Unterschrift des Betreuers

TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

DIPLOMARBEIT

Estimating Actual False Localization Rates for Large-Scale Proteomics Datasets

Ausgeführt am Institut für

Verfahrenstechnik, Umwelttechnik und Technische Biowissenschaften

der Technischen Universität Wien

in Kooperation mit dem

Research Institute of Molecular Pathology (IMP), Wien

unter der Anleitung von

Ao.Univ.Prof. Mag. Dr.rer.nat. Robert Mach

und

Karl Mechtler

durch

Thomas Taus, BSc

Albrechtsgasse 28a, 2500 Baden, Österreich

Datum

Unterschrift (Student)

Statutory Declaration

I hereby declare and confirm that this thesis is entirely the result of my own original work. Where other sources of information or resources have been used, they have been indicated as such and properly acknowledged. I further declare that this or similar work has not been submitted for credit elsewhere.

Vienna, July 16^{th} 2014

Thomas Taus, BSc

Contents

Acknowledgment

I am deeply grateful to my two supervisors Prof. Robert Mach and Karl Mechtler for their continuous support, enthusiasm, confidence and for involving me into such a fascinating project. Moreover, Karl Mechtler enabled me to present my scientific work at various conferences across Europe and the United States of America, which provided me with a unique insight into state-of-the-art research.

Furthermore, I am indebted to all my working colleagues at the Protein Chemistry Facility of the Research Institute of Molecular Pathology (IMP) for their enthusiastic support and interesting discussions. It is an honor for me that I was part of this great team, which provided me with such a pleasant and motivating environment. Particularly I wish to thank Etienne Beltzung and Gerhard Dürnberger, my working colleagues and friends, for assisting me both privately and professionally.

I am also thankful to Boehringer Ingelheim for financially supporting my research at the IMP.

Finally, I would like to express my deepest gratitude to my family, Silvia and Walter Taus my parents, and Gertrude and Wilhelm Taus my grandparents, who have supported me tirelessly and with their full strength, encouraged me and trusted in me during the entirety of my life. Neither my academic education nor my achievements would have been possible without their great help. I would like to greatly thank Katherine Mavro my partner for her inspiring commitment in every situation.

Kurzfassung

In den letzten Jahren hat sich Massenspektrometrie zur Technologie der Wahl für Proteincharakterisierung, sowie für die Analyse von post-translationalen Modifikationen (PTMs), wie etwa Phosphorylierung, entwickelt. Neben der Identifikation und Quantifizierung von modifizierten Peptiden und Proteinen stellt die präzise Lokalisierung der Modifizierungsstellen innerhalb der Aminosäuresequenz eine kritische Aufgabe bei der Beantwortung biologischer Fragestellungen dar. Es wurde eine Vielzahl an Software Lösungen entwickelt, welche in der Lage sind, die wahrscheinlichste PTM Zuordnung zu bestimmen und die Konfidenz der berichteten Resultate abzuschätzen. Dies ermöglichte Hochdurchsatzlokalisierung von Modifikationsstellen. Allerdings kann die sogenannte False Localization Rate (FLR) der berechneten Modifikationsstellen bis heute nicht akkurat mittels eines allgemein anerkannten Ansatzes abgeschätzt werden, weshalb das Ziel dieser Arbeit die Entwicklung einer ebensolchen Methode war. Die neuartige Methode basiert auf das Hinzufügen von Modifikationsstellen, welche *a priori* als inkorrekt bekannt sind, zum Suchraum. Diese so genannten Köderstellen werden generiert, indem entscheidende Fragmentionen entlang der m/z Achse verschoben werden. Um den Ansatz validieren zu können, wurde eine Peptidbibliothek mit rund 60.000 individuellen Phospho-Peptiden erstellt. Basierend auf mehr als 700.000 hoch zuverlässige Peptide Spectrum Matches, welche mit verschiedensten Dissoziationsmethoden aufgezeichnet und unter unterschiedlichsten Instrumenteneinstellungen gemessen wurden, deuten die erhaltenen Ergebnisse an, dass der neue Ansatz im Stande ist, die FLR unter allen untersuchten Bedingungen zuverlässig abzuschätzen.

Abstract

In recent years, mass spectrometry has emerged as the technology of choice for protein characterization, including the analysis of post-translational modifications (PTMs), such as phosphorylation. Besides identification and quantification of modified proteins and peptides, the precise localization of the exact site within the amino acid sequence is critical for the biological questions addressed. A variety of software tools have been introduced that determine the most probable PTM assignment and provide scores estimating the confidence of reported results, thereby enabling high-throughput site localization. However, the false localization rate (FLR) of obtained sites cannot yet be estimated accurately by a generally accepted approach and, thus, it was the aim to develop such a statistical method.The novel approach involves addition of proper candidates that are *a priori* known to be incorrect (decoys) to the search space. Generation of decoy sites is achieved by m/z shifting of site determining fragment ions. In order to validate the method, a peptide library was constructed that comprises roughly 60,000 distinct phospho-peptides. Based on more than 700,000 high confident peptide to spectrum matches that were acquired with diverse dissociation methods and measured under various instrumental conditions, it is assumed that the novel approach is capable of reliable FLR estimation in all scenarios under investigation.

Chapter 1

Introduction

In-depth investigation of living things and biological systems has always been fascinating and inspiring people, thereby evolving into one of the major research areas. Dramatic technological advances during the past century have lead to fundamental discoveries allowing a better understanding of the nature of life. It is now known that an essential feature of every living organism is its complex and dynamic biochemical machinery, enabling both reaction to external stimuli and adaptation to environmental changes. Post-translational modifications (PTMs) play a key role in these molecular dynamics. Protein phosphorylation for example, the most ubiquitous PTM, was discovered to be crucial for a variety of cellular processes, including signal transduction, transcription, proliferation and metabolism [1, 2]. Hence, the detailed study of phosphorylated proteins and their interaction with other proteins, as well as with DNA, is of great interest since the past decades and gave rise to a specialized field termed phospho-proteomics.

The availability of high-throughput and large-scale DNA sequencing methods, which ultimately enabled complete genome analysis [3–8], had a major impact on protein research. Utilizing translated DNA sequences, quick identification of proteins became a feasible task. Mass spectrometry (MS) has matured from a method able to identify a limited number of proteins in rather simple compositions to a high-throughput technology capable of assessing the complexity of whole proteomes, including that of human [9]. Typically, large-scale MS-based proteomics deploys high performance liquid chromatography (HPLC) coupled online via electrospray ionization (ESI) [10] to tandem mass spectrometry (MS/MS). The downstream computational data analysis involves application of sophisticated algorithms aiming at identification, quantification and characterization of peptides and proteins.

Nowadays, MS is the state-of-the-art technology for protein characterization including the analysis of PTMs, such as phosphorylation [11, 12]. Besides identification and quantification of modified peptides, the precise determination of the exact site within the protein carrying the PTM constitutes a critical task for the biological question addressed. Due to the dimension of current phospho-proteomics studies [13–15], which aim at assigning tens of thousands of phospho-sites, software tools have been developed that can tackle the challenging task of high-throughput site localization [13, 16, 17]. However, there exists no generally accepted approach to accurately estimate the error rate of obtain modification assignments, which would be crucial to assure the confidence of reported results and inferred scientific conclusions.

In this work a novel approach for accurate false localization rate (FLR) estimation, applicable to phospho-proteomics datasets and potentially extendable to the analysis of other PTMs, is presented. The method has been validated based on a library comprising roughly 60,000 individual chemically synthesized phospho-peptides with known modification sites. In total, more than 700,000 MS/MS spectra were acquired with distinct sample compositions, diverse fragmentation techniques and altered instrumental setting, aiming at a critical examination of the newly developed FLR estimation procedure.

Chapter 2

Protein Phosphorylation in Biological Systems

Reversible protein phosphorylation, first being considered a rather specialized regulatory mechanism largely confined to glycogen metabolism, is now recognized to be crucial for regulation of nearly every aspect of cellular life [18]. Aberrant phosphorylation can be a cause or consequence of diseases such as cancer [19, 20]. The consequential surge of interest in this research area over recent years culminated in the approval of the first orally active protein-kinase inhibitor as a drug for clinical use [18, 19].

2.1 Targeted Amino Acid Residues

The process of protein phosphorylation involves transfer of the γ-phosphate moiety of adenosine triphosphate (ATP) classically to the hydroxyl group of either serine (Ser), threonine (Thr) or tyrosine (Tyr) residues, forming a phosphoester bond. Enzymes referred to as protein kinases and protein phosphatases catalyze phosphorylation and de-phosphorylation, respectively. In large scale phospho-proteomics studies phospho-Ser is detected predominantly (70-90%), compared to phospho-Thr (10-25%) and the even less frequently observed phospho-Tyr (1-10%). Interestingly, the distribution of these so called O-phosphorylations seems to be highly conserved between different species ranging from bacteria and archaea to highly complex eukaryotes, such as human [13, 21–26].

Besides Ser, Thr and Tyr there exist other amino acids that are ca-

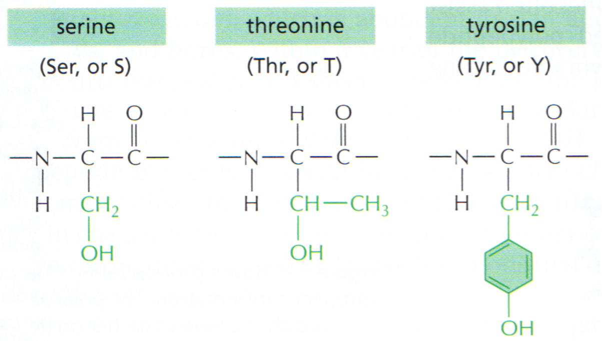


Figure 2.1: Chemical structure of serine, threonine and tyrosine [1].

pable of being phosphorylated and are thought to engage in important biological functions [27–30]. Examples of such non-canonical phosphorylation accepting residues are arginine (Arg), lysine (Lys) and histidine (His), which may covalently bind the phosphate moiety at the side-chain nitrogen forming a phosphoramidate (P-N) bond [31]. However, those so called N-phosphorylations are up to now only scarcely studied, which can be addressed mainly to the acid lability of the P-N bond in comparison to the acid stable O-phosphorylations [27, 29, 32]. Resistance under acidic conditions is critical because commonly employed sample preparation and measurement strategies involve rather low pH.

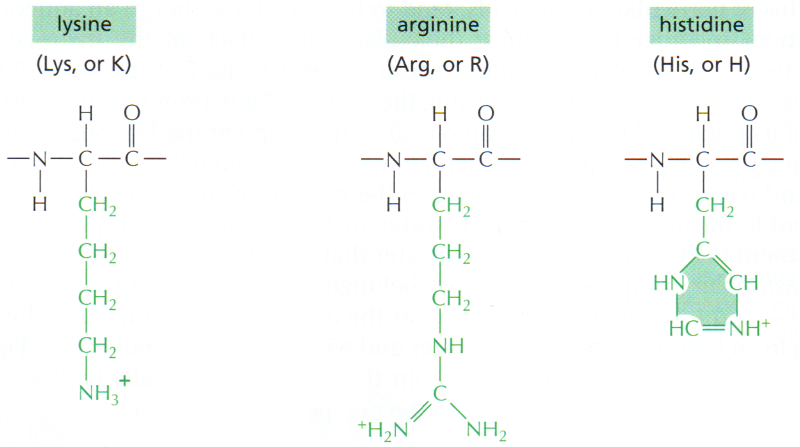


Figure 2.2: Chemical structure of lysine, arginine and histidine [1].

2.2 Consequences

Proteins can be influenced by phosphorylation events in two distinct ways. Firstly, the addition of one or more negatively charged phosphate groups causes a dramatic conformational change, which in turn alters ligand binding capabilities elsewhere on the protein surface, thereby modifying the proteins activity [1]. Secondly, the attached phosphate group itself functions as a part of the binding site [1]. For both of the mentioned possibilities, the phosphorylation state of a certain protein, and thus its activity, is influenced by the combined activities of the protein kinases and phosphatases it is targeted by. As a consequence, phosphate groups are continually turning over, allowing a population of proteins to alter its state of phosphorylation quickly in response to an abrupt change in the rate of attachment of that moiety. Furthermore, protein phosphorylation and dephosphorylation can promote the regulated assembly and disassembly of protein complexes.

