J Mol Model (2005) 11: 431–438
DOI 10.1007/s00894-005-0235-z

ORIGINAL PAPER

**Dariusz Plewczynski · Lukasz Jaroszewski
Adam Godzik · Andrzej Kloczkowski
Leszek Rychlewski**

Molecular modeling of phosphorylation sites in proteins using a database of local structure segments

Received: 12 May 2004 / Accepted: 30 November 2004 / Published online: 11 August 2005

Abstract A new bioinformatics tool for molecular modeling of the local structure around phosphorylation sites in proteins has been developed. Our method is based on a library of short sequence and structure motifs. The basic structural elements to be predicted are local structure segments (LSSs). This enables us to avoid the problem of non-exact local description of structures, caused by either diversity in the structural context, or uncertainties in prediction methods. We have developed a library of LSSs and a profile—profile-matching algorithm that predicts local structures of proteins from their sequence information. Our fragment library prediction method is publicly available on a server (FRAGlib), at http://ffas.ljcrf.edu/Servers/frag.html. The algorithm has been applied successfully to the characterization of local structure around phosphorylation sites in proteins. Our computational predictions of sequence and structure preferences around phosphorylated residues have been confirmed by phosphorylation experiments for PKA and PKC kinases. The quality of predictions has been evaluated with several independent statistical tests. We have observed a significant improvement in the accuracy of predictions by incorporating structural information into the description of the neighborhood of the phosphorylated site. Our results strongly suggest that sequence information ought to be supplemented with additional structural context information (predicted with our segment similarity method) for more successful predictions of phosphorylation sites in proteins.

Keywords Library of protein motifs · Profile–profile sequence similarity (PSI-BLAST · FFAS) · Kinase substrate prediction · Nearest neighbors · Local structure segment (LSS)

Introduction

Protein phosphorylation by various kinases is an important mechanism for signal transduction and the control of intracellular processes. However, only a fraction of protein kinases and their phosphorylation sites have been characterized in detail. This creates a serious need for automatic methods to predict possible posttranslational modification sites in proteins. Such prediction methods should only use protein-sequence information as their input, which is usually the only available information. In order to incorporate additional structural information, we have developed a new method for the prediction of the local structure of proteins from sequence. This method enables us to characterize in better detail local neighborhoods of phosphorylation sites in terms of both local sequence and structural preferences.

Defining the blocks forming global protein structure on the basis of local structural regularity has been a fruitful approach, extensively used in describing and predicting structure from sequence information. Sets of

D. Plewczynski · L. Rychlewski
BioInfoBank Institute, Limanowskiego 24A/16,
60-744 Poznan, Poland

D. Plewczynski (✉)
Interdisciplinary Centre for Mathematical
and Computational Modeling, Warsaw University,
Warsaw, Poland
E-mail: darman@bioinfo.pl
Tel.: +48-61-8653520
Fax: +48-61-8643350

L. Jaroszewski · A. Godzik
Bioinformatics Core JCSG, University of California
San Diego, La Jolla, CA, USA

A. Godzik
The Burnham Institute, La Jolla, CA, USA

A. Kloczkowski
Baker Center for Bioinformatics and Biological Statistics,
Iowa State University, Ames, IA, USA

common structural motifs can be used to describe similarities between distant, non-homologous proteins. The reason for using structural information about proteins is straightforward. Structural comparisons are capable of detecting approximately twice as many distant relationships as sequence comparisons, at similar error rates [1]. The collection of short sequence fragments, even those not having unique structures, can provide the best vocabulary for predicting the structures of proteins from their sequence information.

The idea of using a minimal set of structural units, even those that occur in non-homologous proteins, has been widely utilized for the last few decades. For many years previously, secondary structure elements were popular as building blocks and were used with significant success. In the first efforts at threading and fold recognition [2, 3] secondary structure was used to enhance the recognition of distant homologs. Aurora, Rose and co-workers [4] took this approach to an extreme by comparing proteins represented as strings of only three Q3 symbols (α for alpha-helix, β for beta-strand and *c* for coil) for predicted secondary structures.

