D164–D168 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki027

PSORTdb: a protein subcellular localization database for bacteria

Sébastien Rey, Michael Acab, Jennifer L. Gardy, Matthew R. Laird, Katalin deFays[1], Christophe Lambert[1] and Fiona S. L. Brinkman*

Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, Canada V5A 1S6 and [1]U.R.B.M., F.U.N.D.P Rue de Bruxelles, 61, B-5000, Namur, Belgium

Received August 12, 2004; Accepted September 21, 2004

ABSTRACT

Information about bacterial subcellular localization (SCL) is important for protein function prediction and identification of suitable drug/vaccine/diagnostic targets. PSORTdb (http://db.psort.org/) is a web-accessible database of SCL for bacteria that contains both information determined through laboratory experimentation and computational predictions. The dataset of experimentally verified information (~2000 proteins) was manually curated by us and represents the largest dataset of its kind. Earlier versions have been used for training SCL predictors, and its incorporation now into this new PSORTdb resource, with its associated additional annotation information and dataset version control, should aid researchers in future development of improved SCL predictors. The second component of this database contains computational analyses of proteins deduced from the most recent NCBI dataset of completely sequenced genomes. Analyses are currently calculated using PSORTb, the most precise automated SCL predictor for bacterial proteins. Both datasets can be accessed through the web using a very flexible text search engine, a data browser, or using BLAST, and the entire database or search results may be downloaded in various formats. Features such as GO ontologies and multiple accession numbers are incorporated to facilitate integration with other bioinformatics resources. PSORTdb is freely available under GNU General Public License.

INTRODUCTION

Identification of a bacterial protein's subcellular localization (SCL) provides valuable clues regarding its biological function. For example, surface-exposed or secreted proteins are of primary interest due to their potential as vaccine candidates, diagnostic agents (environmental or medical) and the ease with which they may be accessible to drugs (1–3). Computational SCL analysis of the growing number of completed bacterial genomes or individual proteins allows researchers to screen for vaccine/drug candidates, automatically annotate gene products or select proteins for further study. We have previously developed an SCL predictor for Gram-positive and Gram-negative bacterial proteins called PSORTb [version 2.0; (4)] and here we describe the related database PSORTdb, which contains proteins of experimentally verified localization used in training of multiple SCL predictors, as well as PSORTb's SCL predictions for proteins associated with complete genome sequences.

A database of protein SCL for bacteria is important for numerous reasons: a large, high-quality dataset is required to develop an accurate protein SCL predictor, since a well-curated dataset is a critical component of any machine learning method. Through incorporation of such a dataset into a database, versioning of datasets can be better controlled and researchers can more easily and accurately cite their data source. Novel features influencing protein SCL can also be discovered from analysis of the database contents and its associated additional annotations. An SCL database should also clearly delineate what information is based on sequence similarity or computational predictions (including what computational prediction was used), versus information determined by laboratory experimentation. Such clarification of the information source helps researchers avoid training new computational methods with data that was previously theoretically predicted, a problem which may contribute to a propagation of errors, mispredictions and biases in subsequent analyses.

To date, the most popular source of protein SCL annotations is Swiss-Prot (5) (recently integrated into UniProt (6): the Universal Protein knowledgebase). This source has been widely used by several SCL prediction methods, such SubLoc (7) and its associated database DBSubLoc (8), Proteome

*To whom correspondence should be addressed. Tel: +1 604 291 5646; Fax: +1 604 291 5583; Email: brinkman@sfu.ca

Analyst (9), and LOCkey (10). Although Swiss-Prot is an excellent resource, it is not geared specifically to the annotation of protein SCL. Therefore, annotation quality/confidence scoring by Swiss-Prot only more generally refers to the annotation as a whole. For example, a protein may be annotated as having a particular function and the annotation noted as being 'potential', 'probable' or 'by similarity', however this confidence classification is primarily oriented to the confidence of protein function annotation, rather than SCL. Furthermore, SCL terminology for bacteria is also not always consistent or may be confusing (e.g. 'membrane-bound' or 'membrane-associated').

In most databases, multiple SCLs for one protein are usually not mentioned. For example, protein Q05740 of Swiss-Prot is annotated as 'Periplasmic. Anchored to the cytoplasmic membrane via its N-terminal signal-like sequence, spans the periplasm'. However, it is only referred to as a membrane protein in the DBSubLoc database, which used a Swiss-Prot SCL annotation keyword-based system.

