Aristotle Biomedical Journal, Vol 1, No 1 e-ISSN: 2653-9748

Original Article

An in silico method for studying the phosphorylation in association to active sites

Panagiota Angeliki Galliou[1], Kleio-Maria Verrou[1,2]

[1]*Laboratory of Biological Chemistry, Medical School, Aristotle University of Thessaloniki, 54124*
[2]*School of Medicine, University of Crete, 71500*

Abstract

Post-translational modifications (PTMs) occur to a vast amount of proteins and the most common post-translational modification (PTM) is phosphorylation. Phosphorylation and dephosphorylation, regulate protein functionality by turning the protein active sites (sites with important biological function) on and off. Therefore, identification of a protein's phosphorylated residues and determination of their role are of paramount importance, especially for proteins driving diseases. Notwithstanding the multiple methodologies for identifying phosphorylated residues, literature lacks of methodologies for determination of their role. For this reason, we created a method that aims to enhance the understanding of a protein's regulation by phosphorylation as well as to aid the design of more directed and lower-cost experiments. Our method uses the PhosphoKin tool, which predicts new phosphorylated residues in a given protein sequence, identifies the possibly responsible kinases for the protein's experimentally observed phosphorylated residues and links all phosphorylated residues as well as their kinases with the protein's active sites. Our method assesses the impact of the examined kinases in the protein's phosphorylation and is suitable for associations between specific group of kinases and active sites. Also, it suggests the illustration of a phosphorylation map of the protein that is useful for further analysis.

Keywords: Phosphorylation, Prediction, Kinases, Active Sites, Bioinformatics

Corresponding author:
Panagiota Angeliki Galliou, Laboratory of Biological Chemistry, Medical School, Aristotle University of Thessaloniki, 54124, Thessaloniki, Greece. E-mail: ag.gal.work@gmail.com, Telephone: 00306982809138

Introduction

Nearly all proteins may undergo post-translational modifications (PTMs), after their biosynthesis, as part of their maturation or cell signaling process. The numerous types of identified PTMs remarkably amplify the functional diversity of the proteome by enzymatically modifying proteins (Nørregaard Jensen, 2004).

The most common post-translational modification (PTM) is phosphorylation (Khoury et al., 2011). Phosphorylation is catalyzed by enzymes named kinases and refers to the addition of a phosphoryl group (PO_3^{2-}) to a serine, threonine or tyrosine amino acid that is transferred from high-energy phosphate-donating molecules, such as ATP. Dephosphorylation, is the removal of a phosphate (PO_4^{3-}) group and is catalyzed by enzymes called phosphatases (Berthet et al., 1957; Krebs et al., 1959; Krebs and Fischer, 1956; Sutherland and Wosilait, 1955).

Phosphorylation and dephosphorylation regulate protein functionality (Cori and Green, 1943) by altering the molecule's stereochemistry. For instance, phosphorylation inhibits the activity of the Matrix Metalloprotease-2 (MMP-2), while dephosphorylation enhances it (Sariahmetoglu et al., 2007). Moreover, phosphorylation and dephosphorylation can turn active sites – particular regions in a protein that have a functional role which can either be undergoing a chemical reaction or binding substrates – on and off (Ashcroft et al., 1999; Olsen et al., 2006).

Given the importance of phosphorylation in the regulation of protein functionality, identification of phosphorylated residues in proteins and determination of their role is of critical significance, especially for proteins driving diseases. For this reason, a branch of proteomic research, called phospho-proteomics, is solely devoted to the above purpose. There are many methodologies for identifying the phosphorylated residues of proteins with the most widely-used nowadays in proteomic research being mass-spectrometry (Han et al., 2008). However, understanding the role of these phosphorylations in regulating the function of a protein is more complex and requires further investigation beyond just simply identifying the phosphorylated residues. Literature lacks of methodologies determining the role of phosphorylated residues as well as of kinases in the regulation of protein function.

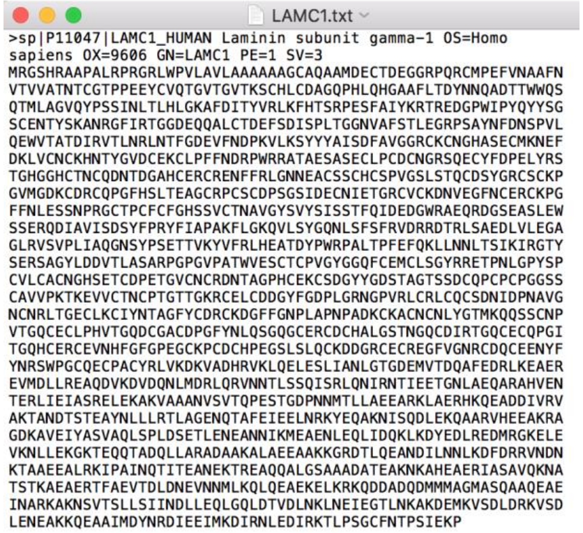
Therefore, we created a method that maps the experimentally observed and predicted phosphorylated residues in a protein relatively to its active sites, aiming to aid the clarification of the protein's regulation via phosphorylation. This method uses the PhosphoKin tool that we created to predict phosphorylated residues in a protein sequence using kinase binding motifs. Then, it combines the predict phosphorylated residues with the protein's already known experimentally observed phosphorylated residues to identify the possibly responsible kinases for each experimentally observed phosphorylated residue. After, this method combines all the phosphorylated residues with the protein's active sites to examine the impact of each kinase in the protein phosphorylation, to make associations and draw conclusions.

