Integrating Bioinformatics Tools to Investigate Protein Phosphorylation

Dimitrios Vlachakis, Elena Bencurova, Mangesh Bhide, Sophia Kossida

Protein phosphorylation is one of the most important protein post-translational modifications and plays a role in numerous cellular processes including recognition, signaling and degradation. It can be studied experimentally by various methodologies, like employing western blot analysis, site-directed mutagenesis, 2 D 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%.

PeerJ PrePrints

Integrating Bioinformatics Tools to Investigate Protein Phosphorylation

Dimitrios Vlachakis[1], Elena Bencurova[2], Mangesh Bhide[2,3], Sophia Kossida[1,4*]

[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]Institute of Neuroimmunology, Slovak Academy of Sciences, 84245 Bratislava, Slovakia
[4]IMGT®, the international ImMunoGeneTics information system®, Institute of Human Genetics, Montpellier, France

*Correspondence to: Sophia Kossida,
Bioinformatics & Medical Informatics Team, Biomedical Research Foundation, Academy of Athens, Soranou Efessiou 4, Athens 11527, Greece
Tel: + 30 210 6597 199, Fax: +30 210 6597 545
E-mail: skossida@bioacademy.gr

PeerJ PrePrints

Abstract

Protein phosphorylation is one of the most important protein post-translational modifications and plays a role in numerous cellular processes including recognition, signaling and degradation. It can be studied experimentally by various methodologies, like employing western blot analysis, site-directed mutagenesis, 2 D 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%.

Keywords

Bioinformatics; Protein Phosphorylation; Post-Translational Modifications; Benchmark; Phosphorylation Prediction

PeerJ PrePrints

Protein Phosphorylation

Protein phosphorylation is one of the most common post-translational modifications (PTMs), illustrating a major cellular reversible process that is performed primarily by the protein kinases (PKs). It plays a crucial role in a variety of biological cellular processes, including signal transduction and cell cycle regulation [1]. Biochemically, PKs play a major role by catalyzing the hydrolysis of adenosine triphosphate (ATP), which, in turn, transfers a phosphate moiety to the appropriate acceptor residue (serine (S) / threonine (T) or tyrosine (Y) in eukaryotes, 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 [2].

Phosphorylation is an important process that 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. The identification of kinase-specific phosphorylation sites and the systematic elucidation of site-specific Kinase-Substrate Relations (ssKSRs) may provide the fundamental basis for better understanding cell plasticity and dynamics underlying the molecular mechanisms of various diseases as well as potential pharmacological targets [3].

The eukaryotic organisms frequently prefer to phosphorylate serine rather than threonine residues, so tyrosine phosphorylation is rarely occurs in eukaryots. On the other hand, histidine phosphorylation constitutes amend 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 PK residue-specific acts as regulatory switch by adding one or more phosphate groups to them. Phosphorylation activity is also detected in cyclins and cyclin-dependent kinases (Cdks), which constitute key regulators of cell cycle progression in eukaryotic cells [4]. It is evident that Cdk activity is detected by phosphorylation at three conserved positions [5]. Another example is the Bcl-2 phosphorylation, which regulates cell apoptosis [6]. Table 1 summarizes some examples of phosphorylated amino acid residues and their function.

PeerJ PrePrints

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 time consuming, as well as the finally output of the detected protein is relatively low. Based of advanced modern techniques the leaders for the identification of phosphorylated sited became the high-throughput methods, such as proteomics and analysis by mass spectrometry [11]. Themass spectrometry methods can be utilized to determine the phosphorylated sites in a wide variety of tissues, indicating a large number of visible phosphorylated sites. 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 are not 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 [12].

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 phosphor-proteins related to the extracellular signal-regulated kinase (ERK) pathway [13].

More advanced techniques for the detection of phosphorylated sites of the proteins are developing of the protein microarrays or protein chips [14]. These arrays use very small amounts of purified proteins in a high-density format. This approach allows the simultaneous determination of a variety of analyses from small amounts of samples within a single experiment. The development of protein microarray technology has revolutionized the identification of enzyme-substrate relationships. Functional protein microarrays are typically prepared by immobilizing individually purified proteins onto a microscope slide using a standard contact chip writer or non-contact microarrayer. A protein microarray can be viewed as a substrate array when an enzymatic reaction is performed on it to identify potential downstream targets. Many types of enzymatic reactions have been developed for various types of PTMs, such as phosphorylation, ubiquitylation, acetylation and deoxyribonucleic acid (DNA) cross-linking. Protein chips offer many advantages for studying protein phosphorylation. Thousands of proteins can be rapidly screened for enzyme substrate relationships in an unbiased fashion with very small amounts of reagents and under a variety of test conditions. In addition, closely related kinases with known redundant functions can be readily differentiated at the molecular level on the basis of their substrate profiles [15].

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 [16]

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 [17].

