SSEP-Domain: protein domain prediction by alignment of secondary structure elements and profiles

Jan E. Gewehr* and Ralf Zimmer
Practical Informatics and Bioinformatics Group, Department of Informatics, Ludwig-Maximilians-University, Amalienstrasse 17, D-80333 Munich, Germany
Received on October 18, 2005; accepted on October 27, 2005
Advance Access publication November 2, 2005
Associate Editor: Anna Tramontano

ABSTRACT

Motivation: The prediction of protein domains is a crucial task for functional classification, homology-based structure prediction and structural genomics. In this paper, we present the SSEP-Domain protein domain prediction approach, which is based on the application of secondary structure element alignment (SSEA) and profile–profile alignment (PPA) in combination with InterPro pattern searches. SSEA allows rapid screening for potential domain regions while PPA provides us with the necessary specificity for selecting significant hits. The combination with InterPro patterns allows finding domain regions without solved structural templates if sequence families define existence.

Results: A preliminary version of SSEP-Domain was ranked among the top-performing domain prediction servers in the CASP 6 and CAFASP 4 experiments. Evaluation of the final version shows further improvement over these results together with a significant speed-up.

Availability: The server is available at http://www.bio.ifi.lmu.de/SSEP/

Contact: jan.gewehr@bio.ifi.lmu.de

Supplementary information: The supplementary data are available at http://www.bio.ifi.lmu.de/SSEP/

1 INTRODUCTION

Protein domains are considered the basic units for protein folding, evolution and function (Heger and Holm, 2003; Vogel et al., 2005). Being able to decompose proteins into their individual domains is therefore a crucial step for functional classification, homology-based structure prediction and structural genomics (Liu and Rost, 2003). The importance of this task is emphasized by the fact that the organizers of the CASP 6 (http://predictioncenter.org/) and the CAFASP 4 (http://www.cs.bgu.ac.il/~dfischer/CAFASP4/) protein structure prediction experiments decided to include a domain prediction category into their evaluations. Recent approaches to domain prediction range from mostly alignment-based algorithms like ADDA (Heger and Holm, 2003), Dompred-DomSSEA (Marsden et al., 2002), and DOMAINATION (George and Heringa, 2002a) over hybrid learning systems (Nagarajan and Yona, 2004), neural networks like PPRODO (Sim et al., 2005), statistics-based approaches like DGS (Wheelan et al., 2000), taxonomy-based approaches (Coin et al., 2004), clustering-based approaches like MKDOM (Gouzy et al., 1999) and DIVCLUS (Park and Teichmann, 1998) to 3D-oriented approaches like SnapDRAGON (George and Heringa, 2002b).

Secondary structure elements of proteins (i.e. the helices, strands and sheets) have been used several times in recent work. They are used to search for similar protein chains by the SSEP server (Shanthi et al., 2003), they can be aligned using e.g. the SSEA server (Fontana et al., 2005), and they are part of the GenTHREADER structure prediction (McGuffin and Jones, 2003) and the MANIFOLD fold prediction method (Bindewald et al., 2003). The DomSSEA protein domain prediction (Marsden et al., 2002) uses secondary structure element alignment (SSEA) for selecting PDB chains of potential templates.

Our approach, SSEP-Domain, is based on the observation that the fold class of a protein domain is often defined by the topology of its secondary structure elements (i.e. the elements and their lengths, the order of these elements, and the contacts between them). For protein domains, this means that their secondary structure element topology may be an indicator for fold class membership even if the sequence differs from all known members of the respective class. If the structure of a protein is unknown, we are still able to make use of its secondary structure elements and their order based on secondary structure prediction. Therefore, the comparison of subsequences of a target protein with known protein domains based on these features may reveal regions of the target that have the potential for being independent folding units.

We combine SSEA with profile–profile alignment (PPA) and InterPro patterns. While the fast SSEA procedure allows rapid screening for potential domain regions, PPA on both sequence and secondary structure profiles provides us with the necessary specificity for selecting significant hits. The combination with InterPro patterns allows finding domain regions without structural templates if sequence family definitions exist. As we use single domains as templates, we can predict multi-domain targets independently of whether the specific domain combination can be found in the protein data bank (PDB) (Berman et al., 2000) or not. Our settings allow predictions in <10 min per target on average on our test set.