2.3 Functional Characteristics

The attachment of phosphate moieties to certain side chains of a protein's amino acids can give rise to cellular dynamic in various ways. It can serve as an on and off switch for the catalytic activity of the targeted enzyme, affect protein-protein or protein-nucleic acid interactions, alter the protein's stability or modify its specific localization within the cell [1]. The combination of these effects enables the conversion of information from beyond the cell membrane into an intracellular chemical change, a process referred to as signal transduction, which ultimately results in a cellular response. Intracellular signaling proteins relay the signal sequentially from one to another until the effector protein that finally alters the cell's behavior. On its way downstream from the primary transduction the signal may be transformed, amplified, integrated with and spread to other signaling pathways leading to highly interconnected regulatory circuits [1]. This emphasizes the requirement of multidisciplinary high-throughput analysis methods – both experimental and computational – in combination with sophisticated mathematical models to uncover the connection between a stimulus and the corresponding cellular response, which is essential for the in-depth study of diseases such as cancer.

Chapter 3

Mass Spectrometry-Based Proteomics

Proteomics in general aims at the analysis of the entire set of proteins expressed by a specific cell or tissue under predefined conditions [33]. In contrast to an organism's genome, which is considered rather constant over time, its proteome seems to be of sheer infinite complexity, since each protein may be present in different forms, in different amounts and at different times. Owing to the availability of whole-genome sequencing data and dramatic technological advancements, MS has over the past decade become the method of choice for large-scale analysis of complex protein samples.

3.1 Bottom-Up versus Top-Down

Each proteomic analysis starts with preparation of the sample, in which proteins are either enzymatically cleaved prior to MS measurement yielding peptides (bottom-up) or they are analyzed intact (top-down). Currently, the most popular method for large-scale investigation of complex samples is the so called bottom-up proteomics approach. Here endoproteases, such as trypsin, are used for proteolytic cleavage. Subsequently, proteins are identified based on peptide masses and sequences, which are inferred using tandem mass spectrometry (see section 3.4) and dedicated software tools (see section 3.5.1). Shotgun proteomics, a commonly applied subtype of bottom-up proteomics, can be seen as the protein equivalent to shotgun genomic sequencing, in which the whole genome is sheared and analyzed in a nontargeted fashion. However, due to the peptide-centric nature of bottom-up proteomics, inference of protein-based information, such as identification of protein isoforms or determination of protein's PTM-state, is impeded [34].

Top-down proteomics on the other side enables identification of protein isoforms [35, 36], allows better characterization of PTMs [37, 38] and provides improved reliability of protein quantification [39–41], compared to the peptide-based analog. However, the method still faces several technological limitations, such as the challenging front-end separation of intact proteins, the requirement for high resolution MS for resolving the proteins' complex isotopic envelopes and the efficient fragmentation of large proteins, which hinder it up to now from widespread application.

This work focuses almost exclusively on the bottom-up proteomics approach.

3.2 Phospho-Peptide Enrichment

Although it is thought that roughly one third of all proteins encoded by the human genome are phosphorylated, the level of phosphorylation at specific sites can vary between less than 1% and greater than 90% [42]. The consequential substoichiometric nature of protein phosphorylation poses a major challenge in phospho-proteomics and necessitates the application of advanced enrichment strategies.

Immobilized metal affinity chromatography (IMAC) is capable of selectively enriching for phospho-peptides based on high-affinity coordination of phosphate groups to certain trivalent metal cations, such as Fe^{3+}, Ga^{3+}, Al^{3+} and Zr^{3+}, under acidic conditions [43, 44]. Phospho-peptides can subsequently be eluted by either increasing the pH or addition of phosphate to the elution buffer.

Alternatively, TiO_2 was discovered to be a chromatographic solid-phase material enabling selective isolation and enrichment of phosphorylated peptides from complex mixtures [45–47]. It provides high chemical stability, physical rigidity and unique amphoteric ion-exchange properties [42]. Phosphate moieties can be trapped under acidic conditions and eluted at high pH.

Strong cation exchange (SCX) chromatography has also been shown to allow separation of phosphorylated and non-phosphorylated peptides, which is based on solution charge state differences caused by addition of a phosphate moiety [48]. Here analytes are eluted applying a gradient of increasing salt concentration. In several large-scale studies [13, 49, 50], SCX followed by either IMAC or TiO_2 has been used successfully for phospho-peptide enrichment of highly complex samples. Further, the combination of titanium enrichment and IMAC has been shown to be highly selective, sensitive and reproducible [51].

Electrostatic repulsion hydrophilic interaction chromatography (ERLIC) [52] is based on hydrophilic-interaction chromatography (HILIC) [53], which deploys a hydrophilic stationary phase and a hydrophobic mobile comprising mostly organic solvents, such as acetonitrile (ACN). It is assumed that HILIC involves partitioning of analytes between the hydrophobic mobile phase and a layer of water-enriched liquid phase, which is partially immobilized on the hydrophilic stationary phase [53]. When using weak anion exchange columns at low pH, negatively charged phospho-peptides can be retained, while positively charged acidic peptides, as well as N-terminally protonated peptides elude [52]. This method is termed ERLIC and can be used for selective isolation of phosphorylated peptides from a tryptic digest.

3.3 Chromatographic Separation

Due to the high complexity of proteomic samples such as whole cell lysates, which is further increased upon proteolytic cleavage of proteins into oligopeptides, and owing to the extraordinary dynamic range, which spans 11 orders of magnitude in case of the human plasma proteome [54], separation of analytes prior to their analysis by MS is indispensable.

Nanoscale reversed-phase high-performance liquid chromatography (nano RP-HPLC) online coupled to MS has become a routinely applied analysis setup for proteomics. This chromatographic technique employs a hydrophobic stationary phase consisting of silica beads with straight octadecyl (C18) side chains and a hydrophilic, aqueous mobile phase. Peptides are eluted based on their hydrophobicity using a gradient of increasingly concentrated organic solvents, such as ACN. Addition of carboxylic acids (e.g. formic acid) improves separation due to their ion pairing capabilities and enhances ionization of analytes required for online MS analysis.

Especially for in-depth analysis of complex large-scale proteomic samples, a further increase in peak capacity can be achieved by multidimensional separation, in which techniques are combined that separate analytes according to distinct orthogonal [55] molecular properties. The two dimensional separation approach deploying SCX followed by RP, also referred to as multidimensional protein identification technology (MudPIT), has become a popular method in shotgun proteomics [56–58].

3.4 Tandem Mass Spectrometric Measurement

MS is the most comprehensive and versatile technology for protein characterization in large-scale proteomics [59]. Measurements are carried out in the gas phase and include determination of both the mass to charge (m/z) ratio and abundance of ionized analytes. Despite front-end chromatographic separation, an average of more than 20 individual peptide species might elude simultaneously each second in a standard LC run of HeLa cell lysate [60], necessitating an advanced measurement procedure.

Peptide-mass fingerprinting, in which proteins are identified based on observed peptide masses, is restricted to the analysis of essentially purified target proteins [33]. Alternatively, selected peptide ions can be fragmented yielding additionally information about the peptide sequence. This method is also referred to as tandem mass spectrometry (MS/MS). Data acquisition in MS/MS is commonly performed in a data-dependent fashion, in which the most intense intact peptide ions of an initial scan (MS^1 scan) are selected for subsequent acquisition of the respective fragmentation spectra (MS^2 scans).

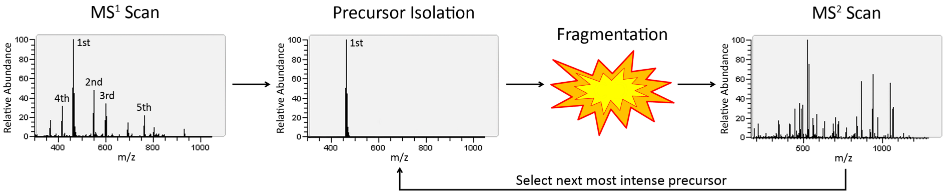


Figure 3.1: MS/MS applying data-dependent acquisition [61].

3.4.1 Fundamentals of Mass Spectrometry

In general, a mass spectrometer consists of the following three parts: an ion source, which is responsible for ionization of analytes, a mass analyzer that is capable of separating ions based on their m/z ratio and, finally, a detector, which determines the number of ionic species at each m/z value.

Ion source: Matrix-assisted laser desorption/ionization (MALDI) [62] and electrospray ionization (ESI) [10] constitute the two most commonly applied techniques to volatize and ionize proteins and peptides for subsequent MS analysis. In MALDI, analytes are sublimated and ionized out of a crystalline matrix by using laser pulses, giving rise to predominantly singly charged species. This is advantageous especially for top-down analysis of high-molecular-weight proteins [59]. Unlike MALDI, ESI is applied to solutions and, thus, can be readily coupled to liquid-based separation techniques, such as chromatography or electrophoresis. ESI is driven by a voltage of a few kV between a metal capillary, which is connected to e.g. a chromatographic system, and a counter electrode that constitutes the entrance of the mass spectrometer. In the created electrically charged spray, droplets are continually de-solvated at atmospheric pressure ultimately creating multiply charged peptides well suited for MS analysis.

Mass analyzer: Ions emanating from the ion source can be subject to detailed investigation by, up to now, 4 distinct basis types of mass analyzers, namely, sector field, time-of-flight (TOF), quadrupole mass filter (Q) and ion traps (IT). Sector field mass analyzers deflect accelerated ions onto m/z-dependent trajectories by a sector-shaped magnetic field. In contrast, TOF discriminates charged species according to their m/z by determining the time period accelerated ions require to travel along a defined distance in a field-less tube. Quadrupole mass filters deploy a set of four linear rods (alternatively 6 or 8 rods in case of hexapole or octupole, respectively) to create an electromagnetic field that allows only ions within a certain m/z range to traverse the device. Finally, ion traps are capable of capturing, measuring and – in most cases – also fragmenting charged species in one device [63]. There exist three subtypes of this kind of mass analyzer, namely, quadrupole ion traps (or Paul traps) [64–66], Penning ion traps [67] and Kingdon ion traps (or orbitraps) [68, 69]. All three are able to separate ions based on their m/z resonance frequencies. Among the most popular mass spectrometers used for proteomic analysis, many employ a combination of different mass analyzer types, such as triple quadrupole (QQQ) [70], quadrupole/time-of-flight (Q-TOF) [71] and hybrid linear ion trap/orbitrap (LTQ-orbitrap) [72].

Detector: After separation of ion according to m/z by the mass analyzer, charged species can be detected and quantified e. g. by electron multipliers, which allow the impact of ions on a surface to be converted into an electric signal. Such a detection principle is deployed in sector field, TOF, quadrupole mass filter and ion trapping instruments. Penning traps and orbitraps, however, require a completely different detection procedure. Oscillating ions in such devices induce an image current in the detector electrodes, which can be converted into a frequency spectrum by applying Fourier transform. The frequencies can in turn be translated into m/z ratios.

3.4.2 Fragmentation Techniques

State-of-the-art mass spectrometers provide a variety of different fragmentation techniques for peptide backbone dissociation, which give rise to distinct complementary fragment ion types. In general, N-terminal fragment ions are referred to as a-, b- or c-type ions and C-terminal ones are called x-, y- or z-type ions, depending on the respective cleavage site [73]. Observed fragment ion types and respective relative abundances vary strongly between the different activation techniques.

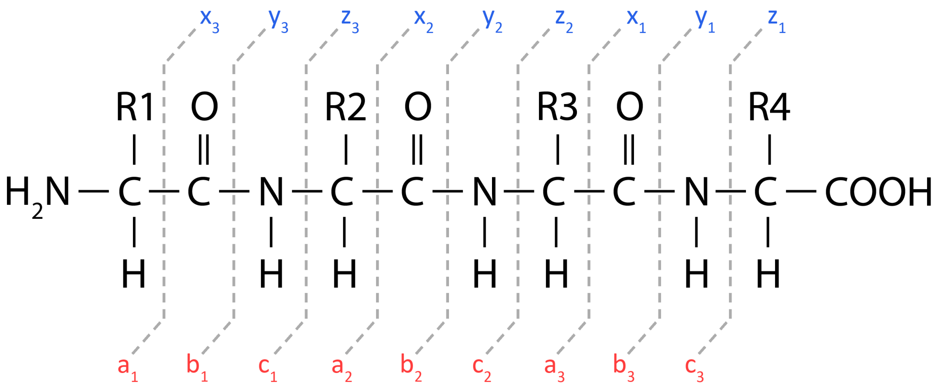


Figure 3.2: Nomenclature of peptide backbone fragments [61].