Methods

Protein sequence

In absence of an *in vitro* protein sequencing, the protein sequence can be taken from various protein databases such as Uniprot (The UniProt Consortium, 2017) and NCBI (Canese and Weis, 2002). However, providing that the proteins entries in protein databases (O'Leary et al., 2016; Pruitt et al., 2009; Zerbino et al., 2018) usually derive from translations of coding sequences (CDS), choosing an entry requires attention.

The entry of a peer reviewed and highly curated protein sequence should be chosen in order to have an adequate similarity with the actual protein in nature. For instance, Uniprot can filter protein entries by their review status. Also, it as has an annotation score and a sequence status, which provide a heuristic measurement of the annotation content and indicate the completion of the canonical sequence, respectively. A protein entry with a complete sequence and the highest annotations score should be considered a good option. Further, the source organism in the protein entry should be the organism of the study. The protein sequence should be downloaded in a fasta format and saved in a .txt document as displayed in Figure 1. The name of the protein should be given before the extension of the file (e.g. LAMC1.txt).



```
>sp|P11047|LAMC1_HUMAN Laminin subunit gamma-1 OS=Homo
sapiens OX=9606 GN=LAMC1 PE=1 SV=3
MRGSHRAAPALRPRGRLWPVLAVLAAAAAAGCAQAAMDECTDEGGRPQRCMPEFVNAAFN
VTVVATNTCGTPPEEYCVQTGVTGVTKSCHLCDAGQPHLQHGAAFLTDYNNQADTTWWQS
QTMLAGVQYPSSINLTLHLGKAFDITYVRLKFHTSRPESFAIYKRTREDGPWIPYQYYSG
SCENTYSKANRGFIRTGGDEQQALCTDEFSDISPLTGGNVAFSTLEGRPSAYNFDNSPVL
QEWVTATDIRVTLNRLNTFGDEVFNDPKVLKSYYYAISDFAVGGRCKCNGHASECMKNEF
DKLVCNCKHNTYGVDCEKCLPFFNDRPWRRATAESASECLPCDCNGRSQECYFDPELYRS
TGHGGHCTNCQDNTDGAHCERCRENFFRLGNNEACSSCHCSPVGSLSTQCDSYGRCSCKP
GVMGDKCDRCQPGFHSLTEAGCRPCSCDPSGSIDECNIETGRCVCKDNVEGFNCERCKPG
FFNLESSNPRGCTPCFCFGHSSVCTNAVGYSVYSISSTFQIDEDGWRAEQRDGSEASLEW
SSERQDIAVISDSYFPRYFIAPAKFLGKQVLSYGQNLSFSFRVDRRDTRLSAEDLVLEGA
GLRVSVPLIAQGNSYPSETTVKYVFRLHEATDYPWRPALTPFEFQKLLNNLTSIKIRGTY
SERSAGYLDDVTLASARPGPGVPATWVESCTCPVGYGGQFCEMCLSGYRRETPNLGPYSP
CVLCACNGHSETCDPETGVCNCRDNTAGPHCEKCSDGYYGDSTAGTSSDCQPCPCPGGSS
CAVVPKTKEVVCTNCPTGTTGKRCELCDDGYFGDPLGRNGPVRLCRLCQCSDNIDPNAVG
NCNRLTGECLKCIYNTAGFYCDRCKDGFFGNPLAPNPADKCKACNCNLYGTMKQQSSCNP
VTGQCECLPHVTGQDCGACDPGFYNLQSGQGCERCDCHALGSTNGQCDIRTGQCECQPGI
TGQHCERCEVNHFGFGPEGCKPCDCHPEGSLSLQCKDDGRCECREGFVGNRCDQCEENYF
YNRSWPGCQECPACYRLVKDKVADHRVKLQELESLIANLGTGDEMVTDQAFEDRLKEAER
EVMDLLREAQDVKDVDQNLMDRLQRVNNTLSSQISRLQNIRNTIEETGNLAEQARAHVEN
TERLIEIASRELEKAKVAAANVSVTQPESTGDPNNMTLLAEEARKLAERHKQEADDIVRV
AKTANDTSTEAYNLLLRTLAGENQTAFEIEELNRKYEQAKNISQDLEKQAARVHEEAKRA
GDKAVEIYASVAQLSPLDSETLENEANNIKMEAENLEQLIDQKLKDYEDLREDMRGKELE
VKNLLEKGKTEQQTADQLLARADAAKALAEEAAKKGRDTLQEANDILNNLKDFDRRVNDN
KTAAEEALRKIPAINQTITEANEKTREAQQALGSAAADATEAKNKAHEAERIASAVQKNA
TSTKAEAERTFAEVTDLDNEVNNMLKQLQEAEKELKRKQDDADQDMMMAGMASQAAQEAE
INARKAKNSVTSLLSIINDLLEQLGQLDTVDLNKLNEIEGTLNKAKDEMKVSDLDRKVSD
LENEAKKQEAAIMDYNRDIEEIMKDIRNLEDIRKTLPSGCFNTPSIEKP
```

Figure 1. The .txt file containing the protein sequence that is inserted as input to PhosphoKin. The sequence of the human laminin γ1-chain (LAMC1) is shown. The sequence was taken from the Uniprot database (entry: P11047).

