A curated compendium of phosphorylation motifs

To the editor:

Protein phosphorylation is a covalent post-translational modification event that is essential for regulation and maintenance of most biological processes in eukaryotes[1]. Conventionally, phosphorylation sites have been discovered using biochemical experiments with the sequence contexts or 'motifs' generally defined by manually aligning the sequences of phosphorylated peptides. Today, high-throughput methods, such as tandem mass spectrometry, allow rapid and directed discovery of hundreds of phosphorylation sites in a single experiment, which necessitates analysis of the 'motifs' in the identified phosphopeptides. Until now, however, there has been no single resource that allows investigators to determine whether a protein/peptide of interest contains a motif that has been previously described in the literature. We describe below a publicly available catalog of phosphorylation motifs that we have created for the community (http://www.hprd.org/PhosphoMotif_finder).

In higher eukaryotes, phosphorylation occurs on serine, threonine and tyrosine residues. Dysregulation of phosphorylation and dephosphorylation events is associated with several diseases and malignancies in humans. The determinants of specificity for kinases are not well understood, although it is understood that both the amino acid sequence motif surrounding the serine/threonine/tyrosine residues and the three-dimensional structure of the substrate proteins contribute to the specificity. The surrounding sequence context, or motif, is in turn responsible for the binding of downstream proteins containing modular phosphoprotein-binding domains. For example, tyrosine phosphorylated proteins have been shown to bind to Src homology 2 (SH2) and phosphotyrosine binding (PTB) domains[2], whereas serine/threonine phosphorylated proteins have been described to bind to WW (two conserved tryptophans), Forkhead, WD40 or 14-3-3 domain-containing proteins[3].

Although several prediction programs for phosphorylation motifs are available[4–8], they do not always offer a simple 'mapping' of all literature-based motifs onto test sequences. For example, Scansite[5] is one resource that predicts short protein sequence motifs that can be recognized by signaling domains or phosphorylated by protein serine/threonine/tyrosine kinases on the basis of results from oriented peptide library and phage display experiments. In addition, NetPhos[8] and KinasePhos[7] predict kinase phosphorylation sites in protein sequences based on a data set of known phosphorylation sites. Phospho.ELM[9] and Phosphosite are databases that contain a list of experimentally determined phosphorylation sites. Significantly, these resources neither provide a listing of consensus motifs nor the exact algorithm used for prediction of motifs.

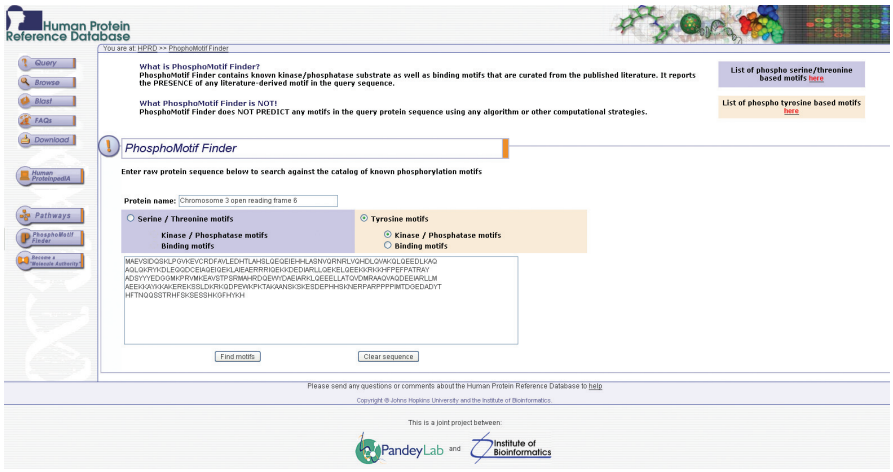
To ensure identification of well-described motifs in any test sequence, one must also go through abstracts of PubMed articles where such motifs have been described—not a trivial task given that there are >130,000 published articles containing the term 'phosphorylation'. Furthermore, the abstracts of most of these papers do not mention if any motif is described or provide the sequence of the motif itself, which precludes automated data mining.

We have assembled a compendium of 324 known phosphorylation motifs based on the literature. These motifs have been categorized as (i) phosphorylation-based substrate motifs (that is, motifs that are recognized by serine/threonine/tyrosine kinases or phosphatases and (ii) phosphorylation-based binding motifs (that is, motifs that get phosphorylated and act as molecular scaffolds for binding to domains that specifically bind to phosphorylated serine/threonine/tyrosine residues). **Supplementary Table 1** online lists 132 phosphotyrosine-based motifs, whereas **Supplementary Table 2** online lists 192 phosphoserine/phosphothreonine-based motifs. Original research articles describing the motif are linked to each motif. These include 24 phosphatase substrate motifs and 80 motifs that bind to domains recognizing phosphorylated serine/threonine/tyrosine residues.