The most successful implementation of a structural units framework is local secondary structure prediction based on segment similarity. This can include two different approaches. The first approach is secondary structure prediction based on nearest-neighbor algorithms [5, 6]. The second is a new method for local structure predictions developed by Baker and co-workers [7], based on a library of sequence-structure motifs called I-sites. The I-sites library was used to build a hidden Markov model for protein sequence [8]. Both approaches take uncertainties about local structure into consideration. The results provide a ranked list of possible Local structure segments (LSSs) for a protein, instead of predicting only a single local structure assignment at each position (predicted local structure). Due to the essential uncertainty of local structures in the absence of tertiary interactions, this diversity of local structures is unavoidable. One main drawback to these methods is their not making further use of this uncertainty. Identified segments are discarded and only secondary structure averages over predicted segments (at each position in the sequence) are kept. The only exception is Baker's de novo protein structure prediction program ROSETTA [9, 10], which combines the predicted segments into a compact tertiary structure. The newest extension of this method predicts conformations of structurally divergent regions in comparative models [11]. Initial conformations for short segments are selected from the protein structure database, whereas longer segments are built up from fragments drawn from the database and combined by the ROSETTA algorithm.

Although current methods can model short loop regions in proteins quite accurately, modeling longer structurally divergent regions remains a difficult problem. The plan here is to predict the local structures of protein main chain, but also include a structural analysis of local neighborhood of phosphorylation sites in proteins. Direct comparison of our method with other state-of-the-art local structure prediction methods has been presented elsewhere [12]. In this article, we focus our attention on local structural descriptions of phosphorylation sites in proteins and give the latest results from our phosphorylation sites predictor [13].

The Swiss-Prot database [14] contains a large number of annotated phosphorylation sites. To develop and test automatic methods for the prediction of phosphorylation sites, we use sequence information from this database. For test purposes we ignore all residues having phosphorylation annotations ‘‘by similarity’’, ‘‘hypothetical’’ or ‘‘predicted’’. We use proteins phosphorylated by PKA and PKC kinases only, these two representing the largest number of cases in the Swiss-Prot database, and which can therefore be used as a test set for our automatic annotation method. We prepare a database of all real (experimentally determined) structures of backbone segments around phosphorylation sites. The structures are collected using the PSI-Blast server running on the PDB database (PDB-Blast) (http://bioinfo.pl/).

Next, in the Materials and Methods section, we provide detailed information about the preparation of the database of short protein fragments. Then, we describe the automatic annotation method for phosphorylation sites. In the Results section, we present benchmarks used for the statistical analysis of the local structure predictions. We describe the local sequence composition of segments around phosphorylated sites together with the predicted structural information. We also include the analysis of background sequence and structural preferences of LSSs unannotated in the Swiss-Prot database. Finally, we present our conclusions and discuss possible future developments.

Materials and methods

Database of local structural segments

We would like to develop a database of short protein segments (motifs), and utilize them for the prediction of local structure of proteins, using only the sequence information [12]. The main idea behind this is to use the most general definition of blocks forming the global protein structure on the basis of local structural regularity. Such a database of secondary structure fragments will be used later to describe local protein structure. The library of fragments is constructed using a representative set of proteins from the ASTRAL database [15, 16] (see Fig. 1). It covers all fragments with well-defined secondary structures that can be found in proteins. We remove from our non-redundant database any identical entries (both in terms of structure and sequence information), so it contains only unique entries.

To define the local structure for each amino acid, we adopt Baker's definition of the local structure's symbolic representation, SLSR consisting of 11 symbols *HGEeBdbLlxc*, each constrained to certain regions of backbone dihedral (ϕ and ψ) angles [7]. Then, each LSS is described as a short string of local-structure symbols. For the database of fragments we use a large collection (over 2000) of short local structure segments (7–19 aa) for which the local structural codes are the same along the chain, except two terminal residues at the beginning, and at the end. We store this large collection of fragments with their sequences, symbolized in local structures representation SLSR codes and parts of the homology profiles from their parent proteins (see Fig. 2). The structural redundancy in the database means that there are many fragments with the same structural codes or sequence profiles (except at the ends of the segment). Each of these fragments comes from a different parent protein, or different parts of a chain in the same parent protein.

