FIEFDom: a transparent domain boundary recognition system using a fuzzy mean operator

Rajkumar Bondugula¹*, Michael S. Lee¹,²,³ and Anders Wallqvist¹

¹Biotechnology HPC Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Materiel Command, Fort Detrick, MD 21702, ²Computational and Information Sciences Directorate, U.S. Army Research Laboratory, Aberdeen Proving Ground, MD 21005 and ³Department of Cell Biology and Biochemistry, U.S. Army Medical Research Institute of Infectious Diseases, Fort Detrick, MD 21702, USA

Received August 1, 2008; Revised November 5, 2008; Accepted November 7, 2008

ABSTRACT

Protein domain prediction is often the preliminary step in both experimental and computational protein research. Here we present a new method to predict the domain boundaries of a multidomain protein from its amino acid sequence using a fuzzy mean operator. Using the nr-sequence database together with a reference protein set (RPS) containing known domain boundaries, the operator is used to assign a likelihood value for each residue of the query sequence as belonging to a domain boundary. This procedure robustly identifies contiguous boundary regions. For a dataset with a maximum sequence identity of 30%, the average domain prediction accuracy of our method is 97% for one domain proteins and 58% for multidomain proteins. The presented model is capable of using new sequence/structure information without re-parameterization after each RPS update. When tested on a current database using a four year old RPS and on a database that contains different domain definitions than those used to train the models, our method consistently yielded the same accuracy while two other published methods did not. A comparison with other domain prediction methods used in the CASP7 competition indicates that our method performs better than existing sequence-based methods.

INTRODUCTION

The 3D structure of a protein holds the key to understanding the detailed function of a protein at the molecular level. However, the cost and time required for experimental structural characterization of larger (genomic) protein sets can be prohibitive, creating a need for developing accurate computational structure prediction approaches (1–3). Proteins can be considered to be built up from domains, where each domain can be thought of as a structural unit of a protein that is compact, local and constitutes a semi-independent unit capable of folding independently (4,5). Delineation of proteins into domains is often the first step in both experimental and computational protein research (6–9). Longhi and co-workers (10) suggest dividing large proteins into domains to increase the yield of protein crystals suitable for X-ray diffraction as large proteins are difficult to crystallize (11,12). Since the initial X-ray structure determinations of proteins were carried out for smaller, one domain proteins, the field of protein structure predictions was focused on one domain proteins. Thus, as a legacy, programs for protein structure prediction are still typically optimized for predicting structures for shorter one domain sequences. Moreover, a majority of eukaryotic proteins are multidomain proteins (13) and predicting the structure of long proteins continues to be a challenge (14). Copley and co-workers (15) present compelling arguments about analyzing genomes at the domain level rather than protein level. Also, reliable identification of domains influences the quality of multiple sequence alignments (16,17). Furthermore, the knowledge of domains is necessary for designing new chimeric proteins (18). Given the above listed applications, protein domain prediction continues to be an important area of research with broad utilities in protein science.

Most of the current approaches for protein domain boundary prediction can be classified into three broad categories (19): domain homology prediction, domain recognition and new domain prediction. Domain homology prediction methods take advantage of the close homology to known domain sequences. In this approach, databases, such as CATH (20), SCOP (21), Pfam (22), CDD (23) or SMART (24), are searched for a close match with the query sequence, and domains are assigned based on sequence similarities. Domain homology prediction is very efficient, provided homologs exist, e.g. the

*To whom correspondence should be addressed. Tel: +1 301 619 1990; Fax: +1 301 619 1983; Email: raj@bioanalysis.org

© Published by Oxford University Press (2008)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
**FIEFDom: a transparent domain boundary recognition system using a fuzzy mean operator**

**Biotechnology HPC Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Materiel Command, Fort Detrick, MD, 21702**

Approved for public release; distribution unlimited

16. **SECURITY CLASSIFICATION OF:**
   - a. report: unclassified
   - b. abstract: unclassified
   - c. this page: unclassified

17. **LIMITATION OF ABSTRACT:** Same as Report (SAR)

18. **NUMBER OF PAGES:** 11

19a. **NAME OF RESPONSIBLE PERSON:**
prediction method CHOP (25) uses this technique. In domain recognition methods, the database of proteins with known structures is searched for sequences that exhibit remote homology with the query sequence (26). In this approach, the remote homologs can be identified using sequence-based methods like PSI-BLAST (27) or using auxiliary information such as the predicted secondary structure (28,29). Both domain homology prediction and domain recognition methods rely on multiple sequence alignments (MSA). These methods, especially those based on artificial neural networks (NNs) (30–32), would still be unreliable for truly novel sequences, i.e. those that do not have detectable homology with protein sequences with known structures. For novel sequences, new domain prediction methods, which only use the amino acid sequence as the input, are often appropriate. Some examples in this direction include SnapDRAGON (33), RosettaDOM (34), DomCut (18) and Armadillo (35). The first two programs infer domains by initially predicting a coarse-grained tertiary structure that can be used to delineate domain boundaries. This methodology often gives good results, but typically requires significant computational resources. Other methods rely on machine learning or statistical models trained on biochemical properties of the amino acids, averaged over a window of the query protein. While these methods are fast and independent of homologs in the databases, they are rarely used because of their limited accuracies. Hybrid methods that combine several sources of information have been proposed in the past, but the performance gains have been modest. For example, in Biozoon (31), the features derived from MSA, physiochemical properties of amino acids, secondary structures, exon boundary information, etc., are integrated using NNs. KemaDom (36) is another hybrid method that uses predicted secondary structure, predicted solvent accessibility, amino acid entropy and physiochemical properties of amino acids as input to an ‘ensemble’ of three support vector machines.

