Acta Biochimica Polonica

Vol. 55 No. 2/2008, 261–267

on-line at: www.actabp.pl

Regular paper

Prediction of signal peptides in protein sequences by neural networks

Dariusz Plewczynski[1]✉, Lukasz Slabinski[2], Krzysztof Ginalski[1]
and Leszek Rychlewski[2]

[1]Interdisciplinary Centre for Mathematical and Computational Modelling, Warsaw University, Warszawa, Poland; [2]BioInfoBank Institute, Poznań, Poland

Received: 25 February, 2008; revised: 09 April, 2008; accepted: 12 May, 2008
available on-line: 26 May, 2008

We present here a neural network-based method for detection of signal peptides (abbreviation used: SP) in proteins. The method is trained on sequences of known signal peptides extracted from the Swiss-Prot protein database and is able to work separately on prokaryotic and eukaryotic proteins. A query protein is dissected into overlapping short sequence fragments, and then each fragment is analyzed with respect to the probability of it being a signal peptide and containing a cleavage site. While the accuracy of the method is comparable to that of other existing prediction tools, it provides a significantly higher speed and portability. The accuracy of cleavage site prediction reaches 73% on heterogeneous source data that contains both prokaryotic and eukaryotic sequences while the accuracy of discrimination between signal peptides and non-signal peptides is above 93% for any source dataset. As a consequence, the method can be easily applied to genome-wide datasets. The software can be downloaded freely from http://rpsp.bioinfo.pl/RPSP.tar.gz.

Keywords: neural networks, protein database, protein sequences, signal peptides

INTRODUCTION

The destination of newly synthesized proteins in a cell is often controlled by their short fragments called signal peptides (SP) (Gierasch, 1989). Most signal peptides comprise N-terminal amino acids that are cleaved off while the protein is being translocated through a membrane. In that way signal peptides modulate various aspects of cellular life, such as the entry of proteins to the secretory pathway, both in eukaryotic and prokaryotic cells (Bruch *et al.*, 1989; Cornell *et al.*, 1989; Gierasch, 1989; Rapoport, 1992). The simplest computational approach for the identification of signal peptides is based on the application of regular expression search, where regular expressions are constructed from experimentally verified signal peptides in proteins (Puntervoll *et al.*, 2003). In order to improve the efficiency of prediction by regular expression search and lower the number of false positives, context-based rules and various logical filters may be applied (Puntervoll *et al.*, 2003). Detection of signal peptides can be also carried out using a weight matrix approach (von Heijne, 1986a, 1986b). This approach is quite efficient in recognition of cleavage sites between a signal sequence and the mature exported protein because many cleavage sites are strongly characterized by a set of simple rules which are quantified by the weight matrix methods (von Heijne 1986a; 1986b; Menne *et al.*, 2000). For instance, the residues at positions –3 and –1 relative to the cleavage site are usually small and neutral. The regular expression search and weight matrix algorithms are now replaced by more sophisticated methods that include various types of machine learning methods such as neural networks (Nielsen *et al.*, 1997), support vector machines (Vert, 2002), hidden Markov models (Nielsen *et al.*, 1999) and many others (Ladunga *et al.*, 1991; Talmud *et al.*, 1996; Nielsen

✉Corresponding author: Dariusz Plewczynski, Interdisciplinary Centre for Mathematical and Computational Modelling, Warsaw University, A. Pawińskiego 5a, 02-106 Warszawa, Poland; tel.: (48) 22 554 0839; fax: (48) 22 554 0801; e-mail: D.Plewczynski@icm.edu.pl

Abbreviations: RPSP, rapid prediction of signal peptide; SP, signal peptide.

et al., 1997a; 1997b; Nielsen & Krogh, 1998; Nielsen *et al.*, 1999; Bendtsen *et al.*, 2004a; 2005b; Menne *et al.*, 2000; Chou 2001; Lao *et al.*, 2002a; 2002b; Vert, 2002; Juncker *et al.*, 2003; Hiller *et al.*, 2004; Kall *et al.*, 2004; Zhang & Henzel, 2004; Liu *et al.*, 2005; Sidhu & Yang, 2006). However, most of these methods classify proteins as secretory or non-secretory but do not provide cleavage site assignment. In addition, high throughput genome sequencing has problems with assigning the 5′-end of genes, so many proteins lack the correct N-terminal end (Reinhardt & Hubbard, 1998). This obviously leads to incorrect prediction of signal peptides.

