AutoMotif Server: a computational protocol for identification of post-translational modifications in protein sequences

Dariusz Plewczynski (D.Plewczynski@icm.edu.pl)

Interdisciplinary Centre for Mathematical and Computational Modeling, University of Warsaw, Pawinskiego 5a Street, 02-106 Warsaw, Poland, Tel: (+48 22) 554-08-39, Fax: (+48 22)554-40-801

Adrian Tkacz

BioInfoBank Institute

Method Article

Keywords: post-translational modifications, phosphorylation, kinase substrate prediction, protein kinases, acetylation, sulfation, amidation, hydroxylation, methylation, pyrrolidone, gamma-carboxyglutamic modification, sequence similarity, database of functional sequence segments, Swiss-Prot database, support vector machine, machine learning

Posted Date: March 23rd, 2007

DOI: https://doi.org/10.1038/nprot.2007.183

Abstract

Introduction

The rapid increase in genomic information requires new automatic techniques to investigate protein functions. The function of proteins is partially determined by short sequence segments. For example the phosphorylation by protein kinases is an important mechanism for controlling intracellular processes. Many kinases are known, but the identification of their potential biological targets is still ongoing research. High substrate specificity of protein kinases ensures correct transmission of signals in cells. The specificity is largely determined by the primary sequence of the target site, but we lack general, efficient and error prune tools for identifying these sites. Most methods designed to predict functional motifs process local sequence information around post-translational modification sites. We present here an advanced computational protocol for rapid identification of post-translational modifications \(PTM) in proteins on the whole genome scale. The AutoMotif Server \(AMS) identifies various types of post-translational modifications in protein sequences. A query protein sequence is dissected into overlapping short segments. Each segment is projected into an abstract space of sequence fragments by 10 different representations. Those projections are compared with the database of representations of known and confirmed by experiments post-translational modification sites using the support vector machine \(SVM) approach [1, 2]. The supervised machine learning approach is able to predict the most of post-translational modification sites in proteins. It is based on the classification of the biological functional information acquired from the Swiss-Prot database version 4.2. The classification models are then used to predict new modification sites in proteins. Users can access a list of sites in proteins annotated as being able to undergo certain post-translational modification in Swiss-Prot database and add new annotated sequence segments from proteins \(positive instances). The AMS server was demonstrated [3, 4] to gain high accuracy in distinguishing short sequence fragments that are post-translational modified from those that are not. The efficiency of the classification for each type of modifications and the prediction power of several versions of the method is estimated using the standardized leave-one-out tests. The sensitivities of the protocol for all types of modifications are in the range of 70%. The AutoMotif Server is freely available at "http://automotif.bioinfo.pl/":http://automotif.bioinfo.pl/. The local version of the software is available on request from the authors. The parameters \(the search type, the number of top models, and the PTM type) are optional and can be easily modified. The following protocol describes how to use AMS server to detect various types of post-translational modifications, and how to understand the resulting score for a given prediction.

Equipment

1. A typical personal computer with Linux, Apple Mac OSX or Windows operating system 2. Input single sequence or a set of sequences in FASTA file format, from experimental data or sequence databases 3. The internet web browser. We suggest using Firefox, but Apple Safari, Microsoft Internet Explorer or Mozilla suite are allowed.

Procedure

1. The AutoMotif Server \(AMS) dissects a query protein sequence into overlapping short sequence segments and identifies selected types of post-translational modification sites. We use supervised SVM classification trained on experimental knowledge for identification of PTM sites. Each sequence segment has assigned a real number calculated by the cost function of SVM classification model. Residues with have the value of cost function, i.e. the score larger than a given cut-off value are identified as possible modification sites. This means that the point representing this sequence segment is located in the region of multidimensional space classified as “positive” by the SVM model’s hyperplane within given cut-off value. In AMS web server we use only single, the most effective type of the kernel, i.e. the polynomial kernel. The one-vote-wins method is used to annotate segments that are predicted as positives by at least one classification model. 2. The AMS server accepts input sequences in the one-letter mode in capital letters: 'ACDEFGHIKLMNPQRSTVWY', with additional letter X for marking empty or unknown positions in a protein sequence, or extension of a sequence segment. Users can input sequences by submiting text file in FASTA file format \(for details see http://en.wikipedia.org/wiki/Fasta_format), or by providing the SWISS-PROT/TrEMBL identifier or accession number in the text box, or simly pasting the amino acids seqeuences. 3. The server predicts by default all types of post-translational modification sites that were precalculated by the authors and which are available with enough statistics in the Swiss-Prot database. The list presently include acetylation, amidation, hydroxylation, methylation, sulfation and phopshorylation \(by PKC, PKA, CK, CK2 and CDC2 protein kinases). The search can be limited by selecting particular type of functional motif from the drop-down menu on the server’s www page \(for example phosphorylation sites in general or by specific kinases). 4. Two types of search procedures are available on the server: the identity search and scan based on SVM classification. The first method identifies identical in terms of sequence 9 residues segments in a query protein and the database of positives for that selected type of modification. The second method runs several versions of SVM predictions that use different projection methods. The registration of a user \(by following the link ”User Site” from the main www page) allows for submiting his or her own list of training instances as a text file with the set of segments dissected from a multiple proteins known to pefrorm certain function. Then the AMS server train the SVM for the new type of functional motif and use it to scan any query protein sequence for potential substrates. This method allows for indroducing new types of biochemical process that are not yet known in public, or that are not contained in Swiss-Prot database . 5. The output www page for a query protein contains two sections. The first section displays results of predictions for each selected model, i.e. the parent protein information \(i.e. the sequence number in a query set), local segment sequences predicted as a modificated sites, their positions \(start, modified central residue, the end position, and the size of a segment) and the output scores. The second section of the output www page describes each used type of post-translational modification, its protein agent, the best SVM method used to classify known instances. Each SVM model is described by the number of positive and negative instances used in training, the precision and recall errors of the classification models. 6. The accuracy of SVM classification models is described by two numbers: the recall R and the precision P. The recall R value measures the percentage of correct predictions \(the probability of correct prediction), whereas

precision P gives the percentage of observed positives that are correctly predicted \(the measure of the reliability of positive instances prediction). The measures of accuracy are calculated separately for each type of PTM using the leave-one-out procedure. The typical recall value is around 30%, and the precision P is over 70% for majority of PTM. 7. In the case of single query protein applying the computational protocol give for each type of PTM the list of predicted modifications for this sequence. When a set of sequences is used as an input, the protocol returns the for each type of modification the list of predicted short sequence fragments that are modified with the parent protein number. The list of predicted modified sites is not ordered. 8. The consensus prediction is also available on the output web page, when several different versions of the method predict the same local sequence fragment to perform given post-translational modification.

