Available online at www.sciencedirect.com
ScienceDirect
Polymer 48 (2007) 5493–5496
polymer
www.elsevier.com/locate/polymer

Polymer Communication

The RPSP: Web server for prediction of signal peptides

Dariusz Plewczynski [a,*], Lukasz Slabinski [b], Adrian Tkacz [b], Laszlo Kajan [b], Liisa Holm [c], Krzysztof Ginalski [a], Leszek Rychlewski [b]

[a] *Interdisciplinary Centre for Mathematical and Computational Modelling, Warsaw University, Pawinskiego 5a, 02-106 Warsaw, Poland*
[b] *BioInfoBank Institute, Limanowskiego 24a, 60-744 Poznan, Poland*
[c] *Bioinformatics Group, Institute of Biotechnology, University of Helsinki, 00014 Helsinki, Finland*

Received 22 May 2007; received in revised form 17 July 2007; accepted 19 July 2007
Available online 24 July 2007

Abstract

The RPSP is a fast web service for detection of signal peptides in proteins. The method uses neural networks trained on known signal peptides from the Swiss-Prot protein database. The web server works either on prokaryotic and eukaryotic proteins or without specifying an organism type. The accuracy of the web server is similar to other available computational prediction web services, yet because of its speed and portability the method can be easily applied to whole proteomes. The RPSP web server is available at http://rpsp.bioinfo.pl.

Keywords: Bioinformatics; Signal peptides; Protein sequences

1. Introduction

Signal peptides determine the destination of newly synthesized proteins in a cell [10] and modulate cellular life, for example, by controlling the entry of proteins to the secretory pathway [6,8,10,27]. Signal peptides are short sequence fragments that are cleaved off while the protein is transported via a membrane. The application of regular expression search provides the simplest computational approach for an identification of signal peptides, where regular expressions are constructed from experimentally verified signal peptides in proteins [26]. The efficiency of prediction may be improved by applying context-based rules and various logical filters [26]. The detection of signal peptides can be also done by a weight matrix approach [20]. In this approach cleavage sites are characterized by a set of simple rules for recognition of differences between a signal sequence and the mature exported protein [32,33,20]. Recently more advanced methods were presented that make use of various machine learning algorithms. The list includes neural networks [22,23], support vector machines [31], hidden Markov models [21] and many others [4,5,7,11–16,18,20–23,25,29–31,34]. Most of these methods correctly classify sequences either as secretory or as non-secretory but do not provide cleavage site assignment, and lack correct assignment of the 5′-end of genes [28].

The most commonly used computational methods for detection of signal peptides are the SignalP [4] and SPEPlip [9]. The SignalP uses three neural networks combined with Hidden Markov Model in order to predict signal peptides in Gram-positive and Gram-negative bacteria, and eukaryotes, whereas SPEPlip applies two neural networks that are trained separately on signal peptides from eukaryotes and prokaryotes. In this manuscript we present similar approach, based on neural networks trained on the newest version of Swiss-Prot database that is significantly faster than previously developed methods and can be used in large-scale predictions of signal peptides. The RPSP algorithm (rapid prediction of signal peptides) is publicly available as a web service at http://rpsp.bioinfo.pl. Our method focuses on the classical types of signal peptides neglecting the non-classically secreted proteins [2,3]. Three types of *in silico* prediction can be performed: for prokaryotic sequences, eukaryotic sequences and without specifying the organism type.

* Corresponding author. Tel.: +48 22 554 08 39; fax: +48 22 554 08 01.
E-mail address: D.Plewczynski@icm.edu.pl (D. Plewczynski).

0032-3861/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.polymer.2007.07.039

