D398–D401 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 11 October 2007*
doi:10.1093/nar/gkm834

AutoPSI: a database for automatic structural classification of protein sequences and structures

Fabian Birzele*, Jan E. Gewehr and Ralf Zimmer

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

Received August 14, 2007; Revised and Accepted September 21, 2007

ABSTRACT

In protein research, structural classifications of protein domains provided by databases such as SCOP play an important role. However, as such databases have to be curated and prepared carefully, they update only up to a few times per year, and in between newly entered PDB structures cannot be used in cases where a structural classification is required. The Automated Protein Structure Identification (AutoPSI) database delivers predicted SCOP classifications for several thousand yet unclassified PDB entries as well as millions of UniProt sequences in an automated fashion. In order to obtain predictions, we make use of two recently published methods, namely AutoSCOP (sequence-based) and Vorolign (structure-based) and the consensus of both. With our predictions, we bridge the gap between SCOP versions for proteins with known structures in the PDB and additionally make structure predictions for a very large number of UniProt proteins. AutoPSI is freely accessible at http://www.bio.ifi.lmu.de/AutoPSIDB.

INTRODUCTION

Structural classifications of proteins as specified by the SCOP (1) database or others provide a view on the space of protein domains that allows for defining different, carefully curated levels of similarity between domains. They are widely used by applications such as template selection for homology modeling and fold recognition, the generation of representative template sets, analysis of fold and function distributions or target selection for structural genomics.

Unfortunately, these databases require a considerable (manual) effort to be updated and therefore cannot be up-to-date with the currently available structures in the PDB (2) even at the time of new releases. Also, they require the availability of protein structure for the target protein, which leads to the fact that most proteins in UniProt are not assigned to any structural classification.

The Automated Protein Structure Identification (AutoPSI) database provides structure and sequence-based SCOP predictions for newly resolved protein structures stored in the PDB as well as purely sequence-based predictions for millions of proteins stored in UniProt (3). In order to predict SCOP classifications, two recently published methods namely AutoSCOP (4) and Vorolign (5) are used which both have been shown to deliver very reliable predictions for their respective task. Therefore, the first application of the AutoPSI database is to bridge the gap between new SCOP releases for structurally resolved proteins in the PDB and yet unclassified by SCOP by providing reliable predictions for many proteins. Second, we also bridge the gap between protein sequences stored in UniProt and known structures in the PDB by providing sequence-based SCOP classification predictions for those proteins. Using a new and automated approach focusing on high specificity (4), AutoPSI extends already existing databases like Swissmodel (6), Modbase (7), (pre-)SCOP or InterPro (8) and helps to further clarify the protein sequence-structure space.