Our new PSORTdb resource addresses all of these issues and incorporates flexible search and output capability, open accessibility to data, and use of various other database identifiers, including GO terms (11), to facilitate incorporation with other bioinformatics resources. It is composed of two parts: ePSORTdb—a dataset of proteins of known SCL (verified by laboratory experimentation) and cPSORTdb—a second dataset of computational predictions of SCL made by PSORTb (4). This database, through its focus on both web-accessible flexibility and use of standard identifiers, will be of use to both researchers with specific questions about a given protein subset, and those wishing to continually incorporate high-quality, continually updated SCL information into their own database schema or SCL prediction research.

CONSTRUCTION OF DATASETS

Manually curated dataset: ePSORTdb

We have developed a high-quality dataset of proteins of known SCL (determined by laboratory experimentation), which is currently the largest dataset of its kind available to date for bacterial proteins. Protein SCL annotations from Swiss-Prot served as a starting point to build the dataset. By automatically parsing bacterial proteins from Swiss-Prot (5), an initial set of annotated proteins were retrieved. To reduce the dataset that would need to be manually curated, any protein whose annotation was noted as 'potential', 'probable', or 'by similarity' by Swiss-Prot was discarded. The resulting dataset was then manually curated through a search of the literature (using PubMed) to verify that the SCL proposed in the Swiss-Prot annotations had been experimentally verified through a laboratory experiment. Often abstracts of publications provided adequate explanation to confirm the SCL experimental determination, although in some cases, this information was only accessible in the full publication. This indicates that text mining of abstracts to obtain such information is not adequate. This procedure was initially performed with release 39 of Swiss-Prot and updated later with release 40.29.

To expand ePSORTdb, other literature sources were used [e.g. microbiology textbooks (12–14) and reference articles] in order to find other proteins for which SCL laboratory determination was reported. In total, 2165 proteins were included in the resulting ePSORTdb dataset.

Table 1. Total number of proteins for each single and multiple SCL site in the ePSORTdb dataset

	Gram-negative	Gram-positive
Single SCL site		
Cytoplasmic (C)	278	194
Cytoplasmic membrane (CM)	309	103
Periplasmic (P)	276	N/A
Outer membrane (OM)	391	N/A
Cell wall (CW)	N/A	61
Extracellular (EC)	190	181
Multiple SCL sites		
C/CM	16	15
CM/P	51	N/A
P/OM	2	N/A
OM/EC	78	N/A
CM/CW	N/A	20
Total	1591	574

Several single and multiple bacterial localization sites are recognized by ePSORTdb. For Gram-negative bacteria, five single sites are included: cytoplasm (C), cytoplasmic membrane (CM), periplasm (P), outer membrane (OM) and extracellular (EC) as well as four multiple localization sites: C/CM, CM/P, P/OM and OM/EC. For Gram-positive bacteria, four single sites are included: cytoplasm (C), cytoplasmic membrane (CM), cell wall (CW) and extracellular (EC); and two multiple localization sites: C/CM and CM/CW. The number of proteins for each single and multiple SCL site for Gram-positive and Gram-negative bacteria is listed in Table 1. Each single SCL term is associated with a unique Gene Ontology (GO) (11) identifier, while multiple SCL localization sites are a combination of their single associated SCL site GO identifiers.

Computationally predicted dataset: cPSORTdb

A total of 140 completely sequenced microbial genomes (96 Gram-negative and 44 Gram-positive organisms) from NCBI's Genomic Biology collection were analyzed by PSORTb v.2.0 (4), and detailed results were assembled together to form the cPSORTdb dataset. Currently, PSORTb is the most precise bacterial localization prediction tool available (4,9), with a measured classification precision of 96% for both Gram-negative and Gram-positive bacteria. Briefly, this predictive tool relies on a series of analytical modules which are designed to identify typical sequence features known to correlate with specific localizations. Based on this information, the program either generates a prediction of the possible localization site for the submitted protein or returns a result of 'unknown' if no accurate prediction can be generated. Thus 277 779 Gram-negative proteins and 133 250 Gram-positive proteins were analyzed by PSORTb and a single SCL prediction was returned for 136 215 and 99 300 of these proteins, respectively. This represents an average coverage of 49.0% for proteins deduced from Gram-negative bacterial genomes and 74.5% for proteins from Gram-positive bacterial datasets. Note that currently PSORTb only flags the possibility of a multiply localized protein rather than returning the two specific sites. The total number of cPSORTdb proteins of each SCL is summarized in Table 2.

