AutoSCOP: automated prediction of SCOP classifications using unique pattern-class mappings

Jan E. Gewehr*, Volker Hintermair and Ralf Zimmer

Practical Informatics and Bioinformatics Group, Department of Informatics, Ludwig-Maximilians-University Munich, Amalienstr. 17, D-80333 Munich, Germany

Received on November 06, 2006; revised on February 28, 2007; accepted on March 05, 2007

Advance Access publication March 22, 2007

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: The sequence patterns contained in the available motif and hidden Markov model (HMM) databases are a valuable source of information for protein sequence annotation. For structure prediction and fold recognition purposes, we computed mappings from such pattern databases to the protein domain hierarchy given by the ASTRAL compendium and applied them to the prediction of SCOP classifications. Our aim is to make highly confident predictions also for non-trivial cases if possible and abstain from a prediction otherwise, and thus to provide a method that can be used as a first step in a pipeline of prediction methods. We describe two successful examples for such pipelines. With the AutoSCOP approach, it is possible to make predictions in a large-scale manner for many domains of the available sequences in the well-known protein sequence databases.

Results: AutoSCOP computes unique sequence patterns and pattern combinations for SCOP classifications. For instance, we assign a SCOP superfamily to a pattern found in its members whenever the pattern does not occur in any other SCOP superfamily. Especially on the fold and superfamily level, our method achieves both high sensitivity (above 93%) and high specificity (above 98%) on the difference set between two ASTRAL versions, due to being able to abstain from unreliable predictions. Further, on a harder test set filtered at low sequence identity, the combination with structure alignment methods improves accuracy and performs comparably even to structure alignment methods. Integrating our method with structure alignment, we are able to achieve an accuracy of 99% on SCOP fold classifications on this set. In an analysis of false assignments of domains from new folds/superfamilies/families to existing SCOP classifications, AutoSCOP correctly abstains for more than 70% of the domains belonging to new folds and superfamilies, and more than 80% of the domains belonging to new families. These findings show that our approach is a useful additional filter for SCOP classification prediction of protein domains in combination with well-known methods such as profile–profile alignment.

Availability: A web server where users can input their domain sequences is available at http://www.bio.ifi.lmu.de/autoscop

Contact: jan.gewehr@ifi.lmu.de

*To whom correspondence should be addressed.

1 INTRODUCTION

One of the subtasks of protein structure prediction, fold recognition, is to predict the fold of a target protein (domain) sequence with unknown structure with respect to the classifications given by domain databases like SCOP (Murzin et al., 1995) and CATH (Orengo et al., 1997). This can be achieved by comparing the sequence itself or, in addition, derived features to a set of representatives for these classifications (so-called templates). We define the problems of family recognition and superfamily recognition analogously as the problem of assigning the correct family/superfamily to a target sequence.

In this article, we describe AutoSCOP, a simple approach for SCOP classification prediction (or simply SCOP prediction) of protein domain sequences. However, the aim of our work is not to come up with a new standalone predictor but instead with a method that can be combined with already existing, well-performing methods for SCOP prediction. We aim at building a method that is highly specific and is at the same time able to make predictions for non-trivial cases, i.e. cases with low sequence identity to the available template sequences. If we cannot be at least relatively sure of a prediction, we abstain and leave the problem to subsequent steps which may be able to correctly predict the SCOP classification of the corresponding instances. AutoSCOP works as a filter where all highly confident predictions are caught and all others are passed on for further processing.

Our data source are sequence patterns as provided by various databases. Recently, a number of sequence pattern databases have been introduced and evaluated in different contexts. Our approach allows for the integration of this data into a single SCOP prediction framework. For an exemplary evaluation, InterPro (Mulder et al., 2003) provides us with a collection of useful databases including Pfam (Bateman et al., 2004) and SUPERFAMILY (Gough and Chothia, 2002). Though e.g. SUPERFAMILY uses structure information for generating libraries of hidden Markov models (HMMs), on the target or query side we only make use of the sequence and do not need the corresponding structure.

