doi:10.1093/nar/gkq1093

PSORTdb—an expanded, auto-updated, user-friendly protein subcellular localization database for Bacteria and Archaea

Nancy Y. Yu, Matthew R. Laird, Cory Spencer and Fiona S.L. Brinkman*

Department of Molecular Biology & Biochemistry, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada

Received September 15, 2010; Revised October 15, 2010; Accepted October 17, 2010

ABSTRACT

The subcellular localization (SCL) of a microbial protein provides clues about its function, its suitability as a drug, vaccine or diagnostic target and aids experimental design. The first version of PSORTdb provided a valuable resource comprising a data set of proteins of known SCL (ePSORTdb) as well as pre-computed SCL predictions for proteomes derived from complete bacterial genomes (cPSORTdb). PSORTdb 2.0 (http://db.psort.org) extends user-friendly functionalities, significantly expands ePSORTdb and now contains pre-computed SCL predictions for all prokaryotes—including Archaea and Bacteria with atypical cell wall/membrane structures. cPSORTdb uses the latest version of the SCL predictor PSORTb (version 3.0), with higher genome prediction coverage and functional improvements over PSORTb 2.0, which has been the most precise bacterial SCL predictor available. PSORTdb 2.0 is the first microbial protein SCL database reported to have an automatic updating mechanism to regularly generate SCL predictions for deduced proteomes of newly sequenced prokaryotic organisms. This updating approach uses a novel sequence analysis we developed that detects whether the microbe being analyzed has an outer membrane. This identification of membrane structure permits appropriate SCL prediction in an auto-updated fashion and allows PSORTdb to serve as a practical resource for genome annotation and prokaryotic research.

INTRODUCTION

Protein subcellular localization (SCL) prediction aids inference of protein function, identifies candidates for drug or vaccine targets, reveals suitable targets for microbial diagnostics and provides directions for experimental design. For biomedical applications, identification of cell surface and secreted proteins from pathogenic bacteria may lead to the discovery of novel therapeutic targets. Characterizing cell surface and extracellular proteins associated with non-pathogenic Bacteria and Archaea can have industrial uses, or play a role in environmental detection.

The first SCL prediction software, called PSORT, was developed in 1991 by Kenta Nakai for bacteria, animals and plants (1, 2). PSORT II, iPSORT and WoLF PSORT were subsequently developed for eukaryotic species (3–5). PSORTb and PSORTb 2.0 were later developed in 2003 and 2005 specifically for Gram-negative and -positive bacterial protein SCL prediction, with a focus on high-precision/specificity predictions (6,7). They have been the most precise SCL prediction methods developed (8). However, recently PSORTb version 3.0 was developed, with 98% precision for Gram-positive bacteria and 97% precision for Gram-negative bacteria, surpassing PSORTb 2.0 and other available prokaryotic SCL predictors (9). PSORTb 3.0 also provides improved genome prediction coverage (higher recall at high precision), as well as the ability to predict a broader range of prokaryotes including Archaea as well as bacteria with atypical membrane/cell wall structures. In addition, PSORTb 3.0 now identifies subcategory localizations for proteins destined to specialized bacterial organelles (such as the flagellum and pilus) as well as host cell destinations.

The speed in which prokaryotic genomes are sequenced has been increasing at a dramatic rate thanks to the availability of sequencing technologies which can decode DNA sequences at a dramatically increased throughput with lower cost. This creates a challenge for maintaining up-to-date functional annotation of these newly sequenced genomes (10). Given the high accuracy of computational SCL prediction for prokaryotes (8), some genome annotation groups have incorporated SCL prediction into their bioinformatics annotation pipeline (11). Instead of having

*To whom correspondence should be addressed. Tel: 778 782 5646; Fax: 778 782 5583; Email: brinkman@sfu.ca

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

many different researchers compute the same prokaryotic protein SCL prediction repeatedly when needed, it would be more efficient to create a centralized database of pre-computed SCL prediction results that is continually updated to incorporate SCL predictions for newly sequenced organisms.