Currently, two of the most commonly used methods for detection of classical signal peptides, SignalP (Bendtsen *et al.*, 2004a; 2004b) and SPEPlip (Fariselli *et al.*, 2003), also predict the actual cleavage site. SPEPlip applies a neural network trained on a set of experimentally verified signal peptides from eukaryotes and prokaryotes, while SignalP uses both neural network and Hidden Markov Model and may work on various types of sequences including those from Gram-positive and Gram-negative bacteria and eukaryotes. In this paper we describe a similar but significantly faster method for identification of signal peptides in proteins based on a neural network trained on the most recent version of Swiss-Prot database. RPSP (rapid prediction of signal peptides) is publicly available as a standalone version together with its source code. The new method focuses on the classical types of signal peptides neglecting the non-classically secreted proteins (Bendtsen *et al.*, 2004a; 2005a), and can be used in large-scale predictions of signal peptides even for heterogeneous sets of sequences. Specifically, three types of prediction can be performed: for prokaryotic sequences, eukaryotic sequences and without specifying the organism type. Similar sequence based approaches were extensively tested by our collaborators in the case of active site prediction and structural motifs for herpes ICP4 protein (Ostrowski *et al.*, 2006; 2007; Plewczynski *et al.*, 2006; 2008; Koczyk *et al.*, 2007; Wyrwicz *et al.*, 2007; 2008; Wyrwicz & Rychlewski, 2007; 2008). Such sequence based methods are able to provide predictions of important short sequence fragments that are phosphorylated, bind ligands, interact with other proteins, RNA molecules, as well as critical residues that stabilize ligands (Plewczynski *et al.*, 2004; 2005a; 2005b; 2005c; 2006; 2008; von Grotthuss *et al.*, 2006).

METHOD

The training and testing datasets were generated using annotated protein sequence information acquired from the Swiss-Prot database (release: 49.4). The initial set was constructed by extracting all Swiss-Prot entries with a keyword: ‘SIGNAL’ in the FT line (20863 entries). All uncertain entries (potential, probable, by similarity, or possessing more than one cleavage site in FT line were removed) (4566 entries left), as well as all archaeal and viral proteins (by accepting entries only with ‘Eukaryota’/’Bacteria’ in the OC line) (4296 entries left). The resulting sequences were split into two sets: eukaryotic (‘Eukaryota’ in OC line) (3331 entries) and prokaryotic (‘Bacteria’ in OC line) (965 entries). For eukaryotes all organelle proteins (entries with the line OG were removed) (3327 entries left), sequences shorter than 15 and longer than 45 amino acids (3294 entries left) and those with residues other than: A, C, G, L, P, Q, S, T at ‘–1’ position (3167 entries left). For prokaryotes we removed all lipoproteins (cross-reference to the PROSITE, keyword: “PROKAR_LIPOPROTEIN”) (894 entries left), sequences shorter than 15 and longer than 50 amino acids (875 entries left) and those with residues other than: A, G, S, T at ‘–1’ position (841 entries left). The datasets of negatives were prepared by extracting N-terminal parts (first 70-residue sequence fragments) of eukaryotic cytoplasmic (2215 entries) and nuclear (3616 entries) proteins for eukaryotes, and N-terminal parts of bacterial cytoplasmic proteins for prokaryotes (6225 entries), removing any potential, probable, fragment and shorter than 70 amino acids entries. All sets were reduced at 60% sequence identity (calculated for whole protein sequences, not for short fragments for which sequence similarity can be misleading) using the CD-HIT clustering tool (Li *et al.*, 2001) (1784 entries left for eukaryotes, 646 for prokaryotes, 987 cytoplasmic and 2265 nuclear for negative eukaryotes and 2040 for negative prokaryotes). This cut-off was optimized as it provides both the best results and moderate memory requirements for training of the neural network. We also checked datasets clustered at 30% and 100% sequence identity, yet the results were worse. The negative datasets were further reduced approximately to the sizes of the positive datasets, to avoid bias during training and testing of the neural network. Finally, the resulting datasets were divided into six approximately equal-size parts (five as training data, one as test data). The neural networks were trained using a standard cross-validation learning procedure on the training sets (separate for eukaryotes, prokaryotes, and mixed) (Baldi & Brunak, 2001) while test sets were used only for final benchmarks. Details of the prepared datasets are shown in Table 1. All training datasets used to build the RPSP method are available on the http://rpsp.bioinfo.pl web pages.

