CELLULAR & MOLECULAR BIOLOGY LETTERS
Volume 10, (2005) pp 73 – 89
http://www.cmbl.org.pl

Received 21 June 2004
Accepted 19 November 2004

A SUPPORT VECTOR MACHINE APPROACH TO THE IDENTIFICATION OF PHOSPHORYLATION SITES

DARIUSZ PLEWCZYŃSKI[1,2]*, ADRIAN TKACZ[3], ADAM GODZIK[4,5] and LESZEK RYCHLEWSKI[1]
[1]BioInfoBank Institute, ul. Limanowskiego 24A/16, 60-744 Poznań, Poland, [2]Interdisciplinary Centre for Mathematical and Computational Modeling, University of Warsaw, ul. Pawińskiego 5a, 02-106 Warsaw, Poland, [3]Bioinformatics Unit, Department of Physics, Adam Mickiewicz University, ul. Umultowska 85, 61-614 Poznań, Poland, [4]The Burnham Institute, La Jolla, CA, USA, [5]Bioinformatics Core JCSG, University of California San Diego, La Jolla, CA, USA

Abstract: We describe a bioinformatics tool that can be used to predict the position of phosphorylation sites in proteins based only on sequence information. The method uses the support vector machine (SVM) statistical learning theory. The statistical models for phosphorylation by various types of kinases are built using a dataset of short (9-amino acid long) sequence fragments. The sequence segments are dissected around post-translationally modified sites of proteins that are on the current release of the Swiss-Prot database, and that were experimentally confirmed to be phosphorylated by any kinase. We represent them as vectors in a multidimensional abstract space of short sequence fragments. The prediction method is as follows. First, a given query protein sequence is dissected into overlapping short segments. All the fragments are then projected into the multidimensional space of sequence fragments via a collection of different representations. Those points are classified with pre-built statistical models (the SVM method with linear, polynomial and radial kernel functions) either as phosphorylated or inactive ones. The resulting list of plausible sites for phosphorylation by various types of kinases in the query protein is returned to the user. The efficiency of the method for each type of phosphorylation is estimated using leave-one-out tests and presented here. The sensitivities of the models can reach over 70%, depending on the type of kinase. The additional information from profile representations of short sequence fragments helps in gaining a higher degree of accuracy in some

*Corresponding author; tel: +48-61-8653520, fax: +48-61-8643350, e-mail: darman@bioinfo.pl

phosphorylation types. The further development of an automatic phosphorylation site annotation predictor based on our algorithm should yield a significant improvement when using statistical algorithms in order to quantify the results.

Key Words: Kinase Substrate Prediction, Profile-Profile Sequence Similarity (PSI-BLAST, FFAS), Library Of Protein Motifs (Local Structure Segments Database), Database of Phosphorylation Sites, Swiss-Prot Database, Support Vector Machine

INTRODUCTION

The rapid increase in the breadth of available genomic information has led to a need for new automatic techniques to predict protein function. Bioinformatic techniques identify signaling domains within protein sequences, but only limited success is achieved in predicting the positions of phosphorylation sites. Here, we describe a machine learning algorithm that, within a support vector machine framework, classifies biological functional information acquired from the Swiss-Prot database. The classification models can be used to predict new, unknown phosphorylation sites in proteins.

Phosphorylation processes are crucial for living processes and whole metabolism in cells; they are an important mechanism for controlling intracellular processes. Many protein kinases are known, but the identification of their potential biological targets is still ongoing research. The appropriate substrate specificity of protein kinases ensures the correct transmission of signals in cells. The primary sequence information in substrate proteins is crucial for determining protein kinase specificity, but we lack an efficient method for identifying these sequences.

There are a number of tools which were designed to predict the functional annotation of proteins based on sequence information (such as ScanSite [1] or the ELM server [2]). Our system does not use the motif methodology, instead focusing on direct residue representation and frequencies. The regular expression search is more permissive (i.e. gives a large number of false positives) because of the inherent difficulties in describing the phosphorylation site using a simple letter pattern. Our tool is more conservative, so it can be used as an additional filter to remove some of the false positives. The databases available on the Internet [3, 4] provide information on a large collection of phosphorylable residues in proteins, and data about peptide phosphorylation by protein kinases. This data suggests that sequence specificity determinants are not that strict; nevertheless, they are located within a 9-amino acid segment around a phosphorylation site. Our tool uses the same 9-amino acid sequence window around the phosphorylation site to predict annotation. Most of the phosphorylation sites are located on the surfaces of the target proteins. The neural network-based tools predicts the position of phosphorylation sites in independent sequences with a sensitivity of over 70% [5]. Our service is a

natural complement to those tools. It uses different machine learning methodology, which makes it possible to build an independent list of plausible phosphorylation sites for a given query protein. Both lists can be compared and used to build a consensus result, which allows for a higher quality of predictions. Our next paper will be a report on a detailed comparison of the SVM results on the PhosphoBase database using this tool, and on the details of the consensus algorithm with sensitivity/specificity scores.

