Published in final edited form as:
Nat Biotechnol. 2008 December ; 26(12): 1339–1340. doi:10.1038/nbt1208-1339.

PhosphoPep – a Database of Protein Phosphorylation Sites for Systems Level Research in Model Organisms

Bernd Bodenmiller[1,2,@], **David Campbell**[3,@], **Bertran Gerrits**[4,@], **Henry Lam**[3], **Marko Jovanovic**[2,5], **Paola Picotti**[1], **Ralph Schlapbach**[4], and **Ruedi Aebersold**[1,3,6,7,*]

[1] Institute of Molecular Systems Biology, ETH Zurich, 8093 Zurich, Switzerland [2] Zurich PhD Program in Molecular Life Sciences, 8057 Zurich, Switzerland [3] Institute for Systems Biology, Seattle, WA 98103, USA [4] Functional Genomics Center Zurich, UZH | ETH Zurich, 8057 Zurich, Switzerland [5] Institute of Molecular Biology, University of Zurich, 8057 Zurich, Switzerland [6] Competence Center for Systems Physiology and Metabolic Diseases, ETH Zurich, 8093 Zurich, Switzerland [7] Faculty of Science, University of Zurich, 8057 Zurich, Switzerland

To the editor

Reversible protein phosphorylation is a universal process that is involved in the control of most biological processes. The comprehensive and quantitative analysis of the protein phosphorylation patterns of cells at different states is therefore of considerable and general interest. Over the past years, mass spectrometry has become the method of choice for the analysis of protein phosphorylation and impressive gains have been realized in the isolation of phosphorylated peptides from complex samples as well as their mass spectrometric and computational analysis. In such studies hundreds to thousands of phosphopeptides and phosphorylation sites are now routinely identified.

Currently, several databases exist which store and disseminate protein phosphorylation data obtained from large scale studies, however, several factors limit their utility. First, and most importantly, the current phosphopeptide databases are human and/or rodent centric. Examples include the human proteome reference database (HPRD)[1], PhosphoElm[2], Phosida[3] and PhosphoSitePlus (www.phosphosite.org). Extensive phosphoproteome data sets for model organisms organized in databases are still missing. Second, the lack of phosphorylation data from diverse species precludes comparative studies, e.g. those that assess whether specific phosphorylation sites or perturbation induced phosphopeptide patterns are conserved between species. For example, the analysis of the evolutionary conservation of the human phosphorylation sites of the Phosida[3] database relies on amino acid sequence conservation, but not on observed phosphorylation sites in other species. Third, none of these databases provides sufficient information to validate, identify and quantify the presented phosphorylation sites by mass spectrometry in independent experiments.

To address these issues and to complement existing protein databases for life science research we describe the PhosphoPep v2.0 database (www.phosphopep.org)[4] which is a significant extension of its first version, PhosphoPep v1.0. In its initial implementation the database contained 12,756 assigned phosphorylation sites identified in *D. melanogaster* Kc167 cells,

[*] Corresponding author: aebersold@imsb.biol.ethz.ch, Phone: +41 44 633 31 70, Fax: +41 44 633 10 51.
[@] These authors contributed equally

Data availability
All data presented in this study are available from PhosphoPep (www.phosphopep.org).

the tandem mass spectra that led to their assignment[4, 5] and a suite of associated software tools supporting the interactive use of the data contained in the database for further experiments and meta-analysis.