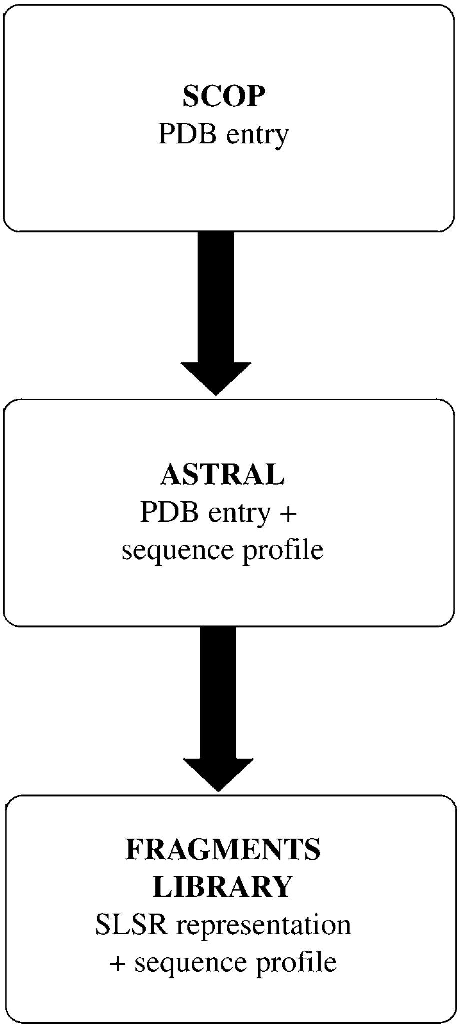

Fig. 1 The fragment database is built from the ASTRAL representative subset of the SCOP database using a 40% sequence similarity threshold. The symbolic local structure representation codes of all fragments together with their homology sequence profiles are cut off from the SLSR codes and homology profiles of a parent protein, respectively and stored in our library

Segment structural similarity prediction method

Our algorithm predicts the local molecular structure of a protein based only on the sequence information. We use the same SLSR representation of the structural information for proteins as in our fragments database. Our method for predicting LSSs is as follows: First, with PSI-Blast [17] we create the sequence profile for the query protein. Then, for all possible overlapping segments of length 7–19 aa, we compare short segments of the profile (taken from the profile of the whole protein), with the database of profiles of short fragments. For each pair of LSSs compared (one from the query protein, and the second from the fragments database) we calculate the profile homology score using the FFAS method [18].

As the result of this procedure, each position in the query protein is represented by a large collection of predicted LSSs. The number of assigned segments depends on the local structural tendency of this residue and its neighbors for a specific local chain conformation. Each predicted fragment is scored with a profile–profile sequence similarity for the original segment of the query protein. For each position of a query protein we retain only the 20 highest scoring fragments in the profile–profile sequence similarity score, and discard the remainder.

We store (for each position in the query sequence) a list of predicted fragments with their lengths, sequence profiles and structural descriptors in terms of SLSR

SEQUENCE
HOMOLOGY
SEQUENCE PROFILE
SLSR codes

Fig. 2 The structural fragment database is presented. We remove all sequences identical in terms of both SLSR codes and sequence homology profiles, to avoid duplication of fragments in the database. Each fragment is represented by its sequence string, a matrix of its sequence profile and by the string of SLSR codes representing the local structure of the C^{α} chain in the ϕ–ψ space of torsional angles

codes. We call the collection of these sets the protein local structural segments (PLSSs).

