Nucleic Acids Research, 2003, Vol. 31, No. 13 ***3613–3617***
DOI: 10.1093/nar/gkg602

PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria

Jennifer L. Gardy, Cory Spencer, Ke Wang[1], Martin Ester[1], Gábor E. Tusnády[2], István Simon[2], Sujun Hua[3], Katalin deFays[4], Christophe Lambert[4], Kenta Nakai[5] and Fiona S.L. Brinkman*

Department of Molecular Biology and Biochemistry and [1]Department of Computing Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6, [2]Institute of Enzymology, Hungarian Academy of Sciences H-1113 Budapest Karolina ut 29, Hungary, [3]Bioinformatics and Computational Biology Track, Biological and Biomedical Sciences Program, Yale University, New Haven, CT 06520, USA, [4]URBM, FUNDP Rue de Bruxelles, 61, B-5000, Namur, Belgium and [5]Human Genome Center, The Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan

Received February 15, 2003; Revised and Accepted April 4, 2003

ABSTRACT

Automated prediction of bacterial protein subcellular localization is an important tool for genome annotation and drug discovery. PSORT has been one of the most widely used computational methods for such bacterial protein analysis; however, it has not been updated since it was introduced in 1991. In addition, neither PSORT nor any of the other computational methods available make predictions for all five of the localization sites characteristic of Gram-negative bacteria. Here we present PSORT-B, an updated version of PSORT for Gram-negative bacteria, which is available as a web-based application at http://www.psort.org. PSORT-B examines a given protein sequence for amino acid composition, similarity to proteins of known localization, presence of a signal peptide, transmembrane alpha-helices and motifs corresponding to specific localizations. A probabilistic method integrates these analyses, returning a list of five possible localization sites with associated probability scores. PSORT-B, designed to favor high precision (specificity) over high recall (sensitivity), attained an overall precision of 97% and recall of 75% in 5-fold cross-validation tests, using a dataset we developed of 1443 proteins of experimentally known localization. This dataset, the largest of its kind, is freely available, along with the PSORT-B source code (under GNU General Public License).

INTRODUCTION

A protein's subcellular localization can provide valuable clues as to its function. As the interpretation of sequenced genomic data becomes increasingly important, so does the need for accurate automated prediction of localization from sequence information alone. Such predictions allow us to screen candidates for drug discovery, automatically annotate gene products (1) and select proteins for further study (2).

Since 1991, several automated subcellular localization predictors have been developed and made available online, including: PSORT (PSORT I) for prokaryotic organisms (3), PSORT II (4), iPSORT (5) and TargetP (6) for eukaryotic organisms, and NNPSL (7) and SubLoc (8) for both organism classes. The programs vary considerably: predictive methods may be manually-constructed rules based on existing knowledge or one of several machine learning approaches, and the methods may recognize single or multiple localization sites. Either single or multiple features of a protein may also be detected, such as a signal peptide or whole protein amino acid composition.

The performance of these programs also varies (9). Datasets used for training vary considerably in size, from 13 to >1000 proteins, and may contain proteins with incorrectly annotated localization sites. Additionally, the lack of data for certain sites contributes to poor performance for specific localizations (9).

*To whom correspondence should be addressed. Tel: +1 6042915646; Fax: +1 6042915583; 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.

Gram-negative bacteria have five primary localization sites—the cytoplasm, the inner membrane, the periplasm, the outer membrane and the extracellular space. For bacterial localization prediction the mostly widely used tool that detects multiple protein features and multiple localizations is PSORT I (3). However, PSORT I has not been updated for the analysis of bacterial proteins since its initial release in 1991. It also does not predict extracellular proteins—proteins which may represent important virulence factors in pathogenic microorganisms. Both SubLoc and NNPSL, other programs available for the analysis of prokaryotic sequences, are limited to predicting three of the five classic subcellular locations present in a Gram-negative bacterial cell (7,8). SignalP (10), another useful method for analysis of bacterial signal peptides, is limited to the identification of the subset of exported proteins that contain this classic targeting signal.

