Computational and Structural Biotechnology Journal 10 (2014) 90–97

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini Review

Integrating phosphoproteomics in systems biology

Yu Liu *, Mark R. Chance

Center for Proteomics and Bioinformatics, Case Western Reserve University, 10900 Euclid Ave, Cleveland, OH 44106, United States

ARTICLE INFO

Available online 1 August 2014

Keywords:
Phosphoproteomics
Systems biology
Kinases
Phosphatases
Phospho-binding proteins
Phosphorylation network
Phospho-signaling networks
Mass spectrometry

ABSTRACT

Phosphorylation of serine, threonine and tyrosine plays significant roles in cellular signal transduction and in modifying multiple protein functions. Phosphoproteins are coordinated and regulated by a network of kinases, phosphatases and phospho-binding proteins, which modify the phosphorylation states, recognize unique phosphopeptides, or target proteins for degradation. Detailed and complete information on the structure and dynamics of these networks is required to better understand fundamental mechanisms of cellular processes and diseases. High-throughput technologies have been developed to investigate phosphoproteomes in model organisms and human diseases. Among them, mass spectrometry (MS)-based technologies are the major platforms and have been widely applied, which has led to explosive growth of phosphoproteomic data in recent years. New bioinformatics tools are needed to analyze and make sense of these data. Moreover, most research has focused on individual phosphoproteins and kinases. To gain a more complete knowledge of cellular processes, systems biology approaches, including pathways and networks modeling, have to be applied to integrate all components of the phosphorylation machinery, including kinases, phosphatases, their substrates, and phospho-binding proteins. This review presents the latest developments of bioinformatics methods and attempts to apply systems biology to analyze phosphoproteomics data generated by MS-based technologies. Challenges and future directions in this field will be also discussed.

Contents

The reversible phosphorylation of amino acid side chains is one of the most common and important post-translational modifications (PTM) of proteins. Current studies are mainly focused on phosphorylation of serine (Ser), threonine (Thr), and tyrosine (Tyr), though other amino acids can also be phosphorylated such as histidine (His), aspartate (Asp), cysteine (Cys), lysine (Lys) and arginine (Arg). The latter side chain modifications are much less studied due to either lower frequency or experimental difficulties. Protein phosphorylation can result in changes of protein structure, activation/inhibition of protein activities, and promotion/prevention of protein–protein interactions [1,2]. Consequently, protein phosphorylation is one of the key regulatory mechanisms inherent in many important cellular processes, such as cell signaling, growth, and proliferation while abnormal phosphorylation can lead to serious diseases, such as cancer, diabetes and rheumatoid arthritis [3–5]. Thus, a better understanding protein phosphorylation will help to improve our knowledge of important cellular processes, provide a better understanding of disease mechanism, and drive the development of efficient treatments and new biomarker strategies.

More than 30% of human proteins are seen to be phosphorylated, these modifications often occur on multiple distinct sites [6]. Most of phosphoproteins are at substoichiometric concentrations with respect

* Corresponding author.
E-mail address: yxl442@case.edu (Y. Liu).

http://dx.doi.org/10.1016/j.csbj.2014.07.003

to site occupancy, and the phosphorylation events are usually dynamic and transient. To systematically study phosphoproteins, several high-throughput phosphoproteomic technologies, such as reverse phase protein array, phospho-specific flow cytometry, and mass spectrometry (MS) based technologies have been developed as summarized in recent reviews [7,8]. Among them, MS based technologies have become major platforms that have been routinely applied to identify phosphoproteins and phosphosites at a global, unbiased, and quantifiable level [9–13]. In a typical MS-based phosphoproteomic experiment, the procedure can be divided in four stages: sample preparation, including cell fractionation and protein digestion; enrichment of phosphopeptides via affinity purification; analysis via liquid chromatography (LC) coupled with tandem MS; and finally, localization and quantification of phosphosites/phosphoproteins using bioinformatics approach (summarized in Fig. 1). Recently, several excellent articles have reviewed the technical details of phosphoproteomic experiments, i.e., the first three steps [14–16]. In this review, we focus on the latest developments of bioinformatics approaches to annotate and integrate phosphoproteomic data, including phosphosites identification, comparative study of phosphoproteins, and construction of phosphorylation networks.

Systems biology is a relative new field that studies the properties and models of biological systems from systematic measurements, i.e., the "omics" data from high-throughput experiments. Systems biology approaches aim to reveal the property and behavior of dynamic and complex biological systems in the system level instead of individual parts [17]. Network construction and modeling are very important parts of systems biology; while signal transduction plays essential roles in the regulation and coordination of networks, and is in many cases mediated by protein phosphorylation [4,5]. Protein kinases and phosphatases are essential components of phosphorylation process: kinases add phosphate groups to their substrates; while protein phosphatases facilitate the reverse reaction. Pathways for protein phosphorylation are large and interconnected networks, involving kinases, phosphatase and their substrates. Large amounts of phosphoproteomics data generated from high-throughput experiments make it possible to construct a range of cell type, tissue, organism, and disease-specific phosphorylation networks, and they have allowed us to investigate the functional phosphorylation associated signaling states. In this review, we will present the recent progress in this exciting field.