Recording of protein's active sites

The position, sequence and function of the protein's active sites should be recorded by an extensive literature search. Overlapping active sites should be considered as one. All active sites described in the literature should be included in the results regardless of their biological function, experimental identification method and efficacy differences between cell types. Nevertheless, a categorization of active sites, according to differences in their functionality (e.g. distinct binding molecules, cell type or cell line specificity and role in physiology or diseases), could be very beneficial for conducting results and will allow focusing on active sites of interest. An example is demonstrated in Table S1.

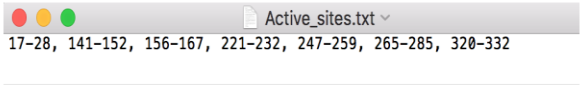
Furthermore, due to various reasons, such as expression and isolation difficulties, certain human proteins are less studied than their orthologous proteins in other species. For example, most literature-derived active sites were identified in the mouse or rat γ1-chain of laminin (lamc1) instead of the human γ1-chain of laminin (LAMC1). In a case that the literature-derived active sites of the protein were identified in a different organism than the one studied, the Basic Local Alignment Search Tool (BLAST) platform (Altschul et al., 1990) could be used. BLAST will align the orthologous sequences of the protein between the different organisms and return the percentage of their sequence identity. A statistically significant, high percentage of sequence identity yields a high level of confidence. The literature-derived active sites should be matched to similar subsequences in the protein of interest in order to obtain the protein of interest's active sites.

Additionally, to increase the level of confidence each active site should be assessed individually for its similarity with the corresponding literature-derived active site using a similarity score (equation 1).

$$\text{Similarity Score} = \frac{\text{Number of same residues}}{\text{Length of active site}}$$

A similarity score of 100% indicates identical subsequences, whereas of 0% completely different ones. Generally, active sites are expected to be highly conserved amongst orthologous proteins, due to evolutionary pressure. However, in case there are a few active sites with low similarity scores (less than 50%), they should be carefully examined and subjected to a more sophisticated alignment algorithm for the validation of their conservation and location in the protein of interest. Also, it is important to keep in mind that not all active sites of a protein are active simultaneously. The active sites should be saved in a .txt file in the following format (Figure 2):

Start_of_active_site_1 -
End_of_active_site_1,
Start_of_active_site_2 -
End_of_active_site_2, etc.



```
17-28, 141-152, 156-167, 221-232, 247-259, 265-285, 320-332
```

Figure 2. The .txt file containing the active sites of the protein that is inserted as input to PhosphoKin. A few active sites of the human laminin γ1-chain (LAMC1) are shown. These active sites derived from an extensive literature search.

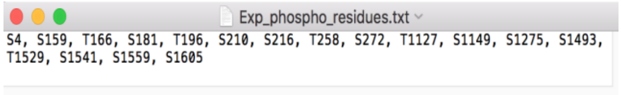
Experimentally observed phosphorylated residues

The phosphorylated residues of the protein that have been experimentally observed can be taken from the PhosphoSitePlus database (Hornbeck et al., 2015). This database contains the post-translational modified residues of mammalian proteins that were experimentally assigned. All recorded residues have been extensive and manually curated. Moreover, PhosphoSitePlus retrieves the sequence of the query protein from Uniptot.

PhosphoSitePlus can filter post-translational modifications (PTMs) by the experimental identification method using two options; Low Throughput Papers (LTP) or High Throughput Papers (HTP). The HTP option returns PTMs identified solely through mass spectrometry, whereas the LTP option returns PTMs identified by any experimental method. Further, for higher level of confidence it provides the option to show PTMs that were referenced more than five times.

Results from PhosphositePlus should be filtered to contain only phosphorylations. Providing that Tyrosine kinases are considered a distinct class of kinases (Pinna and Ruzzene, 1996), depending on the type of kinases the researcher is interested in Serines (S) and Threonines (T) or Tyrosines (Y) could be excluded from the results. Then, the experimentally observed phosphorylated residues should be saved in a .txt file in the following format (Figure 3):

phosphorylated_residue_1,
phosphorylated_residue_2,
phosphorylated_residue_3, etc.



```
S4, S159, T166, S181, T196, S210, S216, T258, S272, T1127, S1149, S1275, S1493,
T1529, S1541, S1559, S1605
```

Figure 3. The .txt file containing the experimentally observed phosphorylated residues in the protein, which is inserted as input to PhosphoKin. The experimentally observed phosphorylated residues in human laminin γ1-chain (LAMC1) are shown. These phosphorylated residues were taken from the PhosphoSitePlus database using the HTP option.

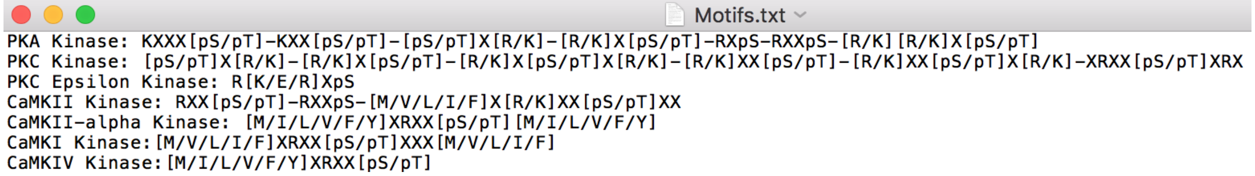
Prediction of phosphorylated and residues and their association with active sites

The prediction of a protein's phosphorylation sites should be implemented using kinase recognition motifs. To obtain the kinase recognition motifs the Phosphomotif Finder database (Amanchy et al., 2007) should be used. This database detects in protein queries literature-derived kinases motifs. The

Serine/Threonine motifs and the Kinase/Phosphotase motifs option should be chosen in the protein query. Phosphomotif Finder will narrow down the list of all known kinases to a few candidate kinases for the phosphorylation of the protein and will display the candidate proteins along with their recognition motifs.