Our approach can be used for any collection of pattern or feature databases, with the InterPro compendium being a convenient example for such a collection which was already
successfully used by other approaches with different prediction aims. We recently used InterPro patterns in our protein domain prediction method SSEP-Domain (Geweehr and Zimmer, 2006). InterPro has further proven to be a valuable resource for EC number prediction using association rule mining (Chiu et al., 2006). Artamonova et al. (2005) have evaluated and successfully used association rules to improve sequence annotation which includes InterPro patterns among other data. In a recent study (Brézellec et al., 2006), the mapping of Pfam annotations to organisms was found to be useful for the identification of dam-associated genes with potential link to DNA maintenance. Especially interesting for SCOP predictions is the mapping between SCOP families and Pfam patterns as investigated by Zhang et al. (2005) This mapping showed that there is a general agreement between these databases, but there still are areas of disagreement as well as unmapped SCOP domains.

Given the latter result, it is obvious that Pfam patterns can be used for SCOP prediction, but it is necessary to discard the disagreeing pattern occurrences and fill the gaps resulting from the unmapped domains with patterns from further data sources. Our approach, which is based on unique mappings from patterns to SCOP classifications, exploits this idea with respect to highly specific SCOP prediction using multiple databases. For maximizing specificity, we introduce a strict criterion for the acceptance of a mapping between a pattern occurrence and a SCOP classification that allows us to discard all mappings that do not clearly match the SCOP hierarchy. Thereby, increasing the number of databases simultaneously increases the coverage on the training data from 64.7% for Pfam alone to 86.2% for all InterPro member databases on family level. On fold and superfamily levels, we achieve a coverage of 99%.

The assignment of patterns to SCOP classifications was trained on one of the most comprehensive datasets available, the ASTRAL compendium (Chandonia et al., 2004). The predictive power was evaluated in a blind test like scenario using three different sets: (1) the complete difference set between two ASTRAL versions (which contains many ‘easy’ predictions due to high sequence identities), (2) a more difficult set with low sequence identities which was used for structure alignment evaluation by Birzele et al., (2007) and (3) the CAFASP4 targets. We made use of an InterPro version that was released before the ASTRAL domains we used in our test set, such that the contained HMMs, profiles and expressions could not have been trained on the SCOP classifications used for testing.

We evaluated the abilities of our method as a filter by combining it with an alignment method well known for its performance in fold recognition, namely log average profile–profile alignment (PPA, von Ohsen et al., 2003). The combination was tested on the second, more difficult dataset. Here, although we do not make use of the target structure, we could achieve results that are comparable even to structure alignment methods. Further, we observe an improvement over the best structure-based method on this set (Vorolign, Birzele et al., 2007) when we combine this method with our method, similarly to the combination with PPA. On the third set, the CAFASP4 targets, we find that we can contribute SCOP predictions for about half of the targets with classifications available in the latest ASTRAL release.

Albeit being simple, the concept of unique patterns is a quite powerful tool for SCOP prediction. The inclusion of unique pattern combinations does not significantly improve performance over unique patterns alone but helps a bit on family level. A possible reason for this is the high co-occurrence of patterns from different databases: for instance, a test domain may be classified correctly by patterns from different databases simultaneously. The extensibility of AutoSCOP was demonstrated successfully by including also ASTRAL HMMs trained on SCOP families into our approach, which increased sensitivity on family level on the complete difference set significantly (already without pattern combinations).

2 METHODS

2.1 Test and training data

As training data, we use the ASTRAL compendium based on SCOP 1.65 (Chandonia et al., 2004), namely the atom-based entries as provided by the corresponding sequence file. This set contains 50979 domains, excluding so-called genetic domains (protein parts glued together from different protein chains). For testing purposes, we use three different sets:

(1) We computed the difference set between the ASTRAL versions 1.65 and 1.67 based on the domain IDs, again under exclusion of genetic domains. The resulting 10039 domains are used to estimate the performance of our approach in a blind test like situation. We aligned all 10039 domains against the ASTRAL 95 subset based on SCOP 1.65 (i.e. a representative set filtered for 95% sequence identity at most) using global sequence alignment. From these runs, we computed the sequence identities in the alignments. We find a large number of highly identical templates in our training data (only ~ 50% of the test domains have below than 95% sequence identity with the training set). Nonetheless, for our evaluation, we kept all 10039 domains, in order to completely reflect the picture given by the ASTRAL/SCOP versions. These domains contain 536 SCOP folds, 804 SCOP superfamilies and 1251 SCOP families. It should be noted that 458 of these folds contain only one superfamily. AutoSCOP’s prediction accuracy for those superfamilies belonging to folds with more than one superfamily in the test set was found to be comparable to the overall prediction accuracy.

