Review

Protein phosphorylation prediction: limitations, merits and pitfalls

Dimitrios Vlachakis[1], Elena Bencurova[2], Louis Papageorgiou[1,3], Mangesh Bhide[2,4] and Sophia Kossida[5]

[1]Bioinformatics & Medical Informatics Team, Biomedical Research Foundation, Academy of Athens, Soranou Efessiou 4, Athens 11527, Greece
[2]Laboratory of Biomedical Microbiology and Immunology, Department of Microbiology and Immunology, University of Veterinary Medicine and Pharmacy, Komenskeho 73, 04181, Kosice, Slovakia
[3]Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, University Campus, Athens, 15784, Greece
[4]Institute of Neuroimmunology, Slovak Academy of Sciences, 84245 Bratislava, Slovakia
[5]IMGT®, the international ImMunoGeneTics information system®, Institute of Human Genetics, Montpellier, France

Received on June 2, 2015; Accepted on June 29, 2015; Published on July 18, 2015

Correspondence should be addressed to Dimitrios Vlachakis; Tel: +30 210 6597 647, Fax: +30 210 6597 545, E-mail: dvlachakis@bioacademy.gr

Abstract

Protein phosphorylation is a major protein post-translational modification process that plays a pivotal role in numerous cellular processes, such as recognition, signaling or degradation. It can be studied experimentally by various methodologies, including western blot analysis, site-directed mutagenesis, 2D gel electrophoresis, mass spectrometry etc. A number of *in silico* tools have also been developed in order to predict plausible phosphorylation sites in a given protein. In this review, we conducted a benchmark study including the leading protein phosphorylation prediction software, in an effort to determine which performs best. The first place was taken by GPS 2.2, having predicted all phosphorylation sites with a 83% fidelity while in second place came NetPhos 2.0 with 69%.

Protein Phosphorylation

Protein phosphorylation is a major post-translational modification, illustrating a major cellular reversible process that is performed primarily by the protein kinases (PKs). It directs a variety of biological cellular processes, including transduction and cellular cycle regulation (Suter *et al.* 2008). Biochemically, PKs play a major role by catalyzing the hydrolysis of adenosine triphosphate (ATP), which in turn, transfers a phosphate group to the appropriate residue (serine (S) / threonine (T) or tyrosine (Y) in eukaryotic organisms, and histidine (H), arginine (Arg) or lysine (K) in prokaryotes. Most importantly, PKs modify a specifically defined subset of substrates, in this way ensuring the signaling fidelity (PK-specific) of the process (Ciesla *et al.* 2011).

Phosphorylation plays a crucial role in cellular regulation, immune response, signaling and energy management of living organisms. Cells communicate with each other and interact with their environment through various signals. These signals represent either mechanical or chemical stimuli, with the latter produced by autocrine, endocrine or paracrine mechanisms. Approximately 2% of the human genome encodes more than five hundred PK genes. Each PK exhibits distinct recognition properties, including short linear motifs (SLMs) flanking the phosphorylation sites (P-sites) that are responsible for attributing primary specificity (Song *et al.* 2012).

The eukaryotic organisms frequently prefer to phosphorylate serine rather than threonine residues, so tyrosine phosphorylation rarely occurs in eukaryotes. On the other hand, histidine phosphorylation constitutes an inherent part of signal transduction within intracellular signaling pathways. However, their frequency is relatively low and occurs in less that 10% of the total transduction events in eukaryotic cells. In all cases, each residue-specific PK acts as regulatory switch by adding one or more phosphate groups to

Table 1. Examples of phosphorylated amino acid residues and their function.