If the researcher is interested only on kinases, the candidate phosphatases should be filtered out of the results. Moreover, some proteins consist of more than one chain, like laminin-111. For proteins with multiple chains, queries on Phosphomotif Finder should be made for all the protein's chains and the results should be merged before or after excluding the candidate phosphatases. The kinases along with their translated recognition motifs in regular expressions should be saved in a .txt file in the following format (Figure 4):

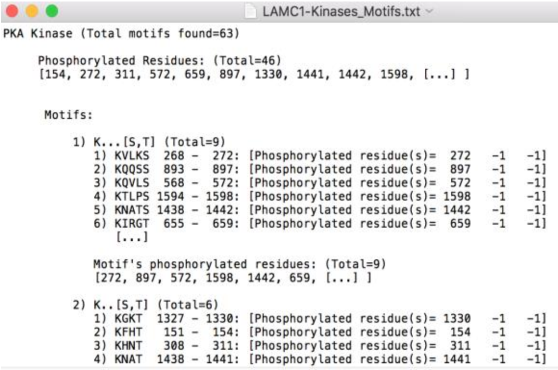
Name_of_kinase_1[space]protein type: motif_1, motif_2, motif_3, etc. [enter]
Name_of_kinase_2[space]protein type: motif_1,motif_2, etc. [enter]
etc.



```
PKA Kinase: KXXX[pS/pT]-KXX[pS/pT]-[pS/pT]X[R/K]-[R/K]X[pS/pT]-RXpS-RXXpS-[R/K][R/K]X[pS/pT]
PKC Kinase: [pS/pT]X[R/K]-[R/K]X[pS/pT]-[R/K]X[pS/pT]X[R/K]-[R/K]XX[pS/pT]-[R/K]XX[pS/pT]X[R/K]-XRXX[pS/pT]XRX
PKC Epsilon Kinase: R[K/E/R]XpS
CaMKII Kinase: RXX[pS/pT]-RXXpS-[M/V/L/I/F]X[R/K]XX[pS/pT]XX
CaMKII-alpha Kinase: [M/I/L/V/F/Y]XRXX[pS/pT][M/I/L/V/F/Y]
CaMKI Kinase:[M/V/L/I/F]XRXX[pS/pT]XXX[M/V/L/I/F]
CaMKIV Kinase:[M/I/L/V/F/Y]XRXX[pS/pT]
```

Figure 4. The .txt file containing the recognition motifs of candidate kinases for the phosphorylation of the protein, which is inserted as input to PhosphoKin. A few candidate kinases for the phosphorylation of the human laminin γ1-chain (LAMC1) along with their motifs are shown. The motifs of kinases were taken from the Phosphomotif Finder database using the Serine/Threonine motifs and the Kinase/Phosphotase motifs option.

The PhosphoKin tool (Methods "The PhosphoKin tool") reads the .txt file and translates the kinases' recognition motifs into regular expressions. Then, it uses the regular expressions to identify the exact subsequences and residues that could be recognized and phosphorylated, respectively, by each motif of each candidate kinase, in a protein sequence (Methods "Protein sequence"). The results are saved in an output .txt file, an example of which is displayed in Figure 5.



```
PKA Kinase (Total motifs found=63)

     Phosphorylated Residues: (Total=46)
     [154, 272, 311, 572, 659, 897, 1330, 1441, 1442, 1598, [...] ]


      Motifs:

          1) K...[S,T] (Total=9)
             1) KVLKS  268 -  272: [Phosphorylated residue(s)=  272   -1   -1]
             2) KQQSS  893 -  897: [Phosphorylated residue(s)=  897   -1   -1]
             3) KQVLS  568 -  572: [Phosphorylated residue(s)=  572   -1   -1]
             4) KTLPS 1594 - 1598: [Phosphorylated residue(s)= 1598   -1   -1]
             5) KNATS 1438 - 1442: [Phosphorylated residue(s)= 1442   -1   -1]
             6) KIRGT  655 -  659: [Phosphorylated residue(s)=  659   -1   -1]
                [...]

             Motif's phosphorylated residues: (Total=9)
             [272, 897, 572, 1598, 1442, 659, [...] ]

          2) K..[S,T] (Total=6)
             1) KGKT  1327 - 1330: [Phosphorylated residue(s)= 1330   -1   -1]
             2) KFHT   151 -  154: [Phosphorylated residue(s)=  154   -1   -1]
             3) KHNT   308 -  311: [Phosphorylated residue(s)=  311   -1   -1]
             4) KNAT  1438 - 1441: [Phosphorylated residue(s)= 1441   -1   -1]
```

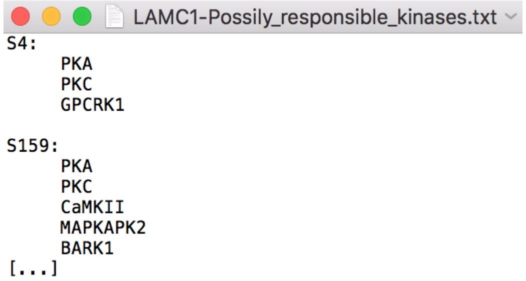
Figure 5. The output .txt file showing the exact protein subsequences and residues that can be bound and phosphorylated, respectively, by the candidate kinases (Figure 4) as predicted by PhosphoKin. As an example, a few phosphorylated residues and binding subsequences by PKA kinase in the human laminin γ1-chain (LAMC1) are presented (unpublished results). The "[…]" indicates that more data are available but not presented. Generally, the numbers refer to the position of the amino acids. The numbers in the Phosphorylated residue(s) show the residue that is predicted to be phosphorylated by the motif. As a rule in PhosphoKin, a motif can show the phosphorylation of only three or less residues. Therefore, the "-1" values in the Phosphorylated residue(s) indicate that there is not another residue predicted to be phosphorylated.

