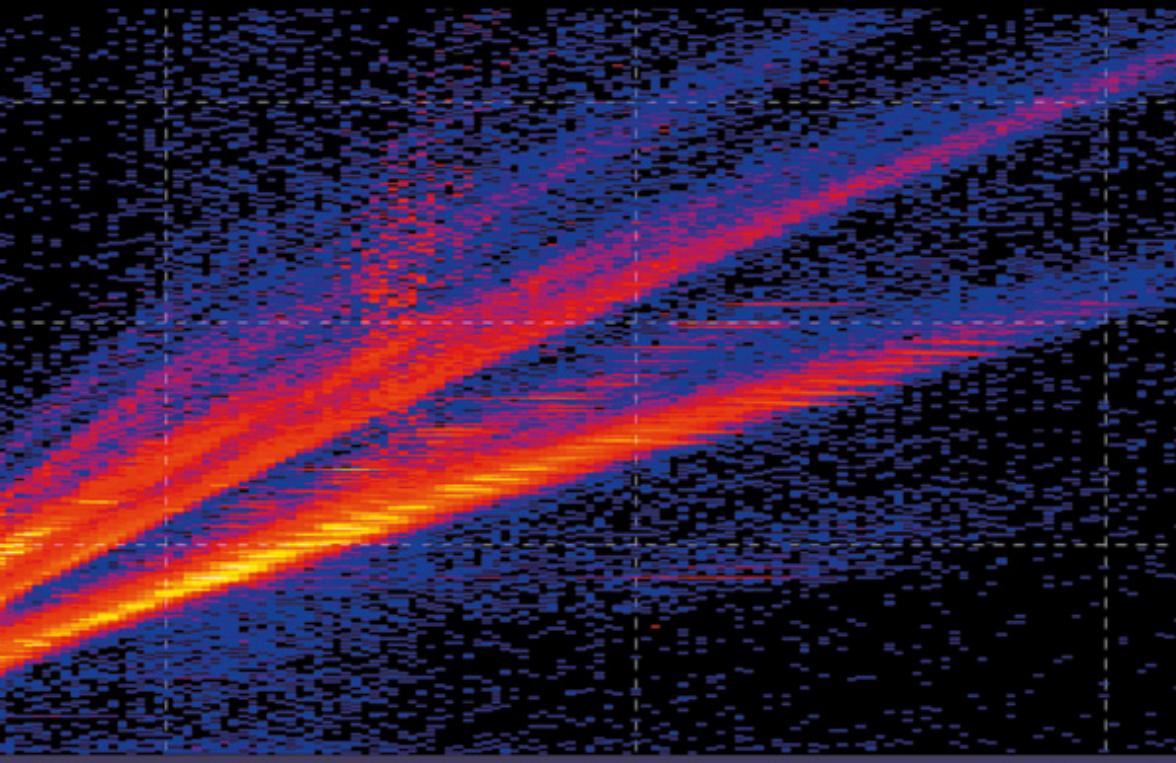
Wiley Series on Mass Spectrometry

Dominic M. Desiderio and Joseph A. Loo, Series Editors



Analysis of Protein Post-Translational Modifications by Mass Spectrometry

John R. Griffiths • Richard D. Unwin

WILEY

Analysis of Protein Post-Translational Modifications by Mass Spectrometry

WILEY SERIES ON MASS SPECTROMETRY

A complete list of the titles in this series appears at the end of this volume.

Analysis of Protein Post-Translational Modifications by Mass Spectrometry

Edited by John R. Griffiths and Richard D. Unwin

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Names: Griffiths, John R., 1964- editor. | Unwin, Richard D., editor.
Title: Analysis of Protein Post-Translational Modifications by Mass Spectrometry / edited by John R. Griffiths, Richard D. Unwin.
Description: Hoboken, New Jersey : John Wiley & Sons, 2016. | Includes bibliographical references and index.
Identifiers: LCCN 2016024928 | ISBN 9781119045854 (cloth) | ISBN 9781119250890 (epub)
Subjects: LCSH: Post-translational modification. | Mass spectrometry.
Classification: LCC QH450.6.A53 2016 | DDC 572/.645--dc23

LC record available at https://lccn.loc.gov/2016024928

Set in 10/12pt Warnock by SPi Global, Chennai, India

10 9 8 7 6 5 4 3 2 1

Contents

List of Contributors

Roland S. Annan
Proteomics and Biological Mass Spectrometry Laboratory
GlaxoSmithKline
Collegeville, PA
USA

Perdita E. Barran
Manchester Institute of Biotechnology
The University of Manchester
Manchester
UK

Navin Chicooree
Cancer Research UK Manchester Institute
The University of Manchester
Manchester, UK

Helen J. Cooper
School of Biosciences
University of Birmingham
Birmingham, UK

Andrew J. Creese
School of Biosciences
University of Birmingham
Birmingham, UK

Dominic M. Desiderio
The Charles B. Stout Neuroscience Mass Spectrometry Laboratory
Department of Neurology
University of Tennessee Health Science Center
Memphis, TN
USA

Sian Estdale
Covance Laboratories
Harrogate
UK

Caroline A. Evans
Department of Chemical and Biological Engineering
University of Sheffield
Sheffield, UK

David Firth
Covance Laboratories Ltd.
Harrogate, UK

Florian Gnad
Proteomics and Biological Resources
Genentech Inc
South San Francisco, CA
USA

John Griffiths
Cancer Research UK Manchester Institute
The University of Manchester
Manchester
UK

David J. Harvey
Department of Biochemistry
University of Oxford
Oxford
UK

Éva Hunyadi- Gulyás
Institute of Biochemistry
Biological Research Centre of the Hungarian Academy of Sciences
Szeged
Hungary

Éva Klement
Institute of Biochemistry
Biological Research Centre of the Hungarian Academy of Sciences
Szeged
Hungary

Jennie R Lill
Proteomics and Biological Resources
Genentech Inc
South San Francisco, CA
USA

Ying Long
Key Laboratory of Cancer Proteomics of Chinese Ministry of Health, Xiangya Hospital
Central South University
Changsha, Hunan
P. R. China

Dean E. McNulty
Proteomics and Biological Mass Spectrometry Laboratory
GlaxoSmithKline
Collegeville, PA
USA

Katalin F. Medzihradszky
Department of Pharmaceutical Chemistry
University of California San Francisco
San Francisco, CA
USA

Kamila J. Pacholarz
Manchester Institute of Biotechnology
The University of Manchester
Manchester
UK

Rebecca Pferdehirt
Proteomics and Biological Resources
Genentech Inc
South San Francisco, CA
USA

Naila Rabbani
Warwick Systems Biology Centre
University of Warwick
Coventry
UK

Timothy W. Sikorski
Proteomics and Biological Mass Spectrometry Laboratory
GlaxoSmithKline
Collegeville, PA
USA

Duncan L. Smith
Cancer Research UK Manchester Institute
The University of Manchester
Manchester
UK

Paul J. Thornalley
Warwick Medical School, Clinical Sciences Research Laboratories
University of Warwick
Coventry
UK

Richard Unwin
Centre for Advanced Discovery and Experimental Therapeutics (CADET)
Central Manchester University Hospitals NHS Foundation Trust
Manchester
UK

Rosie Upton
Manchester Institute of Biotechnology
The University of Manchester
Manchester
UK

Xianquan Zhan
Key Laboratory of Cancer Proteomics of Chinese Ministry of Health, Xiangya Hospital
Central South University
Changsha, Hunan
P. R. China

Preface

While preparing a recent review article in *Mass Spectrometry Reviews* on the analysis of post-translational modifications (PTMs) by mass spectrometry, we realized that, although there is much excellent work and many new tools being developed in this area, the field was lacking a coherent resource where these advances could be easily and readily accessed both by experts and those wishing to begin such studies. We subsequently decided that there was a need for a more comprehensive description of some of these modifications, and their analysis by mass spectrometry, in the form of a textbook. Since a detailed discussion of multiple modifications was required, it rapidly became apparent that this would require the support of experts in their own specialized fields. We are, therefore, grateful that a number of mass spectrometrists from around the world whom we, and others involved in proteomics, consider to be experts in the analysis of specific PTMs, agreed to contribute to this effort.

