BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 5 2005, pages 617–623
doi:10.1093/bioinformatics/bti057

Sequence analysis

PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis

J. L. Gardy[1], M. R. Laird[1], F. Chen[2], S. Rey[1], C. J. Walsh[1], M. Ester[2] and F. S. L. Brinkman[1,*]

[1]Department of Molecular Biology and Biochemistry and [2]Department of Computing Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6

Received on July 5, 2004; revised on September 17, 2004; accepted on September 28, 2004
Advance Access publication October 22, 2004

ABSTRACT

Motivation: PSORTb v.1.1 is the most precise bacterial localization prediction tool available. However, the program's predictive coverage and recall are low and the method is only applicable to Gram-negative bacteria. The goals of the present work are as follows: increase PSORTb's coverage while maintaining the existing precision level, expand it to include Gram-positive bacteria and then carry out a comparative analysis of localization.

Results: An expanded database of proteins of known localization and new modules using frequent subsequence-based support vector machines was introduced into PSORTb v.2.0. The program attains a precision of 96% for Gram-positive and Gram-negative bacteria and predictive coverage comparable to other tools for whole proteome analysis. We show that the proportion of proteins at each localization is remarkably consistent across species, even in species with varying proteome size.

Availability: Web-based version: http://www.psort.org/psortb. Standalone version: Available through the website under GNU General Public License.

Contact: psort-mail@sfu.ca, brinkman@sfu.ca

Supplementary information: http://www.psort.org/psortb/supplementaryinfo.html

INTRODUCTION

Subcellular localization prediction allows researchers to make inferences regarding a protein's function, to annotate genomes, to design proteomics experiments and—particularly in the case of bacterial pathogen proteins—to identify potential diagnostic, drug and vaccine targets. The last year has seen the release of several localization prediction tools, including CELLO (Yu *et al*., 2004) and Proteome Analyst (Lu *et al*., 2004), the only new tools capable of analyzing bacterial proteins. CELLO utilizes an n-peptide composition-based support vector machine (SVM) approach in its analyses, while Proteome Analyst generates predictions using an annotation keyword-based system.

In 2003 we released PSORTb, an open-source tool for localization prediction in Gram-negative bacteria (Gardy *et al*., 2003). PSORTb uses the multiple classification method approach pioneered by PSORT I (Nakai and Kanehisa, 1991), in which several sequence features known to influence localization are analyzed using different computational techniques. By analyzing features including signal peptides, transmembrane helices, homology to proteins of known localization, amino acid composition and motifs, PSORTb v.1.1 attained a classification precision of 97%. However, the method did not extend to Gram-positive organisms and its predictive coverage when applied to whole proteomes—the number of proteins for which a prediction could be made—remained low, at 28%. The goal of the present work was to expand PSORTb's predictive scope by introducing additional classification methods applicable to both Gram-positive and Gram-negative bacteria, while maintaining the existing standard of high precision. SVM was investigated as a potential method for increasing coverage.

An SVM (Vapnik, 1995) is a kernel learning algorithm, in which all the data are mapped as vectors in n-dimensional feature space. Given training data from two classes (positive and negative), an SVM learns the optimal separating hyperplane that separates the two classes and maximizes their distance from the hyperplane. In the previous works on the applicability of SVMs to the localization classification problem, nucleotide or protein sequences have been modeled as vectors representing amino acid composition (Hua and Sun, 2001; Yu *et al*., 2004). We proposed, however, that the precision of an SVM could be improved by utilizing frequently occurring subsequences rather than overall amino acid composition. Such common patterns within a group of proteins may indicate the site of a common biochemical mechanism or a structural motif. In an earlier work, we examined the applicability of this method to the classification of outer membrane proteins (She *et al*., 2003); here, we show that it can be used for high-precision classification of all prokaryotic localization sites.

By the introduction of an SVM-based classifier and expansion of the SCL-BLAST and motif-based analyses, we have significantly improved PSORTb's predictive capacity relative to version 1.1. The program is now capable of generating predictions for Gram-positive bacteria, and is able to make predictions for 75% of a Gram-positive proteome and 57% of a Gram-negative proteome, comparable to the coverage attained by other methods. Five localization sites are predicted for Gram-negative bacteria (cytoplasm, cytoplasmic membrane, periplasm, outer membrane and extracellular) and four for

[*]To whom correspondence should be addressed.