The automatic phosphorylation sites predictor

Figure 3 shows how the automatic phosphorylation sites predictor based on sequence information works. It is an extension of our previous work [13]. It utilizes a knowledge-based database of short segments that are known to be phosphorylated by PKA or PKC kinases. The annotated LSS database is built from segments having known sequence profiles (from the Swiss-Prot database) and known structures (extracted from the PDB database with the PDB-Blast tool). First, for each type of phosphorylation tested, we build the sequence composition preference matrix $[\mathbf{s}^{j+}_{ik}]$, and the structural preference matrix $[\mathbf{q}^{j+}_{is}]$. Index k specifies the type of amino acid, running from 1 to 20, the superscript j is the length of the fragment, and index s refers to the specific structure type from the 11 structural classes. Index I in both matrixes describes the position of the residue in the local segment around the phosphorylation site. Sequence preferences are constructed by simple averaging over all annotated phosphorylation entries in the Swiss-Prot database. The structural preference matrix is calculated similarly by using local structure LSSs for each annotated segment. We also calculate the sequence composition preference matrix $[\mathbf{s}^{j-}_{ik}]$, and the structural preference matrix $[\mathbf{q}^{j-}_{is}]$ for negative instances (known to be non-phosphorylated by

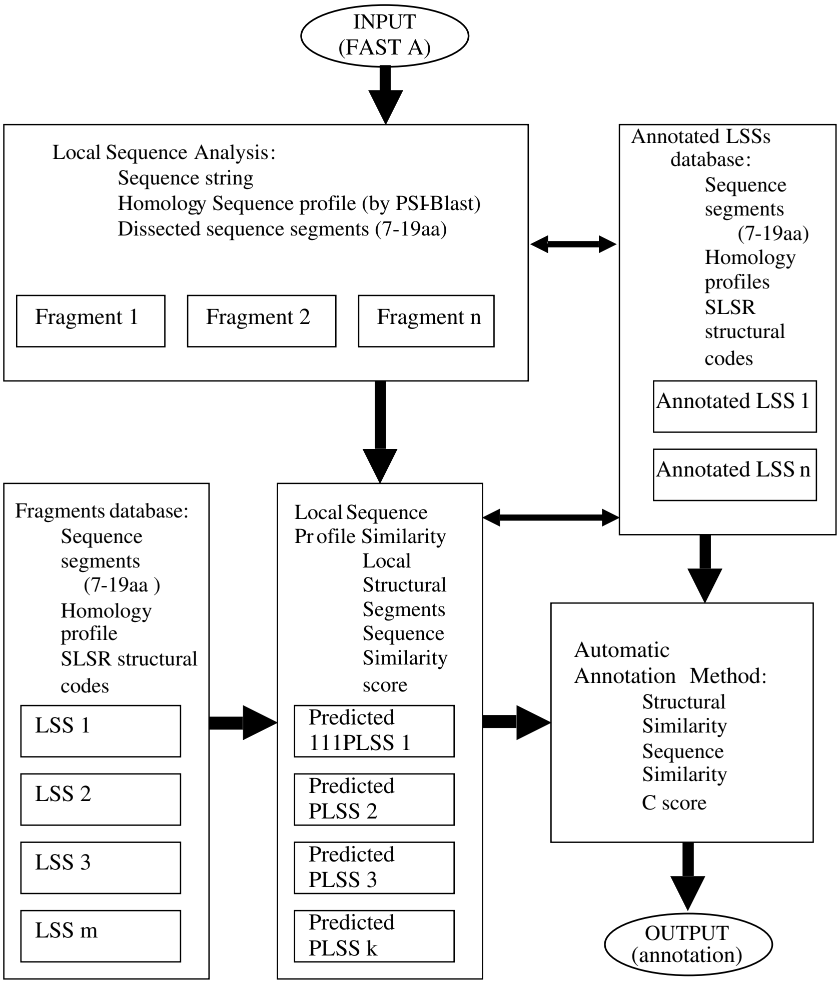

Fig. 3 The automatic annotation service used for predicting posttranslational sites in proteins. Our local prediction method compares sequence profile of the query protein against all members of the fragment database. The query protein is dissected into short parts (7–19 residues long). For each fragment a similarity search is performed. Each member of the fragment database that is similar in terms of the homology sequence profile to the query fragment is added to the list of predicted structures. The list is then sorted and cut to choose 20 results. If the highest score of the predicted fragment is below the user defined cut-off value then the whole prediction is discarded. At the end, some parts of the query protein are covered by the list of 20 fragments from the database (PLSSs). The resulting PLSSs are then compared with the database of segments known to be phosphorylated by PKA or PKC kinases. Each verified segment in the database is stored with the LSS and sequence profile. Our similarity method assigns the probabilities of being phosphorylated (the C scores) to all predicted PLSSs for the input query protein. Sites having scores higher then the cut-off value C_0 (different for PKA and PKC phosphorylations) are predicted to be phosphorylated by a specific kinase

