DOI: 10.1093/nar/gkh378

PrediSi: prediction of signal peptides and their cleavage positions

Karsten Hiller[1], Andreas Grote[1], Maurice Scheer[1,2], Richard Münch[1] and Dieter Jahn[1,*]

[1]Institut für Mikrobiologie, Technische Universität Braunschweig, Spielmannstrasse 7, D-38106 Braunschweig, Germany and [2]Fachbereich für Informatik, Fachhochschule Wolfenbüttel, Am Exer, D-38302 Wolfenbüttel, Germany

Received February 13, 2004; Revised and Accepted March 15, 2004

ABSTRACT

We have developed PrediSi (Prediction of Signal peptides), a new tool for predicting signal peptide sequences and their cleavage positions in bacterial and eukaryotic amino acid sequences. In contrast to previous prediction tools, our new software is especially useful for the analysis of large datasets in real time with high accuracy. PrediSi allows the evaluation of whole proteome datasets, which are currently accumulating as a result of numerous genome projects and proteomics experiments. The method employed is based on a position weight matrix approach improved by a frequency correction which takes in to consideration the amino acid bias present in proteins. The software was trained using sequences extracted from the most recent version of the SwissProt database. PrediSi is accessible via a web interface. An extra Java package was designed for the integration of PrediSi into other software projects. The tool is freely available on the World Wide Web at http://www.predisi.de.

INTRODUCTION

Signal peptides direct proteins to their proper cellular and extracellular locations (1). One major example of such a process is the translocation of proteins across the cytoplasmic membrane via the well-established sec pathway found in both eukaryotic and prokaryotic cells (2). In this secretory pathway, proteins designated for export from the cell are labeled by an N-terminal signal sequence. This signal sequence directs its protein to the secretion apparatus. After translocation of the protein across the cell membrane, the N-terminal signal peptide is usually cleaved off by an extracellular signal peptidase. Signal peptides for the sec pathway generally consist of the following three domains: (i) a positively charged n-region, (ii) a hydrophobic h-region and (iii) an uncharged but polar c-region. The cleavage site for the signal peptidase is located in the c-region (3). However, the degree of signal sequence conservation and length, as well as the cleavage site position, varies significantly between different proteins. Moreover, major differences were observed between eukaryotic and bacterial signal sequences. For various purposes it is desirable to identify signal peptides and their corresponding cleavage positions. For the calculation of sequence length-dependent features such as the molecular weight and the isoelectric point of a protein, the presence or absence of a signal peptide leads to considerably different results. We used the SignalP (4,5) signal peptide prediction tool in combination with the proteomics software JVirGel (6) to improve the calculation of virtual two-dimensional (2D) protein gels with respect to the position of protein spots. However, the resulting application was time consuming and limited to 10 requests of up to 2000 sequences per day using SignalP's free version via the Internet (http://www.cbs.dtu.dk/services/SignalP-2.0/). Moreover, most of the existing prediction tools for the analysis of huge datasets such as whole proteomes are either based on old training datasets or not freely accessible. Finally, a recent evaluation of signal peptide prediction programs revealed that the majority of available tools do not meet today's standards of performance and compatibility (7). Therefore, we set out to develop a new piece of software including the following features: (i) accurate and fast prediction of signal peptides and their corresponding cleavage positions, (ii) a user-friendly web interface, freely available on the World Wide Web for the analysis of unlimited datasets, (iii) presentation of the results in user- as well as computer-friendly formats such as HTML, XML and CSV and (iv) free availability as a Java package for integration into other software projects.

SYSTEM AND METHODS

Dataset of secreted proteins with experimentally determined cleavage positions

For the generation of the position weight matrices (PWMs) of PrediSi, datasets of secreted proteins with experimentally determined cleavage positions were constructed. Three

*To whom correspondence should be addressed. Tel: +49 531 391 5801; Fax: +49 531 391 5854; Email: d.jahn@tu-bs.de

different datasets were employed: one set for eukaryotes, one for Gram-negative and one for Gram-positive bacteria. Amino acid sequences with annotated signal peptides were extracted from the XML version of SwissProt release 42.9 (8). All proteins denoted as 'fragments', 'putative', 'found by similarity', 'probable' or with similar descriptions were removed. Furthermore, all proteins from organelles were excluded. From the prokaryotic datasets, signal peptides which are subject to signal peptidase II cleavage were excluded. The training datasets were aligned according to the annotated experimentally determined cleavage position of each sequence.