Timing

The AMS server is able to predict all types of post-translational modifications for a query sequence in real time, even if multiple classification models are used. When large set of input sequences is used the time needed to perform the prediction is scaling linearly with the size of the set. If two sets are submitted at once, they are run in parallel on our linux cluster, so the time is the same as for single submission. Therefore the critical step for computations is the proper preparation of input data. **Database & Representations** 1. The AMS method for predicting plausible post-translational modification sites classifies known experimental instances. Only the sequence information is used as an input, because in most cases only the potential target protein sequence is known. Our analysis is based on biological information acquired from the Swiss-Prot database [5, 6]. 2. Proteins with acetylation, phosphorylation \(by PKA, PKC, CK, CK2 and CDC2 kinases), sulfation, amidation, hydroxylation, methylation, pyrrolidone and gamma-carboxyglutamic modification sites are selected for our analysis. Those processes have the largest number of known experimental instances. Training cases are taken from proteins experimentally annotated proteins. For each type of post-translational modification the list of proteins with at least one modified site of that particular type is fetched from the Swiss-Prot database. Sites annotated “by similarity”, “partial”, “potential”, “probable” or “predicted” are neglected in the analysis. The remaining list of residues is used as the dataset of positive cases, which includes all short sequence segments dissected from parent proteins with size of 9 amino acids and centered on main annotated residue. If the case of non-symetric segments, where modified residue is not in the center, the lacking positions in a segment are filled with ‘X’ as the type of amino acids. All redundant segments in the database, i.e. with the same sequence, are removed from the training dataset. 3. The “negative” preferences for each position in a short sequence segment for each type of post-translational modification is calculated using the negative instances dataset. We randomly select short sequence segments from proteins of Swiss-Prot database that have appropriate to selected type of modification the central amino acid that are not experimentally annotated to undergo this modification. Those two datasets: positive and negative instances for each type of functional motif are then used for the training of SVM. 4. Sequence segments are projected into the multidimensional space using ten different projections. The first representation \(BIN) encodes each position of a segment into a 20-dimensional vector of binary values 0 or 1. The 1

value denotes that corresponding type of amino acid is present at selected position of a segment and 0 otherwise. Therefore each the vector representing a segment contain 9 coordinates equal to 1, all other dimensions have 0 value. The second representation \(BLOSUM) uses the BLOSUM62 matrix for encoding each position of a segment by a 20 dimensional vector of the substitution scores between the amino acid present in the projected segment at this position and all other 20 types of amino acids. If Arg is found at first position in a segment we represent it by the appropriate Arg column from the BLOSUM62 substitution matrix. In the case of 9 amino acids long segments the representation is in 180 dimensional space \(constructed by 9 columns of the BLOSUM62 matrix for 9 types of amino acids that are present in the projected segment). The LOOKUP method represent each amino acid type in a segment as one dimensional scalar value that is equal to the normalized sequence preference for it. The normalized preferences are pre-calculated earlier for all 9 positions within a segment and for all types of amino acids. For example the Arg amino acid at first position of a segment has the normalization calculated by dividing the probability to find Arg at first position on annotated segments by the probability to find it at the first position of not annotated segments. The profile projection \(PROF) uses similar normalized preferences for each position of 9 residue long segment but storing them as 20-dimensional vectors preserving the information about all types of amino acids for a particular position. The normalized preferences for all types of amino acids at this position is multiplied by appropriate to found amino acid type column from the BLOSUM62 substitution matrix. If in the segment we find Arg the all amino acids preferences are multiplied by the Arg column of the substitution matrix. Each segment is represented as a point in 180 multi-dimensional space. The sparse representation \(SPARSE) takes the normalized preferences for the found type of amino acid at certain position of a segment instead of binary value. All other amino acids are marked by 0 values. In addition, the combinations of above generic representations are used in order to maximize the accuracy and efficiency of both representing the acquired biological knowledge and the training abilities of support vector machine. 5. The bioSQL database using ‘bioperl-db’ perl library is build directly from UniProtKB flat text file. The selection of positives was performed by querying bioSQL database by following MySQL procedure: SELECT count\(sqv.value), sqv.value FROM location l, seqfeature s, term t, seqfeature_qualifier_value sqv, biosequence bs WHERE l.seqfeature_id = s.seqfeature_id AND s.seqfeature_id = sqv.seqfeature_id AND t.term_id = s.type_term_id AND t.name = 'MOD_RES' AND s.bioentry_id = bs.bioentry_id AND bs.alphabet = 'protein' AND sqv.value NOT LIKE '%\(Probable)' AND sqv.value NOT LIKE '%\(Potential)' AND sqv.value NOT LIKE '%\(By similarity)' GROUP BY sqv.value ORDER BY count\(sqv.value) DESC Redundant samples were removed form output data \(except those with different BLAST profile for PROF method).

Critical Steps

1. The optimal way to investigate protein function is to use the complete parent protein sequence, not short parts of it. In that case the interesting non-local multiple modifications sites can be identified. 2. The output score is in the range \[0.000-5.000]. The higher the output score indicate the higher confidence of the predictions. 3. The predicted sequence fragments that are modified for certain type of post-translational modification can repeat in the output page with different reliability scores. Those variants

are predicted by different methods by the use of various projections. If more than one method predicts a site as modified, the prediction is more reliable even if low scores are presented. 4. In all types of post-translational modification sites the best type of a kernel is polynomial one. Representations mixed with LOOKUP projection \(like PROF+LOOKUP and BLOSUM+LOOKUP) are the most efficients. Other projections \(like generic BIN or PROF) have some advantages for particular types of modification sites, but they have lower overall efficiency \(small recall and precision values). When the number of positive instances is large the simple binary method BIN is becoming the most accurate one, whereas in the case of lower statistics profile methods gain better results. The SVM finds more easily proper classification scheme of the test set with simple representations than more complex ones. The linear kernel function in the case of more complicated sequence signatures of post-translational modification sites is not efficient. However in some cases \(PKA phosphorylation with SPARSE+LOOKUP representation) SVM models of this type reach efficiency of the polynomial kernel. In the case of radial basis kernel SVM frequently fails to build the model. In the case of large number of instances the simple LOOKUP method for this type of a kernel is the most accurate. The remarkable cases are acetylation, amidation and pyrrolidone cases, where the system with LOOKUP embedding reaches efficiency of the polynomial kernel.

Anticipated Results

1. The analysis of post-translational modification sites by support vector machine allows for quick and accurate \(very conservative) prediction of a protein function. The high overall precision of best methods allows user to gain deep insight in plausible functional characteristics of unknown new proteins. The recall efficiency ensures that information from previously verified sites will be not lost during automatic scans of known instances. The algorithm can by applied independently from the Web interface in a pipe-line. Large scale genomes analysis is also possible. 2. The main problem for some types of functional modifications is the insufficient number of experimentally verified instances. The number of support vectors for some of our classification models is very large – which is explained by the large dimensionality of the embedding space in such cases and the complicated shape of the separation hyperplane between positive and negative instances. The number of support vectors can be lowered when one chooses low dimensional initial encoding of the amino acids into the general physicochemical properties \(like hydrophobicity, hydrophilicity, polarity, volume, surface area, bulkiness or refractivity). We are working now on incorporating those features for recent update of our service, which will be available within one month.

References

1. Vapnik, V. N. The nature of statistical learning theory. Springer: New York, 1995; p xv, 188. 2. Vapnik, V. N. Statistical learning theory. Wiley: New York, 1998; p xxiv, 736 p. 3. Plewczynski, D.; Tkacz, A.; Wyrwicz, L. S.; Godzik, A.; Kloczkowski, A.; Rychlewski, L. Support-vector-machine classification of linear functional motifs in proteins. _J Mol Model_ 2006, **12**, 453-61. 4. Plewczynski, D.; Tkacz, A.; Wyrwicz, L. S.; Rychlewski, L. AutoMotif server: prediction of single residue post-translational modifications in proteins.

Bioinformatics 2005, **21**, 2525-7. 5. Bairoch, A.; Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. _Nucleic Acids Res_ 1999, **27**, 49-54. 6. Junker, V. L.; Apweiler, R.; Bairoch, A. Representation of functional information in the SWISS-PROT data bank. _Bioinformatics_ 1999, **15**, 1066-7.

Acknowledgements

This work was supported by EC BioSapiens \(LHSG-CT-2003-503265) 6FP project as well as the Polish Ministry of Education and Science \(PBZ-MNiI-2/1/2005 and 2P05A00130).