Gram-positive bacteria (cytoplasm, cytoplasmic membrane, cell wall and extracellular), with the program also able to flag potentially multiply localized proteins. PSORTb remains the most precise tool for localization prediction available, with a measured classification precision of 96% for both Gram-negative and Gram-positive bacteria.

The improved coverage attained by PSORTb v.2.0 allowed us to compare the proportion of proteins resident at each localization site across multiple proteomes. We hypothesized that free-living organisms with more diverse environmental niches may contain more membrane proteins in order to facilitate uptake of a variety of materials. We found, however, that the proportion of proteins at a given localization site remains remarkably constant across species, regardless of lifestyle, environmental niche or proteome size.

SYSTEM AND METHODS

Dataset

PSORTb v.2.0 was trained and evaluated using an expanded version of the original PSORTdb dataset (Gardy *et al*., 2003). This updated dataset, the composition of which is shown in Table 1, includes 150 new Gram-negative proteins and 576 new Gram-positive proteins. Each protein's localization site has been experimentally verified and reported in the literature. The dataset is freely available at http://www.psort.org/dataset.

PSORTb v.2.0 organization

Similar to PSORTb v.1.1, PSORTb v.2.0 consists of a series of analytical modules, each capable of generating predictions for one or more localization sites. However, several significant changes have been made to the modules in the new version. In version 2.0, the SubLocC module has been replaced with a new SVM-based method, as described below. The signal peptide identification module has now been trained with Gram-positive data in addition to the Gram-negative data used in version 1.1 (Nielsen *et al*., http://www.cbs.dtu.dk/ftp/signalp). The SCL-BLAST and Motif modules have been expanded as described below. No changes were made to the HMMTOP transmembrane helix identification module (Tusnády and Simon, 2001) or the OMPMotif module.

As in version 1.1, the modules' predictions are weighted and integrated using a Bayesian network in order to generate the final prediction, which comes in the form of a score distribution. When a single localization site displays a score of 7.5 or greater, that site is returned as a final prediction. New to version 2.0 is the multiple localization flagging—if two sites return high scores, a flag of 'This protein may have multiple localization sites' is appended to the final prediction. This flag is triggered when a site scores between 4.0 and 7.49 for Gram-negative proteins, and between 5.0 and 7.49 for Gram-positive proteins. If no site scores above 4.0 or 5.0, depending on the class, a localization site of 'Unknown' is returned. PSORTb's emphasis is on precision, and returning a result of 'Unknown' when not enough information is available to make a prediction avoids potential false positive results.

SCL-BLAST and SCL-BLASTe

PSORTb's SCL-BLAST module assigns putative localizations based on homology to a protein of known localization. Version 2.0 improves the recall associated with this module by implementing a BLASTP search (Altschul *et al*., 1990) against the expanded PSORTdb database. We have also introduced an exact match filter to detect if a user's query protein is already in the database—if a query protein displays 100% identity to a protein in PSORTdb with a difference between query and subject length of not more than one character (to account for some users' removal of the initial 'f-methionine' residue), the SCL-BLASTe subroutine returns the localization site associated with the subject protein. In cases where an exact match is identified, the query protein is not analyzed by subsequent modules, enabling a result to be returned faster. The SCL-BLAST module is able to generate predictions for each of the five Gram-negative and four Gram-positive localization sites.

Table 1. Composition of the PSORTdb dataset

Localization	Gram-negative	Gram-positive
C	278	194
CM	309	103
P	276	N/A
OM	391	N/A
EC	190	183
CW	N/A	61
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	576

The following abbreviations for localization sites and predictions are used throughout the paper: C, cytoplasm; CM, cytoplasmic membrane; P, periplasm; OM, outer membrane; EC, extracellular; and CW, cell wall. A '/' character indicates a multiply localized/predicted protein.

Motifs and profiles

In PSORTb v.1.1, the Motif module scanned a query sequence for the presence of any 1 of 26 PROSITE motifs indicative of specific Gram-negative localization sites. In PSORTb v.2.0, the module has been expanded to include 44 Gram-negative motifs derived from PROSITE v.18 (Hulo *et al*., 2004), covering all but the cytoplasmic localization site, and 25 Gram-positive motifs covering all 4 localization sites. The complete list of motifs is available at http://www.psort.org/motifs. Each motif has been checked against PSORTdb to ensure that it produces no false positive results. Two motifs used in PSORTb v.1.1 were removed from v.2.0 due to the occurrence of false positives when examined against the expanded PSORTdb.