In parallel, we constructed control datasets of cytoplasmic and nuclear proteins which are clearly devoid of secretory signal peptides for the sec pathway. For this purpose amino acid sequences of proteins with determined appropriate cellular location were extracted from the SwissProt database. Sequences consisting of protein fragments shorter than 70 amino acids or indicated with comments such as 'potential' or 'probable' were excluded.

Identical sequences with regard to the initial 100 N-terminal amino acids were eliminated from all datasets. Integration of similar amino acid sequences that differ only in a few amino acids increased the performance of the self-consistency test. All generated datasets are available for download and as supplementary information (http://www.predisi.de/download.html).

The resulting training datasets consist of 2783 amino acid sequences from eukaryotes, 557 sequences from Gram-negative bacteria and 236 sequences from Gram-positive bacteria. The control datasets consist of 5547 amino acid sequences from eukaryotes, 2013 sequences from Gram-negative bacteria and 1077 sequences from Gram-positive bacteria.

Algorithms

The algorithm employed is based on a position weight matrix approach. We generated three different frequency matrices built on the constructed and aligned datasets described above. The position weight matrices are based on the amino acid frequency of parts of the signal sequences in addition to up to four amino acid residues from the N-terminus. We estimated the optimal size of the PWMs by calculating the accuracy of all meaningful combinations. Before calculating the score, we applied a frequency correction to adjust the amino acid bias present in proteins (9). The score was calculated according to Equation 1. We simplified the frequency correction by determining the amino acid distribution within only one group of organisms (eukaryotes, Gram-negative, Gram-positive bacteria). The group-specific amino acid composition was estimated via calculating the amino acid frequency of all the proteins in the corresponding control dataset.

$$S = \sum_{i=1}^{l_{\mathrm{PWM}}} \log\left(P_i \frac{P_{\mathrm{ideal}}}{P_{\mathrm{obs}}}\right), \qquad \mathbf{1}$$

where S is the score, P_i is the observed amino acid frequency at position i, P_{ideal} is set to 0.05 (statistical ideal amino acid frequency) and P_{obs} is the observed amino acid frequency.

Web interface

The main program for signal peptide prediction was written in Java (http://java.sun.com) to take advantage of its object-oriented technology and to allow integration of its output into dynamic web sites using Java Server Page (JSP) technology. Using this strategy it was possible to smoothly combine and reuse the Java classes with JSP. Jakarta Tomcat was chosen as the servlet container and web server (stable release, version 4.1.29). It is the official reference implementation of the Java Servlet (version 2.3) and JSP (version 1.2) technologies and is available as an open source tool (http://jakarta.apache.org/tomcat). Besides these Java packages the javax.servlet was employed for Tomcat JSP core functionality, and org.apache.commons.fileupload was used for uploading input files. The web server runs on a personal computer (1.8 GHz CPU, 512 MB working memory) with Linux as the operating system (SuSE 9.0, Kernel 2.4.20).

Use of the web interface

The web interface allows the user to easily search a list of sequences (provided in FASTA format) for the presence of potential signal peptides. There are two ways to submit this input list: either pasting the list into the query field or transferring it as a file upload. The user has the option of setting several parameters manually. First, a PWM is selected by taking the organism-specific background into account. For that purpose three matrices for the analysis of sequences from eukaryotes, Gram-positive and Gram-negative bacteria are offered. Second, the user can define the maximal length of the signal peptide. Biologically meaningful values for this parameter lie between 60 and 100 amino acid residues. The default parameter is a length of 70 amino acids. Third, the output format is selectable. Depending on the need for further processing of the resulting data, the user can choose between an HTML table, an easily parseable CSVs (comma separated values) file to port the data to Excel and related applications, and XML format. The output can be shown in the web browser or saved as a file on the local machine. Output parameters given are the overall estimation of whether the investigated amino acid sequence possesses a signal peptide (Y/N), the underlying score and the putative signal peptidase cleavage position.

RESULTS AND DISCUSSION

The prediction of signal peptides has become an important application of genomics and proteomics investigations. SignalP is currently the most efficient and widely used tool for this task. A comparison of most available software in this field underscored the unique performance of this program (7). However, non-commercial utilization of SignalP via the Internet is limited to 10 requests of up to 2000 sequences per day. Response to such requests takes several minutes. This means that SignalP is not suited to fast whole proteome analysis approaches. Finally, the program is not available as public domain software for integration into other software projects. Therefore, we decided to implement an alternative efficient prediction tool which meets the described criteria. The algorithm employed also represents an alternative approach to the neural network and Hidden Markov solutions implemented by SignalP. The fidelity of the employed method was significantly improved by the introduction of a frequency correction in order to adjust the amino acid bias as described by Schneider and Brown (9).