The aim of the book is to provide the reader with an understanding of the importance of the protein modifications under discussion in a biological context, and to yield insights into the analytical strategies, both in terms of sample preparation, chemistry, and analytical considerations required for the mass spectrometric determination of the presence, location, and function of selected important PTMs.

The scene is ably set with a concise introduction to the general strategies employed in PTM analysis by mass spectrometry, covering some of the key technologies which are referred to in more detail in subsequent chapters. Of course, well-known and more thoroughly investigated modifications such as phosphorylation, glycosylation, and acetylation are described in this work in great detail. However, other PTMs are garnering interest within the field and play major roles in protein function both in normal cellular regulation and in the disease setting. These PTMs are generally less well studied to date, and include, for example, tyrosine sulfation, glycation, nitration, and citrullination – the conversion of arginine to citrulline. The analysis of ubiquitination and SUMOylation, both of which involve the addition of a second, small protein to the target in a complex regulation of protein localization, activity, and

stability completes the array of modifications included in this book. In addition, the book rounds off with a description of one of the current "hot topics" in mass spectrometry: that of top-down studies of intact protein structure and modification, using the example of the characterization of monoclonal antibodies.

As editors, it has been our joint pleasure and privilege to have been given the opportunity to read at first hand these works and to compile them into a book of which we are very proud. On behalf of both of us we would like to express our sincere thanks and appreciation for the hard work and generosity given by all of the contributors.

Finally, to you the reader, we hope that you are able to use this book in your research, either as a reference book to dip into from time to time, to introduce you to new methodologies or new ideas to help support your work, or as a means of gaining a greater understanding of the analysis of PTMs by mass spectrometry from some expert scientists.

April 2016

John R. Griffiths
Richard D. Unwin
Manchester, UK

1

Introduction

Rebecca Pferdehirt, Florian Gnad and Jennie R. Lill

Proteomics and Biological Resources, Genentech Inc., South San Francisco, CA, USA

1.1 Post-translational Modification of Proteins

While the human proteome is encoded by approximately 20,000 genes [1, 2], the functional diversity of the proteome is orders of magnitude larger because of added complexities such as genomic recombination, alternative transcript splicing, or post-translational modifications (PTMs) [3, 4]. PTMs include the proteolytic processing of a protein or the covalent attachment of a chemical or proteinaceous moiety to a protein allowing greater structural and regulatory diversity. Importantly, PTMs allow for rapid modification of a protein in response to a stimulus, resulting in functional flexibility on a timescale that traditional transcription and translation responses could never accommodate. PTMs range from global modifications such as phosphorylation, methylation, ubiquitination, and glycosylation, which are found in all eukaryotic species in all organs, to more specific modifications such as crotonylation (thought to be spermatozoa specific) and hypusinylation (specific for EIF5a), which govern more tight regulation of associated proteins. Taken together, over 200 different types of PTMs have been described [5], resulting in an incredibly complex repertoire of modified proteins throughout the cell.

The addition and subtraction of PTMs are controlled by tight enzymatic regulation. For example, many proteins are covalently modified by the addition of a phosphate group onto tyrosine, serine, or threonine residues in a process called phosphorylation [6]. Phosphorylation is catalyzed by a diverse class of enzymes called kinases [7], whereas these phosphomoieties are removed by a second class of enzymes referred to as phosphatases. The tight regulation of kinases and phosphatases often creates "on/off" switches essential for regulation of sensitive signaling cascades. There are some exceptions to this rule however, and the hunt is still underway for the ever-elusive hypusine [8]

Analysis of Protein Post-Translational Modifications by Mass Spectrometry,
First Edition. Edited by John R. Griffiths and Richard D. Unwin.

removing enzyme or putative enzymes responsible for the removal of protein arginine methylation. However, it is also possible that proteins bearing these PTMs are modulated or removed from the cell by other mechanisms of action. For example, proteolysis is rarely (if ever) reversible, and many proteins (e.g., blood clotting factors and digestive enzymes) are tightly governed by irreversible cleavage events where the active form is created after proteolysis of a proenzyme.

While PTMs such as phosphorylation and lysine acetylation exist in a binary "on/off" state, many other PTMs exhibit much more complex possible modification patterns. For example, lysine residues can be modified by covalent attachment of the small protein ubiquitin, either by addition of a single ubiquitin or by addition of ubiquitin polymers. In the latter case ubiquitin itself is used as the point of attachment for addition of subsequent ubiquitin monomers [9]. To add another layer of complexity, ubiquitin has seven lysines (K6, K11, K27, K29, K33, K48, and K63), each of which may be used as the point of polyubiquitin chain linkage, and each of which has a different functional consequence. For example, K63-linked chains are associated with lysosomal targeting, whereas K48-linked chains trigger substrate degradation by the proteasome. Thus, even within one type of PTM, multiple subtypes exist, further expanding the functional possibilities of protein modification.

In addition, many proteins are modified on multiple residues by different types of PTMs. A classic example is the PTM of histones. Histones are nuclear proteins that package and compact eukaryotic DNA into structural units called nucleosomes, which are the basic building blocks of chromatin and essential for regulation of gene expression. The C-termini of histones are composed of unstructured tails that protrude from nucleosomes and are heavily modified by methylation, acetylation, ubiquitylation, phosphorylation, SUMOylation, and other PTMs [10]. Overall, 26 modified residues on a single-core histone have been identified, and many of these residues can harbor multiple PTM types. In a generally accepted theory referred to as the "histone code," the combination of PTMs on all histones comprising a single nucleosome or group of nucleosomes regulates fine-tuned expression of nearby genes.

As we begin to uncover the modified proteome, the importance of the interplay between multiple different PTMs has become increasingly apparent. One classic example is the involvement of both protein phosphorylation and ubiquitylation in the regulation of signaling networks [11]. Protein phosphorylation commonly promotes subsequent ubiquitylation, and the activities of ubiquitin ligases are also frequently regulated through phosphorylation. In a recent study by Ordureau et al., quantitative proteomic studies were employed to describe the PINK1 kinase–PARKIN UB ligase pathway and its disruption in Parkinson's disease [12]. The authors describe a feedforward mechanism where phosphorylation of PARKIN by PINK1 occurs upon mitochondrial damage, leading to ubiquitylation of mitochondria and mitochondrial proteins by PARKIN. These

newly formed ubiquitin chains are then themselves phosphorylated by PINK1, which promotes association of phosphorylated PARKIN with polyubiquitin chains on the mitochondria, and ultimately results in signal amplification. This model exemplifies how intricate interactions between multiple different PTMs regulate protein localization, interactions, activity, and ultimately essential cellular processes.

Recent advances in mass spectrometry methods, instrumentation, and bioinformatics analyses have enabled the identification and quantification of proteome-wide PTMs. For example, it is now a common practice to identify ten thousand phosphorylation sites in a single phosphoproteome enrichment experiment [13]. In addition, precise quantitation allows a deeper understanding of the combinations and occupancy of PTMs within a given protein. Such MS-based PTM analyses have led to previously impossible discoveries, advancing our understanding of the role of PTMs in diverse biological processes.

1.2 Global versus Targeted Analysis Strategies

Detection of PTMs by mass spectrometry can be achieved via global or targeted methods. The biological pathway of interest usually determines the type of PTM to be analyzed and associated methods. In a more targeted approach, researchers decide to investigate PTMs, because a protein of interest shows a higher than expected molecular weight or multiple bands by western blot after application of a stimulus, thus prompting speculation as to whether this could be due to PTM. Either way, the first step in PTM mapping is to determine the type of PTM of interest. In some cases the observed mass shift in a mass spectrometer indicates a certain PTM type. Many PTMs, however, result in the same mass addition (e.g., +42 Da for both acetylation and trimethylation). One powerful strategy in determining PTM identity involves the employment of the enzymes responsible for PTM removal. For example, after antibody enrichment of a modified protein, the antibody-bound protein can be incubated with general phosphatases, deubiquitinating enzymes (DUBs), or deSUMOylating enzymes (SENPs), and PTM removal can be assayed by western blot. Another method for PTM identification is western blotting with PTM specific "pan-antibodies." Many commercially available antibodies exist for this purpose, recognizing common PTMs such as acetylation, methylation, ubiquitylation, and phosphorylation or even more rare PTMs such as crotonyl-, malonyl- or glutaryl-lysine modification. Once the type of PTM that is decorating a protein has been identified, the next step is to attempt to map the amino acid residue(s) that bear this modification.