PSORTb v.2.0 also includes a Profile module, in which localization-specific profiles derived from PROSITE v.18 were selected to generate 100.0% precise predictions against PSORTdb. Each profile is similar to a motif but with position-specific weighting information included, such that more degenerate sequences can be retrieved than via the strict pattern matching of the Motif module. Six profiles were selected, four of which identify both Gram-negative and Gram-positive cytoplasmic proteins and cytoplasmic membrane proteins, and two of which are specific to the Gram-positive cell wall and extracellular sites. The profiles are also available at http://www.psort.org/motifs.

Frequent subsequence-based support vector machines

PSORTb v.2.0 contains a new series of modules utilizing SVMs for classification. Nine SVMs were developed, one for each Gram-negative and Gram-positive localization site. Training data for each SVM consists of a positive class comprising all proteins resident at a specific localization site and a negative class comprising all other proteins of the same Gram category.

Each of the nine positive class datasets was first mined for frequent subsequences using an implementation of the generalized suffix tree (Wang *et al*., 1994). A subsequence was defined as frequent if it occurred in at least X% of proteins in the positive class of training data, where X is a parameter called minimum support, or MinSup. Multiple values of the MinSup parameter were tested.

SVMLight (Joachims, 2002, http://svmlight.joachims.org/) was used to implement nine SVMs whose feature spaces consisted of the frequent subsequences characteristic of a specific localization site. For each localization site, different SVMs were tested using different combinations of MinSup (range: 0.8–13%) and kernel (linear, polynomial with degree = 2, radial basis function with $\gamma = 0.005$). The MinSup/kernel combination giving the highest classification precision combined with a reasonable level of recall (>40%) was selected for inclusion in PSORTb v.2.0. Variations in the margin error penalization parameter C were not evaluated, as our earlier work on the subject showed a negligible effect on precision and recall values (She *et al*., 2003). The final SVMs implemented in PSORTb v.2.0 utilize LibSVM (Lin, 2003, http://www.csie.ntu.edu.tw/∼cjlin/libsvm/).

Evaluation

All evaluations were carried out using 5-fold cross-validation, in each round of which four randomly generated folds of the data were used for training or construction of the module(s) in question and the fifth fold was reserved for testing. Where possible, we have included confusion matrices with our results to aid further evaluation of PSORTb's performance by the researchers using other definitions of accuracy. We have defined precision as TP/(TP + FP) and recall as TP/(TP + FN). In cases where a protein has dual localizations, we count a prediction of either of the two actual sites as a true positive.

IMPLEMENTATION

Expanded PSORTb database and SCL-BLAST

PSORTb's SCL-BLAST module predicts localization of a query sequence based on homology to a protein in the PSORTdb database of proteins of experimentally verified localization. It is therefore expected that a larger and more diverse database will lead to an increase in the program's recall. SCL-BLAST v.2.0 utilizes an updated version of the original PSORTdb database—Gram-positive queries are run against the subset of 576 new proteins of Gram-positive origin, and Gram-negative queries are run against the expanded set of Gram-negative proteins. Furthermore, we investigated whether subsets of the Gram-negative and Gram-positive database could be combined. For example, the cytoplasmic sites and cytoplasmic membrane sites were hypothesized to be functionally equivalent, such that a Gram-negative protein could be searched against a BLAST database containing both Gram-negative proteins and Gram-positive cytoplasmic proteins and cytoplasmic membrane proteins. We examined whether a larger database with such combinations of proteins would increase recall even further.

We tested several combined databases using 5-fold cross-validation and found that higher recall and comparable precision was indeed achieved. For Gram-positive results, a database including Gram-negative cytoplasmic proteins, cytoplasmic membrane proteins and extracellular proteins yielded the best predictions. For Gram-negative queries, optimal results were achieved when the queries were searched against a database that included both Gram-positive cytoplasmic proteins and cytoplasmic membrane proteins—including extracellular proteins in the database resulted in several periplasmic proteins being falsely predicted as extracellular. The results of 5-fold cross-validation testing of SCL-BLAST v.2.0 for each localization site are shown in Table 2. The Gram-negative version of the module retains the 96% precision exhibited in v.1.1, and improves the recall by 8%. The new Gram-positive version also displays a precision of 96% and a recall of 58%, the lower recall most probably due to the smaller Gram-positive database. It is important to note, however, that such recall values are not to be expected when SCL-BLAST is applied to datasets containing a large number of hypothetical proteins, due to their lack of similarity to proteins in the SCL-BLAST database.