Collision-induced dissociation (CID): In CID, also referred to as collision-activated dissociation (CAD), precursor ions are kinetically activated and allowed to collide with neutral non-reactive gas molecules, such as He, Ar or N_2, which results in cleavage of the chemical bond that requires the lowest dissociation energy, namely, the amid bond that connects amino acid residues [74–76]. Heterolytic fission of the protonated amid bond results in formation of b- and y-type fragment ions [75]. Certain PTMs, such as phosphorylation, produce highly abundant neutral molecular losses resulting in poor peptide backbone fragmentation efficiencies [75]. In order to circumvent this issues, a method called multi-stage activation (MSA) has been developed, which is capable of activating both the intact precursor ions and subsequent neutral loss product ions, thereby creating more information-rich spectra [77].

Higher-energy collisional dissociation (HCD): Recently, a fragmentation technique termed HCD has been introduced [78]. It involves energetic injection of selected precursor ions into a dedicated collision cell [79], which results in a similar dissociation mechanism as observed for CID. This so called beam-type CID method produces predominantly y-type and low mass b-type fragment ions [80].

Electron transfer dissociation (ETD): Initiated by protonated peptides taking up low energy electrons that originate from aromatic anions with sufficiently low electron affinity, such as fluoranthene, radical-driven peptide backbone fragmentation is triggered [75]. Homolytic cleavage of N-C_α bonds results in the formation of c- and z-type fragment ions, while a secondary pathway leads to fragment ions of types a and y . This so called ETD, which is based on the previously developed electron capture dissociation (ECD) [81], is particularly well suited for characterization of peptides containing PTMs, such as phosphorylation, because peptide backbone cleavage is more or less independent of both peptide sequence and presence of labile modifications [75]. The insufficient fragmentation of double charged species can be enhanced by application of supplemental collisional activation of undissociated species after electron uptake [82].

Electron-transfer/higher-energy collisional dissociation (EThcD): Lately, a novel fragmentation technique has been introduced, in which ETD is followed by all ion HCD [83]. This so called EThcD method provides unique sequence coverage [83] and seems to be beneficial especially for the analysis of labile PTMs, such as protein phosphorylation [84].

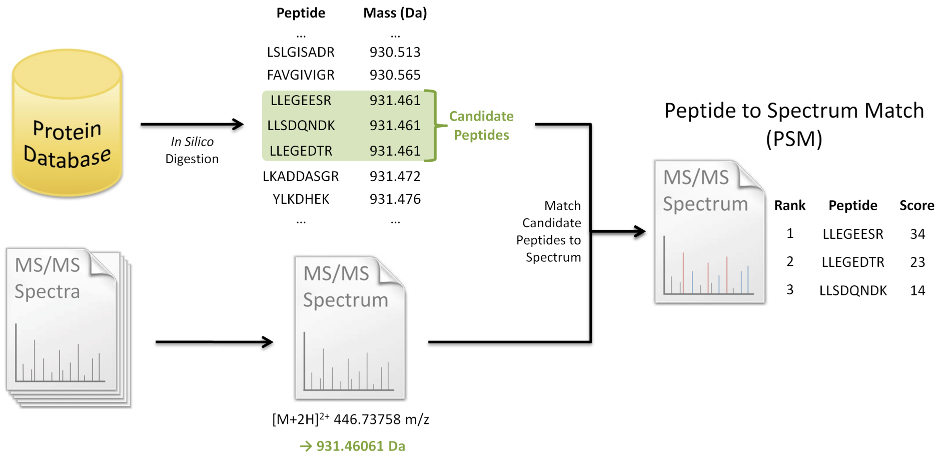
3.5 Computational Data Analysis

Computational analysis of the data obtained by LC-MS/MS measurements has become a crucial procedure in state-of-the-art proteomics experiments, which can mainly be attributed to the dimension of current studies that are trying to assess the sheer boundless complexity of proteomes, especially that of human. Although supervising analysis procedures is still required, most processes are performed in a fully automated fashion. Bioinformatics task performed today include identification, quantification and PTM analysis of peptides and proteins, for each of which specialized algorithms are required.

3.5.1 Peptide and Protein Identification

In general, proteins can be identified applying either one- or two-stage MS. In the first case, proteins are proteolytically cleaved and the masses of obtained peptides are determined by MS in order to infer the identity of the original protein. This procedure is also referred to as peptide mass fingerprinting (PMF) and is intrinsically restricted to the analysis of virtually pure target protein samples [33]. In the second case, not only the intact molecular mass of endoproteolytic peptides is recorded, but also fragmentation spectra are acquired, providing additionally sequence information for each selected peptide.

There exist two distinct approaches for assigning the correct peptide sequence to a given MS^2 spectrum. One aims at deducing the sequence of amino acids directly from the observed fragmentation pattern without any additional information, which is termed *de novo* peptide sequencing, and the other attempts to select the most probable peptide that corresponds to a given MS^2 scan by searching a database containing all possible proteins [85]. The latter requires that the peptide and parent protein sequence in question are known, meaning the genome of the organism under investigation needs to be sequenced. In case such information is not available and, as a consequence, protein database searching is precluded, then *de novo* peptide sequencing constitutes a method to identify unknown peptides and proteins, including novel PTMs [86]. However, *de novo* methods require high mass accuracy MS^2 data as well as sufficient spectral quality, in order to be able to deduce complete or partial sequences. Although there are well-established software solutions for *de novo* sequencing, such as PEAKs (Bioinformatics Solutions Inc.) [87], this approach is still not as broadly applied as database searching.



Peptide	Mass (Da)
...	...
LSLGISADR	930.513
FAVGIVIGR	930.565
LLEGEESR	931.461
LLSDQNDK	931.461
LLEGEDTR	931.461
LKADDASGR	931.472
YLKDHEK	931.476
...	...

Rank	Peptide	Score
1	LLEGEESR	34
2	LLEGEDTR	23
3	LLSDQNDK	14

Figure 3.3: Common workflow of database search engines ([61] modified).

Applying so called database search engines is the prevailing methodology for peptide identification [88]. Their common workflow involves *in silico* digestion of all proteins in the database, determination of candidate peptides for each MS^2 scan according to the individual precursor masses and predefined mass tolerance, matching between theoretical and measured fragment ions and soring of obtained peptide to spectrum matches (PSMs). Such search algorithms can be subdivided into two groups on the basis of the implemented scoring scheme: heuristic and probabilistic ones [89]. Heuristic search engines, such as Sequest [90] and X!Tandem [91], score PSMs by means of similarity (determined e.g. with cross correlation) between a theoretical fragmentation spectrum computed for a certain candidate peptide and the respective acquired MS^2 scan. Probabilistic search tools, in contrast, aim at estimating the likelihood that the observed PSM is a random event. Frequently used search engines of this category include Mascot [92], Phenyx [93], OMSSA [94], ProteinPilot [95], Andromeda [96] and MS Amanda [97].

Alternatively, each acquired tandem mass spectrum can be searched against a spectral library that is compiled from a large collection of previously identified confident PSMs [98, 99]. The potential of this method to complement sequence database searching has been reported several times [100–102]. However, spectral searching is capable of identifying only those peptides, which have already been assessed, unless fragment spectrum prediction algorithms become accurate enough and can replace acquisition of MS^2 spectra.

3.5.2 Phospho-Site Localization

Besides identification of phosphorylated peptides and proteins, the determination of the exact modification position(s) within the sequence constitutes a crucial tasks in phospho-proteomics and is most commonly performed by analysis of LC-MS/MS data. Phospho-site localization is based on detection of so called site-determining ions, which unambiguously pinpoint the modification to a certain sequence position [16]. Only if those fragment ions can be observed explicit assignment of phosphorylation is possible. While this task was previously performed predominantly by manual inspection of tandem mass spectra, which requires expert knowledge and is time-consuming, there has been a shift towards automated high-throughput site assignment using dedicated software tools, owing to the size of current phospho-proteomics

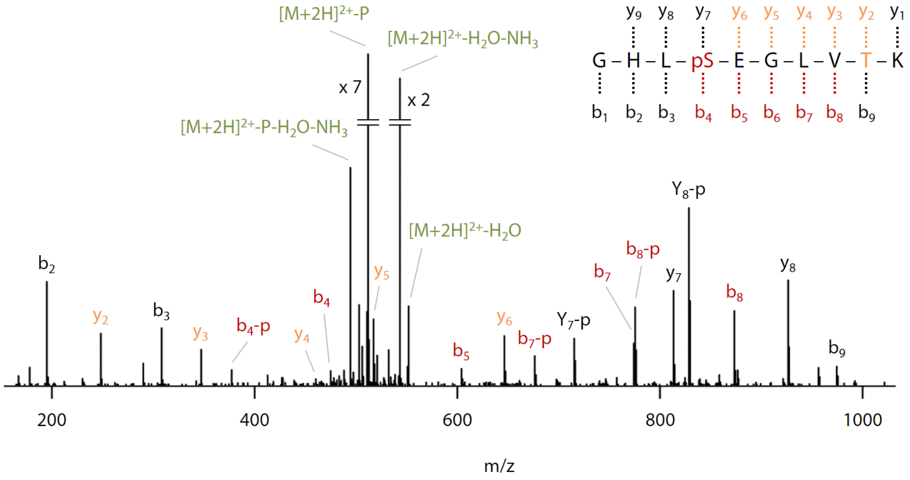


Figure 3.4: Phospho-site assignment based on site determining ions.

datasets [13–15]. These Computer algorithms, such as Ascore [16], PTM score [13], MD-score [103] and phosphoRS [17], assign the most probable sites and provide scores estimating the confidence of reported results.

Chapter 4

Control of Error Rates in High-Throughput Proteomics

Computational data processing is the predominant form of analyzing large-scale proteomics data today. It allows objective and unbiased analysis, additionally reducing working time, when compared to manual inspection of MS data. However, computation steps, such as identifying peptides and proteins, determining significantly regulated proteins, encountering protein-protein interaction partners and localization of PTMs within the protein sequence, are all challenging and occasionally error-prone exercises (even if performed manually instead), owing to a variety of reasons. First, the data can be noisy, containing spurious peaks, which might give rise to incorrect identifications. Second, due to high sample complexity, multiple co-eluting peptides might be selected at once, giving rise to chimeric MS^2 spectra that impede both identification [104] and accurate quantification [105], as well as precise PTM site assignment. Third, low quality tandem mass spectra might lack of sufficient fragment ions for unambiguous peptide identification or localization of chemical modifications. Fourth, the correct interpretation of a spectrum might not be among the candidates considered during the analysis, in case of a certain peptide or protein being absent in the database, an unanticipated PTM, or because a non-peptide species might have been selected for fragmentation.

As a consequence, methods have been developed that enable rigorous estimation of the number of so called false positive observations, regarding especially tasks such as peptide and protein identification and quantification, as well as PTM site assignment. The development of such approaches is of utmost importance in order to assure confidence of inferred scientific conclusions, since those rely on the validity of considered data.

4.1 Confident Peptide and Protein Identification

Owing to the fact that a database search engine generates a peptide match for almost all input tandem mass spectra, given that there is at least one candidate peptide that meets predefined specifications, it is indispensable to distinguish correct from incorrect PSMs [106]. For this purpose, it has previously been state-of-the-art to apply certain constant filter criteria [56], until it has been realized by the scientific community that statistical considerations are required in order to assure quality of reported results [107].

In the following, approaches for estimating the frequency of false identifications, also referred to as false discovery rate (FDR), were developed that can be subdivided into two groups [108]. In what is termed an empirical Bayes approach [109], the observed score distribution, which is usually bimodal, is modeled as a mixture model of two parametric distributions assumed to represent correct and incorrect PSMs. For a given score threshold, the FDR can be estimation based on these distributions, exemplified by the PeptideProphet software [110]. In an alternative approach, proper candidates that are a priori known to be incorrect (decoys) are added to the search space and analyzed together with potentially true peptides (targets) [111, 112]. If performed correctly, this so called target-decoy approach allows to estimate the number of false positive results among target hits by counting decoy observations above the specified score threshold, assuming that incorrect (random) identifications are equally likely to originate from either target or decoy database [113]. (FIGURE: Bayes and target-decoy) Decoy candidates, generated most commonly by reverting of target proteins, should preferably be concatenated to the target ones prior to database searching [113]. For spectral library searching, decoy spectra can be generated by shifting the m/z values of the measured MS^2 scan [114].

The major drawback of empirical Bayes is that score distributions of correct and incorrect PSMs need to be estimated accurately, though they might vary substantially between datasets and search engines [114]. Target-decoy, on the other side, needs also to be treated with caution in certain cases and cannot be applied universally without testing its validity, e.g. in conjunction with a specific search engine [115]. In any case, when collapsing PSMs to unique peptides and proteins, the FDR is continually increasing, due to clustering of true positive PSMs to a smaller set of peptides and proteins, unlike incorrect PSMs that match randomly across the entire database [116]. This circumstance requires use of stringent filter criteria at the PSM-level in order to obtain sufficiently reliable protein identifications.