PhosphoPep v2.0 significantly extends the contents and utilities of the database compared to v1.0. First, PhosphoPep now includes phosphoproteome data from the four species yeast (*S. cerevisiae*), worm (*C. elegans*), fly (*D. melanogaster*) and human (*H. sapiens*) (see Table 1). These data also represent the first large scale phosphorylation data set for *C. elegans*. Second, we implemented a novel function to analyze the conservation of the identified phosphorylation sites across species (Figure 1). Third, for every phosphorylation site we provide, in downloadable form, a mass spectrometric assay based on multiple reaction monitoring to support further experimentation, including accurate quantification of the respective site in complex samples. Fourth, we implemented a dedicated help page which explains all displayed parameters and a downloadable tutorial for those scientists who intend to use the resource but are not trained in the analysis of mass spectrometry data. Specifically, the tutorial describes how the quality of a phosphopeptide and phosphorylation site identification based on fragment ion spectra can be assessed (see Supplementary Material and Methods), and fifth, the pre-existing software tools were adapted for use with the data from all four species. Collectively, these advances significantly expand the available data, support a wider range of queries and make the resource accessible to a wider range of scientists.

The new data added were obtained from focused phosphoproteome mapping experiments carried out in our lab and to some extent by contributions from laboratories making their phosphoproteomic data generously available[6–9]. For the data collected in our group we followed the data collection strategy described for *D. melanogaster* (See Supplementary Material and Methods). For *C. elegans* this generated 5,444 unique high confidence phosphopeptides that could be assigned to 2,959 gene products, comprising 3,545 assigned unique phosphorylation sites. For *S. cerevisiae*, using the same strategy and combining the in house data with a published data set[9] we identified at high confidence 9,554 phosphopeptides that could be assigned to 2,071 gene products, comprising 5,890 assigned unique phosphorylation sites. The assigned proteins cover nearly one third of the predicted yeast proteome with no bias in the range of protein abundance (Supplementary Figure 1A) but with a bias towards proteins involved in signal transduction (Supplementary Figure 1B). For human, we used previously published data from cancer and HELA cells[6–8] that were made accessible to identify at high confidence 3,784 unique phosphopeptides that could be assigned to 5,160 gene product, comprising 2,810 assigned phosphorylation sites. Finally, the contents of the *D. melanogaster* data set include 16,875 phosphopeptides that could be assigned to 5,347 gene products, comprising 12,756 assigned phosphorylation sites.

The ability to support cross species comparisons arose from the inclusion of phosphopeptide data from four species and is a significant new and unique feature of PhosphoPep v2.0. In a first step the user can view the orthologous phosphoproteins (if known) between the species starting from any protein information page[10] (Figure 1). In a second step, the amino acid sequences of the orthologous proteins are aligned and the phosphorylation sites which are stored in PhosphoPep are highlighted on the alignment. In addition, the level of conservation is displayed for each site[11] (Figure 1). This new function will help to assess the conservation of signaling networks and the assignment of phosphorylation sites across species.

In summary, the novel model organism datasets and the unique set of software tools implemented in PhosphoPep v.2.0 support the analysis of single phosphoproteins, the detection of quantitative changes in the state of phosphorylation of whole signaling pathways at different cellular states and the investigations into the evolution of signaling networks from yeast, worm, fly to human. The system has been designed to enable the rapid iterative cycles of experimentation and analysis that are the basis of systems biology research and should therefore find wide application in basic and applied research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We want to acknowledge Steve Gygi, Sean Beausoleil and Cell Signaling Technology for providing us with phosphorylation data as published[6–9]. We also want to thank the Functional Genomics Center Zurich for the generous support with mass spectrometry resources. This project has been funded in part by ETH Zurich, the Swiss National Science Foundation under grant No. 31000-10767, with Federal (US) funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, under contract No. N01-HV-28179, and by the Center for Model Organism Proteomics of SystemsX.ch, the Swiss initiative for systems biology. Work at the FGCZ has been supported by the University Research Priority Program Systems Biology and Functional Genomics of the University of Zurich. RA was supported in part by a grant from F Hoffmann-La Roche Ltd (Basel, Switzerland) provided to the Competence Center for Systems Physiology and Metabolic Disease. BB is the recipient of a fellowship by the Boehringer Ingelheim Fonds.