As the cleavage site position and the amino-acid composition of the signal peptide are known to be highly correlated (Nielsen *et al.*, 1997a; 1997b), the local sequence information is sufficient as an

Table 1. RPSP training datasets

	Signal peptides		Cytoplasmic proteins		Nuclear proteins	
Dataset	Total	Reduced	Total	Reduced	Total	Reduced
Eukaryotes	3167	1784	2215	987	3616	2265
Prokaryotes	841	646	6225	2040	—	—

input to the neural network. RPSP uses two independent neural networks with complex feed-forward, multi-layer architecture and a back-propagation learning algorithm (Fig. 1). The first network is designed to identify if a given residue belongs to a signal peptide or not. Here, we use a symmetric sequence window with 27 amino acids for eukaryotic and mixed (prokaryotic&eukaryotic) sequences and 19 amino acids for prokaryotic sequences as an input for the neural network. We also neglect differences between Gram-negative and Gram-positive bacteria (Nielsen *et al.*, 1997a). The output layer is a single neuron providing the S-score of a prediction. High S-score corresponds to higher probability that the given amino acid belongs to a signal peptide, and low score indicates that the amino acid is rather part of a mature protein. The second neural network identifies the cleavage site (first residue in the mature protein, i.e. position +1). The input for the neural network is an asymmetric sliding window with 24 residues for prokaryotic/eukaryotic and 25 for prokaryotic&eukaryotic sequences. The output layer is also a single neuron that provides the C-score of a prediction. This score describes the cleavage site likelihood for each position in the query sequence. The C-score is higher at the cleavage site than for other parts of the protein sequence. The discrimination between signal peptide and non-signal peptide and cleavage site prediction are guided by Y-score that combines both the S-score and the C-score (similarly to SignalP (Bendtsen *et al.*, 2004a) and SPEPlip (Fariselli *et al.*, 2003)). For the clarity we use the same name convention of various scores as in the SignalP and SPEPlip prediction algorithms. The Y-score is equal to: $Y_i = \sqrt{C_i * \Delta_d S_i}$, where $\Delta_d S_i$ is the difference between the mean S-score for all d amino acids before and d amino acids after position i. The d value of 17 was taken from our benchmarking results. The Y-score provides a better cleavage site prediction than the raw C-score alone, because usually a number of high C-scores can be assigned to amino acids in the query sequence, whereas only one 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 a significant C-score is found. Finally, the D-score is also computed that is an arithmetic mean of Y-score for position i and the mean value of S-score for all amino acids. This score was shown (Bendtsen *et al.*, 2004a; 2004b) to be superior in discrimination between secretory and non-secretory proteins in comparison with the S-mean score used in previous approaches. A protein is expected to contain a signal peptide in the considered position i if the Y-score for this position is larger than 0.35 and D-score is larger than 0.43.