4.2 Reliable PTM Site Assignment

In essence, sites of modification are localized within peptide and protein sequences by applying computer programs, which assign scores to each potential modification assignment (see section 3.5.2). Score cutoffs required for obtaining a reasonable false localization rate (FLR), as it is called, have been determined for some of the tools on the basis of chemically synthesized phospho-peptides, where the correct site is known a priori [16, 17, 103]. Thresholds obtained from such datasets of limited size, are subsequently applied to the analysis of large-scale phospho-proteomics data that aim at determining tens of thousands of sites.

However, it is doubtful that the FLR associated with a specific site localization score can be extrapolated accurately from a confined training set of synthetic phospho-peptides to a dataset that might be up to several hundred times larger and that potentially constitutes of a significantly different population of phosphorylated peptides. Attempts were made to develop statistical approaches that would allow estimation of the FLR [117, 118], similar to those used to assess the error rate of peptide and protein identifications. Especially the large extent of similarity between position isoforms renders the design of such methods a difficult task. Up to now, there exists no generally excepted approach for accurate estimation of the FLR, although it is in demand.

Chapter 5

Results and Interpretation

This chapter comprises two studies, one that presents the potential of EThcD for the analysis of PTMs, such as phosphorylation, and the other introduces a novel approach to accurate FLR estimation. In both cases, the full manuscripts are included as such.

5.1 Unambiguous Phospho-Site Localization Using EThcD

Localization of PTMs, such as phosphorylation, constitutes a difficult task, especially if multiple potential modification sites are present. Successful site assignment requires sufficient sequence coverage of detected fragment ions. A recently developed fragmentation technique, termed EThcD, is shown to enable unambiguous phospho-site localization. Below the respective article published by Frese and co-workers is appended.

Journal of proteome research

Technical Note

pubs.acs.org/jpr

Terms of Use

Unambiguous Phosphosite Localization using Electron-Transfer/Higher-Energy Collision Dissociation (EThcD)

Christian K. Frese,[†,‡] Houjiang Zhou,[†,‡] Thomas Taus,[§] A. F. Maarten Altelaar,[†,‡] Karl Mechtler,[§,∥] Albert J. R. Heck,*[,†,‡] and Shabaz Mohammed*[,†,‡]

[†]Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands.

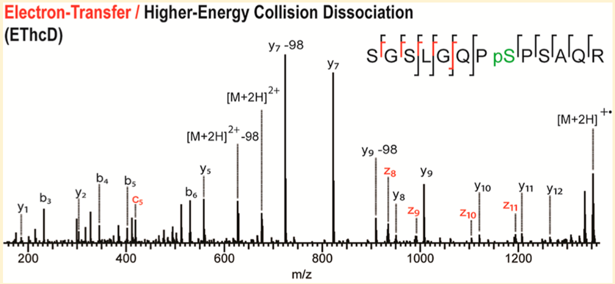
[‡]Netherlands Proteomics Centre, Padualaan 8, 3584 CH Utrecht, The Netherlands

[§]Research Institute of Molecular Pathology (IMP), Dr. Bohrgasse 7, A-1030 Vienna, Austria

[∥]Institute of Molecular Biotechnology (IMBA), Vienna, Austria

Supporting Information

ABSTRACT: We recently introduced a novel scheme combining electron-transfer and higher-energy collision dissociation (termed EThcD), for improved peptide ion fragmentation and identification. We reasoned that phosphosite localization, one of the major hurdles in high-throughput phosphoproteomics, could also highly benefit from the generation of such EThcD spectra. Here, we systematically assessed the impact on phosphosite localization utilizing EThcD in comparison to methods employing either ETD or HCD, respectively, using a defined synthetic phosphopeptide mixture and also using a larger data set of Ti^{4+}-IMAC enriched phosphopeptides from a tryptic human cell line digest. In combination with a modified version of phosphoRS, we observed that in the majority of cases EThcD generated richer and more confidently identified spectra, resulting in superior phosphosite localization scores. Our data demonstrates the distinctive potential of EThcD for PTM localization, also beyond protein phosphorylation.



KEYWORDS: *electron transfer dissociation, ETD, HCD, EThcD, phosphorylation site localization, phosphoRS*

INTRODUCTION

Reversible phosphorylation of proteins is a key regulatory mechanism in living cells.[1] Protein phosphorylation can modulate protein activity, turnover, subcellular localization, complex formation, folding and degradation. Dynamic phosphorylation plays a pivotal role in almost all biological processes including cell division, differentiation, polarization and apoptosis.[2] Moreover, it is an important switch in cellular signal transduction.[3] The importance of this post-translational modification (PTM) for cell biology has driven the development of novel mass spectrometric tools for sensitive and global detection of phosphorylation.[4,5] However, the analysis of phosphorylated peptides by mass spectrometry is still not as straightforward as for "regular", unmodified peptides. One of the major challenges in phosphoproteomics is to improve MS level representation since phosphopeptides are usually present at substoichiometric levels. Hence, an enrichment step is necessary to enable deeper penetration of the phosphoproteome. Enrichment is typically performed by chromatography,[6] antibodies[7] or metal-ion/metal oxide affinity-based[8,9] techniques. Two other main challenges are the identification of phosphopeptides and confident localization of the corresponding phosphosite.[10] The challenge is caused by the higher lability of the phosphate group when compared to the amide bond. A number of strategies have been proposed to circumvent poor fragmentation and improve sequence and site diagnostic fragmentation, including the use of neutral loss-triggered MS/MS/MS[11] and multistage activation (MSA)[12] in ion traps, the use of beam type CID fragmentation,[13] and electron capture/transfer dissociation[14] or a combination of some of these approaches.[9,15]

Once phosphopeptide identification is feasible through sufficient peptide backbone fragments, it can still be challenging to pinpoint the true phosphosite. This becomes more difficult as the number of potential phosphorylation sites within the peptide sequence increases. In principle, unambiguous phosphosite localization requires site-determining fragment ions.[16] Direct validation is feasible through detection of a fragment ion that carries the phosphate group. Neutral loss fragment ions can be used as well; however, since they exhibit the same mass as a water loss from an unmodified residue they do not directly confirm the correct site.[17] Diagnostic phosphosite-specific fragments facilitate pinpointing the correct

Received: December 3, 2012
Published: January 24, 2013

phosphosite.[18−20] Several algorithms and programs have been developed to enable automatic phosphosite localization.[3,16,21−26] These software tools are based on distinct but similar approaches and they all aim to provide a metric that allows for assessment of the confidence in phosphosite localization. Recently, Taus et al. have reported on a new algorithm, coined phosphoRS,[27] which presently is uniquely compatible with CID, HCD and ETD fragmentation and was optimized for both low- and high-resolution MS/MS spectra. phosphoRS provides individual localization probabilities for all potential phosphosites in a given peptide.

Generally, all scoring tools depend on the quality of the MS/MS spectra. The more site-determining ions are detected, the higher the confidence in phosphosite localization. We have recently introduced a novel fragmentation scheme combining electron-transfer and higher-energy collision dissociation, termed EThcD.[28] This method employs dual fragmentation to generate both b/y and c/z ions which leads to very fragment ion- and thus data-rich MS/MS spectra. Compared to HCD and ETD, we found a substantial increase in peptide backbone fragmentation, which translated into a remarkable average peptide sequence coverage of ∼94% for tryptic peptides. We reasoned that localization of post-translational modifications could also highly benefit from EThcD spectra. Here, we systematically assessed the impact on phosphosite localization using EThcD. In this work we evaluate the performance of EThcD in comparison to ETD and HCD using a defined synthetic phosphopeptide mixture and also on a larger data set of Ti^{4+}-IMAC enriched phosphopeptides, all in combination with a modified version of phosphoRS.

■ EXPERIMENTAL SECTION

Materials

All chemicals were purchased from Sigma-Aldrich (Steinheim, Germany) unless otherwise stated. Formic acid and ammonia were obtained from Merck (Darmstadt, Germany). Acetonitrile was purchased from Biosolve (Valkenswaard, The Netherlands).

Sample Preparation

Protein from HeLa cells was harvested and digested with trypsin, as previously described.[29] Ti^{4+}-IMAC beads were prepared as reported elsewhere.[30,31] Phosphopeptides were enriched as previously described.[32] Briefly, Gel-loader tips that were plugged with C8 material (3M, Zoeterwoude, The Netherlands) were filled up to 1 cm with Ti^{4+}-IMAC beads. Columns were equilibrated with loading buffer (80% ACN, 6% TFA). Peptides were reconstituted in loading buffer, loaded onto the columns and washed with washing buffer 1 (50% ACN, 0.5% TFA, 200 mM NaCl) and subsequently washing buffer 2 (50% ACN, 0.1% TFA). Phosphopeptides were eluted with elution buffer 1 (10% NH_3 in H_20) followed by elution buffer 2 (80% ACN, 2% FA). Eluate was acidified and diluted with formic acid to a final acetonitrile concentration of <5%, split into three equal amounts and directly analyzed by single run LC−MS/MS utilizing ETD, HCD and EThcD, respectively.

Mass Spectrometry

All data was acquired on an ETD enabled Thermo Scientific LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). A Thermo Scientific EASY-nLC 1000 (Thermo Fisher Scientific, Odense, Denmark) was connected to the LTQ Orbitrap Velos mass spectrometer. ETD, HCD and EThcD methods were set up as previously described.[28] Briefly, all spectra were acquired in the Orbitrap at a resolution of 7500. For HCD the normalized collision energy was set to 40%. The ETD reaction time was set to 50 ms for ETD and EThcD. Supplemental activation was enabled for ETD. HCD normalized collision energy was set to 30% for EThcD (calculation based on precursor m/z and charge state). The anion AGC target was set to 4e5 for both ETD and EThcD.

Data Analysis

Peak lists were generated using Thermo Scientific Proteome Discoverer 1.3 software (Thermo Fisher Scientific, Bremen, Germany). The nonfragment filter was used to simplify ETD spectra with the following settings: the precursor peak was removed within a 4 Da window, charged reduced precursors were removed within a 2 Da window, and neutral losses from charge reduced precursors were removed within a 2 Da window (the maximum neutral loss mass was set to 120 Da). MS/MS spectra were searched against a database containing the synthetic phosphopeptide sequences and the human Uniprot database (version v2010−12), respectively, including a list of common contaminants using SEQUEST or Mascot (Matrix Science, UK). The precursor mass tolerance was set to 10 ppm, the fragment ion mass tolerance was set to 0.02 Da. Enzyme specificity was set to Trypsin with 2 missed cleavages allowed. Data from the synthetic phosphopeptide mixture was searched with no enzyme specificity. Oxidation of methionine and phosphorylation (S,T,Y) were used as variable modification and carbamidomethylation of cysteines was set as fixed modification. Percolator[33] was used to filter the PSMs for <1% false-discovery-rate. Phosphorylation sites were localized by applying a custom version of phosphoRS[27] (v3.0 − EThcD enabled) that has been expanded to allow analysis of EThcD data.[28] Briefly, the algorithm considers both HCD- and ETD-type fragment ions at the same time. While singly and doubly charged b- and y-type fragment ions including neutral loss of phosphoric acid (H_3PO_4) are considered for site localization, only singly charged c-, z-radical and z-prime ions are scored.

■ RESULTS AND DISCUSSION

Increasing the confidence in phosphosite localization is a key challenge in phosphoproteomics. Site-determining fragment ions are required to unambiguously pinpoint the correct phosphosite. Observing all possible peptide backbone cleavages in a single MS/MS spectrum substantially simplifies phosphosite localization. Recently, we showed that EThcD enables complete peptide sequencing through dual fragmentation.[28] In EThcD, the peptide precursor is initially subjected to an ion/ion reaction with fluoranthene anions in a linear ion trap, which generates c- and z-ions. However, the unreacted precursor and the charge-reduced precursor remain highly abundant after ETD. In the second step HCD all-ion fragmentation is applied to all ETD derived ions. This generates b- and y-ions from the unreacted precursor and simultaneously increases the yield of c- and z-ions by fragmentation of the charge reduced precursor. Since the remaining unreacted precursor population is higher charged than the ETD-derived fragment ions one can apply a level of energy that fragments the precursor but does not induce secondary fragmentation of c- and z-ions. Here, we continue to explore the benefits of this novel fragmentation mode for the analysis of phosphopeptides.