Table 2. Performance of the SCL-BLAST module using an expanded database of proteins

Localization	Performance	
	Precision	Recall
Negative		
C	88.8	39.9
CM	97.4	62.0
P	94.4	68.8
OM	99.4	90.5
EC	97.3	77.4
Total	96.4	68.6
Positive		
C	96.6	58.8
CM	96.8	59.8
CW	91.9	56.7
EC	95.5	57.7
Total	95.7	58.4

Support vector machine-based classification

Our previous work on the applicability of frequent subsequence-based SVM to outer membrane protein prediction (She *et al*., 2003) led us to examine whether the method was applicable to proteins resident at all nine localization sites. We reasoned that frequent subsequences found in proteins resident at each site represented conserved functional and structural motifs that would yield higher precision classification than methods based on overall amino acid composition alone.

By mining frequent subsequences from each of the nine localization sites, again combining the Gram-positive and Gram-negative cytoplasmic sequences and cytoplasmic membrane sequences, we were able to develop nine SVMs, each capable of classifying a protein as likely being resident at a specific localization site or not. Varying numbers of frequent subsequences were tested, as were different kernel functions, and the combination of frequent subsequences and kernel yielding the highest precision as well as a reasonable level of recall were selected for use in PSORTb v.2.0. The performance and parameters associated with each of the nine SVMs is shown in Table 3.

By using a feature space comprising frequent subsequences rather than amino acid composition, we were able to attain high-precision classification across all localization sites. Although the precision values for the two cytoplasmic classifiers are the lowest of the nine values, the 84% precision achieved by the Gram-negative SVM represents a 5% increase relative to the cytoplasmic composition-based SVM SubLocC used in PSORTb v.1.1. We believe that the reduced precision associated with cytoplasmic proteins may be due to the extremely diverse nature of proteins found at this site—proteins found at other sites exhibit more functional and structural constraints, resulting in more unique and characteristic frequent subsequences. This is especially evident when classifying cytoplasmic membrane proteins—the frequent subsequences mined from this structurally and environmentally constrained group of proteins results in high-precision classification.

We observed that as the MinSup value increased for each classifier, the number of frequent patterns decreased, as did precision; recall, however, remained comparatively stable (Chen *et al*., unpublished

Table 3. Parameters and performance of the nine frequent subsequence-based SVM modules

Module	SVM Parameters			Performance	
	MinSup	Frequent patterns	Kernel	Precision	Recall
Negative					
CytoSVM−	0.5	39219	Linear	83.6	68.4
CMSVM−	3	5645	Polynomial	96.9	69.6
PPSVM−	1	27804	Polynomial	96.3	45.3
OMSVM−	1	46688	Linear	94.6	85.3
ECSVM−	2	35380	Polynomial	94.1	56.4
Positive					
CytoSVM+	2	8214	Linear	86.5	79.9
CMSVM+	2	250163	Linear	100.0	63.1
CWSVM+	2	11610	Linear	95.7	55.6
ECSVM+	5	23605	Polynomial	91.7	55.0

Table 4. PSORTb v.2.0 performance as measured by 5-fold cross-validation using the complete subset of singly localized proteins from PSORTdb

Localization	Performance				
	TP	FP	FN	Precision	Recall
Negative					
C	195	15	83	92.9	70.1
CM	286	14	23	95.3	92.6
P	191	9	85	95.5	69.2
OM	371	10	20	97.4	94.9
EC	150	4	40	97.4	78.9
Total	1193	52	251	95.8	82.6
Positive					
C	168	5	26	97.1	86.6
CM	94	3	9	96.9	91.3
CW	54	3	7	94.7	88.5
EC	124	8	59	93.9	67.8
Total	440	19	101	95.9	81.3

data). We also noted that the best performance is not achieved at the smallest MinSup value—when the number of frequent subsequences exceeds a certain level, the performance of the SVM is degraded.

