CMST 9(1) 93-100 (2003) DOI:10.12921/cmst.2003.09.01.93-100

COMPUTATIONAL METHODS IN SCIENCE AND TECHNOLOGY 9(1-2) 93-100 (2003/2004)

Ab Initio SERVER PROTOTYPE FOR PREDICTION OF PHOSPHORYLATION SITES IN PROTEINS*

DARIUSZ PLEWCZYNSKI AND LESZEK RYCHLEWSKI

BiolnfoBank Institute, Limanowskiego 24A/16, 60-744 Poznan, Poland
darman@bioinfo.pl

(Rec. 15 November 2003)

Abstract: We describe an *ab initio* server prototype for prediction of phosphorylation sites. A list ofpossible active sites for a given query protein is build using query protein sequence and the database of proteins annotated for a certain type of activation process by Swiss-Prot DB. All short segments ofa query protein sequence centered around plausible active sites are compared with experimental profiles. Those profiles describe both sequence and structure preferences for each type of active site. Prediction of local conformation ofa query protein chain around examined site is done with the specially prepared library of short local structural segments (LSSs). The short sequence fragments from a query protein are matched with segments in the library using profile with profile alignment. Predicted local structure of a chain near active site qualitatively agrees with experimental data fetched from PDB database. We estimate in this paper the level of improvement over purely sequence based methods gained by incorporating predicted structural information into the local description of phosphorylation sites.

1. INTRODUCTION

Protein phosphorylation is a very important mechanism for signal transduction and control of intracellular processes. Yet only the fraction of protein kinases and phosphorylation sites are characterized in great detail. This creates the need for methods predicting plausible active sites in proteins. Prediction methods should use only sequence information as an input, because in most cases only the sequence of a potential target protein is known. In order to incorporate also the structural description we introduce here simple method for prediction of local structure ofproteins based on sequence.

The reason for utilizing the structural information into a description ofproteins is straightforward. Structural comparison is able to detect approximately twice as many functional relationships between proteins as sequence comparison at the same error rate [1]. Sets ofpossible structural motifs can be used to describe similarity between very distant, non-homologous proteins [2, 3]. This approach can be taken to the extreme by comparing proteins represented as a string of only three Q3 symbols (*α* for alpha-helix, *β* for beta-strand and *c* for coil) of predicted secondary structure [4],

The local (secondary) structure predictions based on segment similarity include two different approaches. The first one is whole branch of secondary structure prediction methods based on the nearest-neighbor algorithms [5, 6] The second one is a new method of local structure

* Dedicated to memory of Professor Jacek Rychlewski

predictions developed by Baker et a1, based on a library of sequence-structure motifs called I-sites [7], As a result both methods obtain a set of possible Local Structure Segments (LSSs) for a protein. Due to the essential uncertainty of local structure formation in the absence of tertiary interactions such diversity of local structures is unavoidable. Similar approach is used here to support by structural information the prediction of phosphorylation sites in proteins.

In order to introduce and test automatic method for prediction of phosphorylation sites we use information contained in the database of phosphorylation sites from Swiss-Prot DB. For initial tests we have selected proteins only with PKA and PKC phosphorylation. Those two types of phosphorylation have the largest number of known experimental instances providing sufficient statistics. Each site was a target for local structure prediction. Our predictions were compared with experimental structural data from the PDB database. Additional test were conducted for 13 amino acid long peptides known from peptide array experiments [8] to be phosphorylated by ABL kinase.

First we present details of our method for predicting local structure of proteins which is used here in extending by structural information local sequence description of protein chain segments around phosphorylated sites. The next section is devoted to the analysis of background sequence and structural preferences (based on the whole family of LSSs not annotated in Swiss-Prot database as active sites). In the last section we present our automatic annotation method for searching phosphorylation sites in a given query proteins.

2. THE METHOD FOR LOCAL STRUCTURE PREDICTION

In our approach we predict local structure of a protein based only on its sequence and the library of short protein motifs. The crucial idea behind it is to use the most general definition of blocks forming the global protein structure on the basis of local structural regularity. Such collection of fragments is constructed here from representative set of proteins from the ASTRAL database [9, 10]. To describe local structure around each amino acid we adopt the idea of symbolized local structure representation SLSR consisting of 11 symbols {HGEeBdbLlxc}, each constrained to a certain backbone dihedral regions [7]. Each LSS from the library can be described as a short string of those local-structure symbols. The library itself is large and very redundant collection (38220 items) of such short local structure segments (7 residues long) for which the local structural codes are the same all along their chain, except offset at the beginning, and at the end of a segment. We store this collection of fragments from ASTRAL database with their sequence, symbolized local structure representation SLSR codes and dissected parts of homology profile for their parent proteins.