To check the accuracy of PrediSi, we performed a self-consistency test. For this purpose we constructed three test datasets containing proteins carrying signal peptides—for eukaryotes, Gram-negative and Gram-positive bacteria. The test datasets consist of all the amino acid sequences from a training dataset extracted from SwissProt and the same number of randomly chosen amino acid sequences without signal peptides from a corresponding control set. We compared the results obtained with the accuracy of SignalP (Table 1). Predictions were only considered as correct if both the existence and the cleavage position of the signal peptide were predicted correctly. The results of the analysis showed that PrediSi was slightly less accurate in the prediction of eukaryotic and Gram-negative signal sequences [85.49% PrediSi versus 90.66% SignalP-Neural Network (NN) and 88.24% SignalP–Hidden Markov Model (HMM); 91.12% versus 91.39% NN and 93.09% HMM, respectively] but slightly better at predicting Gram-positive signal peptides (88.14% versus 85.61% NN and 87.29% HMM) (Table 1). Interestingly, if we allowed a tolerance of two positions between the cleavage position, the accuracy of returning the correct cleavage position increased significantly. Probably some of these falsely predicted cleavage positions are due to database errors as mentioned before (10). PrediSi provides a normalized score on a scale between 0 and 1. A score greater than 0.5 means that the examined sequence very likely contains a signal peptide. The advantage of this user-friendly score is that it is comparable between different weight matrices.

The optimal PWM size differs between the three examined groups of organisms. The optimal size for the eukaryotic PWM