Evaluation of Phosphosite Localization by EThcD using a Defined Phosphopeptide Mixture

To evaluate the potential added value of phosphopeptide analysis by EThcD we initially used a defined mixture of well-characterized synthetic phosphopeptides. This mixture consists of 30 phosphopeptides of varying length with up to four phosphorylated residues (see Supplementary Table 1 for a complete list, Supporting Information). We analyzed this mixture by LC—MS/MS employing ETD, HCD and EThcD fragmentation, respectively. We used identical instrument settings with the only exception being the parameters for peptide dissociation, which were set to the for each method optimized values. The data was searched with SEQUEST and the PSMs were manually validated and filtered (7 ppm peptide mass tolerance, search engine rank 1, absolute Xcorr threshold 0.4). Additionally, we considered only PSMs for which the injection time did not max out (<500 ms), that is, the target number of ions was reached. Note that this precaution was taken to exclude the number of ions as a variable that might impair the quality of fragmentation. We calculated the average precursor ion purity (PIP)[34] for each data set and found similar values, which were approximately 95% for all three techniques. Together, these stringent criteria ensure that the activation technique is the only variable that controls the fragmentation behavior. A summary of the data from this direct comparison is given in Table 1. Similar numbers of PSMs were identified for all three fragmentation techniques. We found that EThcD provided 248 PSMs while these numbers were 237 and 216 for HCD and ETD, respectively. Out of the 30 unique synthetic phosphopeptides injected ETD, HCD and EThCD identified 21, 22 and 24, respectively. We found the average SEQUEST Xcorr being highest for EThcD (2.5) followed by HCD (1.9) and ETD (1.5), which is in line with our previous results for nonmodified peptides.[28] The SEQUEST algorithm correctly annotated the known phosphosites in 79% of ETD and 78% of HCD data. Significantly, for EThcD this was over 95% (of all PSMs), which directly reflects the higher spectral quality, due to the generation of both b/y and c/z ions. This initial data suggests that EThcD provides even more extensive backbone fragmentation of phosphorylated peptides than ETD or HCD alone, facilitating sensitive phosphosite localization with very high confidence. It should be noted that the application of a site localization algorithm would be prudent for real-life samples since the true phosphorylation sites are unknown.

Table 1. Analysis of 30 Synthetic Phosphopeptides

	ETD	HCD	EThcD
#PSM	216	237	248
# unique peptides	21/30	22/30	24/30
average Xcorr	1.5	1.9	2.5
% PSM with correctly localized phosphosite (SEQUEST)	79%	78%	95%
# phosphosites with phosphoRS site probability >99%	478	410	423
% phosphosites with phosphoRS site probability >99%	96%	95%	97%

Recently, Taus et al. described phosphoRS, a novel tool to improve confident localization of phosphosites.[27] The software is based on validated peptide identifications provided by database search engines and calculates site probabilities for each potential phosphosite in the peptide sequence. For this study we used a modified version of phosphoRS that also enables assessment of individual phosphosite probabilities for EThcD fragmentation. We analyzed each data set using phosphoRS and found that it performs equally well for all three fragmentation techniques. Of all true phosphosites, 96% (ETD), 95% (HCD) and 97% (EThcD) were assigned a site probability >99%, which corresponds to a very high confidence in site localization (Table 1). Together, these findings suggest that EThcD generates MS/MS spectra that contain sufficient fragment ions for the unambiguous and sensitive phosphorylation site localization.

Phosphosite Localization of Ti^{4+}-IMAC Enriched Phosphopeptides by EThcD

Next, we assessed the performance of EThcD for phosphosite localization on a larger data set. We used Ti^{4+}-IMAC material for the enrichment of phosphopeptides from a tryptic digest of HeLa cells and analyzed equal amounts (corresponding to enriched phosphopeptides from 100 μg of protein) by LC—MS/MS with ETD, HCD and EThcD, respectively (Supplementary Figure 1A, Supporting Information). All three methods generated a similar number of MS/MS spectra. All spectra were searched with SEQUEST. The ETD data was also searched with Mascot because we found SEQUEST to perform poorly for doubly charged phosphopeptides. Note that other search engines such as OMSSA or SpectrumMill might provide larger number of identifications for ETD data.[35] However, these algorithms are currently not compatible with EThcD data and phosphoRS analysis within the Proteome Discoverer software environment. All identified PSMs were then filtered for <1% FDR using percolator to ensure consistency. In total we identified 2217 (ETD), 4179 (HCD) and 3594 (EThcD) phospho-PSMs (Table 2). Our initial analysis of a defined synthetic phosphopeptide mixture demonstrated that EThcD performs at least on the same level as HCD in terms of peptide identification. However, the overall identification success rate in the Ti^{4+}-IMAC data set was slightly lower for EThcD compared to HCD. This can be attributed to the rigid automatic FDR filtering. The MS/MS spectra from the synthetic phosphopeptide mixture were manually validated whereas the Ti^{4+}-IMAC data set was computationally filtered to <1% FDR. The application of EThcD, in comparison to ETD or HCD alone, significantly increases the number of fragment ions observed in the MS/MS scans. On the one hand EThcD spectra contain more sequence information, which is beneficial for inferring the peptide sequence and PTM localization. On the other hand, these additional fragment ions may also match to random peptide sequences, increasing their score and hampering the differentiation between correct and incorrect matches. Consequently, the chance for a high scoring random match will be elevated. Similar to the increased average score of decoy hits

Table 2. LC—MS/MS Analysis of Ti^{4+}-IMAC Enriched Tryptic Phosphopeptides Originating from a Cellular Lysate using ETD, HCD and EThcD

	ETD	HCD	EThcD
#PSM	2266	4282	3679
ID success rate	25%	51%	44%
average Xcorr	1.9	2.5	3.2
% average peptide sequence coverage	83%	81%	92%
# phospho-PSM	2217	4179	3594
# phospho-sites >99% pRS probability	2002	4291	3942
% phospho-sites >99% pRS probability	81%	89%	95%

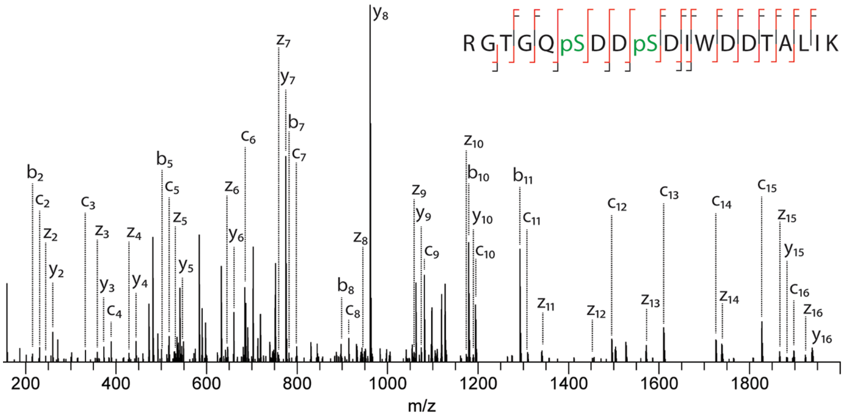


Figure 1. EThcD MS/MS spectrum of a doubly phosphorylated peptide. RGTGQSDDSDIWDDTALIK is doubly phosphorylated and contains in total four potential phosphorylation sites. EThcD generates dual ion series that enable phosphorylation site localization with very high confidence (phosphoRS site probabilities: T(3), 0.0%; S(6), 100.0%; S(9), 100.0%; T(15), 0.0%). SEQUEST Xcorr 7.79.

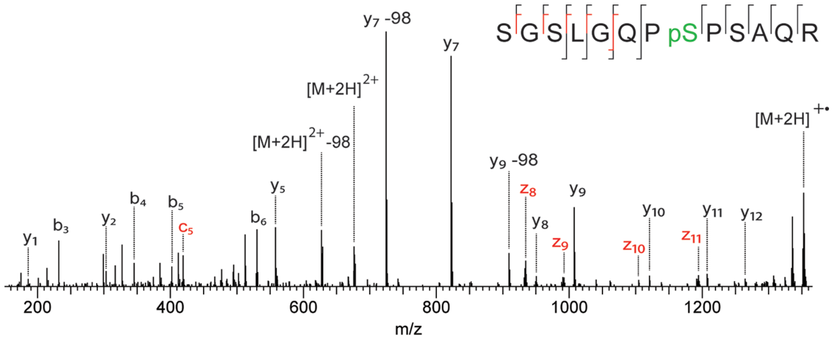


Figure 2. EThcD spectrum of a proline-containing phosphopeptide. This EThcD spectrum of a doubly charged peptide that contains four serine residues, one of which is phosphorylated. ETD does not cleave the N−C$_\alpha$ bond N-terminal to proline and the phosphorylation site probability is only 50% based on c- and z-ions alone. Dual fragmentation by EThcD generates complementary sequence information from c/z- and b/y-ions (SEQUEST Xcorr 4.10). Here, the exact phosphosite is revealed by y-ions that cover the corresponding phosphosite (phosphoRS site probabilitis: S(1): 0.0; S(3): 0.0; S(8): 99.5; S(10): 0.5). SEQUEST Xcorr 4.10.

also the true hits are likely to provide on average higher scores. Depending on whether the distance between the two score distributions decreases or increases, the identification success rate will be higher or lower. Since the ID success rate is slightly lower for EThcD compared to HCD alone, the negative effect of higher-scoring random matches might be more pronounced. Thus, higher score cut-offs need to be applied in order to reach the desired FDR. A standard target-decoy approach[36] against a reversed concatenated database revealed the FDR for EThcD (2.6%) being almost twice as high compared to HCD (1.4%), which provides further evidence for this hypothesis.

Next, we calculated the average peptide sequence coverage for all PSM. As expected, EThcD provided a substantial increase in sequence coverage (92%) compared to HCD (81%) and ETD (83%). Obtaining near-complete peptide sequence coverage tremendously simplifies phosphosite localization. We used the extended phosphoRS algorithm to validate our assumption. Remarkably, EThcD provided for 95% of all phosphosites a confident site localization probability of >99%. In the HCD data set we found that 89% of all phosphosites were assigned with a confident site localization probability >99%, while this was only 81% for ETD data set. We recalculated these number for all peptides that contain >2 residues that can be phosphorylated because singly phosphorylated peptides with only one potential phosphorylation site could bias the results toward HCD. Of all phosphosites from this subset of peptides 97% (ETcaD), 93% (EThcD) and 87% (HCD), respectively, were assigned a localization probability >99%.

For multiply phosphorylated peptides site localization becomes more challenging. Figure 1 shows an MS/MS spectrum of a doubly phosphorylated peptide upon EThcD fragmentation. The overall sequence coverage is 89% taking b/y- and c/z-ions into account. Six out of 18 amino acid bond cleavages are represented by c- and b-ions (referred to as "golden pairs"[37]). Additionally, we observed 11 z/y-ion pairs, which strengthens the argument that EThcD provides extensive sequence information that facilitates pinpointing the correct phosphorylation site. More than 95% of the phosphosites from all doubly phosphorylated peptides were assigned with a site localization probability >99%, highlighting that EThcD performs equally well with singly and doubly phosphorylated peptides. A known limitation of ETD is its inability to cleave the N−C$_\alpha$ bond N-terminal to proline.[38,39] This can hamper phosphosite localization for proline-rich peptides. Generation of dual ion series in EThcD can overcome this issue. Figure 2 shows the EThcD spectrum of a singly phosphorylated peptide that contains four serine residues. The c- and z-ions derived from the ETD step cover only the N-terminal part of the peptide and the site probability is only 50%. The additional y-ions derived from the subsequent HCD activation provide supporting sequence information and cover also the two serine residues next to the prolines which enables unambiguous phosphosite localization.

CONCLUSIONS

Here we have evaluated the potential of EThcD in improving the analysis of phosphopeptides. Our data highlights the benefit of dual ion series as generated by EThcD fragmentation. We observed for a defined phosphopeptide mixture average higher SEQUEST Xcorr values, higher peptide sequence coverage and more confident phosphosite localization in EThcD compared to ETD and HCD. This finding was confirmed when we analyzed a complex phosphopeptide sample resulting from a Ti^{4+}-IMAC enrichment of peptides from a cellular lysate. This is in line with recent reports that showed that confidence in phosphorylation site localization increases when multiple separately acquired MS/MS spectra (e.g., ETD/CID or MSA/ETD) are combined for scoring.[25,26] For this larger data set, we observed that the identification success rate was slightly lower for EThcD compared to HCD. This can be attributed to the use of conventional database search engines that are not optimized for spectra that contain dual ion series.[40] However, the fact that both peptide sequence coverage and the percentage of localized phosphosites are higher for EThcD than for HCD suggests that once a peptide was identified, further analyses such as site localization benefit from the more data-rich EThcD spectra. In EThcD often c/b- and z/y-ion pairs are observed that increase the confidence in a particular peptide backbone cleavage.[41] We speculate that the identification success rate of EThcD for phosphopeptides can be improved by novel or optimized data analysis tools. Finally, we reason that EThcD can also be beneficial and used to improve the localization of other post-translational modifications such as ubiquitination, glycosylation or acetylation.