PERFORMANCE

Our main goal was to develop a fast method for signal peptide detection that could be applied for large scale annotations of heterogeneous sets of sequences and did not necessarily require specifying their origin. The performance analysis of three sets of neural networks trained on prokaryotic, eukaryotic and prokaryotic&eukaryotic sequences was conducted on an independent data set that was not used during the learning procedure. The performance of each NN classification model is described by three measures of accuracy: classification error E, recall R, and precision P:

$$E = \frac{fp + fn}{tp + fp + tn + fn} * 100\% ,$$

$$R = \frac{tp}{tp + fn} * 100\% ,$$

$$P = \frac{tp}{tp + fp} * 100\% , \qquad \text{(Eqn. 1)}$$

where tp is the number of true positives, fp is the number of false positives, tn is the number of true negatives and fn is the number of false negatives. The classification error E provides an overall error measure, whereas recall R measures the percentage of correct predictions (the probability of correct prediction). Precision P gives the percentage of observed positives that are correctly predicted (the measure of reliability of the positive prediction). These measures of accuracy are calculated using a precise but computationally intensive leave-one-out procedure. The leave-one-out test removes from the training dataset one sample, constructs the model on the basis of the remaining training dataset and then tests the prediction of the model on the removed sample. The

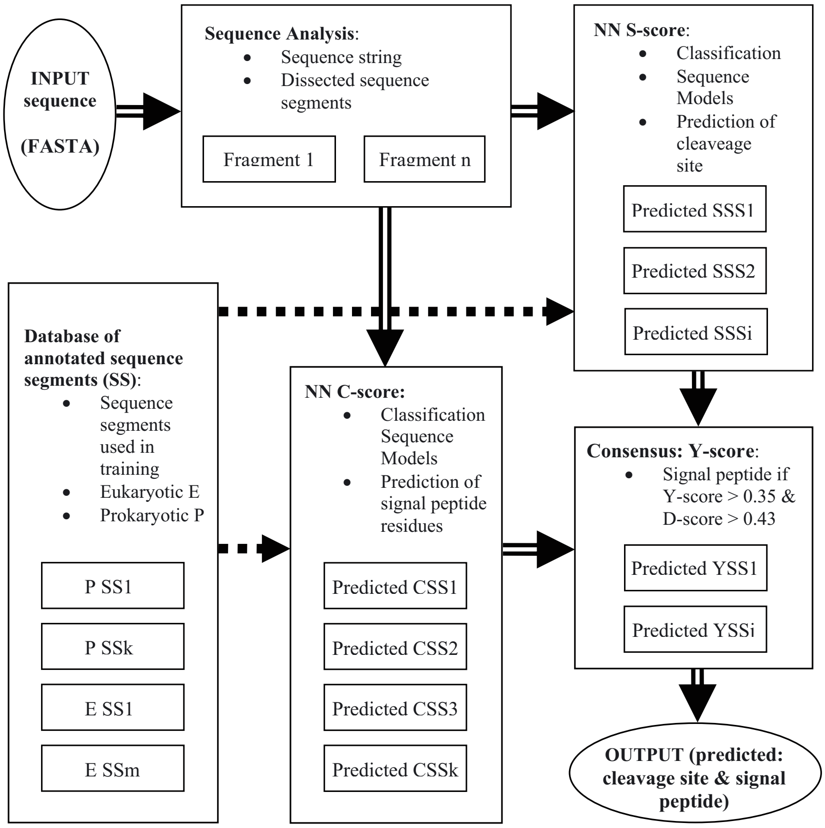

Figure 1. RPSP data flowchart.
Double arrows denote information flow during prediction, dotted lines show connection to the database of annotated sequence segments from Swiss-Prot database, known to be signal peptides from experiments.

resulting error estimators are averaged over all such models (for all positive and all negative instances).

