a given target-template pairing. Correct structural relationships are defined as having >30 equivalent residues. (a) and (b) detail top-hit performances jack knifed at the superfamily and fold levels respectively.

3.1 Detection of structural relationships

Figure 1 charts the mean proportion of equivalent residues predicted (2 Å) for each of the four methods. Additionally we show the best achievable performance using an “ideal method”, which chooses the best result for a given target found by any method. Data are binned according to sequence length. In order to distinguish template selection from alignment accuracy the number of equivalent residues for a template was calculated from the best result for that template-target pair rather than the result for that method.

pGenTHREADER tends to successfully predict more residues for each target across all length ranges. PSI-BLAST performs worse than the other methods particularly over short length ranges, however performs better for longer sequences. PRC and HHP both perform equally well across the range of lengths with HHP doing significantly better for shorter templates of length 50 or less. Selecting the ideal (best result amongst all methods) greatly increased the performance at length ranges of less than 200 amino acids, however only yielded a slight improvement for longer >200 templates.

Figure 2 shows ROC-style plots of the top hits reported for the four methods. We plot absolute numbers of true and false positives instead of the usual rate measures since the different methods recovered different numbers of hits. The definition of a true structural relationship is that at least one method generated a model with the template that had at least 30 equivalent residues at 2 Å.

Clearly the scores produced by PRC and HHpred are more discriminating than pGenTHREADER’s when close relationships are considered. However where detection of distant relationships...
Comparison of alignment accuracies

Table 1. Comparison of alignment accuracies

<table>
<thead>
<tr>
<th>Method</th>
<th>pGT</th>
<th>HHP</th>
<th>PRC</th>
<th>PSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>pGT</td>
<td>X</td>
<td>-73</td>
<td>-33</td>
<td>-73</td>
</tr>
<tr>
<td>HHP</td>
<td>5.5e+07 (-)</td>
<td>X</td>
<td>-55</td>
<td>-10</td>
</tr>
<tr>
<td>PRC</td>
<td>3.9e+07 (-)</td>
<td>4.8e+07 (+)</td>
<td>X</td>
<td>-50</td>
</tr>
<tr>
<td>PSI</td>
<td>5.3e+07 (-)</td>
<td>3.4e+07 (+)</td>
<td>4.4e+07 (+)</td>
<td>X</td>
</tr>
</tbody>
</table>

Values above the diagonal are Z-scores obtained from the normal approximation to the Wilcoxon probability (Sheskin, 1998). Values below the diagonal are the larger of the summed rank values with signs in brackets. Comparisons are reported as the row value—the column value. A negative sign indicates the method in the column produces higher scores than row method. Methods are annotated pGT, HHP, PRC, PSI for pGenTHREADER, HHpred, PRC, and PSI-BLAST, respectively.

is concerned pGenTHREADER is superior to the other methods. Figures 1 and 2 suggest that the overall performance improvement of pGenTHREADER over the other methods is in the detection of very distant relationships (fold recognition) and in improving alignments, but that at moderate distances it is less able to reliably select the best template. Template selection for full-chain multi-domain targets was noticeably poorer for pGenTHREADER although it remains the most sensitive method at greater evolutionary distances (Supplementary Fig. 1).

This seems to be a question of how the probabilities are calculated: in passing we note that the Pearson correlation of pGenTHREADER output scores with the actual number of equivalent residues was 0.95, demonstrating that the raw scores were a highly accurate indication of model quality.

3.2 Alignment accuracy

We assessed alignment accuracy directly by comparing the number of equivalent residues at 2 Å found by each method on a given target-template pairing for the set of template–target pairs identified by all four methods, producing a dataset of 11 364 target-template pairs in total. A Friedman test for the entire set showed a significant difference from 0 (Fisher 2005). Values below the diagonal are the larger of the summed rank values with signs in brackets. Comparisons are reported as the row value—the column value. A negative sign indicates the method in the column produces higher scores than row method. Methods are annotated pGT, HHP, PRC, PSI for pGenTHREADER, HHpred, PRC, and PSI-BLAST, respectively.