One of the first applications of mass spectrometry in protein research was the mapping of a PTM on a single protein [14]. A commonly used approach

involves protein-level immunoprecipitation followed by separating the captured proteins by SDS-PAGE, excising the higher molecular weight band, and performing in-gel tryptic digestion followed by LC-MS/MS. By searching for mass shifts indicative of the suspected modification(s), PTM-containing peptides can be identified and the PTM site mapped back to the protein. The strategy of identifying proteins in complex mixtures by digesting them into peptides, sequencing the resulting peptides by tandem mass spectrometry (MS/MS), and determining peptide and protein identity through automated database searching is referred to as shotgun proteomics and is one of the most popular analysis strategies in proteomics [15]. This protein-level enrichment approach, however, is dependent on sufficient levels of the modified protein compared to unmodified and the availability of protein-specific antibodies for immunoprecipitation. It is also possible that modifications may occur within the antibody epitope, blocking enrichment of the modified form altogether.

Researchers are commonly interested in analyzing PTMs from a complex mixture of proteins rather than on only one substrate. This can be a challenge, since modified peptides often occur in substoichiometric levels compared to unmodified versions and also may ionize less efficiently by electrospray ionization (ESI). However, several enrichment strategies exist, allowing for reduction of sample complexity and easier detection of the modified peptide species. Peptide-level immunoprecipitation using antibodies specific to a given PTM is an increasingly popular method of enrichment prior to MS. While this strategy can be employed for any PTM enrichment, it has been most commonly used for mapping ubiquitination sites. Tryptic digestion of ubiquitinated proteins generates a diglycine remnant attached to the ubiquitinated lysine residue (K-GG) that can be recognized by antibodies. The resulting mass shift of +114.0429 Da can be detected by MS/MS. Not only has K-GG peptide immunoaffinity enrichment enabled the identification of hundreds of ubiquitination sites on a global level but it has also been shown to enhance identification of ubiquitination sites on individual proteins, when compared to protein-level IP coupled with MS/MS [16].

To understand the biological significance of a specific PTM, it is also important to determine the PTM site occupancy or percentage of a protein's total population that is modified. Quantification of site occupancy can be accomplished by combining antibody peptide enrichment with stable isotope-labeled internal standards of the same sequence, a method termed stable isotope standards and capture by anti-peptide antibodies (SISCAPA) [17]. By coupling immunoprecipitation with stable isotope dilution multiple reaction monitoring (SID-MRM), absolute quantitation of both modified and unmodified protein populations can be determined in a high-throughput, multiplexing-compatible fashion [18].

In addition to antibody-based enrichment approaches, several strategies for chemical enrichment of PTM-containing subproteomes have been developed.

These approaches can also be coupled with the use of stable isotope standard peptides and SRM/MRM for accurate quantification of PTM dynamics. The most widely studied PTM, with the most variety of enrichment methods available, is phosphorylation. Global analysis of serine, threonine, and tyrosine phosphorylation can be achieved by a combination of peptide fractionation using strong cation exchange (SCX) followed by further enrichment with immobilized metal affinity chromatography (IMAC). The SCX/IMAC approach allows for enrichment of phosphorylated peptides to over 75% purity and ultimately identification of over 10,000 phosphorylation sites from 5 mg of starting protein [13, 19]. Another common approach for selective enrichment of the phosphoproteome is using metal oxide affinity chromatography (MOAC) such as titanium dioxide (TiO_2) [20] or aluminum hydroxide ($Al(OH_3)$) [21]. MOAC methods have been reported to achieve higher sensitivity than IMAC (at the cost of lower specificity though). The combination of multiple enrichment approaches may ultimately be the best approach.

Phosphopeptide enrichment strategies can also be applied on crude protein extract to enrich for entire phosphoproteins. Enriched fractions are typically separated by two-dimensional gel electrophoresis (2D-GE) or sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE). In either case, each observed protein spot/band is quantified by its staining intensity, and selected spots/bands are excised, digested, and analyzed by MS. The advantage of phosphoprotein enrichment is that intact proteins are separated, and the molecular weight and isoelectric point of proteins can be determined. This greatly aids in protein identification by MS. However, protein-level enrichment has several disadvantages, including loss of small or hydrophobic proteins during precipitation steps, less specific enrichment when compared to phosphopeptides, and difficulty in identifying low-abundance proteins or modifications [22].

In summary, both targeted and global methods for PTM identification have been significantly tuned in recent years but are still facing challenges. The choice of method is usually dictated by the biological question. However, global strategies are becoming increasingly popular due to their versatility, sensitivity, and ability to collect a wealth of data, triggering new hypotheses that ask for validation by targeted experiments.

1.3 Mass Spectrometric Analysis Methods for the Detection of PTMs

Mass spectrometers are powerful, analytical tools that have evolved rapidly over the past few decades to become the instrument of choice for protein and peptide characterization. Mass spectrometry is often used in parallel to other techniques such as western blot analysis or protein microarrays for detecting

and quantifying PTMs. One of the main advantages of mass spectrometry is the ability to rapidly analyze many samples in a high-throughput manner. Mass spectrometric analyses can be divided into three main strategies: "bottom-up," "middle-down," and "top-down" proteomic approaches [23]. Laboratories typically employ bottom-up proteomic methodologies to characterize PTMs. Proteins of interest are purified and proteolytically digested with an enzyme such as trypsin, with resultant peptides being separated by reversed-phase chromatography or another analytical method compatible with mass spectrometric analysis. One of several fragmentation methods and ion detection methodologies can then be employed (see Sections 1.3.1–1.3.4 for description of the various types of bottom-up proteomic analyses). It is common to associate "data-dependent" MS/MS analysis with bottom-up approaches, where resulting peptide spectra are then pieced back together *in silico* to give an overview of the protein and its PTMs.

In top-down proteomics, intact protein ions or large protein fragments are subjected to gas-phase fragmentation for MS analysis. Here, a variety of fragmentation mechanisms can be employed to induce dissociation and mass spectrometric analysis of the protein including collision-induced dissociation (CID), electron transfer dissociation (ETD), and electron capture dissociation (ECD) [24–26]. High-resolution mass detectors such as the quadrupole–time of flight (Q-TOF), Fourier transform ion cyclotron resonance (FT-ICR), or orbitrap mass spectrometers are typically employed as the spectra generated from top-down fragmentation tend to be highly charged and therefore difficult to resolve without high-resolution power. Top-down proteomics to date has been a less popular tool for characterizing PTMs than bottom-up analysis. However, it is an invaluable tool in cases where a bottom-up approach would lose contextual information about combinatorial PTM distribution (e.g., in the case of histone PTM analysis [27]). The middle-down approach has more commonly been employed as a strategy whereby a proteolytic enzyme can be used to generate longer polypeptides from a protein of interest and has shown utility in analyzing complex PTMs such as the histone code [28, 29]. Compared to middle-down and top-down methods, the bottom-up approach often offers better front-end separation of peptides, typically equating to higher sensitivity and selectivity. There are however some limitations to the bottom-up approach including the risk of low sequence coverage, particularly when employing a single proteolytic enzyme such as trypsin where cleavage may result in peptides yielding chemophysical properties with poor analytical attributes, such as size or substandard hydrophobicity.

