Support Vector Machine approach to Active Sites Prediction using Local Sequence Information.

Dariusz Plewczynski[1], Adrian Tkacz[1], Leszek Rychlewski[1*]

Keywords: kinase substrate prediction, nearest neighbor, sequence similarity, database of active sites, Swiss-Prot database, support vector machine

1 Introduction.

The AutoMotif Server (AMS) predicts functional patterns in proteins. A list of possible functional motifs for a given query protein is predicted using only query protein sequence and the database of proteins annotated for certain types of biological processes by Swiss-Prot database [1]. All short segments of a query protein sequence sites are compared with the annotated sequence fragments using the support vector machine SVM approach [3]. Various methods are used here for building models based on different representations (in total 10 different embeddings) of known instances [4]. In order to estimate the efficiency of the classification for each type of functional site and the prediction power of the method the leave-one-out tests are used [2]. User can access all sites annotated by Swiss-Prot database (version 4.2), add new proteins with instances, or new annotation information. All data, constructed models and automatic predictor are updated after each major upgrade of the Swiss-Prot DB.

2 Software and files.

The method is available as an internet server at http://automotif.bioinfo.pl/. The whole database of annotated segments (positive instances), parent proteins (with detailed biological information included) is implemented as https://mysql.bioinfo.pl/ as the mySQL database with phpMyAdmin web interface.

3 Tables.

Table I. The prediction efficiency for various types of phosphorylation. The results are obtained using SVM learning with polynomial kernel ((s a*b+c)^d). Data is collected from Swiss-Prot DB annotation tables (without BY SIMILARITY, PREDICTED, PROBABLE, POTENTIAL or PARTIAL annotations). The first column in the table gives the number of positives and negatives for each type of activation process. The first row describes the dimension for each embedding method.

Results are collected for 8 different methods for preparing SVM input vectors representing each segment (with length 9 or 13 amino acids). The first one is the simplest BIN method uses binary representation of amino acids in the SVM input vector. The BIN+LOOKUP includes additional vector of 9 or 13 values (depending on the size of the segments) of frequency ratios between positives and negatives for these particular amino acid found in the input segment and the position in each predicted segment. The SPARSE method puts instead of 1 the value of frequency ratio between positives and negatives for these particular amino acid found in input segment and the position in each predicted segment. The SPARSE+LOOKUP includes also the frequency ratios for

[1] BioInfoBank Institute, Limanowskiego 24A/16, 60-744 Poznan, Poland, Tel: +48-61-8653520, Fax: +48-61-8643350, E-mail: darman@bioinfo.pl

segments. The LOOKUP vector uses only frequency ratios for amino acids found in a query segment. The BLOSUM+LOOKUP method prints also values for various types of amino acids rescaling them by BLOSUM62 coefficients of the similarity between each type of amino acid and particular type of amino acid found in a query protein. The SUM_PROF uses only sum over the all frequency ratios (dot product of them), and the BLOSUM+SUM_PROF adds also BLOSUM62 similarity matrix. The last two methods uses the whole frequency information calculated on the both (positives and negatives) datasets with, or without separate LOOKUP information.
The most stable method is profile PROF+LOOKUP, SPARSE+LOOKUP or BLOSUM+LOOKUP methods. Other types of methods have lower efficiency (recall / precision).

Recall / precision	Number of positives/ negatives	BIN	BIN +LOOKUP	SPARSE	SPRASE +LOOKUP	BLOSUM +LOOKUP	LOOKUP	BLOSUM +SUM_ PROF	SUM_ PROF	PROF	PROF +LOOKUP
Dim (9/13 frag)		180 264	189 273	180 264	189 273	189 273	9 13	189 273	9 13	180 264	189 273
PKA (9)	86/14353	11.63%	43.02%	36.05%	37.21%	41.86%	41.86%	39.53%	37.21%	41.86%	41.86%
		76.92%	58.73%	55.36%	74.42%	69.23%	85.71%	80.95%	68.09%	75.00%	76.60%
PKC (9)	56/14368	1.79%	16.07%	14.29%	14.29%	17.86%	0%	0%	0%	17.86%	17.86%
		100%	42.86%	44.44%	40.00%	90.91%	0%	-	-	83.33%	62.50%
CDC2 (9)	41/14375	0%	29.27%	21.95%	24.39%	24.39%	21.95%	0%	0%	9.76%	17.07%
		-	31.58%	23.68%	33.33%	28.57%	69.23%	-	-	20.00%	28.00%
SULF (9)	83/6426	39.76%	39.76%	38.55%	39.76%	46.99%	38.55%	13.25%	7.23%	48.19%	57.83%
		97.06%	75.00%	74.42%	73.33%	76.47%	72.73%	100%	100%	86.96%	78.69%
ABL (13)	4/10846	0%	100%	100%	100%	100%	100%	75.00%	75.00%	50.00%	75.00%
		-	100%	100%	100%	100%	100%	100%	100%	100%	100%
CK2 (9)	62/11746	0%	17.74%	19.35%	20.97%	12.90%	14.52%	0%	0%	11.29%	12.90%
		-	47.83%	44.44%	39.39%	50.00%	100%	-	-	53.85%	53.33%
CK (9)	85/11739	0%	10.59%	11.76%	12.94%	8.24%	5.88%	0%	0%	9.41%	9.41%
		-	36.00%	35.71%	40.74%	63.64%	71.43%	-	-	57.14%	36.36%
ABLpept (13)	129/1304	0%	27.13%	6.98%	30.23%	11.63%	34.88%	0%	0%	3.10%	8.53%
		0%	61.40%	64.29%	63.93%	71.43%	77.59%	-	-	57.14%	61,11%

Table 1: The prediction efficiency for predicting phosphorylation by various types of kinases.

4 References.

[1] Bairoch A, Apweiler R. 1999. The Swiss-Prot protein sequence data bank and its supplement TrEMBL in 1999. Nucl Acids Res. 27, pp. 49-54.

[2] Joachims, T. (2000). Estimating the Generalization Performance of a SVM Efficiently. Proceedings of the International Conference on Machine Learning, Morgan Kaufman.

[3] Vapnik,V.N. (1998). Statistical Learning Theory. Wiley, New York.

[4] Zavaljevski, N, Stevens, F.J., Reifman, J. (2002). Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. Bioinformatics. vol. 18(5), pp. 689-696.