BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 10 2005, pages 2522–2524
doi:10.1093/bioinformatics/bti309

Sequence analysis

PSLpred: prediction of subcellular localization of bacterial proteins

Manoj Bhasin, Aarti Garg and G. P. S. Raghava*

Institute of Microbial Technology, Sector 39A, Chandigarh, India

Received on November 5, 2004; revised on February 2, 2005; accepted on February 3, 2005
Advance Access publication February 4, 2005

ABSTRACT

Summary: We developed a web server PSLpred for predicting subcellular localization of gram-negative bacterial proteins with an overall accuracy of 91.2%. PSLpred is a hybrid approach-based method that integrates PSI-BLAST and three SVM modules based on compositions of residues, dipeptides and physico-chemical properties. The prediction accuracies of 90.7, 86.8, 90.3, 95.2 and 90.6% were attained for cytoplasmic, extracellular, inner-membrane, outer-membrane and periplasmic proteins, respectively. Furthermore, PSLpred was able to predict ∼74% of sequences with an average prediction accuracy of 98% at RI = 5.

Availability: PSLpred is available at http://www.imtech.res.in/raghava/pslpred/

Contact: raghava@imtech.res.in

Supplementary information: http://www.imtech.res.in/raghava/pslpred/supl.html

INTRODUCTION

A number of methods such as PSORT I, PSORT-B and NNPSL have been developed for predicting subcellular localization of bacterial proteins based on different datasets and computational techniques (Nakai and Kanehisa, 1991; Gardy *et al*., 2003; Reinhardt and Hubbard, 1998). The accuracies reported by these methods vary between 60 and 81%. Recently, a support vector machines (SVM) based method, CELLO (Yu *et al*., 2004) trained using n-peptide compositions has been developed for predicting subcellular localization of bacterial proteins. This method has achieved an overall accuracy of 89% that is better than existing methods for subcellular localization of prokaryotic proteins. Despite the overall improved performance, CELLO predicts extracellular proteins with a fair accuracy of 78.9%, proteins that may represent important virulence factors in pathogenic microorganisms. Therefore, in the present study, a systematic attempt has been made to achieve higher prediction accuracy for subcellular localizations of prokaryotic proteins using different features of proteins.

The data set used in the present study is the same as that used by Yu *et al*. (2004) and Gardy *et al*. (2003) for developing the methods CELLO and PSORT-B, respectively. This data set has been generated from SWISSPROT release 40.29 (Bairoch and Apweiler, 2000), and consisted of a total of 1443 proteins belonging to different subcellular localizations. We have excluded 141 proteins residing in more than one subcellular location and used the remaining 1302 proteins (248 cytoplasmic, 268 inner membrane, 244 periplasmic, 352 outer membrane and 190 extracellular) for the development of the present method.

In this study, a machine-learning technique, SVM, has been used for the prediction of subcellular localizations of prokaryotic proteins. The prediction of subcellular localizations is a multi-class classification problem. Thus we used a one-versus-rest (1-v-r SVM) strategy, in which the ith SVM has been trained with all the samples in the ith class with a positive label and samples in the remaining classes with a negative label (Hua and Sun, 2001). The performance of the SVM modules developed in the present study was evaluated through 5-fold cross-validation technique. In this technique, the relevant dataset is partitioned randomly into five equal sized sets. The training and testing were carried out five times, using one distinct set for each testing and the remaining (four sets) for the training. In order to assess the predictive performance, accuracy and Matthew's correlation coefficient (MCC) (Matthews, 1975) have been calculated using Equations (1) and (2), respectively.

$$\text{Accuracy}(x) = \frac{p(x)}{\text{Exp}(x)}, \tag{1}$$

$$\text{MCC}(x) = \frac{p(x)n(x) - u(x)o(x)}{\sqrt{[p(x)+u(x)][p(x)+o(x)][n(x)+u(x)][n(x)+o(x)]}}, \tag{2}$$

where x can be any subcellular location, $\text{Exp}(x)$ is number of sequences observed in location x, $p(x)$ is number of correctly predicted sequences of location x, $n(x)$ the number of correctly predicted sequences not of location x, $u(x)$ the number of under-predicted sequences and $o(x)$ the number of over-predicted sequences.