Figures

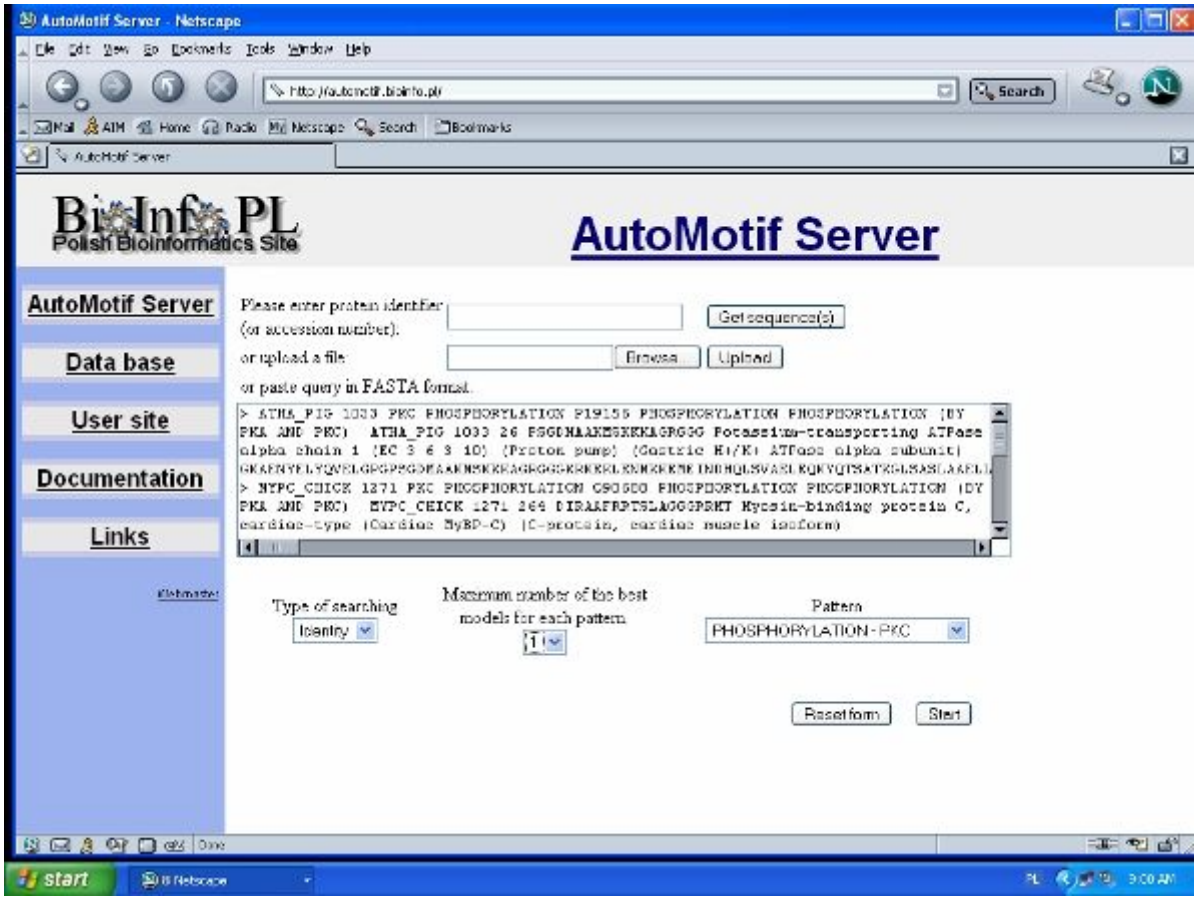


Figure 1

Figure IA

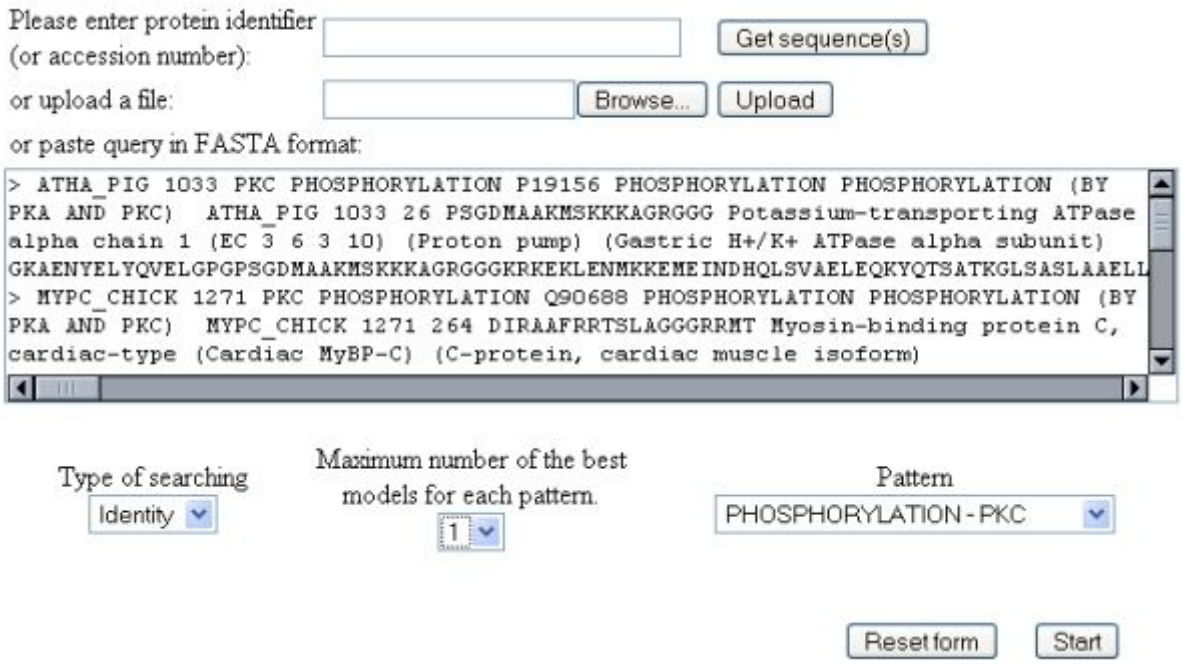


Figure 2

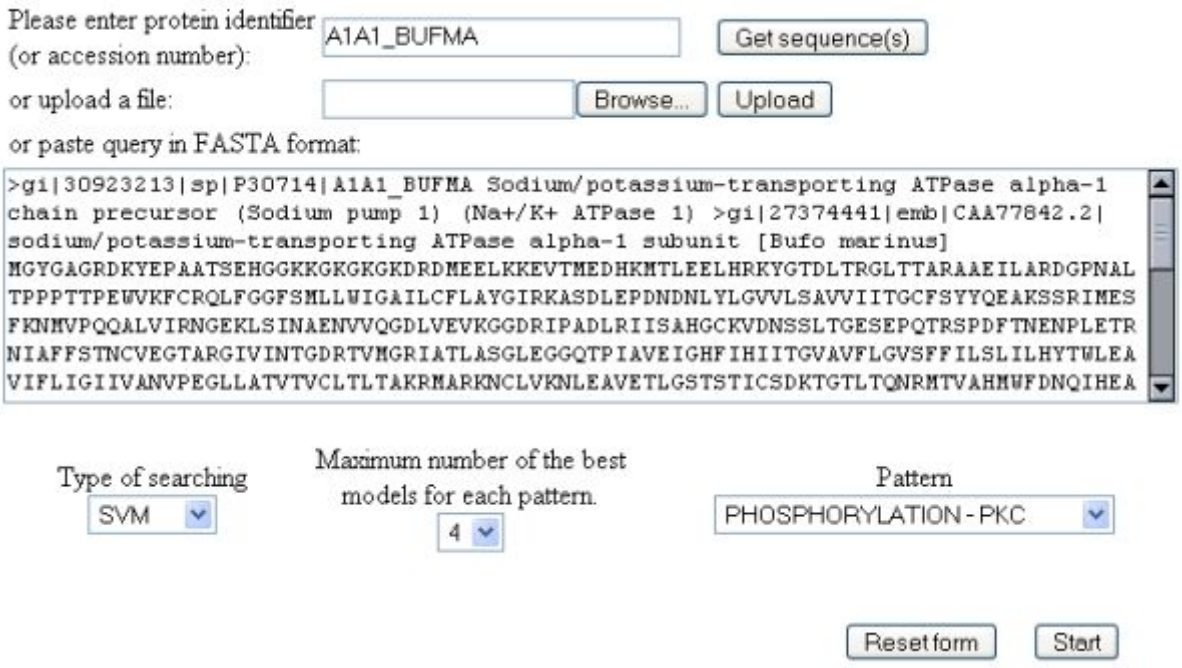


Figure 3

Figure IA. Examples of the server input Web page for a set of sequences (top and middle pictures) and for A1A1_BUFMA protein (bottom). On the middle picture we present the identity search options (scan for identical sequence segments) for PHOSPHORYLATION BY PKC active sites, and the bottom picture presents the Web server page for SVM scan for PHOSPHORYLATION.

Pattern	by	SVM method	LOO Precision	LOO Recall	Positives	Negatives
PHOSPHORYLATION	PKC	BIN	100	1.79	56	10000

Pattern: PHOSPHORYLATION - PKC

Protein	Predicted motif	Start	Centre	End	Size	Score
1	AAKMSKKKA	22	26	30	9	**1.0000**
2	FRRTSLAGG	260	264	268	9	**1.0000**
2	GRRMTSAFL	269	273	277	9	**1.0000**
2	YKPPSYKEH	1164	1168	1172	9	**1.0000**
3	GKSPSKKKK	712	716	720	9	**1.0000**
3	FRTPSFLKK	722	726	730	9	**1.0000**
4	SKSPSKKKK	699	703	707	9	**1.0000**
4	FRTPSFLKK	709	713	717	9	**1.0000**

Figure 4

Figure IB

Pattern	by	SVM method	LOO Precision	LOO Recall	Positives	Negatives