1.3.1 Data-Dependent and Data-Independent Analyses

The type of mass spectrometric analysis performed for PTM detection depends on whether a single protein with a single PTM is being analyzed or if it is a

global approach, such as a global phosphoproteomic analysis. When targeting a single protein or a subset of proteins for a PTM of interest, a straightforward strategy is to perform an enzymatic digestion followed by data-dependent MS/MS analysis of peptides. In this approach, the intact molecular weight of each peptide in the full MS scan is analyzed, and then a selection of the most abundant peptides in the full MS scan are sequentially selected for fragmentation using one of several fragmentation methods. The resulting spectra are then analyzed either through *de novo* sequencing or more commonly using a search algorithm such as SEQUEST [30], Mascot [31], or Andromeda [32]. Peptides are then scored using an algorithm to calculate the false discovery rate or validated through manual spectral interpretation or by incorporation of a synthetic standard.

In traditional data-dependent acquisition (DDA), a proteomic sample is digested into peptides, separated often by reversed-phase chromatography, and ionized and analyzed by mass spectrometry. Typically instruments are programmed to select any ions that fall above a certain intensity threshold in full MS for subsequent MS/MS fragmentation. Although a powerful and highly utilized technique, the method is indeed biased to peptides that are of higher abundance, and lower level moieties such as post-translationally modified peptides may go undetected using DDA. Several years ago an alternative methodology called data-independent acquisition (DIA) was introduced which has slowly been gaining momentum [33]. In DIA analysis, all peptides within a defined mass-to-charge (m/z) window are subjected to fragmentation; the analysis is repeated as the mass spectrometer walks along the full m/z range. This results in the identification of lower level peptides, for example, post-translationally modified species present at substoichiometric levels compared to their nonmodified counterparts. It also allows accurate peptide quantification without being limited to profiling predefined peptides of interest and has proved useful in the biomarker community where quantitation on complex samples is routinely employed. The DIA method has matured in terms of utility over the past few years with the introduction of more user friendly and accurate search algorithms and spectral library search capabilities [34, 35]. Its utility as a tool to identify complex, low level, and isobaric amino acids has also recently been reported [36, 37].

1.3.2 Targeted Analyses

In addition to data-dependent approaches, targeted methods also exist whereby specific ion transitions can be monitored. These various targeted approaches are summarized in Figure 1.1, each of which has been employed to characterize post-translationally modified peptides.

Precursor ion scanning (PIS) is a sensitive mode of mass spectrometric operation primarily performed on triple quadrupole instruments, which has been

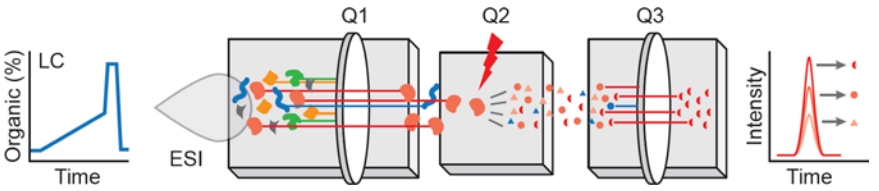


Figure 1.1 The selected reaction monitoring technique. Molecular ions of a specific analyte are selected in Q1 and fragmented in Q2. electrospray ionization (ESI). Molecular ions of one or several contaminants are isolated and fragmented together. A specific fragment ion from the target analyte (transition) is selected in Q3 and guided to the detector. The number of target fragment ions is counted over time, resulting in an SRM trace for each transition. On the far right, cycles through three transitions, corresponding to three different fragments of the target analyte, and the corresponding three SRM traces are shown. Source: Picotti and Aebersold 2012 [38]. Reproduced with permission from Macmillan Publishers Ltd.

employed for the analysis of predefined PTMs. In PIS, the third quadrupole of a triple quadruple mass spectrometer is fixed on a selected *m/z*, typically that being a neutral loss ion, for example, 79 Da for a phosphate anion observed in negative ion mode of detection, whereby the (PO_3) species is derived from the CID of phosphorylated moieties [39]. This method is highly selective and sensitive and has been applied to other PTMs beyond the analysis of phosphopeptides. Another targeted method traditionally employed for identification of post-translationally modified peptides is neutral loss scanning (NLS) [40]. NLS experiments monitor all pairs of precursor ions and product ions that differ by a constant neutral loss consistent with the PTM of interest. However, with the exponential improvements in speed and sensitivity for instruments such as the orbitrap and Q-TOF, these methods are less commonly employed than several years ago.

1.3.3 Multiple Reaction Monitoring

Multiple reaction monitoring (MRM), also known as selected reaction monitoring (SRM), is a targeted mass spectrometric methodology that is not limited to the analysis of PTM modified peptides but has been used extensively as a sensitive method to analyze various types of peptides. In MRM analyses MS/MS is applied to detect and quantify selected peptides of interest, such as those previously identified in differential discovery studies or specific post-translationally modified forms of a known peptide. Here, the specificity of precursor to product transitions is harnessed for quantitative analysis of multiple proteins in a single sample. Software tools such as MRMaid [41] or Skyline [42] allow rapid MRM transition generation and method construction for targeted analyses.

1.3.4 Multiple Reaction Monitoring Initiated Detection and Sequencing

Multiple reaction monitoring-initiated detection and sequencing (MIDAS™) [43] has been a well-utilized method for the analysis of PTM modified peptides with application to acetylated [44], phosphorylated [45], and ubiquitinated [46] species. MIDAS is a hypothesis-driven approach that requires the primary sequence of the target protein to be known and a proteolytic digest of this protein to be performed. MIDAS allows one to perform a targeted search for the presence of post-translationally modified peptides with detection based on the combination of the predicted molecular weight (measured as mass–charge ratio) of the PTM modified proteolytic peptide and a diagnostic fragment which is generated by specific fragmentation of modified peptides during CID performed in MS/MS analysis. Sequence information is subsequently obtained which enables PTM site assignment.

1.4 The Importance of Bioinformatics

The ultimate goal of proteomics is to obtain a picture of the entire complement of proteins without gaps. Genomics has already achieved this goal at the level of DNA and RNA by mapping complete genotypes. Proteomics, however, aims to describe phenotypes that display a significantly more complex functional diversity in a dynamic environment. Historically, proteomics tried to approach this challenge by establishing comparably primitive approaches such as two-dimensional gels, which gave the genomics field a competitive edge. In the last decade, however, mass spectrometry has become the method of choice, and recent advances allow the measurement of expression and modification states of thousands of proteins in a single experiment. In the last few years, the number of identified PTM sites, in particular, phosphorylation sites, has increased up to 100-fold [47]. Furthermore, mass spectrometry enables the reconstruction of protein interactions in networks and complexes. Shotgun proteomics is the most widely used approach generating thousands of spectra per hour. Therefore computational methods have to face a huge amount of generated data and a combinatorial explosion in the number of potential molecular states of proteins. In the early era of mass spectrometry as a high-throughput technology, computational analysis was commonly considered the "Achilles heels of proteomics" [48] because of the alarmingly high false discovery rates accompanied with the absence of adequate statistical methods. Fortunately, the establishment of stringent standards by the community [49] and the development of robust computational methods dragged the false discovery rates down to one percent and reduced the fraction of unassigned spectra to 10% [50].

The primary problem that all computational approaches try to solve is to assign a given MS/MS spectrum to a peptide sequence within the shortest amount of time. The most common approach is to generate theoretical fragment masses for candidate peptides from a specified protein sequence database and map these against experimental spectra. The pool of possible peptides is mainly defined by the proteolytic enzyme, mass tolerance, and specified PTM. Numerous software tools have been developed to this end [51], and they mainly differ in scoring the similarities between calculated and experimental spectra and in the statistical validation of results. SEQUEST [30] is one of the first and most commonly used tools for MS/MS-based proteomics. Its scoring scheme is based on spectral correlation functions that basically count "matched peaks," defined as the number of fragment ions common between the computed and experimental spectra. Mascot [31] extends this approach by estimating the probability of observing the shared peak count by chance. Because Mascot is a commercial software, the underlying algorithms are not provided. The search engine Andromeda [32], which is integrated into the freely available MaxQuant platform [52], also employs probabilistic scores. Notably, because selection of precursor ion for fragmentation is performed with low resolution to ensure high sensitivity, coeluting peptides with similar masses are frequently cofragmented. While the resulting "chimerical" MS/MS spectra [53] usually distort the detection and quantitation of peptides, Andromeda includes an algorithm that detects the "second" peptide and uses this information to increase the identification rate.