THE AUTOPSI DATABASE CONTENT AND METHODS

Protein content

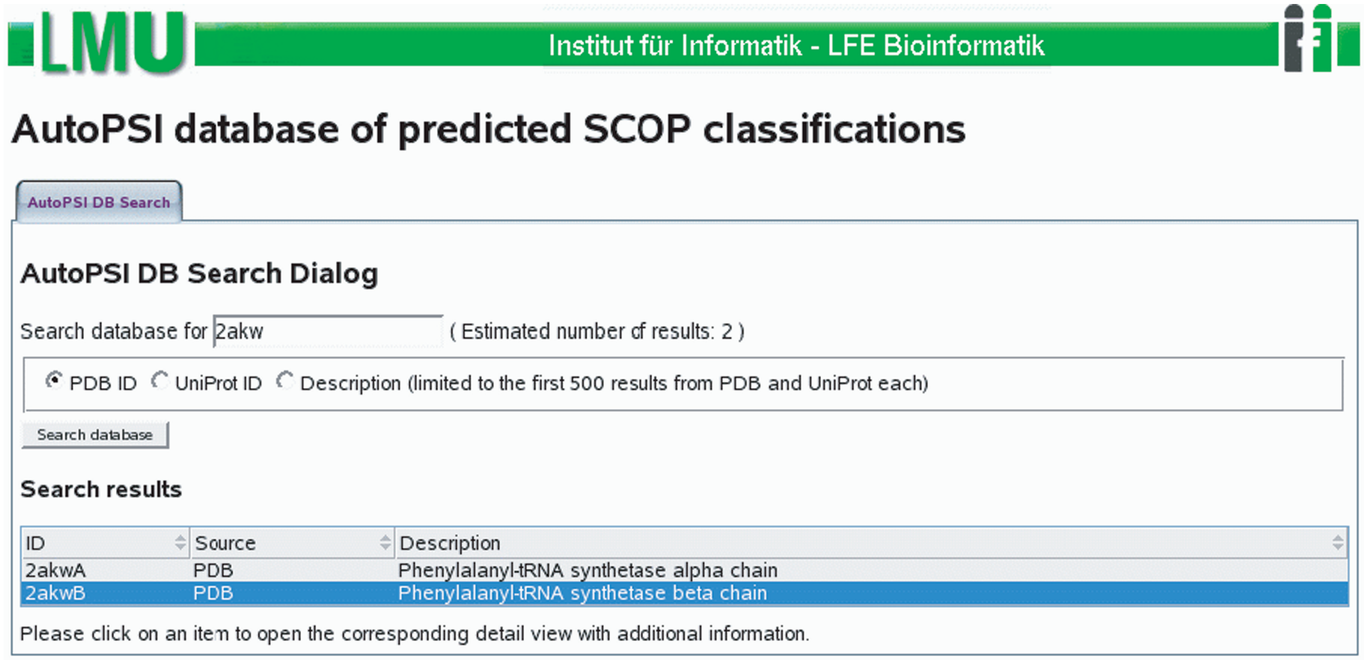
The protein data available in our database is based on PDB and UniProt. In particular, we provide all PDB and UniProt sequences for which we could make SCOP classification predictions or find already existing (pre-) SCOP annotations using any of our methods. Currently, the database contains more than 40 000 unique PDB protein ids containing over 100 000 chains.

Pattern content and AutoSCOP

AutoSCOP identifies patterns of the InterPro member databases, which are unique for a certain level of the

*To whom correspondence should be addressed. Tel: +49 (0) 89 21804064; Fax: +49 (0) 89 21804054; Email: fabian.birzele@bio.ifi.lmu.de

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.



AutoPSI database of predicted SCOP classifications

AutoPSI DB Search Dialog

Search database for 2akw (Estimated number of results: 2)

● PDB ID ○ UniProt ID ○ Description (limited to the first 500 results from PDB and UniProt each)

Search results

ID	Source	Description
2akwA	PDB	Phenylalanyl-tRNA synthetase alpha chain
2akwB	PDB	Phenylalanyl-tRNA synthetase beta chain

Please click on an item to open the corresponding detail view with additional information.

Figure 1. Search dialog and entry page of the AutoPSI database.

SCOP hierarchy (class, fold, superfamily or family). For the current version of the AutoPSI database the ASTRAL (9) 1.69 distribution was scanned with InterProScan (10) on the InterPro databases of version 12.1. Given those unique patterns, SCOP classifications can be made for proteins with unknown classification and even unknown structure. Regions that match a unique pattern are likely to belong to the corresponding level of the SCOP hierarchy (given the training data). Therefore, all PDB sequences in the database were scanned and the identified patterns and their locations were stored. For each PDB sequence, we know the location of the patterns and, based on AutoSCOP, the assigned SCOP classification on each SCOP level (if available) for the specified region. For UniProt sequences, we made use of the corresponding Protein2IPR.dat file, which was downloaded from InterPro (August 2007). The file contains precomputed pattern occurrences on UniProt sequences as computed by InterProScan. Again, whenever we find a pattern for a UniProt sequence, we know its location and annotate a classification if the corresponding pattern is known to be unique for a certain SCOP level. About 2.6 million UniProt proteins harbor unique AutoSCOP patterns.

Structures, domains and Vorolign

For proteins with known structure we annotate the structural domains. If available, we make use of the SCOP definitions directly. In SCOP 1.71, about 27 000 protein structures are structurally classified. If not, we use PDP (11) to automatically assign potential structural domains and then run the Vorolign structural alignment method for these potential domains against a template database of structural domains (the ASTRAL40 1.69), in order to predict the SCOP classification based on structural similarity. The database contains more than 30 000 predictions for about 13 000 unique PDB ids.