We propose a different method, which we call FIEFDom (Fuzzy Integration of Extracted Fragments for Domains), for predicting the domain boundaries of proteins from a given sequence and its sequence profile (a 2D matrix that represents the likelihood of each amino acid occurring at every position along the protein sequence) using a fuzzy mean operator (FMO). A FMO represents a special case of the fuzzy nearest neighbor algorithm (37), with the number of classes set to one. The choice of FMO was motivated by its simplicity, transparency, ease of updating the method and more abstractly for its asymptotic error bounds. FIEFDom is transparent, i.e. the choice of the program to designate a region as a domain boundary can be traced back to all proteins in the local database that contributed to the decision, offering additional insight. Also, our model need not be trained or tuned whenever new examples of domain boundaries become available. The sequences of newly determined boundaries can just be appended to the reference database file. In addition, the users can choose the domain definitions (e.g. CATH or SCOP) to suit their needs, just by replacing the reference protein set (RPS). As the available data approaches infinity, the upper bound of the maximum error rate is at most twice the optimal Bayes’ error rate (38). We show that our procedure works well for a wide range of proteins: from ones with many close homologs to ones with only remote homologs. We illustrate the effects of redundancy and the number of reference proteins in the database on the accuracy of our method. We compare the performance of our method with two other methods, PPRODO (32) and DOMpro (30), adjusting our reference database as necessary to ensure impartial comparisons of the underlying algorithms. Finally, we compare the performance of our method with six sequence-based domain prediction methods that participated in CASP7 (39), both in domain number prediction accuracy and domain position prediction accuracy. An executable of the FIEFDom software is freely available for download at http://www.bhsai.org/downloads/fiefdom.

**METHODS AND MATERIALS**

**Databases**

SCOP is a manually curated database that contains structural domains defined by Alexei Murzin and his colleagues. This database is generally accepted as a standard for protein structure classification (40). For analysis of various aspects of FIEFDom, we use the following ASTRAL SCOP (41) databases: SCOP 1.65 (30%) (i.e. the ASTRAL SCOP version 1.65 database containing domain sequences with 30% maximum sequence identity), SCOP 1.69 (20%), SCOP 1.69 (30%), SCOP 1.69 (40%), SCOP 1.73 (30%) and SCOP 1.73 (95%). Table 1 shows the domain compositions of the above

<table>
<thead>
<tr>
<th>Number of domains</th>
<th>SCOP database version (maximum percentage sequence identity)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.65 (30%)</td>
</tr>
<tr>
<td>One</td>
<td>3145</td>
</tr>
<tr>
<td>Two</td>
<td>533</td>
</tr>
<tr>
<td>Three</td>
<td>107</td>
</tr>
<tr>
<td>Four</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>3805</td>
</tr>
</tbody>
</table>

Data in the first row indicate the number of one-domain proteins in each database. The second row contains the number of two-domain proteins, etc. The last row indicates the total number of proteins included in each database.

**Table 1. Domain composition of proteins contained in the SCOP databases used in this work**
databases. Since ASTRAL SCOP databases contain sequences of individual domains, we concatenate domain sequences from the same protein chain to reconstruct the original multidomain proteins. Due to the relative scarcity of proteins with more than four domains in the SCOP database, we only consider proteins that contain up to four domains in this study. Each of these databases, with domain and domain boundary residues labeled, constitutes a RPS. We choose every other version of the SCOP database for analysis to provide a larger increment in the number of newly observed domains as opposed to using consecutive versions. For multidomain proteins, 20 residues before and after the true domain boundary (as defined by SCOP) are designated as boundary residues. We use this widely used (19,28,30,32,33,36) labeling protocol to facilitate a fair comparison with other methods. The method developed is not strongly dependent on the number of boundary residues picked. Note that we do not address the issue of predicting domains with non-contiguous sequences and consequently we discard such proteins. We found that less than 7% of the domains in SCOP have non-contiguous sequences.

Procedure

We use a three-step procedure to predict domain boundaries. First, we generate the specific scoring matrix (PSSM, a profile generated by PSI-BLAST program) (27) of the query sequence using a large database of known sequences. Second, we use the generated profile to search for similar fragments in the RPS. Third, the matches with the proteins in RPS are parsed, and the domain boundary propensity \( P_{B} \) (the likelihood of an amino acid to be in domain boundary) of the query protein is predicted using a FMO. These steps are detailed below.