PeerJ PrePrints

136

137

138

PeerJ PrePrints

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. Such techniques require a protein sequence as an input, and in consequence possible numerical measurements are applied for each phosphorylated serine, threonine or tyrosine residues (S / T / Y) in sequence.

There is variability in the bioinformatics tools used for phosphorylation prediction, depending on the number of residues, which surround the phosphorylated point. Additionally, several different learning methods are employed, including artificial neural networks (ANNs), decision trees, genetic algorithms, position-specific scoring matrices (PSSMs), support vector machines (SVMs) and bayesian probability (BP). ANNs and SVMs are the most popular and frequently used methodologies applied by phosphorylation prediction tools. Some methods strike a balance between the simplicity of PSSMs and the opaqueness of ANNs. 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 to use or not use the information structure. Finally, the tools also stand out from their specificity, if they are non-kinase-specific tools or kinase-specific tools. So, the tool makes provisions for specific kinase or kinase families or not kinase-specific [18].

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. The initial computational step requires the implementation of algorithms in order to match the obtained spectrum with the known spectra databases. A number of software packages can be used for this step including Mascot [19], SEQUEST [20], OMSSA [21], X! Tandem [22], GutenTag [23], InsPecT [24] and Spectral Networks Analysis [25]. Due to the high false-positive rate of matching the complex MS spectra, a second computational step is required to filter out known false positives, using algorithms such as DTASelect and PhosphoPIC, followed by a final computational step for further curation and confirmation.

For quantitative proteomics using differential gel electrophoresis (DIGE), a variety of computational tools exist in order to primarily analyze biomarkers by quantifying individual proteins and indicate the separation between one or more protein spots on a scanned image of a DIGE gel. Additionally, these tools match spots between gels of similar samples to show the differences. Software packages include BioNumerics 2D, Delta2D, ImageMaster, Melanie, PDQuest, Progenesis and REDFIN, among others. While such software packages are widely utilized, they are still far from perfect. For instance, while PDQuest and Progenesis tend to agree on the quantification and analysis of well-defined well-separated protein spots, they deliver different results and analysis tendencies with less-defined less-separated spots.

In addition, bioinformatics tools are often used together with protein arrays in order to finalize the result outcome. These tools, namely TM4 software BASE and its Web-based system have been developed at Lund University [26].

PeerJ PrePrints

Finally, RNA interference (RNAi) technology represents a convenient method for inhibiting expression of specific proteins. A Bayesian network has been employed to identify the best network model fitting all perturbation data available. In a large-scale study where genetic perturbation by RNAi was performed, several hundred human kinases were targeted by RNAi and around 11% of the kinome was found essential for promoting cell survival. Additionally, many new kinase substrate pairs were also identified [38].

Table 2 demonstrates the main aspects of phosphorylation site prediction tools available. It is evident that a) different machine learning methods have been utilized; b) widely varying amounts of information (in terms of number of residues surrounding the phosphorylation site) have been incorporated into predictive models; c) a number of different methods has been proposed for both structure-based and sequence based categories; d) several tools exist for both kinase-specific and non-kinase specific predictions and e) many sources of training and testing data have been utilized. However, questions focused on determining the optimum number of residues surrounding the phosphorylation site and developing improved structure based methods as well as additional meta-predictors, still remain open [18].

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.

PeerJ PrePrints

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 protein 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 (see Table 3), 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 [27], NetPhosK 1.0 [28], Musite.net [29], ScanSite [30], SMALI [31], PPSP [32], GPS 1.10 [33] [34] [35] and Phospho.ELM [37]. The tables summarizing the findings of this benchmark can be found in Supplementary Material section of this article.

It was found that the 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 Table 3.

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 shown 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 their predicted potential, as these packages failed almost completely to predict phosphorylation sites in our benchmark, with the 7 predictions, what represent only 3% match with the real data (Figure 1).

Conclusions

PeerJ PrePrints

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. Therefore, studying protein phosphorylation is very important as it could be linked to almost all manifestations of life. 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 a 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.

PeerJ PrePrints

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.

PeerJ PrePrints

References

1. Suter B, Graham C, StagljarI. Exploring protein phosphorylation in response to DNA damage using differentially tagged yeast arrays. Biotechniques. 2008 Nov;45(5):581-4. doi: 10.2144/000112949.

2. Cieśla J, Frączyk T, Rode W. Phosphorylation of basic amino acid residues in proteins: important but easily missed. ActaBiochim Pol. 2011;58(2):137-48. Epub 2011 May 27.

3. Song C, Ye M, Liu Z, Cheng H, Jiang X, Han G, Songyang Z, Tan Y, Wang H, Ren J, Xue Y, ZouH. Systematic analysis of protein phosphorylation networks from phosphoproteomicdata. Mol Cell Proteomics. 2012 Oct;11(10):1070-83. Epub 2012 Jul 13.