Protein phosphorylation affects most cellular processes. The main question in the area is how specificity in substrate recognition is achieved. Herein, we present a machine learning approach, which predicts the position of phosphorylation sites with a sensitivity of over 70%, depending on the type of kinase. In our approach, the method for predicting a plausible phosphorylation site's position is based on the classification of known experimental instances. It uses only sequence information as the input, because in most cases, only the sequence of a potential target protein is known. The Swiss-Prot database [6] contains a large number of annotated phosphorylation sites. That is the main reason for our developing and testing this method of prediction of phosphorylation site positions using sequence information from this database. For our initial tests, we selected proteins that are phosphorylated by PKA, PKC, CK, CK2 and CDC2 kinases. Those types of phosphorylation processes have the largest number of known experimental instances, and thus could provide sufficient statistical data. We neglect here all those residues with phosphorylation annotated: “by similarity”, “hypothetical” or “predicted”.

In the “Materials & Methods” section, we provide detailed information about the preparation of the database of short protein fragments and describe the automatic annotation algorithm for the prediction of post-translational modification sites in proteins. In the “Results & Discussion” section, we present the benchmarks used for the statistical analysis of local structure prediction quality. We describe the local sequence composition of segments around phosphorylated sites together with predicted structural information. We also include the analysis of the background sequence and structural preferences of LSSs not annotated in the Swiss-Prot database. Finally, we present conclusions and discuss possible future developments.

MATERIALS AND METHODS

The dataset of proteins known to be phosphorylated by various kinases

As the training dataset for our automatic prediction method, we used proteins that are from the Swiss-Prot database and that have at least one site experimentally verified to be phosphorylated by any kinase. In order to maximize the classification accuracy of models, we neglected all sites annotated “by similarity”, “partial”, “potential”, “probable” or “predicted”. The remaining phosphorylation sites were used to create a positive instance dataset which includes all the sequence segments from the parent proteins dissected within a 9-amino acid window around the phosphorylation site. All the redundant segments

with the same sequence were removed from the database. We used 67 proteins with PKA phosphorylation (86 different segments), 49 proteins with PKC phosphorylation (56 different segments), 18 proteins with CDC2 (41 different segments), 35 proteins with CK2 (62 different segments) and 44 proteins phosphorylated by CK kinase (85 different segments).

Those sequence segments that have the proper central residue according to the type of phosphorylation process, but which are not annotated as functional ones, were used as the negative cases for our method. For example, in the case of PKA and PKC phosphorylation, in order to obtain background preferences for sites with known structures, we extracted 14353 PKA-negative and 14369 PKC-negative sequence segments with the correct central residue (S or T amino acids). Those negative instances were randomly chosen as sequence segments from proteins found in the Swiss-Prot database, and annotated to have at least one site phosphorylatable by any type of kinase.

Both the positive and negative datasets of segments were projected (see "Materials & Methods" section) into one abstract multidimensional space in order to build a detailed sequence model for each type of kinase. Then, the statistical learning theory was used to classify all the cases, and to construct the separation border between the positives and negatives.

Local segment sequence and structure preferences around phosphorylated sites

In our previously published findings [7], we developed a library of local structural segments and a profile-profile matching algorithm that predicts the local structure of proteins from their sequence information. The fragment library prediction method server (FRAGlib, publicly available at http://ffas.ljcrf.edu/Servers/frag.html) allows for prediction of the local structural conformation of sequence segments around phosphorylated sites. This algorithm has also been successfully applied to the characterization of the local structure around phosphorylation sites in proteins [8, 9]. Our results strongly suggest that sequence information is the crucial source of information for the successful predictions of phosphorylation site positions in proteins. It can be supplemented by additional structural context information, predicted using our segment similarity method. Unfortunately, only proteins phosphorylated by PKA and PKC kinases represent the largest number of instances in the Swiss-Prot database, and therefore only they can be used as the benchmark and test dataset for the automatic annotation method. The structural counterpart of the prediction is evaluated using the database of all real and experimentally confirmed structures of parts of the main Cα chain around the phosphorylation sites. The real structures are collected using the PSI-Blast server running on the PDB database (PDB-Blast) (http://www.bioinfo.pl/).

In order to quantify the local sequence and structural preferences around the phosphorylation sites in proteins, we used the collected database of sequence segments. For PKA and PKC phosphorylation, we obtained real structures of proteins around phosphorylation sites with the PDB-Blast server developed by

our group (http://www.bioinfo.pl/). It is a PSI-Blast program that compares the sequence of a query protein with all the sequences from the PDB database within very strict thresholds in order to get the one true structure of a protein, and it makes use of crystallized protein data. We collected models for 56 proteins with PKA phosphorylation sites and 38 with PKC phosphorylation sites. However, we found only 11 structural segments crystallized around sites with both PKA and PKC phosphorylations. Most of the phosphorylation sites are located in unstructured parts of proteins; these are difficult to crystallize, and frequently, those coordinates are missing in the PDB.