ASSOCIATED CONTENT

Supporting Information

Additional information as noted in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: a.j.r.heck@uu.nl, s.mohammed@uu.nl.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Mathias Madalinski for peptide synthesis of the primeXS phosphopeptide mixture. This research was performed within the framework of the PRIME-XS project, grant number 262067, funded by the European Union 7th Framework Program. Additionally, The Netherlands Proteomics Centre, a program embedded in The Netherlands Genomics Initiative, is kindly acknowledged for financial support as well as The Netherlands Organization for Scientific Research (NWO) with the VIDI grant (700.10.429). Work in the Mechtler lab was supported by the European Commission via the FP7 projects MeioSys and PRIME-XS, the Austrian Science Fund via the Special Research Program Chromosome Dynamics (SFB-F3402).

REFERENCES

(1) Cohen, P. The regulation of protein function by multisite phosphorylation–a 25 year update. *Trends Biochem. Sci.* **2000**, *25* (12), 596−601.

(2) Hunter, T. Signaling–2000 and beyond. *Cell* **2000**, *100* (1), 113−27.

(3) Olsen, J. V.; 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* (3), 635−48.

(4) Eyrich, B.; Sickmann, A.; Zahedi, R. P. Catch me if you can: mass spectrometry-based phosphoproteomics and quantification strategies. *Proteomics* **2011**, *11* (4), 554−70.

(5) Mann, M.; Jensen, O. N. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* **2003**, *21* (3), 255−61.

(6) Di Palma, S.; Hennrich, M. L.; Heck, A. J.; Mohammed, S. Recent advances in peptide separation by multidimensional liquid chromatography for proteome analysis. *J. Proteomics* **2012**, *75* (13), 3791−813.

(7) Rush, J.; Moritz, A.; Lee, K. A.; Guo, A.; Goss, V. L.; Spek, E. J.; Zhang, H.; Zha, X. M.; Polakiewicz, R. D.; Comb, M. J. Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat. Biotechnol.* **2005**, *23* (1), 94−101.

(8) Ficarro, S. B.; McCleland, M. L.; Stukenberg, P. T.; Burke, D. J.; Ross, M. M.; Shabanowitz, J.; Hunt, D. F.; White, F. M. Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. *Nat. Biotechnol.* **2002**, *20* (3), 301−5.

(9) Zhou, H.; Low, T. Y.; Hennrich, M. L.; van der Toorn, H.; Schwend, T.; Zou, H.; Mohammed, S.; Heck, A. J. Enhancing the identification of phosphopeptides from putative basophilic kinase substrates using Ti (IV) based IMAC enrichment. *Mol. Cell. Proteomics* **2011**, *10* (10), M110 006452.

(10) Boersema, P. J.; Mohammed, S.; Heck, A. J. Phosphopeptide fragmentation and analysis by mass spectrometry. *J. Mass Spectrom.* **2009**, *44* (6), 861−78.

(11) Beausoleil, S. A.; Jedrychowski, M.; Schwartz, D.; Elias, J. E.; Villen, J.; Li, J.; Cohn, M. A.; Cantley, L. C.; Gygi, S. P. Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (33), 12130−5.

(12) Schroeder, M. J.; Shabanowitz, J.; Schwartz, J. C.; Hunt, D. F.; Coon, J. J. A neutral loss activation method for improved phosphopeptide sequence analysis by quadrupole ion trap mass spectrometry. *Anal. Chem.* **2004**, *76* (13), 3590−8.

(13) Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **2007**, *4* (9), 709−12.

(14) Chi, A.; Huttenhower, C.; Geer, L. Y.; Coon, J. J.; Syka, J. E.; Bai, D. L.; Shabanowitz, J.; Burke, D. J.; Troyanskaya, O. G.; Hunt, D. F. Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (7), 2193−8.

(15) Molina, H.; Matthiesen, R.; Kandasamy, K.; Pandey, A. Comprehensive comparison of collision induced dissociation and electron transfer dissociation. *Anal. Chem.* **2008**, *80* (13), 4825−35.

(16) Beausoleil, S. A.; Villen, J.; Gerber, S. A.; Rush, J.; Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **2006**, *24* (10), 1285−92.

(17) Stensballe, A.; Jensen, O. N.; Olsen, J. V.; Haselmann, K. F.; Zubarev, R. A. Electron capture dissociation of singly and multiply phosphorylated peptides. *Rapid Commun. Mass Spectrom.* **2000**, *14* (19), 1793−800.

(18) Kelstrup, C. D.; Hekmat, O.; Francavilla, C.; Olsen, J. V. Pinpointing phosphorylation sites: Quantitative filtering and a novel site-specific x-ion fragment. *J. Proteome Res.* **2011**, *10* (7), 2937−48.

(19) Shin, Y. S.; Moon, J. H.; Kim, M. S. Observation of phosphorylation site-specific dissociation of singly protonated phosphopeptides. *J. Am. Soc. Mass Spectrom.* **2010**, *21* (1), 53−9.

(20) Gehrig, P. M.; Roschitzki, B.; Rutishauser, D.; Reiland, S.; Schlapbach, R. Phosphorylated serine and threonine residues promote

site-specific fragmentation of singly charged, arginine-containing peptide ions. *Rapid Commun. Mass Spectrom.* **2009**, *23* (10), 1435−45.

(21) Lu, B.; Ruse, C.; Xu, T.; Park, S. K.; Yates, J. R., 3rd Automatic validation of phosphopeptide identifications from tandem mass spectra. *Anal. Chem.* **2007**, *79* (4), 1301−10.

(22) Bailey, C. M.; Sweet, S. M.; Cunningham, D. L.; Zeller, M.; Heath, J. K.; Cooper, H. J. SLoMo: automated site localization of modifications from ETD/ECD mass spectra. *J. Proteome Res.* **2009**, *8* (4), 1965−71.

(23) Savitski, M. M.; Mathieson, T.; Becher, I.; Bantscheff, M. H-score, a mass accuracy driven rescoring approach for improved peptide identification in modification rich samples. *J. Proteome Res.* **2010**, *9* (11), 5511−6.

(24) Ruttenberg, B. E.; Pisitkun, T.; Knepper, M. A.; Hoffert, J. D. PhosphoScore: an open-source phosphorylation site assignment tool for MSn data. *Journal of Proteome Research* **2008**, *7* (7), 3054−9.

(25) Hansen, T. A.; Sylvester, M.; Jensen, O. N.; Kjeldsen, F. Automated and high confidence protein phosphorylation site localization using complementary collision-activated dissociation and electron transfer dissociation tandem mass spectrometry. *Anal. Chem.* **2012**, *84* (22), 9694−9.

(26) Vandenbogaert, M.; Hourdel, V.; Jardin-Mathe, O.; Bigeard, J.; Bonhomme, L.; Legros, V.; Hirt, H.; Schwikowski, B.; Pflieger, D. Automated phosphopeptide identification using multiple MS/MS fragmentation modes. *J. Proteome Res.* **2012**, *11* (12), 5695−703.

(27) Taus, T.; Kocher, T.; Pichler, P.; Paschke, C.; Schmidt, A.; Henrich, C.; Mechtler, K. Universal and confident phosphorylation site localization using phosphoRS. *J. Proteome Res.* **2011**, *10* (12), 5354−62.

(28) Frese, C. K.; Altelaar, M.; van den Toorn, H. W.; Nolting, D.; Griep-Raming, J.; Heck, A. J.; Mohammed, S. Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. *Anal. Chem.* **2012**, *84* (22), 9668−73.

(29) Altelaar, A. F.; Frese, C. K.; Preisinger, C.; Hennrich, M. L.; Schram, A. W.; Timmers, H. T.; Heck, A. J.; Mohammed, S. Benchmarking stable isotope labeling based quantitative proteomics. *J. Proteomics* **2012**, DOI: 10.1016/j.jprot.2012.10.009.

(30) Zhou, H.; Ye, M.; Dong, J.; Han, G.; Jiang, X.; Wu, R.; Zou, H. Specific phosphopeptide enrichment with immobilized titanium ion affinity chromatography adsorbent for phosphoproteome analysis. *J. Proteome Res.* **2008**, *7* (9), 3957−67.

(31) Yu, Z.; Han, G.; Sun, S.; Jiang, X.; Chen, R.; Wang, F.; Wu, R.; Ye, M.; Zou, H. Preparation of monodisperse immobilized Ti(4+) affinity chromatography microspheres for specific enrichment of phosphopeptides. *Anal. Chim. Acta* **2009**, *636* (1), 34−41.

(32) Zhou, H.; Ye, M.; Dong, J.; Corradini, E.; Cristobal, A.; Heck, A. J. R.; Zou, H.; Mohammed, S. Robust phosphoproteome enrichment using monodisperse microspheres-based immobilized titanium (IV) ion affinity chromatography. *Nat. Protoc.* **2012**, accepted.

(33) Kall, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4* (11), 923−5.

(34) Mertins, P.; Udeshi, N. D.; Clauser, K. R.; Mani, D. R.; Patel, J.; Ong, S. E.; Jaffe, J. D.; Carr, S. A. iTRAQ labeling is superior to mTRAQ for quantitative global proteomics and phosphoproteomics. *Mol. Cell. Proteomics* **2012**, *11* (6), M111 014423.

(35) Kandasamy, K.; Pandey, A.; Molina, H. Evaluation of Several MS/MS Search Algorithms for Analysis of Spectra Derived from Electron Transfer Dissociation Experiments. *Anal. Chem.* **2009**, *81* (17), 7170−80.

(36) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4* (3), 207−14.

(37) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. Automated de novo sequencing of proteins by tandem high-resolution mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97* (19), 10313−7.

(38) Cooper, H. J.; Hudgins, R. R.; Hakansson, K.; Marshall, A. G. Secondary fragmentation of linear peptides in electron capture dissociation. *Int. J. Mass Spectrom.* **2003**, *228* (2−3), 723−8.

(39) Li, W.; Song, C.; Bailey, D. J.; Tseng, G. C.; Coon, J. J.; Wysocki, V. H. Statistical analysis of electron transfer dissociation pairwise fragmentation patterns. *Anal. Chem.* **2011**, *83* (24), 9540−5.

(40) Kim, M. S.; Zhong, J.; Kandasamy, K.; Delanghe, B.; Pandey, A. Systematic evaluation of alternating CID and ETD fragmentation for phosphorylated peptides. *Proteomics* **2011**, *11* (12), 2568−72.

(41) Nielsen, M. L.; Savitski, M. M.; Zubarev, R. A. Improving protein identification using complementary fragmentation techniques in fourier transform mass spectrometry. *Mol. Cell. Proteomics* **2005**, *4* (6), 835−45.

5.2 Accurate FLR Estimation in Phospho-Proteomics

Especially in case of high-throughput localization of phosphorylation sites it is important to estimate the fraction of false positive assignments. Until now, there exists now commonly accepted approach for this task. In the following a study is presented that introduces accurate FLR estimation based on fragment ion offsetting. The appended manuscript is planned to be published in a peer-reviewed scientific journal.

Accurate Estimation of False Localization Rates in Phospho-Proteomics

*Thomas Taus[1,2], Etienne Beltzung[3], Thomas Köcher[1,2], Gerhard Dürnberger[4] and Karl Mechtler[1,2]**

1. Research Institute of Molecular Pathology (IMP), Vienna, Austria.
2. Institute of Molecular Biotechnology (IMBA), Vienna, Austria.
3. Max F. Perutz Laboratories (MFPL), Vienna, Austria.
4. Gregor Mendel Institute of Molecular Plant Biology (GMI), Vienna, Austria.

* Author for correspondence

karl.mechtler@imp.ac.at

Post-translational modifications, such as phosphorylation, play a key role in a variety of important cellular processes. In-depth characterization of phosphorylated peptides and proteins, including assignment of the exact site of modification, is therefore of great importance for the biological questions addressed. Based on tandem mass spectrometry data phosphorylation sites can be localized in high-throughput deploying dedicated computational methods. These algorithms assess the most probable assignments and provide scores estimating the confidence of reported results. In most cases, site localization tools have been validated on a limited set of chemically synthesized phospho-peptides. Nevertheless, they are commonly applied to complex biological samples comprising potentially a significantly different population of phospho-peptides. Until now there exists no generally accepted approach for estimating accurately so called false localization rates (FLR), although two methods have been suggested recently. Here we present a strategy for FLR estimation that is based on fragment ion offsetting. In order to rigorously validate the method we generated chemically synthesized phospho-peptide libraries comprising almost 60,000 distinct species. In total, more than 700,000 highly confident PSMs of different fragmentation techniques were acquired that can serve as a valuable resource for a broad range of applications. Obtained results suggest that fragment ion offsetting enables accurate estimation of FLR even if challenged with low quality MS/MS spectra.

