BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 10 2005, pages 2525–2527
doi:10.1093/bioinformatics/bti333

Sequence analysis

AutoMotif server: prediction of single residue post-translational modifications in proteins

Dariusz Plewczynski[1,2,*], Adrian Tkacz[1], Lucjan Stanisław Wyrwicz[3] and Leszek Rychlewski[1]

[1]BioInfoBank Institute, Limanowskiego 24A/16, 60-744 Poznan, Poland, [2]Interdisciplinary Centre for Mathematical and Computational Modeling, University of Warsaw, Warsaw, Poland and [3]Bioinformatics Unit, Department of Physics, Adam Mickiewicz University, Poznan, Poland

Received on November 24, 2004; revised on January 12, 2005; accepted on February 16, 2005
Advance Access publication February 22, 2005

ABSTRACT

Summary: The AutoMotif Server allows for identification of post-translational modification (PTM) sites in proteins based only on local sequence information. The local sequence preferences of short segments around PTM residues are described here as linear functional motifs (LFMs). Sequence models for all types of PTMs are trained by support vector machine on short-sequence fragments of proteins in the current release of Swiss-Prot database (phosphorylation by various protein kinases, sulfation, acetylation, methylation, amidation, etc.). The accuracy of the identification is estimated using the standard leave-one-out procedure. The sensitivities for all types of short LFMs are in the range of 70%.

Availability: The AutoMotif Server is available free for academic use at http://automotif.bioinfo.pl/

Contact: darman@bioinfo.pl

INTRODUCTION

The post-translational modifications (PTMs) modulate all aspects of cellular life, such as transient modifications in signal transduction pathways (Pawson, 2004). These modifications predominantly occur in protein linear functional motifs (LFMs), but only a small fraction of modified sites has been identified. By the prediction of PTM sites from a protein sequence, we can obtain valuable information that can form the basis for further research. We present here our attempt to identify PTM sites by support vector machine (SVM) trained on proteins of the Swiss-Prot database (Bairoch and Apweiler, 1999).

A simple approach for the identification of PTM sites is based on the application of regular expression search. Regular expressions are constructed from experimentally verified functional sites in proteins. In order to improve the efficiency of prediction by regular expression search and to lower the number of false-positives, context-based rules and logical filters are applied in the ELM resources (Puntervoll *et al*., 2003). The Sulfinator (Monigatti *et al*., 2002) uses hidden Markov model (HMM) to recognize sulfated residues and it is built on the basis of multiple sequence alignments of 25-amino acid long segments. The NetPhos server utilizes neural networks trained on PhosphoBase database (Kreegipuu *et al*., 1999) in order to characterize a 9-amino acid neighborhood of the serine, threonine and tyrosine phosphorylation sites in eukaryotic proteins (Blom *et al*., 1999). The PredPhospho server predicts phosphorylation sites and the type of protein kinase acting at each site using SVMs (Kim *et al*., 2004). Another web program Scansite uses sequence profiles derived from experimental data for identification of PTM sites. This motif-based scanning approach is applied for genome-wide prediction of signaling pathways (Yaffe *et al*., 2001). The eMOTIF (Huang and Brutlag, 2001) discovers conserved sequence motifs in families of proteins derived from the multiple sequence alignments with a wide range of specificities and sensitivities. The PROSITE database (Falquet *et al*., 2002) allows the inferring of a function and the classification of a protein using a set of local sequence similarity tools.

METHODS

We trained SVM for each type of PTM separately (http://automotif.bioinfo.pl//about.htm) on proteins of the Swiss-Prot database (version 42) (Bairoch and Apweiler, 1999). We created the dataset of experimentally verified segments including the 9-amino acid long sequence fragments (positive instances). We built the dataset of negative cases (negative instances) in order to calculate the probability of finding an amino acid at a certain position in the LFM. The negative instances were chosen randomly from those that do not include experimentally verified PTM of any type. These two datasets (positive and negative instances) were projected as sets of points in a multidimensional space (http://automotif.bioinfo.pl/embedding.htm). The SVM (Yu-Dong *et al*., 2002; Vapnik, 1998; Joachims, 1999) constructed the separation border between the sets of positives and negatives and used it for further predictions. The prediction score measures the distance between a point representing short-sequence segment and this separation plane. The detailed description of the method with a list of references is presented in (Plewczynski *et al*., to be published; http://automotif.bioinfo.pl/help_index.php). The general model of the information flow and components of the AutoMotif server (AMS) are presented in Figure 1.

By default the AMS identifies PTM sites of 12 types in a query protein. Users can limit the search by choosing the particular type of PTM from the drop-down list on the server's main page (like a phosphorylation in general or by specific protein kinase). Users can also submit their own list of positives to create a new SVM model and use it to scan protein sequences. The output format of the service is a HTML page with a list of potential PTM sites with similarity score, residue name and residue position within the sequence of a query protein. The higher the output score the higher is the confidence of

*To whom correspondence should be addressed.