SSEP-Domain performed well in both the CASP 6 and the CAFASP 4 experiments. Since then, we have significantly sped up our algorithm. In the evaluation of the final version of SSEP-Domain under CAFASP conditions, we also observe a slightly improved performance. We provide a web-based user interface for SSEP-Domain. Our service may also be used as part of the domain prediction meta-server META-DP (Saini and Fischer, 2005).

The SSEP-Domain approach has been previously mentioned as part of the CASP 6 (http://predictioncenter.org/casp6/abstracts/).
Fig. 1. Visualization of the different stages of the SSEP-Domain domain prediction pipeline.

We define the domain prediction task as the problem of decomposing a target protein sequence into subsequences, each of which represents one protein domain of the target. Like Nagarajan and Yona (2004), we consider only continuous subsequences as domain regions.

Our approach to domain prediction consists of three consecutive steps. We search for potential domain boundaries on the target, i.e., positions where domains most likely start or end, using SSEA. From these boundaries we deduce subsequences that may represent protein domains. In the second step these are scored using a combination of SSEA and PPA. We also include InterPro patterns as potential domain regions. In the final step, we combine the predicted regions recursively. This workflow is shown in Figure 1.

2 THE DOMAIN PREDICTION PIPELINE

2.1 Basic alignment methods and data

2.1.1 Secondary structure element alignment The SSEA procedure as used in this work was described in the supplementary material for McGuffin et al., 2001. In short, one reduces each secondary structure sequence to the sequence of its elements after discarding leading and trailing coils. These elements are then aligned using a simple dynamic programming protocol, where the scores are based on the type and the length of the aligned elements.

The main advantage of SSEA may be its simplicity which makes it extremely fast and thus ideal for screening and preselection purposes. It has been shown that, for a fold recognition benchmark set with low sequence homology, SSEA significantly outperforms sequence-based alignment methods (Bindewald et al., 2003).

2.1.2 Log average profile–profile alignment The PPA approach has recently become popular as it has proven to provide superior alignment quality as well as high fold recognition performance (Yona and Levitt, 2002; Rychlewski et al., 2000). The log average scoring function was developed as an extension of the amino acid similarity score for sequences (von Öhsen and Zimmer, 2001). PPA using the log average scoring function has been shown to compare favorably against other profile–profile approaches and against several alignment-based fold recognition methods (von Öhsen and Zimmer, 2001; von Öhse et al., 2003). In SSEP-Domain, we combine sequence and secondary structure profiles as described for the Arby structure prediction server (von Öhse et al., 2004), which performed well in the CAFASP 3 experiment (Fischer et al., 2003).

2.1.3 Target and template data For each target sequence, we run PSIPRED (Jones, 1999) against the NR database of April 2004. From this run we obtain not only the secondary structure prediction, but also the sequence and the secondary structure profile of the target. The same was done for each domain in our template library.

We use the atom-based ASTRAL compendium (Chandonia et al., 2004) based on SCOP (Murzin et al., 1995) version 1.65 (released in December 2003) and the corresponding subsets filtered for 95 and 25% sequence identity without genetic domains. Domains for which we obtained only zeros in the frequency profiles for the sequences were discarded. Furthermore, the ASTRAL compendium provides us with the classification of the templates into fold classes. The template library Domains contains the ASTRAL 95 subset.

2.1.4 Parameter calibration Some parameters were fitted to statistical evaluations on ASTRAL (length filter, significance filter, score normalization and gap penalties). All other parameters were calibrated such that SSEP-Domain achieves 100% accuracy with respect to the predicted domain number together with <10 min average runtime per target on a training set of about 500 randomly chosen PDB chains available in ASTRAL 1.65.

2.2 Step 1: finding potential domain boundaries

We regard all centers of predicted coil regions on the target sequence t as potential boundaries. Since the number of boundaries may affect the complexity of the method quadratically (see Step 2), we employ a heuristic to select only a reasonably small number of these centers for further evaluation (Algorithm 1).

2.2.1 SSEA and length filtering of templates Firstly, we collect all centers of predicted coil regions on t together with the start and the end of t in the set Centers. For each template domain sequence d in our template database Domains, we align d against all subsequences rij between coil centers ci and cj in Centers using SSEA. These subsequences may differ in length from |d| at maximum (|d| ≈ |rij|). The highest-scoring rij we call the domain image of d.