PSORTb v.2.0 performance

The new SVM modules, as well as the updated motif, profile and signal peptide modules, were incorporated into PSORTb v.2.0. As in version 1.1, a Bayesian network was constructed in order to integrate the predictions of all modules to generate a final prediction. Multiple weighting values were tested, and the values yielding the highest precision were used in the final version of PSORTb v.2.0. Five-fold cross-validation was then used to evaluate the Gram-negative and Gram-positive versions of the complete program. The resulting confusion matrices are available as Supplementary Tables S1a and S1b. From the confusion matrices, we calculated the precision and recall values for each localization site for both proteins annotated as having a single localization site (Table 4) and those annotated as having dual localization sites (Table 5).

On single localization proteins, PSORTb v.2.0 attained precision values of 96% for both classes of organisms, and recall of 83 and 81% for Gram-negative and Gram-positive proteins, respectively. We observed that the precision values remained relatively constant across localization sites, while the recall was highest for membrane proteins, most probably due to their conserved structural motifs readily identifiable by the frequent subsequence-based SVMs, HMMTOP and OMPMotif modules. The Gram-negative version of PSORTb v.2.0 exhibits a 0.7% decrease in precision relative to PSORTb v.1.1; however, an 8% increase in recall is observed.

Performance of the program on proteins annotated as having dual localization sites is comparable to the performance for singly localized proteins with respect to Gram-negative organisms, with a precision of 95% and a recall of 84%. However, we noted that the overall precision for Gram-positive multiply localized proteins was only 75%. Upon inspection, we realized that this was due to six annotated cytoplasmic membrane/cell wall proteins being predicted as cytoplasmic proteins. Noting that singly localized cytoplasmic membrane and cell wall proteins were infrequently mispredicted as cytoplasmic proteins, we investigated these six proteins further. Although experimental evidence supporting a possible cytoplasmic localization was found for only one of the six protiens—*Bacillus subtilis* ComGG (Chung *et al*., 1998)—the other five proteins include enzymes and heat-shock proteins, for which cytoplasmic- or peripheral membrane-associated localizations are not uncommon. It may be that rather than making mispredictions, PSORTb is detecting a more complex pattern of localization for certain proteins.

Table 5. PSORTb v.2.0 performance as measured by 5-fold cross-validation using the complete subset of multiply localized proteins from PSORTdb[a]

Localization	Performance				
	TP	FP	FN	Precision	Recall
Negative					
C/CM	11	2	5	84.6	68.8
CM/P	34	1	17	97.1	66.7
P/OM	2	2	0	50.0	100.0
OM/EC	76	1	2	98.7	97.4
Total	123	6	24	95.3	83.7
Positive					
C/CM	12	6	3	66.7	80.0
CM/CW	6	0	14	100.0	30.0
Total	18	6	17	75.0	51.4

[a] For a protein resident at X and Y localization sites, a true positive (TP) is a prediction of either X, Y or X/Y. A false positive (FP) is all multiply localized proteins not resident at X or Y that are predicted as X, Y or X/Y. A false negative (FN) is all X/Y proteins not predicted as neither X, Y nor X/Y.

Proteome coverage

The measured recall of a program when evaluated using 5-fold cross-validation does not give an accurate reflection of the predictive coverage when the program is applied to the analysis of whole proteomes. Because the training and testing data consist of a number of well-characterized proteins, a large number of predictions is possible. However, hypothetical proteins—which make up a notable proportion of a proteome—often do not contain enough information

Table 6. Comparison between PSORTb v.1.0, CELLO v.2.0 and Proteome Analyst v. using a set of 144 Gram-negative proteins not used in training of any program[a]

Localization	PSORTb v.2.0					CELLO v.2.0					Proteome Analyst v.1.0				
	TP	FP	FN	Precision	Recall	TP	FP	FN	Precision	Recall	TP	FP	FN	Precision	Recall
C	22	1	8	95.7	73.3	27	9	3	75.0	90.0	22	1	8	95.7	73.3
CM	39	1	3	97.5	92.9	35	4	7	89.7	83.3	40	6	2	87.0	95.2
P	26	0	6	100.0	81.3	16	6	16	72.7	50.0	29	1	3	96.7	90.6
OM	34	1	5	97.1	87.2	24	3	15	88.9	61.5	34	0	5	100.0	87.2
EC	1	0	0	100.0	100.0	1	19	0	5.0	100.0	1	6	0	14.3	100.0
Total	122	3	22	97.6	84.7	103	41	41	71.5	71.5	126	14	18	90.0	87.5

[a] See also supplementary table S3.

for a prediction to be generated. We therefore set out to measure PSORTb v.2.0's performance when applied to whole proteomes, with the expectation that we would see an increase in the 28% average coverage of version 1.1.