Several databases containing prokaryotic SCL information have been developed over the years (see http://www.psort.org for a list), such as DBSubLoc, PA-GOSUB and UniProt (12–14). Some are developed specifically for certain types of bacteria; for example, LocateP Database and Augur contain localization predictions specific to Gram-positive bacteria (15,16); others like DBMLoc are specific for multiple SCLs (17). Some incorporate predictions from multiple SCL-prediction tools like CoBaltDB (18). However, none of them are reported, or observed, to be continually updated in a frequent, regular fashion to accommodate newly sequenced genomes, nor do they contain high-precision predictions suitable for handling diverse prokaryotic cellular structures.

PSORTdb (19) is a database initially developed in 2005 to contain experimentally determined (ePSORTdb) and computationally predicted (cPSORTdb) protein SCLs for Bacteria. The computational predictions in cPSORTdb were originally generated by PSORTb 2.0, the most precise bacterial SCL predictor of its time (7). It is widely used by researchers wishing to identify the SCL of specific proteins, verify high-throughput experimental results, as well as those who need a training data set to develop novel SCL prediction software.

To keep up with the increasing rate of prokaryotic genomes sequenced, and a new version of PSORTb (version 3.0), we have now developed a new version of PSORTdb (version 2.0; http://db.psort.org) that automatically computes PSORTb 3.0 SCL prediction results as new prokaryotic genomes become available through NCBI each month. A largely expanded training data set of proteins with known SCLs have also been added to ePSORTdb. We have also improved the user interface to facilitate easier browsing and searching of the database, as well as convenient download options for filtered search results. Old PSORTb 2.0 prediction results are still maintained in this database for archival purposes. The entire database is freely available through the web or for download, and highlighted features are briefly described below.

New database with largely expanded content

ePSORTdb has significantly expanded and now contains over 9100 entries for Gram-negative bacteria, 2980 entries for Gram-positive bacteria and 800 entries for archaeal proteins (previously ePSORTdb contained a total of only 2171 bacterial proteins). This data set came from manual literature search as well as Swiss-Prot annotations as described in the paper describing PSORTb 3.0 (9). The data set can be used by SCL-prediction software developers to test novel machine algorithms and build better SCL prediction tools. For computationally predicted SCLs in cPSORTdb, to date more than 1000 proteomes deduced from sequenced prokaryotic genomes have been analyzed and the results are available for access on the web server as well as for download. At the time of writing, cPSORTdb contains 1286 Gram-negative bacterial replicons, 508 Gram-positive bacterial replicons, 126 archaeal replicons, 30 replicons belonging to Gram-negative bacteria without outer membrane and 11 replicons that belong to Gram-positive bacteria with an outer membrane. All together, SCL data for more than 3 700 000 proteins are currently stored in cPSORTdb. This database is now set to be continually updated, with whole-proteome SCL predictions added as newly available sequenced prokaryotic genomes become available through NCBI's microbial genome database each month.

Automatic database update using a computational 'outer membrane detection' procedure