PHOSPHORYLATION	PKC	BIN	100	1.79	56	10000
PHOSPHORYLATION	PKC	BLOSUM SUM PROF	83.33	17.86	56	10000
PHOSPHORYLATION	PKC	PROF LOOKUP	62.5	17.86	56	10000
PHOSPHORYLATION	PKC	SPARSE	44.44	14.29	56	10000

Pattern: PHOSPHORYLATION - PKC

Protein	Predicted motif	Start	Centre	End	Size	Score
3	GKSPSKKKK	712	716	720	9	**0.3369**
4	SKSPSKKKK	699	703	707	9	**0.1910**
2	YKPPSYKEH	1164	1168	1172	9	**0.9111**
3	GKSPSKKKK	712	716	720	9	**1.1473**
3	FRTPSFLKK	722	726	730	9	**0.8664**
4	SKSPSKKKK	699	703	707	9	**1.0688**
4	FRTPSFLKK	709	713	717	9	**0.8664**
1	AAKMSKKKA	22	26	30	9	**0.2109**
2	YKPPSYKEH	1164	1168	1172	9	**0.9333**
3	GKSPSKKKK	712	716	720	9	**1.4074**
3	FRTPSFLKK	722	726	730	9	**0.8610**
4	SKSPSKKKK	699	703	707	9	**1.1527**
4	FRTPSFLKK	709	713	717	9	**0.8610**
1	AAKMSKKKA	22	26	30	9	**1.1678**

Figure 5

Figure IB. Examples of the server output Web page for “PHOSPHORYLATION BY PKC” scan in set of proteins and A1A1_BUFMA protein. On the top the output for Identity search, on the bottom for the SVM search.

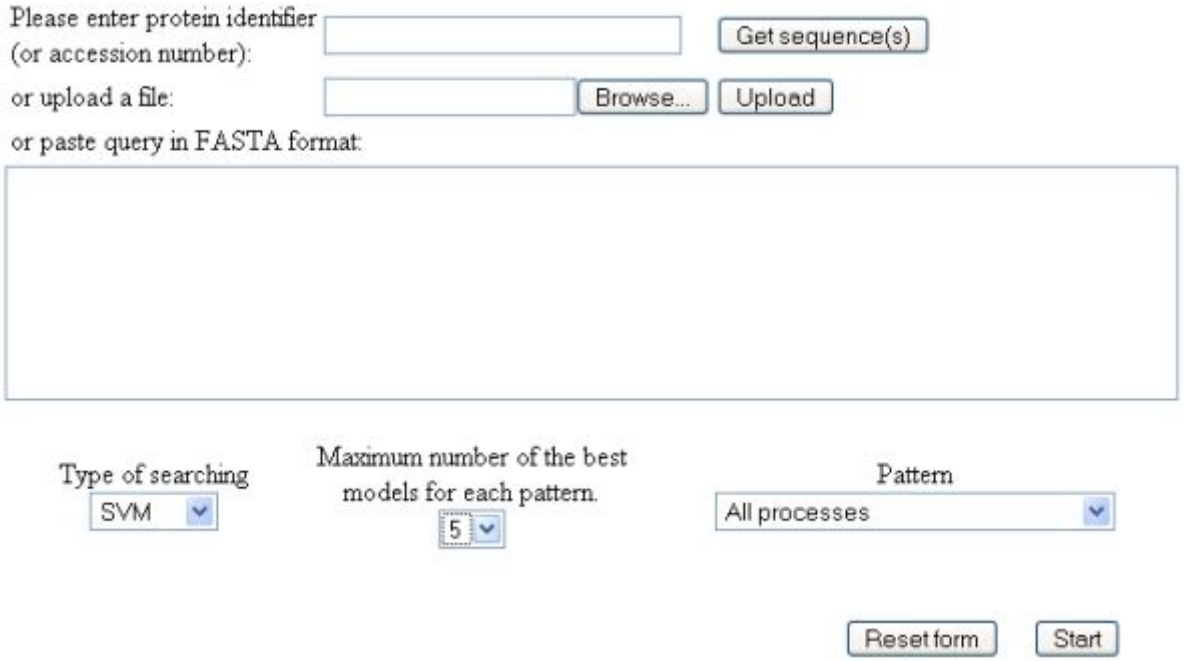
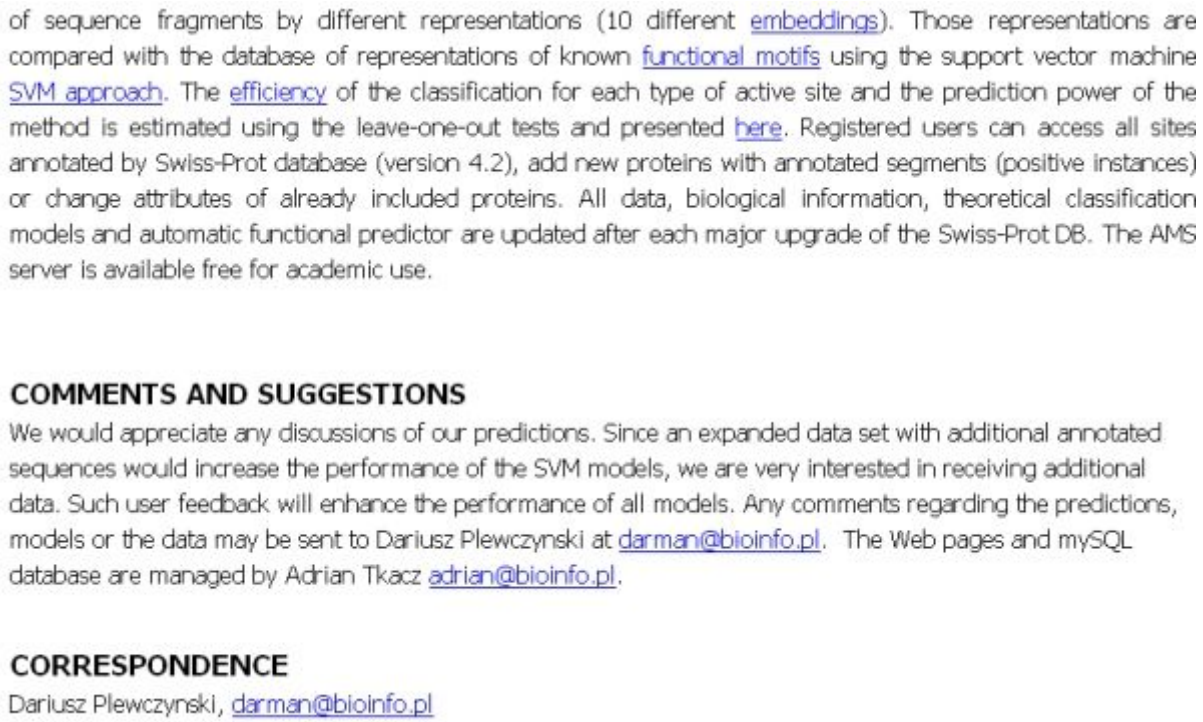