We chose the threshold of 5% based on all ASTRAL domains in version 1.65. For this set, the mean coil length at either end of a domain according to DSSP (Kabsch and Sander, 1983) applied to the coordinate files provided by ASTRAL is about 4.5 amino acids, and the mean length of a domain is about 188 amino acids. Using a threshold of 5%, for a potential region of length 188, we allow templates to differ by the average coil length at either end, i.e. by 9 amino acids at maximum. In addition, using a scaled threshold, we assume that with increased domain length the length differences between homolog domains are also increased.

2.2.2 Significance filtering of domain images For filtering out unlikely domain images, we compare the SSEA score of a hit s_max(d) against a threshold s_setup(d) derived from the all-against-all SSEA alignment score distribution of the fold class the template belongs to. These distributions were computed for each fold class by aligning all members against each other, based on ASTRAL 95.
Algorithm 1 Domain Boundary Search (Step 1)

1: // initialization
2: Centers ⇔ centers of coil regions predicted on target t 
3: Regions ⇔ \{ r_{ij} = [l; c_{i}, c_{j}] | c_{i}, c_{j} \in Centers \land c_{i} < c_{j} \}
4: Images ⇔ \{ \}
5: Domains ⇔ ASTRAL95 
6: 
7: // generation of domain images
8: for all template domains d ∈ Domains do
9: // get highest scoring region of similar length
10: s_{max}(d) ⇔ max\_{i,j\in Regions} s_{raw}(d, r_{ij})
11: 
12: // significance filtering: score high enough?
13: if s_{max}(d) > s_{thresh}(d) then
14: add corresponding region r_{ij} to Images
15: with score(r_{ij}) ⇔ s_{max}(d)
16: end if
17: end for
18: 
19: // accumulative scoring of coil centers
20: ∀ c ∈ Centers : score(c) ⇔ 0
21: for the top-scoring r_{ij} ∈ Images do
22: score(c_{i}) ⇔ score(c_{i}) + score(r_{ij})
23: score(c_{j}) ⇔ score(c_{j}) + score(r_{ij})
24: end for
25: select the top-scoring coil centers

Only hits having a score higher than the mean of the corresponding distribution are accepted and thus added to the set of domain images (Images). For classes having only one member, we use the mean of all computed means as threshold.

2.2.3 Accumulative boundary scores The score of each of the top-scoring 100 accepted domain images is then added to the corresponding coil centers. We select the ends of the target sequence as well as the four top scoring coil centers with respect to this accumulative score as domain boundaries.

2.3 Step 2: scoring of domain regions
A potential domain region is defined as a subsequence of the target that starts and ends at potential boundaries and contains at least 50 residues. In the second phase, we take a closer look at these regions r ∈ Regions (Algorithm 2).

2.3.1 Alignment-based region scores All fold classes are ranked by their highest-scoring member d (under the restriction that |d| ≈ |r|) with respect to the SSEA scores against r (see Step 1), and the highest-scoring 20 classes are selected. In order to find distant homologs in the members of these classes with matching secondary structures and similar lengths (D_{top}), we align each of them with r using PPA on both sequence and secondary structure profiles. The largest score of these alignments is assigned as score_{raw}(r) to r.

Algorithm 2 Scoring of Domain Regions (Step 2)

1: Regions ⇔ potential domain regions
2: 
3: for all r ∈ Regions do
4: // score fold classes by highest-scoring members
5: for all fold classes Fold ∈ Domains do
6: score(Fold) ⇔ max\_{d \in Fold} SSEA(r, d)
7: end for
8: 
9: // select members of potential fold classes
10: D_{top} ⇔ members of top-scoring fold classes
11: 
12: // score normalization for multiplicative scoring
13: score_{raw}(r) ⇔ max\_{d \in D_{top}, |d| \approx |r|} PPA(r, d)
14: score_{final}(r) ⇔ score_{raw}(r) / (10 log |r|)
15: end for

Fig. 2. Score histogram of fold recognition evaluation on an ASTRAL subset. For better readability, the histograms were cut at a score of 100, as few hits exceeded this score. The optimal threshold for partitioning scores into hits and misses lies at about score_{raw}(r) / (10 log |r|) = 10.

2.3.2 Score normalization We compute the final score of a potential domain region r as score_{final}(r) = score_{raw}(r) / (10 log |r|). Since we assume that the raw scores grow stronger than logarithmically with increasing region length, the denominator penalizes shorter regions. The factor of 10 results from fold recognition experiments using the combined PPA scores divided by the logarithm of the corresponding domain lengths on an ASTRAL subset filtered for 25% sequence identity. In this evaluation, we find that the optimal threshold to discriminate between hits and misses is score_{raw}(r) / (10 log |r|) ≈ 10 (Fig. 2).