References

1. Peri S, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 2003;13:2363–2371. [PubMed: 14525934]

2. Diella F, et al. Phospho. ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. BMC Bioinf 2004;5:79.

3. Gnad F, et al. PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biol 2007;8:R250. [PubMed: 18039369]

4. Bodenmiller B, et al. PhosphoPep--a phosphoproteome resource for systems biology research in Drosophila Kc167 cells. Mol Syst Biol 2007;3:139. [PubMed: 17940529]

5. Bodenmiller B, Mueller LN, Mueller M, Domon B, Aebersold R. Reproducible isolation of distinct, overlapping segments of the phosphoproteome. Nat Methods 2007;4:231–237. [PubMed: 17293869]

6. Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol 2006;24:1285–1292. [PubMed: 16964243]

7. Rikova K, et al. Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. Cell 2007;131:1190–1203. [PubMed: 18083107]

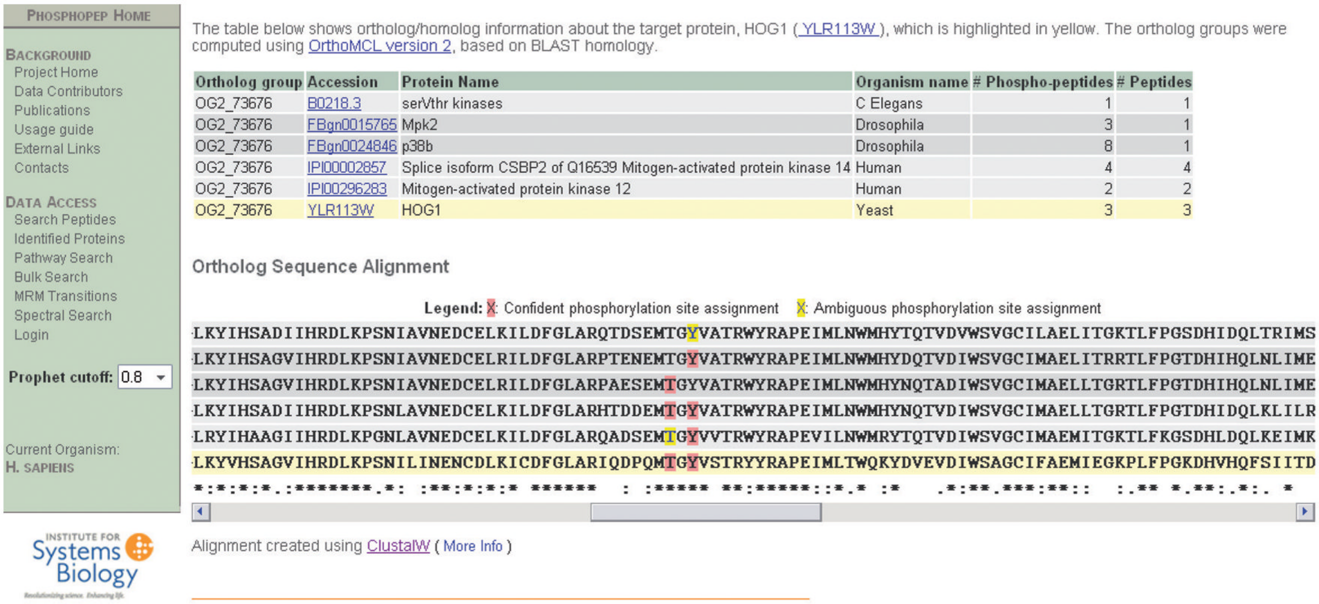
8. Beausoleil SA, et al. Large-scale characterization of HeLa cell nuclear phosphoproteins. P Natl Acad Sci USA 2004;101:12130–12135.