It should be noted that the number of the predicted phosphorylated residues might be much higher than the number of experimentally observed phosphorylated residues. The higher number of predicted phosphorylated residues could be attributed to the generality of kinase motifs. Nonetheless, it could reflect the phenomenon of "hierarchical phosphorylation", according to which a single experiment may not detect all nature occurring-protein phosphorylations since not all phosphorylations occur simultaneously in a protein at a given time (Roach, 1991). In addition, other factors, like the protein isolation tissue and parameters of the experimental identification method, could impact the number of predicted phosphorylations. Phospho-proteomics is a complex field and PTM's of many proteins still remain unknown.

However, it is important to stress that the PhosphoKin tool predicts phosphorylated residues based only on the linear sequence of a protein. The 3D structure of proteins may represent features that enable phosphorylation (Small et al., 1977). Thus, prediction of phosphorylated residues using the linear sequence could miss important phosphorylations for the function of the protein. Nevertheless, only a portion of human proteins have their 3D structure solved (by X-Ray Crystallography). Therefore, in the absence of a protein's 3D structure, motifs found in the linear sequence are the best predictive features.

Identification of possibly responsible kinases for the experimentally observed phosphorylated residues

The PhosphoKin tool reads the .txt file with the experimentally observed phosphorylated residues in the protein (Methods "Experimentally observed phosphorylated residues") and based on the predicted phosphorylation sites and phosphorylated residues for each candidate kinase (Methods "Prediction of phosphorylated and residues and their association with active sites"), it identifies the possibly responsible kinases for each experimentally observed phosphorylated residue in the protein. The results are saved in an output .txt file, an example of which is displayed in Figure 6.



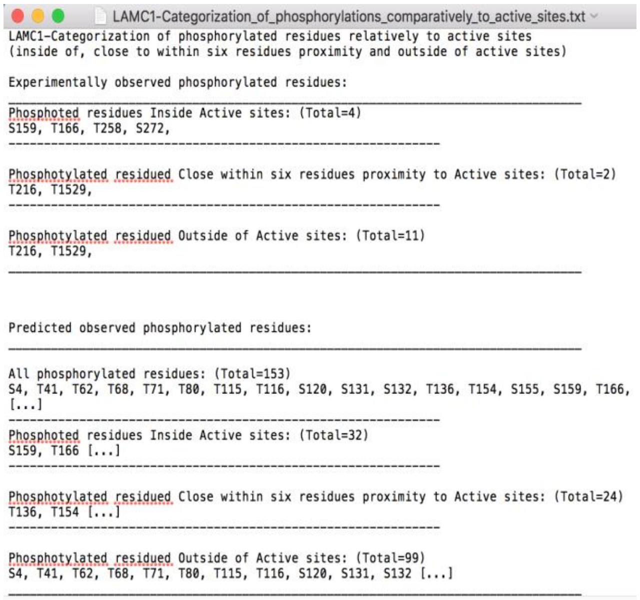
```
S4:
      PKA
      PKC
      GPCRK1

S159:
      PKA
      PKC
      CaMKII
      MAPKAPK2
      BARK1
[...]
```

Figure 6. The output .txt file showing the possibly responsible kinases for the experimentally observed phosphorylated residues of the protein, as predicted by PhosphoKin. As an example, the possibly responsible kinases for two experimentally observed phosphorylated residues in the human laminin γ1-chain (LAMC1) are shown (unpublished results). The "[…]" indicates that more data are available but not presented.

The possibly responsible kinase(s) for an experimentally observed phosphorylated residue in a protein are the kinase(s) that can theoretically bind to the surrounding area and catalyze the phosphorylation of this residue. Yet, the exact kinase(s) that catalyze this phosphorylation in nature remain in the scope of further research. Also, it is possible that this phosphorylation in nature could be catalyzed by a yet unknown kinase or a yet unknown kinase motif. Nonetheless, the PhosphoKin tool significantly narrows down the list of all known kinases to a few possible ones for the phosphorylation of each experimentally observed phosphorylated residue in a protein. Thus, it provides an essential aid to the design of more directed and lower-cost experiments.

Categorization of phosphorylated residues according to their location relatively to active sites and association of possibly responsible kinases with active sites

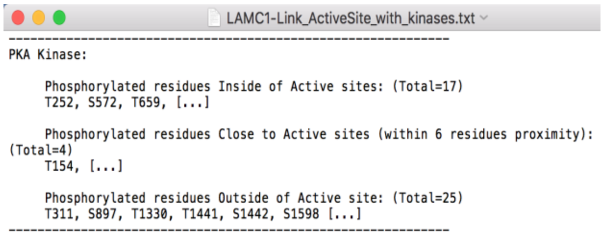
The PhosphoKin tool reads the .txt file with the active sites (Methods “Recording of protein’s active sites”) and categorizes the experimentally observed and predicted phosphorylated residues of all kinases according to their location relatively to active sites. The location categories are inside active sites, outside of and close to them within six residues proximity. The third category (close to active sites) was included in order to eliminate possible literature contradictions and losses derived from the identification method (commonly synthetic peptides) regarding the length of active sites. Further, due to the reasoning that phosphorylations to residues located close within a few residues proximity to an active site could significantly influence the function of the active site. The results are saved in an output .txt file, an example of which is demonstrated in Figure 7.