By dividing the region scores by 10 log |r|, we obtain scores where the border between hits and misses is at about 1.0, a neutral score in our multiplicative approach to domain combination as described in Step 3. Potential hits have a score above 1.0 and therefore augment the score of a combination of domain regions, potential misses are below 1.0 and therefore diminish the final scores.

2.3.3 Patterns as domain regions We also add InterPro patterns (Mulder et al., 2003) found by the InterProScan program on the target sequence to the list of potential domain regions. The score of a single pattern is computed as one minus the E-value returned by InterProScan for the pattern. The maximum score of 1.0 allows patterns with highly confident hits to fill gap regions without affecting the overall multiplicative score of a domain combination as.
described in Step 3. In other words, we regard patterns as neutral domain regions against the background of the score transformations for alignment-based region scores. We discard members from PRINTS and PROSITE, since these databases contain many short-ranged patterns.

2.4 Step 3: combining multiple domain regions

Finally, for combining potential domain regions, we recursively try every possible non-overlapping combination.

2.4.1 Multiplicative scoring We score each combination $c$ based on the scores obtained for the regions in the previous steps and gap penalties for unassigned parts on the target sequence:

$$s(c) = \prod_{i=1}^{p} s(r_i) \prod_{j=1}^{p+1} g_j,$$

where $r_i, i \in \{1, \ldots, p\}$ denotes a participating region and $g_j$ denotes the factor for the unassigned region between $r_{j-1}$ and $r_j$, with $g_1$ being the gap at the beginning, and $g_{p+1}$ being the gap at the end of the target sequence.

2.4.2 Gap costs We assume that gaps may only contain coils. Furthermore, we assume that all known domains may be combined with each other independently of whether they occur in single- or multi-domain chains. Therefore, we analyzed the coil lengths at both ends of the DSSP (Kabsch and Sander, 1983) secondary structures of all ASTRAL domains. We do not penalize gaps of length less than 10 (see Section 2.2.1), and we allow only gaps shorter than the minimal domain length of 50. All gaps of length 10–49 are penalized with the empirically estimated probability of observing combined coils (the coil region at the end of the first domain plus the coil region at the beginning of the second domain) longer than 9.

This coarse-grained setup with only three different gap states (0, 9, 10–49, and >49) allows pattern boundaries to diverge from alignment boundaries within a range of the minimal domain length while favoring short gaps. If we find gaps after having scored the candidates for the final output, we elongate all regions equally until all gaps are closed. Thus, like many other predictors, we concentrate on boundaries between domains and do not predict inter-domain linker regions.

3 RESULTS

3.1 CAFASP 4 and CASP 6 results

CAFASP (critical assessment of fully automated structure prediction) (Fischer et al., 2003) is a blind test experiment for protein structure prediction. Experimentalists are contacted and asked for sequences of protein structures that will be solved shortly after the end of the prediction season. Then predictor groups try their servers in predicting these structures before they become available within one week after the release of a target. CAFASP 4 was held from May 2004 to September 2004, containing domain prediction as subcategory of the experiment. Here, SSEP-Domain was ranked among the top domain prediction servers (Saini and Fischer, 2005). The best performance was observed on so-called homology targets, targets for which templates having a high sequence identity were available.

In parallel to the CAFASP 4 experiment, the CASP 6 (critical assessment of structure prediction) experiment (Moult et al., 2005) was performed. The SSEP-Align structure prediction server that participated in CASP 6 contained SSEP-Domain as a first step for its predictions. SSEP-Align submitted domain predictions for 60 of the 63 evaluated targets along with the predicted protein structure models. SSEP-Align is ranked among the top ten predictors (both human and server groups) for all criteria, the best result being rank 6 on a set of multi-domain targets. Among the servers, SSEP-Align is ranked fourth on all targets and third on multi-domain targets with respect to NDO score (Tai et al., 2005; see Table 3 for comparison). In addition, SSEP-Align submitted the top-scoring prediction for the difficult multi-domain target T0237 (Tai et al., 2005; not evaluated in CAFASP). It should be noted that the CASP evaluation, when compared with CAFASP, is mainly based on overlap score, using different target sets and different domain definitions.

