Phosphoproteomics-based network medicine

Zexian Liu, Yongbo Wang and Yu Xue

Department of Biomedical Engineering, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei, China

Keywords
kinase activity; phosphoproteomics; phosphorylation; phosphorylation-mediated signaling network; phospho-signature

Correspondence
Y. Xue, Department of Biomedical Engineering, College of Life Science and Technology, Huazhong University of Science and Technology, 1037 Luoyu Road, Wuhan, Hubei 430074, China
Fax: +86 27 87793172
Tel: +86 27 87793903
E-mail: xueyu@hust.edu.cn

(Received 2 March 2013, revised 10 May 2013, accepted 5 June 2013)
doi:10.1111/febs.12380

One of the major tasks of phosphoproteomics is providing potential biomarkers for either diagnosis or drug targets in medical applications. Because most complex diseases are due to the actions of multiple genes/proteins, the identification of complex phospho-signatures containing multiple phosphorylation events within phosphoproteomics-based networks generates more efficient and robust biomarkers than a single, differentially phosphorylated substrate or site. Here, we briefly summarize the current efforts and progress in this newly emerging field of phosphoproteomics-based network medicine by reviewing the computational (re)construction of phosphorylation-mediated signaling networks from unannotated phosphoproteomic data, the discovery of robust network phospho-signatures and the application of these signatures for classifying cancers and predicting drug responses. The challenges as well as the potential advantages are evaluated and discussed. Although the current techniques are at present far from mature, we believe that such a systematic approach as we describe can generate more useful and robust biomarkers for biomedical usage, even at the current stage of development.

Introduction

The past decade has witnessed rapid progress in phosphoproteomics, which employs state-of-the-art high-throughput mass spectrometry for the purpose of large-scale profiling of phosphopeptides in vivo [1–7]. The simultaneous identification and quantification of thousands of phosphopeptides from one sample has thus become something of a ‘routine’ assay [2,4,7]. It is widely believed that phosphoproteomics can serve as a powerful approach to clinical trials in personalized medicine by providing highly specific biomarkers and drug targets [1,5,8–10]. However, phosphoproteomic identification of single biomarkers is like searching for a ‘needle in a haystack’ for two major reasons. First, the coverage of phosphoproteomics is quite low, and any single phosphoproteomic assay can only monitor a small proportion of bona fide phosphorylation events in vivo [11–14]. Even for the same sample, the overlapping rate of two workflows with different processing requirements is usually lower than 50% [12–14]. Thus, even if one disease can actually be attributed to a single phosphorylation event, such a biomarker is difficult to reproduce technically and thus makes little practical sense for clinical trials. Second, although there are single-gene diseases, most diseases are quite complicated and multiple genes are implicated [15–17]. Thus, sophisticated approaches are needed to identify dependable and efficient biomarkers from the phosphoproteomic data.

Recently, ‘network medicine’ has emerged as a promising strategy that takes into consideration both

Abbreviations
AML, acute myeloid leukemia; DDR, DNA damage response; ESC, embryonic stem cell; GPS, Group-based Prediction System; iPSC, induced pluripotent stem cells; mTOR, mammalian target of rapamycin; NSCLC, non-small cell lung cancer; Plk, Polo-like kinase; PPI, protein–protein interaction; ssKSR, site-specific kinase–substrate relation; TK, tyrosine kinase.
key genes/proteins and their relationships in specific modules, pathways and processes [15–17]. Accumulating evidence suggests that the phosphorylation-mediated signaling network is not static but can be dynamically rewired in different samples and diseases or upon undergoing different treatments [9,18–22]. Thus, elucidating the key features of such a network would be expected to provide phospho-signatures that would be highly useful in further biomedical applications. Here, we summarize the cutting-edge advances in phosphoproteomics-based network medicine by reviewing the computational methodologies for (re)constructing phosphorylation-mediated signaling networks from phosphopeptides, discovering efficient phospho-signatures from the networks and using the signatures as potential biomarkers and/or drug targets. Although this newly emerging field is still in its infancy, we anticipate that such a systematic method will prove to be an indispensable approach for personalized medicine.