4. Masumoto H, Muramatsu S, Kamimura Y, Araki H.S-Cdk-dependent phosphorylation of Sld2 essential for chromosomal DNA replication in budding yeast. Nature. 2002 Feb 7;415(6872):651-5. Epub 2002 Jan 23.

Decker T, KovarikP. Serine phosphorylation of STATs. Oncogene. 2000 May 15;19(21):2628-37.

5. Lew DJ, KornbluthS. Regulatory roles of cyclin dependent kinase phosphorylation in cell cycle control. CurrOpin Cell Biol. 1996 Dec;8(6):795-804.

6.Ruvolo PP, Deng X, May WS. Phosphorylation of Bcl2 and regulation of apoptosis. Leukemia. 2001 Apr;15(4):515-22.

7. Decker T, KovarikP. Serine phosphorylation of STATs. Oncogene. 2000 May 15;19(21):2628-37.

8. Hunter T, Ling N, Cooper JA. Protein kinase C phosphorylation of the EGF receptor at a threonine residue close to the cytoplasmic face of the plasma membrane. Nature. 1984 Oct 4-10;311(5985):480-3.

9. Witthuhn BA, Quelle FW, Silvennoinen O, Yi T, Tang B, Miura O, Ihle JN.JAK2 associates with the erythropoietin receptor and is tyrosine phosphorylated and activated following stimulation with erythropoietin. Cell. 1993 Jul 30;74(2):227-36.

10. Wolanin PM, Thomason PA, Stock JB. Histidine protein kinases: key signal transducers outside the animal kingdom. Genome Biol. 2002 Sep 25;3(10):REVIEWS3013. Epub 2002 Sep 25.

11. St-Denis N, GingrasAC. Mass spectrometric tools for systematic analysis of protein phosphorylation. ProgMolBiolTransl Sci. 2012;106:3-32. doi: 10.1016/B978-0-12-396456-4.00014-6.

12. Sundstrom JM, Sundstrom CJ, Sundstrom SA, Fort PE, Rauscher RL, Gardner TW,AntonettiDA. Phosphorylation site mapping of endogenous proteins: a combined MS and bioinformatics approach. J Proteome Res. 2009 Feb;8(2):798-807. doi: 10.1021/pr8005556.

13. Lovrić J, Dammeier S, Kieser A, Mischak H, KolchW. Activatedraf induces the hyperphosphorylation of stathmin and the reorganization of the microtubule network. JBiol Chem. 1998 Aug 28;273(35):22848-55.

PeerJ PrePrints

14. 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.Global analysis of protein activities using proteome chips.Science. 2001 Sep 14;293(5537):2101-5. Epub 2001 Jul 26.

15. Lin J, Xie Z, Zhu H, QianJ. Understanding protein phosphorylation on a systems level. BriefFunct Genomics. 2010 Jan;9(1):32-42. doi: 10.1093/bfgp/elp045. Epub 2010 Jan 7.

16. Leitner A, Sturm M, Lindner W. Tools for analyzing the phosphoproteome and other phosphorylated biomolecules: a review. AnalChimActa. 2011 Oct 3;703(1):19-30. doi: 10.1016/j.aca.2011.07.012. Epub 2011 Jul 19.

17. MacKeigan JP, Murphy LO, BlenisJ. SensitizedRNAi screen of human kinases and phosphatases identifies new regulators of apoptosis and chemoresistance. Nat Cell Biol. 2005 Jun;7(6):591-600. Epub 2005 May 1.

18. Trost B, KusalikA. Computational prediction of eukaryotic phosphorylation sites. Bioinformatics. 2011 Nov 1;27(21):2927-35. doi: 10.1093/bioinformatics/btr525. Epub 2011 Sep 16.

19. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999 Dec;20(18):3551-67.

20. Yates JR 3rd, Eng JK, McCormack AL, SchieltzD. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal Chem. 1995 Apr 15;67(8):1426-36.

21. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. Open mass spectrometry search algorithm. J Proteome Res. 2004 Sep-Oct;3(5):958-64.

22. Craig R, Beavis RC.TANDEM: matching proteins with tandem mass spectra. Bioinformatics. 2004 Jun 12;20(9):1466-7. Epub 2004 Feb 19.

23. Tabb DL, Saraf A, Yates JR 3rd.GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. Anal Chem. 2003 Dec 1;75(23):6415-21.

24. Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, BafnaV. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra.Anal Chem. 2005 Jul 15;77(14):4626-39.

25. Bandeira N, Tsur D, Frank A, PevznerPA. Protein identification by spectral networks analysis. ProcNatlAcadSci U S A. 2007 Apr 10;104(15):6140-5. Epub 2007 Apr 2.

26. Dudoit S, Gentleman RC, QuackenbushJ. Open source software for the analysis of microarray data. Biotechniques. 2003 Mar;Suppl:45-51.