Fig. 3. Domain Superfamily detection Receiver Operating Characteristic (ROC) like curves. The performance of pGenTHREADER (mGT), PRC (PRC), HHpred (HHP) and PSI-BLAST (PB) are plotted for all Hits using actual true positives (x-axis) against actual false positives (y-axis). The performance for each method was jack-knifed by repeated assessments leaving out an entire superfamily at a time and averaged to produce final statistics.

3.4 Domain superfamily detection

We compared the accuracy of the pDT scores and profile–profile scores in recognizing true superfamily members by reporting actual true and false positives obtained for the CATH S35 dataset. Receiver Operating Characteristic (ROC) like curves were plotted using all reported hits from each method (Fig. 3) for a threshold of 1000 false positives.

Overall pDT and PRC recovered most true positive relationships, almost double the amount obtained by HHpred and pGT. At fewer than 147 false positives, pGT detected more true positive relationships than the other methods, however at all subsequent numbers of false positives, PRC outperformed all other methods reflecting its superior calibration of e-values across the range of superfamilies. PSI-BLAST represented the middle ground, outperforming HHpred and pGenTHREADER at >300 FP but discriminating fewer positive relationships at more significant e-values (close to 0).

Each method recovered different numbers of predictions at the respective score cut-offs (Table 2) reflecting their ability to generate alignments and power to discriminate true relationships. Notably fewer hits were returned at e-values of equivalent magnitude using HHpred than the other algorithms, perhaps a function of the different HMM profile calibration steps.

The poorer performance of the pGenTHREADER algorithm is likely due to the use of whole PDB chains for the threading template library as oppose to domain delineated templates. For multi-domain proteins, PSI-BLAST profiles can be biased by inclusion of many sequence relatives that possess just one of the domains. False positive assignments resulted from incorrect whole chain alignments due to over-extension of a partially correct alignment between equivalent single domains from a multi-domain sequence. However, a significant advantage of using whole chain templates is evident in fold recognition where maintenance of an up to date template library is a key determinant of the accuracy of the method. Despite recent improvements in both CATH and SCOP databases there inevitably
An important challenge to the scientific community is to rapidly recognize and superfamily annotations can be rapidly produced hours (McGuffin et al., 2006). Consequently high throughput fold recognition and superfamily annotations can be rapidly produced.

### Table 2. Discriminatory power of the scores in superfamily detection.

<table>
<thead>
<tr>
<th>Method</th>
<th>Number positives</th>
<th>1000 FP</th>
<th>100 FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>pDT</td>
<td>60431</td>
<td>23128</td>
<td>11935</td>
</tr>
<tr>
<td>PRC</td>
<td>49142</td>
<td>30421</td>
<td>10300</td>
</tr>
<tr>
<td>HHP</td>
<td>29220</td>
<td>18376</td>
<td>9917</td>
</tr>
<tr>
<td>PSI</td>
<td>32039</td>
<td>24389</td>
<td>7393</td>
</tr>
<tr>
<td>pGT</td>
<td>29493</td>
<td>14318</td>
<td>8232</td>
</tr>
</tbody>
</table>

Actual true positives are reported at 1000 false positives and 100 false positives to represent performance at low and very low error rates. Performance estimates have been jack-knifed at the superfamily level and represent averages rounded to the nearest integer. Methods are annotated pGT, HHP, PRC, PSI for pGenTHREADER, HHPred, PRC and PSI-BLAST, respectively.

### Table 3. Performance statistics using top hit annotations

<table>
<thead>
<tr>
<th>Method</th>
<th>Positives</th>
<th>Coverage at 0.05 EPQ</th>
<th>Coverage at 0.01EPQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>pDT</td>
<td>3374</td>
<td>0.956</td>
<td>0.821</td>
</tr>
<tr>
<td>PRC</td>
<td>3284</td>
<td>0.960</td>
<td>0.472</td>
</tr>
<tr>
<td>HHP</td>
<td>2871</td>
<td>0.918</td>
<td>0.196</td>
</tr>
<tr>
<td>PSI</td>
<td>3149</td>
<td>0.953</td>
<td>0.293</td>
</tr>
<tr>
<td>pGT</td>
<td>3052</td>
<td>0.715</td>
<td>0.101</td>
</tr>
</tbody>
</table>

The last two columns represent coverage obtained at different error per query rates. For 4008 chains there were 2872 representatives with ≥1 nr35 representative in the CATH S35 library covering a total of 1295 homologous superfamilies. The maximum number of positive domain matches that could be obtained was 3541.

exists a significant lag period between the release of a PDB structure and its classification into domains.