In addition, a reliability index (RI) assignment (Bhasin and Raghava, 2004; Hua and Sun, 2001), which measures the level of certainty in the prediction, has been calculated using Equation (3).

$$\text{RI} = \begin{cases} \text{INT}(\Delta^* 5/3 + 1, & \text{if } 0 \le \Delta < 4, \\ 5, & \text{if } \Delta \ge 4, \end{cases} \tag{3}$$

where Δ is the difference between the highest and second highest SVM output scores.

The detailed performance of all the SVM modules developed in the present study is shown in Table 1. The first SVM module 'A', based on amino acid compositions, has achieved an overall accuracy

*To whom correspondence should be addressed.

Table 1. Detailed performance of SVM modules developed using different features of proteins and PSI-BLAST[a]

	Amino acid composition (A)		Dipeptide composition (B)		Physico-chemical properties composition (C)		ProPSI-BLAST (D)		Hybrid1 (A + B + C)		Hybrid2 (A + B + C + D)	
	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
Cytoplasm	87.1	0.80	87.1	0.78	83.5	0.77	34.3	—	85.9	0.81	90.7	0.86
Extracellular	77.9	0.81	73.7	0.79	75.8	0.77	79.5	—	80.0	0.83	86.8	0.88
Inner-membrane	86.9	0.87	85.8	0.89	85.8	0.83	59.7	—	89.2	0.88	90.3	0.90
Outer-membrane	93.5	0.87	93.8	0.88	87.8	0.82	93.8	—	94.3	0.89	95.2	0.95
Periplasmic	79.9	0.76	84.0	0.77	78.3	0.73	65.5	—	82.8	0.78	90.6	0.84
Overall	86.1	0.83	86.1	0.83	83.0	0.79	68.1	—	87.4	0.84	91.2	0.89

[a]ACC: accuracy; MCC: Matthew's correlation coefficient.

Table 2. Comparison of the performance of the present method (PSLpred) with CELLO and PSORT-B[a]

	PSLpred		CELLO		PSORT-B	
	ACC	MCC	ACC	MCC	ACC	MCC
Cytoplasm	90.7	0.86	90.7	0.85	69.4	—
Extracellular	86.8	0.88	78.9	0.82	70.0	—
Inner-membrane	90.3	0.90	88.4	0.92	78.7	—
Outer-membrane	95.2	0.95	94.6	0.90	90.3	—
Periplasmic	90.6	0.84	86.9	0.80	57.6	—
Overall	91.2	0.89	88.9	—	74.8	—

[a]ACC: accuracy; MCC: Matthew's correlation coefficient.

of 86.1%. However, amino acid composition provides information only about frequency but not about local order of residues. Therefore, in order to incorporate information about frequency as well as local order of residues of proteins, SVM module 'B' based on dipeptide compositions was constructed. Dipeptide composition is considered to be an improved feature in comparing amino acid composition as it encapsulates global (residue frequency) as well as local information (residue order) of the protein sequence (Bhasin and Raghava, 2004). Interestingly, in the present study, a dipeptide composition-based SVM module was unable to enhance the prediction accuracy beyond 86.1%, which is similar to the accuracy of the amino acid composition-based SVM module. Furthermore, another SVM module 'C' based on 33 physico-chemical properties, where the input vector has 33 scalar values, each representing an average value of a distinct physico-chemical property of a protein (Bhasin and Raghava, 2004) has also been constructed. An overall accuracy of 83% was achieved using module 'C' which is lower in comparison to modules 'A' and 'B'.

In this study, a similarity search-based module ProPSI-BLAST 'D' has also been constructed, in which the query sequence is searched against a database of bacterial proteins using PSI-BLAST. PSI-BLAST is used instead of the normal standard BLAST because it has the capability to detect remote homologies (Altschul *et al*., 1990). Three iterations of PSI-BLAST have been carried out at a cut-off E-value of 0.001. This module can predict any of the five localizations depending upon the similarity of the query sequence to the proteins present in the database and returns 'unknown subcellular localization', if no significant similarity is obtained. However, the performance of this module was found to be poor compared to composition-based SVM modules (Table 1).