Table 2. Total number of proteins for each PSORTb (version 2.0) computationally predicted SCL site in the cPSORTdb dataset

SCL site	Gram-negative	Gram-positive
Cytoplasmic	94 592	68 948
Cytoplasmic membrane	47 284	26 586
Periplasmic	5 536	N/A
Outer membrane	6 097	N/A
Cell wall	N/A	1 132
Extracellular	1 203	4 228
Potential multiple SCLs	5 776	1 064
Total	160 488	101 958

Since the same protein sequence may appear in both cPSORTdb and ePSORTdb, links between both datasets were manually examined to ensure pairs of similar sequences belong to the same organism. In ePSORTdb, organisms are defined at both the genus and species level, because strain information is frequently not available. cPSORTdb, however, contains further information regarding the strain. For this reason, linking was performed at the species level. ePSORTdb contains 1301 proteins that match at this level to proteins in cPSORTdb.

Many other excellent SCL computational prediction methods for bacteria are currently available, such as Proteome Analyst (9) or CELLO (15). Unfortunately precomputed analyses of whole genomes by such programs are not yet available. However, if such precomputed data becomes accessible we will incorporate it into cPSORTdb, upon permission of the associated authors of the programs, and will clearly acknowledge their contribution as a collaboration on the PSORTdb website. In the meantime, PSORTb predictions do remain the most precise available and represent a high-quality computational analysis. We will continue to update this dataset and make new versions of these predictions available, while continuing to make the older versions of predictions accessible through this database.

DATABASE SCHEMA

The PSORTdb database consists of the two independent, but linked, databases ePSORTdb and cPSORTdb. Although they essentially appear to be simple lists of proteins and their SCL, their differences in size and content that have made them difficult to merge. For example, module predictions and SCL scores exist only in cPSORTdb, while only proteins of ePSORTdb have literature references. Furthermore, as mentioned above, protein organisms are defined differently between cPSORTdb and ePSORTdb. Thus it was not possible to merge both datasets. However, some cPSORTdb proteins have a link (as explained above) to the ePSORTdb dataset providing more information on them, such as literature references. A more comprehensive look at the data fields is presented below.

NCBI's GI number (16) is used as the primary identifier for proteins in both datasets. Because of GI's popularity as an identifier, this also facilitates linkage to other databases. When available, the following data fields enhance protein identification: protein name, gene name, alternate protein and gene names, Swiss-Prot accession number (only for ePSORTdb) and RefSeq accession number. The addition of these fields makes it more likely that the proteins can be easily found through a user's query and linked to supplementary databases. Other fields further define the proteins in a broader sense: source organism name, phylum, class, NCBI taxonomy identifier (17), chromosome accession identifier of NCBI (only for cPSORTdb) and Gram stain class (positive or negative). In cases where the same protein from the same species is present in both datasets, an additional field provides a link between both entries of ePSORTdb and cPSORTdb. Finally, amino acid sequences and their length are also made available.

Table 3. Fields in ePSORTdb, the database of proteins of experimentally verified SCL

Common[a]	Specific
GI number[b]	Experimental subcellular localization (terse)[b]
Swiss-Prot accession number	Experimental subcellular localization (verbose)[b]
Protein name[b]	GO accession ID[c]
Alternate protein name	GO accession definition[c]
Gene name[b]	PubMed ID reference
Alternate gene name	Reference title
Taxonomy ID (from NCBI)	ISBN number reference
Organism[b]	WWW reference
Phylum	Reference comments
Class	Reference summary[b,d]
Gram stain	
Amino acid sequence	
Sequence length	
cPSORTdb GI link	

[a]Common fields are those that are shared with cPSORTdb.
[b]Field displayed by default in the result page.
[c]GO references of the experimental SCL.
[d]Includes all fields relevant to references (i.e. PubMed ID, title, ISBN number, WWW and comments).

Experimentally verified SCLs contained in ePSORTdb are accessible in three formats: a terse definition of the SCL which is machine readable, its associated GO identifier number and a verbose definition (e.g. CytoplasmicMembrane; 0005886; cyptoplasmic membrane integral membrane protein). Additional literature references like PubMed identifiers (PMID), book titles and associated ISBN numbers are stored in the ePSORTdb dataset.

PSORTb prediction results in cPSORTdb are classified into three groups:

(i) predicted SCL site, its GO identifier and its score;
(ii) prediction scores for each possible site;
(iii) detailed results from each analytical module.

Tables 3 and 4 provide a complete list of fields available for ePSORTdb and cPSORTdb, respectively.