Detailed benchmark results are shown in Table 2. One of the main assets of the program is its high efficiency of signal peptide prediction even without specifying the organism of a protein. The precision of the method operating without distinguishing between prokaryotic and eukaryotic proteins is not significantly lower (only 4% difference) than using separate neural networks trained either on eukaryotic or prokaryotic sequences. The accuracy of cleavage site prediction reaches 73% on heterogeneous source data that contains both prokaryotic and eukaryotic sequences, while the accuracy of discrimination between signal peptides and non-signal peptides is above 93% for any source dataset. Thus, the results are comparable to those that can be obtained with other prediction tools such as SignalP 3.0 (Bendtsen *et al.,* 2004) or SPEPlip (Fariselli *et al.,* 2003).

Another crucial advantage of RPSP is that the method is very fast. For instance, analysis of 959 proteins (our full benchmark set) takes about 2 seconds on a Linux machine with 2 GHz CPU and 512 MB RAM. Examples of genome-wide signal peptide predictions, with the time of execution and SignalPv3.0 and SPEPlip results for comparison, are shown in Table 3. Most importantly the RPSP server is designed for high-throughput analyses and, in addition, the method is also available as a standalone program. This is a significant advantage over other tools that provide only standard web server interfaces. Those services usually cannot accept an input of more than around a thousand proteins having certain limits of residues per sequence, number of residues in total and number of jobs accepted from a single IP internet address. Altogether, this makes RPSP the method of choice in high throughput studies, such as massive analyses of whole proteomes in

Table 2. Results of independent benchmarks

	Discrimination (SP/non-SP)				Cleavage site
Dataset	Sensitivity	Specificity	Accuracy	CC*	Accuracy
Eukaryotes	90%	98%	95%	0.91	77%
Prokaryotes	91%	98%	95%	0.91	78%
Euk & Pro	86%	98%	93%	0.87	73%

*Matthews correlation coefficient

Table 3. Results of genome analyses.

RPSP was run on Linux machine with 2 GHz CPU and 512 MB RAM. The running times for SignalP and SPEPlip were taken from the web servers as their local versions are not readily available

Genome	Sequences	SignalP 3.0 Predicted SP	SPEPlip Predicted SP	RPSP Predicted SP	RPSP local version time [s]	RPSP Web server time [s]	SignalP 3.0 Web server time [s]	SPEPlip Web server time [s]
Plasmodium falciparum	5365	474	501	519	7	70	263	203
Chlamydophila pneumoniae	1113	164	134	112	3	30	127	97

the context of function prediction or detailed characterization of proteins.

CONCLUSIONS

In this paper we describe a new, fast method for identification of signal peptides in proteins. The method uses two neural networks trained on experimental data from the Swiss-Prot database. We would like to stress that our training dataset, in comparison to previous approaches, is based on the most recent version of the Swiss-Prot database, so present the significant update of the earlier signal peptide prediction methods. Experimental laboratories provide each year a number of new confirmed signal peptides, therefore updating of the training sets is very important.

The main advantage of RPSP is its ease of use, i.e. its local version with source code and precompiled binary is available. The method is very fast. We were able to optimize the source code and the binary during the compilation. The availability of a 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 the web server technical design. The existing signal peptide prediction servers cannot accept input of more than around a thousand proteins, all have to contain less than a certain limit of residues per sequence (few thousand) and there is also the limit for the number of residues in total. Additionally, the existing servers have also the limit of a few thousand lines in the input file, and the number of jobs accepted from a single IP internet address is limited. The splitting of whole proteome sequences into smaller bunches by hand is time consuming and cannot be done in a high-throughput manner. In addition, the local version of SignalP does not contain the source code (you cannot modify it in accordance with your specific needs), and cannot be used in commercial applications. We were also not able to get its binary compiled version from the authors of SignalP, so we could not perform a detailed comparison of both tools on independent benchmarks. Apparently the availability of the local version of SignalP is limited. There is no local version, either binary or source code, of SPEPlip predictor, you are allowed to use only the web service.

