BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 23 2014, pages 3410–3411
doi:10.1093/bioinformatics/btu572

Sequence analysis Advance Access publication August 24, 2014

PHOXTRACK–a tool for interpreting comprehensive datasets of post-translational modifications of proteins

Christopher Weidner[†], Cornelius Fischer[†] and Sascha Sauer[*]

Otto Warburg Laboratory, Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany

Associate Editor: Igor Jurisica

ABSTRACT

Summary: We introduce PHOXTRACK (PHOsphosite-X-TRacing Analysis of Causal Kinases), a user-friendly freely available software tool for analyzing large datasets of post-translational modifications of proteins, such as phosphorylation, which are commonly gained by mass spectrometry detection. In contrast to other currently applied data analysis approaches, PHOXTRACK uses full sets of quantitative proteomics data and applies non-parametric statistics to calculate whether defined kinase-specific sets of phosphosite sequences indicate statistically significant concordant differences between various biological conditions. PHOXTRACK is an efficient tool for extracting post-translational information of comprehensive proteomics datasets to decipher key regulatory proteins and to infer biologically relevant molecular pathways.

Availability: PHOXTRACK will be maintained over the next years and is freely available as an online tool for non-commercial use at http://phoxtrack.molgen.mpg.de. Users will also find a tutorial at this Web site and can additionally give feedback at https://groups.google.com/d/forum/phoxtrack-discuss.

Contact: sauer@molgen.mpg.de.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 21, 2014; revised on August 1, 2014; accepted on August 19, 2014

Comprehensive analysis of the entire set of kinases and other enzymes inducing post-translational protein modification assists in providing a holistic view of cellular states, and supports in generating unbiased hypotheses for experimental testing. In particular, phosphorylation of proteins plays a pivotal role in the regulation of many cellular signaling and disease processes. Mass spectrometry technologies allow for comprehensive quantification of the proteome (Sauer *et al.*, 2005), including the detection of several 10 000 phosphosite sequences of cellular proteins. However, the current standard data analysis strategies applied for the discovery of important regulatory kinases, such as sequence motif analyses, are insufficient for gaining specific insights (Olsen and Mann, 2013). Moreover, these conventional approaches suffer from severe inefficiencies, as they make use of only a minor fraction of the information contained in large-scale phosphoproteome datasets (see Supplementary Background Information and results in Supplementary Figs S1–S4). A major limitation of all current phosphoproteomic data analyses consists in the need of setting a threshold for determining significantly hyper- and hypophosphorylated phosphosite (increased or decreased phosphorylation of a particular amino acid relative to the control experiment). This generally restricts the analyses on only few strongly regulated phosphosite sequence sites and completely ignores the information of the major part of the many thousands of phosphopeptides detected in biological samples (Supplementary Fig. S5).

PHOXTRACK (PHOsphosite-X-TRacing Analysis of Causal Kinases) largely overcomes these current limitations of phosphosite sequence data analysis. Notably, PHOXTRACK considers the full set of phosphosite sequences detected by mass spectrometry, resulting in powerful and largely unbiased enrichment of regulated kinases. PHOXTRACK further applies non-parametric Kolmogorov–Smirnov enrichment statistics (Meierhofer *et al.*, 2013; Subramanian *et al.*, 2005) to calculate whether defined kinase-specific sets of phosphosites indicate statistically significant concordant differences between various biological samples under investigation (Fig. 1, Supplementary Fig. S6 and Tutorial).

We initially defined experimentally validated kinase-specific phosphosite sets using data from the public databases PhosphoSitePlus (Hornbeck *et al.*, 2012), Swiss-Prot, the Human Protein Reference Database (Keshava Prasad *et al.*, 2009) and Phospho.ELM (Dinkel *et al.*, 2011), containing thousands of substrate sequence/kinase interactions derived from thousands of human and murine studies (Fig. 1 and Supplementary Table S1). Notably, the user can extend the list of kinase-specific phosphosite sets by using alternative databases including disease- and treatment-associated information.

We then used the prior defined sets of kinase-specific phosphosites and computed an enrichment score that reflects the degree to which the set of kinase-specific phosphosites is overrepresented at the top or bottom of a ranked list of detected phosphorylation ratios (Fig. 1). As shown in Supplementary Figs S7–S10, we successfully benchmarked the results obtained by PHOXTRACK using biological data from technically diverse proteomics studies focusing on cell cycle analysis, chemotherapeutic treatment of embryonic stem cells, and phosphoprofiling of human breast cancer tissues of high- versus low-risk recurrence groups. In all these studies, conventional phosphosite data analysis only roughly hinted to rather broadly defined kinase groups. Notably, PHOXTRACK indicated specifically causal kinases, which could be confirmed by additional experiments including, for example, kinase/kinase substrate immunoblotting, leading to fruitful network analyses and biological interpretation (Supplementary Results).

[*] To whom correspondence should be addressed.

