Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W79–W83***
doi:10.1093/nar/gkl216

ProSAT2—Protein Structure Annotation Server

R. R. Gabdoulline*, S. Ulbrich, S. Richter and R. C. Wade

Molecular and Cellular Modeling Group, EML Research, Schloss-Wolfsbrunnenweg 33, 69118 Heidelberg, Germany

Received February 14, 2006; Revised and Accepted March 24, 2006

ABSTRACT

ProSAT2 is a server to facilitate interactive visualization of sequence-based, residue-specific annotations mapped onto 3D protein structures. As the successor of ProSAT (Protein Structure Annotation Tool), it includes its features for visualizing SwissProt and PROSITE functional annotations. Currently, the ProSAT2 server can perform automated mapping of information on variants and mutations from the UniProt KnowledgeBase and the BRENDA enzyme information system onto protein structures. It also accepts and maps user-prepared annotations. By means of an annotation selector, the user can interactively select and group residue-based information according to criteria such as whether a mutation affects enzyme activity. The visualization of the protein structures is based on the WebMol Java molecular viewer and permits simultaneous highlighting of annotated residues and viewing of the corresponding descriptive texts. ProSAT2 is available at http://projects.villa-bosch.de/mcm/database/prosat2/.

INTRODUCTION

Suppose you have just read a paper in which it is stated that mutation of residue 249 to alanine results in reduced activity of enzyme X in yeast, and you have found in a database that a variant exists with residue 249 of enzyme X in humans mutated to phenylalanine. Now you are interested in understanding how these mutations may affect enzyme X's stability and function. This means putting this residue-specific data in the context of the protein structure and what it is known about the location of the active site and catalytic residues. To do this, you need to download the protein structure from the Protein Data Bank (PDB), display it with your favorite molecular graphics program, and highlight residue 249. However, it is not this simple. What if the residue 249 referred to in the paper is not the same as that in the PDB structure file? What if the numbering is different in yeast and human enzymes? And where is the active site? It is quite an arduous and sometimes tedious process to address these and other similar issues. The goal of ProSAT2 is to automate this process so that sequence-specific functions or effects of mutations can be studied by simultaneous, interactive exploration of protein structure and functional annotations. This is illustrated by the screenshot of a ProSAT2 session shown in Figure 1.

A flowchart for ProSAT2 is shown in Figure 2. There are two main steps.

The first step is the collection and processing of data about functional residues that are associated with the protein. This is performed by the *ProSAT Miner*. The assembled data and appropriate protein structure files are supplied by the user or downloaded from on-line servers. The residue numbering in the data from different sources is brought into correspondence by performing sequence alignments. A *ProSAT2 annotation* file is written in a format that conforms to the Extended Mark-up Language (xml) (1) standard. SwissProt (2) and PROSITE (3) functional annotations from ProSAT 1 (Protein Structure Annotation Tool) (4) are added to the annotation file.

The second step is the visualization of the acquired residue-specific data on the protein structure. To facilitate handling of the data, which can become voluminous, the *ProSAT2 Annotation Selector* is invoked interactively to allow the user to choose subsets of the data that she/he is interested in visualizing.

In the next sections, we describe the ProSAT Miner, the ProSAT2 annotation file, the ProSAT2 Annotation Selector, and the visualization capabilities of ProSAT2. We then give a description of its current usage for investigating residue-specific data from the UniProt and Brenda databases and end with a discussion of future directions for the development of ProSAT2.

PROSAT MINER

ProSAT Miner retrieves, for a user-specified protein, residue-specific information from a data server. It is a script written in the Python language. As input, ProSAT Miner requires an identifier, for example, UniProt ID, of a protein or the identifier of a protein family (as currently implemented—the Enzyme Commission number, EC number). It accepts an additional input of a list of IDs that refer to structure files in the Brookhaven pdb (5). ProSAT Miner extracts data

*To whom correspondence should be addressed. Tel: +49 6221 533 266; Fax: +49 6221 533 298; Email: razif.gabdoulline@eml-r.villa-bosch.de