Amino acid (Physicochemical properties)	Single letter code	Function	Information	References
Serine (Aliphatic and polar groups)	S	Biosynthesis of purines and pyrimidines and other metabolites Example: The serine 727 which is located in the amino acid sequence of protein STAT1 of STAT proteins, is phosphorylated by a phosphorylating kinase. The stimulus is an INF-γ and the pathways which are triggered by this stimulus are JAK2-dependent, RAS-independent. The result from these pathways is over-expression of dominant-negative and constitutively active Ras.	It is known that the STAT signal transduction factors and activators of transcription require serine phosphorylation by bSTAT serine kinase to their C-terminus, before activation. Prior to this, a tyrosine residue phosphorylation occurs in cytokine-stimulated cells by the receptor-associated Janus Kinases (JAKs), contributing to STATS' dimerization. These reactions are necessary for the activation of the well known JAK-STAT signaling pathway	(Decker & Kovarik 2000)
Threonine (Aliphatic and polar groups)	T	Isoleucine precursor Related Diseases: Irritability, difficult personality	Threonine phosphorylation occurs in the human epidermal growth factor (EGF) receptor. Threonine is located in a very basic sequence of 9 residues of the cytoplasmic area of the plasma membrane and is located in the area near the kinase. Its location helps the phosphorylation and consequently the modification of signaling between the inner region and the external EGF-binding area.	(Hunter *et al.* 1984)
Tyrosine (Aromatic side chains)	Y	Signal transduction processes *Tyrosine hydroxylase -> levodopa *Tyrosine-> Thyroid hormones Related Diseases: brain neural problems	A representative example of tyrosine phosphorylation occurs in the erythropoietin receptor (EPOR). Erythropoietin (EPO) is a glycoprotein hormone that regulates erythropoiesis, through interactions with the EPOR receptor. Tyrosine phosphorylated EPOR triggers the JAK/STAT5 signaling cascade and is related to gene transcription and mitogenesis.	(Witthuhn *et al.* 1993)
Histidine (Basic side chains)	H	Histamine precursor, carbon atoms-source in purines	Histidine phosphorylation occurs in several platelet proteins and it is necessary for the platelet activation. For example P-selectin is phosphorylated in a cytoplasmatic tail after platelets activation by thrombin and collagen. The stimulation by thrombin increase the kinetics of phosphohistidine and disappearance of P-selectin is very fast. Activated platelets are exhibiting high production of phosphohistidine. This situation shows the induction of rapid and reversible phosphorylation of histidine in mammalian cells, after the activation of the cells, a situation that concerns the cell signaling by a protein histidine kinase.	(Crovello *et al.*1995, Wolanin *et al.* 2002)

them. Phosphorylation activity is also detected in cyclins and cyclin-dependent kinases (Cdks), which constitute key regulators of the cell cycle progression in eukaryotic cells (Masumoto *et al.* 2002). It is known that Cdk activity is detected by phosphorylation at three conserved positions (Lew & Kornbluth 1996). Another example is the Bcl-2 phosphorylation, which regulates cell apoptosis (Ruvolo *et al.* 2001). Table 1 summarizes some examples of phosphorylated amino acid residues and their function.

Detection of phosphorylated points with biological techniques

The most common methods for detecting and characterizing phosphorylated residues include experimental approaches supported mainly by western blot analysis and site-directed mutagenesis. Nevertheless, such experimental approaches are usually limited to specific tissues or cells and are time consuming. Based on new technologies, the leading techniques for the identification of phosphorylated sites became the high-throughput methods, such as proteomics and analysis by mass spectrometry (St-Denis & Gingras 2012). The mass spectrometry method can be utilized to determine the phosphorylated sites in a wide variety of tissues. However, it suffers from certain limitations and disadvantages. For example, the identification of kinases responsible for the phosphorylation catalysis is limited due to sensitivity. In addition, a number of important proteins cannot be detected by this technique due to their low abundance. Furthermore, many phosphorylated sites are changed to hypo-stoichiometrical levels, which usually prevent their detection. In general, this technology requires very expensive instruments and high levels of expertise, not always available (Sundstrom *et al.* 2009).

Another high-throughput approach is two-dimensional gel electrophoresis (2-DE), which can be used to separate protein mixtures and detect phosphorylation changes. This approach was successfully used for the identification of several phosphorproteins related to the extracellular signal-regulated kinase (ERK) pathway (Lovric *et al.* 1998).

More advanced techniques for the detection of phosphorylated sites are the protein microarrays or protein chips (Zhu *et al.* 2001). New immunoassay techniques can also be used by high throughput approaches, mainly based on the use of phosphor-specific monoclonal antibodies that have been developed against different phosphorylated amino acids (Leitner *et al.* 2011).

In addition, down regulating or knocking out a target kinase in vitro and observing the resulting phenotype is another way to identify substrates. This methodology has been used in small- as well as large-scale studies (MacKeigan *et al.* 2005).

Bioinformatics phosphorylation tools

The use of bioinformatics is one of the most used techniques for detecting phosphorylation due to its ability to eliminate the disadvantages of the above techniques, as it is based on methodology that relies on computational approaches (Table 2). For example, the method that is based on bayesian probability is more expressive than PSSMs, but is more easily interpreted biologically and mathematically than ANNs. These bioinformatics tools also use other information, which is based on whether or not to use the information structure. Finally, the tools also stand out from their specificity, if they are non-kinase or kinase-specific tools.