Our method predicts the local structure of a protein based only on sequence information. First we create the sequence profile for a query protein using PSI-Blast [11, 12]. Then for all overlapping segments of length 7aa we compare those short segments of the profile (dissected from the profile of whole query protein) with profiles of fragments from the library. For each pair of compared short segments (one from a query protein, and the second one from the fragments database) we calculate the profile homology score [13, 14]. We use here simplified rescaled scalar product of two vectors representing the sequence profiles of two segments.

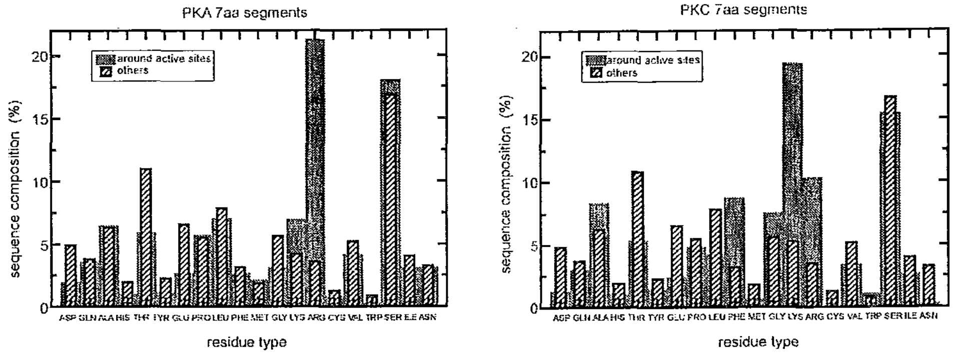
We keep for later use only the average structural preference calculated using certain number of predicted LSS with the highest profile-profile similarity score. We abandon all differences between predicted local structure segments. Our tests show that averaging over larger number of predicted LSSs is not improving the quality of predictions of local structural preference around phosphorylated sites.

3. LOCAL SEQUENCE AND STRUCTURE PREFERENCES AROUND PHOSPHORYLATED SITES

In order to maximize the local description accuracy based on sequence and structure of active sites we took the list of proteins from Swiss-Prot database with at least one experimentally verified phosphorylation site. We neglect all phosphorylation sites annotated "by similarity", "hypothetical" or "predicted". For detailed analysis we used 67 proteins with PKA phosphorylation (98 sites) and 49 proteins with PKC phosphorylation (73 sites). Apart from predicting the local structure of each protein using our method we took experimental structures of the part of the main chain around the phosphorylated residue. We collected models for 56 proteins with PKA and 38 with PKC phosphorylation sites. However we found only 11 structural segments around the sites for both the PKA and the PKC, mainly because many phosphorylation sites are found in unstructured parts of a proteins, which are difficult to crystallize.

To sample the background preferences we took 17718 sites not annotated as PKA phosphorylated and 18799 sites not annotated as PKC phosphorylated with appropriate central amino acids. In order to obtain background preferences for sites with known structure we extracted 340 PKA-negative and 141 PKC-negative sites from protein segments with assigned coordinates and correct central residue (5 or *T* amino acid). We analyzed the sequence and local structure composition of those positive and negative instances. Preferences are displayed using four structural classes [4]: helixes (structural codes H and G describing strong helical preference), extended (B and E), loops (L) and the last class for all others structural codes. For each type of phosphorylation the structural preferences of those four classes were calculated for true instances and segments not annotated in Swiss-Prot database. While the sequence composition of both types of instances shows clear differences (see Fig. 1), smaller differences could be observed between local structures (predicted and experimental, see Fig. 2).

Additional tests were conducted on a set of 1433 peptides with known susceptibility to phosphorylation by the ABL kinase determined by experiments [8]. The experimental data were collected using microarrays, which enables the direct measurement of the individual phosphorylation state of each member in a large peptide library, even for non-substrates.