(2) The first, complete difference set contains many easy targets, as described above. In order to compare our approach with well-known alignment methods, we further made use of the more difficult subset of these domains where such trivial targets were filtered out as described in (Birzele et al., 2007): All domains from the test set that belong to a family that already existed in ASTRAL 1.65 were selected. Then, all domains having more than 30% sequence identity with more than 30 identically aligned residues to any domain in the training data were removed, resulting in 979 remaining test domains. This set is very interesting for our evaluations, as we especially aim at a good performance in non-trivial cases. The 979 domains are classified in 129 different folds, 169 different superfamilies and 208 different families.

(3) In addition, we evaluated our method on the 58 CAFASP4 targets (see Section 3.3).
2.2 The AutoSCOP approach

We use the InterProScan program (Quevillon et al., 2005) against the InterPro 7.2 databases for searching sequence patterns on the amino acid sequences in our training and test data. In our training dataset (the ASTRAL compendium version 1.65), we find 6702 different patterns, using the InterPro member databases ProDom (Bru et al., 2005), Pfam (Bateman et al., 2004), PIRSF (Wu et al., 2004), PRINTS (Atwood 2002), PROSITE (Hulo et al., 2004), SMART (Letunic et al., 2004), SUPERFAMILY (Gough and Chothia, 2002) and TIGRFAMs (Haft et al., 2003) as well as ‘seg’ (Wootton, 1994) and ‘coil’ (Lupas et al., 1991) indicators as provided by InterProScan.

It was observed before for the mouse secretome (Grimmond et al., 2003), that some InterPro patterns as well as SUPERFAMILY predictions were exclusively found in secretome proteins and that such occurrences might be used as an alternative approach to identifying putative secretome proteins. The authors of the PANDORA system (Kaplan et al., 2003) also suggest to analyze protein sets as given by GO (Camon et al., 2003) or SCOP by studying shared keywords.

In a similar fashion, we assign those patterns that occur in only one subtree of the classification hierarchy (in our training data) with respect to a particular SCOP level as so-called unique patterns to the corresponding SCOP subtree. Thus, for instance, a superfamny may be described by a set of unique patterns, each of which covers a subset of the superfamily’s members.

Given a sequence with unknown classification for prediction we again use InterProScan to detect patterns. We then compare the found patterns to our database of unique patterns. If any unique pattern as defined on the training data is found, we assign the corresponding classification to the sequence. If we find unique patterns with differing SCOP classification assignments, we abstain. In cases where we do not find unique patterns, we look for unique combinations of common patterns.

2.2.1 Stage 1: unique patterns

Let $P$ denote the set of all patterns that have been found in the training data. For a classification task, let $C$ define a set of classes. In particular, for SCOP prediction, let $l \in \{\text{fold, superfamny, family}\}$ denote a SCOP level and $C_l$ denote the set of SCOP classifications on level $l$, e.g. $C_{\text{fold}} = \{'a.1', 'a.2', \ldots\}$.

Definition: pattern-class graph. The pattern-class graph $G_l = (V_l, E_l)$ for a level $l$ is defined as a bipartite graph using patterns $P$ and classifications $C_l$ as nodes $V_l$. An edge $e \in E_l \subseteq P \times C_l$ exists between a pattern $p \in P$ and a class $c \in C_l$ iff pattern $p$ occurs in at least one member (i.e. one protein domain sequence) of $c$.

Definition: unique patterns. A pattern $p \in V_l$ is called a unique pattern for a pattern-class graph $G_l$ iff $\deg_{G_l}(p) = 1$, i.e. $p$ has exactly one adjacent edge in $G_l$. We define $P^*_l = \{p \in P | \deg_{G_l}(p) = 1\}$ as the set of unique patterns on pattern-class graph $G_l$.

Thus, we obtain functions $f_l : P^*_l \rightarrow C_l$ that return the corresponding classification for a unique pattern (see Fig. 1 for an illustration). For domain sequences, we can now define prediction functions $f : \text{Sequences} \rightarrow C_l$ that map a sequence to a SCOP classification on level $l$ if $f$ maps all unique patterns found on the sequence to the same classification, and we abstain otherwise.