27. Nikolaj Blom , Steen Gammeltoft and Sùren Brunak. Sequence and Structure-based Prediction of Eukaryotic Protein Phosphorylation SitesJ. Mol. Biol.1999;294, 1351-1362

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

PeerJ PrePrints

29. Jianjiong Gao, Jay J. Thelen, A. Keith Dunker and Dong Xu. Musite, a Tool for Global Prediction of General and Kinase-specific Phosphorylation Sites. Molecular & Cellular Proteomics 2010. 9,2586 –2600

30. John C. Obenauer, Lewis C. Cantley and Michael B. Yaffe.Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Research 2003; 13, 3635–3641

31. Lei Li, Chenggang Wu, Haiming Huang, Kaizhong Zhang, Jacob Gan andShawn S.-C. Li. Prediction of phosphotyrosine signaling networks using a scoring matrix-assisted ligand identification approach. Nucleic Acids Research 2008;10,3263–3273

32. Yu Xue, Ao Li, Lirong Wang, Huanqing Feng and Xuebiao Yao. PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory.BMC Bioinformatics 2006;7:163

33. Feng-Feng Zhou, Yu Xue, Guo-Liang Chen, Xuebiao Yao. GPS: a novel group-based phosphorylation predicting and scoring method. Biochemical and Biophysical Research Communications 2004; 325,1443–1448

34. Yu Xue, Fengfeng Zhou, Minjie Zhu, Kashif Ahmed, Guoliang Chen and Xuebiao Yao. GPS: a comprehensive www server for phosphorylation sites prediction. Nucleic Acids Research, 2005;33

35. Yu Xue, Jian Ren, Xinjiao Gao, Changjiang Jin, Longping Wen, and Xuebiao Yao. GPS 2.0, a Tool to Predict Kinase-specific Phosphorylation Sites in Hierarchy. Molecular & Cellular Proteomics 7.9 2008;1598-1608

36. Lilia M. Iakoucheva, Predrag Radivojac1, Celeste J. Brown, Timothy R. O'Connor, Jason G. Sikes, Zoran Obradovic1 and A. Keith Dunker. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Research 2004; 3 1037-1049

37. Holger Dinkel, Claudia Chica, Allegra Via, Cathryn M. Gould, Lars J. Jensen, Toby J. Gibson and Francesca Diella. Phospho.ELM: a database of phosphorylation sites—update 2011.Nucleic Acids Research 2011, D261–D267

38. Jimmy Lin, Zhi Xiw, Heng Zhu and Jiang Qian. Understanding protein phosphorylation on a system level. Briefings in functional genomics 2010; 9(1): 32–42.

39. Colleen Sweeney Crovello, Barbara C. Furie and Bruce Furie. Histidine phosphorylation of P-Selectin upon stimulation of huma platelets: A novel pathyay for activation – dependent signal transduction. Cell 1995, 84 270-286

PeerJ PrePrints

343

344

PeerJ PrePrints

345 **Table & Figure Legends:**

346

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

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

351 Table 3. Experimentally verified phosphorylated amino acids on specific proteins together with the
352 corresponding predictions of various phosphorylation prediction tools.

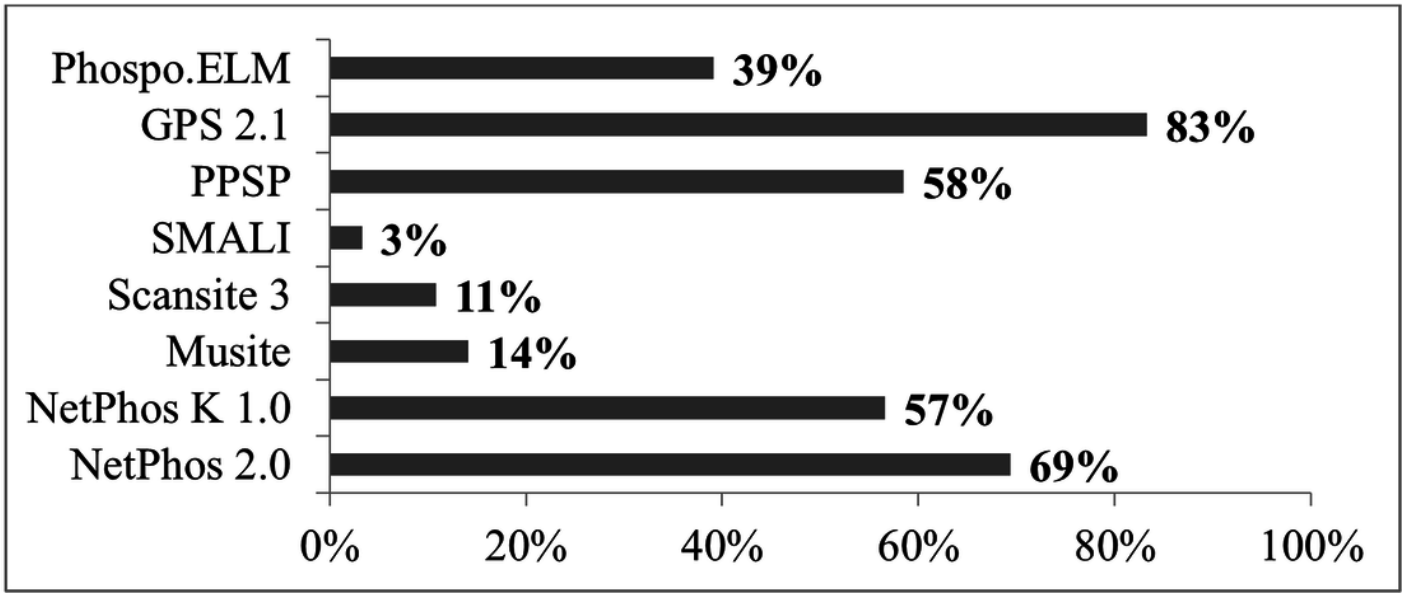
353 Figure 1. Graphic representation of positive prediction of phosphorylated sites by selected software