In the first step, the profile of the query sequence is calculated using the PSI-BLAST program and the non-redundant or nr (ftp://ftp.ncbi.nih.gov/blast/db) database (42). We generate the profile by running the PSI-BLAST program for three iterations. Default values are used for the remaining parameters. In the second step, we perform profile-sequence alignment between the query profile and the proteins in the RPS to search for matching fragments by running the PSI-BLAST program a second time. During this step, the expectation value threshold (e-value) is set at 10,000. This high threshold ensures that the alignments retrieved contain both large and small protein fragments. The parameters for this two-stage PSI-BLAST protocol were optimized in a previous work on secondary structure prediction (43). In the third step, the matching fragments found in the second step are parsed and scored using the following scoring scheme (43):

\[
S = \max\{1.7 + \log_{10}(e\text{-value})\}
\]

The score, \( S \), is formulated as a ‘dissimilarity’ measure. For instance, the fragments of proteins in the RPS that have high sequence similarity with the subsequences of the query protein have high statistical significance (or low e-value), and therefore have low scores. Finally, the domain boundaries (if any) are predicted using the scored fragments. For each residue, the \( P_{B} \) is calculated from the domain boundary memberships \( (B) \) of the residues in the fragments that are aligned with the current residue. The \( P_{B} \) of the query protein is calculated using the following expression for the FMO:

\[
P_{B}(r) = \frac{\sum_{j=1}^{K} B_{j}(r) (1/S_{j}^{2/(m-1)})}{\sum_{j=1}^{K} (1/S_{j}^{2/(m-1)})}
\]

where, \( r \) is the current residue identifier, \( K \) is the number of fragments that have a residue aligned with the current residue \( r \), \( B_{j}(r) \in \{0 \text{ if the residue lies in the domain and 1 if the residue lies on the domain boundary}\} \) is the domain boundary membership of the residue in the \( j \text{th} \) fragment that has a residue aligned with the current residue \( r \), \( S_{j} \) is the score for the \( j \text{th} \) fragment defined in Equation 1, and \( m \) is a fuzzifier (37) that controls the weight of the dissimilarity measure, \( S \). The value of \( m \) was set to 1.5 based on previous work on secondary structure predictions (43). The boundary prediction results are not very sensitive to this parameter (data not shown). The values of \( P_{B}(r) \) range from 0 to 1, whereas a value of 0 indicates that it is unlikely that \( r \) lies on a domain boundary, whereas a value of 1 indicates a strong likelihood that the residue is located in a boundary region. A typical alignment produced while searching for matching fragments in the RPS (step 3) is shown in Figure 1.

Figure 1. The fragments retrieved when the RPS is searched for matching fragments with a typical protein. The fragments shown are labeled using their SCOP definitions. Residues labeled ‘D’ lie in protein domains, whereas residues labeled ‘B’ lie on the domain boundary; ‘-’ is used to indicate that no residue in the current fragment is aligned with the query sequence. For the Alanine residue (A) in the shaded box, the domain boundary propensity is calculated using Equation 2 based on the five aligned residues \( (K = 5) \), four of which are found in non-boundary regions and one is found in a boundary region. The importance of these contributions is inversely weighted by their respective scores, \( S \), shown on the right, as detailed in Equation 2. In this case, the likelihood \( P_{B} \) that the alanine residue belongs to domain boundary is 0.0804.

Postprocessing

The values of \( P_{B} \) are smoothed by averaging over a window of length \( W \) (\( W = 5 \), in this work) around each amino acid position in the query sequence. In the termini, the average is based only on those residues that are
actually present in the window. The potential regions that contain domain boundaries are obtained by selecting those regions that have a $P_B$ value above a threshold value $T$, where $T$ was set to 0.4. The details and the statistical measures underlying this choice are given in the next subsections. Once the potential regions are identified, the area under each identified sequence segment is calculated. We use this area to represent the confidence in the predicted domain boundary. If two regions lie within 40 residues of each other, the region with lower confidence is removed from further consideration. Also, predicted domain boundaries that fall within 40 residues of either the COOH or NH$_2$ termini are discarded. The midpoint of each region is returned as the location of the domain boundary. As an example, the raw $P_B(r)$ output is illustrated for the *Escherichia coli* MurF protein [PDB: 1GG4, Chain A] in Figure 2. The predicted domain boundaries (residues 91 and 314) within two potential regions of interest are marked with dotted lines, agreeing very well with the actual boundaries centered on residues 98 and 313.