To sample the background sequence preferences, we took 17718 sites not annotated as PKA phosphorylated, and 18799 sites not annotated as PKC phosphorylated, with the appropriate central amino acids. In order to obtain the background preferences for the sites with known structures, we also extracted 340 PKA-negative and 141 PKC-negative sites from protein segments with assigned coordinates and correct central residues (S or T). We analysed the sequence and local structure composition of those positive and negative cases. While the sequence composition of both types of instance displays clear differences, much less significant differences could be observed between the local structures of each type [8]. The predicted local structure of both types is in qualitative agreement with the real structures. A comparison with other available structure prediction was also performed [7]. The differences between the results of those methods and our results in the modelling of local structural preferences around phosphorylation sites are within the accuracy of our method.

Local structural preferences

Our test shows [9] that, in the case of PKA and PKC phosphorylation, our method has large recall efficiency. In the test cases, almost all the actual positions of the phosphorylation sites were predicted using our algorithm. There is also a clear and significant difference between the mean prediction score values for true predictions and those for false ones. The proper cut-off value, which depends on the type of phosphorylation process, can provide a better percentage rate of precision, losing only a small subset of annotated sites, i.e. those with lower recall values. However, there are more false predictions than true ones. That is why more refined statistical methods, such as the support vector machine approach to classification and prediction, described in the following section, are needed to further improve the overall benchmark results for our method. This will also help to discriminate between false positives and true ones. The structural part of the prediction score is helpful in predictions, but the main difference for those two types of phosphorylation (by PKA and PKC kinases) is observed for only the sequence part. This is the reason for utilizing the SVM statistical learning theory, using only sequence information and skipping the local structural part of it. The results for other types of kinases (CK, CK2, CDC2 not included here) show that the short protein fragment has a greater structural preference towards kinase than the sequence one does.

However, the statistics for those cases are rather poor, so a strict statement cannot be made.

Local sequence fragments representation

In order to use the Support Vector Machine approach to the classification of various types of phosphorylation processes, we should represent short sequence fragments using abstract multidimensional space. This representation does not change the information content of the database.

There are at least six basic ways to represent the sequence of a short protein segment. The first one is a binary representation (called here BIN), which encodes each position of the segment into a long 20-dimensional vector of 0 and 1. The 1 value is taken if the corresponding type of amino acid is present at a certain position of the segment. The single residue Tyr (T) is represented here as a vector with the coordinates [1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0], assuming that T is marked as the first dimension of the space. For 9-residue long segments, the representation space has a dimension equal to 180.

The second method (BLOSUM) is a simple extension of the previous one. It uses the BLOSUM62 matrix, which evaluates the similarity between amino acids. Therefore, each position of the segment is represented by the 20-dimensional vector of the substitution scores of similarity between the amino acid found in the represented segment at this position and all 20 amino acids. The dimension for this embedding is the same as for the previous binary method.

For each amino acid found at certain position of a segment, the LOOKUP method uses the scalar value describing the normalized sequence preference for it. Normalized preferences are calculated by dividing the frequencies for all the types of amino acid and all their possible positions within a segment for the positives dataset by the frequency for the negatives dataset. The background preferences for negative instances are calculated for proteins with the considered type of kinase. The dimension of this method is only 9 in the case of 9aa segments; one normalized sequence preference for each position of a segment.

The same dimension of the abstract space of embedding is gained for another method, here called SUM_PROF. Instead of the normalized preference for only one amino acid found at a certain position of the query sequence segment, this embedding uses the sum over all the normalized preferences of amino acids, each multiplied by the BLOSUM62 similarity factor between it and the actual amino acid type found in the projected segment.

The profile method (PROF) uses the same normalized preferences, but taken as a single 20-dimensional vector of normalized preferences instead of summing them up. In that case the dimension of the embedding space is equal to 180 (9aa segments).

The last method (SPARSE) is similar to the BIN method, but instead of binary values, it takes real values equal to the normalized preferences for each position of a segment – the ratio of preference for positive instances and the preference of negative ones – and otherwise, the 0 value. Therefore, it has the same dimension as the first two methods (180).

We also tested various combinations of the above methods, such as BIN+LOOKUP and SPARSE+BLOSUM, adding to the first method the additional dimensions from another one, i.e. the Cartesian product of two vector spaces merged into a larger one. This resulting projection has additional information which may help us to achieve higher test accuracy.