(Re)construction of phosphorylation-mediated signaling networks from phosphoproteomic data

In general there are two types of computational methodologies for phosphorylation-mediated signaling networks. The initial step for both is the mapping of all the identified phosphopeptides in order to obtain the benchmark sequence data and to designate the integrated phosphoproteins together with their exact phosphorylation sites (Fig. 1A,B).

The first approach involves directly mapping phosphoproteins to pathways or protein–protein interaction (PPI) networks (Fig. 1A). For example, Matsuoka et al. identified over 900 DNA damage response (DDR) related phosphorylation sites from > 700 proteins using an S/T-Q motif, which is the consensus sequence recognized by ATM/ATR (Ataxia telangiectasia mutated/Ataxia telangiectasia and Rad3-related) [23]. Using the two pathway analysis tools Ingenuity Pathway Analysis [24] and PANTHER [25], they obtained a number of network modules that potentially participate in DDR [23]. In 2010, Huttlin et al. systematically identified nearly 36 000 phosphorylation sites in 6296 proteins from nine different murine tissues [18] and mapped them to the STRING database, which is the most comprehensive resource of both the known and predicted PPIs [26]. To model a more integrative signaling pathway, they incorporated the phosphoproteomic data together with the known mitogen-activated protein kinase (MAPK) pathway, which was obtained from the KEGG PATHWAY database [27]. Recently, Weigand et al. quantitatively...
detected 12,669 phosphorylation sites from human MDA-MB-231 tumor xenografts treated with a humanized antibody (RG7356) for the CD44 receptor [28]. The subextractor algorithm was used to integrate the phosphorylation sites with PPI information through a Bayesian probabilistic model [29], while further analysis predicted and confirmed that CD44-expressing tumors respond to RG7356 therapy, mainly by exerting an effect on the MAPK pathway [28].

The phosphorylation of modifiable sites is the result of the actions of upstream regulatory kinases [2]. A kinase having either a high level of activity or low recognition specificity may modify a greater number of phosphorylation sites. Thus, the kinase activity and kinase–substrate relations should also be included for a more precise integrative analysis. Recently, attention has been focused on the construction of kinase-substrate networks inferred from the phosphoproteomic data [3,30,31] (Fig. 1B). First, potential regulatory kinases of phosphoproteins are predicted with a sequence-based kinase-specific predictor [3], such as SCANSITE [32], NETPHOS [33] OF Group-Based Prediction System (GPS) [34], by directly inputting protein sequences. From the result, potential site-specific kinase–substrate relations (ssKSRs) are determined and further filtered using the identified phosphorylation sites as a reference. Then, a number of contextual factors such as PPI and co-localization information on the kinase and substrate can be used to greatly reduce the false positive hits [35–38]. Because a given kinase can also be phosphorylated by another kinase, all of the predicted ssKSRs are potentially involved in the kinase-substrate networks that can be visualized using CYTOSCAPE [39] or similar tools. Linding et al. in 2007 developed a seminal algorithm, termed NETWORKIN, by integrating both motif-based predictions and direct or indirect PPIs, thus constructing a high confidence human phosphorylation network containing 7143 ssKSRs among 1759 substrates and 68 kinases for 4488 phosphorylation sites [35,36,40]. Using this powerful tool, they painstakingly modeled the phosphorylation-mediated DDR network by linking together several different biological processes involved in the response to DNA double-stranded breaks, cell cycle checkpoints and apoptosis. This method was rapidly adopted by mainstream researchers for constructing a variety of networks, including the JNK phosphorylation network [41], autophagy-associated phosphorylation networks [42], and the dynamic phosphorylation networks during human embryonic stem cell (ESC) differentiation [43] and mouse skin carcinogenesis [44].