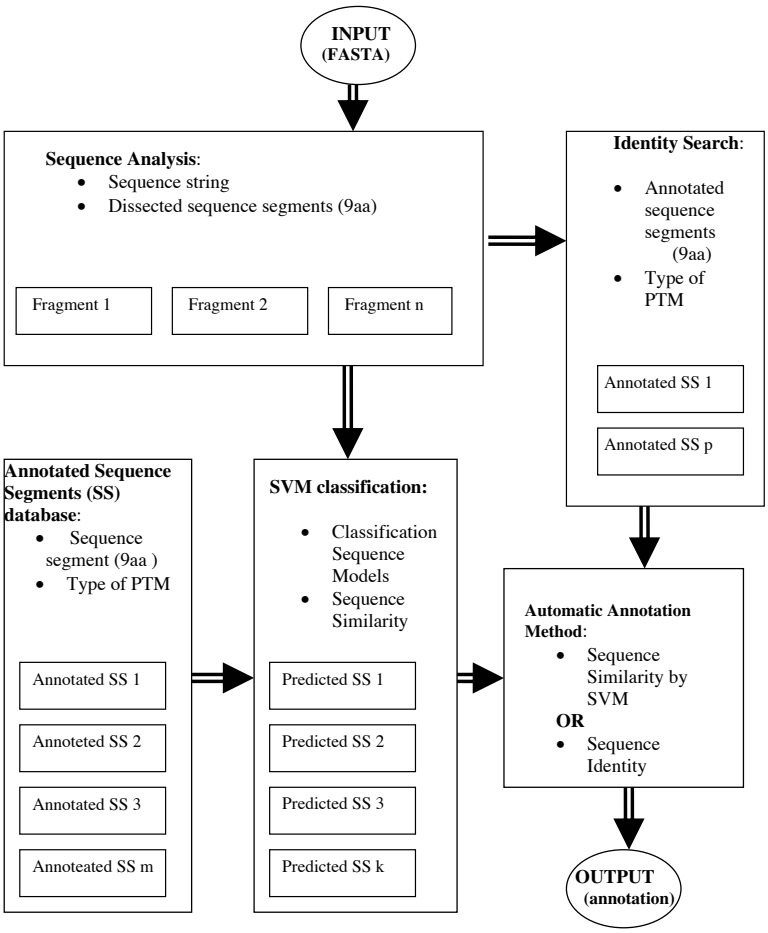

Fig. 1. The AutoMotif server data flowchart for prediction of PTM sites in proteins. The general model of the information flow and components of the automatic annotation service for prediction of post-translational single residue modification sites in proteins. The identification of PTM sites in a query protein is as follows. The query protein is divided into a set of overlapping sequence segments. Two search procedures are then used to annotate these segments: identity scan or SVM classification. The first one identifies the segments that are identical to one of positives from the database. The second method uses SVMs to classify the segments of the query protein into two groups: potential PTM sites and negatives. It is based on 10 classification models (http://automotif.bioinfo.pl/embedding.htm) for each type of PTM. Each classification model uses different representation of short-sequence segments. The list of PTM sites predicted by at least one classification model in the query protein is then returned to a user.

the predictions, i.e. sequence segments around predicted PTM sites are more similar to stored LFMs.

RESULTS

The performance of SVM models for each type of LFM is described by the recall R and the precision P. The R value measures the percentage of correct predictions, whereas P gives the percentage of observed positives that are correctly predicted. These measures of accuracy are calculated separately for each type of PTM using the leave-one-out procedure. The results are presented in http://automotif.bioinfo.pl/prediction.htm for all used projection methods and all types of LFMs considered. The typical R value is $\sim$30%, and P is $>$70% for majority of PTMs.

As an example of application of AMS we tested the modifications in 14-3-3 protein family. The sequences were obtained from Pfam resources (16 sequences from HMM seed; PF00244). The results are presented online as a biological tutorial at the web pages of the server, http://automotif.bioinfo.pl/tutorial1.html

ACKNOWLEDGEMENTS

This work was supported by EC and MNiI within the ELM (QLRT-CT2000-00127) 5FP project, BioSapiens (LHSG-CT-2003-503265) and GeneFun (LSHG-CT-2004-503567) 6FP projects. LSW is a fellow of the foundation for Polish Science.

REFERENCES

Bairoch,A. and Apweiler,R. (1999) The Swiss-Prot protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.*, **27**, 49–54.

Blom,N. *et al*. (1999) Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.*, **294**, 1351–1362.

Falquet,L. *et al*. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.

Huang,J.Y. and Brutlag,D.L. (2001) The eMOTIF database. *Nucleic Acids Res.*, **29**, 202–204.

Joachims,T. (1999) Making large scale SVM learning practical. In Scholkopf,B., Burges,C. and Smola,A. (eds), *Advances in Kernel Methods—Support Vector Learning*. MIT Press, Cambridge.

Kim,J.H. *et al*. (2004) Prediction of phosphorylation sites using SVMs. *Bioinformatics*, **20**, 3179–3184.

Kreegipuu,A. *et al*. (1999) PhosphoBase, a database of phosphorylation sites: release 2.0. *Nucleic Acids Res*., **27**, 237–239.

Monigatti,F. *et al*. (2002) The sulfinator: predicting tyrosine sulfation sites in protein sequences. *Bioinformatics*, **18**, 769–770.

Pawson,T. (2004) Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems. *Cell*, **116**, 191–203.

Plewczynski,D. *et al*. (2005) Support vector machine approach to phosphorylation sites prediction. *Cell. Mol. Biol. Lett*., **10**, 73–89.

Puntervoll,P. *et al*. (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res*., **31**, 3625–3630.

Vapnik,V.N. (1998) *Statistical Learning Theory*. Wiley, New York.

Yaffe,M.B. *et al*. (2001) A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nat. Biotechnol*., **19**, 348–353.

Yu-Dong,C. *et al*. (2002) Support vector machines for predicting the specificity of GalNAc-transferase. *Peptides*, **23**, 205–208.