2. Method

The machine learning algorithm is trained here on protein sequences acquired from the Swiss-Prot database (release: 49.4) that is known to contain the signal peptides. Firstly all Swiss-Prot entries with a keyword 'SIGNAL' in FT line (20 863 entries) were extracted. Uncertain entries marked in FT line as potential, probable or by similarity were removed (4566 entries left), then archaeal and viral proteins were also removed (4296 entries left). The resulting dataset was split into two groups: eukaryotic sequences (3331 entries) and prokaryotic ones (965 entries). In the case of eukaryotes all untypical entries were discarded (organelle proteins, signal peptide sequences shorter than 15 and longer than 45 amino acids and those with residues other than A, C, G, L, P, Q, S, T at '−1' position). In the case of prokaryotic sequences all lipoproteins, signal peptide sequences shorter than 15 and longer than 50 amino acids, or those with residues other than A, G, S, T at '−1' position were removed. Resulting datasets comprised the positives used in the training, whereas the negatives were prepared by extracting N-terminal parts (70-residue long sequence fragments) of eukaryotic cytoplasmic and nuclear proteins for eukaryotes and N-terminal parts of bacterial cytoplasmic proteins for prokaryotes. The datasets were reduced later at 60% sequence identity for whole protein sequences using the CD-HIT clustering tool [17]. The selected cut-off was selected as it provides both the best results [24] and moderate memory requirements for training of neural network. The resulting training datasets contain 1784 positives for eukaryotes and 646 for prokaryotes, 987 eukaryotic negatives from cytoplasm and 2265 from nucleus, and 2040 prokaryotic negatives. In order to avoid the bias during training and testing of the neural networks, the negative datasets were reduced approximately to the sizes of positive datasets. The neural networks were trained with six-fold cross-validation on three training sets (separately for eukaryotes, prokaryotes and mixed) [1]. All training datasets used to build the RPSP method are available on the server web pages and can be used for training of different types of machine learning algorithms such as support vector machine.

Because the cleavage site position is strongly correlated with the amino acid composition of the signal peptide [22,23] the local sequence information is sufficient as an input to the neural network. Two neural networks with feed-forward, multi-layer architecture and back-propagation learning algorithm are used here. The first network determines if a given residue belongs to the signal peptide or not. A symmetric sliding window with 27 amino acids for eukaryotic and mixed origin sequences, and 19 amino acids for prokaryotic sequences is used as an input for the neural network. We neglect differences between Gram-negative and Gram-positive bacteria [22,23] as this information is not readily available especially in bioinformatics screening of large sequence datasets. The output layer is built from a single neuron calculating the S-score of a prediction. High score corresponds to higher probability that a given amino acid belongs to a signal peptide, and low score indicates that the amino acid is a part of a mature protein. The second neural network recognizes the cleavage site. The input for this neural network is given as an asymmetric sliding window with 24 residues for prokaryotic/eukaryotic and 25 amino acids for mixed origin sequences. The output layer (the single neuron) provides the C-score of a prediction. This score describes the cleavage site likelihood for each residue in the query sequence. This score is higher at the cleavage site than for other parts of protein sequence. The final discrimination between signal peptide and non-signal peptide together with cleavage site prediction is given by Y-score that combines both the previously mentioned scores. This procedure is similar to SignalP [4] and SPEPlip [9] algorithms and we use the same name convention of various scores as in the SignalP and SPEPlip papers. The final Y-score is equal to: $Y_i = \sqrt{C_i \times \Delta_d S_i}$, where $\Delta_d S_i$ is the difference between the mean S-score for all d amino acids before and after position i. The d value of 17 was taken from our benchmarking results. The high values of C-score can be assigned to various amino acids in the query sequence, whereas only the single residue can be the true cleavage site. As a consequence, cleavage site is predicted for the highest Y-score, which means that the slope of the S-score is steep and significant C-score is found. Therefore Y-score provides a better cleavage site prediction than the raw C-score alone. In addition, the D-score is calculated as the arithmetic mean value of Y-score for position i and mean value of S-score for all amino acids before it. This score was shown previously [4] to be superior in discriminating between secretory and non-secretory proteins in comparison with the S-mean score used in previous approaches. Protein is expected to contain a signal peptide in considered position i if Y-score for this position is larger than 0.35 and D-score is larger than 0.43 [4].

3. Results

We developed a rapid method for signal peptide detection that can be applied for large-scale annotations of heterogeneous sets of sequences and that does not require specifying their origin. The performance of three neural networks trained separately on prokaryotic, eukaryotic and those of mixed origin sequences were conducted on independent test sets that were not used during the learning procedure. Detailed benchmark results are shown in Table 1. The high efficiency of signal peptide prediction even without specifying the organism of a protein is a strong asset of our approach. The precision of the method operating without distinguishing between prokaryotic and eukaryotic proteins is not significantly lower than that using separate neural networks trained either on eukaryotic or prokaryotic sequences. The overall classification error of cleavage site prediction reaches 0.7 on heterogeneous data that contains both prokaryotic and eukaryotic sequences, while the accuracy of discrimination between signal peptides and non-signal peptides is above 0.9. As one can see from Table 1 our results are comparable with those that can be obtained with other prediction tools such as SignalP 3.0 [4] or SPEPlip [9].