2.2.2 Stage 2: pattern combinations

In stage 2, we also include unique combinations of common patterns (patterns that are not unique with respect to the chosen SCOP level), but only if no unique patterns are found. We analyze combinations of common patterns by searching for consensus classifications. Each common pattern occurs in a number of different SCOP classifications on the chosen hierarchy level. If the intersection of the sets of possible classifications for all found common patterns contains exactly one remaining classification assignment, we predict this classification for the target. If the intersection is empty or contains more than one possible classification, we abstain from a prediction.

2.2.3 AutoSCOP*: inclusion of further data sources

In order to show the extensibility of the AutoSCOP approach, we also included predictions made by HMMs trained on SCOP families as provided by ASTRAL for SCOP version 1.65. Predictions were made using HMMer 2.3.2 (S.R.Eddy, http://hmmer.janelia.org) against the complete HMM library. For each target, the top hit was used like any InterPro pattern whenever the e-value obtained by the hit was below an e-value of 0.1, which is proposed as a useful cutoff in HMMer’s user’s guide. We will refer to AutoSCOP including ASTRAL Family HMMs as AutoSCOP* in the results section.

3 RESULTS

3.1 Mapping of training domains

Table 1 shows the number of unique patterns for each individual database in our training data. In fact, most of the patterns (6410 of 6702, 95.64%) are unique on at least the fold level, which is an indicator for the high quality of the database scan results. Here, we can also assess the performance of our pattern-based approach on the training data. With unique patterns alone, we can correctly assign folds, superfamilies and families to 99.12, 98.99 and 86.20% of all domains, respectively. Unique combinations of common patterns can add nearly 3% on family level.

3.2 ASTRAL difference set—new SCOP domains

In our test set, we find 97 new SCOP folds, 163 new SCOP superfamilies and 326 new SCOP families. On fold level, 433 of the 10 039 domains in our test set belong to new folds and are therefore considered as ‘new’ in our framework, i.e. targets for which we do not have a template with a similar fold in our database. Accordingly, the remaining 9606 domains are considered ‘known’, i.e. targets for which a correct prediction would be possible. On superfamny level, we have 601 ‘new’ and
9438 ‘known’ domains, and on family level, we have 1167 ‘new’ domains and 8872 ‘known’ domains. We evaluate prediction accuracy on the ASTRAL difference set for each SCOP level by means of:

- sensitivity: the number of correct predictions on ‘known’ domains divided by the number of all ‘known’ domains, and
- specificity: the number of correct predictions divided by the number of all predictions, including wrongly predicted ‘new’ domains.

### 3.2.1 Individual contributions of InterPro databases
It is interesting to see how the individual contributions of the databases differ (Table 1). On fold level, the SUPERFAMILY database already covers 96.5% of all domains (the best individual result on fold level). A similar performance can be observed on superfamily level with 94.6%. However, on family level, SUPERFAMILY only achieves 30% coverage, whereas here Pfam achieves the best result with 65%. This shows that some patterns are good for certain levels of the SCOP hierarchy (e.g. SUPERFAMILY for fold and superfamily), but none performs best on all SCOP levels.

Table 2 shows the performance of AutoSCOP on the test data after leaving out individual databases. As could be expected from the coverage analysis, SUPERFAMILY and Pfam are especially important. Some databases are nearly completely covered by the remaining InterPro members for our purpose. PRINTS and PROSITE are interesting, as these databases increase performance on family level but slightly decrease performance on superfamily and fold level.

Table 2. Influence of InterPro databases as measured by the prediction performance of AutoSCOP (i.e. our approach on InterPro data) after leaving out individual databases