any kinase with the same central amino acid). Then, the entries in the normalized preference matrix are ratios defined as $[\mathbf{S}^{\mathrm{j}}_{\mathrm{ik}}]=[\mathrm{s}^{\mathrm{j+}}_{\mathrm{ik}}/\mathrm{s}^{\mathrm{j-}}_{\mathrm{ik}}]$, $[\mathbf{Q}^{\mathrm{j}}_{\mathrm{is}}]=[\mathrm{q}^{\mathrm{j+}}_{\mathrm{is}}/\mathrm{q}^{\mathrm{j-}}_{\mathrm{is}}]$. This allows us to define a composite score C, combining sequence and structure information for phosphorylation processes for which there is sufficient experimental data.

Our prediction method is: first, we dissect a query protein into short overlapping segments having lengths from 7 to 19 aa. For each segment we calculate the combined sequence and structure probability score C

$$C=\sum_{j=7}^{19}\left(\frac{1}{j}\sum_{i=1}^{j}S^{j}_{ik}Q^{j}_{is}\right).$$

Here S^{j}_{ik} is the normalized sequence preference for the k-th type of amino acid at the I-th residue in a segment around a phosphorylation site ($I=j/2+1$). Q^{j}_{is} represents the corresponding normalized structural preference for the s -th structural state at the I-th residue in a segment. The S^{j}_{ik} and Q^{j}_{is} are separately computed for each type of phosphorylation processes (in our case, for sites formed acted upon by PKA and PKC kinases). The C score gives the likeness for phosphorylation of a given residue. We assume that a potential phosphorylation site should have a C score higher than a specified cutoff value C_0. The cutoff values for different phosphorylation processes would be different. We define the cutoff in a simple way as

$$C_0=C_{\mathrm{mean}}-C_{\mathrm{dev}},$$

where C_{mean} is the mean value of the C score calculated for the set of segments that are phosphorylated (true positives) and C_{dev} is the standard deviation of the C score for the same set.

The output of the prediction contains three columns with the residue number, the corresponding amino acid and the C score, respectively.

Results

Results for the local structure prediction method

We use our own statistical benchmark for a more detailed analysis of accuracy and quality of predictions of PLSSs for various proteins. We selected a set of 645 proteins with different folds from the ASTRAL database (all these proteins differ from those used in the development of the database of fragments). For each protein we predict PLSSs using our tool. Then we collect results of all such predictions and analyze them in terms of their various statistical properties. All results are summarized in Table 1. The predictions in Table 1 for each length of fragment are averaged over all proteins from the benchmark dataset. The measure of the quality of predictions is defined as the percentage of correctly predicted (identical between real LSS and PLSS) symbolic local structures representation codes (SLSRs) divided by the number of residues in the fragment. The Q3 structural score is defined for three (α-helices, β-strands and coil c) secondary structure states [4]. In this case we

Table 1 Statistical description of the prediction quality for the benchmark dataset constructed from the SCOP fold level proteins for different fragment lengths