Another crucial advantage of the RPSP is that the method is very fast. For example, the analysis of 959 proteins takes about 2 s on a Linux machine with 2 GHz CPU and 512 MB RAM.

Table 1
Results of neural network prediction on independent benchmarking datasets

Organism type	Discrimination					Cleavage site prediction [error]	SignalP2[a] cleavage site prediction [ratio of correct predictions]
	Precision	Recall	Accuracy	Matthews correlation	SignalP2[a] correlation		
Eukaryotes	0.9	1.0	0.9	0.9	0.9	0.8	0.7
Prokaryotes	0.9	1.0	0.9	0.9	0.8	0.8	0.7
Eukaryotes and prokaryotes	0.9	1.0	0.9	0.8	–	0.7	–

[a] Results of SignalP2 performance on independent benchmarking dataset from Swiss-Prot version 29 with signal peptides and non-secretory, i.e. cytoplasmic or nuclear, proteins after redundancy reduction. No results for both eukaryotes and prokaryotes (mixed organism type) are available.

The genome-wide signal peptide predictions run on Linux machine with 2 GHz CPU and 512 MB RAM on two genomes *Plasmodium falciparum* and *Chlamydophila pneumoniae* take, respectively, 7 s and 3 s. For those two genomes SignalP 3.0 web server needs 263 s and 127 s, whereas SPEPlip web server needs 203 s and 97 s. The availability of free local version with the source code is the crucial advantage over the other previously developed algorithms that provide only web server interfaces. The web server technology has some inherent limits due to internet architecture and technical design. The existing signal peptide prediction servers cannot accept input of more than around 1000 proteins, all have to contain less than certain limit of residues per sequence (few thousands) and there is also the limit for the number of residues in total. Additionally existing servers have also the limit of few thousand lines in the input file, and the number of jobs accepted from single IP internet address. For example, in the case of both *P. falciparum* (5365 proteins) and *C. pneumoniae* (1113 proteins) proteomes existing servers return ''webserver error'', i.e., job is rejected due to exceeded sequence and memory limits. On the contrary, the RPSP server is designed for high-throughput analyses and, in addition, the method is also available as a stand-alone program. Therefore it can be compiled in much more effective way on user workstation using standard C/C++ compiler. User thus can get easier and faster way to perform high-throughput screening of large sets of proteins. Altogether, these make RPSP the method of choice in high-throughput studies, such as massive analyses of whole proteomes in the context of function prediction or detailed characterization of proteins.

4. Web server

The RPSP web server is publicly available at http://rpsp.bioinfo.pl. The RPSP accepts protein sequences in FASTA format, with additional letter X for marking empty and unknown positions in a sequence or positions that extend a sequence segment outside chain ends. User can input sequences by submitting text file or by pasting the sequences in the text box. Three types of prediction can be performed: for prokaryotic sequences, eukaryotic sequences and without specifying the organism type. The prediction results are sent to a user by e-mail or provided on the server's output web pages. The output page contains, for each submitted protein, its name and detected signal peptide with its sequence and length. In case when input contains a single sequence, an annotated figure highlighting the signal peptide and predicted secondary structure elements (by PSIPRED [19]) is shown. For multiple FASTA records, annotations for the first 50 sequences are only presented. A link to the full RPSP prediction is provided for download and subsequent automated parsing.

The RPSP source code in C programming language together with LINUX precompiled binary can be freely downloaded from http://bioinfo.pl/RPSP.tar.gz. Consequently, predictions can be run locally on any typical workstation and can be used in large-scale analyses. The training dataset, based on the new version of the Swiss-Prot database, is freely available at http://rpsp.bioinfo.pl/training/RPSPdata.tar.gz web page, and can be used by users to train their own machine learning algorithms.

Acknowledgements

This work was supported by EC within BioSapiens (LHSG-CT-2003-503265) and SEPSDA (SP22-CT-2004-003831) 6FP projects, EMBO Installation Grant, Polish Ministry of Education and Science (PBZ-MNiI-2/1/2005), and Foundation for Polish Science (FOCUS).