<table>
<thead>
<tr>
<th>Method</th>
<th>Family sensitivity</th>
<th>Family specificity</th>
<th>Superfamily sensitivity</th>
<th>Superfamily specificity</th>
<th>Fold sensitivity</th>
<th>Fold specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>AutoSCOP</td>
<td>84.20%</td>
<td>96.77%</td>
<td>94.40%</td>
<td>98.17%</td>
<td>93.36%</td>
<td>98.13%</td>
</tr>
<tr>
<td>W/o Pfam</td>
<td>74.57%</td>
<td>97.06%</td>
<td>92.77%</td>
<td>97.73%</td>
<td>92.33%</td>
<td>98.11%</td>
</tr>
<tr>
<td>W/o PIR</td>
<td>84.19%</td>
<td>96.77%</td>
<td>94.40%</td>
<td>98.17%</td>
<td>93.36%</td>
<td>98.13%</td>
</tr>
<tr>
<td>W/o PRINTS</td>
<td>82.96%</td>
<td>96.65%</td>
<td>94.51%</td>
<td>98.17%</td>
<td>93.47%</td>
<td>98.14%</td>
</tr>
<tr>
<td>W/o PRODOM</td>
<td>83.60%</td>
<td>96.80%</td>
<td>94.31%</td>
<td>98.16%</td>
<td>93.27%</td>
<td>98.13%</td>
</tr>
<tr>
<td>W/o PROSITE</td>
<td>82.78%</td>
<td>97.13%</td>
<td>94.56%</td>
<td>98.19%</td>
<td>93.47%</td>
<td>98.16%</td>
</tr>
<tr>
<td>W/o SMART</td>
<td>77.70%</td>
<td>96.52%</td>
<td>94.39%</td>
<td>98.14%</td>
<td>93.35%</td>
<td>98.11%</td>
</tr>
<tr>
<td>W/o SUPERFAMILY</td>
<td>79.76%</td>
<td>97.17%</td>
<td>84.49%</td>
<td>99.06%</td>
<td>83.26%</td>
<td>99.08%</td>
</tr>
<tr>
<td>W/o TIGRFAMs</td>
<td>82.79%</td>
<td>96.64%</td>
<td>94.40%</td>
<td>98.17%</td>
<td>93.36%</td>
<td>98.13%</td>
</tr>
</tbody>
</table>

Note: This analysis shows that many predictions are made by more than one database. As in the coverage analysis, again SUPERFAMILY and Pfam have the highest impact on the final method. The lowest value for each column was highlighted.
all three evaluated levels of the SCOP hierarchy. For the family level, unique patterns on InterPro data achieve a specificity of more than 96%, for superfamiliy and fold more than 98%. With respect to sensitivity, this approach performs best on fold and superfamily level, achieving values of more than 93 and 94%, respectively. As expected from the mapping results, the less complete coverage of the family level is reflected in the prediction performance for SCOP families in a sensitivity of only ~80%.

The inclusion of pattern combinations has practically no effect on fold and superfamily predictions, but, on family level, slightly increases sensitivity and slightly decreases specificity. More importantly, the inclusion of ASTRAL Family HMM predictions (AutoSCOP*) rises sensitivity up to 95% on family level even for unique patterns alone and also slightly improves performance on superfamily and fold level while keeping high specificity.

Comparison with our reference methods shows that AutoSCOP* achieves the highest sensitivity of all compared sequence-based methods on superfamily and fold level (its sensitivity being only second to the structure alignment method Vorolign). On family level, where sequence similarity is most important, AutoSCOP*, PSI-BLAST (Altschul et al., 1997) and Vorolign are close together and achieve sensitivities of 95.25, 95.57 and 95.91%, respectively, but with clearly lower specificity for Vorolign and PSI-BLAST as compared to AutoSCOP*. Asteroids and Pfam generally achieve higher specificity but significantly lower sensitivity.

## False assignments of domains from new classifications

Errors often result from the assignment of test domains from ‘new’ classifications to ‘known’ classifications. Using unique InterPro patterns on fold level, we make 170 false assignments, 115 of which are wrongly assigned new fold domains (67.64%). On superfamiliy level, of the 164 false assignments, 135 fall into this category (82.31%). On family level, of the 226 false assignments, we have 192 targets from new families (84.95%). Correspondingly, we correctly abstain from a prediction for 73.44% of the test domains belonging to new folds (318 of 433), 77.53% of the test domains belonging to new superfamilies (466 of 601) and 83.54% of the test domains belonging to new families (975 of 1167).

On superfamiliy level, we further analyzed the wrong assignments made by unique InterPro patterns from new superfamilies to already existing superfamilies (135): ~70% have corresponding PSI-BLAST hits with e-values less than 1E−5, and nearly 50% have PSI-BLAST hits with e-values less than 1E−20. This shows that, as judged by sequence similarity, many of these assignments are reasonable. All errors with clear PSI-BLAST hits could be attributed to changes in the classification or in the domain definition between ASTRAL 1.65 and newer versions.