Fig, 1. Sequence composition for proteins' segments centered around PKA and PKC phosphorylation sites (as annotated in Swiss-Prot database) together with the background preference. For phosphorylation instances we use 67 proteins with 98 sequence (PKA), and 49 proteins with 73 active sites (PKC). The background preferences was build using 17718 (PKA) and 18799 (PKC) sequence segments centered do the proper amino acids *(S* or *T)*

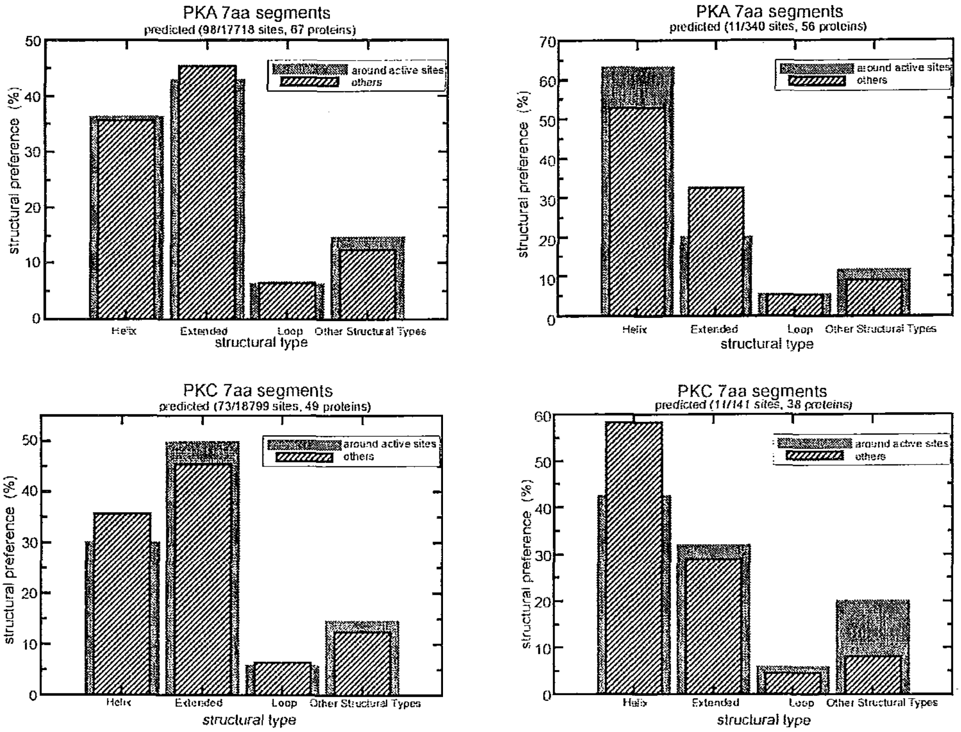


Fig. 2. Structural preference for proteins' segments of length 7aa centered around PKA and PKC phosphorylation sites (left - predicted, and right - experimentally verified). For predicting we use 67 proteins with 98 sequence segments (PKA), and 49 proteins with 73 active sites (PKC). The real structure is known for 56 proteins with 11 3D structural segments around active sites (PKA) and 38 proteins with also 11 active sites for PKC. The background preferences include 17718 (PKA) and 18799 (PKC) sequence segments for predictions, and 340 (PKA) and 141 (PKC) structural segments for real data

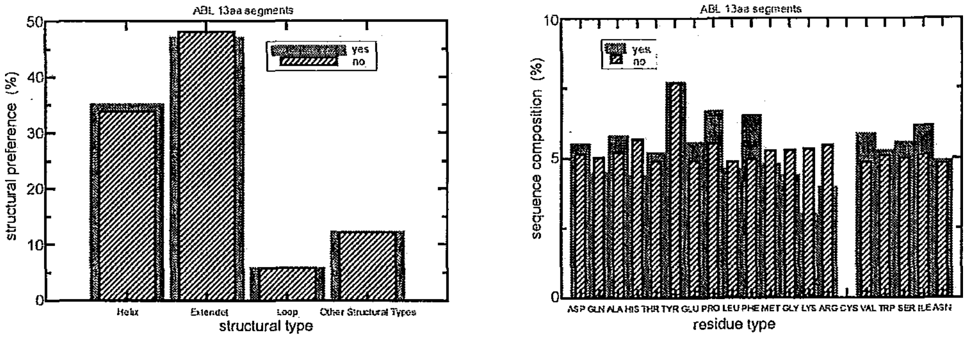


Fig. 3. Sequence composition and structural preferences for the list of short (13 residue long) peptides tested using the phosphorylation by ABL kinase. For detailed analysis we consider two subsets (based on experimental confirmation): phosphorylated 129 short peptides (YES), and not-phosphorylated 1304 short peptides (NO). Both types of information (sequence composition and structural preference) for YES instances differs only weakly from NO instances. This make it difficult to automatically discriminate between them

Based on the phosphorylation status peptides were divided into 129 positive and 1304 negative sites. In this case there is almost no visible differences between the predicted local structure of both types of sites and very small differences in the sequence compositions between those two sets (see Fig. 3). The structure of the peptides is unknown so no comparison with real 3D structure of peptide chains could be performed.