Figure 6

Figure IIA.


of sequence fragments by different representations (10 different embeddings). Those representations are compared with the database of representations of known functional motifs using the support vector machine SVM approach. The efficiency of the classification for each type of active site and the prediction power of the method is estimated using the leave-one-out tests and presented here. Registered users can access all sites annotated by Swiss-Prot database (version 4.2), add new proteins with annotated segments (positive instances) or change attributes of already included proteins. All data, biological information, theoretical classification models and automatic functional predictor are updated after each major upgrade of the Swiss-Prot DB. The AMS server is available free for academic use.

COMMENTS AND SUGGESTIONS

We would appreciate any discussions of our predictions. Since an expanded data set with additional annotated sequences would increase the performance of the SVM models, we are very interested in receiving additional data. Such user feedback will enhance the performance of all models. Any comments regarding the predictions, models or the data may be sent to Dariusz Plewczynski at darman@bioinfo.pl. The Web pages and mySQL database are managed by Adrian Tkacz adrian@bioinfo.pl.

CORRESPONDENCE

Dariusz Plewczynski, darman@bioinfo.pl

Figure 7

Figure IIA. The AutoMotif Server main Web pages. On the top we present the main input page, and on the bottom there is Documentation section of the AMS

Figure 8

Figure IIB

Pattern	by	Filter	Amount	Status	Instances
2-AMINO-3-OXOPROPIONIC ACID	-			Waiting	
ACETYLATION	-	ON	552	OK	Display
ADP-RIBOSYL	-			Waiting	
ADP-RIBOSYL	CTX			Waiting	
ADP-RIBOSYL	IAP			Waiting	
ALKYLATION	-			Waiting	
ALKYLATION	SH-1			Waiting	
ALKYLATION	SH-2			Waiting	
AMIDATION	-	ON	723	OK	Display
AMIDATION	G-108			Waiting	
AMIDATION	G-29			Waiting	
AMIDATION	G-34			Waiting	
AMIDATION	G-63			Waiting	
AMIDATION	G-65			Waiting	
AMIDATION	G-93			Waiting	
AMIDE-LINKED TO CELL WALL	-			Waiting	
BLOCKED	-			Waiting	
CITRULLINE	-			Waiting	
CONVERTED TO A PYRUVOYL GROUP	-			Waiting	
D-ALANINE	-			Waiting	
DEAMIDATION	-			Waiting	

Figure 9

Figure IIB. The AutoMotif Server links and “User Site” Web pages. On the top we present all available links to functional prediction sites, databases and servers, and on the bottom there is Swiss-Prot based set of functional motifs used in service.

To access for AutoMotif Server Users Site please sign in first.

Login	
Password	
Sign in	

Figure 10

Figure IIIA

Figure 11

Hello Dariusz Plewczynski!

Please wait a while... (up to 5 min.)

Loading positives ... OK
Creating profile for positives ... OK
Creating and saving negatives... OK
Saving example file for positives... OK
Saving example file for negatives... OK
svm_learn has been started... OK
svm_classify has been started... OK

The SVM model for Active Site has following accurancy:
Leave-One-Out estimate of the error: 10.95%
Leave-One-Out estimate of the recall: 17.86%
Leave-One-Out estimate of the precision: 100.00%
Test classification (all instances) results:
Accuracy on test set: 99.29% (417 correct, 3 incorrect, 420 total)
Precision/recall on test set: 100.00%/94.64%

Finished!

Figure 12

Figure IIIA. The AutoMotif Server “User Site” section of the service. On the top picture we present user’s login screen, and on the middle the uploading of a set of positives from the text file. It is important that input file should include only segments with the same length centered on the functional site, no empty lines. The input file format uses only single-upper case letters to mark amino acids, and cannot include more than 1000 positives. Below we present the output page of building the user’s own model from the set of supplied positives.

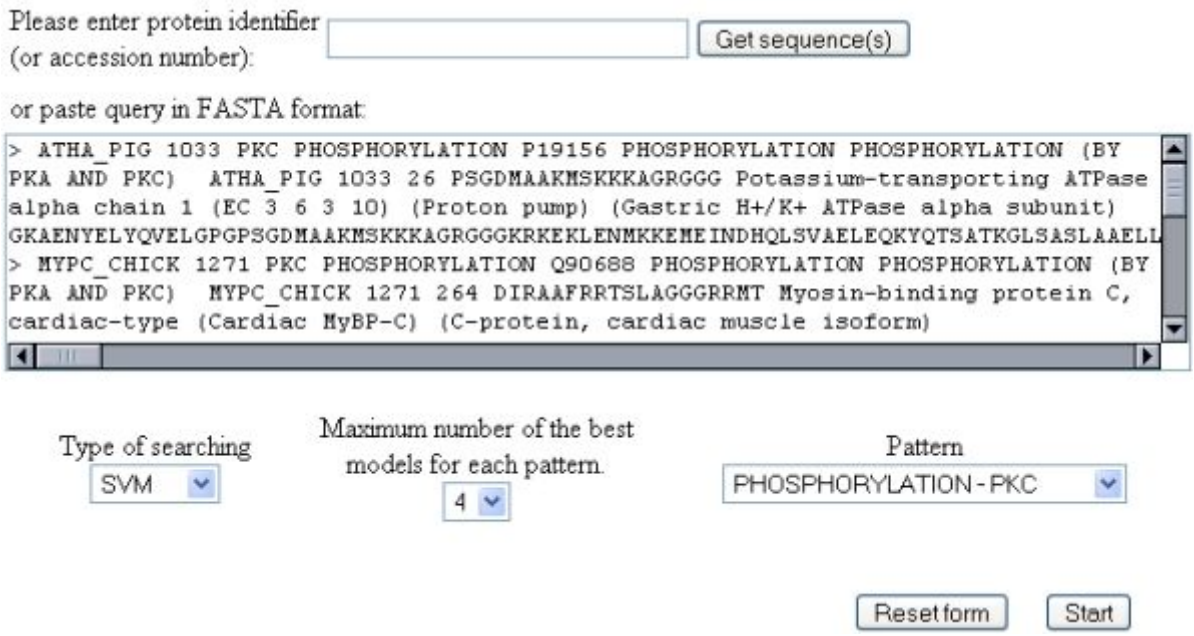


Figure 13

Figure IIIB

Pattern	by	SVM method	LOO Precision	LOO Recall	Positives	Negatives
MySetOfPos	-	BIN	100	17.86	56	364