PeerJ PrePrints

1

Figure 1

Graphic representation of positive prediction of phosphorylated sites by selected software



PeerJ PrePrints

Table 1(on next page)

Table 1.

Examples of phosphorylated amino acid residues and their function

PeerJ PrePrints

2 **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	[7]
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.	[8]
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.	[9]
Histidine	H	Histamine precursor, carbon atoms-source in	Histidine phosphorylation occurs in several platelet	[10, 39]

PeerJ PrePrints

(Basic side chains)		purines	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.	

PeerJ PrePrints

Table 2(on next page)

Table 2

Phosphorylation detection tools together with their 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.

PeerJ PrePrints

Table 2. Phosphorylation detection tools together with their 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

PeerJ PrePrints

Table 3(on next page)

Table 3.

Experimentally verified phosphorylated amino acids on specific proteins together with the corresponding predictions of various phosphorylation prediction tools.

PeerJ PrePrints

Table 3. Experimentally verified phosphorylated amino acids on specific proteins together with the corresponding predictions of various phosphorylation prediction tools.

Accession number	Protein, *organism*	Experimentally verified phosphorylated amino acids	NetPhos 2.0	NetPhos K 1.0	Musite	Scansite3	SMA LI	PPSP	GPS 2.1	Phospho. ELM
E0J4T6	GTP-binding protein, *E. coli*	S16	-	-	-	-	-	-	-	-
E8VA72	Enolase, *Bacillus subtilis*	Y281	-	-	-	-	YES	-	YES	-
O15530	3-phosphoinositide-dependent protein kinase 1, *Homo sapiens*	S241	YES	YES	-	-	-	YES	YES	YES
		S394	YES	-	-	-	-	-	YES	-
		S398	YES	-	-	-	-	-	YES	-
		S501	-	YES	-	-	-	-	YES	-
		S529	-	-	-	-	-	-	YES	-
		T354	YES	YES	-	-	-	YES	YES	-
		Y9	YES	-	-	-	-	YES	YES	YES
		Y37	-	-	-	-	-	-	-	-
		Y373	YES	YES	-	YES	-	YES	YES	YES
O34507	Serine/threonine-protein kinase PrkC, *Bacillus subtilis*	S214	-	YES	-	-	-	YES	-	-
		T165	-	-	-	-	-	YES	YES	-
		T167	-	-	-	-	-	YES	YES	-
O34824	Phosphoglucosamine mutase, *Bacillus subtilis*	S88	-	-	-	-	-	-	-	-
		S102	-	-	-	-	-	-	-	-
O95997	Securin, *Homo sapiens*	S165	YES	YES	YES	-	-	-	YES	YES
P04049	RAF proto-oncogene serine/threonine-	S29	-	YES	-	-	-	-	YES	-
		S43	YES	YES	-	-	-	YES	YES	YES
		S259	YES	YES	-	YES	-	YES	YES	YES

PeerJ PrePrints

	protein kinase, *Homo sapiens*	S269	-	-	-	-	-	-	-	-
		S289	YES	YES	YES	-	-	YES	YES	YES
		S296	YES	YES	YES	-	-	YES	YES	YES
		S301	YES	YES	YES	-	-	-	YES	YES
		S338	YES	YES	-	YES	-	YES	YES	YES
		S339	YES	YES	-	YES	-	YES	YES	YES
		S494	-	-	-	-	-	-	YES	YES
		S621	YES	YES	YES	YES	-	YES	YES	YES
		S642	YES	YES	-	-	-	YES	YES	YES
		T431	-	-	-	-	-	-	-	-
		Y341	YES	YES	-	-	-	YES	YES	YES
P04083	Annexin A1, *Homo sapiens*	S5	-	-	-	-	-	-	YES	-
		S27	YES	YES	-	-	-	YES	YES	YES
		S37	YES	YES	-	-	-	YES	YES	-
		Y21	-	-	-	-	-	-	-	-
P04792	Heat shock protein beta-1, *Homo sapiens*	S78	YES	YES	-	-	-	YES	YES	YES
		S82	YES	-	-	-	-	YES	YES	YES
P0A5N2	Cell wall synthesis protein Wag31, *Mycobacterium tuberculosis*	T73	-	YES	-	-	-	YES	YES	-
P0A6N2	Elongation factor Tu, *E. coli*	T382	-	-	-	-	-	-	-	-
P0A763	Nucleoside diphosphate kinase, *E. coli*	T93	-	YES	-	-	-	-	YES	-
P10636	Tau protein, *Homo sapiens*	S214	YES	YES	YES	-	-	YES	YES	-
		S548	-	-	-	-	-	-	-	-
		S554	YES	-	-	-	-	YES	YES	-