4. THE AUTOMATIC PHOSPHORYLATION SITE PREDICTOR

In this section we introduce the automatic phosphorylation sites predictor, which is based on sequence and structure information. First we build for each tested phosphorylation type the sequence composition preference [S_{ik}] and the predicted local structural preference [Q_{is}]. The k-th index of the first matrix runs from 1 up to 20 (items describe the preference for different type of amino acid), and the s-th index of the second matrix runs over all 11 structural classes. The z'-th index in both matrixes describes position of a residue in short local segment around active site central residue. The sequence preferences are constructed by simple averaging over all known phosphorylation instances found in Swiss-Prot database. The structural preference matrix is computed in similar way using predicted local structure LSSs for each annotated instance.

The prediction method is as follows. First we dissects a query protein into overlapping short segments of length 7*aa*. For each segment we calculate the probability score given by:

$$SQ = \frac{1}{7}\sum_{i=1}^{7} S_{ik} Q_{is} ,$$

where $[S_{ik}]$ is the k-th coordinate of the matrix representing the normalized sequence preference for i-th residue of a segment centered around active site (i = 4). The k-th index represents the type of amino acid found in a query protein's segment at i-th position. The $[Q_{is}]$ factor represents the normalized similarity between the predicted local structure of a segment of a query protein and the local structural preference for each position of a segment, and each type of phosphorylation process. Using this score we describe the reliability of a prediction for a certain type of phosphorylation at specified residue. As the output of our method we take only those sites, which have the score *SQ* larger than given cut-off value.

In order to qualify the significance of information contained in $[S_{ik}]$ and $[Q_{is}]$ matrixes we perform predictions on the same set of all proteins with known phosphorylated sites verified by experiments. For each type of phosphorylation we collect the predictions using *SQ* score. Then we divide the resulting set of predicted as phosphorylated residues into two groups: confirmed "1" (by experiments and therefore annotated in Swiss-Prot database), and "0" (not confirmed, or false predictions). For both subsets we analyze scores and present them as histograms. We used *SQ* = 0.1 as the cut-off for our automatic method.

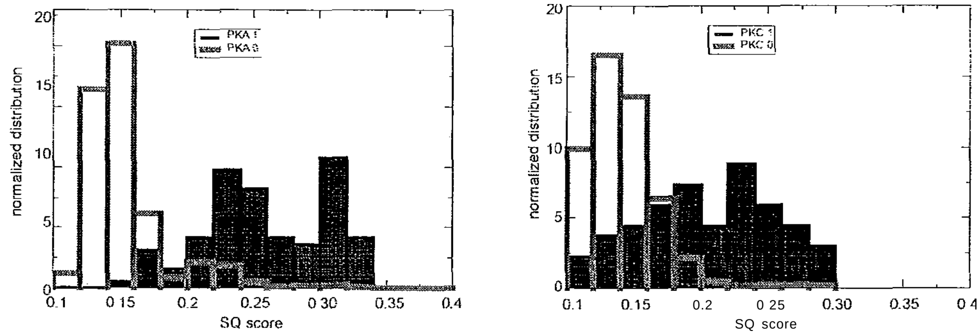


Fig. 4. The results of prediction method for automatic assignment of phosphorylation site based on sequence and structure similarity between the query proteins and the library of known annotated in Swiss-Prot database short segments around phosphorylated sites. Those normalized histograms shows distributions of *S*Q scores for predicted segments (true instances marked by "1", and false ones marked by "0"). One can notice the clear difference of mean value between those two sets of predictions

On Figure 4 we illustrate the results for PICA and PKC phosphorylation, which shows the ordering efficiency of our method. In the test cases almost all real phosphorylation sites were predicted using our algorithm, but some of them with low value of *SQ* score. In spite of that there is a clear difference between both histograms and mean values for true predictions and false ones. The proper cut-off value (which depends on the type of phosphorylation process) can provide good percentage rate of efficiency, loosing only the small subset of annotated sites. Yet the more advanced and refined statistical methods (for example similar to those used in PSI-blast tool) are needed to improve the overall benchmark results for our method. In accordance with results of the previous section the sequence term of the *SQ* score (described by the matrix *S*) is the core part in our prediction method. The structural part of the algorithm is the most time consuming step, so if the best efficiency in practical applications of our method is not needed it can be neglected. The level of improvement is marginal, and Fig. 4 is hold also for the data with only sequence part of *SQ* score. The difference between those two scores is almost negligible in most cases.