Length of fragments	Number of fragments for each length found in predictions	Average sequence similarity	Quality of prediction	Q3 quality of prediction	Average profile–profile score	Average number of positive alignments for each length	Total number of positive alignments for each length
7	38220	0.2288	0.1792	0.3877	−1.1140	824	1062038
8	21242	0.2474	0.1762	0.4261	−1.1050	491	364555
9	15964	0.2798	0.1720	0.4683	−1.0859	329	188058
10	12454	0.3120	0.1758	0.5125	−1.0603	232	101393
11	9464	0.3424	0.1664	0.5382	−1.0416	174	58662
12	6474	0.3939	0.1651	0.5886	−1.0173	141	29513
13	5538	0.4443	0.1594	0.6354	−0.9987	108	19205
14	4576	0.4530	0.1609	0.6427	−0.9865	75	12000
15	4472	0.5321	0.1595	0.7078	−0.9824	66	9793
16	3900	0.5387	0.1546	0.7087	−0.9583	48	6219
17	3744	0.5759	0.1478	0.7496	−0.9453	33	4330
18	3588	0.5722	0.1464	0.7489	−0.9363	26	3428
19	16640	0.5784	0.1436	0.7663	−0.9165	32	10487

The second column shows the number of predicted fragments for a given segment length obtained from the database of secondary structure segments. The third column is the average sequential similarity of predicted fragments to the corresponding fragment of the query protein. An average coverage of good fragments (with structural similarity larger than 0.75 in terms of SLSR codes to the predicted segment of a query protein) over the output list of predicted fragments is quite uniform. The average position of the good fragments within the list of the 20 (chosen arbitrarily) best predicted fragments for each position in the query protein divided by the number of fragments in the list lies between 0.49 and 0.54, i.e. in the middle of the list. The fourth and the fifth column show the average quality of the structural predictions. As the measure of quality we use the percent of properly predicted symbolized local structures representation SLSR codes, or in the case of Q3 only the three secondary structure elements (helices, beta strands, and loops). The sixth column gives the average profile–profile score for the predicted fragments. The last two columns describe the average number of positive alignments, and the total number of accepted fragment alignments for each fragment length convert Baker's codes *HGEeBdbLlxc* into the three Q3 classes. The quality of predictions measured by Q3 reaches 73%; however, the average coverage of well-predicted fragments is quite uniform (the structural similarity to true query protein structure is above 75%) within the output list of the first 20 predicted fragments. However, the best PLSSs are almost never the first on our list.

Local segments' sequence and structure preferences around phosphorylated sites

We take the list of proteins from Swiss-Prot database as those having at least one experimentally verified phosphorylation site. We neglect all uncertain phosphorylation sites annotated "by similarity", "hypothetical" or "predicted". For our detailed analysis, we have 67 proteins with PKA phosphorylation and 98 sites in total and 49 proteins with PKC phosphorylation and 73 sites. We obtained structures of proteins around phosphorylation sites with the PDB-Blast server developed in our group (http://bioinformatics.burnham-inst.org/pdb_blast/). (It uses PSI-Blast program comparing the sequence from a query protein against all sequences from the PDB database with strict thresholds in order to obtain only one true (from crystal protein data) structure for the protein segment.) We collect these segments for the 56 proteins with PKA and the 38 with PKC phosphorylation sites. We found only 11 structural segments around sites with both PKA and PKC phosphorylations. It appears that most of the phosphorylation sites occur in unstructured parts of proteins, which are difficult to crystallize and that those coordinates are consequently often missing in the PDB [19].

To sample the background sequence preferences, we take 17,718 sites not annotated as being PKA phosphorylated and 18,799 sites not annotated as being PKC phosphorylated with the appropriate central amino acids. In order to obtain background preferences for sites with known structures, we also extract 340 PKA-negative and 141 PKC-negative sites from protein segments with known coordinates and the correct central residues (either Ser or Tyr). We analyze the sequences and local structure compositions of those positive and negative cases. While the sequence composition of both types of phosphorylations exhibit a clear difference, much less significant differences could be observed between the local structures of the two types (see Fig. 4). The predicted local structures of both types are in qualitative agreement with the real structures. A comparison with other available structure prediction tools like ROSETTA [9, 11] or HMMstr/I-sites [10] was also performed. The differences between the results of those methods and ours for modeling the local structural preferences around phosphorylation sites are within the accuracy of our method. The *C* scores show similar tendencies for all methods (see Fig. 4).