**Performance metrics**

The performance is assessed in terms of three metrics: accuracy, specificity and sensitivity (29,35,44). These metrics are defined as follows:

\[
\text{Accuracy} = \frac{TP}{TP + FP + FN}, \quad \text{Specificity} = \frac{TP}{TP + FP}, \quad \text{Sensitivity} = \frac{TP}{TP + FN}
\]

where $TP$ denotes true positives (domain boundaries correctly predicted as domain boundaries), $FP$ stands for false positives (regions incorrectly predicted as domain boundaries) and $FN$ stands for false negatives (missed domain boundaries). Here we assume that if the predicted domain boundary is within 20 residues designated as boundary residues, the prediction is a true positive. Our definition of accuracy is appropriate since the term ‘true negative’ (all non-domain boundaries correctly predicted as non-domain boundaries) is not a practical concept in the context of domain boundary prediction. Also, for one-domain proteins, the accuracy is defined as the fraction of proteins in which no domain boundary is predicted.

**Choice of threshold value, $T$**

In this subsection we investigate the effect of the threshold, $T$, above which the regions on the $P_B$ curve are designated as potential regions containing domain boundaries. The post-processing step for the domain boundary prediction procedure involves applying a threshold $T$ to filter the background noise and to designate potential regions that contain domain boundaries. We used SCOP 1.73 (30%) to study the effect of $T$ on the sensitivity, specificity and accuracy of the domain boundary prediction. We systematically varied the value of $T$ from 0 to 1 in increments of 0.1 and recorded the performance metrics as shown in Figure 3. We found that values of $T$ in the range between 0.0 and 0.3 strongly influenced sensitivity, specificity and accuracy. For larger values, these measures remained relatively constant or had a plateau-like behavior in the region $\sim 0.3$–0.5. Figure 3a illustrates the receiver operating characteristic (ROC) curve of the average multidomain predictions by varying $T$ while Figure 3b illustrates the influence of $T$ on the accuracy of one, two, three, four and all domain boundary predictions. Based on the plots in Figure 3, we fixed the value of $T$ at 0.4 for all further analysis.

**RESULTS**

In this section, we analyze the performance of our method with varying levels of sequence/structure information availability in an attempt to simulate practical, real-life conditions. First, we present the results of the program under various conditions of homologous sequence availability for building a profile. Second, we investigate how growth of the RPS database affects accuracy. Third, we increase the redundancy of protein sequences (structure availability of related sequences) in the RPS and study its effect on our system’s performance. We then compare the performance of our method with existing methods. We present results using a jack-knife procedure on the RPS, where each sequence in the RPS is used as a query protein, while the remaining proteins are used as the domain database for fragment searches.

**Availability of homologs**

In the *nr* database, some proteins have more homologs than others. The experiments described in this paragraph emulate various conditions under which homolog availability varies for the query protein using the SCOP 1.73
(30%) database. At one extreme, for query proteins that have many homologs in the nr database, the profile is rich in evolutionary information. Use of such profiles leads to more sensitive fragment searches in the RPS, resulting in higher prediction accuracy. The performance metrics when the query profile is used to indentify matching fragments are shown in Table 2 (first row, top section). On the other extreme, for proteins that do not have any homologs in nr, the profile returned is merely the scoring matrix [i.e. BLOSUM62 (45)] used in the alignment algorithm. A profile-sequence alignment in such a case is the same as a sequence-sequence alignment. To simulate the above scenario, for each protein, we perform sequence-sequence alignment using the query sequence directly (no profile is generated; only the second PSI-BLAST run is performed). The results are presented in Table 2 (second row, top section). These results help us draw the bottom line performance of our system, when the query sequences are truly novel and appear to have no known homologs. We can also infer that our system does not completely fail under these conditions; it only performs with reduced accuracy. The average accuracies on the SCOP 1.73 (30%) database using profile-sequence alignments for finding matching fragments in the RPS for one domain proteins and

![Graph](image)

**Figure 3.** The effect of threshold on the performance of FIEFDom for the SCOP 1.73 (30%) dataset. (a) Receiver operating characteristic (ROC) curve averaged over all of the domain sets is plotted as the threshold (T) is varied from 0 to 1 in intervals of 0.1. (b) One-domain (blue solid line), two-domain (pink dashed line), three-domain (black dotted line), four-domain (red dashed-dotted line) and the average domain boundary prediction accuracy are plotted as a function of the threshold value, T. Based on the maximum and slow variability of the accuracy values over a range of T values, we selected T = 0.4 as the appropriate value to be used in our model.

**Table 2.** Studying the effect of homolog availability for building profiles, the number of proteins in the RPS and the effect of maximum sequence identity among the sequences in the RPS on the performance of FIEFDom

<table>
<thead>
<tr>
<th>Database</th>
<th>Alignment</th>
<th>Number of domains</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>One</td>
<td>Two</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>Sp</td>
</tr>
<tr>
<td>Homolog availability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCOP 1.73 (30%)</td>
<td>PS</td>
<td>97</td>
</tr>
<tr>
<td>SCOP 1.73 (30%)</td>
<td>SS</td>
<td>99</td>
</tr>
<tr>
<td>Number of proteins in RPS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCOP 1.65 (30%)</td>
<td>PS</td>
<td>97</td>
</tr>
<tr>
<td>SCOP 1.69 (30%)</td>
<td>PS</td>
<td>97</td>
</tr>
<tr>
<td>SCOP 1.73 (30%)</td>
<td>PS</td>
<td>97</td>
</tr>
<tr>
<td>Maximum sequence identity in RPS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCOP 1.69 (20%)</td>
<td>PS</td>
<td>97</td>
</tr>
<tr>
<td>SCOP 1.69 (30%)</td>
<td>PS</td>
<td>97</td>
</tr>
<tr>
<td>SCOP 1.69 (40%)</td>
<td>PS</td>
<td>97</td>
</tr>
</tbody>
</table>