5. CONCLUSIONS

The main problem we faced in this research project is the insufficient number of experimentally verified structures from PDB database for main chain around phosphorylated sites. Because of the very poor statistics of experimental data structural part of our method has only limited use. In many cases even though coordinates for the phosphorylated proteins are available coordinates for the actual sites are missing indicating that a structure disorder tool like GlobPlot [15] could improve the predicting efficiency by filtering our predictions.

In order to avoid the problem of lacking experimental structural information we use the fragments database library for predicting a local structure of proteins using only sequence information. Our local structure prediction method has very low precision in studied cases but its results are in qualitative agreement with experimental data. The predicted structural preference for annotated instances is very similar to not annotated ones. Including the structural information is improving only slightly results of automatic prediction of phosphorylation sites over only sequence based method.

We conclude our work stating that further development of automatic phosphorylation sites annotation predictors should gain a significant improvement using statistical algorithms based mostly on sequence analysis. Our preliminary results shows that the automated server for predicting the active sites for various biologically significant processes (for example phosphorylation by kinases) based only on analyzing the sequence composition for short fragments in proteins is able to provide good efficiency in search for phosphorylation sites. This strongly supports further developing of automatic predictors for predicting unknown localization of phosphorylation active sites.

Acknowledgment

This work was supported by the 6FP EU grant "ELM" (QLRI-CT2000-00127) and polish SPUB-M 438/2003-2004 grant.

References

[1] M. Levitt, and M. Gerstein, *A unified statistical framework for sequence comparison and structure comparison*, Proc. Natl. Acad. Sci. **95**, 5913-5920 (1998).

[2] R. Luthy, A. D. McLachlan, and D. Eisenberg, *Secondary structure-based profiles: use of structure-conserving scoring tables within protein super families*, Bioinformatics, **16**, 1111-1119, (1991).

[3] D. Fischer and D. Eisenberg, *Protein fold recognition using sequence-derived predictions. Protein Sci.,* **5,** 947-955 (1996).

[4] H. Xu, R. Aurora, G. D, Rose, and R. H. White, *Identifying two ancient enzymes in archaea using predicted secondary structure alignment,* Nature Structural Biology, **6**, 750-754 (1999).

[5] L. Rychlewski, and A. Godzik, *Secondary structure prediction using segment similarity. Protein Engineering,* **10,** 1143-1153 (1997).

[6] T. M. Yi, and E. S. Lander, *Protein secondaiy structure prediction using nearest-neighbor methods*, J. Mol. Biol., **232,** 1117-1129 (1993).

[7] C. Bystroff, and D. Baker, *Prediction of local structure in proteins using a library of sequence-structure motifs,* J. Mol. Biol., **281,** 565-577 (1998).

[8] L. Rychlewski, M. Kschischo, L. Dong, M. Schutkowski, and U. Reimer, *Target specificity analysis of the Abl kinase using peptide microarray data,* Submitted to JMB (2003).

[9] J. M. Chandonia, N, S. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S. E. Brenner, *ASTRAL compendium enhancements,* Nucleic Acids Research, **30,** 260-263 (2002).

[10] S. E. Brenner, P. Koehl, and M. Levitt, *The ASTRAL compendium for sequence and structure analysis,* Nucleic Acids Research **28**, 254-256 (2000).

[11] S. F. Altschul. W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, *Basic local alignment search tool,* J. Mol. Biol., **215,** 403-410 (1990).

[12] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs,* Nucleic Acids Res. **25,** 3389-3402 (1997).

[13] L. Rychlewski, L. Jaroszewski, W. Li, A. Godzik, *Comparison of sequence profdes. Strategies for structural predictions using sequence information,* Protein Science, **9**, 232-241 (2000).

[14] L. Jaroszewski, W. Li, and A. Godzik, *Improving the quality of twilight-zone alignments,* Prot. Sci., **9**, 1487-1496 (2001).

[15] R. Linding, R. B. Russell, V. Neduva, and T. J. Gibson, *GlobPlot: exploring protein sequences for globularity and disorder.* Nucleic Acids Research. **31.** 3701-3708 (2003).