In 2008 we developed a sequence-based tool, termed GPS, which has the capacity to hierarchically predict kinase-specific phosphorylation sites for 408 human kinases [34]. Other tools, such as SCANSITE [32] and NETPHOS [33], only predict ~30 kinases. Thus, by comparison, the performance of GPS 2.0 is better than analogous predictors [34,45]. Combining GPS predictions, PPIs and protein complexes, Bensimon et al. constructed an alternative DDR phosphorylation network and observed that ~40% of the associated phosphorylation events were ATM-independent [46]. Recently, we integrated the GPS algorithm together with the PPI information to develop a software package called in vivo GPS (in vivo GPS) for the prediction of in vivo ssKSRs [38]. We constructed eukaryotic phosphorylation networks and predicted a total of 186,922 ssKSRs among 1079 protein kinases and 9247 substrates for 44,290 phosphorylation sites in five different species (Saccharomyces cerevisiae, Caenorhabditis elegans, Drosophila melanogaster, Mus musculus and Homo sapiens). The results suggest that the PPI filter removes up to 95% of any false positive hits predicted by the GPS sequence-based algorithm [38]. In addition, by analyzing Polo-like kinase (Plk) mediated phosphorylation, it was demonstrated that the subcellular co-localization information is also an efficient contextual filter [37].

Network-based discovery of phospho-signatures

Because only a small proportion of phosphorylation sites can be identified in any single phosphoproteomic profiling, statistical approaches have been widely adopted to test whether phosphoproteins are over-represented or under-represented in distinct pathways or significantly modified by specific kinases. Thus, phosphoproteomic profiling can be described as a ‘near random sampling’. Using this approach, one pathway with more actual phosphorylation sites will be characterized by more phosphopeptides than another pathway containing fewer phosphorylation sites (Fig. 2A). Also, one kinase with a greater number of modified sites will be more frequently identified, i.e. with more hits, than another kinase having fewer modified sites (Fig. 2A). Although there are a number of intrinsic biases in the currently available techniques, this framework is nevertheless fundamental for the performance of statistical analysis.

Enrichment analysis of phosphorylation-associated pathways provides efficient and robust network phospho-signatures. For example, by comparing the phosphoproteomic data sets of ESCs and induced pluripotent stem cells (iPSCs), Phanstiel et al. observed that a number of somatic-cell-related processes are
significantly over-represented in iPSCs, and these results were shown to be consistent with additional analyses at both the transcript and protein levels [47]. Thus, it was discovered that somatic cell programs are incompletely silenced in iPSCs [47]. Also, it is possible to predict individual molecular biomarkers from the pathways or sub-networks if some of their neighbors are known disease genes [15] (Fig. 2B). For example, Matsuoka et al. determined that DDR-associated phosphorylated proteins are significantly enriched in the AKT-insulin pathway. Based on the results, they confirmed that an insulin responsive site in 4E-BP1, Ser111, was phosphorylated by ATM [23].

The available evidence suggests that the phosphoproteomic data faithfully reflect the dynamics of kinase activity in vivo. For example, by a comparison of the phosphoproteome in the presence or absence of Plk1 activity, three independent studies respectively identified 390 [19], 1071 [48] and 752 [20] differentially regulated phosphorylation sites. In total, 1979 non-redundant sites were identified, while only 26 (~1.3%) phosphorylation sites were identified in all three experiments [37]. However, the Plk1 consensus sequence N/D/E-pS was significantly over-represented in all three studies [19,20,48]. In this regard, although the overlap rate in phosphoproteomic studies is low, kinase activity analysis can be used to generate consistent results which can serve as a robust phospho-signature (Fig. 2C). Based on the hypothesis that there

is higher kinase activity when there is a greater number of modified sites, we systematically analyzed the human liver phosphoproteome and demonstrated that the activities of 60 and 67 kinases were significantly upregulated (i.e. more sites modified) and downregulated (fewer sites modified), respectively [38]. At least for the upregulated kinases, these results are highly consistent with the known data [38]. Also, Bennetzen et al. used two autophagy inducers, resveratrol and spermidine, to quantitatively identify the phosphoproteome regulated in the autophagic response [42]. Using NETWORKIN [35,36] and MOTIF-X [49], a highly effective tool for phosphorylation motif discovery, they detected the two enriched motifs S/T-P and RXXS that are recognized by CDK2 and PAK4/PAK7/DMPK/CLK1, respectively [42]. More recently, Casado et al. formally described a kinase-substrate enrichment analysis approach for predicting activated kinases in acute myeloid leukemia (AML) cells by comparing the phosphoproteome-based kinase-substrate networks obtained from control and test samples [50]. The predictions were successfully validated in cell lines by western blotting analysis of the activity-correlated autophosphorylation sites in the predicted kinases. With this method, they also determined that certain kinases, such as CDC7, PDK1 and ERK, are more active in drug-resistant primary AML cells, while Abl, Lck, Src and CDK1 are more active in drug-sensitive cells [50]. In addition, this strategy was used for