We now present an updated version of PSORT for the prediction of protein subcellular localization for Gram-negative bacteria. PSORT-B combines several methods, including homology analysis, identification of sorting signals and other motifs, and machine learning methods into an expert system for prediction of five subcellular localizations, given a complete amino acid sequence. This initial version is designed to favor high precision over high recall and returns a probability score for each of the five possible localization sites. As part of the development of PSORT-B, we have also constructed the largest dataset to date of bacterial proteins of experimentally known subcellular localization.

MATERIALS AND METHODS

Dataset of proteins of experimentally determined subcellular localization

For training of certain PSORT-B analytical modules described below, and for 5-fold cross-validation of PSORT-B's accuracy, a manually curated dataset of proteins of experimentally known subcellular localization was constructed. Gram-negative bacterial sequences with an annotated subcellular localization were extracted from SWISS-PROT release 40.29 (11). All proteins denoted as fragments were removed, as were all proteins whose annotations were listed as 'by similarity' or 'putative'. Further filtering removed those sequences with ambiguous annotation in the subcellular localization field. The proteins were manually checked against the literature for experimental verification of the annotated localization site. The final dataset consists of 1443 proteins of experimentally determined localization and is available online at http://www.psort.org/dataset. The dataset comprises 1302 proteins resident at a single localization site: 248 cytoplasmic, 268 inner membrane, 244 periplasmic, 352 outer membrane and 190 extracellular; and contains a further 141 proteins resident at multiple localization sites: 14 cytoplasmic/inner membrane, 50 inner membrane/periplasmic and 77 outer membrane/extracellular.

PSORT-B analytical modules

PSORT-B utilizes six analytical modules in generating an overall prediction of localization site. A query protein undergoes each of the analyses and the results from each module are combined using a Bayesian Network. Certain modules are multi-category predictors, while others are binary predictors designed to predict one localization site only. The modules are also designed with the ability to return a localization site of 'unknown' if no features can be reliably identified within the query sequence. Modules included in this initial version of PSORT-B include: SCL-BLAST for homology analysis, a PROSITE motif-based analysis for detection of localization-specific motifs, HMMTOP for detection of transmembrane alpha-helices, a novel outer membrane protein motif analysis, a type II secretion signal peptide predictor and a variation of SubLoc. Each of these modules is described briefly below with more information available in PSORT-B's documentation at www.psort.org.

SCL-BLAST. Subcellular localization tends to be evolutionarily conserved (12), thus homology to a protein of known localization appears to be a good indicator of a protein's actual localization site. We therefore constructed a module entitled SCL-BLAST (for SubCellular Localization BLAST), in which a BLAST search of a submitted protein is carried out against our database of 1443 proteins of known localization, using an E-value cutoff of $10\,e^{-10}$. A length restriction is placed on resulting high scoring pairs, such that the length of the high scoring pair must be within 80–120% of the length of the subject. This reduces the potential for misprediction of localization based on similarity to a single domain of a protein in the database, a protein whose domains may reside in different localization sites. The module returns the localization site and SWISS-PROT accession number of any hits fulfilling the above criteria and can generate a prediction for any of the five sites.

Motifs. A protein's functional description is often indicative of its subcellular localization (2). Therefore certain sequence patterns corresponding to function may also correlate with a specific subcellular localization. PROSITE release 17.0 (13) was searched for such potential patterns and the resulting list was tested on the dataset of 1443 proteins of known localization. Twenty six motifs, available at http://www.psort.org/motifs, capable of identifying subcellular localization with 100% precision were retained. The module returns the localization site and PROSITE accession number of any pattern found within the sequence and can generate a prediction for any of the five sites.

HMMTOP. Integral inner membrane proteins are characterized by the presence of alpha-helical transmembrane regions (14) and this feature has been used as a reliable indicator of localization at the inner membrane in past predictors, including PSORT I (3). PSORT-B utilizes the Hidden Markov Model-based method HMMTOP (15,16) to identify potential transmembrane alpha helices, assigning a localization of inner membrane if three or more helices are found.