Consistent with other reports (Madera, 2002; Muller, 1999; Reid et al., 2007) PSI-BLAST performed well despite the fact that the e-value scores had not undergone a calibration step in the same way as the HMM-HMM profile comparison methods.

### 3.5 Performance in whole genome annotation

An important challenge to the scientific community is to rapidly and computationally characterize newly predicted sequences arising from genome projects with structural or functional information. As standard practice in these annotation methods, often only the best match by score is considered for a query sequence (Reid et al., 2007). To reflect this practice, in this assessment, top non-overlapping hits were considered for each method over the respective regions of each query chain. Similar numbers of top hits were reported for each of the different methods (Table 3). The pDomTHREADER algorithm outperformed all other methods at very low error rates achieving more than 82% coverage at 0.01 EPQ (Fig. 4).

This result suggests room for improvement in better discriminating true positive matches at the low end of the scale for pDomTHREADER implementation without compromising its ability to accurately rank high scoring pairs. PRC and PSI-BLAST both outperformed HHPred and pGenTHREADER which suffered greater numbers of false positives at the lower end of the score scale. This suggests these approaches might be better suited to recognizing folds rather than discriminating homologous superfamily matches.

#### 3.6 High scoring cross architecture matches

This study highlighted several high scoring matches between superfamilies of different architectures in addition to those that were not part of the SAS8 curated exceptions defined in Reid et al., 2007 (Supplementary Table S4). Some of these links occur between short sequence/template matches. Others contain regions of low sequence complexity or are incorrect alignments over discontinuous domains. Longer matches obtained by multiple methods might represent genuine evolutionary links; a common structural core or secondary structure motif between architectures that warrant further investigation as part of a future study.

#### 3.7 Domain boundary predictions

pDomTHREADER obtained the most accurate residue boundary predictions of all methods (Table 4); on average the predicted residues deviated by just seven residues in the nr35 benchmark set. The ordering of methods by accuracy for boundary predictions is pDT <= PRC <= HHP <= PSI <= pGT. All reported P-values for each test were highly significant (P < 0.001) except for the comparison between PRC and HHP. Again the performance of pGT is attributed to the use of whole chain templates. During profile building, the most conserved portions of the query sequence receive high coverage providing anchor points for accurate alignments. If unrelated sequences are incorporated into the profile, sharp distinctions between conserved and variable parts become blurred affecting the accuracy of predicted domain boundaries. The pGenTHREADER protocol is susceptible to these effects during both template and query profile construction. Consequently the average residue deviation from actual boundaries was more than double those obtained for pDomTHREADER.

With increasing availability of grid computing services and affordable commodity computing equipment, it is now practical to generate sequence profiles for entire proteomes in a matter of days (McGuffin et al., 2006). Consequently high throughput fold recognition and superfamily annotations can be rapidly produced.
Table 4. Domain boundary performance

<table>
<thead>
<tr>
<th>Method</th>
<th>pDT</th>
<th>PRC</th>
<th>HHP</th>
<th>PSI</th>
<th>pGT</th>
</tr>
</thead>
<tbody>
<tr>
<td>pDT</td>
<td>7.06</td>
<td>2.02</td>
<td>2.51</td>
<td>-3.70</td>
<td>-7.44</td>
</tr>
<tr>
<td>PRC</td>
<td>4.07e-51</td>
<td>9.08</td>
<td>-0.49</td>
<td>-1.37</td>
<td>-6.12</td>
</tr>
<tr>
<td>HHP</td>
<td>2.48e-53</td>
<td>1.35e-01</td>
<td>9.57</td>
<td>-1.19</td>
<td>-5.63</td>
</tr>
<tr>
<td>PSI</td>
<td>2.10e-86</td>
<td>3.03e-07</td>
<td>4.72e-04</td>
<td>10.76</td>
<td>-4.44</td>
</tr>
<tr>
<td>pGT</td>
<td>2.23e-111</td>
<td>1.47e-33</td>
<td>6.28e-21</td>
<td>1.87e-15</td>
<td>15.20</td>
</tr>
</tbody>
</table>