Tyrosine kinases and tyrosine phosphatases, as well as SH2 and PTB domains, have been investigated in some detail over the past several years. This is both because of their relevance to growth factor–receptor signaling and the availability of good antiphosphotyrosine antibodies as tools to dissect phosphotyrosine-mediated interactions and pathways. In this compendium, we have catalogued 132 phosphotyrosine-based motifs consisting of 50 tyrosine kinase substrate motifs, 19 tyrosine phosphatase substrate motifs, 56 SH2 domain and 7 PTB domain-binding motifs (**Supplementary Table 1**).

Serine/threonine kinases phosphorylate proteins on serine/threonine residues to regulate the activity of the proteins as well as to make them available for binding to 14-3-3 proteins, WW-binding domains and Forkhead-associated (FHA), WD40 and PBD domains. We have catalogued 192 serine/threonine phosphorylation-based motifs consisting of 170 serine/threonine kinase substrate motifs, 5 serine/threonine phosphatase substrate motifs and 17 motifs that mediate binding to serine/threonine binding proteins (**Supplementary Table 2**).

To make the list of phosphorylation motifs easily accessible to the community, we have also developed a search tool designated 'PhosphoMotif Finder', which we incorporated into the Human Protein Reference Database developed earlier by



The PhosphoMotif Finder search tool at Human Protein Reference Database (<http:// www.hprd.org/ PhosphoMotif_finder>).

our group (http://www.hprd.org/)[10],which currently contains ~14,000 phosphorylation sites described in the literature. PhosphoMotif Finder pinpoints the phosphorylation-based substrate or binding motifs in any input sequence and provides the original citation in the literature that described the motif in question. It does not use any matrix-based scoring methods or algorithm-based predictions of phosphorylation sites making it different from other prediction programs mentioned above. This should be a useful resource for investigators to quickly determine whether any motif that they have discovered has been previously reported or not. PhosphoMotif Finder is especially useful for researchers who identify phosphorylation sites and would like to know if there is any specific enzyme associated with this function. The complete list of known phosphorylation motifs assembled here can also be used to refine various algorithms used to predict phosphorylation sites. These motifs can also be used to systematically test and identify binding partners[11] and for generation of phosphorylation motif–specific antibodies.

We anticipate that this resource will become an encyclopedia of known phosphorylation motifs and we will update it on a regular basis. We encourage community participation and welcome any suggestions to make this resource as comprehensive as possible.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This work was supported by a contract N01-HV-28180 from the National Heart Lung and Blood Institute and grants U54RR020839 and TR01CA106424T from the National Institutes of Health. We thank Shashikanth for assistance with the curation.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Ramars Amanchy[1], Balamurugan Periaswamy[1,2], Suresh Mathivanan[1,2], Raghunath Reddy[1,2], Sudhir Gopal Tattikota[2] & Akhilesh Pandey[1]

[1]McKusick-Nathans Institute of Genetic Medicine and the Departments of Biological Chemistry, Oncology and Pathology, Johns Hopkins University, Baltimore, Maryland 21205, USA. [2]Institute of Bioinformatics, International Tech Park, Bangalore 560 066, India.
e-mail: pandey@jhmi.edu

1. Cohen, P. *Trends Biochem. Sci.* **17**, 408–413 (1992).
2. Pawson, T., Gish, G.D. & Nash, P. *Trends Cell Biol.* **11**, 504–511 (2001).
3. Yaffe, M.B. & Elia, A.E. *Curr. Opin. Cell Biol.* **13**, 131–138 (2001).
4. Kobe, B., Kampmann, T., Forwood, J.K., Listwan, P. & Brinkworth, R.I. *Biochim. Biophys. Acta* **1754**, 200–209 (2005).
5. Obenauer, J.C., Cantley, L.C. & Yaffe, M.B. *Nucleic Acids Res.* **31**, 3635–3641 (2003).
6. Yaffe, M.B. *et al. Nat. Biotechnol.* **19**, 348–353 (2001).
7. Huang, H.D., Lee, T.Y., Tzeng, S.W. & Horng, J.T. *Nucleic Acids Res.* **33**, W226–W229 (2005).
8. Blom, N., Sicheritz-Ponten, T., Gupta, R., Gammeltoft, S. & Brunak, S. *Proteomics* **4**, 1633–1649 (2004).
9. Diella, F. *et al. BMC Bioinformatics* **5**, 79 (2004).
10. Peri, S. *et al. Genome Res.* **13**, 2363–2371 (2003).
11. Jones, R.B., Gordus, A., Krall, J.A. & MacBeath, G. *Nature* **439**, 168–174 (2006).