### Fold prediction of CAFASP4 targets

CASP (Moult et al., 2005) and CAFASP (Fischer et al., 2003) are community-wide blind test experiments for protein structure prediction. Our domain prediction method SSEP-Domain (Geweher and Zimmer, 2006) (which includes an InterProScan run on a target protein) participated in CAFASP4 during 2004 (Saini and Fischer, 2005). We analyzed the InterPro pattern occurrences that were found by SSEP-Domain during this experiment. The databases for all evaluations presented in this article were chosen such that we make use of data already

---

**Table 3. Sensitivity and specificity of Auto SCOP and reference methods**

<table>
<thead>
<tr>
<th>Method</th>
<th>Family sensitivity</th>
<th>Family specificity</th>
<th>Superfamily sensitivity</th>
<th>Superfamily specificity</th>
<th>Fold sensitivity</th>
<th>Fold specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>AutoSCOP*</td>
<td>95.25%</td>
<td>96.28%</td>
<td>94.96%</td>
<td>98.07%</td>
<td>93.85%</td>
<td>98.13%</td>
</tr>
<tr>
<td>AutoSCOP*, unique patterns only</td>
<td>95.01%</td>
<td>96.43%</td>
<td>94.95%</td>
<td>98.08%</td>
<td>93.85%</td>
<td>98.13%</td>
</tr>
<tr>
<td>AutoSCOP</td>
<td>84.20%</td>
<td>96.77%</td>
<td>94.40%</td>
<td>98.17%</td>
<td>93.36%</td>
<td>98.13%</td>
</tr>
<tr>
<td>AutoSCOP, unique patterns only</td>
<td>80.34%</td>
<td>96.92%</td>
<td>94.39%</td>
<td>98.36%</td>
<td>93.36%</td>
<td>98.13%</td>
</tr>
<tr>
<td>Asteroids</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>ASTRAL Family HMMs</td>
<td>91.45%</td>
<td>96.74%</td>
<td>86.35%</td>
<td>99.39%</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>BLAST</td>
<td>93.59%</td>
<td>96.26%</td>
<td>89.65%</td>
<td>99.09%</td>
<td>88.17%</td>
<td>98.19%</td>
</tr>
<tr>
<td>Pfam</td>
<td>61.14%</td>
<td>97.67%</td>
<td>78.57%</td>
<td>99.10%</td>
<td>77.40%</td>
<td>99.12%</td>
</tr>
<tr>
<td>PSI-BLAST</td>
<td>95.57%</td>
<td>93.06%</td>
<td>94.31%</td>
<td>97.69%</td>
<td>93.07%</td>
<td>98.13%</td>
</tr>
<tr>
<td>SUPERFAMILY</td>
<td>27.22%</td>
<td>97.06%</td>
<td>89.31%</td>
<td>97.86%</td>
<td>90.22%</td>
<td>98.20%</td>
</tr>
<tr>
<td>Vorolign</td>
<td>95.91%</td>
<td>89.52%</td>
<td>96.60%</td>
<td>95.91%</td>
<td>96.66%</td>
<td>97.68%</td>
</tr>
</tbody>
</table>

**Note:** In total, we were able to match 97.5% of all test domains to Asteroids regions using BLAST. Sensitivity for Asteroids was computed relatively to the number of domains coming from known superfamilies in this set. Further, we included the predictions made by the ASTRAL Family HMMs based on ASTRAL 1.65 directly (see Section 6.2.3) and computed individual unique pattern results for the two InterPro databases with the highest coverage (SUPERFAMILY and Pfam). We added PSI-BLAST and BLAST predictions with e-value cutoffs that give comparable specificity to AutoSCOP on fold level (7.0 and 0.2, respectively). For PSI-BLAST, we used five iterations and an inclusion threshold of 0.001 against an NR database downloaded from the NCBI in April 2004 (which was available at the same time as SCOP 1.65) together with ASTRAL 1.65, using the best ASTRAL hit below the threshold in the last iteration as prediction. For BLAST, we directly used ASTRAL 1.65 as database.