Outer membrane protein motifs. The identification of outer membrane proteins is of particular interest, both due to the difficulty in predicting their characteristic beta-barrel structure and their high potential for use as drug targets. A data mining

approach was used to identify frequent sequences occurring only in beta-barrel proteins, both integral outer membrane proteins and autotransporter proteins, which possess a beta-barrel transport domain. A total of 279 frequent sequences, available at http://www.psort.org/motifs, were generated and used to build a classifier. A user-submitted sequence is screened for the presence of three or more of the frequent sequences and is classified as either outer membrane or non-outer membrane based on the result.

SubLocC—amino acid composition. Support Vector Machine (SVM) has been successfully applied to overall amino acid composition-based subcellular localization prediction in the SubLoc program (8). Using the software LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm), a similar SVM was trained on 248 cytoplasmic sequences and 1054 non-cytoplasmic sequences. A query protein's amino acid composition is analyzed and used to assign the protein to one of the two categories.

Signal peptides. Signal peptides, short sequences present at the amino-terminus of many proteins, direct a protein for transport across the inner membrane (17). Thus the presence of a signal peptide implies that a protein is not resident in the cytoplasm. The prokaryotic SignalP training data, available at http://www.cbs.dtu.dk/ftp/signalp/, was used to train a Hidden Markov Model (HMM) to identify signal peptide cleavage sites within the first 70 residues of a sequence. A probability value is assigned to the cleavage site and, if it exceeds a pre-assigned cutoff, a prediction of non-cytoplasmic is returned. If the *p*-value of the predicted cleavage site falls within a 'twilight zone', the signal peptide is then passed to an SVM trained on the same data, also capable of identifying signal peptides. If the SVM returns a result of signal peptide, a non-cytoplasmic prediction is returned. If no signal peptide is identified, the module returns an output of 'unknown', as the lack of a predicted signal peptide does not necessarily imply a cytoplasmic localization.

Final prediction: scoring the results

The predictive power of each module was assessed and this data was used to construct a Bayesian Network capable of generating a final probability value for each localization site. This final score reflects the likelihood of a protein actually being resident at a specific localization, given the predictions of the individual modules. A score out of 10 is produced for each of the five possible localization sites, representing the probability value calculated multiplied by a factor of 10, with a high value reflecting high confidence that the given protein is resident in that subcellular location. The sites are ranked in descending order of probability. If none of the localization sites have a score >7.5, a prediction of 'unknown' is returned. A distribution of scores heavily favoring one site indicates the protein is likely to be resident there, while a distribution favoring two sites may indicate the protein has domains residing in more than one localization site. An even distribution of scores indicates an unknown localization.

Evaluation of the accuracy of PSORT-B

All evaluations, with the exception of PROSITE motif module analysis, were carried out using 5-fold cross-validation. In *k*-fold cross-validation, the relevant dataset is partitioned randomly into k equally sized partitions and module development and evaluation is carried out k times, each time using one distinct partition as the testing set and the remaining $k-1$ partitions as the training set. The precision and recall is computed as the average of the total runs, the procedure prevents artificially inflated performance values. The choice of $k=5$ implies that 80% of the sequences are used for training of a module and 20% for testing. SCL-BLAST was evaluated using the dataset of 1443 proteins of known subcellular localization. PROSITE motifs were selected to yield a 100% precision value over the same dataset. Outer membrane protein motifs were evaluated using the 425 beta-barrel proteins in the dataset and the remaining 1018 non-outer membrane proteins. SubLocC was evaluated using the 248 cytoplasmic proteins in the dataset and the remaining 1054 non-cytoplasmic proteins—cytoplasmic proteins with a second, dual localization were not included in either class. The predictive power of HMMTOP was evaluated on the 268 integral inner membrane and the remaining 1175 non-inner membrane proteins in the dataset. The signal peptide module was trained using the SignalP dataset mentioned above and evaluated with the Menne *et al.* (18) dataset of 426 signal peptides and 433 non-signal peptides. The overall performance of PSORT-B was assessed using the dataset of 1443 proteins of known localization. Precision and recall values were calculated per localization site and the overall precision and recall of PSORT-B was calculated based on the total true-positives (TP), false-positives (FP) and false-negatives (FN) over the five sites. For the purposes of evaluation, predictions were considered to have been made if the PSORT-B scoring system gave a score for a particular localization site of >7.5, as we found this to be a useful cutoff in our evaluation of accuracy. Proteins resident at dual localization sites were considered to have been predicted correctly if one of their localization sites scored >7.5. For all evaluations, precision or specificity, was calculated as $TP/(TP + FP)$ (i.e. how often results are correct) and recall or sensitivity, was calculated as $TP/(TP + FN)$ (i.e. how well all true results are retrieved). For modules capable of predicting multiple localization sites, the reported precision and recall values are averaged across the relevant localization sites.