1. Phosphosite localization

MS/MS spectra generated from phosphoproteomics experiments are used to identify phosphopeptides by applying database search tools, such as MASCOT [18] and SEQUEST [19]. However, most search engines are not designed or optimized for identification of phosphopeptides, and they don't provide reliable confidence levels for the exact localization of possible phosphosites (i.e., the identification of exact amino acids phosphorylated). This becomes crucial if there are two or more potential phosphosites in detected peptides. The situation becomes even more complicated when the phosphopeptides have low abundance and when intense neutral loss peaks in the MS/MS spectra dominate signals of interest. Statistical methods have been developed to score the reliability of phosphosite localization using one of two following strategies as reviewed in [20]: estimating the probability of each candidate phosphosite based on site-determining ions in MS/MS spectra, examples include the popular A-score [21] and PTM score [9]. Alternate methods include calculating a score based on the difference of database search outputs between different site assignments for a given phosphopeptide, as implemented in Mascot Delta Score [22] and SLIP score [23]. In this section, we present details of these methods, recent developments, and results from the comparison of different phosphosite localization methods.

Gygi and colleagues proposed the A-score, a measurement of the confidence for the correct phosphosite localization in a given peptide that has two or more potential phosphosites [21]. First, based on the peptide sequence, fragment ions that are able to uniquely assign a specific phosphosite are identified and termed "site-determining ions"; then, in the corresponding MS/MS spectrum, the site-determining ions matched peaks are identified and counted for each possible phosphosite; the cumulative binomial probability is calculated using the total number of site-determining ions and the number of matched ions; finally, the probabilities for the top two candidates are used to calculate A-score by formula $-10 \times \log(P_1) + 10 \times \log(P_2)$, where P_i is the probability for the best two candidates [21]. The PTM sore is calculated in a very similar way [9], the major difference is the selection of ions: all detected ions are used to calculate PTM score; in the case of A-score only "site-determining ions" are used. Taus et al., observed that the density of peaks in different regions is different across each

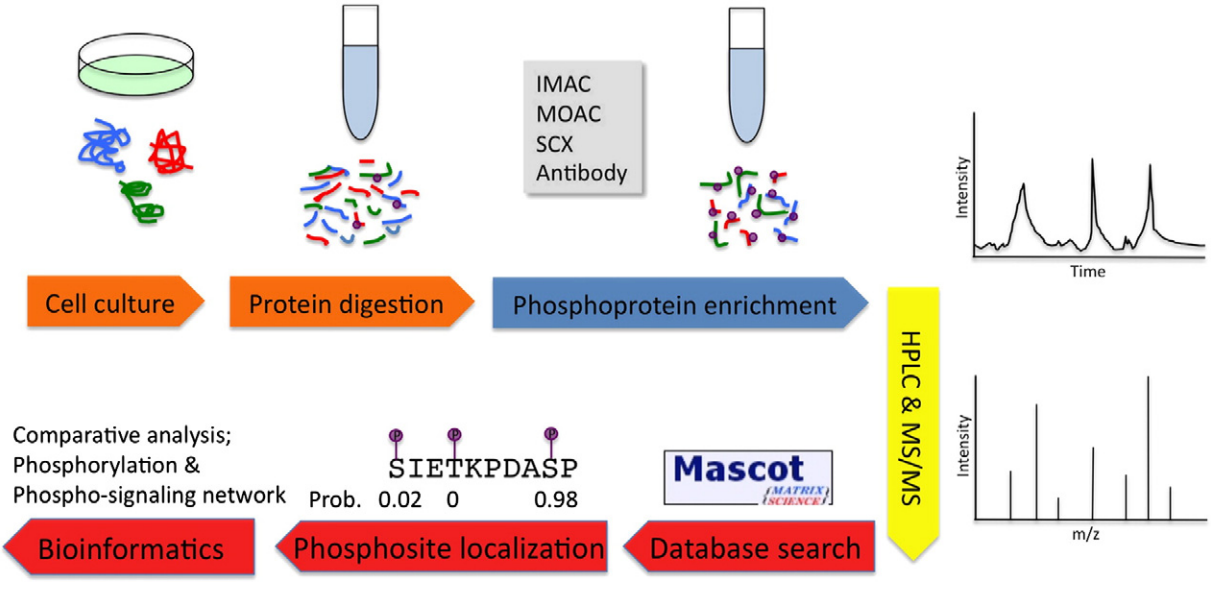

Fig. 1. Summary of MS-based phosphoproteomics experiments.