The diagonal values represent the average residue deviations obtained by the method compared to actual boundaries. Values above the diagonal represent the difference between average residue deviations between the row method and column method. Values below the diagonal are the significance P-values from a one-tailed Wilcoxon unpaired signed rank test of residue deviations from the row and column methods. Methods are annotated pDT, PRC, HHP, PSI and pGT for pDomTHREADER, PRC, HHPred, PSI-BLAST and pGenTHREADER, respectively.

for large volumes of sequence data using the pGenTHREADER and pDomTHREADER. For accurate superfamily annotations and domain boundaries, we recommend pDomTHREADER over other profile–profile matching algorithms. For distant fold recognition and high accuracy modeling alignments, we recommend pGenTHREADER. Where CPU resources are limited, the power of both threading approaches can be leveraged where other methods fail.

4 DISCUSSION

The problem of mapping the 3D structures of proteins to genome sequences is a key challenge of the post-genomic era. The structure of a protein sequence can be used to infer evolutionary relationships not detectable at the amino-acid level. These relationships frequently suggest common function and provide template information for the construction of high quality structural models. For functional inference it is important to determine specific structural domain based assignments since function is most strongly related to domain architecture (Bashkin and Chothia, 2003; Heyyg and Gerstein 2001). However, for powerful detection of distant sequence-structure relationships, and subsequent homology modeling, correct recognition of the fold of the protein presents a more realistic goal (Jaroszewski et al., 2002).

Our results suggest that the two separate application areas provide a practically useful distinction, in agreement with related studies (Cheng et al., 2008; Reid et al., 2007). The methods also encapsulate differences between aspects of structure determination which are important for structure prediction purposes (since this permits improved template selection; Sadowski and Jones, 2007) yet irrelevant for discriminating structural homologies.

Generating separate related methods permits other technical details to be addressed: the fold libraries used in each case derived from different underlying sequence sets: full chains for structure prediction as opposed to domains for superfamily detection. This allows structure prediction to be as comprehensive as possible given the limitations of current techniques for assembling multiple domain predictions from independent single-domain predictions whilst ensuring the accuracy of domain boundary predictions for generating superfamily assignments and domain architectures.

Local alignment methods are necessary to detect distant homologies since evolutionary modules are preserved on any scale, provided that they are useful. Practically, this carries the risk of identifying very short high scoring matches that generate false-positives. The use of structural data to assess the likelihood of homology between two proteins provides more information than sequences alone (Swanson et al., 2009). However substructures recur at above the level expected by chance, which may be attributable to physical properties of proteins, or to the generation of functional diversity through re-use of structural scaffolds (Redfern et al., 2008). Unfortunately this generates problems for distant homology recognition by invalidating the underlying assumptions of the implied evolutionary models.

We have described two separate fold recognition based solutions which meet the challenges of both sensitive fold recognition (pGenTHREADER) and domain superfamily discrimination (pDomTHREADER) that can be applied to whole proteomes, both of which outperform sequence profile based methods. Future improvements in these areas may arise through inclusion of other informative features, exploring other methods to optimally combine features or implementing new machine-learning methodologies. However, it is most likely that the continued growth of sequence and structural databases will be the greatest source of improvements as more intermediates generate links between, what are at present, isolated groups of proteins.

5 AVAILABILITY

The pGenTHREADER and pDomTHREADER algorithms can be freely downloaded for academic use from http://bioinf.cs.ucl.ac.uk/downloads/pGenTHREADER and have been incorporated into the suite of servers at UCL for proteome annotation and structure prediction (http://bioinf.cs.ucl.ac.uk/psipred/psiform.html).

ACKNOWLEDGEMENTS

The authors would like to thank Dr Ollie Redfern for providing the CORA structural alignments for the CATH S35 representatives, and Mr Tony Lewis for providing CATH domain boundaries.

Funding: Biosapiens Network of Excellence, funded by the European Commission within its FP6 Programme, under the thematic area Life sciences, Genomics and Biotechnology for Health, contract number LSHG-CT-2003-503265 (MIS, DTI) and by a BBSCc case studentship in collaboration with BioFocus DPI (AL).

Conflict of Interest: none declared.

REFERENCES


Ried, A.J. et al. (2007) Methods of remote homology detection can be combined to increase coverage by 10% in the twilight zone. Bioinformatics, 23, 2353–2360.