KEYWORDS: mass spectrometry, protein phosphorylation, tandem mass spectrometry, false localization rate.

Introduction

The complex and dynamic biochemical machinery, which enables both reaction to external stimuli and adaptation to environmental changes, constitutes an essential feature of all living organisms. Post-translational modifications (PTMs) form extensively interconnected regulatory circuits[1] and allow rapid molecular dynamics by modulating activity, stability, spatial localization and complex formation of proteins[2]. The actual sites bearing the modification can be of crucial importance for the biological function conducted[3]. Reversible protein phosphorylation, probably the most ubiquitous PTM, is controlled by a finely coordinated network of kinases and phosphatases, and aberrant phosphorylation can be a cause or consequence of diseases such as cancer[4,5]. Hence, systematic analysis of this widespread PTM is of major importance.

Mass spectrometry (MS) enables unbiased protein characterization[6] and has evolved from a method capable of identifying a limited number of proteins to a high-throughput technology enabling the assessment of whole proteomes, including that of human[7]. Nowadays, MS is the method of choice for protein characterization, including the analysis of PTMs, such as phosphorylation[8,9]. Typically, proteins are digested with specific proteases and resulting proteolytic peptides are separated with nano-flow high performance liquid chromatography (LC) online coupled via electrospray ionization[10] to tandem mass spectrometry (MS/MS).

Acquired spectral data is then analyzed computationally by a series of sophisticated algorithms aiming at identification[11–16], quantification[17–19] and characterization of peptides and proteins, including the detection of their interaction partners[20] and localization of PTMs[21–24]. Nevertheless, such *in silico* tasks are challenging and occasionally error-prone exercises, owing for example to noisy data, high sample complexity giving rise to chimeric MS/MS spectra[25,26], low-quality MS^2 scans lacking sufficient sequence information or the correct interpretation of a spectrum might not have been considered during the analysis. Regarding that for instance extensively applied protein database search engines generate peptide to spectrum matches (PSMs) for almost all input MS/MS scans, given that there is at least one peptide candidate that meets predefined requirements, it is indispensable to sort out incorrect identifications[27]. For this task, constant score thresholds had been applied to a variety of significantly different datasets, until the claim was raised that advanced statistical considerations are required, in order to ensure quality of reported results[28]. It can be assumed that the situation is quite similar for PTM localization, since most of the commonly used tools were validated on a limited set of chemically synthesized phospho-peptides and the cutoff values derived from this validation phase are now widely applied to datasets that might comprise a substantially different population of phospho-peptides.

Regarding peptide and protein identification, two major categories of approaches have been introduced that enable the estimation of false positive matches, also referred to as false discovery rate (FDR). In what is termed an empirical Bayes approach[29], the observed score distribution is approximated by a mixture model of two distributions that are assumed to represent correct and incorrect PSMs in order to estimate the FDR of matches exceeding a certain score[30]. The alternative strategy is termed target-decoy approach[31,32]. In brief, PSMs are searched against a set of potentially correct peptides (target) and entrapment sequences (decoy), which are *a priori* known to be incorrect. If decoy peptides are generated in an appropriate way, then the FDR of identified PSMs can be assessed by the ratio of decoy and target hits above a specified score threshold[33]. In case of phospho-site assignment, the so called false localization rate (FLR) could in principle also be estimated by a target-decoy-based approach. However, generation of adequate decoy sites poses a major challenge, since the target-decoy approach requires that random matches are equally likely to originate from either the target or decoy set. Considering additionally entrapment amino acids, such as glutamic acid or proline[34], as potential targets of phosphorylation distorts observed S, T and Y patterns, which in turn affects site localization. This can be addressed to changes regarding the number of and distance between putatively phosphorylated amino acids within a peptide sequence. The described effect might be even more pronounced if all but S, T and Y amino acids[35] are considered as decoy sites. Thus, we believe that addition of entrapment amino acids might be disadvantageous and aimed at developing another approach to generate decoy sites.

Here we present a method for accurate estimation of the FLR, which is based on fragment ion offsetting. The unique feature of this novel approach is that intrinsically both the patterns of and distances between phosphorylation targets are maintained, allowing random (incorrect) assignments to originate equally likely from either the target or decoy set. In order to rigorously validate our method, we designed chemically synthesized peptide libraries[36] and subjected them to LC-MS/MS analysis applying collision induced dissociation (CID), electron transfer dissociation (ETD)[37], higher energy collisional dissociation (HCD)[38] and the recently introduced electron transfer/higher energy collisional dissociation (EThcD)[39], which has been suggested to have favorable figures of merit especially in case of labile PTMs, such as phosphorylation[40]. Obtained results indicate that using fragment ion offsetting the FLR can be estimated accurately, even when challenged by the analysis of low quality MS/MS spectra. Further, the acquired set of in total roughly 730,000 highly confident (1% FDR) PSMs from synthetic phospho-peptides constitute a valuable resource for a variety of versatile applications.

Materials and Methods

Sample Preparation. Phosphorylated peptides were chemically synthesized applying solid-phase Fmoc-chemistry (Novabiochem) using a Syro instrument (MultiSynTech). Starting from the C-terminus, single amino acids were sequentially concatenated, except at permutation positions, where an equimolar mixture of 18 amino acids was added, in order to create a variety of distinct phosphorylated peptides in a single synthesis run. Based on 188 seed peptides with two permutation positions each, this strategy gave rise to theoretically 59,688 individual phospho-peptides. For subsequent LC-MS/MS analysis, phospho-peptides were dissolved first in an aqueous solution containing 30% acetonitrile (ACN) and diluted to a concentration of 2 pmol/mL in 0.1% trifluoroacetic acid (TFA), assuming a yield of 2 µmol. Finally, peptide libraries were combined into mixes of 5 by randomly choosing from distinct peptide length bins, enhancing retention time distribution of eluting analytes. It was required that phospho-site isomers were distributed to distinct mixes.

LC-MS/MS analysis. Peptide library mixtures were subjected to nano-HPLC-MS/MS analysis, deploying an UltiMate 3000 RSLCnano system (Dionex Thermo Fisher) online coupled to an LTQ Orbitrap Velos/VelosPro ETD (Thermo Fisher Scientific). Separation of peptides was carried out on a C18 column (Acclaim PepMap 100, nanoViper, 50 cm x 75 µm, 2 µm, 100 Å, Dionex Thermo Fisher) using the following solvent system: A: 0.1% TFA and B: 80% ACN, 0.08% TFA. Synthetic phospho-peptides were separated using a 120 min gradient from 2% to 35% B, followed by a 5 min gradient to 90% B.

MS 1 survey scans were performed in the orbitrap mass analyzer, recording a window between 300 and 1800 m/z at a resolution of 60,000 with the automatic gain control (AGC) set to 10^6 and a maximal injection time of 500 ms. For both MS and MS/MS acquisition one microscan was recorded and internal recalibration of mass spectra was performed by enabling the lock mass option based on polydimethylcyclosiloxane ions (protonated $(Si(CH_3)_2O)_6$; 445.120025 m/z). Tandem MS was performed in a data-dependent fashion, selecting the three most abundant precursor ions for CID, ETD, HCD and EThcD. Acquisition of CID and ETD spectra in the linear ion trap was performed with an AGC target value of 10,000 and a maximal ion inject time of 200 ms. For CID a normalized collision energy (NCE) of 35%, a Q value of 0.25 and an activation time of 10 ms was used. ETD was performed with fluoranthene as electron donor, applying supplemental activation and using a precursor charge state-dependent reaction time with 2+ as default charge state and 90 ms corresponding activation time. Both HCD and EThcD spectra were acquired in the orbitrap mass analyzer with an AGC target value of 200,000 and a maximal ion inject time of 250 ms. The NCE was set to 28% and 27% for HCD and EThcD, respectively. For acquisition of low quality spectra, AGC target values were set to 100 and 20,000 for ion trap and orbitrap readout, respectively.

Data analysis. Recorded MS/MS spectra were analyzed using Proteome Discoverer (v.1.4.0.288, Thermo Fisher Scientific) applying Mascot (v.2.2.07, Matrix Science) for peptide identification. All searches were performed against an in-house generated database comprising all human and *Bacillus subtilis* entries from SwissProt (release November 2012) and synthetic phospho-peptides as individual sequence entries. Subsequently, SequenceReverser.exe (v.1.0.13.13, Max Planck Institute of Biochemistry)[18] was deployed to generate a concatenated forward/reverse database, including a list of common contaminants. For all searches, a precursor ion mass tolerance of 10 ppm was specified, allowing up to four missed cleavage sites for trypsin. In case of CID and ETD, the fragment ion mass tolerance was set to 0.5 Da, whereas for HCD and EThcD it was limited to 0.02 Da. Phosphorylation of serine, threonine and tyrosine, as well as oxidation of methionine were specified as variable modifications. All peptide to spectrum matches (PSMs) being a rank one identification with a minimal peptide length of 7 amino acids were filtered to a false discovery rate (FDR) of 1% applying the target/decoy approach[33]. Further, peptides were required to correspond to any of the phospho-peptides of the generated synthetic libraries.

For calculation of site probabilities and subsequent FLR estimation an in-house version of phosphoRS[24] was used. In brief, based on estimating the likelihood that the observed match between a positional phospho-isoform and the respective MS/MS spectrum has occurred just by chance, phosphoRS aims at deriving individual site probability values for each putative phosphorylation site. In order to model the frequency of incorrect site localizations for a phospho-proteomic dataset, target and decoy sites, which are generated by offsetting the theoretical m/z values of site determining fragment ions, are analyzed together. Finally, the FLR of obtained site assignments can be estimated by dividing decoy by target sites that both exceed a specified site probability threshold. All downstream calculations of obtained phosphoRS output were performed using R programming language (v.2.15.1., R Foundation for Statistical Computing, www.R-project.org).

Results and discussion

The ability to estimate accurately the FLR of reported phosphorylation site assignments is of utmost importance in order to provide a solid basis for continuative biological studies. To the best of our knowledge, there exists still no generally accepted approach for this task, although two methods have been suggested recently[34,35]. Aiming at filling this gap, we developed a novel approach based on fragment ion offsetting that enables estimation of the global FLR, given a set of peptide sequences identified at a specified FDR. This method should allow accurate assessment of the FLR for MS/MS data generated with different fragmentation regimes, such as CID, ETD, HCD and EThcD, and it should be applicable for variable mass accuracy.

Fragment ion offsetting basis on generating entrapment sites and analyzing them with a custom version of phosphoRS together with putatively correct (target) ones (Figure 1). Given a phospho-PSM, the software analyzes first the target sites as described previously[24]. In a next step, site determining ions that discriminate between the positional isoforms are identified and for each isoform those theoretical fragment ions are shifted, which distinguish it from the rank one isoform. In case of the top scoring isoform itself, m/z values of theoretical fragment ions that distinguish it from the rank two hit are offset. Decoy sites generated in this manner are then scored and individual localization probabilities are assessed. After comparing the highest-ranking target and decoy isoforms in terms of peptide score, the phosphoRS site assignment of the superior, originating from either the target (Figure 1A) or decoy (Figure 1B) set, is returned. For each PSM a different m/z offset is selected randomly from the set of applicable ones.

This way, decoy localizations can only occur by chance and, owing to the preserved S, T and Y patterns and the competition between target and decoy sites, random assignments should be distributed equally between target and decoy sites. The FLR of obtained sites exceeding a certain site probability threshold is finally estimated by dividing the number of decoy sites by the number of target sites (Figure 1C).

Notably, this approach is intrinsically limited to modelling random site assignments, whereas systematic localization errors cannot be estimated. Such systematic mistakes can be addressed for example to the ambiguity between a phosphorylated fragment ion after neutral loss of phosphoric acid (+80 Da – 98 Da = –18 Da) and the non-phosphorylated homolog after neutral loss of water (-18 Da). Thus, the custom version of phosphoRS used in this study was modified not to score such neutral loss fragments ions.