The support vector machine approach for the classification of short sequence fragments

Our algorithm utilizes various types of projection of short sequence segments into one abstract multidimensional feature space. The classification of all known instances is done within the support of the vector machine SVM framework. SVM is a statistical learning method with a good performance record, and it is easier to implement than neural networks. This method was proposed as an effective machine learning approach by Vapnik and Cristianini [10, 11]. The theory of SVM in the case of pattern recognition with the discussion of regression and the learning of a ranking function was extensively reported on by Vapnik [10, 12]. The SVM method was successfully applied to various problems including text classification [13, 14], image recognition [15] and medical applications [16, 17]. The SVM approach was also used in bioinformatics [18, 19], especially in the analysis of gene expression data [20], the classification of microarray data [21, 17], the inference of gene functional classification [22-24] and the analysis of proteins [25-27].

Most of those tasks have the property of sparse instance vectors. The SVM approach has the ability to construct predictive models with a large generalization power, even in the case of a large dimensionality of the data when the number of observation available for training is low. SVM always seeks a globally optimized solution and avoids over-fitting, so the large number of features, as in our binary representation of sequence segments, is permitted. Our work is based on the SVMlight implementation code in C language created by Thorsten Joachims [28, 29], and also used widely in the field of bioinformatics [30]. It uses sparse instance vector properties to obtain compact and efficient representation. The efficiency of the constructed prediction models is estimated here using the leave-one-out method.

In order to extract the relevant information from the heterogeneous biological data, we used statistical learning theory in terms of the SVM approach. SVM tries to separate a given set of binary labeled training vectors with an optimal hyperplane. The optimum is reached for the hyperplane that maximizes the separating margin between the two classes of the training vectors with a relatively small number of support vectors.

The output of the training phase for each type of kinase for the phosphorylation site is a classification function (model). It consists of the set of D support vectors, T_j and a_i, which are nonzero, positive real numbers. Those constants are obtained from the optimization procedure, called the quadratic programming problem (QP problem) used to find the maximal margin hyperplane. For any

embedding *T* of the input space of segments *[x]* into the representations space, all models are given in the form of the cost function:

$$f(T[x]) = \sum_{i=1}^{i=D} l_i a_i K(\Omega\{T[x]\}, \Omega\{T_i\}), \quad (1)$$

where $K(T, T_i)$ is the proper kernel function that defines the feature space, Ω is a nonlinear mapping function from the embedding space into the feature space, and l_i are known *a priori* class labels for the support vectors. We used l_i = +1 for positive cases and l_i = -1 for negative cases. The kernel function is a positive defined function reflecting the similarity between a given input sample and the set of support vectors T_i. We built all the models of the phosphorylation sites using three kinds of kernel in SVM learner:

- a linear one, given by the linear inner product in the feature space: $K(T, T_i) = \langle \Omega\{T\}, \Omega\{T_i\} \rangle$,
- a polynomial one, described by the kernel function: $K(T, T_i) = (a\langle \Omega\{T\}, \Omega\{T_i\} \rangle + c)^d$,
- a radial basis kernel: $K(T, T_i) = \exp(-g \langle \Omega\{T\}, \Omega\{T_i\} \rangle \langle \Omega\{T\}, \Omega\{T_i\} \rangle)$.

Those types of kernel are the most standard ones, and have also been extensively studied in the field of bioinformatics [18, 19, 30].

The number of free parameters of the QP problem is equal to the number of all instances in the training dataset. The non-zero parameters a_i describe the strength of this particular *i*-th support vector in the decision function. SVM chooses as support vectors those points that lie closest to the separating hyperplane. The mapping function Ω need not be explicitly defined, because in the kernel function, only the inner product of it is used.

The automatic phosphorylation site position predictor

The sequence-based automatic phosphorylation site position predictor is a complement to our previous work [9]. It uses the knowledge database of short sequence segments phosphorylated by various types of kinases. This database is built from segments with known sequence profiles (from the Swiss-Prot database). The automatic annotating service receives a sequence from a query protein as an input, and predicts the positions of its phosphorylation sites for kinases of a certain type. It uses the SVM classification models constructed as described in the previous section.

The prediction method is as follows. First, we dissect a query protein into overlapping short segments of length 9aa. For each segment x_j, we assign a label using an SVM-constructed model according to the decision function given by:

$$f(T[x]) = \mathrm{sgn}\left(\sum_{i=1}^{i=D} l_i a_i K(\Omega\{T[x]\}, \Omega\{T_i\}) + b \right), \quad (2)$$

where $K(T_j, T_i)$ is the proper kernel function that defines a similarity in the feature space, T is the embedding, i.e. mapping, function from the input space of the segments to the representation space, and b is a bias value (score cut-off value). $\{a_i\}$ are nonzero, positive real numbers that define the maximal margin hyperplane, with T_j as the set of D support vectors of the model. All those constants are separately computed during the optimization phase on a training set for each type of phosphorylation process by a certain kinase.