A total of 106 Gram-negative and 45 Gram-positive proteome files from the NCBI's Microbial Genomes page were analyzed (one organism may have multiple proteome files, each representing a different chromosome), and a complete summary of the results can be found in Supplementary Tables S2. The average coverage when PSORTb v.2.0 is applied to Gram-negative proteomes is 56.7%, and a maximum coverage of 78.8% (*Thermotoga maritima*) was achieved. When applied to Gram-positive proteomes, the average coverage increases to 74.8%, with a maximum of 83.2% (*Bacillus halodurans*).

The Gram-positive version of the program displays higher predictive coverage than the Gram-negative version due to the higher recall associated with the Gram-positive cytoplasmic SVM. Cytoplasmic proteins represent the largest class of proteins within the cell, and an improved ability to identify these results in high overall coverage. The level of coverage for each proteome appears to be irrespective of phylogenetic grouping, with predictions being generated as readily for organisms such as spirochetes or mollicutes as for the proteobacteria.

Comparison to other methods

We next set out to compare PSORTb v.2.0's performance with that of other comprehensive web-based predictive tools. Proteome Analyst (Lu *et al*., 2004) is capable of generating predictions for five Gram-negative localization sites and three Gram-positive sites—it does not differentiate between cell wall and extracellular proteins. CELLO (Yu *et al*., 2004) generates predictions for the five Gram-negative localization sites only. SubLoc (Hua and Sun, 2001) was not evaluated as it does not predict membrane proteins, and a comparison between PSORT I (Nakai and Kanehisa, 1991) and PSORTb appears in our earlier work (Gardy *et al*., 2003).

Because both Proteome Analyst and CELLO were trained using the original PSORTb dataset of 1443 Gram-negative proteins, a fair method of assessment was to use proteins not included in these programs' training data. A total of 144 singly localized new Gram-negative proteins in the version of PSORTdb described here were submitted to the Proteome Analyst and CELLO web servers for analysis. For a comparable evaluation of PSORTb v.2.0, the predictions generated for these proteins during the earlier 5-fold cross-validation procedure were used, such that the new proteins were not included in the PSORTb training data. A comparison of the performance of the three programs is given in Table 6, and the associated confusion matrices are available as Supplementary Table S3. A Gram-positive comparison was not carried out, as we were unsure whether Proteome Analyst's Gram-positive training data and that used in PSORTb overlapped.

In terms of precision, PSORTb v.2.0 outperforms both Proteome Analyst and CELLO by 7.6 and 26.1%, respectively. The significant difference between PSORTb and CELLO is due to the fact that unlike the other two programs, CELLO forces predictions for each query protein. While this does lead to a prediction generated for every protein in a proteome, the cost in terms of reliability of these predictions is significant. This decreased precision may not be apparent when evaluations are reported using the accuracy measure, in which high recall is able to compensate for lower precision, and illustrates that reporting confusion matrices leads to, epigrammatically enough, the least confusion when comparing the performance of multiple programs.

We also wished to compare the predictive coverage of PSORTb v.2.0 to that of the other programs when applied to the analysis of whole proteomes. Because CELLO generates a prediction in every case, it was not included in the present analysis. In Lu *et al*. (2004), the authors of Proteome Analyst report predictive coverage for two proteomes—one Gram-negative and one Gram-positive. Proteome Analyst displayed a coverage of 75.6% for the Gram-negative bacterium *Pseudomonas aeruginosa* and 67.2% for the Gram-positive bacterium *B.subtilis*. When PSORTb v.2.0 was used to analyze the same organisms, it attained a coverage of 68.1% for *P.aeruginosa* and 76.5% for *B.subtilis*.