Summarizing, we would like to highlight three features of our machine learning software, namely: i) new training set based on updated version of Swiss-Prot database; ii) speed of the software and the availability of the local version with source code and precompiled binary for LINUX; and iii) good efficiency even without specifying the origin organism class. These points make RPSP, a tool for rapid prediction of signal peptides, the method of choice in high throughput environments, such as massive analysis of whole proteomes in the context of function prediction or detailed characterization of proteins. The prediction of signal peptides in both example proteomes are performed in real time. The RPSP source code in C programming language together with the LINUX precompiled binary can be downloaded freely 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.

Acknowledgements

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

REFERENCES

Baldi P, Brunak S (2001) *Bioinformatics: The Machine Learning Approach.* 2nd edn., MIT Press, Cambridge, MA.

Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S (2004a) Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng Des Sel* **17**: 349–356.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004b) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**: 783–795.

Bendtsen JD, Kiemer L, Fausboll A, Brunak S (2005a) Non-classical protein secretion in bacteria. *BMC Microbiol* **5**: 58.

Bendtsen JD, Nielsen H, Widdick D, Palmer T, Brunak S (2005b) Prediction of twin-arginine signal peptides. *BMC Bioinformatics* **6**: 167.

Bruch MD, McKnight CJ, Gierasch LM (1989) Helix formation and stability in a signal sequence. *Biochemistry* **28**: 8554–8561.

Chou KC (2001) Prediction of signal peptides using scaled window. *Peptides* **22**: 1973–1999.

Cornell DG, Dluhy RA, Briggs MS, McKnight CJ, Gierasch LM (1989) Conformations and orientations of a signal peptide interacting with phospholipid monolayers. *Biochemistry* **28**: 2789–2797.

Fariselli P, Finocchiaro G, Casadio R (2003) SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. *Bioinformatics* **19**: 2498–2499.

Gierasch LM (1989) Signal sequences. *Biochemistry* **28**: 923–930.

Hiller K, Grote A, Scheer M, Munch R, Jahn D (2004) PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res* **32** (Web Server issue): W375–W379.

Juncker AS, Willenbrock H, Von Heijne G, Brunak S, Nielsen H, Krogh A (2003) Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci* **12**: 1652–1662.

Kall L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* **338**: 1027–1036.

Koczyk G, Wyrwicz LS, Rychlewski L (2007) LigProf: a simple tool for in silico prediction of ligand-binding sites. *J Mol Model* **13**: 445–455.

Ladunga I, Czako F, Csabai I, Geszti T (1991) Improving signal peptide prediction accuracy by simulated neural network. *Comput Appl Biosci* **7**: 485–487.

Lao DM, Arai M, Ikeda M, Shimizu T (2002a) The presence of signal peptide significantly affects transmembrane topology prediction. *Bioinformatics* **18**: 1562–1566.

Lao DM, Okuno T, Shimizu T (2002b) Evaluating transmembrane topology prediction methods for the effect of signal peptide in topology prediction. *In Silico Biol* **2**: 485–494.

Li W, Jaroszewski L, Godzik A (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**: 282–283.

Liu L, Li J, Tian X, Ren D, Lin J (2005) Information theory in prediction of cleavage sites of signal peptides. *Protein Pept Lett* **12**: 339–342.

Menne KM, Hermjakob H, Apweiler R (2000) A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics* **16**: 741–742.

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

Nielsen H, Engelbrecht J, Brunak S, von Heijne G (1997b) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int J Neural Syst* **8**: 581–599.

Nielsen H, Brunak S, von Heijne G (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng* **12**: 3–9.

Nielsen H, 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.