Pattern: MySetOfPos

Protein	Predicted motif	Start	Centre	End	Size	Score
1	AAKMSKKKA	22	26	30	9	**0.9074**
2	LKKPSVKWF	173	177	181	9	**0.1724**
2	FRRTSLAGG	260	264	268	9	**0.2607**
2	GRRMTSAFL	269	273	277	9	**0.1692**
2	YKPPSYKEH	1164	1168	1172	9	**0.4420**
3	GKSPSKKKK	712	716	720	9	**1.0000**
3	FRTPSFLKK	722	726	730	9	**0.6034**
3	FLKKSKKKS	727	731	735	9	**0.3512**
4	SKSPSKKKK	699	703	707	9	**0.9994**
4	FRTPSFLKK	709	713	717	9	**0.6034**
4	FLKKSKKKE	714	718	722	9	**0.1147**

Figure 14

Figure IIIB. The AutoMotif Server “User Site” search section of the service. On the top we present the input page with set of proteins and on the bottom picture the scan results for user supplied set of positives representing “MySetOfPos” functional motif.

Example I

The AutoMotif Server (AMS) accepts input sequences in the one-letter code in capital letters: 'ACDEFGHIKLMNPQRSTVWY', with additional letter X for marking empty unknown positions in a protein chain or positions extending a sequence chain outside its ends. There are three methods for a user to input sequences. He or she can: • submit a file in FASTA format with single or multiple sequences directly from local disk.
• enter SWISS-PROT/TrEMBL protein identifier (or accession number) and press a "Load sequence" button. The protein sequence will appear below in the text field.
• or simple paste a single or several sequences in FASTA format directly into the input text field.
If user pastes more than one sequence in FASTA format the '>' delimiter must be used (see Example IA). In the case of single sequence user can omit description line and use only bare sequence. It is also possible to submit short peptide sequences (see Example IB). If sequences are shorter than 9 amino acids user should use X to mark empty positions, or positions outside ends of peptide. User must submit all 9 positions always puting the main residue to evaluate in the center and including 4 flanking residues on both sides. The optimal way to investigate sequences is to use the complete parent protein sequence.

Example IA

Examples of input sequences in FASTA format. The detailed description of the FASTA format is available at NCBI web page http://www.ncbi.nlm.nih.gov/BLAST/fasta.html
Example for four real protein sequences known to be phosphorylated by PKC kinase:
> ATHA_P1G 1033 PKC PHOSPHORYLATION P19156 PHOSPHORYLATION PHOSPHORYLATION (BY PKA AND PKC) ATHA_P1G 1033 26 PSGDMAAKMSKKKKAGRGGGKRKEKLENMKKEMEINDHQLS VAELEGKYQTSATKGLSASLAAELLLRDGPNALRPPRGTPEYVKFARQLAGGLQCL MWVAAAICLIAFAIQASEGDLTTDDNLYLALALIAVVVVTGCFGYYQEFKSTNIIA SFKNLVPQQATVIRDGDKFQINADQLVVGDLVEMKGGDRVPADIRILQAQGRKVDN SSLTGESEPQTRSPECTHESPLETRNIAFFSTMCLEGTAQGLVVNTGDRTIIGRIA SLASGVENEKTPIAIEIEHFVDIIAGLALFGATFFIVAMCIGYTFLRAMVFFMAI VVAYVPEGLLATVTVCLSLTAKRLASKNCVVKNLEAVETLGSTSVICSDKTGTLTQ NRMTVSHLWFDNHIHSADTTEDQSGQTFDQSSETWRALCRVLTLCNRAAFKSGQDA VPVPKRIVIGDASETALLKFSELTLGNAMGYRERFPKVCEIPFNSTNKFQLSIHTL EDPRDPRHVLVMKGAPERVLERCSSILIKGQELPLDEQWREAFQTAYLSLGGLGER VLGFCQLYLSEKDYPPGYAFDVEAMNFPTSGLSFAGLVSMIDPPRATVPDAVLKCR TAGIRVIMVTGDHPITAKAIAASVGIISEGSETVEDIAARLRVPVDQVNRKDARAC VINGMQLKDMDPSELVEALRTHPEMVFARTSPQQKLVIVESCQRLGAIVAVTGDGV NDSPALKKADIGVAMGIAGSDAAKNAADMILLDDNFASIVTGVEQGRLIFDNLKKS IAYTLTKNIPELTPYLIYITVSVPLPLGCITILFIELCTDIFPSVSLAYEKAESDI MHLRPRNPKRDRLVNEPLAAYSYFQIGAIQSFAGFTDYFTAMAQEGWFPLLCVGLR PQWENNHHLQDLQDSYGQEWTFGQRLYQQYTCYTVFFISIEMCQIADVLIRKTRRLS AFQQGFFRNRILVIAIVFQVCIGCFLCYCPGMPNIFNFMPIRFQWWLVPMPFGLLI FVYDEIRKLGVRCCPGSWWDQELYY
> MYPC_CHICK 1271 PKC PHOSPHORYLATION Q90688 PHOSPHORYLATION PHOSPHORYLATION (BY PKA AND PKC) MYPC_CHICK 1271 264 DIRAAFRRTSLAGGGRRMT Myosin-binding protein C, cardiac-type (Cardiac MyBP-C) (C-protein, cardiac muscle isoform) 2 PEPAWKAVSAFTKRKTTEVAAGSTAVFEAETEKTGIKVKWQRAGTEITDSEKYAI KAEGNKHSLTISNVGKDDEVTYAVIAGTSKVKFELKVKEPEKSEPVAPAEASPAPA ASELPAPPVESNGNPEVPPAETQPEEPVDPIGLFVTRPQDGEVTVGGNITFTAKVA GESLLKKPSVKWFKGKWMDLASKVGKHLQLHDNYDRNNKVYTFEMEIEANMTFAG GYRCESVSTKDKFDSSNFNLIVNEAPVSGEMDIRAAFRRTSLAGGGRRMTSAFLSTE GLEESGELNFSALLKKRDSFLRTANRGDGKSDSQPDVDVWEILRKAPPSEYEKIAF QYGITDLRGMLKRLKRIKKEEKKKSTAFLKKLDPAYQVDKGQKIKLMVEVANPDADV KWLKNGQEIQVSGSKYIFEAIGNKRILTINHCSLADDAAYECVVAEEKSFTELFVK EPPILITHPLEDQMVMVGERVEFECEVSEEGATVKWEKDGKVELTREETFKYRFKKD GKKQYLIINESTKEDSGHYTVKTNGGVSVAELIVQEKKLEVYQSIADLTVKARDQA VFKCEVSDENVKGWLKNGKEVVPDERIKIGRIHKLTIEDVTPGDEADYSFIP QGFAYNLSAKLQFLEVKIDFVPREEPPKIHLDCLGQSPDTIVVVAGNKLRLDVPIS GDPTPTVWWGKVNKKGELVHQSNEDSLTFSENSSDLSTDSKLLFESEGRVRVEKHE DHCVFIIEGAEKEDEGVYRVIVKNPVGEDKADITVKVIDVPDPPEAPKISNIGEDY CTVQWQPPTYDGGGPVLGYILERKKKKSYRWMRLNFDLLKELTYEAKRMIEGVVYE MRIYAVNSIGMSRPSPASQRFMPIAPPSEPTHFTVEDVSDTTVALKWRPPERIGAG GLDGYIVEYCKDGSAEWTPALPGLTERTSALIKDLVTGDKLYFRVKAINLAGESGA AIIKEPVTVQEIMQRPKICVPRHLRQTLVKKVGETINIMIPFGGKPRPKISWMKDG QTLDSKDVGIRNSSTDTILFIRKAELHHSGAYEVTLQIENMTDTVAITIQIIDKPG PPQNIKLADVWGFNVALEWTPPQDDGNAQILGYTVQKADKKTMEWYTVYDHYRRTN CVVSDLIMGNEYFFRVFSEMLCGLSETAATTKNPAYIQKTGTTYKPPSVKEHDFSE PPHFTHPLVNRSVIAGYNTTLSCAVRGIPKPKIFWYKNKVDLSGDAKYRMFSKQSV LTLEIRKPTPLDGGFYTCKAVNERGEAEIECRLDVRVPQ
> ADDA_HUMAN 737 PKC PHOSPHORYLATION P35611 PHOSPHORYLATION PHOSPHORYLATION (BY PKC AND PKA) ADDA_HUMAN 737 726 KKKKKKFRTPSFLKKSKKKS Alpha adducin (Erythrocyte adducin alpha subunit) 3 MNGDSRAAVVTSPPPTTAPHKERYFDRVDENNPEYLRERNMAPDLRQDFNMMEGKK RVSMILQSPAFCEELESMIGEGFKKIGKNPTGLLALQGIADFMTTNVPNVYPAAPQG GMAALNMSLGMVTPVNDLRGGDSIAYDKGEKLLRCKLAAFYRLADLFGWSQLIYNH ITTRVNSEQEHFLVPFGLLYSEVTASSLVKINLQGDIVDRGSTNLGVNQAGFTLH SAIYAARPDVKCVVHIHTPAGAAVSAMKCGLLPISPEALSLGEVAYHDYHGILVDE EEKVLIQKNLGPKSKVLILRNHGLVSVGESVEEAFYYIHNLVVACEIQVRTLASAG GPDNLVLLNPEKYKAKSRSPGSPVGEGTGSPPKWGIQEQEFEALMRMLDNLGYRTG YPYRYPALREKSKKYSDVEVPASVTGYSFASDGDSGTCSPLRHSFGKQGREKTRWL NSGRGDEASEEGQNGSSPKSKTKWTKEDGHRTSTSAVPNLFVPLNTNPKEVQEMRN KIREQNLQDIKTAGPQSQVLCGVVMDRSLVQGELVTASKAIEKEYQPHWIVSTTG PNPFTTLTDRELEEYRREVERQKGSEENLDEAREQKEKSPPDQPAVPHPPSTPI KLEEDLVPEPTTGDDSDAATFKPTLPDLSPDEPSEALGFPMLEKEEEAHRPPSPTE APTEASPEPAPDPAPVAEEAAPSAVEEGAAADPGSDGSPGKSPSKKKKKFRTPSFL KKSKKKSDS
> ADDB_HUMAN 726 PKC PHOSPHORYLATION P35612 PHOSPHORYLATION PHOSPHORYLATION (BY PKC AND PKA) ADDB_HUMAN 726 713 KKKKKFRTPSFLKKSKKKE Beta adducin (Erythrocyte adducin beta subunit) 4 MSEETVPEAASPPPPQGGPYFDRFSEDDPEYMRLRNRAADLRQDFNLMEGKKRVTM ILQSPSFREELGLIGEQMKGKMNSSNWALRQIADFMAGTSHAVFPTSGMNVSMM TPINDLHTADSLNLAKGERLMRCKISSVYRLLDLYGWAQLSDTYVTLRVSKEQDHF LISPKGVSCSEVTASSLIKVNILGEVVEKGSSCFPVDTTGFCLHSAIYAARPDVRC IIHLHTPATAAVSAMKWGLLPVSHNALLVGDMAYYDFNGEMEQEADRINLQKCLGP TCKILVLRNHGVVALGDTVEEAFYKIFHLQAACEIQVSALSAGVGVENLILLEQEK HRPHEVGSVQWAGSTFGPMQKSRLGEHEFEALMRMLDNLGYRTGYTYRHPFVQEKT KHKSEVEIPATVTAFVFEEDGAPVPALRQHAGKQGKEKTRWLNTPNTYLRVNVADE VQRSMGSPRPKTTVMKADEVEKSSSGMPIRIENPNGFVPLYTDPQEVLEMRNKIRE QNRQDVKSAGPQSQLLASVIAEKSRSPSTESQLMSKGDEDTKDSEETVPNPFSQL TDQELEEYKKEVERKKLELDGEKETAPEEPGSPAKSAPASPVQSPAKEAETKSPLV SPSKSLEEGTKKTETSKAATTEPETTQPEGVVVNGREEEQTAEEILSKGLSQMTTS ADTDVDTSKDKTESVTSGPMSPEGSPSKSPSKKKKKFRTPSFLKKSKKKEKVES