In addition, the structure alignment method Vorolign was applied using its default settings. The highest sequence-based value for each column was highlighted. Upper part: Sensitivity and specificity on different SCOP levels on test set number 1, the ASTRAL 1.67-1.65 difference set, for AutoSCOP and AutoSCOP*. Lower part: Reference methods. As a reference for the prediction quality of AutoSCOP, we used ASTRAL’s Asteroids (version 1.65-040809) predictions on superfamily level (Asteroids’ most complete level).
available before the beginning of CAFASP4 only. For 46 of the 58 CAFASP4 targets, we can find SCOP annotations in ASTRAL 1.71. For 23 of these targets, we can make AutoSCOP predictions on the InterPro data (50%). Of these 23, 21 are completely correct (91.3%), including two two-domain targets for which AutoSCOP finds the correct fold for both domains. One target is a new fold but is wrongly predicted as belonging to a known fold. For the remaining target, two possible folds were found, one of which is correct.

3.4 Performance in the sequence twilight zone

We compared our approach to well-known alignment methods using the test set of non-trivial targets defined by Birzele et al. (2007) (see Section 2.1). Vorolign and CE results were quoted from Birzele et al. (2007). For profile–profile alignment, we aligned the target domains against the ASTRAL 25 compendium (Version 1.65) as a representative template set as described by Birzele et al. but without using Vorolign’s secondary structure element-based filtering. PSI-BLAST hits were computed as described for the complete difference set (Table 3). For all alignment methods, the classification of the top scoring template was used as the predicted classification of the target sequence.

Table 4 shows the results. We find that AutoSCOP* performs better on superfamily and fold than on family level. On these SCOP levels, sensitivity is slightly worse than for global log average PPA on both sequence and secondary structure profiles, which has been shown to be a very sensitive and accurate approach for fold recognition (von Öhlsen et al., 2003, 2004). When using only InterPro patterns (AutoSCOP), we lose 0.4% on superfAMILY and fold level and 8.1% on family level as compared to AutoSCOP*. Further, AutoSCOP* achieves specificity rates of 99.9% (fold), 99.6% (superfamily) and 97.0% (family) due to being able to abstain from predictions, which matches our filtration goal.

3.5 Using AutoSCOP* as a filter

3.5.1 Sequence-based prediction in combination with PPA For evaluation of AutoSCOP’s ability as a filter, we combine AutoSCOP with PPA as follows: We predict the SCOP classification using AutoSCOP*. For all abstentions, we then align the corresponding targets against the ASTRAL 25 using PPA as described above. The corresponding results are shown in Table 4, labeled as AS* + PPA. We find that, in combination, we achieve ~4% improvement over the best individual method on fold and superfamily level, and ~7% on family level. We also find that using AutoSCOP* as a filter clearly increases accuracy over using PSI-BLAST or Asteroids as a filter. Comparison with the results of structure alignment methods on the same test set as an upper bound to accuracy shows that our combination outperforms the well-known CE method (Shindyalov and Bourne, 1998) on all levels and the best structure alignment method in our comparison (Vorolign, Birzele et al., 2007) on both superfamily and family level. In this setup, the inclusion of Astral Family HMM predictions only slightly improves accuracy: using AutoSCOP instead of AutoSCOP* in combination with PPA we get 0.6% less accuracy on family level and identical accuracy on superfAMILY and fold level, as most of the additional predictions are covered by PPA.

3.5.2 Inclusion of structure information by combination with Vorolign. Using AutoSCOP together with Vorolign, e.g. for the purpose of assigning a classification to a newly resolved structure, we achieve a clear improvement over Vorolign alone. In other words, on this set, we can correct some false assignments made by structure alignments. Again, using AutoSCOP instead of AutoSCOP* in combination with Vorolign decreases accuracy only by up to 0.6% (on family level).