A major new feature of PSORTdb version 2 is the ability to automatically determine what 'Gram-stain' or cellular structure a given bacterial proteome should be analyzed under, by identifying through a sequence analysis whether the proteome is consistent with an outer membrane-containing bacterium or not. Previously, when microbial genomes have been released by NCBI, we manually determined the Gram-stain and cell structure based on bacterial phylum recorded and literature review. We noticed that neither the Gram-stain in the NCBI microbial database nor the bacterial phylum were 100% accurate in indicating the cell structure, due in part to the increasing diversity of bacterial genomes being sequenced. Gram-positive organisms traditionally have a cytoplasmic membrane surrounding the cell, and a thick cell wall composed of peptidoglycan that encircles the cytoplasmic membrane. Gram-negative organisms typically have a much thinner cell wall within the periplasm and an asymmetrical outer membrane surrounding the entire cell that is in addition to the cytoplasmic membrane. However, the traditional Gram-staining procedure does not always accurately denote the structure of all bacteria. For example, *Deinococcus spp.* stain Gram-positive because they have a thick cell wall, yet they also have an outer membrane (20). *Mycoplasma spp.*, on the other hand, stain Gram-negative because they have no cell wall, but they also only have one cell membrane (i.e. no classical outer membrane) (21). The former should really be analyzed like a Gram-negative, to identify proteins in its outer membrane, while the latter should not have proteins predicted in non-existing outer membrane SCLs since it contains no such structure. Using taxonomy alone is also insufficient in detecting cell structure. Most bacteria within one phylum tend to have the same cellular structure, but *Halothermothrix orenii* of the phylum Clostridia has both characteristics of Gram-positive organisms yet also has an outer membrane (22). Hence we developed a novel automatic cell-structure determination method which we report here. Through research of different possibilities, we have determined that the presence of an outer membrane in a bacterium can be accurately determined by detecting the presence of the outer membrane protein Omp85, or more accurately the *omp85* gene or its orthologs, in a microbial genome. Omp85 is essential for outer membrane biogenesis and is the only known essential outer membrane protein for the viability of bacteria (the latter based on high-resolution analyses of saturated transposon mutagenesis of classic Gram-negative bacteria such as *Pseudomonas aeruginosa* (23,24). Using Omp85 proteins from four divergent genera of bacteria, *Neisseria gonorrhoeae*, *Thermosipho africanus*, *Synechococcus sp. PCC 7002* and *Thermus thermophilus*, we use BLAST to search for homologs of each in the sequenced bacterial genome to be analyzed (*E*-value cut off of 10^{-3}). We found this was necessary to ensure high recall/sensitivity as simply using one Omp85 protein or ortholog did not detect all bacteria that we had manually confirmed as having an outer membrane. Using a data set of 813 diverse bacterial proteomes, curated regarding their outer membrane status, we also determined the appropriate *E*-value cutoff for this analysis and were able to easily obtain 100% precision and 100% recall using the diverse set of 4 Omp85 query sequences. We then compare the results with the phylum taxonomy of the bacteria, which are usually good indicators of bacterial membrane structure except in a few unusual cases. If the two methods have agreeing results, the bacterial structural category is automatically assigned. If the two results disagree, then manual examination is used to assign one of the categories: classic Gram-negative, classic Gram-positive, Gram-positive with an outer membrane and Gram-negative without an outer membrane. In this way, the majority of the 1000+ prokaryotic genomes completely sequenced to date have been automatically assigned a cell structure and a corresponding PSORTb prediction module pipeline, with a few atypical bacteria flagged for manual inspection. This analysis not only aids appropriate automatic prediction of bacterial SCL prediction, but may also serve as a useful resource for microbiologists in general wishing to quickly determine the membrane structure of a given bacterial genome. Of course, there is the possibility of yet other atypical second membranes being discovered such as the unusual mycobacterial membrane (25), but manual curation of these cases should be possible and we do aim to increase the capability of the PSORTb family of software to identify such bacteria more accurately in the future.

Advanced search and filter functions

PSORTdb 2.0 is available for access with an updated web interface that has maintained the flexible search and browse functionalities, but with improved usability. One may search in either the database of proteins of known SCLs (ePSORTdb) or the database of computationally predicted SCLs (cPSORTdb) for specific proteins by organism domain, taxonomy based on NCBI's Taxonomy Database, Gram-stain, localization, secondary localization, protein name, GenBank ID number or a combination of any of these categories. There is an option to download the list of proteins that match the search criteria for the user's convenience. For example, one may obtain a list for all predicted extracellular proteins in all of the *Escherichia coli* strains by selecting 'Extracellular' for the Localization search category and '*Escherichia coli*' for Organism name search category.

If the user has the sequence of a protein that lacks an ID corresponding to PSORTdb's GenBank locus tag, our web interface provides a BLAST search function, which allows the user to find the localization of the query protein as well as homologous proteins. In the vast majority of the cases, SCL is highly conserved between highly similar, homologous proteins (26). In some cases proteins that are peripherally attached to the cell membrane or the cell wall will have only one of its two main localizations predicted depending on its sequence. These can be noted in the database with text 'this protein may have multiple localizations' or 'unknown/multiple localization' (the 'unknown' designation being because the scores for a given localization are split over multiple localizations, and we feel such cases should require more manual inspection to deduce their most probable localization/localizations). In some rare cases, proteins with homologous sequences have changed localizations (27). The BLAST option allows for the detection and examination of these cases.