Example IB
The set of peptides phosphorylated by PKC kinase can be also pasted into the browser text window in the same format:
> pept 1
EPAATSEHG
> pept 2
PAATSEHGG
> pept 3
PAAVSEHGD
> pept 4
GDKKSKKAK
> pept 5
GKSPSKKKK
> pept 6
FRTPSFLKK
> pept 7
SKSPSKKKK
> pept 8
FRTPSFLKK
> pept 9
EYIKSVKGG
> pept 10
SAYGSVKAY
> pept 11
SAYATVKAY
> pept 12
SAYGSVKAY
> pept 13
SAYGSVKPY
> pept 14
SKLGSVKAA
> pept 15
AKGGTVKAA
> pept 16
NRIQTQMDV
> pept 17
AAKMSKKKA
> pept 18
AARTSPLRP
> pept 19
TKKQSFKQT
> pept 20
KTTASTRKV

Figure 15

Example I

By default the server predicts all available in the database types of biochemical processes, posttranslational modifications. User can limit his or her scan by choosing the particular process from the drop-down list (for example phosphorylation sites in general or by specific kinases). The available functional motifs up to now include:

- ACETYLATION
- AMIDATION
- GCG_ACID
- HYDROXYLATION
- METHYLATION
- SULFATION
- PHOSPHORYLATION
- PHOSPHORYLATION BY PKC
- PHOSPHORYLATION BY PKA
- PHOSPHORYLATION BY CK
- PHOSPHORYLATION BY CK2
- PHOSPHORYLATION BY CDC2
- PHOSPHORYLATION BY ABL

The user can choose two types of scan – either identity search or SVM method scan (see Figure IA). The first one performs a simple search over the database of collected from Swiss-Prot instances for functional motifs.

When it finds the exact matches in terms of short (9aa) sequence strings it displays them. The second one runs SVM search with various embedding methods in order to scan a query sequence for certain type of process. The maximum number of the best models for each pattern is fixed at 5. One can scan a query sequence with smaller number of best models by choosing the preferable option from the drop-down list on the main page of the server.

After preparing a sequence(s) user starts a scan by pressing the button labelled 'Start'. The server is working on-line, so one need to wait for a moment to see results of his or her query directly in the browser window.

The output from AutoMotif contains two main parts (see examples below).

The first part is a large table with information about all used in a scan types of patterns and SVM models. Each method is constructed for certain type of embedding, type of pattern (for example PHOSPHORYLATION) and modification "BY" of it (for example BY PKC). The number of positives and negatives used in training of SVM for this partivcular type is provided in the last two columns. The precision and recall errors for used methods calculated automatically during the training phase by Leave-One-Out test is presented in the middle two columns.