3.2 Final version under CAFASP conditions

In the following section, we will concentrate on the CAFASP 4 evaluation, since there SSEP-Domain participated as individual server and did not miss any submission. Furthermore, CAFASP evaluated more servers that did not participate in CASP than vice versa.

3.2.1 Changes after CAFASP 4 At the beginning of the experiment, we detected domain boundaries by sliding for each template domain a window of roughly the size of the domain along the target sequence. For each window position, we performed SSEAs similarly to the final method described in Step 1. InterPro patterns as additional domain regions were introduced shortly after the start of the experiments. Since the CAFASP version of SSEP-Domain needed up to several hours for one target, after the end of CASP we added length filtering in order to reduce the number of potential templates and replaced the exhaustive sliding window approach by the coil-center-based domain boundary search (see Step 1). Thus, the main difference between the final version and the preliminary version is the speed of the predictions.

This speed-up can be understood by looking at the number of performed PPAs in Step 2, which is the most time-consuming part of SSEP-Domain. Naively implemented, each potential domain region would be aligned against more than 9000 templates in the ASTRAL95. The preselection of potential fold classes using the SSEA scores without length filtering reduces the number of alignments per potential region to ~11% of the original number of templates. The additional length filter then reduces the number of alignments per region to <2% of the number of available templates. So we achieve a speedup of two orders of magnitude due to preselection and length filtering, resulting in <10 min average runtime per target.

Further, the final version yields slightly different predictions as indicated by the performance analysis (Table 1): two more targets are predicted correctly. One reason for this are the new, coil-centered boundaries. Using a sliding window as in the CAFASP version, there may be low-scoring predicted boundaries near to each other, while the new approach combines such blurred boundaries in the coil centers. This results in an accumulated score for each coil and thus a clearer picture of whether a coil may contain a linker region or not. In addition, the length filtering (see Section 2.2.1)
For comparison, the CAFASP values on all 58 targets are given (based on the values given on the official evaluation website for the numbers of correctly predicted targets on single- and two-domain targets). The percentage given for a predictor for a certain set is computed as the relative fraction of correct predictions in the corresponding set. Predictors were ranked by sensitivity on the common subset (Common) descendingly.

### 3.2.2 Experimental setup
For this work, we evaluated the final, sped-up version of our domain prediction under CAFASP 4 conditions in order to compare with our own CAFASP results as well as with the CAFASP performance of other servers. This means that the domain database we used as template data and for parameter calibration does not contain any of the test targets, since it was available before the start of CAFASP 4.

We quote the CAFASP 4 results from the official evaluation website (downloaded October 1, 2005) for the following methods: ADDA (Heger and Holm, 2003), Armadillo (Dumontier et al., 2005), BIOZON (Nagarajan and Yona, 2004), Dompred-DomSSEA (Marsden et al., 2002), Dompred-DPS (Marsden et al., 2002), DOMPRO (Cheng et al., 2005), DOPRO (N. von Øhsen), GLOBPLOT (Linding et al., 2003), MATEO (M. Lexa), Robetta-GINZU (Chivian et al., 2003) and Robetta-RosettaDOM (Kim et al., 2005).

Further, we quote the results of the CAFASP 4 domain prediction consensus method (Saint and Fischer, 2005) and InterPro (Mulder et al., 2003). To our CAFASP results we will refer as SSEP-CAFASP in the tables.

The CAFASP 4 test set contains 58 targets. Some servers had missing predictions during CAFASP 4, namely Armadillo (7), DomSSEA (5), DPS (5), GLOBPLOT (4) and BIOZON (1). For consistency reasons, in the tables we kept the values for all targets from the CAFASP 4 evaluation for sensitivity, specificity and average overlap score. These count missing submissions as wrong, ignore them or assign 0%, respectively. In addition, we computed the common subset of targets for which all servers sent predictions.