LAMC1-Categorization_of_phosphorylations_comparatively_to_active_sites.txt

```
LAMC1-Categorization of phosphorylated residues relatively to active sites
(inside of, close to within six residues proximity and outside of active sites)

Experimentally observed phosphorylated residues:
________________________________________________________________________
Phosphoted residues Inside Active sites: (Total=4)
S159, T166, T258, S272,
-------------------------------------------------------

Phosphotylated residued Close within six residues proximity to Active sites: (Total=2)
T216, T1529,
-------------------------------------------------------

Phosphotylated residued Outside of Active sites: (Total=11)
T216, T1529,
________________________________________________________________________


Predicted observed phosphorylated residues:
________________________________________________________________________

All phosphorylated residues: (Total=153)
S4, T41, T62, T68, T71, T80, T115, T116, S120, S131, S132, T136, T154, S155, S159, T166,
[...]
-------------------------------------------------------
Phosphoted residues Inside Active sites: (Total=32)
S159, T166 [...]
-------------------------------------------------------

Phosphotylated residued Close within six residues proximity to Active sites: (Total=24)
T136, T154 [...]
-------------------------------------------------------

Phosphotylated residued Outside of Active sites: (Total=99)
S4, T41, T62, T68, T71, T80, T115, T116, S120, S131, S132 [...]
________________________________________________________________________
```

Figure 7. The output .txt file of PhosphoKin, showing the categorization of the phosphorylated residues in a protein according to their location relatively to active sites (inside, outside, close within six residues proximity). As an example, the categorization of the experimentally observed and predicted phosphorylated residues in the human laminin γ1-chain (LAMC1) is shown (unpublished results). The “[…]” indicated that more data are available but not presented.

Additionally, the PhosphoKin tool associates the phosphorylation activity of each kinase with the active sites of the protein based on the categorization of phosphorylated residues according to their location relatively to active sites. The results are saved in an output .txt file, an example of which is demonstrated in Figure 8.



```
------------------------------------------------------------
PKA Kinase:

     Phosphorylated residues Inside of Active sites: (Total=17)
     T252, S572, T659, [...]

     Phosphorylated residues Close to Active sites (within 6 residues proximity):
(Total=4)
     T154, [...]

     Phosphorylated residues Outside of Active site: (Total=25)
     T311, S897, T1330, T1441, S1442, S1598 [...]
------------------------------------------------------------
```

Figure 8. The output .txt file of PhosphoKin that links the examined kinases with the protein’s active sites. As an example, the predicted phosphorylated residues by PKA kinase in the human laminin γ1-chain (LAMC1) is shown (unpublished results). The “[…]” indicated that more data are available but not presented. The phosphorylated residues of PKA kinase are categorized according to their location relatively to active sites (inside, outside and close). This categorization suggest whether the activity of the kinase is more oriented inside, outside or close to the protein’s active sites.

The PhosphoKin tool

The tool in written in python. It uses as input four .txt files; one containing the protein sequence (Methods “Protein sequence”), one containing the protein’s active sites (Methods “Recording of LAMC1 Active Sites”), one containing the experimentally assigned phosphorylated residues in the protein (Methods “Experimentally observed phosphorylated residues”) and one containing the kinases along with their recognition motifs (Methods "Prediction of phosphorylated and residues and their association with active sites"). As output the tool produces four .txt files; one for the prediction of phosphorylation sites and phosphorylated residues in the protein (Methods "Prediction of phosphorylated and residues and their association with active sites"), one for the identification of possibly responsible kinases for the experimentally observed phosphorylated residues in the protein (Methods "Identification of possibly responsible kinases for the experimentally observed phosphorylated residues"), one for the categorization of phosphorylated residues according to their location relatively to active sites (Methods "Categorization of phosphorylated residues according to their location relatively to active sites and association of possibly responsible kinases with active sites") and one for the association of kinases with active sites (Methods "Categorization of phosphorylated residues according to their location relatively to active sites and association of possibly responsible kinases with active sites"). The tool, which can be found here: https://github.com/AngelikiGal/PhosphoKin, should be downloaded and run in the terminal by the following command: "python3 PhosphoKin.py".

However, the tool has some limitations. The four input files should have a specific format (Figure 1-4 and Methods "Protein sequence", "Recording of LAMC1 Active Sites", "Experimentally observed phosphorylated residues" and "Prediction of phosphorylated and residues and their association with active sites") in order to be correctly processed by the tool. Also, the tool accepts only motifs that indicate three or less residues for phosphorylation.

Mapping of LAMC1 phosphorylation sites

An overview of the results could derive from illustrating them all together in a phosphorylation map of the protein. The phosphorylation map should contain the protein sequence in rows of sixty residues. The active sites, as well as the experimentally observed phosphorylated residues and the predicted phosphorylation sites, should be marked in the protein sequence using specific annotation. The active sites should be marked in bold font-weight, the experimentally observed phosphorylated residues in red coloring and the predicted phosphorylation sites in yellow highlight. To display the interaction of each kinase with residues of the protein sequence, a list of kinases should be presented under each protein sequence row. The list of kinases should contain all the kinases that interact with residues of the above protein sequence row and the interaction of each kinase should be attributed by a line in the list of kinases. The binding to a residue should be displayed with the "X" letter and the phosphorylation of a residue with the "P" letter. On the contrary, the absence of an interaction with a residue should be indicated with the "-" character. An example of a phosphorylation map is presented in Figure S1.