To browse an entire replicon of a genome, user can now start by simply entering an organism name. An auto-complete function allows the user to select from a list of possible organism/strain names once they start typing a name. Each genome replicon now has a proteome summary page. A pie chart of PSORTb 3.0 protein localization proportion distributions within the proteome, and numeric break down of the number of proteins in each localization category for both PSORTb 3.0 and PSORTb 2.0 predictions, and general taxonomy and cellular-structural information are provided (i.e. under what PSORTb analysis category the organism was analyzed, based on the 'outer membrane detector' analysis that had been performed). By clicking on a specific localization label, a list of all proteins from that localization within the organism proteome will be returned. This is useful if a researcher wants to get a list of, for example, all outer-membrane proteins in a pathogenic bacterium for drug or vaccine target research.

CONCLUSION

We have developed a new version of PSORTdb which contains a greatly expanded data set of experimentally verified SCLs, as well as pre-computed highly precise SCL predictions for all prokaryotes, including now Archaea and bacterial organisms with atypical membrane/cell structures. The web server has been redesigned to facilitate user-friendly search and browsing of the database. It is continuously updated as newly sequenced prokaryotic genomes are released, using a novel computationally based cell-structure analysis which we developed. This allows PSORTdb to remain useful for researchers for analysis of novel species as the number of microbial genome sequences grow at a rapid pace. The contents of PSORTdb can be easily incorporated into whole-genome annotations and the entire database is open access, so it may be a valuable and convenient tool for a wide range of bioinformatics, genomics and microbiology researchers.

ACKNOWLEDGEMENTS

We thank NCBI for providing a compilation of complete microbial genomes data and gratefully acknowledge Raymond Lo for comments regarding this article.

FUNDING

Natural Sciences and Engineering Research Council of Canada (NSERC) (with the support of the Cystic Fibrosis Foundation and Simon Fraser University Community Trust); Michael Smith Foundation for Health Research Senior Scholarship (to F.S.L.B.). Funding for open access charge: NSERC.

Conflict of interest statement. None declared.