A: accuracy; Sp: specificity; Sn: sensitivity. Alignment: PS profile-sequence, SS- sequence-sequence alignment. All values are percentages. Top: The availability of homology information for query sequences is simulated by using either the query profile (profile-sequence consistent with high availability) or the query sequence itself (sequence-sequence consistent with low availability) to search for identical fragments in the RPS. For multidomain proteins, the profile-sequence yields on average 13% higher overall accuracy, compared to the sequence-sequence alignment method. Middle: Every other version of the SCOP database, with 30% maximum sequence identity among the proteins, is used to study the effect of number of proteins in the RPS. The larger the size of the RPS (see Table 1 for the detailed breakdown in number of proteins and domain compositions), the higher is the average domain boundary prediction accuracy for multidomain proteins, presumably because the additional structure/sequence information uncovered as additional novel structures are added to the database. Bottom: Three simulations were conducted by experimenting with databases of three different maximum sequence identities among the reference proteins. The maximum sequence identity among the reference proteins varies from 20% to 40%.
multidomain proteins are 97% and 58%, respectively while the average (specificity, sensitivity) for multidomain proteins is (91%, 61%). For sequence-sequence alignments, one and multidomain protein accuracies are 99% and 45%, respectively, and the average (specificity, sensitivity) for multidomain proteins is (92%, 48%). Note that, although average specificities of the two methods are comparable, the sensitivities of the method that uses profiles is significantly higher, reiterating the importance of evolutionary information (in the form of profiles) while searching for fragments. While the results clearly demonstrate the advantage of using a profile to aid the fragment search, they also indicate that the absence of profile, on average, reduces the multidomain accuracy of our method by 13%.

Variability of the RPS database

We now turn to the performance of our method as new information is added to the RPS in the form of new protein sequences (for example, from newly sequenced genomes). We run our method on every other version of SCOP at the same sequence identity level, i.e. on SCOP 1.65 (30%), SCOP 1.69 (30%) and SCOP 1.73 (30%) databases. The same program is used to generate the alignments, parse the matches and calculate the PB curves. The only difference among the three experiments is the text file containing different RPSs, emphasizing the feature that updating the program amounts to merely appending (or replacing) the RPS text file. This advantage is unique to our approach due to the FMO-based model. The performance metrics of FIEFDom on various data-sets for one and multidomain proteins are presented in Table 2 (middle section). The averages (specificity, sensitivity) for SCOP 1.65 (30%), SCOP 1.69 (30%) and SCOP 1.73 (30%) are (92%, 52%), (91%, 55%) and (87%, 63%), respectively. Note that as we move from an older database [SCOP 1.65 (30%)] to a newer database [SCOP 1.73 (30%)], the average specificity decreases while the average sensitivity increases. Concomitant with this trend, the average multidomain prediction accuracies increase from 50% for SCOP 1.65 to 58% for the SCOP 1.73 database, while the accuracy for one domain prediction remains at 97%. Quantitatively, we observed that, for every 1000 new protein sequences added to the RPS (while maintaining maximum sequence identity level), the overall accuracy (one domain and multidomain) increases roughly by 2.3%. Figure 4a shows one, two, three, four and average domain prediction accuracies plotted as a function of the database version. It is clear from Table 2 (middle section) and Figure 4a that, as time progresses, i.e. as additional sequence/structure information becomes available, the accuracy of FIEFDom increases due to availability of novel sequences that can be added to the RPS, without the need for retraining the model per se.

The effect of protein sequence redundancy

Next, we study the dependency of the domain boundary likelihood, $P_B$, on the redundancy of protein sequence information. This redundancy can be modeled by using RPSs of the same ASTRAL SCOP version, but with different sequence identity thresholds. Raising the maximum sequence identity among the sequences increases the number of available sequences in the RPS, thereby improving the chances of finding fragments in the RPS that are similar to the subsequences of the query sequence. We also simulate a real-life scenario where the RPS contains the sequences of all SCOP family members, but not the sequences that belong to same family as the query sequence. In this experiment, we run the jack-knife procedure with SCOP 1.69 (20%), SCOP 1.69 (30%) and SCOP 1.69 (40%). We did not experiment further with higher-identity thresholds for three reasons: higher thresholds might lead to bias in favor of highly sequenced protein families, 40% sequence identity is the lower limit after which comparative modeling for protein structure prediction becomes reliable (46), and the jack-knife procedure may not be objective beyond this threshold.
The first comparison is aimed at understanding how the programs trained on SCOP domain definitions perform on proteins whose domain definitions are derived from the CATH database (20). PPRODO is an NN-based domain prediction system in which the profile extracted by the PSI-BLAST program is used as input to NNs for domain boundary prediction. A continuous signal is generated as output by the system, and the authors suggest a threshold of 0.25 above which an amino acid is designated as a domain boundary residue. DOMpro combines information from profiles, predicted secondary structures, and predicted relative solvent accessibility using recursive NNs. PPRODO was trained on two-domain proteins derived from SCOP 1.65 (released August 2003), and DOMpro was trained on the multidomain proteins in the CATH database version 2.5.1 (released January 2004). To make a fair comparison of different methodologies, we use FIEFDom with a RPS derived from the SCOP 1.65 (30%) (released August 2003) database.