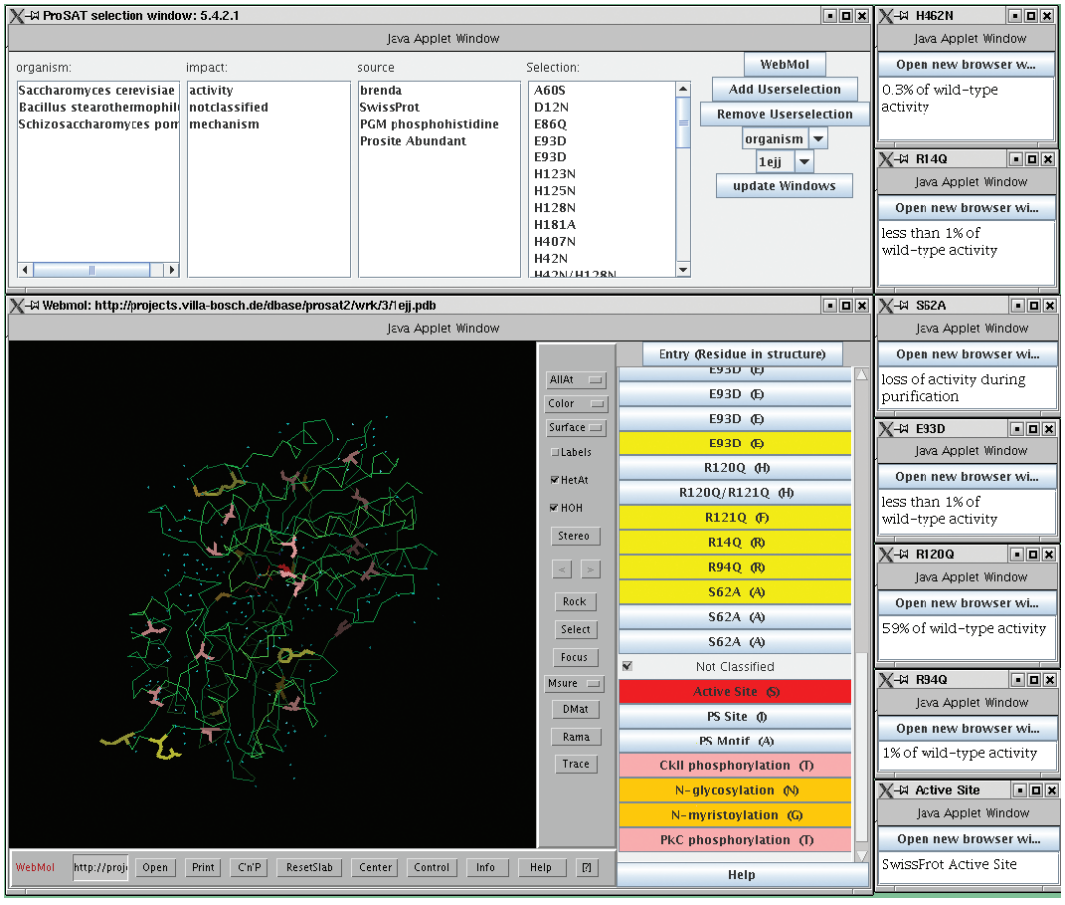

Figure 1. Screenshot showing a ProSAT2 session. At upper left is the ProSAT2 Selector window, at lower left the WebMol 3D structure visualization window. On the right hand side are the windows that pop up after requesting detailed information from the residue-highlighting panel (in the middle with coloured buttons).

about variants or mutations and writes this to an annotation file readable by ProSAT2. This task is achieved by ProSAT Miner in several steps (see flowchart in Figure 2):

(i) *Protein structure retrieval*. Protein structure files are retrieved from the PQS (6) mirror in the ProSAT1 (4) server or from the RCSB database (5). If IDs for the protein structure are not input by the user, or if the user requests it, the script can use the protein structure files (up to a maximum of 10 different structures) that are referred to in the data source for variants or mutations, e.g. as currently implemented, from UniProt (7) or BRENDA (8) database entries.

(ii) *Sequence retrieval from structure files*. The sequences of all polypeptide chains in the structure files are extracted.

(iii) *Residue-specific information retrieval*. ProSAT Miner finds the descriptions of variants or mutants in the source data, extracts them and organizes them into a storage table containing:

(a) mutation name
(b) organism to which the protein belongs
(c) short description of the mutations and their effect
(d) link to a web resource with a more detailed description, e.g. a Pubmed entry.

The mutations are defined by the type of the original wild-type residue, its position number in the sequence, and the new amino acid type. The impact of the mutations is categorized by comparing the description of the mutation in the database with a list of keywords (Table 1). Searching is done with the support for regular expressions in the Python language.