PeerJ PrePrints

		S579	YES	YES	-	-	-	YES	YES	YES
		S602	YES	YES	-	-	-	YES	YES	-
		S606	-	YES	-	-	-	YES	YES	-
		S610	-	YES	-	-	-	YES	YES	-
		S622	-	-	-	-	-	YES	YES	-
		S641	YES	YES	-	-	-	YES	YES	-
		S669	YES	YES	-	-	-	YES	YES	-
		S673	-	YES	-	-	-	YES	YES	-
		S717	YES	-	-	-	-	YES	YES	-
		T534	-	YES	YES	-	-	YES	YES	YES
		T548	-	YES	YES	YES	-	YES	YES	YES
		Y18	-	-	-	-	-	YES	YES	-
P13796	Plastin-2, Homo sapiens	S5	YES	YES	-	-	-	-	YES	-
P18159	Phosphoglucomutas, *Bacillus substilis*	S100	-	-	-	-	-	-	-	-
P23528	Cofilin-1, *Homo sapiens*	S3	-	-	-	-	-	-	YES	YES
		S156	-	-	-	-	-	YES	YES	YES
		T25	YES	YES	-	-	-	YES	YES	YES
		Y68	YES	-	-	-	-	-	YES	YES
		Y140	YES	-	-	-	YES	YES	YES	YES
		Y156	-	-	-	-	-	-	-	-
P29320	Ephrin type-A receptor 3, *Homo sapiens*	Y596	-	-	-	-	-	YES	YES	YES
		Y602	YES	YES	-	YES	-	YES	YES	YES
		Y701	YES	-	-	-	-	YES	YES	-
P30307	M-phase inducer phosphatase 3, Homo *sapiens*	S38	YES	YES	-	-	-	YES	YES	YES
		S57	-	YES	-	-	-	YES	YES	YES
		S61	-	-	-	-	-	-	YES	YES
		S64	-	YES	-	-	-	-	YES	YES

PeerJ PrePrints

		S168	-	-	-	-	-	YES	YES	YES
		S191	-	YES	-	-	-	YES	YES	YES
		S198	YES	YES	-	-	-	YES	YES	YES
		S212	-	-	-	-	-	-	-	-
		S472	YES	YES	YES	-	-	-	YES	-
		T48	YES	YES	-	YES	-	YES	YES	YES
		T67	YES	YES	-	YES	-	YES	YES	YES
P31103	Nucleoside diphosphate kinase, *Bacillus substilis*	T91	-	-	-	-	-	-	-	-
		T92	-	-	-	-	-	-	YES	-
P31120	Phosphoglucosamine mutase, *E. coli*	S102	-	-	-	-	-	YES	-	-
P31751	RAC-beta serine/threonine-protein kinase, *Homo sapiens*	S473	-	-	-	-	-	-	-	-
		S474	-	-	-	-	-	-	-	-
		T309	-	YES	-	-	-	YES	YES	YES
P35568	Insulin receptor substrate 1, *Homo sapiens*	S270	YES	YES	-	-	-	YES	YES	YES
		S307	YES	YES	YES	YES	-	YES	YES	YES
		S312	YES	YES	YES	-	-	YES	YES	YES
		S348	YES	YES	YES	-	-	YES	YES	YES
		S636	YES	YES	YES	-	-	-	YES	YES
		S1101	YES	YES	-	-	-	YES	YES	YES
		Y896	YES	-	-	YES	-	YES	YES	YES
		Y941	YES	-	-	YES	-	YES	YES	-
P37840	Alpha-synuclein, *Homo sapiens*	S129	YES	YES	-	-	-	-	YES	YES
		Y125	YES	YES	-	YES	-	YES	YES	YES
P41685	Cellular tumor antigen p53, *Felis catus*	S20	YES	-	-	-	-	YES	-	-
		S33	YES	-	-	-	-	-	YES	-
		S46	-	YES	-	-	-	-	YES	-