[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

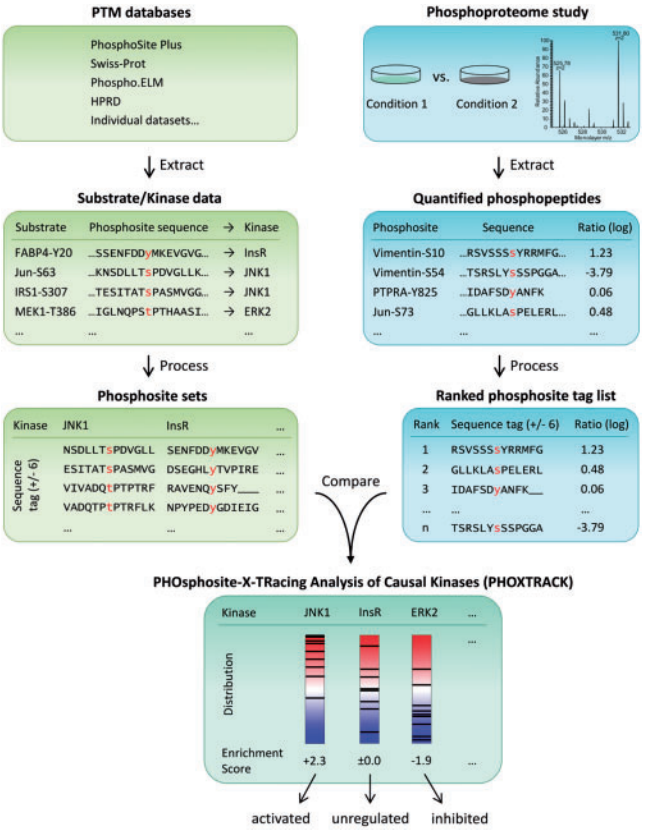
PTM databases
PhosphoSite Plus
Swiss-Prot
Phospho.ELM
HPRD
Individual datasets...
Phosphoproteome study
Condition 1
vs.
Condition 2
Extract
Extract
Substrate/Kinase data
Substrate Phosphosite sequence → Kinase
FABP4-Y20 ...SSENFDDyMKEVGVG... → InsR
Jun-S63 ...KNSDLLTsPDVGLLK... → JNK1
IRS1-S307 ...TESITATsPASMVGG... → JNK1
MEK1-T386 ...IGLNQPStPTHAASI... → ERK2
... ...
Quantified phosphopeptides
Phosphosite Sequence Ratio (log)
Vimentin-S10 ...RSVSSSsYRRMFG... 1.23
Vimentin-S54 ...TSRSLYsSSPGGA... -3.79
PTPRA-Y825 ...IDAFSDyANFK 0.06
Jun-S73 ...GLLKLAsPELERL... 0.48
... ...
Process
Process
Phosphosite sets
Kinase JNK1 InsR ...
Sequence tag (+/- 6)
NSDLLTsPDVGLL SENFDDyMKEVGV ...
ESITATsPASMVG DSEGHLyTVPIRE ...
VIVADQtPTPTRF RAVENQySFY____ ...
VADQTPtPTRFLK NPYPEDyGDIEIG ...
... ...
Compare
Ranked phosphosite tag list
Rank Sequence tag (+/- 6) Ratio (log)
1 RSVSSSsYRRMFG 1.23
2 GLLKLAsPELERL 0.48
3 IDAFSDyANFK__ 0.06
... ...
n TSRSLYsSSPGGA -3.79
PHOsphosite-X-TRacing Analysis of Causal Kinases (PHOXTRACK)
Kinase JNK1 InsR ERK2 ...
Distribution
...
Enrichment Score +2.3 ±0.0 -1.9 ...
activated
unregulated
inhibited

Fig. 1. Principle of PHOXTRACK. (Left) Thousands of experimentally verified sequence-specific kinase/substrate data were extracted from public databases. The 13-mer sequence window around the phosphorylated site of the peptide (±6) is used as identifier tag. In the current version PHOXTRACK contains 7560 human and 3177 murine unique phosphosite tags with kinase information, distributed to 1241 human and 651 murine kinase/substrate sets with up to 500 substrate phosphosites per kinase. (Right) Large-scale phosphoproteome data commonly contain several thousands of identified phosphosites of protein sequences that were relatively quantified between two conditions (e.g. treated versus non-treated samples). The user can submit a list of sequence tags (±6) assigned to measured peptide ratios (logarithmized). (Bottom) PHOXTRACK then compares the user list with the phosphosite set databases and computes an enrichment score, which reflects the degree to which the kinase-specific set of phosphosites is overrepresented at the top (red) or bottom (blue) of the ranked phosphosite list submitted by the user. PHOXTRACK thus predicts activation or inhibition of kinase activity

Additional sensitivity will be achieved by steadily increasing the quantity and quality of experimental data of phosphosite sequences and of public databases containing information on experimentally validated substrates of kinases. Currently, PHOXTRACK makes use of comprehensive assembled sequence-specific kinase substrate data derived from human and mouse samples (Supplementary Fig. S11). But the PHOXTRACK tool also enables the user to upload similar databases for other organisms.

Furthermore, extension of the PHOXTRACK tool for further post-translational modifications will significantly benefit from community efforts to add experimentally validated protein modification data (for example acetylation, methylation or ubiquitination data) to maintained publically available databases. By loading these data, PHOXTRACK can in principle be extended for inferring the biological roles of a large variety of post-translational modifications in complex datasets.

ACKNOWLEDGEMENTS

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding: The study was supported by the BMBF (0315082/01EA1303).

Conflict of interest: none declared.

REFERENCES

Dinkel,H. *et al.* (2011) Phospho.ELM: a database of phosphorylation sites—update 2011. *Nucleic Acids Res.*, **39**, D261–D267.

Hornbeck,P.V. *et al.* (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.*, **40**, D261–D270.

Keshava Prasad,T.S. *et al.* (2009) Human protein reference database—2009 update. *Nucleic Acids Res.*, **37**, D767–D772.

Meierhofer,D. *et al.* (2013) Protein sets define disease States and predict in vivo effects of drug treatment. *Mol. Cell Proteomics*, **12**, 1965–1979.

Olsen,J.V. and Mann,M. (2013) Status of large-scale analysis of post-translational modifications by mass spectrometry. *Mol. Cell. Proteomics*, **12**, 3444–3452.

Sauer,S. *et al.* (2005) Miniaturization in functional genomics and proteomics. *Nat. Rev. Genet.*, **6**, 465–476.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.