<table>
<thead>
<tr>
<th>Server</th>
<th>Single (%)</th>
<th>Two (%)</th>
<th>Common (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSENSUS</td>
<td>82.76</td>
<td>66.67</td>
<td>81.82</td>
</tr>
<tr>
<td>SSEP-Domain</td>
<td>82.76</td>
<td>66.67</td>
<td>81.82</td>
</tr>
<tr>
<td>SSEP-CAFASP</td>
<td>79.31</td>
<td>60.00</td>
<td>75.00</td>
</tr>
<tr>
<td>RosettaDOM</td>
<td>79.31</td>
<td>60.00</td>
<td>75.00</td>
</tr>
<tr>
<td>DOPRO</td>
<td>75.86</td>
<td>60.00</td>
<td>75.00</td>
</tr>
<tr>
<td>InterProScan</td>
<td>72.41</td>
<td>60.00</td>
<td>75.00</td>
</tr>
<tr>
<td>DOMPRO</td>
<td>70.69</td>
<td>60.00</td>
<td>75.00</td>
</tr>
<tr>
<td>DPS</td>
<td>62.07</td>
<td>60.00</td>
<td>75.00</td>
</tr>
<tr>
<td>GINZU</td>
<td>74.12</td>
<td>75.00</td>
<td>86.67</td>
</tr>
<tr>
<td>DomSSEA</td>
<td>65.52</td>
<td>75.00</td>
<td>86.67</td>
</tr>
<tr>
<td>GLOBPLOT</td>
<td>63.79</td>
<td>75.00</td>
<td>86.67</td>
</tr>
<tr>
<td>ADDA</td>
<td>65.52</td>
<td>75.00</td>
<td>86.67</td>
</tr>
<tr>
<td>MATEO</td>
<td>39.66</td>
<td>75.00</td>
<td>86.67</td>
</tr>
<tr>
<td>BIOZON</td>
<td>17.24</td>
<td>75.00</td>
<td>86.67</td>
</tr>
<tr>
<td>Armadillo</td>
<td>13.79</td>
<td>75.00</td>
<td>86.67</td>
</tr>
</tbody>
</table>

### Table 1. Sensitivity of predictions on single-domain, two-domain and all targets of the common subset of targets for which all servers submitted predictions

For comparison, the CAFASP values on all 58 targets are given (based on the values given on the official evaluation website for the numbers of correctly predicted targets on single- and two-domain targets). The percentage given for a predictor for a certain set is computed as the relative fraction of correct predictions in the corresponding set. Predictors were ranked by sensitivity on the common subset (Common) descendingly.

### Table 2. Specificity of predictions (based on the values given on the official evaluation website), rounded to full percentages

All submitted predictions were used, so e.g. the single-domain class may contain up to 41 correct predictions (all available single-domain targets in CAFASP 4). As in CAFASP 4, missing, split-, and multi-domain predictions were not considered. For each server, we give the fraction of correct predictions within a certain class followed by the number of all predictions made by the server for this class shown in brackets, in order to be able to distinguish between servers with high specificity but low sensitivity and vice versa. For instance, for SSEP-CAFASP on the two-domain set, 82% (11) means that 9 out of 11 submitted two-domain predictions were correct. For consistency with the CAFASP 4 evaluation, for the single and two set, specificity is computed as the number of correct predictions divided by the number of single- and two-domain predictions. Predictors were ranked by specificity on all targets descendingly.

This set is the basis for our rankings and plots. The following sets are used in our evaluation:

1. **CAFASP** contains all 58 targets, including those which were missed by some servers.
2. **Common** contains the 44 targets for which all servers submitted predictions (see above).
3. **Single** contains the 29 single-domain targets from the Common set.
4. **Two** contains the 15 two-domain targets from the Common set.

### 3.2.3 Sensitivity and specificity
For our first evaluation, we concentrate on the predicted number of domains. This assessment does not penalize situations where predicted boundaries are far from being correct, as long as the number of predicted domains equals the native domain definition. In CAFASP 4, sensitivity and specificity of the predicted domain numbers were evaluated. Sensitivity is defined as TP/(TP + FN), and specificity is defined as TP/(TP + FP). TP denotes the number of true positives, FP the number of false positives and FN the number of false negatives, each with respect to the evaluated category (e.g. single-domain). Furthermore, in CAFASP 4, split-domain predictions were considered as wrong predictions for the sensitivity evaluation and left out for the specificity evaluation. Therefore, in addition to our CAFASP-like evaluation, we computed the corresponding values for the affected servers (RosettaDOM and GINZU) after including split-domain predictions with respect to the number of predicted domains (see below).
CAFASP sets. We observe an increase of average overlap score on the CAFASP (91.28%, Table 3) and good specificity (83%, Table 2). SSEP-Domain achieves high sensitivity (81.82%, Table 1), high average overlap score (91.28%, Table 3) and good specificity (83%, Table 2).