PeerJ PrePrints

		S156	-	-	-	-	-	-	-	-
		S385	-	YES	YES	-	-	-	YES	-
		T181	-	-	-	-	-	-	-	-
P49841	Glycogen synthase kinase-3 beta, *Homo sapiens*	S9	YES	YES	-	YES	-	YES	YES	YES
		T390	YES	YES	YES	-	-	YES	YES	YES
		T402	-	YES	-	-	-	YES	YES	YES
P51593	E3 ubiquitin-protein ligase HUWE1, *Rattus norvegicus*	Y219	YES	-	-	-	-	YES	YES	-
P51636	Caveolin-2, *Homo sapiens*	S23	YES	YES	-	-	-	-	YES	YES
		S36	YES	YES	-	-	-	-	YES	YES
		S727	-	-	-	-	-	-	-	-
		Y19	YES	YES	-	-	-	YES	YES	YES
		Y27	YES	-	-	-	YES	YES	YES	YES
P55008	Allograft inflammatory factor 1, *Homo sapiens*	Y54	YES	-	-	-	-	YES	YES	-
		Y124	-	-	-	-	-	YES	YES	-
P55211	Caspase-9, *Homo sapiens*	S302	-	-	-	-	-	-	-	-
		S307	YES	YES	YES	-	-	YES	YES	YES
		T125	-	-	-	-	-	-	-	-
		Y153	-	-	-	-	YES	-	YES	YES
P61012	Cardiac phospholamban, *Canis familiaris*	S16	YES	YES	-	YES	-	YES	YES	-
		T17	-	-	-	-	-	-	-	-
P62753	40S ribosomal protein S6, *Homo sapiens*	S235	YES	YES	YES	YES	-	YES	YES	YES
		S236	YES	YES	YES	YES	-	YES	YES	YES
		S240	YES	YES	-	-	-	YES	YES	YES
		S244	YES	YES	-	-	-	-	YES	YES

PeerJ PrePrints

		S247	YES	YES	-	-	-	-	YES	YES
P65728	Serine/threonine-protein kinase PknG, *Mycobacterium tuberculosis*	S65	YES	YES	-	-	-	-	YES	-
		T23	-	-	-	-	-	YES	YES	-
		T32	-	YES	-	-	-	YES	YES	-
		T63	YES	-	YES	-	-	YES	YES	-
		T64	YES	YES	-	-	-	-	YES	-
P80885	Pyruvate kinase, Bacillus substilis	S36	YES	YES	-	-	-	-	-	-
P95078	Serine/threonine-protein kinase PknK, *Mycobacterium tuberculosis*	T179	-	-	-	-	-	-	YES	-
		T181	-	-	-	-	-	-	YES	-
Q00969	Cyclic AMP-dependent transcription factor ATF-2, *Rattus norvegicus*	S44	YES	YES	-	-	-	YES	YES	-
		S103	YES	YES	-	-	-	-	-	-
		S472	YES	YES	-	-	-	YES	YES	-
		S480	-	YES	-	-	-	-	YES	-
		T51	YES	YES	-	-	-	-	YES	-
		T53	YES	YES	-	-	-	YES	YES	-
		T55	-	-	-	-	-	-	YES	-
Q02750	Dual specificity mitogen-activated protein kinase kinase 1, *Homo sapiens*	S218	YES	-	-	-	-	YES	YES	YES
		S222	YES	-	-	-	-	YES	YES	YES
		S298	YES	YES	-	-	-	YES	YES	YES
		T286	YES	YES	YES	-	-	YES	YES	YES
Q06752	Cysteine--tRNA ligase, *Bacillus substilis*	S270	YES	YES	-	-	-	YES	-	-
Q12778	Forkhead box protein O1, *Homo sapiens*	S54	-	-	-	-	-	-	-	-
		S212	-	YES	-	-	-	YES	YES	-
		S256	YES	YES	YES	YES	-	YES	YES	YES

PeerJ Preprints

		S319	YES	YES	-	YES	-	YES	YES	YES
		S322	YES	-	-	-	-	-	YES	YES
		S325	-	YES	-	-	-	-	YES	YES
Q12968	Nuclear factor of activated T-cells, cytoplasmic 3, *Homo sapiens*	T24	-	-	-	-	-	-	-	-
Q13541	eIF4E-binding protein, *Homo sapiens*	S65	YES	-	-	-	-	YES	YES	YES
		S83	YES	YES	YES	-	-	YES	YES	YES
		S101	YES	YES	YES	-	-	YES	YES	YES
		S112	YES	YES	YES	-	-	YES	YES	YES
		T37	YES	YES	-	-	-	YES	YES	YES
		T41	-	-	-	-	-	YES	YES	-
		T46	YES	YES	-	-	-	YES	YES	YES
		T50	-	-	-	-	-	YES	YES	-
		Y54	-	-	-	-	YES	YES	YES	-
		T70	YES	YES	YES	-	-	YES	YES	YES
Q16236	Nuclear factor erythroid 2-related factor 2, *Homo sapiens*	S40	YES	YES	-	-	-	-	YES	-
Q5S007	Leucine-rich repeat serine/threonine-protein kinase 2, *Homo sapiens*	S40	-	-	-	-	-	-	-	-
		S910	YES	YES	-	-	-	YES	YES	-
		S935	-	YES	-	-	-	YES	YES	-
Q61083	Mitogen-activated protein kinase kinase kinase 2, *Mus musculus*	T524	YES	YES	-	-	-	YES	YES	-
Q62074	Protein kinase C iota type, *Mus*	T411	YES	YES	-	-	-	YES	YES	-