An analysis based on these two proteomes suggests that while Proteome Analyst attains higher coverage on a Gram-negative organism, PSORTb v.2.0 generates more predictions for a Gram-positive proteome. Because Proteome Analyst relies on SWISS-PROT annotation keywords for classification, we believe that PSORTb's sequence feature-based method may yield higher coverage for organisms with little database annotation available.

Comparative proteome analysis

Using the data generated during our analysis of whole proteome predictive coverage (Supplementary Tables S2a and S2b), we investigated our hypothesis that free-living organisms might exhibit a higher than normal proportion of membrane proteins. The proportion

Table 7. Analysis of the proportion of predicted proteins at each localization site for 106 Gram-negative and 45 Gram-positive proteome files

Localization	Statistical analysis Average percentage of predictions	SD	Average percentage of proteome	SD	Correlation coefficient[a]
Negative					
C	59.3	5.4	33.3	5.5	0.97
CM	30.2	3.5	16.9	2.4	0.97
P	2.9	1.7	1.7	1.1	0.84
OM	4.5	3.1	2.4	1.6	0.72
EC	0.7	0.5	0.4	0.3	0.77
Positive					
C	68.5	3.9	50.7	3.5	0.99
CM	26.5	3.1	19.7	2.6	0.95
CW	1.2	0.6	0.9	0.5	0.44
EC	3.8	2.1	2.8	1.5	0.81

[a]Calculated between the number of proteins at localization X and the total number of proteins in the proteome.

of proteins at each localization site was determined, both as a fraction of the total predictions and as a fraction of total proteome size. Table 7 shows the average and standard deviation for both types of calculation, as well as the correlation coefficient between the number of proteins predicted at a given site and overall proteome size.

Cytoplasmic proteins and cytoplasmic membrane proteins represent the largest fractions of the proteome, and the large sample size yields a high correlation coefficient. This indicates that proteome size and not lifestyle or other factors is the primary determinant for the number of proteins at a given site. This correlation is evident for most other localization sites as well, with only the cell wall showing variable values. Because the cell wall represents a comparatively tiny fraction of the proteome, however, this variability may be attributed to a small sample size.

When the data are visualized as a scatter plot (Supplementary Figure S1), these constant proportions are more easily observable. Several points of interest also become obvious. Two Gram-negative organisms, *T.maritima* and *Aquifex aeolicus*—organisms that are found near the base of the tree of life—appear to have unusually high proportions of cytoplasmic proteins. The mycoplasmas, noted for their smaller genomes and membrane protein variability, exhibit higher than normal proportions of outer membrane proteins.

DISCUSSION

We have developed PSORTb v.2.0, an updated version of the PSORTb tool for the prediction of bacterial protein subcellular localization. Version 2.0 improves significantly upon the original release of the program, with its predictive capability extended to include Gram-positive organisms and its predictive coverage increased. A flag indicating potentially multiply localized proteins has also been added. PSORTb's existing standard of high precision is maintained, and with a measured precision of 96%, the program continues to be the most precise tool for bacterial localization prediction available.

We attribute the 2-fold increase in predictive coverage primarily to the incorporation of nine frequent subsequence-based SVM modules. With a higher precision than the amino acid composition-based SVMs such as SubLoc and CELLO, our method allows for the classification of proteins based on characteristic patterns that might not have been detected through conventional methods, such as multiple sequence alignment. The SVMs have allowed us to address concerns raised with the first release of PSORTb, namely how to identify a larger number of cytoplasmic proteins, as well as cytoplasmic membrane proteins with three or fewer predicted helices. Cytoplasmic proteins in particular are a large and diverse group of proteins and represent the majority of proteins encoded for by a genome. The ability to identify these proteins is the key to attaining a high predictive coverage rate, and we are interested in pursuing ways of increasing our SVM's ability to detect these proteins, particularly in the case of Gram-negative organisms.

Comparison with other available predictive tools shows that PSORTb remains the most precise predictive method available. We believe this is the result of two aspects of the program. First, a recent review has highlighted the importance of utilizing multiple methods for localization prediction (Schneider and Fechner, 2004), and PSORTb is one of the few localization prediction methods to take such an approach. Second, PSORTb does not force predictions—if we are unable to generate a confident prediction, the program will return a result of 'Unknown'. Instead of optimizing precision and recall, we have always chosen to emphasize precision in the development of PSORTb, reasoning that biologists are searching for correct results and, given the choice of an incorrect result or an 'Unknown' result, prefer 'Unknown'.