(iv) *Sequence retrieval from UniProt and alignment*. The ProSAT Miner fetches the sequences of the protein from the listed organisms from UniProt (7). As the residue numbering may differ between different publications, it is necessary to first align the mutations to a reference sequence. This is not straightforward and is implemented in the following way to achieve reasonable matching. For a single point mutation, the mutated sequence is filled with placeholders to constrain the mutation to be near the position listed in the database. Mutations that originate from the same publication and are applied to the same protein from the same organism are likely to be based on the same residue numbering scheme. They are aligned starting from the position listed in the database entry but using only the sequence built from the list of mutated residues. Therefore they are subsumed and aligned together to the sequence.

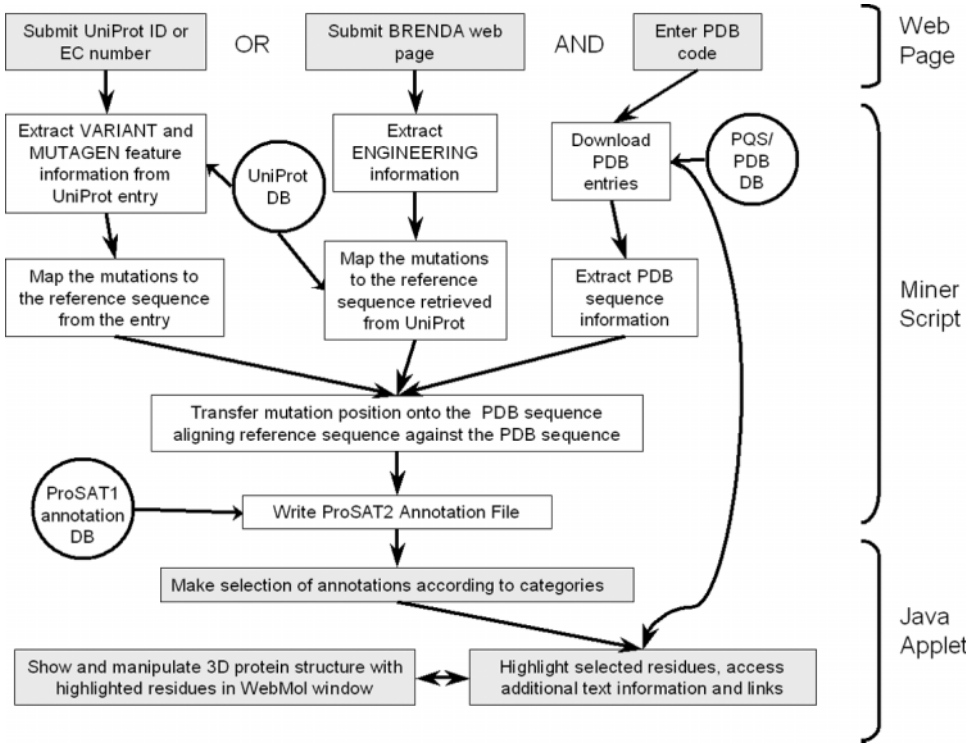

Figure 2. Flowchart illustrating how ProSAT2 extracts residue-specific data from databases and permits the user to sort this and highlight residues on 3D protein structures along with information on their function or the impact of their mutation. The steps with user intervention are in grey. The databases used are shown in circles.

Table 1. Keyword search for the impact of a mutation

Impact type	Keywords
Kinetics	Turnover-number
	K_M
	K_{Cat}
	K_{Cat}
	V_{max}
Mechanism	Kinetic mechanism
	Abnormal kinetics
Activity	Activity
	Loss of function
Expression level	Expression
Stability	Thermostability
	Thermolability
Disease	Lethal
	Severe
	Cancer
	Patient
	Sickness
Polymorphism	Polymorphism
Binding	Binding
	Affinity
	Interaction
	K_D
	K_i
	Association
Unclassified	

(v) *Aligning mutations to the sequence in the structure files*. Once all mutations have been aligned to their reference sequences, they are aligned to the sequences that have been extracted from the structure files (step ii). This is achieved first by aligning the reference sequence to a structure file's sequence and then transferring mutation positions to the structure file's sequence.

(vi) *Writing to annotation file*. The featured residues are written to a ProSAT2 annotation file, along with the name of the protein and the present date.

PROSAT2 ANNOTATION FILE

The transfer of annotations from the ProSAT2 Miner to the visualization components of ProSAT2 is done via an XML file. To support the grouping and selecting feature of the ProSAT2 Selector interface (described below), the annotation file allows for flexible grouping of the annotations in supergroups and subgroups. By default, the ProSAT Miner will create groups for the organisms, the impact of a mutation, and the information source of the annotation (e.g. BRENDA or UniProt). The XML file also supports the transfer of the annotation positions to the 3D structure and the ProSAT2 visualization components.