Table 2. Phosphorylation detection tools together with the corresponding machine learning technique they employ, the number of phosphorylated residues and the sequence structural information. The K-spec/No-spec column indicates whether the tools are kinase or non-kinase specific.

Tool	Machine learning technique	Number of phosylated residues for each tool	1D/3D Sequence/ structural info	K-spec/No-spec
NetPhos	ANN	9-33	3D	No-spec
NetPhosK	ANN	9-33	3D	K-spec
PHOSIDA	SVM	13	1D	No-spec
Musite	SVM	Exact range of lengths not explicitly stated	1D	K-spec
ScanSite	PSSM	15	1D	K-spec
SMALI	PSSM	7	1D	K-spec
GPS 1.0	PSSM,Markov Clustering	7	1D	K-spec
PPSP	BP	9	1D	K-spec

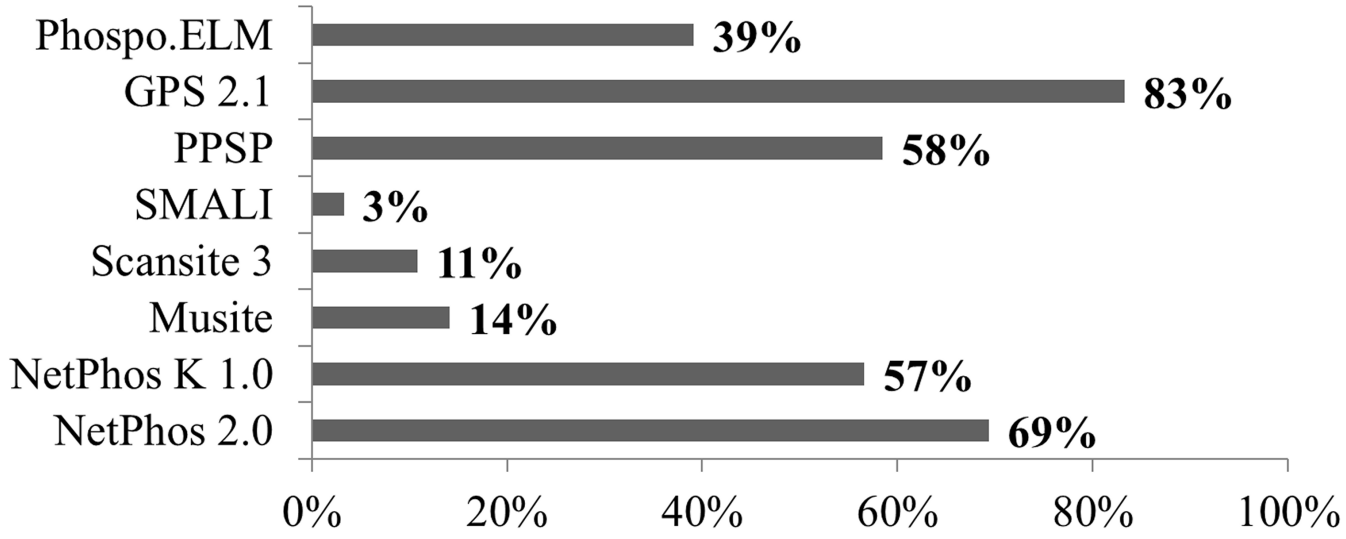

Figure 1. Graphical representation of positive prediction percentages of phosphorylated sites by selected software.

In other word, the tool makes provisions for specific kinases or kinase families or is not kinase-specific (Trost & Kusalik 2011).

For optimal results, experimental techniques are often facilitated by the simultaneous use of bioinformatics tools. For example, extensive computational analysis is needed before performing phosphor-peptide identification by mass spectrometry, due to the complexity of the latter. A number of software packages can be used for this step including Mascot (Trost & Kusalik 2011), SEQUEST (Yates *et al.* 1995), OMSSA (Geer *et al.* 2004), X! Tandem (Craig & Beavis 2004), GutenTag (Tabb *et al.* 2003), InsPecT (Tanner *et al.* 2005) and Spectral Networks Analysis (Bandeira 2011).

One of the problems observed in predicting phosphorylation sites is related to sensitivity and specificity. Phosphorylation prediction appears to be more sensitive when the detected regions are located in a single protein, whereas higher specificity appears when detected areas are in an entire proteome.