In the first comparison, we use the SCOP 1.73 (30%) (released September 2007) database as a test set, which was released about four years later than their respective training databases (PPRODO and DOMpro) or RPS (FIEFDom). Table 3 summarizes the performance characteristics of the three systems. The average multidomain prediction accuracy of FIEFDom on the SCOP 1.73 (30%) database is 80%, while the one domain prediction accuracy is 97%. The average multidomain accuracies of PPRODO and DOMpro are 36% and 13%, respectively. Their one domain accuracies are 56 and 80%, respectively. While testing PPRODO, we extracted the raw signal from the PPRODO output file and applied the cutoff suggested by the authors. One might argue that PPRODO used only two-domain proteins for training, and DOMpro used only multidomain proteins for training; hence, it is not fair to compare the results directly. To resolve these

Table 3. The performance metrics of the three programs on a dataset that is about four years further in time from the training or reference data

<table>
<thead>
<tr>
<th>Method</th>
<th>One A</th>
<th>Two Sp</th>
<th>Two Sn</th>
<th>Three A</th>
<th>Three Sp</th>
<th>Three Sn</th>
<th>Four A</th>
<th>Four Sp</th>
<th>Four Sn</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIEFDom</td>
<td>97</td>
<td>93</td>
<td>77</td>
<td>73</td>
<td>96</td>
<td>85</td>
<td>82</td>
<td>94</td>
<td>88</td>
</tr>
<tr>
<td>PPRODO</td>
<td>56</td>
<td>53</td>
<td>54</td>
<td>37</td>
<td>50</td>
<td>38</td>
<td>26</td>
<td>78</td>
<td>51</td>
</tr>
<tr>
<td>DOMpro</td>
<td>80</td>
<td>32</td>
<td>12</td>
<td>10</td>
<td>34</td>
<td>14</td>
<td>11</td>
<td>55</td>
<td>23</td>
</tr>
<tr>
<td>FIEFDom (only two-domains)</td>
<td>91</td>
<td>94</td>
<td>73</td>
<td>70</td>
<td>80</td>
<td>39</td>
<td>36</td>
<td>90</td>
<td>35</td>
</tr>
<tr>
<td>FIEFDom (only multidomains)</td>
<td>89</td>
<td>91</td>
<td>76</td>
<td>71</td>
<td>95</td>
<td>86</td>
<td>82</td>
<td>96</td>
<td>88</td>
</tr>
</tbody>
</table>

All values are percentages. Five prediction sets were generated to understand how FIEFDom (with three versions of the same RPS), PPRODO and DOMpro perform on the SCOP 1.73 (30%) database. The first row shows the performance of FIEFDom that uses the SCOP 1.65 (30%) database as the RPS. The second and third rows show the performance of PPRODO and DOMpro, respectively. The fourth and fifth rows show the performance of FIEFDom that uses a RPS containing only two-domain proteins or multidomain proteins, respectively.
issues, we repeated the comparison twice with modified
RPSs, once with the RPS containing only two-domain
proteins and second time with the RPS containing only
multidomain proteins. We summarize the results in
Table 3. Thus, when FIEFDom uses the RPS that con-
tains only two-domain proteins, the average multidomain
prediction accuracy is 46%, and, when it uses the RPS
with only multidomain proteins, the average accuracy
is 79% while the respective one domain accuracies are 91% 
and 89%. From these results, it is clear that FIEFDom
successfully maintains higher performance levels com-
pared to these two programs when tested on a database
that is more recent and even when a systematically
domain-biased RPS is used. Note that PPRODO was opti-
mized for predicting two-domain proteins only, and hence
it has a tendency to divide many one-domain proteins into
two-domain proteins. This tendency to overpredict
domain boundaries is one of the main reasons for its
lower accuracy compared to FIEFDom. On the other
hand, the lower accuracies observed in the DOMpro
model are due to its tendency to underpredict domain
boundaries.