References

[1] Baldi P, Brunak S. Bioinformatics: the machine learning approach. 2nd ed. Cambridge, MA: MIT Press; 2001.
[2] Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S. Feature-based prediction of non-classical and leaderless protein secretion. Protein Eng Des Sel 2004;17(4):349–56.
[3] Bendtsen JD, Kiemer L, Fausboll A, Brunak S. Non-classical protein secretion in bacteria. BMC Microbiol 2005;5:58.
[4] Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 2004;340(4):783–95.
[5] Bendtsen JD, Nielsen H, Widdick D, Palmer T, Brunak S. Prediction of twin-arginine signal peptides. BMC Bioinformatics 2005;6:167.
[6] Bruch MD, McKnight CJ, Gierasch LM. Helix formation and stability in a signal sequence. Biochemistry 1989;28(21):8554–61.
[7] Chou KC. Prediction of signal peptides using scaled window. Peptides 2001;22(12):1973–9.
[8] Cornell DG, Dluhy RA, Briggs MS, McKnight CJ, Gierasch LM. Conformations and orientations of a signal peptide interacting with phospholipid monolayers. Biochemistry 1989;28(7):2789–97.
[9] Fariselli P, Finocchiaro G, Casadio R. SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. Bioinformatics 2003;19(18):2498–9.
[10] Gierasch LM. Signal sequences. Biochemistry 1989;28(3):923–30.
[11] Hiller K, Grote A, Scheer M, Munch R, Jahn D. PrediSi: prediction of signal peptides and their cleavage positions. Nucleic Acids Res 2004;32(Web Server issue):W375–9.

[12] Juncker AS, Willenbrock H, Von Heijne G, Brunak S, Nielsen H, Krogh A. Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci 2003;12(8):1652–62.
[13] Kall L, Krogh A, Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. J Mol Biol 2004;338(5):1027–36.
[14] Ladunga I, Czako F, Csabai I, Geszti T. Improving signal peptide prediction accuracy by simulated neural network. Comput Appl Biosci 1991; 7(4):485–7.
[15] Lao DM, Arai M, Ikeda M, Shimizu T. The presence of signal peptide significantly affects transmembrane topology prediction. Bioinformatics 2002;18(12):1562–6.
[16] Lao DM, Okuno T, Shimizu T. Evaluating transmembrane topology prediction methods for the effect of signal peptide in topology prediction. In Silico Biol 2002;2(4):485–94.
[17] Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 2001;17(3):282–3.
[18] Liu L, Li J, Tian X, Ren D, Lin J. Information theory in prediction of cleavage sites of signal peptides. Protein Pept Lett 2005;12(4):339–42.
[19] McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics 2000;16(4):404–5.
[20] Menne KM, Hermjakob H, Apweiler R. A comparison of signal sequence prediction methods using a test set of signal peptides. Bioinformatics 2000;16(8):741–2.
[21] Nielsen H, Brunak S, von Heijne G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. Protein Eng 1999;12(1):3–9.
[22] Nielsen H, Engelbrecht J, Brunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 1997;10(1):1–6.
[23] Nielsen H, Engelbrecht J, Brunak S, von Heijne G. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Int J Neural Syst 1997;8(5–6): 581–99.
[24] Nielsen H, Engelbrecht J, von Heijne G, Brunak S. Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site. Proteins 1996;24(2):165–77.
[25] Nielsen H, Krogh A. Prediction of signal peptides and signal anchors by a hidden Markov model. Proc Int Conf Intell Syst Mol Biol 1998;6:122–30.
[26] Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, et al. ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 2003;31(13):3625–30.
[27] Rapoport TA. Transport of proteins across the endoplasmic reticulum membrane. Science 1992;258(5084):931–6.
[28] Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 1998;26(9):2230–6.
[29] Sidhu A, Yang ZR. Prediction of signal peptides using bio-basis function neural networks and decision trees. Appl Bioinformatics 2006;5(1): 13–9.
[30] Talmud P, Lins L, Brasseur R. Prediction of signal peptide functional properties: a study of the orientation and angle of insertion of yeast invertase mutants and human apolipoprotein B signal peptide variants. Protein Eng 1996;9(4):317–21.
[31] Vert JP. Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings. Pac Symp Biocomput 2002;649–60.
[32] von Heijne G. Net N–C charge imbalance may be important for signal sequence function in bacteria. J Mol Biol 1986;192(2):287–90.
[33] von Heijne G. A new method for predicting signal sequence cleavage sites. Nucleic Acids Res 1986;14(11):4683–90.
[34] Zhang Z, Henzel WJ. Signal peptide prediction based on analysis of experimentally verified cleavage sites. Protein Sci 2004;13(10):2819–24.