RESULTS

The precision and recall of each analytical module were assessed individually (Table 1—see also Methods for details) and used to construct the probabilistic system for the generation of final probability values. PSORT-B's overall precision and recall for each localization site was evaluated using the dataset of 1443 proteins and the results are shown in Table 2.

PSORT-B is available online at http://www.psort.org. This web site also contains links to other PSORT programs and additional resources for subcellular localization prediction. For PSORT-B, the user is asked to submit one or more amino acid sequences in FASTA format. Results are returned on a

Table 1. Evaluation of PSORT-B's analytical modules

Module	Precision	Recall
SubLocC	78.6	74.2
HMMTOP	99.4	65.3
Motif	100.0	6.5
OMP Motif	100.0	23.6
SCL-BLAST	96.7	60.4
Signal	87.0	98.2

Table 2. Comparison of PSORT I and PSORT-B's performance

Localization	PSORT I Precision	Recall	PSORT-B Precision	Recall
Cytoplasmic	59.7	75.4	97.6	69.4
Inner membrane	55.4	95.1	96.7	78.7
Periplasmic	60.9	66.4	91.9	57.6
Outer membrane	65.3	54.5	98.8	90.3
Extracellular	0.0	0.0	94.4	70.0
Overall	59.6	60.9	96.5	74.8

new web page. For each query protein, a list of the five possible localization sites is returned with the corresponding final probability value (a value out of 10). The sites are ranked in descending order of probability.

As Table 1 illustrates, the precision and recall of the different modules vary greatly, however they are designed to favor precision—the focus is on predicting results correctly rather than generating a prediction in every case. The overall accuracy of PSORT-B also reflects this, as shown in Table 2. The accuracy of PSORT-B is a significant improvement over the PSORT I program, which according to our analyses had an overall precision and recall of 59.6 and 60.9%, respectively, when evaluated using our dataset of 1443 proteins. A large increase in precision can be observed for each localization site, however recall is lessened in certain cases, reflecting our goal of returning an accurate prediction rather than a prediction we are not confident in. This is especially evident for inner membrane proteins, where a 16.4% decrease in recall is compensated for by a 41.3% increase in precision.

DISCUSSION

PSORT-B represents a powerful tool for prediction of protein subcellular localization for Gram-negative bacteria. High precision allows for confident predictions, and prevents propagation of erroneous predictions. Allowing the user to view the outputs and the probability values for each query enables them to incorporate their own specific knowledge about the query to arrive at their own conclusion. Additionally, PSORT-B is able to handle situations where no prediction is possible by assigning equal probabilities to each of the localization sites, again avoiding propagation of erroneous predictions.

PSORT-B also introduces novel analytical modules. SCL-BLAST is the first publicly available localization predictor to utilize homology information and the outer membrane motif module represents the only currently available web-based outer membrane protein classifier. Its frequent sequence-based approach allows for higher precision than other unreleased classifiers relying on machine learning approaches (Gardy, unpublished results).

In future versions of PSORT-B we propose expanding its capability to include more complex localizations and to better handle proteins that are resident in more than one localization over time and/or have domains present in more than one of the five classic localization sites. Presently PSORT-B attempts to handle such cases by often providing a high score for two localizations for the protein. In the present evaluation, proteins with dual localization sites were reported as correctly predicted if one of their localization sites scored higher than our assigned cutoff of 7.5. In future versions of PSORT-B, a more appropriate cutoff may be determined after analysis of this class of proteins.