Using the cost function (see Eq. 1), we described the reliability of the predictions. As the output of our method, we took only those sites which have a score (the value of the cost function from Eq. 1) larger than b. This means that points representing sequence segments centered on those sites lie in the region classified as positive by the SVM model's hyperplane with a given b as the margin value.

For the purposes of the Web server, we use polynomial kernels. As the output, all the predictions with the appropriate sign for the decision function are listed (Eq. 2). Our method is a simple one-vote wins approach, where we take the best model for each predicted segment. The S_k score for each segment is given by the cost function (Eq. 1) for the k-th method:

$$S_k\lfloor x_j \rfloor = f_k\left(T\lfloor x_j \rfloor\right), \quad (3)$$

The overall reliability of predictions is also described by the numeric values of the precision P_k and the recall value R_k (see below, Eq. 4) for the k-th method.

RESULTS AND DISCUSSION

The SVM approach provides the fast optimization algorithm with working set selection based on the steepest feasible descent. It uses a "shrinking" heuristic caching of kernel evaluations and folding in the linear case. The overall efficiency of the constructed model is described by solving classification, regression, and ranking problems. XiAlpha estimates are computed at essentially no computational expense, but they are conservatively biased [29, 31]. Almost unbiased estimates are provided by leave-one-out testing. The results of most leave-one-outs (often more than 99%) are predetermined and need not be computed; this is exploited by the SVMlight code [29]. The generalization performance efficiency is described here by the second efficient estimation method in terms of the error rate, the precision and the recall. The algorithm includes the learning ranking functions [32], which learn a function from preference examples, so that it orders a new set of objects as accurately as possible. It also handles several hundred-thousands of training examples and many thousands of support vectors, which is crucial in the case of large datasets of positive and negative instances. It supports standard kernel functions like linear, polynomial or radial ones.

The performance of our tool is described here with three measures of accuracy: classification error E, recall R and precision P:

$$E = \frac{fp + fn}{tp + fp + tn + fn} * 100\%, \quad (4a)$$

$$R = \frac{tp}{tp + fn} * 100\%, \quad (4b)$$

$$P = \frac{tp}{tp + fp} * 100\%, \quad (4c)$$

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, while the recall R measures the percentage of correct predictions, i.e. the probability of obtaining a correct prediction, and the precision P gives the percentage of observed positives

Tab. 1. SVM learning results with a linear kernel.

Recall / Precision	Number of positives/ negatives	BIN	BIN +LOOKUP	SPARSE	SPARSE +LOOKUP	BLOSUM +LOOKUP	LOOKUP	BLOSUM +SUM_PROF	SUM_PROF	PROF	PROF +LOOKUP
Dim (9aa)		180	189	180	189	189	9	189	9	180	189
PKA (9)	86/14353	0%	36.05%	13.95%	37.21%	17.44%	38.37%	0%	0%	0%	0%
		-	86.11%	75.00%	91.43%	88.24%	80.49%	-	-	-	-
PKC (9)	56/14368	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		-	-	-	-	-	-	-	-	-	-
CDC2 (9)	41/14375	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		-	-	-	-	-	-	-	-	-	-
CK2 (9)	62/11746	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		-	-	-	-	-	-	-	-	-	-
CK (9)	85/11739	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		-	-	-	-	-	-	-	-	-	-
PHOSPH (9)	1101/10000	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		-	-	-	-	-	-	-	-	-	-

The prediction efficiency for predictions of the sites of various types of phosphorylation processes obtained using SVM learning with a linear kernel. Here, we present two error estimators for all the types of projection and post-translational modification. The first one is the recall R, which measures the percentage of correct predictions (the probability of correct prediction). The second one is the precision P, which gives the percentage of observed positives that are correctly predicted (the measure of the reliability of positive instances prediction). Recall equals 0% and precision is not well defined (marked by "-") if the SVM training phase cannot be finished. For some types of phosphorylation sites or types of projections, the training procedure fails. In such cases, we chose recall equal to 0% (no positives found), and precision is marked by "-". The most stable methods are the simple LOOKUP or mixed SPARSE+LOOKUP approaches. Other types of method have some advantages in particular types of phosphorylation, but they have a lower efficiency (recall/precision).

Tab. 2. SVM learning results with a polynomial kernel.