Consensus predictions

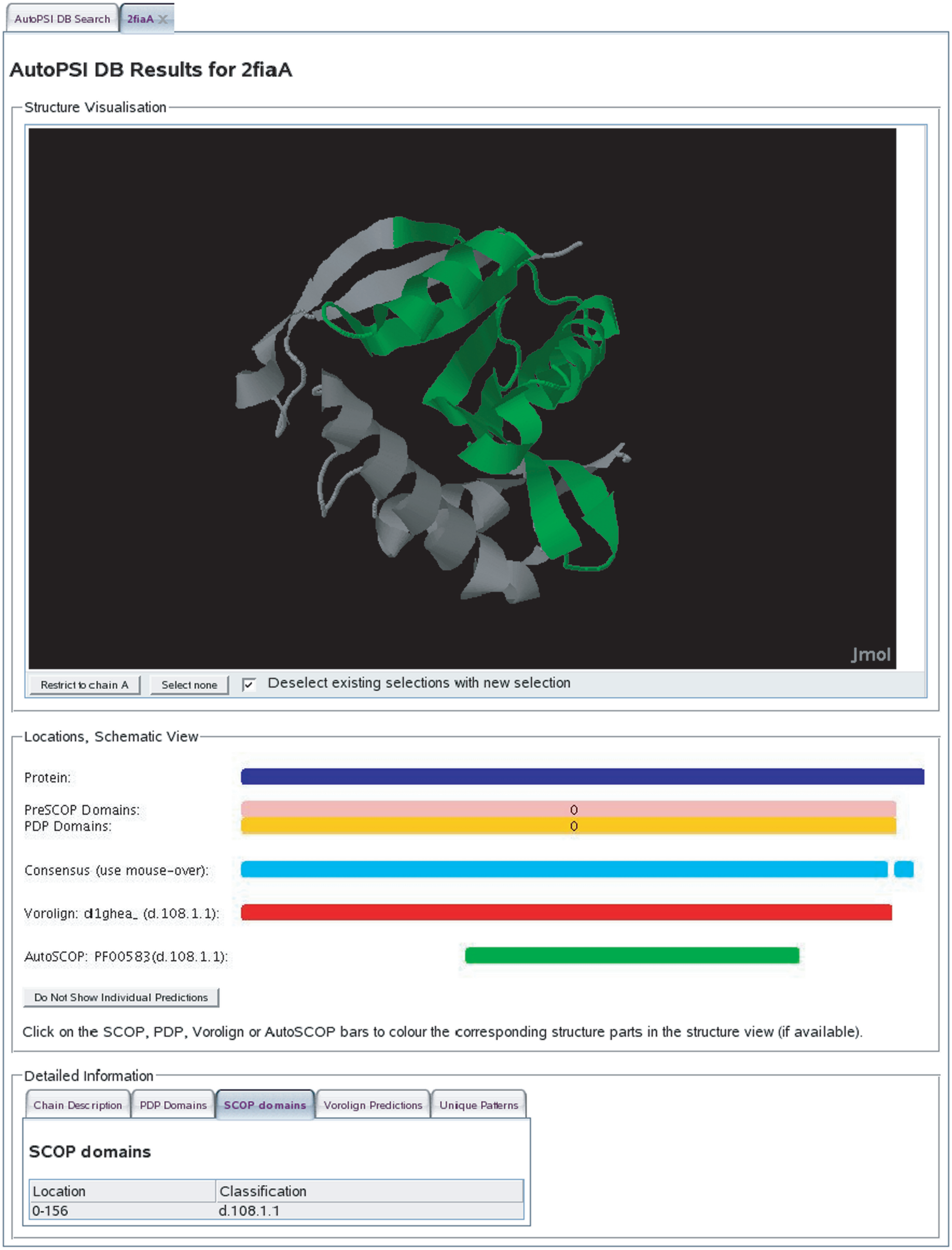
In cases where more than one prediction is available, we compute a simple consensus for the corresponding SCOP classifications (residue-wise by choosing the finest level of agreement between them). As an example, if we have two predictions, namely a.1.1 from position 1 to 200 (delivered by AutoSCOP) and a.1.1.2 (predicted by Vorolign) from position 50 to 150, our consensus yields a prediction of a.1.1 for the regions 1–49 and 151–200, and a prediction of a.1.1.2 for positions 50 to 150. This works also as a control mechanism, because in situations such as finding two completely differing predictions a.1 and b.3, the consensus abstains for the corresponding region and therefore excludes potentially unreliable assignments.