For the second comparison, we predict the domain
boundaries in the dataset used to develop DOMpro. The
rationale here is to check how well the models trained on
SCOP databases (FIEFDom and PPRODO) perform on
proteins derived from the CATH database. The CATH-
derived database used to train the DOMpro program con-
tains 963 one-domain proteins and 354 multidomain
proteins. Table 4 summarizes the results. Similar to the
previous comparison, Table 4 also includes the perfor-
mance of FIEFDom when using the RPS containing
only two-domain proteins or multidomain proteins. The
average domain prediction accuracies of FIEFDom,
PPRODO and DOMpro on the CATH-derived database
are 77%, 64% and 55%, respectively. If a RPS containing
only two-domain proteins is used, then the accuracy of
FIEFDom drops to 69%; when the RPS contains only
multidomain proteins, the accuracy becomes 74%. It is
clear from Table 4 that the application of FIEFDom on
either of three different training sets (a RPS with one
and multidomain proteins, a RPS with only two-domain
proteins, and a RPS with multidomain proteins) yields, on
average, better results compared with PPRODO and
DOMpro. In this test, the slight variations (35,40,48,49)
in domain definitions of the test database compared to the
training database did not adversely affect the performance
of our procedure.

Comparison with other sequence-based methods in CASP7

We compared the domain number prediction accuracy of
FIEFDom with six sequence-based methods (methods
that do not use protein-fold information or ab initio
processing) used in CASP7. The performance was
measured across the 97 targets (70 one-domain proteins
and 27 multidomain proteins) included in CASP7. In
addition to domain number prediction accuracy, we
also compared the ability of the methods to correctly pre-
dict both the domain number as well as the position of
the domain boundary. For one-domain proteins we consider
accuracy (A), and for multidomain number predictions,
specificity (Sp), sensitivity (Sn) and accuracy (A) were
determined. To rank the methods used in CASP7 we
determined the average prediction accuracy of both one-
and multidomain proteins for each method. If the
position of at least one domain in a multidomain protein
is not correctly predicted, the prediction is counted as a
‘partial’ success. If the positions of all domains in a multi-
domain protein are predicted correctly, it is counted as a
‘complete’ success. The results in Table 5 demonstrate that
FIEFDom has comparable or better accuracy when
compared to other methods. However, we caution that
analyses based on small data sets, such as the target set
used in CASP7, are less informative when compared to the
large scale analyses shown in the previous section.

Table 4. The performance metrics of the three programs on a dataset
that uses domain definitions derived from the CATH database

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of domains</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>One</td>
<td>Multi</td>
<td>A</td>
<td>Sp</td>
<td>Sn</td>
</tr>
<tr>
<td>FIEFDom</td>
<td>92</td>
<td>91</td>
<td>65</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>PPRODO</td>
<td>90</td>
<td>58</td>
<td>51</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>DOMpro</td>
<td>91</td>
<td>58</td>
<td>21</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>FIEFDom (only two domains)</td>
<td>89</td>
<td>91</td>
<td>50</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>FIEFDom (only multidomain)</td>
<td>89</td>
<td>91</td>
<td>62</td>
<td>58</td>
<td></td>
</tr>
</tbody>
</table>

All values are percentages. Five prediction sets were generated to
understand how FIEFDom (with three versions of the same RPS),
PPRODO and DOMpro perform on a database that derives its
domain definitions from the CATH database (version 2.5.1). The
results for two-, three- and four-domain proteins have been averaged
and are shown under ‘Multi’. The first row shows the performance of
FIEFDom that uses the SCOP 1.65 (30%) database as the RPS. The
second and third rows show the performance of PPRODO and
DOMpro, respectively. The fourth and the fifth rows show the perfor-
mance of FIEFDom that uses a RPS containing only two-domain
proteins or multidomain proteins, respectively.

DISCUSSION AND CONCLUSION

We propose a new and transparent method to predict
the domain boundaries for a given protein sequence.
The method is based on finding fragments similar to the
subsequences of the query sequence in the RPS and using
a FMO to infer domain boundaries from these fragments.
The query can either be a sequence or a sequence profile.
Our algorithm provides a domain recognition method that
mainly detects alignments to the super-family members
(SCOP classification) of the query sequence in the RPS.

For sequences that have few or no homologs in
the database, the profile of the sequence simply corre-
spends to the amino acid substitution matrix used in the
construction of the profile. Use of such profiles in the
profile-sequence alignment then becomes equivalent to
performing the sequence-sequence alignment in the
search of overlapping fragments. This, in effect, draws
the lower boundary of our prediction accuracy in these
cases. Conversely, if a query has a number of homologs
in the database of known sequences, then the profile is
well defined. Using a well-defined profile leads to more sensitive searches, resulting in higher prediction accuracy. A more rigorous implementation, using profile-profile alignment for finding similar fragments, is possible at the cost of increased computational time. In this way, our method can accommodate sequences that only have remote homologs with known boundaries (FIEFDom becomes a domain recognition method) and sequences that have many homologs with known domain boundaries (FIEFDom becomes a domain homology method).

One of the advantages of our approach is the transparency of the system. All of the processing is done using plain text files. The PSI-BLAST algorithm returns a text file (default output format) that contains all of the information about matching fragments. This human readable file is parsed by our program for modeling domain boundaries. Looking into the PSI-BLAST output file, the user can trace the sequences whose fragments matched with stretches of the query protein and contributed to the current decision. Since each neighbor (match) is weighted by its e-value, the relative contribution of each neighbor is apparent. This is contrary to black-box models in which the decision made by the model cannot be attributed to specific training data. Regardless of the alignment strategy (sequence-sequence or profile-sequence), the PSI-BLAST program produces similar output, and the actual prediction algorithm is independent of the alignment method used.