DATABASE ACCESS AND WEB INTERFACE

PSORTdb's data is housed in a MySQL database. Using PHP and JavaScript, the web database application (freely accessible at http://db.psort.org/) was developed allowing access to this data without prior knowledge of SQL, relational databases and specifics of the PSORTdb database schema. The browsing and dynamic textbox features of the web interface also make it easier for the user to search the data if they are unfamiliar with how the data is stored. Each dataset has three searching tools

Table 4. Fields in cPSORTdb, the database of proteins of computationally predicted SCL

Common[a]	PSORTb module predictions details
Chromosome accession ID (from NCBI)[b]	SCL-BLAST predicted localization
GI number[b]	SCL-BLAST details
RefSeq accession ID[b]	Motif predicted localization
Protein name[b]	Motif details
Gene name[b]	OMPMotif predicted localization
Alternate gene name[b]	OMPMotif details
Taxonomy ID (from NCBI)	HMMTOP predicted localization
Organism[b]	HMMTOP details
Phylum	SignalP predicted localization
Class	SignalP details
Gram stain	Profile predicted localization
Amino acid sequence	Profile details
Sequence length	SCL-BLASTe predicted localization
ePSORTdb GI link	SCL-BLASTe details
PSORTb predicted localization	CytoSVM predicted localization
PSORTb prediction version	CMSVM predicted localization
Predicted localization[b]	PPSVM predicted localization
GO accession ID[c]	CWSVM predicted localization
GO accession definition[c]	OMSVM predicted localization
Predicted localization score[b]	ECSVM predicted localization
PSORTb prediction scores	SublocC predicted localization
Cytoplasmic score	
Cytoplasmic membrane score	
Periplasmic score	
Cell wall score	
Outer membrane score	
Extracellular score	

[a]Common fields are those that are shared with ePSORTdb.
[b]Field displayed by default in the result page.
[c]GO references of the predicted localization.

that all perform in the same manner (but simply act on different datasets).

Text search tool

One or more keywords or other values suitable for a given field can be used to query the database. The keywords are searched against one or more data fields. Boolean operators are available to the user to make complex queries. A dynamic textbox displays a description and/or example of what type of text can be entered for that particular search field. This feature was designed to be an aid to users in choosing their queries. For example, if a user wishes to search under a localization field, all possible localizations are presented in a dynamic textbox that the user can then choose from, to ensure they spelt the term correctly.

Browse tool

This tool allows the user to explore the dataset in a hierarchical fashion similar to browsing the NCBI Taxonomy Database (17) or Gene Ontologies (11). The text used to populate the browsing function is dynamically generated from the MySQL database. The tool is very flexible and permits exploration of the data by SCL, phylum, class, Gram stain and genome in every possible logical combination.

BLAST search

Sequences may be submitted and searched against the database using a BLAST search (18), which searches against a file system rather than the MySQL database. FASTA formatted sequences are currently required as input. This tool facilitates analysis of SCL for those proteins not in the database, since there has been recent evidence suggesting that SCL may be more conserved than originally thought (19).

Text search and Browsing produce an HTML table of results that can be viewed page by page. Initially, a default set of fields are displayed but there are numerous options available that will allow the user to customize their result listing. The user can change the number of records viewed per page, simultaneously sort the table on up to three fields (in ascending or descending order), choose which fields they would like displayed and rearrange the order of the fields. In addition to viewing the results as an HTML document on a web browser, the user may also download the data as a tab-delimited file or a FASTA formatted file.

From the result list and the BLAST search, the user can click on the protein (via the GI accession number) to obtain detailed annotations. The annotation page displays all the information made available from the result list. It also provides the protein's amino acid sequence, links to the protein historical data (e.g. between PSORTb versions in cPSORTdb) and links to external databases. Due to its use of common accession numbers, such as GI or RefSeq accession number, it is easily linked to other existing databases.

Lastly, a web form is also available through which researchers can submit proposed updates/corrections to the database (subject to manual review). We feel this is an important component of this database, to facilitate researchers' participation and inclusion of their data. We have opted to make this form as simple as possible, to encourage participation. However, we will not depend on such submissions to maintain this database, but rather will be continuing to use literature search approaches, including text mining and review of complete papers, to increase the number of proteins in this database over time.

PSORTDB: MANY USES

Applications of PSORTdb are numerous in both bioinformatics and related biology fields. Individual researchers wishing to identify the SCL of a given protein or proteins can easily find such information in our database, through its flexible, intuitive web interface. The database may also be useful for microbiologists wishing to identify targets in bacterial genomes for surface proteins for environmental diagnostics, medical diagnostics, vaccines, antimicrobials and other uses. Information contained in the PSORTdb database can also assist in the annotation of newly sequenced bacterial genomes, and proteomic researchers may also utilize such information to formulate experiments of proteomes or subproteomes.