The efficiency of the phosphorylation sites predictor

Our tests were performed on all proteins with PKA or PKC phosphorylated sites verified by experiment. For

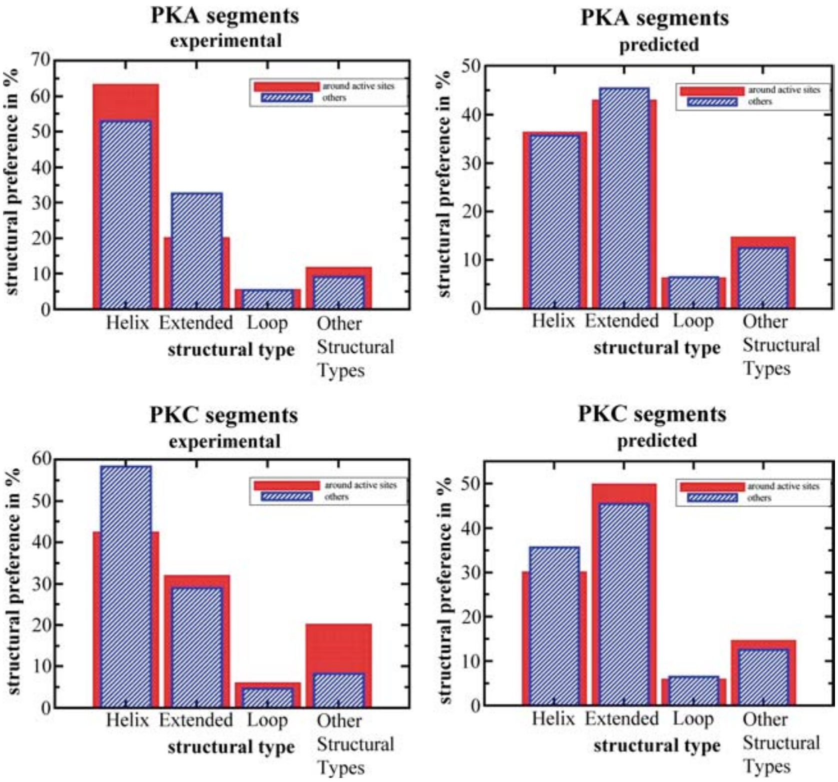

Fig. 4 Structural preferences for segments of protein backbone around the sites phosphorylated by PKA and PKC kinases sites (*left*, experimental, *right*, predicted by our method)

Table 2 The mean and standard deviation of the *C* score for the set of segments confirmed or not confirmed to be phosphorylated by PKA or PKC kinases

Phosphorylation type	Mean value		Standard deviation		Minimal/maximal values	
	Confirmed	Not confirmed	Confirmed	Not confirmed	Confirmed	Not confirmed
PKA	0.26	0.154	0.032	0.074	0.144/0.338	0.1/0.331
PKC	0.205	0.142	0.028	0.049	0.111/0.295	0.1/0.295

In calculations (to remove noise) we use only segments having a C score larger than the cut-off value of 0.1. We also give the minimum and the maximum value obtained for each dataset

each type of phosphorylation, we collect the predictions with their *C* scores. Then we divide the resulting set of predicted residues into two groups: confirmed by experiments and annotated in the Swiss-Prot database (true positives), and not confirmed (false positives). Then we analyze the *C* scores for all segments, calculating the mean values for each subset. Results are presented in Table 2.