Other computational tools such as Protein Prospector [54] employ empirical scoring schemes that incorporate the number of matched peaks as well as the fraction of total peak intensities that can be explained by them. But when it comes to the identification of PTM sites, all methods face the same issue of the combinatorial explosion of theoretical peptides in cases where too many variable modification types are allowed. Consequently, spectra-to-peptide searches are usually restricted to up to three modifications. However, Byonic [55], which is also based on the principle of matching experimental to theoretical spectra, allows a larger number of modification types by setting an upper limit on the total occurrence of each modification. Furthermore, Byonic provides "wild-card" searches that allow the detection of unanticipated modifications by searching within specified mass delta windows.

In addition to the combinatorial explosion of theoretical peptides, another challenge in the analysis of PTMs is the precise localization of PTMs within peptides. Since PTM sites of the same protein commonly display distinct behaviors [56], it is imperative to determine their exact localizations. To this end, Ascore [57] assesses the probability of correct site localization based on the presence and intensity of site-determining ions. The corresponding algorithm essentially reflects the cumulative binomial probability of identifying site-determining ions. The same concept is used by the "localization probability score," [56] which is integrated into MaxQuant.

After the identification of peptides and associated PTMs, output scores of database search tools are translated into estimated false discovery rates. To this end, "target-decoy searching" [58] is commonly applied. The main idea of this approach is to search MS/MS spectra against a target database that contains protein sequences and reversed counterparts. Under the assumption that false matches to sequences from the original database and matches to decoy peptide sequences follow the same distribution, peptide identifications are filtered using score cutoffs corresponding to certain FDRs.

Taken together, technological advances and accompanied developments of computational methods now allow the routine identification of thousands of proteins, including PTM sites, giving a global and hopefully soon a complete picture of the proteome. Bioinformatics approaches have mastered many problems in the analysis of proteomics data but are still facing several challenges including the decryption of unmatched spectra. The accumulation of detected PTM sites across studies has been managed by various databases, including UniProt (www.uniprot.org) [2], PhosphoSite (www.phosphosite.org) [59], and PHOSIDA (www.phosida.com) [60].

Acknowledgements

We thank Allison Bruce from Genentech for help with the illustration.

References

1 Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A* 2007;**104**:19428–19433.

2 Consortium TU. UniProt: a hub for protein information. *Nucleic Acids Res* 2015;**43**:D204–D212.

3 Ayoubi TA, Ven WJVD. Regulation of gene expression by alternative promoters. *FASEB J* 1996;**10**:453–460.

4 Jensen ON. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Curr Opin Chem Biol* 2004;**8**:33–41.

5 Walsh C. *Post-translational Modification of Proteins: Expanding Nature's Inventory*. Roberts and Company Publishers; 2006.

6 Johnson LN. The regulation of protein phosphorylation. *Biochem Soc Trans* 2009;**37**:627–641.

7 Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. *Science* 2002;**298**:1912–1934.

8 Dever TE, Gutierrez E, Shin B-S. The hypusine-containing translation factor eIF5A. *Crit Rev Biochem Mol Biol* 2014;**49**:413–425.

9 Komander D, Rape M. The ubiquitin code. *Annu Rev Biochem* 2012;**81**:203–229.

10 Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. *Cell Res* 2011;**21**:381–395.

11 Hunter T. The age of crosstalk: phosphorylation, ubiquitination, and beyond. *Mol Cell* 2007;**28**:730–738.

12 Ordureau A, Sarraf SA, Duda DM, Heo J-M, Jedrychowski MP, Sviderskiy VO, Olszewski JL, Koerber JT, Xie T, Beausoleil SA, Wells JA, Gygi SP, Schulman BA, Harper JW. Quantitative proteomics reveal a feedforward mechanism for mitochondrial PARKIN translocation and ubiquitin chain synthesis. *Mol Cell* 2014;**56**:360–375.

13 Villén J, Gygi SP. The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nat Protoc* 2008;**3**:1630–1638.

14 Stenflo J, Fernlund P, Egan W, Roepstorff P. Vitamin K dependent modifications of glutamic acid residues in prothrombin. *Proc Natl Acad Sci U S A* 1974;**71**:2730–2733.

15 Nesvizhskii AI. Protein identification by tandem mass spectrometry and sequence database searching. In: *Mass Spectrometry Data Analysis in Proteomics*. New Jersey: Humana Press; 2006. p 87–120.

16 Anania VG, Pham VC, Huang X, Masselot A, Lill JR, Kirkpatrick DS. Peptide level immunoaffinity enrichment enhances ubiquitination site identification on individual proteins. *Mol Cell Proteomics* 2014;**13**:145–156.

17 Anderson NL, Anderson NG, Haines LR, Hardie DB, Olafson RW, Pearson TW. Mass spectrometric quantitation of peptides and proteins using Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA). *J Proteome Res* 2004;**3**:235–244.

18 Kuhn E, Addona T, Keshishian H, Burgess M, Mani DR, Lee RT, Sabatine MS, Gerszten RE, Carr SA. Developing multiplexed assays for troponin I and interleukin-33 in plasma by peptide immunoaffinity enrichment and targeted mass spectrometry. *Clin Chem* 2009;**55**:1108–1117.

19 Gruhler A, Olsen JV, Mohammed S, Mortensen P, Færgeman NJ, Mann M, Jensen ON. Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol Cell Proteomics* 2005;**4**:310–327.

20 Larsen MR, Thingholm TE, Jensen ON, Roepstorff P, Jørgensen TJD. Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. *Mol Cell Proteomics* 2005;**4**:873–886.

21 Wolschin F, Wienkoop S, Weckwerth W. Enrichment of phosphorylated proteins and peptides from complex mixtures using metal oxide/hydroxide affinity chromatography (MOAC). *Proteomics* 2005;**5**:4389–4397.

22 Fíla J, Honys D. Enrichment techniques employed in phosphoproteomics. *Amino Acids* 2012;**43**:1025–1047.

23 Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature* 2003;**422**:198–207.

24 McLafferty FW, Horn DM, Breuker K, Ge Y, Lewis MA, Cerda B, Zubarev RA, Carpenter BK. Electron capture dissociation of gaseous multiply charged ions by Fourier-transform ion cyclotron resonance. *J Am Soc Mass Spectrom* 2001;**12**:245–249.

25 Syka JEP, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A* 2004;**101**:9528–9533.

26 Wells JM, McLuckey SA. Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol* 2005;**402**:148–185.

27 Han J, Borchers CH. Top-down analysis of recombinant histone H3 and its methylated analogs by ESI/FT-ICR mass spectrometry. *Proteomics* 2010;**10**:3621–3630.

28 Moradian A, Kalli A, Sweredoski MJ, Hess S. The top-down, middle-down, and bottom-up mass spectrometry approaches for characterization of histone variants and their post-translational modifications. *Proteomics* 2014;**14**:489–497.

29 Sidoli S, Lin S, Karch KR, Garcia BA. Bottom-up and middle-down proteomics have comparable accuracies in defining histone post-translational modification relative abundance and stoichiometry. *Anal Chem* 2015;**87**:3129–3133.

30 Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 1994;**5**:976–989.

31 Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999;**20**:3551–3567.

32 Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* 2011;**10**:1794–1805.

33 Egertson JD, Kuehn A, Merrihew GE, Bateman NW, MacLean BX, Ting YS, Canterbury JD, Marsh DM, Kellmann M, Zabrouskov V, Wu CC, MacCoss MJ. Multiplexed MS/MS for improved data independent acquisition. *Nat Methods* 2013;**10**:744–746.

34 Distler U, Kuharev J, Navarro P, Levin Y, Schild H, Tenzer S. Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. *Nat Methods* 2014;**11**:167–170.

35 Röst HL, Rosenberger G, Navarro P, Gillet L, Miladinović SM, Schubert OT, Wolski W, Collins BC, Malmström J, Malmström L, Aebersold R. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat Biotechnol* 2014;**32**:219–223.