Table 1 shows the sensitivity of the CAFASP 4 predictions together with the results of SSEP-Domain. With 48 of all 58 targets predicted correctly (82.76%), SSEP-Domain achieves the highest number of correct predictions of all individual servers. Only the CAFASP consensus method also achieves 48 correct predictions. Sensitivity evaluation on the Common set shows a similar picture: CONSENSUS and SSEP-Domain perform best, followed by SSEP-CAFASP, RosettaDOM and DOPRO. While RosettaDOM, CONSENSUS and DOPRO find more native two-domain proteins, SSEP-Domain achieves the highest number of correct predictions for single-domain proteins together with InterProScan.

Table 2 shows the corresponding specificity values. With 82%, SSEP-Domain achieves the highest specificity on two-domain targets. However, while we observe high overall sensitivity for SSEP-Domain, RosettaDOM, GINZU and CONSENSUS achieve higher overall specificity on all targets (Fig. 3, upper panel).

If we include split-domain predictions, we get different values for RosettaDOM and GINZU: RosettaDOM now achieves 79.55% sensitivity on Common and 86% specificity (of 56 counted predictions) on both single-domain and two-domain targets; GINZU achieves 72.73 and 80% (of 55 counted predictions), respectively.

3.2.4 Overlap score The second major part of the CAFASP evaluation is the assessment of the correct boundary placement using a so-called overlap score. We follow the CAFASP evaluators in using the algorithm described in Jones et al. (1998) for the predictions of the final version. The values for the CAFASP participants were taken from the official website. For this evaluation, split-domain predictions were included already in the original CAFASP 4 evaluation. Table 3 shows the overlap scores for all evaluated servers on the different sets. SSEP-Domain achieves the highest score of all evaluated predictors on the Single, Common, and CAFASP sets. We observe an increase of average overlap score on the CAFASP set of ~3% for SSEP-Domain over the CAFASP predictions (91.87–88.86%). Figure 3 (lower panel) shows all CAFASP 4 participants in a plot of sensitivity versus average overlap score on the Common set.

3.3 Influence of InterPro Direct comparison between InterPro as evaluated by CAFASP 4 and SSEP-Domain on CAFASP shows that we gain ~16% in average overlap score and ~10% in sensitivity by combining InterPro with our alignment-based approach. For our evaluation, we used InterProScan on InterPro 7.2 (March 25, 2004), which contains member databases with dates ranging from September 2003 to March 2004. A pattern occurs as part of the highest-scoring domain combination for 19 of the targets, and patterns lead to different predictions with respect to the predicted domain number than alignments alone for 2 of the 58 targets. In both cases, the alignment-based prediction would have been wrong.

4 DISCUSSION SSEP-Domain is an alignment-based approach to domain prediction (Table 4 gives an overview of the contained algorithms). We combine SSEA and direct boundary placement to detect potential

**Fig. 3.** Sensitivity on the Common set versus specificity on the single and two class (upper panel), and sensitivity versus average overlap score (lower panel) on the Common set. Each ‘o’ in gray represents one CAFASP participant. The results of SSEP-Domain in its final version are marked by a circle in black. SSEP-Domain achieves high sensitivity (81.82%, Table 1), high average overlap score (91.28%, Table 3) and good specificity (83%, Table 2).

**Table 3.** Average overlap score of the CAFASP 4 predictors and SSEP-Domain for the CAFASP, Single, Two, and Common sets