Our test shows that, in the case of PKA and PKC phosphorylation, the recall efficiency of our method is high. In the test cases, almost all real phosphorylation sites were predicted using our algorithm, but some of them with a low value of the *C* score. There is also a clear discrepancy between the mean values for true predictions and false ones. The proper cutoff value C_0 (which depends on the type of phosphorylation process) can provide a better percentage rate of precision, losing only a small subset of annotated sites (lower recall values). Yet the number of false predictions is much larger than true ones. This is why more refined statistical methods (for example similar to those used in the PSI-Blast tool, or support vector machine approach to classification and prediction—see Ref. [20]) are needed to improve the overall benchmark results for our method. This would also be likely to help discriminate between false positives and true ones. The structural part of the *C* score (described by the matrix Q) improves predictions, but the main difference between the two types of phosphorylation sites (PKA and PKC kinases) occurs in the sequence part (S matrix). The results for other types of kinases (CK, CK2, CDC2 not included here) show larger structural preferences than sequence ones. However, the statistics for these cases are quite poor, so no definitive statement about this is possible.

Conclusions

The main problem faced in the biological application of our local structure prediction method is a lack of experimental structure segments from the PDB database for protein main chains around phosphorylation sites. For these cases, we have poor statistics. It is also clear that our local structure prediction method has low precision for the cases studied because the phosphorylation sites often occur in the unstructured regions of proteins [19]. The lack of experimentally determined structures at phosphorylation sites greatly impairs any analysis of the potential structural preferences for the kinase towards their targets. In many cases, even though coordinates for the phosphorylated proteins themselves are available, the coordinates for the actual sites within these are missing, indicating that a structure disorder prediction tool (such as GlobPlot [21]) might improve the predicting efficiency and benchmark results by filtering our predictions.

Our prediction service incorporates diversities of local molecular structure predictions, which are crucial for proper description of conformation of a protein chain around posttranslational modification sites. We conclude our work stating that further development of local structure prediction methods and applying them to automatic phosphorylation sites annotation methods should include more refined preparation of the fragments library. This can be done effectively using not only statistical, purely bioinformatics based tools, but also traditional molecular-modeling based on ab initio approach. This will be presented in our next paper.

Acknowledgments This work was supported by the USA grant ("SPAM" GM63208), BioSapiens (LHSG-CT-2003-503265), ELM (QLRT-CT-2000-00127) and GeneFun (LSHG-CT-2004-503567) projects within 6FP EU program.

References

1. Levitt M, Gerstein M (1998) Proc Natl Acad Sci 95:5913–5920
2. Luthy R, McLachlan AD, Eisenberg D (1991) Bioinformatics 16:1111–1119
3. Fischer D, Eisenberg D (1996) Protein Sci 5:947–955
4. Xu H, Aurora R, Rose GD, White RH (1999) Nat Struc Biol 6:750–754
5. Rychlewski L, Godzik A (1997) Protein Eng 10:1143–1153
6. Yi TM, Lander ES (1993) J Mol Biol 232:1117–1129
7. Bystroff C, Baker D (1998) J Mol Biol 281:565–577
8. Bystroff C, Thorsson V, Baker D (2000) J Mol Biol 301:173–190
9. Simons KT, Bonneau R, Ruczinski I, Baker D (1999) Proteins 37:171–176
10. Bystroff C, Shao Y (2002) Bioinformatics 18:S54–61
11. Rohl CA, Strauss CE, Chivian D, Baker D (2004) Proteins 55:656–677
12. Plewczynski D, Rychlewski L, Ye Y, Jaroszewski L, Godzik A (2004) BMC Bioinformatics 5:98
13. Plewczynski D, Rychlewski L (2003) Comput Methods Sci Technol 9:93–100
14. Bairoch A, Apweiler R (1999) Nucleic Acids Res 27:49–54

15. Chandonia JM, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE (2002) Nucleic Acids Res 30:260–263
16. Brenner SE, Koehl P, Levitt M (2000) Nucleic Acids Res 28:254–256
17. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Nucleic Acids Res 25:3389–3402
18. Rychlewski L, Jaroszewski L, Li W, Godzik A (2000) Protein Sci 9:232–241
19. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK (2004) Nucleic Acids Res 32:1037–1049
20. Zavaljevski N, Stevens FJ, Reifman J (2002) Bioinformatics 18:689–696
21. Linding R, Russell RB, Neduva V, Gibson TJ (2003) Nucleic Acids Res 31:3701–3708