How to analyze the results

The total number of the kinases' phosphorylated residues should be examined both for the experimentally observed and predicted phosphorylations, while more emphasis should be given on the former for conducting conclusions. The kinases could be grouped based on their total phosphorylated residues in order to speculate the role of each kinase in the phosphorylation of the protein. The kinases with the higher total phosphorylated residues (top kinases) are more likely to contribute to the protein's regulation via phosphorylation. Additionally, the top kinases in the experimentally observed phosphorylated residues should be compared to the top kinases in the predicted phosphorylations. This comparison will strengthen the importance of the top kinases in the phosphorylation of the protein as they are expected to be the same in both categories.

Furthermore, the location of the top kinases' phosphorylated residues should be investigated relatively to active sites (inside, outside and close). This, should be examined for both experimentally observed and predicted phosphorylated residues and will reveal whether the phosphorylation activity of a kinase is more oriented inside, outside or close to active sites. Phosphorylations that are located inside of as well as close within a few residues proximity to active sites are more likely to regulate the function of an active site and thus, the function of the protein. However, a top kinase with an activity that is oriented more outside the active sites could still greatly impact the regulation of the protein.

Moreover, examining whether all the possibly responsible kinases for an experimentally observed phosphorylated residue located inside or close to an active site, belong to the same group of kinases (e.g. ecto-kinases), could lead to associations between the group of kinases and the active site. The higher the number of the experimentally observed phosphorylated residues inside and close to an active site, for which their possibly responsible kinases fall under the same kinases' group, the stronger the association is between the group of kinases and the active site.

Similarly, exploring the association between a certain group of kinases and a group of active sites, such as important active sites in diseases, could yield very interesting results. For instance, inside the active site "RPESFAIYKRTR" in LAMC1, which binds to cancer cell lines (Nomizu et al., 1997), the residues S159 and T166 were experimentally found to be phosphorylated (Rush, 2008). CKII, PKA and PKC, which are known for an ecto-phosphorylation activity (Bohana-Kashtan et al., 2005; Hogan et al., 1995; Kondrashin et al., 1999), were the only possibly responsible kinases for the phosphorylation of T166, while PKA and PKC were among the possibly responsible kinases for the phosphorylation of S159. Therefore, the activity of PKA and PKC was associated with the active site "RPESFAIYKRTR" (unpublished data).

Conclusions

Our method enhances the understanding of the role of phosphorylation in a protein's regulation by combining the active sites with the phosphorylated residues (experimentally observed and predicted). Moreover, this method aids the design of more directed and lower-cost experiments by identifying the possibly responsible kinases of the experimentally observed phosphorylated residues and by predicting new phosphorylation sites and phosphorylated residues. Furthermore, it gives ground for associations between specific active sites and group of kinases as well as suggests the illustration of a detailed and helpful protein's phosphorylation map. However, further research *in vitro* is needed to demonstrate the exact kinases catalyzing the experimentally observed phosphorylated residues, the importance of the top kinases in the protein's regulation and any active sites-group of kinases association.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. Journal of Molecular Biology 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amanchy, R., Periaswamy, B., Mathivanan, S., Reddy, R., Tattikota, S.G., Pandey, A., 2007. A curated compendium of phosphorylation motifs. Nature Biotechnology 25, 285–286. https://doi.org/10.1038/nbt0307-285

Ashcroft, M., Kubbutat, M.H.G., Vousden, K.H., 1999. Regulation of p53 Function and Stability by Phosphorylation. Molecular and Cellular Biology 19, 1751–1758. https://doi.org/10.1128/MCB.19.3.1751

Berthet, J., Rall, T.W., Sutherland, E.W., 1957. The relationship of epinephrine and glucagon to liver phosphorylase. IV. Effect of epinephrine and glucagon on the reactivation of phosphorylase in liver homogenates. J. Biol. Chem. 224, 463–475

Bohana-Kashtan, O., Pinna, L.A., Fishelson, Z., 2005. Extracellular phosphorylation of C9 by protein kinase CK2 regulates complement-mediated lysis. European Journal of Immunology 35, 1939–1948. https://doi.org/10.1002/eji.200425716

Canese, K., Weis, S., n.d. PubMed: The Bibliographic Database. 2002 Oct 9 [Updated 2013 Mar 20]. In: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013-.

Cori, G.T., Green, A.A., 1943. CRYSTALLINE MUSCLE PHOSPHORYLASE: II. PROSTHETIC GROUP. J. Biol. Chem 151, 1–38

Han, X., Aslanian, A., Yates, J.R., 2008. Mass spectrometry for proteomics. Current Opinion in Chemical Biology 12, 483–490. https://doi.org/10.1016/j.cbpa.2008.07.024

Hogan, M.V., Pawlowska, Z., Yang, H.A., Kornecki, E., Ehrlich, Y.H., 1995. Surface phosphorylation by ecto-protein kinase C in brain neurons: a target for Alzheimer’s beta-amyloid peptides. J. Neurochem. 65, 2022–2030

Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., Skrzypek, E., 2015. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Research 43, D512–D520. https://doi.org/10.1093/nar/gku1267