REFERENCES

1. Nakai,K. and Kanehisa,M. (1991) Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins*, **11**, 95–110.
2. Nakai,K. and Kanehisa,M. (1992) A Knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, **4**, 897–911.
3. Horton,P. and Nakai,K. (1997) Better prediction of protein cellular localization sites with the k nearest neighbors classifier. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 147–152.
4. Bannai,H., Tamada,Y., Maruyama,O., Nakai,K. and Miyano,S. (2002) Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, **18**, 298–305.
5. Horton,P., Park,K.J., Obayashi,T., Fujita,N., Harada,H., Adams-Collier,C.J. and Nakai,K. (2007) WoLF PSORT: Protein localization predictor. *Nucleic Acids Res.*, **35**, W585–W587.
6. Gardy,J.L., Spencer,C., Wang,K., Ester,M., Tusnády,J.L., 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.
7. Gardy,J.L., Laird,M.R., Chen,F., Rey,S., Walsh,C.J., Ester,M. and Brinkman,F.S. (2005) PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics*, **21**, 617–623.
8. Gardy,J.L. and Brinkman,F.S.L. (2006) Methods for predicting bacterial protein subcellular localization. *Nat. Rev. Micro.*, **4**, 741–751.
9. Yu,N.Y., Wagner,J.R., Laird,M.R., Melli,G., Rey,S., Lo,R., Dao,P., Sahinalp,S.C., Ester,M., Foster,L.J. *et al.* (2010) PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, **26**, 1608–1615.
10. Lagesen,K., Ussery,D.W. and Wassenaar,T.M. (2010) Genome update: The 1000th genome–a cautionary tale. *Microbiology*, **156**, 603–608.
11. Vallenet,D., Engelen,S., Mornico,D., Cruveiller,S., Fleury,L., Lajus,A., Rouy,Z., Roche,D., Salvignol,G., Scarpelli,C. *et al.* (2009) MicroScope: a platform for microbial genome annotation and comparative genomics. *Database (Oxford)*, **2009**, bap021.
12. Guo,T., Hua,S., Ji,X. and Sun,Z. (2004) DBSubLoc: Database of protein subcellular localization. *Nucleic Acids Res.*, **32**, D122–D124.
13. Lu,P., Szafron,D., Greiner,R., Wishart,D.S., Fyshe,A., Pearcy,B., Poulin,B., Eisner,R., Ngo,D. and Lamb,N. (2005) PA-GOSUB: A searchable database of model organism protein sequences with their predicted gene ontology molecular function and subcellular localization. *Nucleic Acids Res.*, **33**, D147–D1453.
14. Hinz,U. and UniProt Consortium (2010) From protein sequences to 3D-structures and beyond: The example of the UniProt knowledgebase. *Cell Mol. Life Sci.*, **67**, 1049–1064.
15. Zhou,M., Boekhorst,J., Francke,C. and Siezen,R.J. (2008) LocateP: Genome-scale subcellular-location predictor for bacterial proteins. *BMC Bioinformatics*, **9**, 173.
16. Billion,A., Ghai,R., Chakraborty,T. and Hain,T. (2006) Augur–a computational pipeline for whole genome microbial surface protein prediction and classification. *Bioinformatics*, **22**, 2819–2820.
17. Zhang,S., Xia,X., Shen,J., Zhou,Y. and Sun,Z. (2008) DBMLoc: A database of proteins with multiple subcellular localizations. *BMC Bioinformatics*, **9**, 127.
18. Goudenege,D., Avner,S., Lucchetti-Miganeh,C. and Barloy-Hubler,F. (2010) CoBaltDB: Complete bacterial and archaeal orfeomes subcellular localization database and associated resources. *BMC Microbiol.*, **10**, 88.
19. Rey,S., Acab,M., Gardy,J.L., Laird,M.R., deFays,K., Lambert,C. and Brinkman,F.S. (2005) PSORTdb: A protein subcellular localization database for bacteria. *Nucleic Acids Res.*, **33**, D164–D168.
20. Thompson,B.G. and Murray,R.G. (1981) Isolation and characterization of the plasma membrane and the outer membrane of *Deinococcus radiodurans* strain sark. *Can. J. Microbiol.*, **27**, 729–734.
21. Miyata,M. and Ogaki,H. (2006) Cytoskeleton of mollicutes. *J. Mol. Microbiol. Biotechnol.*, **11**, 256–264.
22. Mavromatis,K., Ivanova,N., Anderson,I., Lykidis,A., Hooper,S.D., Sun,H., Kunin,V., Lapidus,A., Hugenholtz,P., Patel,B. *et al.* (2009) Genome analysis of the anaerobic thermohalophilic bacterium *Halothermothrix orenii. PLoS One*, **4**, e4192.
23. Voulhoux,R., Bos,M.P., Geurtsen,J., Mols,M. and Tommassen,J. (2003) Role of a highly conserved bacterial protein in outer membrane protein assembly. *Science*, **299**, 262–265.
24. Tashiro,Y., Nomura,N., Nakao,R., Senpuku,H., Kariyama,R., Kumon,H., Kosono,S., Watanabe,H., Nakajima,T. and Uchiyama,H. (2008) Opr86 is essential for viability and is a potential candidate for a protective antigen against biofilm formation by *Pseudomonas aeruginosa. J. Bacteriol.*, **190**, 3969–3978.
25. Hoffmann,C., Leis,A., Niederweis,M., Plitzko,J.M. and Engelhardt,H. (2008) Disclosure of the mycobacterial outer membrane: Cryo-electron tomography and vitreous sections reveal the lipid bilayer structure. *Proc. Natl Acad. Sci. USA*, **105**, 3963–3967.
26. Nair,R. and Rost,B. (2002) Sequence conserved for subcellular localization. *Protein Sci.*, **11**, 2836–2847.
27. Li,T., Huang,X., Zhou,R., Liu,Y., Li,B., Nomura,C. and Zhao,J. (2002) Differential expression and localization of Mn and Fe superoxide dismutases in the heterocystous cyanobacterium *Anabaena sp.* strain PCC 7120. *J. Bacteriol.*, **184**, 5096–5103.