Benchmark of state of the art, current bioinformatics tools

In this study, a series of current state-of-the-art phosphorylation prediction tools were investigated and benchmarked in regards to their accuracy in detecting actually phosphorylated amino acids. In an effort to use a wide repertoire of test proteins the RCSB-PDB database was harvested for phosphorylated structures of proteins that have been determined by X-ray crystallography at low resolution (i.e. high fidelity). More specifically we used the proteins with accession numbers: E0J4T6, E8VA72, O15530, O34507, O34824, O95997, P04049, P04083, P04792, P0A5N2, P0A6N2, P0A763, P10636, P13796, P18159, P23528, P29320, P30307, P31103, P31120, P31751, P35568, P37840, P41685, P49841, P51593, P51636, P55008, P55211, P61012, P62753, P65728, P80885, P95078, Q00969, Q02750, Q06752, Q12778, Q12968, Q13541, Q16236, Q5S007, Q61083, Q62074, Q64010, Q6J1J1, Q6P2N0, Q8BZ03, Q8HXW5, Q93V58, Q95207, Q9H2X6, Q9MZA9, Q9UD71, Q9UMF0, 2VX3, 1U54, 1T15, 2ERK and 2IVV.

The phosphate groups on the selected crystal structures have been co-crystallized alongside the main protein crystal. All phosphorylated residues in the selected structures (Supplementary Table 1) confirm that these amino acids are capable of being phosphorylated under the right circumstances. Non-phosphorylated residues could either be unable to be phosphorylated or were just unable to get phosphorylated under the given experimental conditions. Therefore, our benchmark mainly focuses on the ability of each software package to accurately predict the residues that have been experimentally shown to be phosphorylated in the crystal structure.

All major phosphorylating software programs were examined; namely NetPhos 2.0 (Blom *et al.* 1999), NetPhosK 1.0 (Blom *et al.* 2004), Musite.net (Gao *et al.* 2010), ScanSite (Obenauer *et al.* 2003), SMALI (Li *et al.* 2008), PPSP (Xue *et al.* 2006), GPS 1.10 (Xue *et al.* 2005, 2008, Zhou *et al.* 2004) and Phospho.ELM (Dinkel *et al.* 2011). The raw data output files from the above programs are included in the supplementary data. A table summarizing the findings of this benchmark has also been generated (Supplementary Table 1).

It was found that each software comes with its

strengths and weaknesses. Some are better at detecting serine phosphorylation, whereas some are more suitable for correctly predicting Tyrosine or Threonine phosphorylation. The actual phosphorylated residues and the programs that correctly predicted each particular phosphorylation *in silico* are summarized in Supplementary Table 1.

Collectively, it was found that GPS 2.2 was the most accurate phosphorylation prediction package. NetPhos 2.0 came in second place, having succeeded in 147 out of 212 phosphorylation sites. PPSP (124 correct predictions) and NetPhosK 1.0 (120 correct predictions) came in third place, while Phospo.ELM showed a 39% successful prediction of phosphorylated sites. Musite and ScanSite3 performed quite average having predicted only 30 and 23 out of 212 phosphorylation sites, respectively. Finally SMALI proved to be quite poor in its prediction potential, as it failed almost completely to predict phosphorylation sites in our benchmark, with only 7 predictions that represent only a 3% match with the real data (Figure 1).

Conclusions

Protein phosphorylation is one of the most important post-translational modifications that proteins undergo. Many biological functions, such as recognition, signaling and degradation are linked to signals that arrive through protein phosphorylation. In this regard, a series of *in silico* tools have been developed to help scientists predict plausible phosphorylation sites on a given protein. Herein, a benchmark was conducted amongst the leading protein phosphorylation prediction software, in an effort to determine which tool performs best. Conclusively, the best prediction tool for protein phosphorylation was found to be GPS 2.2, having predicted all phosphorylation sites with an 83% fidelity. NetPhos 2.0 came in second place, while PPSP and NetPhosK 1.0 were found to perform reasonably well with an approximately 57% prediction potential in our benchmark.

Acknowledgements

The authors are grateful to Georgia-Ioanna Kartalou, Nikitas Papangelopoulos and Spyridon Champeris Tsaniras from Bioinformatics & Medical Informatics Team, Biomedical Research Foundation, Academy of Athens and Maria G Roubelakis from Gene Therapy Laboratory, Biomedical Research Foundation, Academy of Athens for the help with experimental work.