36 Parker BL, Yang G, Humphrey SJ, Chaudhuri R, Ma X, Peterman S, James DE. Targeted phosphoproteomics of insulin signaling using data-independent acquisition mass spectrometry. *Sci Signal* 2015;**8**:rs6–rs6.

37 Sidoli S, Lin S, Xiong L, Bhanu NV, Karch KR, Johansen E, Hunter C, Mollah S, Garcia BA. Sequential window acquisition of all theoretical mass spectra (SWATH) analysis for characterization and quantification of histone post-translational modifications. *Mol Cell Proteomics* 2015;**14**:2420–2428.

38 Picotti P, Aebersold R. Selected reaction monitoring-based proteomics: workflows, potential pitfalls and future direction. *Nat Methods* 2012;**9**(6):555–566.

39 Annan RS, Carr SA. The essential role of mass spectrometry in characterizing protein structure: mapping post-translational modifications. *J Protein Chem* 1997;**16**:391–402.

40 Casado-Vela J, Ruiz EJ, Nebreda AR, Casal JI. A combination of neutral loss and targeted product ion scanning with two enzymatic digestions facilitates the comprehensive mapping of phosphorylation sites. *Proteomics* 2007;**7**:2522–2529.

41 Mead JA, Bianco L, Ottone V, Barton C, Kay RG, Lilley KS, Bond NJ, Bessant C. MRMaid, the web-based tool for designing multiple reaction monitoring (MRM) transitions. *Mol Cell Proteomics* 2009;**8**:696–705.

42 MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* 2010;**26**:966–968.

43 Unwin RD, Griffiths JR, Leverentz MK, Grallert A, Hagan IM, Whetton AD. Multiple reaction monitoring to identify sites of protein phosphorylation with high sensitivity. *Mol Cell Proteomics* 2005;**4**:1134–1144.

44 Evans CA, Griffiths JR, Unwin RD, Whetton AD, Corfe BM. Application of the MIDAS approach for analysis of lysine acetylation sites. In: Hake SB, Janzen CJ, editors. *Protein Acetylation*. Totowa, NJ: Humana Press; 2013. p 25–36.

45 Unwin RD, Griffiths JR, Whetton AD. A sensitive mass spectrometric method for hypothesis-driven detection of peptide post-translational modifications: multiple reaction monitoring-initiated detection and sequencing (MIDAS). *Nat Protoc* 2009;**4**:870–877.

46 Mollah S, Wertz IE, Phung Q, Arnott D, Dixit VM, Lill JR. Targeted mass spectrometric strategy for global mapping of ubiquitination on proteins. *Rapid Commun Mass Spectrom RCM* 2007;**21**:3357–3364.

47 Choudhary C, Mann M. Decoding signalling networks by mass spectrometry-based proteomics. *Nat Rev Mol Cell Biol* 2010;**11**:427–439.

48 Patterson SD. Data analysis–the Achilles heel of proteomics. *Nat Biotechnol* 2003;**21**:221–222.

49 Carr S, Aebersold R, Baldwin M, Burlingame A, Clauser K, Nesvizhskii A. The need for guidelines in publication of peptide and protein identification data

working group on publication guidelines for peptide and protein identification data. *Mol Cell Proteomics* 2004;**3**:531–533.

50 Cox J, Mann M. Quantitative, high-resolution proteomics for data-driven systems biology. *Annu Rev Biochem* 2011;**80**:273–299.

51 Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods* 2007;**4**:787–797.

52 Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 2008;**26**:1367–1372.

53 Houel S, Abernathy R, Renganathan K, Meyer-Arendt K, Ahn NG, Old WM. Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *J Proteome Res* 2010;**9**:4152–4160.

54 Clauser KR, Baker P, Burlingame AL. Protein prospector role of accurate mass measurement (±10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal Chem* 1999;**71**:2871–2882.

55 Bern M, Kil YJ, Becker C. Byonic: advanced peptide and protein identification software. *Curr Protoc Bioinforma* 2012;Chapter 13:Unit13.20.

56 Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, Mann M. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* 2006;**127**:635–648.

57 Beausoleil SA, Villén 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.

58 Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 2007;**4**:207–214.

59 Hornbeck PV, Chabra I, Kornhauser JM, Skrzypek E, Zhang B. PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics* 2004;**4**:1551–1561.

60 Gnad F, Ren S, Cox J, Olsen JV, Macek B, Oroshi M, Mann M. PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol* 2007;**8**:R250.

2

Identification and Analysis of Protein Phosphorylation by Mass Spectrometry

Dean E. McNulty, Timothy W. Sikorski and Roland S. Annan

Proteomics and Biological Mass Spectrometry Laboratory, GlaxoSmithKline, Collegeville, PA, USA

2.1 Introduction to Protein Phosphorylation

Much of the activity in the cellular proteome is under the control of reversible protein phosphorylation. Phosphorylation-dependent signaling regulates differentiation of cells, triggers progression of the cell cycle, and controls metabolism, transcription, apoptosis, and cytoskeletal rearrangements. Signaling via reversible protein phosphorylation also plays a critical role in intracellular communication and immune response. Phosphorylation can function as a positive or negative switch, activating or inactivating enzymes. It can serve as a docking site to recruit other proteins into multiprotein complexes or serve as a recognition element to recruit other enzymes that add other post-translational modifications (PTMs) or additional phosphorylation sites. Phosphorylation can trigger a change in the three-dimensional structure of a protein or initiate translocation of the protein to another compartment of the cell. Disruption of normal cellular phosphorylation events is responsible for a large number of human diseases [1–3]. From the discovery of the first functionally relevant phosphorylation site in 1955 [4], the ability to analyze protein phosphorylation has exploded in the last five years to the point where it is now possible to quantitate changes in tens of thousands of phosphorylation sites in response to a cell receiving an external stimulus or undergoing a normal change in the physiology [5]. While phosphorylation is known to occur on histidine, aspartate, cysteine, lysine, and arginine residues, this chapter focuses on the more commonly modified and well-studied amino acids: serine, threonine, and tyrosine.

The first evidence for protein phosphorylation was uncovered in 1906 when Phoebus Levene identified phosphate in the amino acid composition of the egg yolk protein vitellin [6]. While there was evidence in the 1920s to suggest the

Analysis of Protein Post-Translational Modifications by Mass Spectrometry,
First Edition. Edited by John R. Griffiths and Richard D. Unwin.

phosphate was on the amino acid serine [7], it was not until 1932 that Levene and Fritz Lipmann isolated phosphoserine from vitellin [8]. Prior to the 1950s, research on phosphoproteins was focused mainly on abundant proteins found in egg yolk (such as vitellin) and milk (casein), and the biological function, if any, of the phosphorylation was unknown. But by the early 1950s, it was being shown that in tumor cells the phosphorus in phosphoproteins was being turned over rapidly and that tumors contained high levels of phosphoserine [9, 10], together suggesting that this modification must have some function. In 1954 Kennedy and Burnett, using labeled ATP, demonstrated that an enzyme from rat liver mitochondria was responsible for catalyzing the phosphorylation of serine on both alpha and beta casein [11]. A year later Fischer and Krebs provided the first evidence that protein phosphorylation had a biological function. They demonstrated that inactive phosphorylase b could be converted to active phosphorylase a in the presence of ATP and Mg [4], and in the next few years they identified phosphorylase kinase as the enzyme responsible for the activation and showed that it phosphorylated a specific serine residue on phosphorylase b [12].

It is now widely recognized that cascades of protein phosphorylation transmit signals from the extracellular environment to trigger a biological response within the cell. The first evidence that kinases worked in series came in 1968 with the discovery of cAMP-dependent protein kinase A (PKA) and the fact that it phosphorylated and activated phosphorylase kinase [13]. It quickly became clear that PKA had many substrates in multiple tissues [14], and the idea that protein phosphorylation was a widespread phenomenon began to take hold. Throughout the 1970s and 1980s many additional serine/threonine (S/T) protein kinases were discovered, and in 1983 Tony Hunter showed that the v-Src protein was a tyrosine kinase (TK) [15]. The difficulty in detecting phosphotyrosine in these early years arises from the fact that we now know it constitutes only a few percent of the total phosphoamino acid pool [5, 16] and that it comigrated with the much more abundant phosphothreonine in the standard electrophoretic systems used in the late 1970s to detect ^{32}P-labeled phosphoamino acids [17].