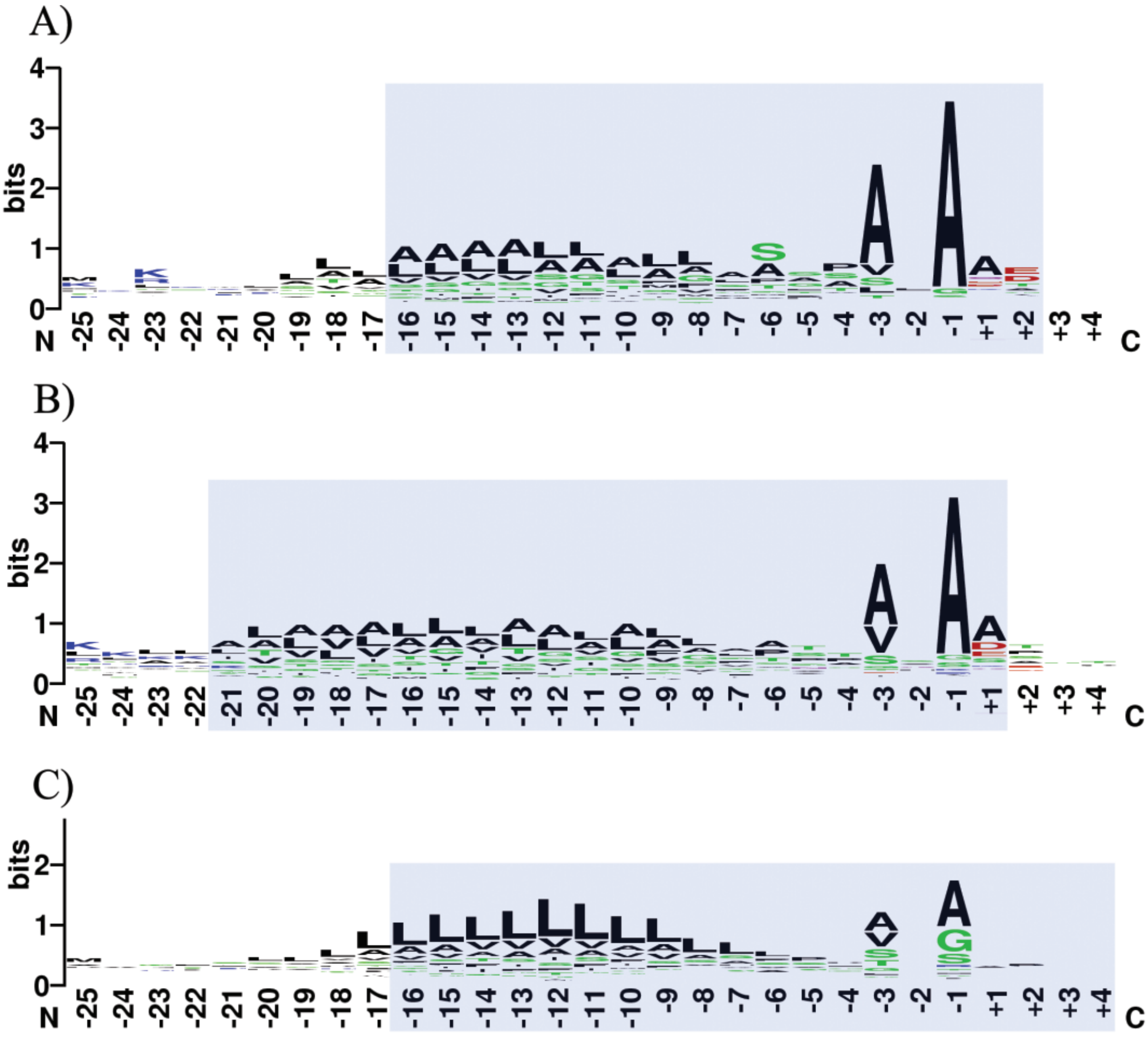
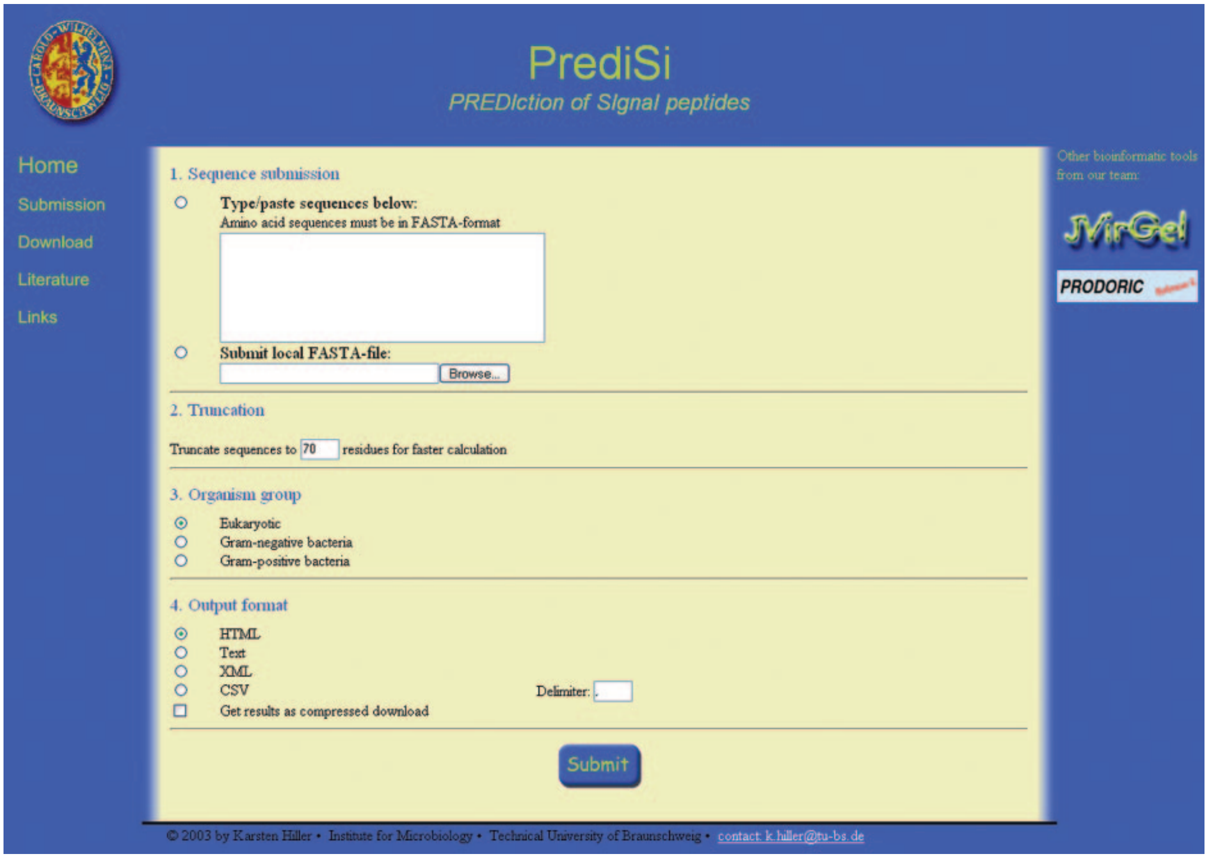


Figure 1. Sequence logos based on the aligned amino acid sequences of signal peptides. The signal peptide is cleaved off between position −1 and 0. (**A**) Gram-negative bacteria, (**B**) Gram-positive bacteria, (**C**) eukaryotes. Shaded area represents PWM region.