4 CONCLUSION

AutoSCOP is a simple yet effective sequence-pattern-based approach to SCOP prediction on different SCOP levels. For the domains in the test set with known folds/superfamilies, we achieve sensitivity values of more than 93%. Here, especially the specificity values of ~98% are striking. On the Vorolign test set, AutoSCOP even achieves specificity rates of up to 99.9% (fold level). This means that, if a prediction is made, it is indeed very reliable. A test on CAFASP4 targets also shows that the predictions made by AutoSCOP can provide useful information in blind-test protein structure prediction scenarios.
The combination with PPA underlines the potential of the AutoSCOP approach by improving the sensitivity of our predictions over the best individual method by ~ 4%. On family and superfamily level, this combination even outperforms the structure alignment methods in our comparison. Therefore, AutoSCOP can be used as a filter for template selection and fold or superfamily recognition in addition to alignment-based recognition methods.

The inclusion of unique pattern combinations does not significantly improve the performance over unique patterns alone. One possible reason for this is the high redundancy between the InterPro member databases. In Table 2 we observe that, with the exception of Pfam and SUPERFAMILY, leaving out one database does not change the performance very much, especially on superfamily and fold level. Even after leaving out the SUPERFAMILY database, we still observe sensitivity values well above 80% for these levels.

The low coverage and the low sensitivity of AutoSCOP on family level can be explained by the focus of the pattern databases, many of which concentrate on less fine-grained similarities (an obvious example is the SUPERFAMILY database). This implies that many patterns that are unique on coarse levels can be found in more than one SCOP family, and therefore the prediction of the family is not possible. It seems that most patterns work best on superfamily level, which also explains the similar performance of AutoSCOP on superfamily and fold level, as all unique patterns on superfamily level have to be unique on fold level by definition. Therefore, especially for the family level, inclusion of specialized data sources such as ASTRAL Family HMMs is useful.

One problem for AutoSCOP as well as for many other SCOP predictors is the handling and recognition of domains belonging to new folds, superfamilies or families. As we have seen, many predictions for such cases could be traced back to changes in the SCOP versions. However, sometimes we observe only low sequence identities, and in such cases it remains difficult to discriminate between known and new classifications. Discarding such targets can increase specificity but comes with the loss of many good predictions in the twilight zone of sequence identities. Here, further work is necessary.

The proposed method can easily be extended by including sequence patterns from other data sources, which we have shown here by including predictions from ASTRAL Family HMMs. It is further applicable to any protein domain hierarchy with SCOP being one very popular example. For the time intervals between releases of such hierarchies, reliable predictions of potential protein classifications are important also for proteins with already available structures. The combination with Vorolign (Table 4) shows that there is potential to detect and avoid errors in assignments made on the basis of structure alignments. AutoSCOP may also be a useful additional component for systems like SCOPmap (Cheek et al., 2004) that combine both sequence-based and structure-based predictors into a larger system.

Please note that this approach works on domain sequences. This means that, in order to work properly, multi-domain sequences have to be split into individual domains before applying AutoSCOP. A number of methods are available for this purpose, e.g. our own SSEP-Domain method (Gewehr and Zimmer, 2006, sequence based) or PDP (Alexandrov and Shindyalov, 2003, structure based). In any case, given the positions of patterns on an unsplit amino acid sequence, it is always possible to map potential SCOP classes to regions indicated by these positions.

If an InterProScan run is necessary, the runtime of AutoSCOP was found to be about half the runtime of a PPA run (with included profile generation for the target), but slightly longer than a Vorolign scan as described in (Birzele et al., 2007), using up to a few minutes per target on an AMD Athlon XP with 1.8 GHz. If annotated patterns are available, as it is the case for millions of protein sequences, the whole AutoSCOP process is mainly reduced to a database lookup which can be done in a few seconds.

Therefore, the AutoSCOP approach can directly be applied to map millions of sequences to SCOP for which no structures are available yet, using pattern searches for new sequences or precomputed InterPro data, though in the latter case it may be desirable to also include e.g. an ASTRAL Family HMM search (AutoSCOP*) for family level predictions. Regarding the steady growth of the necessary databases, we expect the power of our simple approach to grow with time, too.

We provide a web server for the AutoSCOP method available at http://www.bio.ifi.lmu.de/autoscop, where users can submit domain sequences in order to obtain SCOP predictions.

ACKNOWLEDGEMENTS

We thank Alessandro Macri for helpful discussions. Niklas von Öhsen kindly provided the PPA software. Fabian Birzele and Stefan Kramer made helpful comments on the manuscript. Part of this work was funded by DFG under grant PROSEQO II (Zi 616/2).

Conflict of Interest: none declared.

REFERENCES