With the development by Hunter and Sefton of a two-dimensional (2D) separation method for phosphoamino acids [15], it quickly became clear that phosphorylation on tyrosine was also widespread. In 1981 the EGF receptor (EGFR) was shown to have TK activity and that stimulation of cells with EFG led to rapid tyrosine phosphorylation on multiple proteins [18, 19]. By the end of the 1980s more than 10 receptor tyrosine kinases (RTKs) had been identified. The realization that growth factor receptors had intrinsic TK activity connected intracellular signaling through (largely) serine/threonine (S/T) kinases with external signals communicated via ligand binding to transmembrane receptors. In many cases, nonreceptor tyrosine kinases (NRTK) constitute the next step in the signaling cascade, transmitting signals from the intracellular

domains of the RTK to downstream S/T protein kinases [20, 21]. Vast amounts of research in the 1980s and 1990s encompassing all areas of cellular biology would discover many more kinases and their substrates and add much fine detail to the mechanism of phosphorylation-dependent signaling.

The identification of all human kinase genes was made possible with the complete sequencing of the human genome [22]. Bioinformatic analysis has identified 478 protein kinases (see Figure 2.1, right), belonging to a large superfamily that shares a eukaryotic protein kinase (ePK) domain. There are an additional 40 atypical protein kinases (aPK), which have been demonstrated to have protein kinase activity, but do not share the ePK domain. Altogether the 518 protein kinases make up one of the largest families of eukaryotic genes (see Figure 2.1). All major kinase groups and most kinase families are shared across metazoans, and many are shared in yeast [23]. Protein tyrosine kinases (PTK) of which 90 have been identified are found only in metazoans [24]. More than half of these (58) are RTKs, involved in regulating the multicellular aspects of an organism via cell-to-cell communication. It is surprising how little is actually known about most of these 518 protein kinases (termed the "kinome").

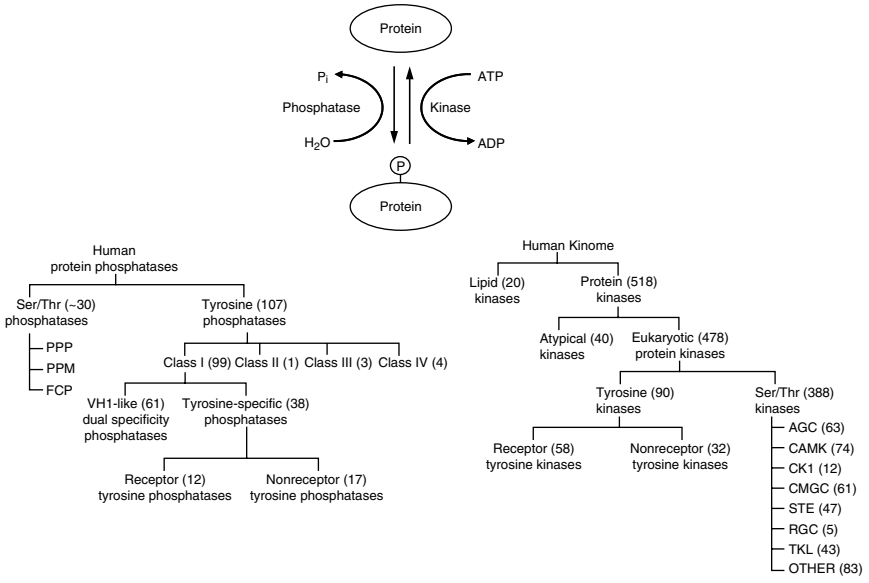


Figure 2.1 Protein phosphorylation is governed by two large superfamilies of enzymes. Protein kinases (right) add phosphate to (primarily) serine, threonine, and tyrosine residues. Protein phosphatases (left) remove the phosphate group. There are similar numbers of tyrosine kinases and phosphatases. The very small number of serine/threonine phosphatases achieves selectivity by forming combinatorial enzyme complexes with a large number of regulatory subunits.

More than 100 of the kinases have absolutely no known function, and 50% are largely uncharacterized [25]. A very small percentage of the kinome accounts for most of the published literature. This lack of knowledge about most of the human kinome is reflected in the fact that, of the twenty approved kinase therapeutics, they address only nine different kinases as their primary targets [26]. This is in spite of the fact that kinases are characterized as excellent drug targets in cancer and many other diseases. Kinase gene profiling shows distinct expression pattern differences between healthy and disease tissues for large clusters of the kinome [27].

Given the wide range of processes that are under the control of reversible protein phosphorylation and the large number of protein kinases in the metazoan genomes, it is not surprising that the extent of phosphorylation in higher-order organisms is massive. Current phosphosite databases [28, 29] list more than 150,000 sites on over 18,000 human proteins, many more than were previously predicted. The large majority of these sites have been identified in high-throughput phosphoproteomics studies utilizing MS. Large-scale phosphoproteome studies suggest that the overall phosphoamino acid composition of any cell is approximately 75–85% phosphoserine, 10–20% phosphothreonine, and 1–6% phosphotyrosine [5, 30–33]. This composition likely reflects the biology of the cell and not some bias of the mass spectrometer, as it has been shown using a large-scale synthetic phosphopeptide library that peptides containing all three types of phosphoamino acids are detected equally [32].

In 15–25% of phosphoproteins only a single site has been identified. The functional significance of these single sites is to act, in many cases, as a simple switch. Glycogen phosphorylase, for instance, contains only a single phosphoserine that drives it from the inactive to the active state [34]. The majority of proteins, however, are phosphorylated on more than one site and by more than one kinase. The spliceosome protein Srrm2 was found to contain anywhere between 177 and 300 sites [30, 33]. As might be expected, a weak but significant correlation exists between a protein's abundance and the number of sites identified in an analysis [5]. However, it is clear that multisite phosphorylation is the rule rather than the exception. It has been suggested that the multiplicity of phosphorylation on proteins might just be background noise. However, it is equally likely that given the wide variety of biological functions under the control of protein phosphorylation and the wide variety of mechanisms by which it occurs, the functional significance of most of the complex hyperphosphorylation that occurs on proteins is not yet understood. What is emerging, however, is just how intricately this multisite phosphorylation is coordinated. While some phosphorylation clusters share a common biological function, in many cases each site or a combination of sites has distinct and separable roles in that function.

The budding yeast transcription factor Pho4 controls the expression of genes needed by the organism to survive under conditions of phosphate starvation.

In a normal phosphate-rich environment, PHO4 is phosphorylated on 5 cyclin/Cdk sites and exported out of the nucleus. When yeasts are deprived of phosphate, these sites are unoccupied, and Pho4 accumulates in the nucleus and activates expression of phosphate-responsive genes. Four of the five cyclin/Cdk sites have distinct roles to play in the regulation of this function, with two being required for nuclear export, one for blocking nuclear import, and one for blocking promoter binding [35, 36]. To add complexity to this mechanism, under intermediate conditions of phosphate availability, PHO4 is phosphorylated on only one of the sites, allowing it to bind differentially to its target promoters and trigger expression of only a subset of the phosphate-responsive genes [37].