The second part of the output provides in tables results of the predictions for each type of a pattern (see Figure IB). The name of a type of process (functional pattern) is displayed in the front of each table. Following it the predictions are printed in tables with 7 columns containing:

- Column 1: The sequence position in a list (with accordance to submitted list of sequences: the first protein or a peptide is marked by 1, the second one by 2 etc.).
- Column 2: The predicted segment with accordance to the pattern. The central residue is provided with sequence context (shown as a 9-residue sequence string centered on the residue being analyzed).
- Column 3: The start position of the segment predicted as a functional motif .
- Column 4: The center position of the segment that is the position of a modificated residue in predicted functional motif.
- Column 5: The end position of the segment predicted as a functional motif in a query protein(s).
- Column 6: The size of the segment (or functional motif).
- Column 7: The output score for a fragment with value in the range [0.000-5.000].

The potential functional motifs sometimes are repeated when predicted by various methods (with different scores). Each method predicts a little different set of sequence segments as the functional motifs. Our automatic predictor uses identity search or SVM scan.

Figure 16

Example II

Pattern	PHOSPHORYLATION
by	PKC
SVM method	BIN
LOO Precision	100
LOO Recall	1.79
Positives	56
Negatives	10000

Pattern: **PHOSPHORYLATION - PKC**

Protein	Predicted motif	Start	Centre	End	Size	Score
1	AAKMSKKKA	22	26	30	9	**1.0000**
2	FRRTSLAGG	260	264	268	9	**1.0000**
2	GRRMTSAFL	269	273	277	9	**1.0000**
2	YKPPSYKEH	1164	1168	1172	9	**1.0000**
3	GKSPSKKKK	712	716	720	9	**1.0000**
3	FRTPSFLKK	722	726	730	9	**1.0000**
4	SKSPSKKKK	699	703	707	9	**1.0000**
4	FRTPSFLKK	709	713	717	9	**1.0000**

The SVM scan gives the following list of predicted motifs (using 4 best methods):

SVM method	LOO Precision	LOO Recall	Positives	Negatives
BIN	100	1.79	56	10000
BLOSUM_SUM_PROF	83.33	17.86	56	10000
PROF_LOOKUP	62.5	17.86	56	10000
SPARSE	44.44	14.29	56	10000

Pattern: **PHOSPHORYLATION - PKC**

Figure 17

Example IIA

Protein	Predicted motif	Start	Centre	End	Size	Score
3	GKSPSKKKK	712	716	720	9	**0.3369**
4	SKSPSKKKK	699	703	707	9	**0.1910**
2	YKPPSYKEH	1164	1168	1172	9	**0.9111**
3	GKSPSKKKK	712	716	720	9	**1.1473**
3	FRTPSFLKK	722	726	730	9	**0.8664**
4	SKSPSKKKK	699	703	707	9	**1.0688**
4	FRTPSFLKK	709	713	717	9	**0.8664**
1	AAKMSKKKA	22	26	30	9	**0.2109**
2	YKPPSYKEH	1164	1168	1172	9	**0.9333**
3	GKSPSKKKK	712	716	720	9	**1.4074**
3	FRTPSFLKK	722	726	730	9	**0.8610**
4	SKSPSKKKK	699	703	707	9	**1.1527**
4	FRTPSFLKK	709	713	717	9	**0.8610**
1	AAKMSKKKA	22	26	30	9	**1.1678**
2	FRRTSLAGG	260	264	268	9	**0.1762**
2	YKPPSYKEH	1164	1168	1172	9	**0.9680**
3	GKSPSKKKK	712	716	720	9	**1.9094**
3	FRTPSFLKK	722	726	730	9	**0.8428**
4	SKSPSKKKK	699	703	707	9	**1.0625**
4	FRTPSFLKK	709	713	717	9	**0.8428**

Figure 18

The results for a both types of scans for a set of four proteins. All of them are known to be phosphorylated by PKC kinase (which is verified by experimental results and annotated in Swiss-Prot database). The higher the score indicate the higher confidence of the predictions. This means that potential segments are more similar to one or more of the phosphorylation functional motifs used in training of SVM method.

Pattern	PHOSPHORYLATION
by	PKC
SVM method	BIN
LOO Precision	100
LOO Recall	1.79
Positives	56
Negatives	10000

Pattern: **PHOSPHORYLATION - PKC**

Protein	Predicted motif	Start	Centre	End	Size	Score
1	EPAATSEHG	1	5	9	9	**1.0000**
2	PAATSEHGG	1	5	9	9	**1.0000**
3	PAAVSEHGD	1	5	9	9	**1.0000**
4	GDKKSKKAK	1	5	9	9	**1.0000**
5	GKSPSKKKK	1	5	9	9	**1.0000**
6	FRTPSFLKK	1	5	9	9	**1.0000**
7	SKSPSKKKK	1	5	9	9	**1.0000**
8	FRTPSFLKK	1	5	9	9	**1.0000**
9	EYIKSVKGG	1	5	9	9	**1.0000**
10	SAYGSVKAY	1	5	9	9	**1.0000**
11	SAYATVKAY	1	5	9	9	**1.0000**
12	SAYGSVKAY	1	5	9	9	**1.0000**
13	SAYGSVKPY	1	5	9	9	**1.0000**
14	SKLGSVKAA	1	5	9	9	**1.0000**
15	AKGGTVKAA	1	5	9	9	**1.0000**
16	NRIQTQMDV	1	5	9	9	**1.0000**
17	AAKMSKKKA	1	5	9	9	**1.0000**
18	AARTSPLRP	1	5	9	9	**1.0000**
19	TKKQSFKQT	1	5	9	9	**1.0000**
20	KTTASTRKV	1	5	9	9	**1.0000**

The output of SVM method search for phosphorylation by PKC kinase sites provides following result:

SVM method	LOO Precision	LOO Recall	Positives	Negatives
BIN	100	1.79	56	10000
BLOSUM_SUM_PROF	83.33	17.86	56	10000
PROF_LOOKUP	62.5	17.86	56	10000
SPARSE	44.44	14.29	56	10000

Pattern: **PHOSPHORYLATION - PKC**

Figure 19

Example IIB

Protein	Predicted motif	Start	Centre	End	Size	Score
3	GKSPSKKKK	712	716	720	9	**0.3369**
4	SKSPSKKKK	699	703	707	9	**0.1910**
2	YKPPSYKEH	1164	1168	1172	9	**0.9111**
3	GKSPSKKKK	712	716	720	9	**1.1473**
3	FRTPSFLKK	722	726	730	9	**0.8664**
4	SKSPSKKKK	699	703	707	9	**1.0688**
4	FRTPSFLKK	709	713	717	9	**0.8664**
1	AAKMSKKKA	22	26	30	9	**0.2109**
2	YKPPSYKEH	1164	1168	1172	9	**0.9333**
3	GKSPSKKKK	712	716	720	9	**1.4074**
3	FRTPSFLKK	722	726	730	9	**0.8610**
4	SKSPSKKKK	699	703	707	9	**1.1527**
4	FRTPSFLKK	709	713	717	9	**0.8610**
1	AAKMSKKKA	22	26	30	9	**1.1678**
2	FRRTSLAGG	260	264	268	9	**0.1762**
2	YKPPSYKEH	1164	1168	1172	9	**0.9680**
3	GKSPSKKKK	712	716	720	9	**1.9094**
3	FRTPSFLKK	722	726	730	9	**0.8428**
4	SKSPSKKKK	699	703	707	9	**1.0625**
4	FRTPSFLKK	709	713	717	9	**0.8428**

Figure 20

The output of ”phosphorylation by PKC kinase” scan for a set of peptides. The predicted to be phosphorylated peptides are repeated in the case of SVM scan because they are predicted by various methods (with different scores). Each method provides a different, although ovelapping set of peptides to be phosphorylated by PKC kinase. The identity search provides the results above.