It has been observed earlier that a hybrid approach-based SVM module performed better than individual feature-based modules (Bhasin and Raghava, 2004). Hence, we have also adopted the same strategy and developed an SVM module hybrid1 using features that have been used to develop modules 'A', 'B' and 'C'. The performance of the hybrid1 module has been found to be better than individual feature-based SVM modules (Table 1). Further to improve the performance, another SVM module hybrid2 was developed based on all features of a protein and the output of PSI-BLAST. In the hybrid2 module, SVM was provided with an input vector of 459 dimensions that consisted 20 for amino acid composition, 33 for physico-chemical properties, 400 for dipeptide composition and 6 for PSI-BLAST output. The hybrid2 module achieved an overall accuracy of 91.2% with a minimum of 86.8% for extracellular proteins (Table 1). It proved that the hybrid2 module was able to encapsulate more comprehensive information, which successfully improved the prediction accuracy. These results confirmed that the prediction accuracy of subcellular localization of proteins can be increased using a wide range of information about it. In order to provide confidence in predictions, RI was also computed for the hybrid2 SVM module. It was observed that this method is able to predict subcellular localization of ~74% of protein sequences with an accuracy of 98% for RI = 5.

As shown in Tables 1 and 2, the accuracy achieved by our method is nearly 2 and 16% higher than the CELLO (Yu *et al*., 2004) and PSORT-B (Gardy *et al*., 2003) methods, respectively; hence,

the PSLpred method is better than existing methods for predicting prokaryotic subcellular localizations. In general, for all the five subcellular localizations, the performance of our method is better than the existing methods. The prediction accuracy achieved for cytoplasmic proteins by our method is similar to CELLO but 22% higher than that of PSORT-B. Noticeably, for the inner membrane, our method has achieved 90.3% accuracy, which is ~2 and 12% higher than CELLO and PSORT-B, respectively. In case of periplasmic proteins, the prediction accuracy of our method is 4 and 33% higher compared to the accuracy of the CELLO and PSORT-B methods, respectively. However, for the outer membrane, the prediction accuracy of our method is similar to the CELLO method but it is 5% higher compared to the accuracy of PSORT-B. For extracellular proteins, accuracy has been increased from 78.9% (CELLO) to 86.8% for PSLpred.

Recently, the SubLoc method, which uses amino acid composition as an input to SVM, has been evaluated on the current data set; it achieved an overall accuracy of 78.5% (Yu *et al*., 2004). In comparison, we have been able to achieve a highest accuracy of 86.1% with an amino acid composition-based SVM module. In order to verify the results, we again analyzed and re-examined the performance of the amino acid composition-based SVM module at default as well as at different parameters of various SVM kernel functions. The detailed results can be obtained from Tables S1, S2A and S2B of supplementary material (http://www.imtech.res.in/raghava/pslpred/supl.html). SignalP is another powerful tool for predicting secretory proteins (Bendtsen *et al*., 2004). Recently, the developers of SignalP have checked the performance of the SignalP 3.0 method on the current data set, by categorizing all proteins as secretory except cytoplasmic proteins and reported an overall accuracy of 95% (Bendtsen *et al*., 2004). Though the performance of SignalP has been found to be better than PSLpred, one cannot compare PSLpred and SignalP, as the number of classes predicted by these two methods is different.

In summary, PSLpred can complement all the existing subcellular localization prediction methods and can assist in the development of automated genome annotation tools. All the SVM modules constructed in the present study have been implemented as web server (www.imtech.res.in/raghava/pslpred/) using CGI/Perl script. The present version of PSLpred predicts only a single localization site. The information about various SVM modules constructed in the present study can be obtained from supplementary material (www.imtech.res.in/raghava/pslpred/supl.html).

ACKNOWLEDGEMENTS

The authors are thankful to Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology, Government of India, for financial assistance. This report has IMTECH communication number 52/2004.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequences database and its supplement TrEMBL in 2000. *Nucl. Acids Res.*, **28**, 45–48.

Bendtsen,J.D. *et al.* (2004) Improved prediction of signal peptides-SignalP 3.0. *J. Mol. Biol.*, **16**, 783–795.

Bhasin,M. and Raghava,G.P.S. (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucl. Acids Res.*, **32**, 415–419.

Gardy,J.L. *et al*. (2003) PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucl. Acids Res.*, **31**, 3613–3617.

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

Matthews,B.W. (1975) Comparison of predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

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

Reinhardt,A. and Hubbard,T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucl. Acids Res.*, **26**, 2230–2236.

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