Recall / Precision	Number of positives/ negatives	BIN	BIN+ LOOK UP	SPARSE	SPRASE +LOOK UP	BLO SUM+ LOOK UP	LOOK UP	BLO SUM+ SUM_ PROF	SUM_ PROF	PROF	PROF+ LOOK UP
Dim (9aa)		180	189	180	189	189	9	189	9	180	189
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%
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%
PHOSPH (9)	1101/10000	26.88%	25.07%	3.36%	19.71%	29.43%	5.99%	0%	0%	33.79%	34.42%
		77.49%	69.17%	68.52%	68.45%	73.64%	75.86%	-	-	71.95%	72.88%

The results for the best kernel type of SVM method for all the considered types of phosphorylation sites. A detailed description of the two error estimators is given in the legend to Tab. 1. The results are obtained using SVM learning with a polynomial kernel ((s a*b+c)^d). We collected results for 10 different embeddings. 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. The most stable methods are the profile PROF+LOOKUP, SPARSE+LOOKUP or BLOSUM+LOOKUP methods. Other types of methods have a lower efficiency (recall/precision). For some types of phosphorylation site, or types of projection, the training procedure fails. In such cases, we chose a recall value equal to 0% (no positives found), and precision is marked by "-".

that are correctly predicted, i.e. the measure of the reliability of positive instance prediction. Those measures of accuracy, as mentioned before, can be computed using conservative but easy to compute Xi-Alpha estimates and using the more precise but computationally intensive leave-one-out procedure. The leave-one-out test removes one sample from the training data, constructs the model on the basis of the remaining training dataset, and then tests the prediction of the model on the removed sample. The resulting error estimators are averaged for all such models for all positive and all negative instances.

The results of predictions for phosphorylation sites by various kinases for different projection methods and kernel functions are presented in Tabs. 1, 2 and 3. We collected the results for the 10 different methods described in the previous section for preparing SVM input vectors representing sequence fragments. The first type of kernel function is linear, and is not efficient in the case of more complicated sequence signatures of phosphorylation sites. In some cases (PKA

Tab. 3. SVM learning results with a radial kernel.

Recall / Precision	Number of positives/ negatives	BIN	BIN+ LOOK UP	SPARSE	SPRASE +LOOK UP	BLO SUM+ LOOK UP	LOOK UP	BLO SUM+ SUM_ PROF	SUM_ PROF	PROF	PROF+ LOOK UP
Dim (9aa)		180	189	180	189	189	9	189	9	180	189
PKA (9)	86/14353	0%	0%	0%	0%	0%	11.63%	0%	0%	0%	0%
		-	-	-	-	-	76.92%	-	-	-	-
PKC (9)	56/14368	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		-	-	-	-	-	-	-	-	-	-
CDC2 (9)	41/14375	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		-	-	-	-	-	-	-	-	-	-
CK2 (9)	62/11746	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		-	-	-	-	-	-	-	-	-	-
CK (9)	85/11739	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		-	-	-	-	-	-	-	-	-	-
PHOSPH (9)	1101/10000	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
		-	-	-	-	-	-	-	-	-	-

The results of various methods for the predictions of phosphorylation sites using SVM learning with a radial kernel. Two estimators describing the efficiency of the method are described in Tab. 1. In most cases, the SVM learner fails to construct a model with a radial kernel. For those cases, we chose a recall value equal to 0% (no positives found), and precision is marked by "-". When the number of positives is large, the LOOKUP method is the best one, reaching the efficiency of the models with polynomial kernels.

Tab. 4. The classification error for three generic embeddings.

Phosphorylation kinase	#positives	BIN	LOOKUP	PROF
PKA phosphorylation	86	0.55%	0.39%	0.43%
PKC phosphorylation	56	0.38%	0.40%	0.33%
CDC2 phosphorylation	41	0.28%	0.25%	0.37%
CK2 phosphorylation	62	0.53%	0.45%	0.52%
CK phosphorylation	85	0.72%	0.69%	0.70%
Phosphorylation (all types)	1101	8.03%	9.51%	7.87%

The classification error for three generic embeddings (binary, lookup and profile) for each type of phosphorylation process. The first column presents the number of positive instances found in Swiss-Prot DB using annotation information (without BY SIMILARITY, PREDICTED, PROBABLE, POTENTIAL or PARTIAL annotations). The classification error E provides an overall error measure. This measure of accuracy is computed using the leave-one-out procedure, which removes from the training data one sample, constructs the model on the basis of remaining training dataset and then tests the prediction of the model on the removed sample. The resulting error estimator is averaged for all such models (for all positive and all negative instances).

phosphorylation with SPARSE+LOOKUP embedding), the models reach the efficiency of the polynomial kernel, which is the best one in all the tested types of phosphorylation. In the case of the polynomial kernel, the most stable methods are the PROF+LOOKUP method and BLOSUM+LOOKUP approach. Both yield excellent results for all the types of phosphorylation functional motif. Other projection methods have some advantages for some particular types of kinases, but they have a lower overall efficiency (recall/precision). In the case of a radial basis kernel, SVM frequently fails to build the model, other than for LOOKUP embedding.