Fig. 2. Phosphoproteomic profiling and network-based discovery of phospho-signatures. (A) The 'near random sampling' hypothesis that the pathway or kinase with a greater number of phosphorylation sites will be sampled with more phosphopeptides. (B) Potentially novel molecular biomarkers can be selected in the event that their neighbors are known disease genes in the same sub-networks or pathways [15]. (C) Based on the hypothesis of higher kinase activity with a greater number of modified sites, differentially activated kinases can be statistically calculated by comparing the phosphoproteome-based kinase-substrate networks under different conditions.
analyzing the phosphoproteomic dynamics during mouse skin carcinogenesis, and the deregulated activities of PAK4, PKC and SRC were determined to be major drivers of malignancy [44].

Biomedical usage of network phospho-signatures

It is believed that phosphorylation-mediated signaling networks are rewired in a variety of diseases, and that phosphoproteomic assays have the capacity to detect at least a considerable proportion of aberrant phosphorylation events [1,10,51,52]. Thus, the identification of stable phospho-signatures is important for both early diagnosis and effective treatment. Conventional studies have mainly focused on identifying single phosphorylation events as potential biomarkers. For example, by analyzing the mTOR (the mammalian target of rapamycin) regulated phosphoproteome, Grb10 was validated as a key substrate of mTORC1, which inhibits tumor growth by phosphorylating Grb10 [53,54]. However, we speculated that a single substrate might not be sufficient as an efficient biomarker, because one kinase usually modifies multiple substrates and may therefore synergistically participate in diseases or the response to drugs. Thus, the dynamics of only one gene/protein may not be adequate for accurately classifying a given disease or predicting a drug response [9,55].

Recently, the identification of network phospho-signatures with multiple targets has emerged as an important endeavor. Drug-related genome analyses have suggested that kinases represent nearly 20% of all potential drug targets [56,57]. Indeed, a number of both serine/threonine kinases and tyrosine kinases (TKs) have been shown to be potent targets [6,10,51,54,55], whereas kinase inhibitors have come to be regarded as some of the most potentially useful drugs for cancer treatment [52,58]. Thus, the first approach is to monitor the dynamics of kinase activity in diseases or upon drug administration (Fig. 3A). Because the autophosphorylation of TKs is related to their activity, Rikova et al. performed tyrosine phosphoproteomic analyses on 41 non-small-cell lung cancer (NSCLC) cell lines and 150 NSCLC tumor samples [9]. The TK activity was estimated from the number of observed spectra and then used for the classification of tumors. One of the major findings in the study was that different combinations of activated TKs usually exist in different NSCLCs, so individualized treatments must be considered for any effective therapy [9]. However, because of the low reproducibility of the phosphoproteomic technique, the phosphopeptides of most TKs can only be detected in a small number of samples [9]. Thus, as described above, the kinase activity may be accurately estimated from the phosphorylated substrates.

The second method for obtaining network phospho-signatures is monitoring the limited substrates in specific pathways (Fig. 3B). For example, dasatinib, a multiple kinase inhibitor, has been approved as a potent drug for two types of leukemia tumors and is still undergoing evaluation for use in other cancers [55]. With a phosphoproteomic approach, Klammer et al. identified 12 phosphorylation sites in non-kinase proteins as a network phospho-signature which accurately

![Fig. 3. Biomedical use of two types of network phospho-signatures. (A) By western blotting analysis of activity-correlated autophosphorylation sites, the kinase activities can be measured and correlated with the phenotypic data. (B) The expression or phosphorylation levels of a subset of key proteins can be monitored and correlated with the phenotype data.](image-url)
predicts the dasatinib response in NSCLCs [55]. Also, Lee et al. developed a novel strategy of order- and time-dependent drug combinations for more efficiently killing triple-negative breast cancer cells [59]. The expression or phosphorylation levels of 35 proteins in related signaling pathways were quantitatively monitored and correlated with phenotype data by means of a mathematical model which accurately predicted drug-induced apoptosis in breast cancer cells [59]. In addition, this method was further refined and used for modeling of the genotoxic-stress-induced DDR network [60].