References

Bandeira N 2011 Protein identification by spectral networks analysis. *Methods Mol Biol* **694** 151-168

Blom, N, Gammeltoft S & Brunak S 1999 Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol* **294** 1351-1362

Blom N, Sicheritz-Pontén T, Gupta R, Gammeltoft S & Brunak S 2004 Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* **4** 1633-1649

Ciesla J, Fraczyk T & Rode W 2011 Phosphorylation of basic amino acid residues in proteins: important but easily missed. *Acta Biochim Pol* **58** 137-148

Craig R & Beavis RC 2004 TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20** 1466-1467

Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ & Diella F 2011 Phospho.ELM: a database of phosphorylation sites--update 2011. *Nucleic Acids Res* **39** D261-267

Gao J, Thelen JJ, Dunker AK & Xu D 2010 Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Mol Cell Proteomics* **9** 2586-2600

Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W & Bryant SH 2004 Open mass spectrometry search algorithm. *J Proteome Res* **3** 958-964

Leitner A, Sturm M & Lindner W 2011 Tools for analyzing the phosphoproteome and other phosphorylated biomolecules: a review. *Anal Chim Acta* **703** 19-30

Lew DJ & Kornbluth S 1996 Regulatory roles of cyclin dependent kinase phosphorylation in cell cycle control. *Curr Opin Cell Biol* **8** 795-804

Li L, Wu C, Huang H, Zhang K, Gan J & Li SS 2008 Prediction of phosphotyrosine signaling networks using a scoring matrix-assisted ligand identification approach. *Nucleic Acids Res* **36** 3263-3273

Lovrić J, Dammeier S, Kieser A, Mischak H & Kolch W 1998 Activated raf induces the hyperphosphorylation of stathmin and the reorganization of the microtubule network. *J Biol Chem* **273** 22848-22855

MacKeigan JP, Murphy LO & Blenis J 2005 Sensitized RNAi screen of human kinases and phosphatases identifies new regulators of apoptosis and chemoresistance. *Nat Cell Biol* **7** 591-600

Masumoto H, Muramatsu S, Kamimura Y & Araki H 2002 S-Cdk-dependent phosphorylation of Sld2 essential for chromosomal DNA replication in budding yeast. *Nature* **415** 651-655

Obenauer JC, Cantley LC & Yaffe MB 2003 Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*

31 3635-3641
Ruvolo PP, Deng X & May WS 2001 Phosphorylation of Bcl2 and regulation of apoptosis. *Leukemia* **15** 515-522
Song C, Ye M, Liu Z, Cheng H, Jiang X, Han G, Songyang Z, Tan Y, Wang H, Ren J, Xue Y & Zou H 2012 Systematic analysis of protein phosphorylation networks from phosphoproteomic data. *Mol Cell Proteomics* **11** 1070-1083
St-Denis N & Gingras AC 2012 Mass spectrometric tools for systematic analysis of protein phosphorylation. *Prog Mol Biol Transl Sci* **106** 3-32
Sundstrom JM, Sundstrom CJ, Sundstrom SA, Fort PE, Rauscher RL, Gardner TW & Antonetti DA 2009 Phosphorylation site mapping of endogenous proteins: a combined MS and bioinformatics approach. *J Proteome Res* **8** 798-807
Suter B, Graham C & Stagljar I 2008 Exploring protein phosphorylation in response to DNA damage using differentially tagged yeast arrays. *Biotechniques* **45** 581-584
Tabb DL, Saraf A & Yates JR 3rd 2003 GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal Chem* **75** 6415-6421
Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA & Bafna V 2005 InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem* **77** 4626-4639
Xue Y, Li A, Wang L, Feng H & Yao X 2006 PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinformatics* **7** 163
Xue Y, Ren J, Gao X, Jin C, Wen L & Yao X 2008 GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Mol Cell Proteomics* **7** 1598-1608
Xue Y, Zhou F, Zhu M, Ahmed K, Chen G & Yao X 2005 GPS: a comprehensive www server for phosphorylation sites prediction. *Nucleic Acids Res* **33** W184-187
Yates JR 3rd, Eng JK, McCormack AL & Schieltz D 1995 Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem* **67** 1426-1436
Zhou FF, Xue Y, Chen GL & Yao X 2004 GPS: a novel group-based phosphorylation predicting and scoring method. *Biochem Biophys Res Commun* **325** 1443-1448
Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, Mitchell T, Miller P, Dean RA, Gerstein M & Snyder M 2001 Global analysis of protein activities using proteome chips. *Science* **293** 2101-2105