The overall predictive power of SVM models for phosphorylation by various types of kinases is illustrated on Tab. 4. In our tests, we used a large number of negatives for phosphorylation by various types of kinases, which allowed us to approximate the comparison in terms of the calculated precision and recall values between the various methods. The numbers of support vectors for those cases are large, as explained by the large dimensionality of the embedding space and the complicated shape of the separation hyperplane between the positive and negative instances. The number of support vectors can be lowered when we choose lower dimensional initial encoding of the amino acids into the physicochemical properties, such as hydrophobicity, hydrophilicity, polarity, volume, surface area, bulkiness or refractivity. In that case, for each position of a projected segment, instead of 20 dimensions for each type of amino acid, SVM will use only a few variables representing those properties [33].

The potential functional motifs are sometimes repeated with different scores when predicted by various methods. Each method predicts a different set of peptides as the phosphorylation functional motifs. Our automatic predictor uses an identity search or SVM scan. The user should analyze sequences using both methods in order to investigate the wider set of possibilites. The higher the output score, the higher the confidence of the predictions. This means that the potential segments are more similar to one or more of the functional motifs stored in the database used in the training of SVM methods.

CONCLUSIONS

Our approach guarantees the conservative description of the available biological data. The phosphorylation site analysis by support vector machine (SVM) allows for quick and accurate prediction of phosphorylation site position in new sequences. The algorithm can by applied independently from the Web interface via a pipe-line, so massive, large-scale genome analysis is also possible. The main problem we faced in the case of some phosphorylation site types is the insufficient number of experimentally verified instances in Swiss-Prot. In such cases, to make the set of training segments larger, PSI-BLAST can be used [34, 35]. It allows for building the larger list of positives using the sequence similarity between known instances and the new ones. One can also use more refine sequence similarity in terms of profiles [36], or segments from multiple alignments (by BLAST) instead of single sequences from a parent proteins. All

those tools help in preparing large enough set of positives for a phosphorylation by certain type of kinase. In order to improve the quality of those added positives a structure disorder tool like GlobPlot [37] can be used to filter them. Then on such prepared set of positives the SVM learning procedure can be applied giving the classification model.

Our service is natural complement of NetPhos tool which uses Neural Networks methodology. For a given query protein one can predict two sets of predicted phosphorylation sites. Both sets can be compared and then used to build a consensus result. The consensus approach allows for higher quality of predictions. The SVM results on PhosphoBase database, details of the consensus algorithm with sensitivity/specificity scores will be presented in our next paper.

We test various combinations of generic methods like BIN+LOOKUP, SPARSE+BLOSUM etc. as a Cartesian product of two vector spaces merged into larger one. This additional information sometimes helps in gaining the higher accuracy in some phosphorylation types, but in general is not providing the higher efficiency of predictions. The further development of automatic phosphorylation sites annotation predictor based on our algorithm should get a significant improvement when using statistical algorithms in order to quantify the results. In our next paper we will present details of the internet web server allowing for remote access to all models of different types of phosphorylation sites. In the future our tool will be used also for prediction of other types of single amino acids post-translational modification processes, cell signaling networks within proteomes, and will aid in the identification of drug targets for the treatment of human diseases.

Acknowledgement. This study was supported by a USA grant ("SPAM" GM63208), and the BioSapiens (LHSG-CT-2003-503265), the ELM (QLRT-CT2000-00127), GeneFun (LSHG-CT-2004-503567) projects within 5FP and 6FP EC program.

REFERENCES

1. Obenauer, J.C., Cantley, L.C. and Yaffe, M.B. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. **Nucleic Acids Res.** 31 (2003) 3635-3641.
2. Puntervoll, P., Linding, R., Gemünd, C., Chabanis-Davidson, S., Mattingsdal, M., Cameron, S., Martin, D. M. A., Ausiello, G., Brannetti, B., Costantini, A., Ferrè, 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., Küster, B., Helmer-Citterich, M., Hunter, W. N., Aasland, R. and Gibson, T.J. ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. **Nucleic Acids Res.** 31 (2003) 3625-3630.
3. Kreegipuu, A., Blom, N., Brunak, S. and Jarv, J. Statistical analysis of protein kinase specificity determinants. **FEBS Lett.** 430 (1998) 45-50.