Our datasets can also be useful for bioinformaticians developing new SCL predictors. Versions of the experimentally based dataset have been previously used by others to train automated SCL predictors for bacterial proteins such as CELLO (15), Proteome Analyst (9) or PRED-TMBB (20). This database, which will be continually updated, allows us to provide a central access point for researchers wishing to obtain all, or a queriable subset of, this valuable dataset. We also permit users to download the entire database, for those bioinformaticians who may require access to continually updated, high-quality SCL annotations for incorporation into their own databases directly.

By incorporating flexible and open access functionality, with both curated and computationally derived data, this resource should be useful to many. However, in the future, we hope to extend this database further in terms of its functionality, what organisms it covers, and how much literature is referenced, as we aim to continue to expand PSORTdb into an increasingly valuable resource.

ACKNOWLEDGEMENTS

This work was funded by Natural Sciences and Engineering Research Council of Canada (NSERC) with some additional funding provided by the Functional Pathogenomics of Mucosal Immunity (FPMI) Project which is supported by Genome Prairie and Genome BC of Genome Canada, and Inimex Pharmaceuticals. S.R. is a Swiss National Science Foundation Scholar. J.L.G. and F.S.L.B. are a Michael Smith Foundation for Health Research Trainee and Scholar, respectively.

REFERENCES

1. Allan,E. and Wren,B.W. (2003) Genes to genetic immunization: identification of bacterial vaccine candidates. *Methods*, **31**, 193–198.
2. Mora,M., Veggi,D., Santini,L., Pizza,M. and Rappuoli,R. (2003) Reverse vaccinology. *Drug Discov. Today*, **8**, 459–464.
3. Paine,K. and Flower,D.R. (2002) Bacterial bioinformatics: pathogenesis and the genome. *J. Mol. Microbiol. Biotechnol.*, **4**, 357–365.
4. Gardy,J.L., Spencer,C., Wang,K., Ester,M., Tusnady,G.E., Simon,I., Hua,S., deFays,K., Lambert,C., Nakai,K. *et al.* (2003) PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res.*, **31**, 3613–3617.
5. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
6. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H.Z., Lopez,R., Magrane,M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
7. Hua,S.J. and Sun,Z.R. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.
8. Guo,T., Hua,S.J., Ji,X.L. and Sun,Z.R. (2004) DBSubLoc: database of protein subcellular localization. *Nucleic Acids Res.*, **32**, D122–D124.
9. Lu,Z., Szafron,D., Greiner,R., Lu,P., Wishart,D.S., Poulin,B., Anvik,J., Macdonell,C. and Eisner,R. (2004) Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics*, **20**, 547–556.
10. Nair,R. and Rost,B. (2002) Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics*, **18** (Suppl. 1), S78–S86.
11. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
12. Fischetti,V.A., Novick,R.P., Ferretti,J.J., Portnoy,D.A. and Rood,J.I. (2000) *Gram-Positive Pathogens.* ASM Press, Washington, DC.
13. Sonenshein,A.L., Hoch,J.A. and Losick,R.M. (2001) Bacillus subtilis *and Its Closest Relatives: From Genes to Cells.* ASM Press, Washington, DC.
14. Neidhardt,F.C., Curtiss III,R., Ingraham,J.L., Lin,E.C.C., Low,K.B., Magasanik,B., Reznikoff,W.S., Riley,M., Schaechter,M. and Umbarger,H.E. (eds) (2004) Escherichia coli *and* Salmonella: *Cellular and Molecular Biology.* ASM Press, Washington, DC.
15. Yu,C.S., Lin,C.J. and Hwang,J.K. (2004) Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci.*, **13**, 1402–1406.
16. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2000) GenBank. *Nucleic Acids Res.*, **28**, 15–18.
17. Wheeler,D.L., Chappey,C., Lash,A.E., Leipe,D.D., Madden,T.L., Schuler,G.D., Tatusova,T.A. and Rapp,B.A. (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **28**, 10–14.
18. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
19. Nair,R. and Rost,B. (2002) Sequence conserved for subcellular localization. *Protein Sci.*, **11**, 2836–2847.
20. Bagos,P., Liakopoulos,T., Spyropoulos,I. and Hamodrakas,S. (2004) A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins. *BMC Bioinformatics*, **5**, 29.