**Future perspectives**

Currently, the major bottleneck of phosphoproteomics is its low level of coverage. A single phosphorylation event that is identified in the course of one phosphoproteomic profiling may be technically difficult to repeat in another assay [11–14]. Also, multiple genes/proteins may participate in complex diseases or the response to drug treatment [15–17]. Thus, the identification of single biomarkers for accurately classifying diseases or predicting drug responses is an impractical task, at least at the present time. Since we are undoubtedly entering into an era in which network medicine will play an important role, robust network phospho-signatures are obviously needed.

For biomedical applications, recent analyses suggested that the expression or phosphorylation levels of a subset of key proteins in specific pathways or processes can be monitored through low-throughput assays [55,59,60]. Then mathematical models can be constructed to correlate the genotypes with phenotypes. These network signatures were proved to be more efficient and robust than single biomarkers [55,59,60]. However, ‘low-throughput’ usually means labor-intensive and time-consuming processes, even if they are indeed more accurate. Also, to construct such models, at least a certain number of genes/proteins should have been identified as being involved in the targeted pathways or processes. These candidates are usually selected based on already existing knowledge or high-throughput profiling results. Thus, for a less well studied sample, at least a two-step procedure should be performed, starting with a large-scale detection of potentially regulated substrates followed by small-scale quantification of their expression patterns and/or phosphorylation dynamics. If robust network phospho-signatures turn out to be retrievable from less reproducible high-throughput profiling, the procedure would be greatly simplified. Indeed, although directly monitoring kinase activity from its phosphopeptides is less dependable [9], the results of the kinase activity analysis can indeed be robustly estimated from phosphoproteome-based kinase-substrate networks, at least for Plk1 [19,20,48].

(Re)construction of kinase-substrate phosphorylation networks from the phosphoproteomic data is heavily dependent on the accurate characterization or prediction of sKSRs [3,45]. Because proteins can physically interact with kinases and yet not become phosphorylated, direct or indirect PPI information is not an optimal filter for narrowing down candidates for kinase-substrate relations. Recently, Newman et al. incubated 289 human kinases with protein chips and identified 3656 kinase-substrate relations in vitro [61]. Whether these kinase-substrate phosphorylation events really occur in vivo still remains to be determined. An accurate determination of kinase-substrate relations is still a great challenge. Also, it is estimated that there are up to 518 kinase genes encoded in the human genome, and the exact recognition motifs/patterns for a considerable number of these kinases have yet to be determined. Even our improved gcrs algorithm can still only predict kinase-specific phosphorylation sites for 408 (~79%) human kinases [34]. Thus, the prediction performance obviously needs to be improved. However, although the current approaches for the analysis of kinase activity are quite immature, they are still highly useful for elucidating phosphorylation at a system-wide level and can provide robust phospho-signatures [19,20,38,42,48].

Taken together, we anticipate that network phospho-signatures, and not single phosphorylation events, will come to be accepted as the most appropriate biomarkers and/or drug targets for biomedical usage. Kinase activity is not the only important feature that can be derived from the less reproducible phosphoproteomic data, and it is anticipated that more efficient and robust network indicators will be discovered in the near future. In addition, while the current analysis of kinase activity only processes qualitative phosphopeptide data, incorporating quantitative information will further improve the accuracy of biomarker identification.

**Acknowledgements**

The authors would like to express thanks for helpful discussions with Dr Luonan Chen (SIBS, CAS), Dr Daniel Figeys (University of Ottawa), Dr Rune Linding (C-SIG), Dr Dong Li (BPRC) and Dr Ping Xu (BPRC). We also thank Lisa Jenkins in Dr Ettore Appella’s laboratory (NIH/NCI) for carefully revising the manuscript. The authors are supported by the...
References


insights into the heterogeneity of signaling pathway activation in leukemia cells. Sci Signal 6, rs6.