Table 1. Statistical examination of the accuracy of the different models promoted by SignalP and the accuracy of the new weight matrix approach

Dataset	Eukarya			Gram-positive			Gram-negative		
	Positive	Control	Overall	Positive	Control	Overall	Positive	Control	Overall
PrediSi	72.66	98.31	85.49	78.39	97.89	88.14	86.54	95.7	91.12
NN (SignalP)	82.11	99.21	90.66	77.97	93.25	85.61	86.54	96.24	91.39
HMM (SignalP)	78.73	97.74	88.24	75.42	99.16	87.29	87.07	99.1	93.09

Scores for the various predictions are given separately for Gram-positive bacteria, Gram-negative bacteria and eukaryotes. The values provided are the percentage of correctly identified signal peptides including the correct positions of their cleavage site. The positive dataset consists of proteins carrying signal peptides; the control consists of proteins without signal peptides. The overall score combines the obtained values for the positive and control datasets.



1. Sequence submission

○ **Type/paste sequences below:**
Amino acid sequences must be in FASTA-format

○ **Submit local FASTA-file:**

2. Truncation

Truncate sequences to 70 residues for faster calculation

3. Organism group

● Eukaryotic
○ Gram-negative bacteria
○ Gram-positive bacteria

4. Output format

● HTML
○ Text
○ XML
○ CSV Delimiter: .
□ Get results as compressed download

© 2003 by Karsten Hiller • Institute for Microbiology • Technical University of Braunschweig • contact: k.hiller@tu-bs.de

Figure 2. Screenshot of the PrediSi web interface.

is −16/+4 (with the cleavage position between positions −1 and +1), for Gram-negatives −16/+2 and for Gram-positives −21/+1. Figure 1 depicts sequence logos (11) of signal peptides for the three different groups. The estimated matrix size correlates well with the information content of the observed sequences. Agreeing with earlier analysis, signal peptides of Gram-positives are larger than those of other organisms (12). In summary, accuracy of prediction with PrediSi is similar to that with SignalP.

The use of a very fast algorithm for the prediction of the signal peptides enables our web interface to finish the necessary calculations nearly in real time. For example, the analysis of 20 000 eukaryotic sequences takes only about 10 s and is, therefore, limited only by the data transfer via the Internet. To our knowledge, this is the fastest public method available for predicting signal peptides. Using PrediSi it is not necessary to deliver the results by email or to install queues, because the results are directly presented in the web browser (Figure 2). Other methods such as Markovian models and neural networks need much more calculation time to perform such a task.

ACKNOWLEDGEMENTS

We would like to thank Dr Barbara Schulz for critical proofreading of the manuscript. This work was funded by the German Bundesministerium für Bildung und Forschung (BMBF) for the Bioinformatics Competence Center Intergenomics' (Grant No. 031U110A/031U210A).

REFERENCES

1. Zheng,N. and Gierasch,L.M. (1996) Signal sequences: the same yet different. *Cell*, **86**, 849–852.

2. Rapoport,T.A., Jungnickel,B. and Kutay,U. (1996) Protein transport across the eukaryotic endoplasmic reticulum and bacterial inner membranes. *Annu. Rev. Biochem.*, **65**, 271–303.
3. von Heijne,G. (1985) Signal sequences. The limits of variation. *J. Mol. Biol.*, **184**, 99–105.
4. Nielsen,H. and Krogh,A. (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 122–130.
5. 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.
6. Hiller,K., Schobert, M., Hundertmark, C., Jahn, D. and Münch, R. (2003) JVirGel: calculation of virtual two-dimensional protein gels. *Nucleic Acids Res.*, **31**, 3862–3865.
7. 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.
8. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
9. Schreiber,M. and Brown,C. (2002) Compensation for nucleotide bias in a genome by representation as a discrete channel with noise. *Bioinformatics*, **18**, 507–512.
10. Nielsen,H., Engelbrecht,J., von Heijne,G. and Brunak,S. (1996) Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site. *Proteins*, **24**, 165–177.
11. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
12. Tjalsma,H., Bolhuis,A., Jongbloed,J.D., Bron,S. and van Dijl,J.M. (2000) Signal peptide-dependent protein transport in *Bacillus subtilis*: a genome-based survey of the secretome. *Microbiol. Mol. Biol. Rev.*, **64**, 515–547.