4. Kreegipuu, A., Blom, N. and Brunak, S. PhosphoBase, a database of phosphorylation sites: release 2.0. **Nucleic Acids Res.** 27 (1999) 237-239.
5. Blom, N., Gammeltoft, S. and Brunak, S. Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites. **J. Mol. Biol.** 294 (1999) 1351-1362.
6. Bairoch, A. and Apweiler, R. The Swiss-Prot protein sequence data bank and its supplement TrEMBL in 1999. **Nucleic Acids Res.** 27 (1999) 49-54.
7. Plewczyński, D., Rychlewski, L., Ye, Y., Jaroszewski, L. and Godzik, A. Integrated web service for improving alignment quality based on segments comparison. **BMC Bioinformatics** 5 (2004) 98-105.
8. Plewczyński, D. and Rychlewski, L. Ab Initio server prototype for prediction of phosphorylation sites in proteins. **Computational Methods in Science and Technology** 9 (2003) 93-100.
9. Plewczyński, D., Jaroszewski, L., Godzik, A., Kloczkowski, A. and Rychlewski, L. Molecular modelling of phosphorylation sites in proteins using database of local structure segments. **J. Mol. Model.** (2004), in press.
10. Vapnik, V.N. Statistical learning theory. (1998) Wiley, New York.
11. Cristianini, N. and Shawe-Taylor, J. Support vector machines. (2000) Cambridge, UK.
12. Vapnik, V.N. The nature of statistical learning theory. (1995) Springer.
13. Joachims, T. Text categorization with support vector machines: learning with many relevant features. **Proceedings of the European Conference on Machine Learning** (1998) Springer.
14. Joachims, T. Transductive inference for text classification using support vector machines. **International Conference on Machine Learning** (1999).
15. Vojtech, F. and Vaclay, H. An iterative algorithm learning the maximal margin classifier. **Pattern Recognition.** 36 (2003) 1985-1996.
16. Valentini, G. Gene expression data analysis of human lymphoma using support vector machines and output coding ensembles. **Artif. Intell. Med.** 26 (2002) 281-304.
17. Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. Gene selection for cancer classification using support vector machines. **Mach. Learn.** 46 (2002) 389-422.
18. Kim, H. and Park, H. Protein secondary structure prediction by support vector machines and position-specific scoring matrices. **Protein Engin.** 16 (2003) 553-560.
19. Minakuchi, Y., Satou, K. and Konagaya, A. Prediction of protein-protein interaction sites using supprot vector machnes. **Proceedings of the international conference on mathematics and engineering techniques in medicine and biological sciences.** (2003) 22-28.
20. Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M. and Haussler, D. Knowledge-based analysis of microarray gene expression data by using support vector machines. **Proc. Natl Acad. Sci. USA** 97 (2000) 262-267.

21. Furey, T.S., Cristianini, N, Duffy, N., Bednarski, D.W., Schummer, M. and Huassler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. **Bioinformatics** 16 (2000) 906-914.
22. Pavlidis, P., Weston, J., Cai, J. and Grundy, W.N. Gene functional classification from heterogeneous data. **Proceedings of the 5th International Conference on Computational Molecular Biology** (2001) 242-248.
23. Krishnan, V.G. and Westhead, D.R. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. **Bioinformatics** 19 (2003) 2199-2209.
24. Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lengauer, T. and Muller, K.R. Engineering supprot vector machine kernels that recognize translation initiation sites. **Bioinformatics** 16 (2000) 815-824.
25. Jaakkola, T., Diekhans, M. and Haussler, D. A discriminative framework for detecting remote protein homologies. **J. Comput. Biol.** 17 (2000) 95-114.
26. Hua, S. and Su, Z. A novel method of protein secondary structure prediction with segment overlap measure: support vector machine approach. **J. Mol. Biol.** 308 (2001) 397-407.
27. Ding, C.H.Q. and Dubchak, I. Multi-class protein fold recognition using support vector machines and neural networks. **Bioinformatics** 17 (2001) 349-358.
28. Joachims, T. Making large scale SVM learning practical. in: **Advances in Kernel Methods - Support Vector Learning (**Scholkopf, B., Burges, C. and Smola, A. Eds), MIT Press, Cambridge, USA, 1999.
29. Joachims, T. Learning to classify text using support vector machines. **Dissertation** (2002) Kluwer, Germany.
30. Zavaljevski, N., Stevens, F.J. and Reifman, J. (2002). Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. **Bioinformatics** 18 (2002) 689-696.
31. Joachims, T. Estimating the generalization performance of a SVM efficiently. **Proceedings of the International Conference on Machine Learning** (2000) Morgan Kaufman.
32. Joachims, T. Optimizing search engines using clickthrough data. **Proceedings of the ACM Conference on Knowledge Discovery and Data Mining** (2002) ACM.
33. Lohman, R., Schneider, G., Nehrens, D. and Wrede, P. A neural network model for the prediction of membrane-spanning amino acid sequences. **Protein Sci.** 3 (1994) 1597-1601.
34. Altschul, S.F., Madden, T., Shaffer, A., Zhang, J. and Zhang, Z. Gapped blast and psi-blast: a new generation of protein database search programs. **Nucleic Acids Res.** 25 (1997) 3389-3402.
35. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. Basic local alignment search tool. **J. Mol. Biol.** 215 (1990) 403-410.

36. Rychlewski, L., Jaroszewski, L., Li, W. and Godzik, A. Comparison of sequence profiles. Strategies for structural predictions using sequence information. **Protein Sci.** 9 (2000) 232-241.
37. Linding R., Russell, R.B., Victor Neduva, V. and Gibson, T.J. GlobPlot: exploring protein sequences for globularity and disorder. **Nucleic Acids Res.** 31 (2003) 3701-3708.