PeerJ PrePrints

	musculus	Y264	-	YES	-	-	-	YES	YES	-
		Y279	-	-	-	-	-	-	YES	-
		Y333	YES	-	-	-	-	-	YES	-
Q64010	Adapter molecule crk, *Mus musculus*	Y221	YES	-	-	YES	-	YES	YES	YES
Q6J1J1	Baculoviral IAP repeat-containing protein 5, *Bos taurus*	T117	-	-	-	-	-	-	-	-
Q6P2N0	Myosin light chain kinase, *Homo sapiens*	S19	YES	YES	YES	-	-	YES	YES	-
		T181	-	-	-	-	-	-	-	-
Q8BZ03	Serine/threonine-protein kinase D2, *Mus musculus*	S244	YES	YES	-	-	-	-	YES	-
		S707	YES	YES	-	-	-	-	YES	-
		S711	YES	YES	-	-	-	YES	YES	-
		S873	-	YES	-	-	-	-	YES	-
		Y438	-	-	-	-	-	-	YES	-
Q8HXW5	Presenilin-1, *Macaca fascicularis*	S346	YES	YES	-	-	-	YES	YES	-
Q93V58	Serine/threonine-protein kinase GRIK1, *Arabidopsis thaliana*	S261	-	YES	-	-	-	-	YES	-
		T154	-	-	-	-	-	-	-	-
Q95207	Interferon alpha/beta receptor 2, *Ovis aries*	Y340	YES	-	-	-	-	YES	YES	-
		Y525	-	YES	-	-	-	YES	YES	-
Q9H2X6	Homeodomain-interacting protein	S16	-	YES	-	-	-	YES	YES	-
		S118	-	YES	-	-	-	-	YES	-

PeerJ PrePrints

	kinase 2, *Homo sapiens*	S135	YES	YES	-	-	-	-	YES	-
		S441	YES	YES	-	-	-	-	YES	-
		S634	-	-	-	-	-	-	YES	-
		S668	YES	YES	-	YES	-	YES	YES	-
		S815	YES	-	-	-	-	-	YES	-
		S827	YES	YES	YES	YES	-	YES	YES	-
		S934	YES	YES	-	-	-	YES	YES	-
		S1041	-	YES	-	-	-	-	YES	-
		S1155	YES	YES	-	-	-	-	YES	-
		S1188	-	YES	-	-	-	-	YES	-
		T252	-	-	-	-	-	-	YES	-
		T273	-	-	-	-	-	YES	YES	-
		T361	-	-	-	-	-	-	-	-
		T482	-	-	-	-	-	YES	YES	-
		T517	YES	YES	-	-	-	YES	YES	-
		T566	YES	-	-	-	-	-	YES	-
		T687	-	YES	-	-	-	-	YES	-
		T991	YES	-	-	-	-	-	YES	-
		Y36	-	-	-	-	-	-	-	-
Q9MZA9	Vimentin, *Ovis aries*	S55	YES	-	-	-	-	YES	YES	-
		S56	-	YES	YES	-	-	YES	YES	-
Q9UD71	Protein phosphatase 1 regulatory subunit 1B, *Homo sapiens*	T34	YES	YES	YES	YES	-	YES	YES	YES
Q9UMF0	ICAM-5, Homo sapiens	Y361	-	-	-	-	-	-	-	-
2VX3	Dual specificity tyrosine-phosphorylation-regulated kinase	Tyr218	-	-	-	-	-	YES	-	YES

PeerJ PrePrints

	1a, *Homo sapiens*									
1U54	Activated CDC42 kinase 1, *Homo sapiens*	Tyr178	YES	-	-	-	YES	YES	YES	YES
1T15	Breast cancer type 1 susceptibility protein, *Homo sapiens*	Ser10	-	YES	-	-	-	YES	YES	-
2ERK	Extracellular signal-regulated kinase 2, *Homo sapiens*	Thr190	YES	YES	-	-	-	YES	YES	YES
		Tyr192	YES	-	-	-	-	YES	YES	YES
2IVV	Proto-oncogene tyrosine-protein kinase receptor ret precursor, *Homo sapiens*	Tyr206	YES	-	YES	-	YES	YES	-	YES

(- *the phosphorylated site was not found*)

PeerJ PrePrints