An eventual goal is to have a probability value associated with each residue in a protein, rather than the whole protein itself, however more experimental data is required before this goal can be realized. Another future goal is to have variable cutoffs for some of the analytical modules, to permit more flexible analysis. For example, different Expect value cutoffs could be offered for SCL-BLAST to permit the user to increase recall at the expense of precision. We would also like to improve prediction of certain classes of proteins that we wished to mention here, since they are currently poorly predicted: cytoplasmic membrane-associated proteins that have two or less transmembrane alpha-helices, outer membrane-associated proteins that do not have a classic beta-barrel structure and proteins secreted by non-type II secretion pathways.

PSORT-B has purposefully been constructed in a modular form with Perl, to permit the introduction of additional analyses targeted to such issues in the future. Additionally, BioPerl modules for each analysis or related analyses are being developed (i.e. we have released an SVM and SubLoc BioPerl module—see the CPAN archive www.cpan.org).

Though we plan to expand PSORT-B to make predictions for other bacteria, this initial version focused on the analysis of Gram-negative bacteria, as they are presently the most poorly analyzed by current localization predictors. None of the other predictors available make predictions for all five localization sites of Gram-negative bacteria and the original PSORT I program did not predict certain important localization classes—such as outer membrane proteins—very accurately (65.3% precision, 54.5% recall). As Table 2 shows, PSORT-B currently predicts outer membrane proteins most accurately of all the localizations. This was a particular focus of ours because outer membrane proteins—as primary cell surface components of Gram-negative bacteria—are attractive potential vaccine targets, diagnostic agents and drug targets of medical, agricultural and environmental interest. Periplasmic proteins are most poorly predicted by PSORT-B (91.9% precision, 57.6% accuracy). This is due in part to the lack of experimental study of this class of proteins. Hopefully, as more periplasmic proteins are identified through proteomic studies, this data may be incorporated into particular modules like SCL-BLAST and Motifs, and more accurate predictions obtained. One of the powers of PSORT-B is that it will only

increase its predictive capability over time: We will ensure that precision does not drop as we update the modules and we expect recall to increase as, for example, more proteins of known localization are entered in the SCL-BLAST database and more motifs (purposefully chosen with 100% precision) are incorporated into the analysis.

PSORT-B is being developed under an open source license (GNU GPL) to encourage the open development and expansion of this resource, although one module of the program, which must be obtained separately from the PSORT-B source code, remains under another license and is free for academics. From the psort.org website, users can also link to other open source resources being developed under the PSORT umbrella, as well as other computational tools that may aid a researcher in prediction of protein subcellular localization.

ACKNOWLEDGEMENTS

We wish to thank Dr Oliver Schulte (Simon Fraser University, Canada) for his helpful comments regarding the implementation of a probabilistic scoring system and Shannan Ho Sui for her work on the initial development of SCL-BLAST. J.L.G., C.S. and F.S.L.B. were funded by the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

1. Nakai,K. (2000) Protein sorting signals and prediction of subcellular localization. *Adv. Protein. Chem.*, **54**, 277–344.
2. Eisenhaber,F. and Bork,P. (1998) Wanted: subcellular localization of proteins based on sequence. *Trends Cell Biol.*, **8**, 69–70.
3. Nakai,K. and Kanehisa,M. (1991) Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins*, **11**, 95–110.
4. Horton,P. and Nakai,K. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, **24**, 34–36.
5. 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.
6. Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
7. Reinhardt,A. and Hubbard,T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.*, **26**, 2230–2236.
8. Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.
9. Emanuelsson,O. (2002) Predicting protein subcellular localisation from amino acid sequence information. *Brief. Bioinform.*, **3**, 361–376.
10. Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Syst.*, **8**, 581–599.
11. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
12. Nair,R. and Rost,B. (2002) Sequence conserved for subcellular localization. *Protein Sci.*, **11**, 2836–2847.
13. Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.
14. von Heijne,G. (1994) Signals for protein targeting into and across membranes. *Subcell. Biochem.*, **22**, 1–19.
15. Tusnady,G.E. and Simon,I. (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.*, **283**, 489–506.
16. Tusnady,G.E. and Simon,I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849–850.
17. Bernstein,H.D. (1998) Membrane protein biogenesis: the exception explains the rules. *Proc. Natl Acad. Sci. USA*, **95**, 14587–14589.
18. Menne,K.M.L., Hermjakob,H. and Apweiler,R. (2000) A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics*, **16**, 741–742.