Comparative analysis also highlights the importance of publishing confusion matrices—different tools use different metrics in their reporting, and often the definition of a particular metric varies between groups. Without access to actual predictions, it can be difficult to objectively assess multiple predictive tools. Still, we show here that of the other localization tools currently available, Proteome Analyst offers an excellent complement to PSORTb—despite a slightly lower precision, Proteome Analyst's recall of Gram-negative proteins is especially good.

PSORTb's significant increase in predictive coverage allowed us to examine the distribution of proteins across localization sites on a proteome-wide scale. We found that with very few exceptions the proportions of proteins found at each localization site remained notably consistent across species, regardless of lifestyle, physiology or proteome size. This may reflect the nature of biological networks. When a new gene is introduced into an organism, its product will carry out certain new functions. In order for these new functions to be of any benefit to the organism, however, more new genes and gene products might be necessary—the proteins that will form a complete new functional pathway capable of interacting with other pathways in the cell. The new proteins that constitute the pathway will likely span different cellular compartments. For example, an organism taking up residence in a new environment may develop a series of membrane transporters to take up nutrients from and sense its surroundings; however, it must also develop the cytoplasmic components necessary for processing the incoming nutrients and signals.

PSORTb version 2.0 is available on the Web at http://www.psort.org/psortb. Users can submit one or multiple query sequences over the Web for analysis, selecting one of three possible output formats, or can download a standalone version of the program under the GNU General Public License. All of the completed microbial genomes listed on the NCBI have precomputed results available for download at http://www.psort.org/genomes/, and this Genomes page will be updated as new genomes are released. The PSORTdb dataset used in the development of the program is also available on the website, along with other resources for subcellular localization prediction, including links to other tools and datasets of interest. We believe that PSORTb v.2.0 and psort.org are valuable resources to the microbiology and localization prediction communities. We offer an open source, flexible and high coverage predictive tool, which is presently the most precise localization prediction method available. Utilizing the program for comparative proteome analysis has already generated interesting results regarding the proportion of proteins in different cellular compartments, and we are presently investigating this area further.

ACKNOWLEDGEMENTS

The authors would like to thank Cory Spencer for his work on the first version of PSORTb, as well as Rong She, Ke Wang and other members of the data mining group at Simon Fraser University's Department of Computing Science. J.L.G. and F.S.L.B. are a Michael Smith Foundation for Health Research Trainee and Scholar, respectively. Funding for this work was provided by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Chung,Y.S., Breidt,F. and Dubnau,D. (1998) Cell surface localization and processing of the ComG proteins, required for DNA binding during transformation of *Bacillus subtilis*. *Mol. Microbiol.*, **29**, 905–913.

Gardy,J.L., Spencer,C., Wang,K., Ester,M., Tusnády,G.E., Simon,I., Hua,S., deFays,K., Lambert,C., Nakai,K. and Brinkman,F.S.L. (2003) PSORTb: improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res.*, **31**, 3613–3617.

Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.

Hulo,N., Sigrist,C.J.A., Le Saux,V., Langendijk-Genevaux,P.S., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P. and Bairoch,A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.*, **32**, D134–D137.

Joachims,T. (2002) SVMLight.

Lin,C. (2003) LibSVM.

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.

Nakai,K. and Kanehisa,M. (1991) Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins*, **11**, 95–110.

Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, **10**, 1–6.

Schneider,G. and Fechner,U. (2004) Advances in the prediction of protein targeting signals. *Proteomics*, **4**, 1571–1580.

She,R., Chen,F., Wang,K., Ester,M., Gardy,J.L. and Brinkman,F.S.L. (2003) Frequent-subsequence-based prediction of outer membrane proteins. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, NY, pp. 436–445.

Tusnády,G.E. and Simon,I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849–850.

Vapnik,V. (1995) *The Nature of Statistical Learning Theory*. Springer, NY.

Wang,J.T., Chirn,G., Marr,T.G., Shapiro,B., Shasha,D. and Zhang,K. (1994) Combinatorial pattern discovery for scientific data: some preliminary results. In *Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data*. ACM Press, NY, pp. 115–125.

Yu,C., Lin,C. and Hwang,J. (2004) Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci.*, **13**, 1402–1406.