An annotation format was developed to serve as a basis for 3D visualization and handling a large amount of information on protein mutations. The ProSAT2 Annotation Selector expects as input an XML format annotation with a list of items belonging to an arbitrary number of subgroups, that are organized into supergroups. This makes it easy to dynamically generate selection categories according to supergroups. It is also easy to add new items to the file, e.g. with a script that adds information from a different source. New groups can be created in ProSAT2 Annotation Selector by defining a selection as a new group.

The same type of XML ProSAT2 annotation file is also used for another feature of the ProSAT2 server: the uploading of user-created annotations together with PDB structure files for visualization. A detailed format description together with the DTD of the ProSAT2 annotation file can be found on the ProSAT2 website.

PROSAT2 ANNOTATION SELECTOR

In the ProSAT2 Selector Window (upper left window in Figure 1), a selection box appears for each supergroup (category) specified in the annotation file. Its entries are the subgroups that the supergroup contains. The last field shows which items are currently selected (all of the annotation file's items upon startup). Once a set of mutations have been selected, the user can choose the structure file to be displayed with annotations on the right-hand side of the Selector Window. The user can also choose a supergroup that can be used to sort items. This sorting will appear in a 3D visualization window. For example, it may be useful to sort mutations by the organism they have been applied to.

Selections of annotations are organised in the following way. Selection of the entries of one box will mark the items that belong to the selected subgroups of the supergroup that the box represents The selections from different boxes are linked by a logical AND. This means that the intersection of the sets of selected items will be the resulting choice, e.g. choosing *Saccharomyces cerevisae* in the organism category, and 'activity' in the impact section, will result in the selection of mutations that were applied to yeast and resulted in changed activity of the protein (Figure 1). Once created, the selection can be used to annotate the structure in the 3D visualization, or be stored permanently in the annotation file. This allows the restoration of the selections after the application has closed. The storage process will create the category 'user selection', along with an entry whose name the user must specify. This will dynamically alter the interface by adding a new selection box, representing the new category. Stored user selections can also be removed with the help of this interface.

VISUALIZATION OF PROTEIN STRUCTURES AND ANNOTATION TEXTS

Webmol (9) has been chosen as the 3D structure visualization tool for ProSAT2 because of the richness of its graphical features. Webmol is written in Java and can be run either as a stand-alone program or as a browser applet. It can load protein structure files in PDB format. For ProSAT2, the Webmol applet was enhanced with a separate selection frame that enables the user to highlight the functional residues specified in the annotation file. Several WebMol instances can be created with the ProSAT2 Selector, each for the selection made before starting 3D visualization. The fact that it is possible to have several WebMol instances open at the same time enables the user to compare mutations in different proteins of the same protein family.

The WebMol window has an additional embedded control panel. The control panel is populated with buttons representing all the items that result from the user selection, and are sorted according to the user choice in the Selector Window. Only the items that have a valid *range* entry and thus were successfully mapped to a given structure will appear as buttons in the WebMol window.

The control panel offers buttons, each representing an item defined in the annotation file. A click with the left mouse button on one of these will highlight the associated mutated residue. By default, the coloring depends on the charge of the residue. The highlighting color can be selected in the context menu, which shows up by pressing the right mouse button. The context menu also gives access to the stored descriptions and links to the publications about the mutation. A new window containing this data will open by selecting the 'Detail' entry. The buttons are labeled by the item's name, which is normally the type of original residue and its numbered position in the sequence, as well as the new amino acid code e.g. R124K means the amino acid arginine at position 124 has been changed to lysine. Additionally, the actual residue at this position on the structure is shown in brackets. This is useful, because if the mutation is displayed on a different organism, the original residue quite often differs from the original amino acid.

IMPLEMENTATION OF PROSAT2

Parser tools have been written for the UniProt and BRENDA databases. These hold, respectively, naturally occurring variant or mutagenesis information and protein engineering information. Both of these databases have a large amount of data about residue modification or variation. Out of 4599 entries in BRENDA (release 5.2, January 2006), there are 1372 entries having engineering information (average 11 mutations and up to 134 for the K^+-transporting ATPase). 205780 entries of SwissProt (release 47) can be related to 2260 enzymes with uniquely assigned EC numbers. Among these, 809 have variant information (at least one feature keyword VARIANT), with an average of 17 variants per enzyme and up to 531 for a Tyrosine kinase.