9. Li X, et al. Large-scale phosphorylation analysis of alpha-factor-arrested Saccharomyces cerevisiae. J Proteome Res 2007;6:1190–1197. [PubMed: 17330950]

10. Chen F, Mackey AJ, Vermunt JK, Roos DS. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS ONE 2007;2:e383. [PubMed: 17440619]

11. Larkin MA, et al. Clustal W and Clustal X version 2.0. Bioinformatics 2007;23:2947–2948. [PubMed: 17846036]

12. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 2002;74:5383–5392. [PubMed: 12403597]



Prophet cutoff: 0.8

Current Organism:
H. SAPIENS

The table below shows ortholog/homolog information about the target protein, HOG1 (YLR113W), which is highlighted in yellow. The ortholog groups were computed using OrthoMCL version 2, based on BLAST homology.

Ortholog group	Accession	Protein Name	Organism name	# Phospho-peptides	# Peptides
OG2_73676	B0218.3	ser/thr kinases	C Elegans	1	1
OG2_73676	FBgn0015765	Mpk2	Drosophila	3	1
OG2_73676	FBgn0024846	p38b	Drosophila	8	1
OG2_73676	IPI00002857	Splice isoform CSBP2 of Q16539 Mitogen-activated protein kinase 14	Human	4	4
OG2_73676	IPI00296283	Mitogen-activated protein kinase 12	Human	2	2
OG2_73676	YLR113W	HOG1	Yeast	3	3

Ortholog Sequence Alignment

Legend: X: Confident phosphorylation site assignment X: Ambiguous phosphorylation site assignment

```
·LKYIHSADIIHRDLKPSNIAVNEDCELKILDFGLARQTDSEMTGYVATRWYRAPEIMLNWMHYTQTVDVWSVGCILAELITGKTLFPGSDHIDQLTRIMS
·LKYIHSAGVIHRDLKPSNIAVNEDCELRILDFGLARPTENEMTGYVATRWYRAPEIMLNWMHYDQTVDIWSVGCIMAELITRRTLFPGTDHIHQLNLIME
·LKYIHSAGVIHRDLKPSNIAVNEDCELRILDFGLARPAESEMTGYVATRWYRAPEIMLNWMHYNQTADIWSVGCIMAELLTGRTLFPGTDHIHQLNLIME
·LKYIHSADIIHRDLKPSNLAVNEDCELKILDFGLARHTDDEMTGYVATRWYRAPEIMLNWMHYNQTVDIWSVGCIMAELLTGRTLFPGTDHIDQLKLILR
·LRYIHAAGIIHRDLKPGNLAVNEDCELKILDFGLARQADSEMTGYVVTRWYRAPEVILNWMRYTQTVDIWSVGCIMAEMITGKTLFKGSDHLDQLKEIMK
·LKYVHSAGVIHRDLKPSNILINENCDLKICDFGLARIQDPQMTGYVSTRYYRAPEIMLTWQKYDVEVDIWSAGCIFAEMIEGKPLFPGKDHVHQFSIITD
 *:*:*:*.:*******.*: :**:*:*:* ******  : :***** **:*****::*.* :*   .*:**.***:**::  :.** *.**:.*:. *
```

Alignment created using ClustalW (More Info)

Figure 1.

Figure 1A. Analysis of phosphorylation site conservation. As an example the yeast protein Hog1 is used. In the upper half, the orthologous proteins of Hog1 in worm, fly and human are displayed[10]. In the lower half, the amino acid sequences of the proteins are shown and the identified phosphorylation sites are highlighted[11]. It can be seen that the phosphorylation of the TXY motif, which is known to activate MAP kinases, is conserved between all species.

Table 1

Organism	**Phosphopeptides with P>0.8**[a]	**Total phosphorylation sites**	**Phosphopeptides with assigned phosphorylation site(s)**[b]
D. melanogaster	16,875	16,608	12,756
S. cerevisiae	9,554	8,901	5,890
C. elegans	5,444	4,986	3,545
Human	3,784	3,980	2,810

[a] PeptideProphet Score as computed by PeptideProphet[12]

[b] A phosphopeptide was considered to have an unassigned/assigned site if a dCn threshold was not reached/exceeded (See Supplementary Material and Methods)