Ostrowski J, Mikula M, Karczmarski J, Rubel T, Wyrwicz LS, Bragoszewski P, Gaj P, Dadlez M, Butruk E, Regula J (2007) Molecular defense mechanisms of Barrett's metaplasia estimated by an integrative genomics. *J Mol Med* **85**: 733–743.

Ostrowski J, Rubel T, Wyrwicz LS, Mikula M, Bielasik A, Butruk E, Regula J (2006) Three clinical variants of gastroesophageal reflux disease form two distinct gene expression signatures. *J Mol Med* **84**: 872–882.

Plewczynski D, Pas J, Von Grotthuss M, Rychlewski L (2004) Comparison of proteins based on segments structural similarity. *Acta Biochim Pol* **51**: 161–172.

Plewczynski D, Jaroszewski L, Godzik A, Kloczkowski A, Rychlewski L (2005a) Molecular modeling of phosphorylation sites in proteins using a database of local structure segments. *J Mol Model* **11**: 431–438.

Plewczynski D, Tkacz A, Godzik A, Rychlewski L (2005b) A support vector machine approach to the identification of phosphorylation sites. *Cell Mol Biol Lett* **10**: 73–89.

Plewczynski D, Tkacz A, Wyrwicz LS, Rychlewski L (2005c) AutoMotif server: prediction of single residue post-translational modifications in proteins. *Bioinformatics* **21**: 2525–2527.

Plewczynski D, Tkacz A, Wyrwicz LS, Godzik A, Kloczkowski A, Rychlewski L (2006) Support-vector-machine classification of linear functional motifs in proteins. *J Mol Model* **12**: 453–461.

Plewczynski D, Tkacz A, Wyrwicz LS, Rychlewski L, Ginalski K (2008) AutoMotif Server for prediction of phosphorylation sites in proteins using support vector machine: 2007 update. *J Mol Model* **14**: 69–76.

Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A, Ferre F, Maselli V, Via A, Cesareni G, Diella F, Superti-Furga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Kuster B, Helmer-Citterich M, Hunter WN, Aasland R, Gibson TJ (2003) ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res* **31**: 3625–3630.

Rapoport TA (1992) Transport of proteins across the endoplasmic reticulum membrane. *Science* **258**: 931–936.

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

Sidhu A, Yang ZR (2006) Prediction of signal peptides using bio-basis function neural networks and decision trees. *Appl Bioinformatics* **5**: 13–19.

Talmud P, Lins L, Brasseur R (1996) 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* **9**: 317–321.

Vert JP (2002) Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings. *Pac Symp Biocomput*: 649–660.

von Grotthuss M, Plewczynski D, Ginalski K, Rychlewski L, Shakhnovich EI (2006) PDB-UF: database of predicted enzymatic functions for unannotated protein structures from structural genomics. *BMC Bioinformatics* **7**: 53.

von Heijne G (1986a) Net N-C charge imbalance may be important for signal sequence function in bacteria. *J Mol Biol* **192**: 287–290.

von Heijne G (1986b) A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res* **14**: 4683–4690.

Wyrwicz LS, Rychlewski L (2007) Fold recognition insights into function of herpes ICP4 protein. *Acta Biochim Polon* **54**: 551–559.

Wyrwicz LS, Gaj P, Hoffmann M, Rychlewski L, Ostrowski J (2007) A common cis-element in promoters of protein synthesis and cell cycle genes. *Acta Biochim Polon* **54**: 89–98.

Wyrwicz LS, Rychlewski L (2008) Cytomegalovirus immediate early gene UL37 encodes a novel MHC-like protein. *Acta Biochim Polon* **55**: 67–74.

Wyrwicz LS, Koczyk G, Rychlewski L (2008) Homologues of HSV-1 nuclear egress factor UL34 are potential phosphoinositide-binding proteins. *Acta Biochim Polon* **55**: 207–213.

Zhang Z, Henzel WJ (2004) Signal peptide prediction based on analysis of experimentally verified cleavage sites. *Protein Sci* **13**: 2819–2824.