Although the sensitivity of FIEFDom is comparatively higher than the programs we compared with, we note that an even higher sensitivity would be desirable. However, when we compared the results of the runs that used RPS with/without labeled termini, we found that the sensitivity of the RPS-only run is increased at the cost of the specificity. Consequently, including the termini in the RPS results in lower one-domain accuracy and slightly higher multidomain accuracy. When we used the termini-included RPS on the SCOP 1.73 (30%) database, we obtained 81% one domain accuracy and (specificity, sensitivity, accuracy) of (78%, 71%, 59%) for multidomain proteins.

One of the problems of many data-driven bioinformatics tools is that they quickly become outdated if developers do not take time to update or make use of new data that become available after the tool is released. Updating a tool generally involves training and fine tuning the system with new data. In our case, the implementation of the algorithm is separate from the data used by algorithm. Consequently, FMO in FIEFDom does not need any training. For example, a new sequence representing a novel fold, can be easily added to the system by appending to the existing sequence file, and such new information is readily accounted for in the subsequent queries. There are many other advantages of keeping the RPS separate from the algorithm itself. First, the user can add/remove sequences from the RPS, altering the number of homologous sequences available to the algorithm. Second, the user can define the domain boundaries using a different database (for example, CATH database). Third, the user may choose whether or not to label the termini of the proteins in the RPS as domain boundaries. One of the benefits of including N- and C-termini into the RPS is that domain boundaries can be recognized for proteins that contain segments similar to experimentally determined structural domains. For example, the structurally-characterized zinc-binding RING finger domain, which is typically 40–60 residues in length (50), is present in proteins from many eukaryotic and viral genomes. FIEFDom, with labeled termini in the RPS, can detect these domains within larger proteins and assign domain boundaries before and after the identified segment (results not shown). However, when we compared the results of the runs that used RPS with and without labeled termini, we found that the sensitivity of the termini-included run is increased at the cost of the specificity. Consequently, including the termini in the RPS results in lower one-domain accuracy and slightly higher multidomain accuracy.

### Table 5: The performance of various sequence-based domain prediction methods on the 97 (70 one-domain proteins and 27 multidomain proteins) CASP7 targets

<table>
<thead>
<tr>
<th>Methods</th>
<th>One Number</th>
<th>Domain Position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>Sp</td>
</tr>
<tr>
<td>FIEFDom</td>
<td>100</td>
<td>88.9</td>
</tr>
<tr>
<td>CHOP (25)</td>
<td>55.8</td>
<td>37.5</td>
</tr>
<tr>
<td>DomSSEA (28)</td>
<td>92.9</td>
<td>100</td>
</tr>
<tr>
<td>DPS⁴</td>
<td>80.5</td>
<td>100</td>
</tr>
<tr>
<td>HHPred1⁵</td>
<td>95.6</td>
<td>100</td>
</tr>
<tr>
<td>HHPred3⁶</td>
<td>95.7</td>
<td>100</td>
</tr>
<tr>
<td>NNPutLab⁷</td>
<td>78.5</td>
<td>80.0</td>
</tr>
</tbody>
</table>

All values under the domain number prediction are percentages. Sequence-based domain prediction methods that were used in the CASP7 are listed on left. For one domain number prediction, the accuracy (A) is listed. For multidomain number prediction, accuracy (A), specificity (Sp) and sensitivity (Sn) are listed. The domain number prediction accuracy for all targets in CASP7 set is listed under the ‘Combined’ heading. For the domain position prediction of multidomain proteins, the actual count of the proteins whose domain boundaries are predicted completely correct and partially correct is listed.

³http://predictioncenter.org/casp7/meeting_docs/abstractsd.pdf.
accuracy compared to our stated results because the RPS that we are using is heavily weighted by protein sequences that have been amenable to experimental structural determination.

FIEFDom is a flexible tool that can predict domain boundaries for both proteins that have only remote homologs and proteins from highly sequenced families with high accuracy. The transparent model of FIEFDom provides insight into the problem in contrast to the current machine learning-based models. Due to rapid improvements in sequencing technologies, many new complete genomes are available every year, and, since our method can readily absorb new information without the need for model training, FIEFDom should maintain its relevance in the future.

ACKNOWLEDGEMENTS

We thank Drs. Jaques Reifman and Dong Xu for their critical review of the manuscript. We thank the anonymous reviewers for their valuable insight and suggestions.

FUNDING

This work was supported by the U.S. Department of Defense High Performance Computing Modernization Program, under the High Performance Computing Software Applications Institutes initiative. Funding for open access charge is same as funding for work.

Conflict of interest statement. The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the US Army or of the US Department of Defense. This paper has been approved for public release with unlimited distribution.

REFERENCES