Khoury, G.A., Baliban, R.C., Floudas, C.A., 2011. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. Scientific Reports 1. https://doi.org/10.1038/srep00090

Kondrashin, A., Nesterova, M., Cho-Chung, Y.S., 1999. Cyclic Adenosine 3‘:5‘-Monophosphate-Dependent Protein Kinase on the External Surface of LS-174T Human Colon Carcinoma Cells. Biochemistry 38, 172–179. https://doi.org/10.1021/bi982090e

Krebs, E.G., Fischer, E.H., 1956. The phosphorylase b to a converting enzyme of rabbit skeletal muscle. Biochim. Biophys. Acta 20, 150–157

Krebs, E.G., Graves, D.J., Fischer, E.H., 1959. Factors affecting the activity of muscle phosphorylase b kinase. J. Biol. Chem. 234, 2867–2873

Nomizu, M., Kuratomi, Y., Song, S.-Y., Ponce, M.L., Hoffman, M.P., Powell, S.K., Miyoshi, K., Otaka, A., Kleinman, H.K., Yamada, Y., 1997. Identification of Cell Binding Sequences in Mouse Laminin γ1 Chain by Systematic Peptide Screening. Journal of Biological Chemistry 272, 32198–32205. https://doi.org/10.1074/jbc.272.51.32198

Nørregaard Jensen, O., 2004. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. Current Opinion in Chemical Biology 8, 33–41. https://doi.org/10.1016/j.cbpa.2003.12.009

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research 44, D733–D745. https://doi.org/10.1093/nar/gkv1189

Olsen, J.V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., Mann, M., 2006. Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks. Cell 127, 635–648. https://doi.org/10.1016/j.cell.2006.09.026

Pinna, L.A., Ruzzene, M., 1996. How do protein kinases recognize their substrates? Biochimica et Biophysica Acta (BBA) - Molecular Cell Research 1314, 191–225. https://doi.org/10.1016/S0167-4889(96)00083-3

Pruitt, K.D., Harrow, J., Harte, R.A., Wallin, C., Diekhans, M., Maglott, D.R., Searle, S., Farrell, C.M., Loveland, J.E., Ruef, B.J., Hart, E., Suner, M.-M., Landrum, M.J., Aken, B., Ayling, S., Baertsch, R., Fernandez-Banet, J., Cherry, J.L., Curwen, V., DiCuccio, M., Kellis, M., Lee, J., Lin, M.F., Schuster, M., Shkeda, A., Amid, C., Brown, G., Dukhanina, O., Frankish, A., Hart, J., Maidak, B.L., Mudge, J., Murphy, M.R., Murphy, T., Rajan, J., Rajput, B., Riddick, L.D., Snow, C., Steward, C., Webb, D., Weber, J.A., Wilming, L., Wu, W., Birney, E., Haussler, D., Hubbard, T., Ostell, J., Durbin, R., Lipman, D., 2009. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Research 19, 1316–1323. https://doi.org/10.1101/gr.080531.108

Roach, P.J., 1991. Multisite and Hierarchal Protein Phosphorylation. J Biol Chem 266, 14139–14142.

Rush, J., 2008. CST Curation Set: 4488; Year: 2008; Biosample/Treatment: cell line, MKN-45/untreated &'||' normal; Disease: gastric carcinoma; SILAC: -; Specificities of Antibodies Used to Purify Peptides prior to LCMS: RXXp[ST] Antibodies Used to Purify Peptides prior to LCMS: Phospho-Akt Substrate (RXRXXS/T) (110B7) Rabbit mAb Cat#: 9614, PTMScan(R) Phospho-Akt Substrate Motif (RXXS*/T*) Immunoaffinity Beads Cat#: 1978

Sariahmetoglu, M., Crawford, B.D., Leon, H., Sawicka, J., Li, L., Ballermann, B.J., Holmes, C., Berthiaume, L.G., Holt, A., Sawicki, G., Schulz, R., 2007. Regulation of matrix metalloproteinase-2 (MMP-2) activity by phosphorylation. The FASEB Journal 21, 2486–2495. https://doi.org/10.1096/fj.06-7938com

Small, D., Chou, P.Y., Fasman, G.D., 1977. Occurrence of phosphorylated residues in predicted β-turns: Implications for β-turn participation in control mechanisms. Biochemical and Biophysical Research Communications 79, 341–346. https://doi.org/10.1016/0006-291X(77)90101-2

Sutherland, E.W., Wosilait, W.D., 1955. Inactivation and Activation of Liver Phosphorylase. Nature 175, 169–170. https://doi.org/10.1038/175169a0

The UniProt Consortium, 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Research 45, D158–D169. https://doi.org/10.1093/nar/gkw1099

Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O.G., Janacek, S.H., Juettemann, T., To, J.K., Laird, M.R., Lavidas, I., Liu, Z., Loveland, J.E., Maurel, T., McLaren, W., Moore, B., Mudge, J., Murphy, D.N., Newman, V., Nuhn, M., Ogeh, D., Ong, C.K., Parker, A., Patricio, M., Riat, H.S., Schuilenburg, H., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Zadissa, A., Frankish, A., Hunt, S.E., Kostadima, M., Langridge, N., Martin, F.J., Muffato, M., Perry, E., Ruffier, M., Staines, D.M., Trevanion, S.J., Aken, B.L., Cunningham, F., Yates, A., Flicek, P., 2018. Ensembl 2018. Nucleic Acids Research 46, D754–D761. https://doi.org/10.1093/nar/gkx1098