<table>
<thead>
<tr>
<th>Server name</th>
<th>CAFASP (%)</th>
<th>Single (%)</th>
<th>Two (%)</th>
<th>Common (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSEP-Domain</td>
<td>91.87</td>
<td>98.50</td>
<td>77.32</td>
<td>91.28</td>
</tr>
<tr>
<td>CONSENSUS</td>
<td>90.97</td>
<td>94.63</td>
<td>81.14</td>
<td>90.03</td>
</tr>
<tr>
<td>GINZU</td>
<td>89.31</td>
<td>93.36</td>
<td>81.48</td>
<td>89.31</td>
</tr>
<tr>
<td>RosettaDOM</td>
<td>89.88</td>
<td>92.51</td>
<td>82.55</td>
<td>89.12</td>
</tr>
<tr>
<td>SSEP-CAFASP</td>
<td>88.86</td>
<td>94.36</td>
<td>76.69</td>
<td>88.33</td>
</tr>
<tr>
<td>DomSSEA</td>
<td>81.50</td>
<td>95.48</td>
<td>73.66</td>
<td>88.04</td>
</tr>
<tr>
<td>DOMPRO</td>
<td>87.65</td>
<td>95.47</td>
<td>70.15</td>
<td>86.84</td>
</tr>
<tr>
<td>DPS</td>
<td>78.23</td>
<td>92.80</td>
<td>73.39</td>
<td>86.18</td>
</tr>
<tr>
<td>ADDA</td>
<td>85.04</td>
<td>93.53</td>
<td>69.19</td>
<td>85.23</td>
</tr>
<tr>
<td>DOPRO</td>
<td>84.96</td>
<td>88.00</td>
<td>76.59</td>
<td>84.11</td>
</tr>
<tr>
<td>GLOBPLOT</td>
<td>75.35</td>
<td>88.10</td>
<td>64.54</td>
<td>80.07</td>
</tr>
<tr>
<td>InterProScan</td>
<td>75.99</td>
<td>83.88</td>
<td>61.19</td>
<td>76.14</td>
</tr>
<tr>
<td>MATEO</td>
<td>73.22</td>
<td>78.67</td>
<td>66.65</td>
<td>74.57</td>
</tr>
<tr>
<td>BIOZON</td>
<td>62.16</td>
<td>61.32</td>
<td>66.99</td>
<td>63.25</td>
</tr>
<tr>
<td>Armadillo</td>
<td>49.70</td>
<td>49.57</td>
<td>63.19</td>
<td>54.21</td>
</tr>
</tbody>
</table>

Predictors were ranked by average overlap score on the Common set descendingly.

**Table 4.** Main algorithmic ingredients of SSEP-Domain and the corresponding purpose in the domain prediction pipeline

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSEA</td>
<td>Search for topological similarities; speeds up scoring</td>
</tr>
<tr>
<td>Coil centers</td>
<td>Avoids close-by, low-scoring potential boundaries</td>
</tr>
<tr>
<td>Length filter</td>
<td>Restricts search to templates of similar length; improves precision and speeds up the scoring</td>
</tr>
<tr>
<td>Score filter</td>
<td>Discards low-scoring SSEAs in Step 1</td>
</tr>
<tr>
<td>PPA</td>
<td>Final scoring of potential domain regions</td>
</tr>
<tr>
<td>Normalization</td>
<td>Prepares scores for multiplicative scoring</td>
</tr>
<tr>
<td>InterPro</td>
<td>Finds members of known sequence families</td>
</tr>
</tbody>
</table>

Details are given in the text in the corresponding sections.
domain boundaries on a target sequence. Domain regions are deduced from these boundaries and an InterPro pattern search. They are evaluated using a combination of SSEA and PPA on both sequence and secondary structure. The combination of multiple domain regions is done using a simple recursive algorithm based on the scores of the individual regions. For this pipeline, we observe an average runtime of <10 min per target on the CAFASP set with a maximum of 18 min on an Intel Xeon with 2.8 Ghz. This is a significant speed-up compared with the preliminary version of SSEP-Domain (see Section 3.2.1) which needed up to several hours per target. The evaluation of the influence of InterPro patterns shows that the combination of our alignment-based approach with InterPro patterns is indeed beneficial for domain prediction.

SSEP-Domain has been tested in the blind test scenario of CAFASP successfully, being part of the top group of domain predictors. Since features were added to the server during and after the experiment, we evaluated the final version under CAFASP 4 conditions. This gives us the opportunity to compare our results with the CAFASP predictors. SSEP-Domain performs well, achieving high sensitivity, high overlap scores and good specificity. The final version yields the best overall accuracy of domain predictions as measured by overlap score due to an improved performance over the preliminary version. Direct comparison with other CAFASP participants shows that SSEP-Domain performs very well on single-domain proteins, but 3 of the other 14 methods (RosettaDOM, GEANFAMMER package that finds homologous domains in single- and multi-domain proteins, and Inugo and RosettaDOM) have higher overlap scores on two-domain proteins.

On our website, we have made SSEP-Domain available to the community. In future work, we will concentrate on further speeding up the algorithm in order to make it more suitable for structural genomics purposes. Furthermore, we will work on achieving higher accuracy for multi-domain proteins.

ACKNOWLEDGEMENTS

Niklas von Oehsen kindly provided the PPA software. This work was funded by the DFG under grant PROSEQO II (Zi 616/2).

Conflict of Interest: none declared

REFERENCES