Support of UniProt and BRENDA is implemented as separate modules of the ProSAT2 Miner. These modules differ only up to the step of aligning mutation or variance positions to the reference sequence. In the case of the BRENDA parser module, the reference sequence is retrieved from the UniProt database using the reference in the BRENDA page and the position of the mutation is aligned to this sequence. When the VARIANT or MUTAGEN feature information is taken from UniProt, the position of the mutation is not aligned to the reference sequence since for UniProt, the sequence information is taken from the same database entry as the variant information. To avoid the need to retrieve all UniProt entries with a given EC number, a query against the EBI server (http://srs.ebi.ac.uk) is made to restrict the entries to those that contain VARIANT or MUTAGEN feature information. Support of other databases can be implemented by writing additional parser modules to ProSAT Miner.

The user needs to provide the ProSAT2 web server with two types of input: one containing residue-specific information (UniProt ID or accession code, EC number or BRENDA web page) and the other containing information about the related protein structure (PDB ID code). In cases when one input can be derived from the other (PDB ID code found in UniProt entry or BRENDA web page, UniProt entry for a given PDB ID code retrieved from the EBI server) the server

tries to proceed with only one input giving the user a possibility to select missing information later.

CONCLUSIONS

The ProSAT2 server allows the user to access various combinations of residue-specific information mapped to 3D structures, i.e. a user-defined structure file can be annotated from a user-defined residue-specific information source. Currently, two very different types of information sources are supported: the mutation information from BRENDA and the variant information from UniProt. Other types of annotations are currently supported on a user-supplied basis, i.e. the user submits the annotation file in ProSAT2 format. This annotation format is intended to be used in the application of ProSAT2 to other databases holding residue modification/variation data—namely, separate parsers can be written to extract data from the databases and map residues to relevant structures. There are many mutant databases, e.g. PMD (10), MutDB (11) or mutation data resulting from text mining activities (12,13). These all can gain from showing the modification sites not only on the sequence but also on the protein 3D structure.

ACKNOWLEDGEMENTS

We thank Christopher Baker, Rene Witte, Jiri Damborsky, Fridtjof Feldbusch and Dirk Walther for helpful discussions during the development of ProSAT2. We gratefully acknowledge financial support from the Klaus Tschira Foundation and the Center for Modeling and Simulation in the Biosciences (BIOMS), Heidelberg, as well as NATO Collaborative Linkage Grant 980504. Funding to pay the Open Access publication charges for this article was provided by BIOMS and Klaus Tschira Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Bray,T., Paoli,J., Sperberg-McQueen,C.M., Maler,E., Yergeau,F. and Cowan,J. (2004) Extensible Markup Language (XML) 1.1. W3C Recoomendation. .
2. Boeckmann,B., Blatter,M.C., Famiglietti,L., Hinz,U., Lane,L., Roechert,B. and Bairoch,A. (2005) Protein variety and functional diversity: Swiss-Prot annotation in its biological context. *C. R. Biol.*, **328**, 882–899.
3. Hugo,N., Bairoch,A., Bullard,V., Cerruti,L., De_Castro,E., Langendijk-Genevaux,P.S., Pagni,M. and Sigrist,C.J. (2006) The PROSITE database. *Nucleic Acids Res.*, **34**, D227–230.
4. Gabdoulline,R.R., Hoffmann,R., Leitner,F. and Wade,R.C. (2003) ProSAT: functional annotation of protein 3D structures. *Bioinformatics*, **19**, 1723–1725.
5. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L., Feng,Z. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.*, **33**, D233–D237.
6. Henrick,K. and Thornton,J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.*, **23**, 358–361.
7. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.
8. Schomburg,I., Chang,A., Ebeling,C., Gremse,M., Heldt,C., Huhn,G. and Schomburg,D. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.*, **32**, D431–D433.
9. Walther,D. (1997) WebMol—a Java based PDB viewer. *Trends Biochem. Sci.*, **22**, 274–275.
10. Kawabata,T., Ota,M. and Nishikawa,K. (1999) The protein mutant database. *Nucleic Acids Res.*, **27**, 355–357.
11. Danzer,J., Moad,C., Heiland,R. and Mooney,S. (2005) MutDB services: interactive structural analysis of mutational data. *Nucleic Acids Res.*, **33**, W311–W314.
12. Horn,F., Lau,A.L. and Cohen,F.E. (2004) Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. *Bioinformatics*, **20**, 557–568.
13. Baker,C.J.O. and Witte,R. (2006) Mutation mining—a prospector's tale. *Inf. Syst. Front.*, **8**, 47–57.