ACCESS TO THE AUTOPSI DATABASE

The data stored in the AutoPSI database is accessible in two different ways. First, we provide a flat file distribution of the database, which contains the individual predictions made for PDB proteins as well as UniProt proteins. Second, the data is accessible by a web interface, which will now be described in some more detail.

Searching the database

The AutoPSI database may be searched for entries by means of PDB identifiers, UniProt identifiers and descriptions. The entry dialog is shown in Figure 1. Search results are presented in a table, and database entries can be selected for detailed inspection in the so-called ‘detail view’.



AutoPSI DB Results for 2fiaA

Click on the SCOP, PDP, Vorolign or AutoSCOP bars to colour the corresponding structure parts in the structure view (if available).

SCOP domains

Location	Classification
0-156	d.108.1.1

Figure 2. Detail view of the AutoPSI database for protein 2fiaA. Patterns matching the protein are shown in green, domain predictions based on PDP are shown in yellow, SCOP assignments in yellow or pink (preSCOP). Pattern locations may be visualized on the structure by clicking on the corresponding regions (as the match of PF00583 on the structure shown). Structures are visualized using Jmol (http://www.jmol.org) if available.

Detail information

For a selected entry, a detail view (for an example see Figure 2) can be requested, which opens as a new tab on the database website. This page allows to view detailed information about SCOP classifications annotated to the proteins, shows the protein structure (if available) using Jmol (http://www.jmol.org). The image below the structure visualization shows the protein, its domains and consensus predictions. By clicking on the button 'Show Individual Predictions', a user can get a larger image which contains detailed information about matching InterPro patterns and Vorolign predictions as well as their locations on the sequence (Figure 2). The locations of all patterns and Vorolign predictions can be visualized on the structure by clicking on the corresponding pattern in the image. Below the overview image and the structure visualization, a box containing tabs for the sequence itself and the outputs of the different predictions mechanisms and annotations found for it provides further information as well as links to external databases.

Conclusion and Future Direction

Structurally classified protein sequences and structures are a useful basis for research in protein structure prediction, and the lack thereof can possibly result in worse performance than necessary. An example is fold recognition and related methods, which often rely on structurally annotated domain templates. We therefore provide a database of predicted SCOP annotations for new PDB entries based on two reliable predictors, namely Vorolign and AutoSCOP. Additionally, we provide sequence-based SCOP classification predictions for about 2.6 million UniProt sequences. The database is a resource that can help in two ways: researchers with an interest in specific proteins may get a clue on the structural classification and, associated with this, possible further properties such as a general function; method developers can use the database to derive and compare larger scale data for their purpose. In future, we plan to integrate more structure prediction methods and pattern sources into the SCOP assignment process. We will also provide additional information for sequence-based predictions of UniProt proteins including sequence alignments with potential templates and others.

ACKNOWLEDGEMENT

Funding to pay the Open Access publication charges for this article was provided by Ludwig-Maximilians-Universität (LMU) München.

Conflict of interest statement. None declared.

REFERENCES

1. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.
2. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
3. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.
4. Gewehr,J.E., Hintermair,V. and Zimmer,R. (2007) AutoSCOP: automated prediction of SCOP classifications using unique pattern-class mappings. *Bioinformatics*, **23**, 1203–1210.
5. Birzele,F., Gewehr,J.E., Csaba,G. and Zimmer,R. (2007) Vorolign – fast structural alignment using Voronoi contacts. *Bioinformatics*, **23**, e205–e211.
6. Kopp,J. and Schwede,T. (2006) The SWISS-MODEL Repository: new features and functionalities. *Nucleic Acids Res.*, **34**, D315–D318.
7. Pieper,U., Eswar,N., Davis,F.P., Braberg,H., Madhusudhan,M.S., Rossi,A., Marti-Renom,M., Karchin,R., Webb,B.M. *et al.* (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.*, **34**, D291–D295.
8. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.
9. Chandonia,J.M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) The ASTRAL compendium in 2004. *Nucleic Acids Res.*, **32**, D189–D192.
10. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
11. Alexandrov,N. and Shindyalov,I. (2003) PDP: protein domain parser. *Bioinformatics*, **19**, 429–430.