In contrast to PHO4, whose function is regulated by multisite phosphorylation via a single kinase, Sic1 is regulated by a multisite phosphorylation cascade that involves a complex dance of two different kinases. Sic1 controls the G1/S phase transition in budding yeast by inhibiting the S-phase Clb5–Cdk1 kinase. Ubiquitin-mediated destruction of Sic1 releases Clb5–Cdk1 and allows the cell to proceed to S phase (Figure 2.2a). In one of the first examples of how phosphorylation regulates ubiquitin-mediated proteolysis, Sic1 was shown to be phosphorylated on at least nine different sites and required a combination of at least three of six to trigger degradation [38]. In fact it was later shown that some phosphorylation on at least six of the nine Cdk sites is required for destruction [39]. Five of the nine Cdk-dependent sites form three pairs of high-affinity recognition elements termed phosphodegrons (see Figure 2.2b), which are recognized by ubiquitin ligases [40]. These nine sites are phosphorylated by two different cyclin/Cdks, with each showing preference for different sites. At the transition to S phase, Cln2–Cdk1 phosphorylates Sic1 on a subset of the nine sites, but with no fully formed degrons (Figure 2.2b, top). This cluster of phosphorylation sites, however, is an excellent docking platform for the slowly released Clb5–Cdk1 (Figure 2.2b, bottom), which goes on to complete phosphorylation of the residues critical for the formation of the degrons [41]. The ordered phosphorylation by two different kinases imposes a tight regulation on the G1/S transition in which Cln2–Cdk1 is not allowed to trigger the change until sufficient levels of Clb5–Cdk1 accumulate.

For both Pho4 and Sic1, phosphorylation drives the protein's biological function by regulating protein–protein interactions. In the case of Pho4, it blocks the interaction of Pho4 with nuclear import and export transport proteins and the transcriptional coactivator protein that allows promoter binding. In the case of Sic1, phosphorylation of the priming sites facilitates binding of cyclin/Cdk complexes through their regulatory subunit Cks1. Phosphorylated sites within the three degrons of Sic1 then serve as docking sites for the SCF ubiquitin ligase. Indeed while the earliest examples of the biological significance of protein phosphorylation were in the conformation-induced stimulation of enzymatic activity, it has since become clear that much of protein

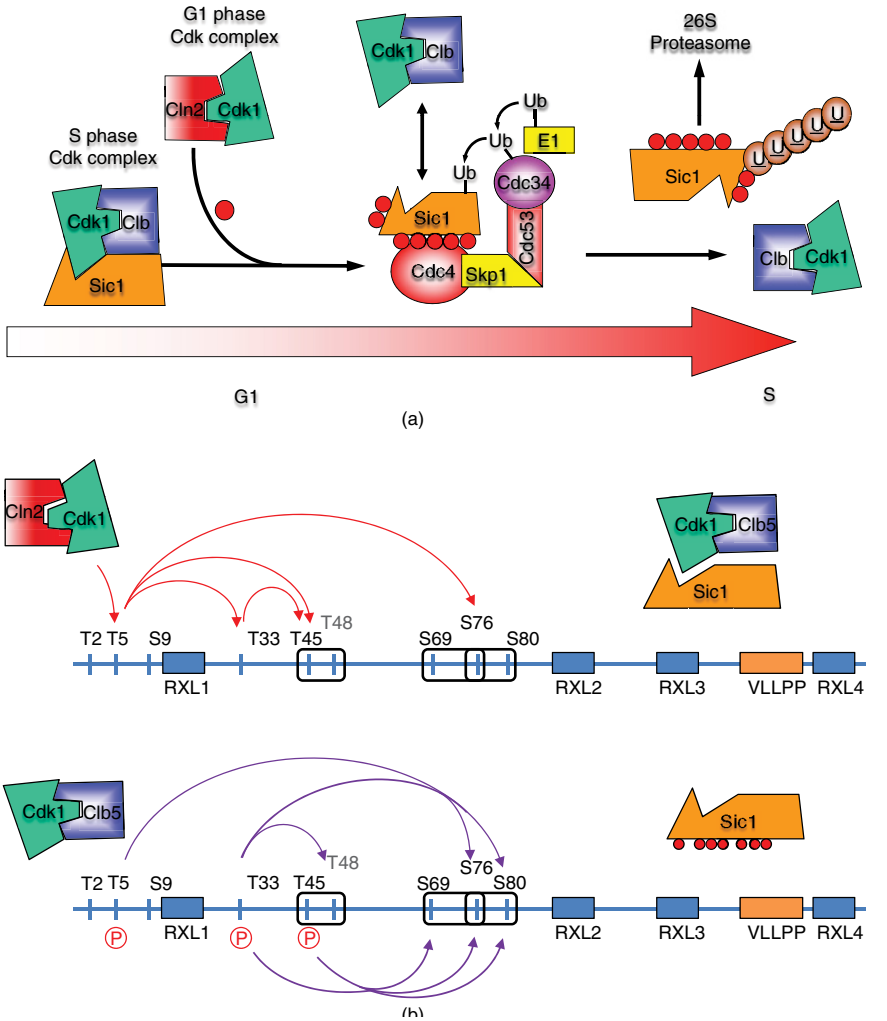


Figure 2.2 Cascades of multisite phosphorylation regulate biological function. (a) Sic1 controls the G1/S phase transition in budding yeast by inhibiting the S-phase Clb5–Cdk1 kinase. Phosphorylation-dependent ubiquitin-mediated destruction of Sic1 releases Clb5–Cdk1 and allows the cell to proceed to S phase. (b) In the first wave of phosphorylation (top), a subset of required sites are sequentially modified, but no fully formed binding sites (▭) for the ubiquitination machinery are formed. These initial sites act as priming sites for the second wave of phosphorylation (bottom), which is being carried out by the slowly released Clb5–Cdk1. The now fully formed phosphodegrons bind the ubiquitination machinery, initiating destruction of Sic1. Without further sequestration of Clb5–Cdk, the cells can transition into S phase.

phosphorylation serves to either recruit or block the recruitment of other proteins. The first example of this came with the discovery of SH2 domains. The search for TK substrates in the early 1980s revealed that growth factor receptor TKs preferred themselves as substrates. This raised the question "How do RTK transmit signals to drive cellular behavior?" In 1986 Tony Pawson identified a region in the oncogenic NRTK v-Fes that was conserved in all cytoplasmic tyrosine kinases and influenced their kinase activity [42]. Termed Src homology domain 2 (SH2), it was later shown that SH2 domain-containing proteins bind other proteins, including growth factor receptors, that are phosphorylated on tyrosine [43, 44]. The recruitment of SH2 domain-containing proteins to phosphotyrosine-containing residues on growth factor receptors thus provides a mechanism by which RTKs can cascade signals into the cytoplasm. There are 120 SH2 domains on 115 proteins in the human genome. They occur on proteins that link tyrosine phosphorylation to intracellular signaling, including all NRTKs, some tyrosine phosphatases, some lipid kinases, and many adaptor proteins [45]. While the SH2 domain remains the prototype for phosphorylation-mediated protein–protein interactions, other phosphosite-dependent binding domains have since been discovered, including the PTB domain that also binds phosphotyrosine [46]. More than ten phosphoserine and phosphothreonine binding domains have also been discovered [47] including WD40 domains, which are part of the F-box proteins that act as the substrate recognition element of SCF E3 ubiquitin ligases including the one that mediated the destruction of Sic1 as described earlier.

Along with the reality that multisite phosphorylation is the norm for eukaryotic proteins, it has also now become clear that most of this phosphorylation occurs in intrinsically disordered regions of proteins [48]. Nearly all eukaryotic proteins contain disordered regions, and some proteins are predicted to be entirely disordered [49]. Intrinsically disordered proteins (IDP) play a central role in mediating protein–protein interactions and the assembly of complex protein interaction networks [50]. The disordered regions contain multiple conserved sequence motifs that serve as docking sites for other proteins, including protein kinases. The flexibility of the disordered regions makes them accessible to PTM, including but not limited to phosphorylation. With the addition of these PTMs, it is estimated that perhaps a million sequence-specific interaction motifs exist with the disordered regions of the proteome [51]. In addition to Sic1, two other well-studied examples of phosphorylation (and other PTM) clusters in disordered regions that control function are p53 [52] and RNA polymerase II [53]. The latter protein contains 52 YSPTSPS repeats in the disordered C-terminal tail that are phosphorylated on the second and fifth serines in the motif, recruiting splicing factors, chromatin modifiers, termination machinery, and other protein modules to the elongation machinery. Interestingly, the phosphorylation of intrinsically disordered regions often brings about a disordered to ordered transition in the