Synthetic phospho-peptide dataset In order to evaluate our approach, we aimed at designing and chemically synthesizing phospho-peptide libraries with known phosphorylation sites. The generated dataset should comprise a large collection of distinct phospho-peptides that are representative for other phospho-proteomic studies in terms of S, T and Y frequencies, number of phosphorylation sites per peptide, observed missed cleavage sites of trypsin and peptide length distribution. Therefore, we selected randomly 188 phospho-peptides from a previous study[23] and used them as seeds for generation of peptide libraries (Figure 2A). For each seed peptide we introduced two sequence positions that were permutated with 18 different amino acids. These permutation sites were positioned preferably in between S, T and Y residues so that they can have most effect on fragmentation behavior at site determining sequence positions, thereby generating a more diverse dataset. Taken together, the resulting libraries comprise a total of 59,688 distinct phospho-peptides, which were subjected to LC-MS/MS analysis applying CID, ETD, HCD and EThcD.

After analysis of in total 745,852 recorded MS/MS scans we could identify 419,668 PSMs at 1% FDR on PSM-level, comprising 92,361 CID-type PSMs, 107,170 ETD-type PSMs, 109,635 HCD-type PSMs and 110,502 EThcD-type PSMs (Figure 2B). The identified PSMs correspond to 41,316 distinct phospho-peptides for CID, 44,782 with ETD, 45,322 with HCD and 46,461 with EThcD (Figure 2C). Notably, these numbers are not intended as a comparison between the applied fragmentation techniques but should rather show the dimension of the acquired phospho-peptide dataset. To the best of our knowledge, this set represents the largest MS/MS collection of chemically synthesized phospho-peptides identified at 1% FDR.

In order to illustrate the diversity of the acquired dataset, we compared identified library seeds to phospho-peptides obtained from titanium dioxide enrichment of a HeLa whole cell lysate[24], with respect to peptide mass and hydrophobicity estimated by the Gravy score (Figure 2C). It could be shown, that identified seed phospho-peptides span over a comparable range as those of HeLa. This indicates that the acquired dataset is of similar diversity as complex biological samples.

Applicable m/z offset values For FLR estimation, decoy sites are generated by offsetting theoretical masses of site determining ions along the m/z axis. The actual values, by which fragments are shifted, need to be assessed specifically due to the following reasons. Using an offset that is equal to for example the neutral loss of water or the addition of a phosphoryl group would result in overestimation of incorrect assignments, because shifted site determining ions will erroneously be match to systematically observed fragment ions. Further, especially in the case of high mass accuracy MS/MS data, the distribution of peaks along the m/z axis is not continuous, owing to the discrete masses of the analytes' elementary building blocks. As a consequence, there exist distances between peaks that are hardly ever observed. Using such an m/z value as offset for site determining ions, would result in underestimation of false positive site assignments.

In order to evaluate, which m/z shifts are applicable and which ones are not, we investigated the average surrounding of correct and incorrect site determining ions by applied the following procedure. For every confidently identified PSM all correct site determining ions are identified and for each of those an artificial spectrum is created by aligning the fragment ion to zero m/z (Figure 3A). Resulting shifted spectra were combined to one collection of peaks for each fragmentation technique separately and signal abundances were summed up using a sliding window corresponding to the fragment ion mass accuracy of the deployed mass analyzer (Figure 3B). For CID and ETD spectra, acquired in the linear ion trap, a window of 1 m/z was used, whereas for HCD and EThcD, both recorded in the orbitrap, 0.04 m/z were specified. The same procedure was applied using incorrect site determining ions (Figure 3C). Resulting spectra are hereinafter referred to as m/z distance spectra.

Appropriate m/z offsets were required to provide an average abundance similar to that of incorrect site determining ions, which are centered around zero in the m/z distance spectrum of incorrect site determining ions. Further, average intensities were requested to be comparable between the m/z distance spectra of correct and incorrect site determining ions. The absolute value of applicable m/z shifts was restricted to 100 and for comparison of average abundance a tolerance of ±25% was used. In total 14,086 m/z offsets at a step width of 0.01 m/z were decided to be applicable for generation of decoy sites using CID and 13,307 for ETD. In case of HCD and EThcD using a step width of 0.001 m/z 13,683 and 8,863 shifts were selected, respectively.

Accuracy of FLR estimation Next, we evaluated how well estimated and actual FLR correspond to each other for the individual fragmentation techniques applied. Based on the knowledge of the correct phosphorylation sites within the synthetic peptides, the actual FLR could be determined for a given site probability threshold. Similarly, FLR estimates could be assessed by considering the quantity of decoy assignments exceeding varying site probability cutoffs. The comparison of estimated and actual FLR is illustrated (Figure 4). Despite minor deviations from identity, obtained results suggest that applying the fragment ion offsetting strategy the FLR can be estimated with reasonable accuracy over the entire range for CID, ETD, HCD and EThcD.

In order to rigorously validate the approach, we extended the comparison of estimated and actual FLR to a dataset comprising low quality MS/MS scans, which are expected to render site localization substantially difficult. Overall identification rates dropped from 56% to 34%, indicating that altered instrumental conditions indeed influenced spectral quality significantly. In total, 305,580 PSMs could be identified at 1% FDR. Still, results suggest that the FLR can be estimated accurately for all activation types applied, when using the fragment ion offsetting method (Figure 5).

Conclusions

The ability to ensure high quality of reported phospho-sites or, generally speaking, PTM assignments becomes increasingly important, since the scope of current proteomics studies has expanded to assess the complexity of whole proteomes. However, to the best of our knowledge, there exists no commonly used approach that would enable accurate estimation of the FLR, although two methods have been suggested recently[34,35]. In this study we presented a strategy, which allows to assess the global FLR for a given dataset by offsetting theoretical m/z values of site determining ions. In this way, STY patterns and respective frequencies are maintained, whereby random assignments are equally likely to occur among target or decoy sites, an essential requirement for target-decoy analysis. This distinguishes our approach from existing ones. Based on the presented results it can be assumed that fragment ion offsetting-based FLR estimation provides reasonable accuracy for all fragmentation techniques under investigation. The development of such a method could have an impact on the scientific community comparable to that of the target-decoy approach for peptide and protein identification. The latter has been used extensively for benchmarking advancements in sample preparation, instrumentation and computational data analysis. Similarly, accurate FLR estimation enabled by offsetting site determining fragment ions could provide the basis for subsequent experimental and computational developments in PTM analysis.

Further, we acquired an extensive set of high confident PSMs from newly generated synthetic phospho-peptide libraries. It was designed to be representative of other phospho-proteomics studies and to contain spectra acquired with a variety of fragmentation techniques, such as CID, ETD, HCD and recently introduced EThcD. The data can serve as valuable resource for versatile applications, including study of phospho-peptides' fragmentation behavior, retention time prediction, alternative identification and characterization algorithms and other downstream computational tasks.

Taken together, our FLR estimation strategy and synthetic phospho-peptide dataset can aid experimental as well as computational improvements and developments, thereby advancing the proteomics toolbox further.

Acknowledgement

This work was funded by Boehringer Ingelheim. The technical support of all other members of the Mechtler group, especially Mathias Madalinski, Susanne Opravil and Johannes Fuchs, is gratefully acknowledged.

FIGURE CAPTIONS

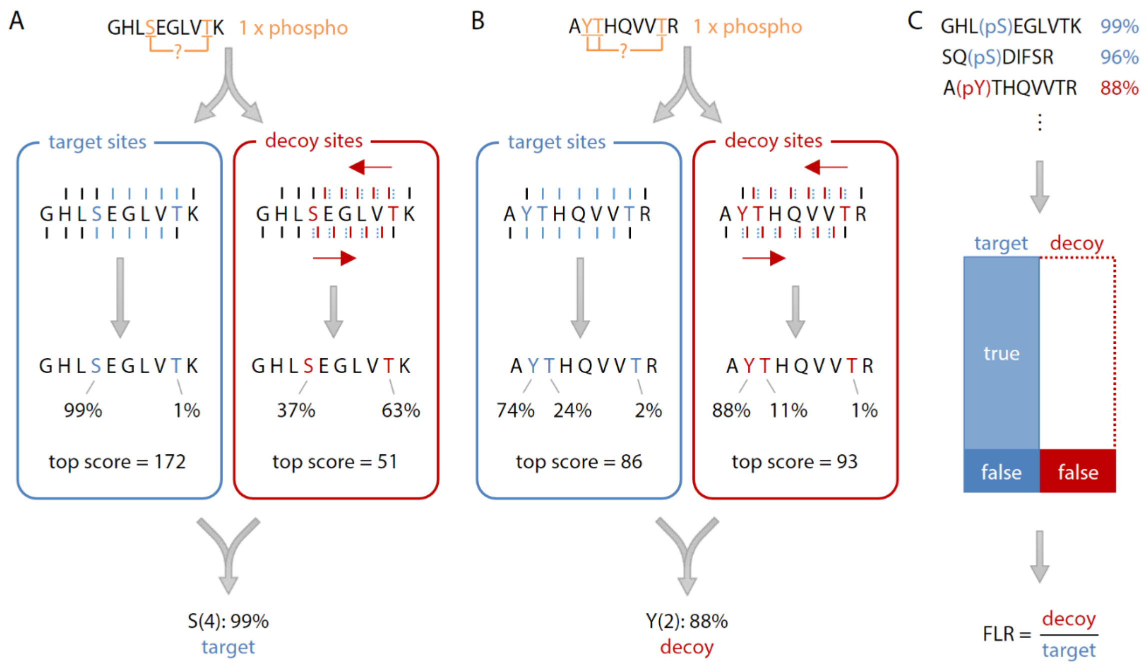


Figure 1. Description of FLR estimation approach. For each PSM, potentially correct (target) sites are analyzed in parallel to respective entrapments sites. Decoy sites are generated by offsetting m/z values of site determining fragment ions, whereby site assignments will only be observed by chance. After comparing target and decoy site assignments, the best result for each PSM, which might be either (A) target or (B) decoy, is chosen on the basis of isoform score and site probability. (C) Assuming that incorrect localizations are equally likely to originate from either the target or entrapment set, the FLR can be estimated by the ratio between number of decoy and target sites above a specified site probability threshold.

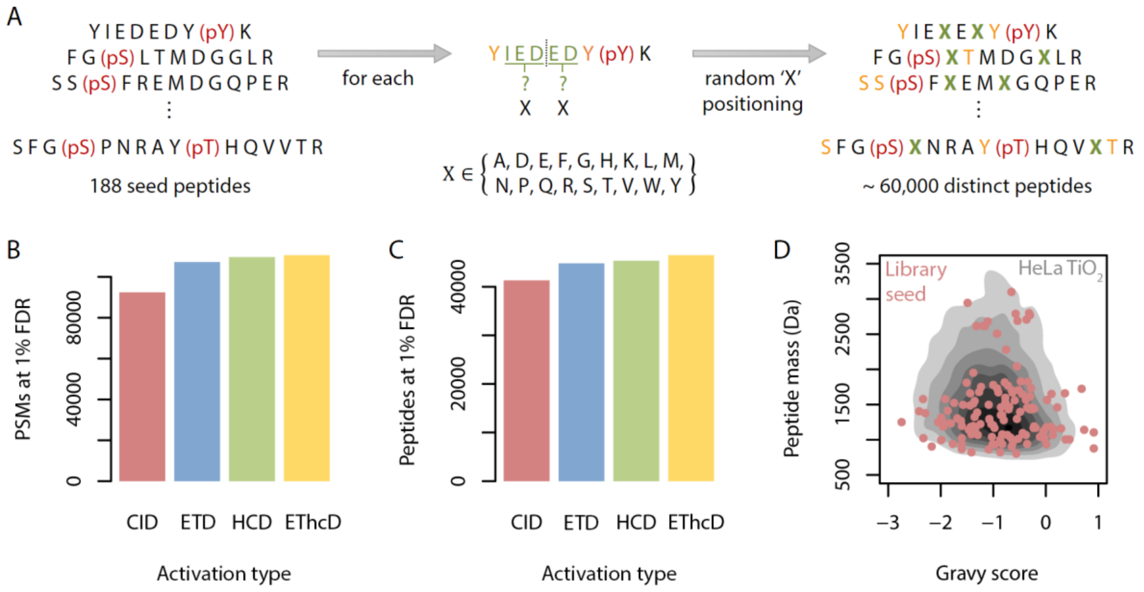


Figure 2. Synthetic phospho-peptide dataset. (A) In total, 188 seed peptides that are representative for other phospho-proteomic studies[23] were selected in order to generate synthetic phospho-peptide libraries. Within every seed peptide two sites, permutated with 18 amino acids each, were randomly positioned, giving rise to 59,688 distinct phospho-peptides. The numbers of identified (B) PSMs and (C) peptides at 1% FDR are illustrated for CID (red), ETD (blue), HCD (green) and EThcD (yellow). (D) Further, the range in terms of peptide mass and hydrophobicity, estimated by Gravy score, which is covered by identified library seeds (red) and HeLa whole cell lysate after titanium dioxide enrichment of phospho-peptides[24] (black) is compared.

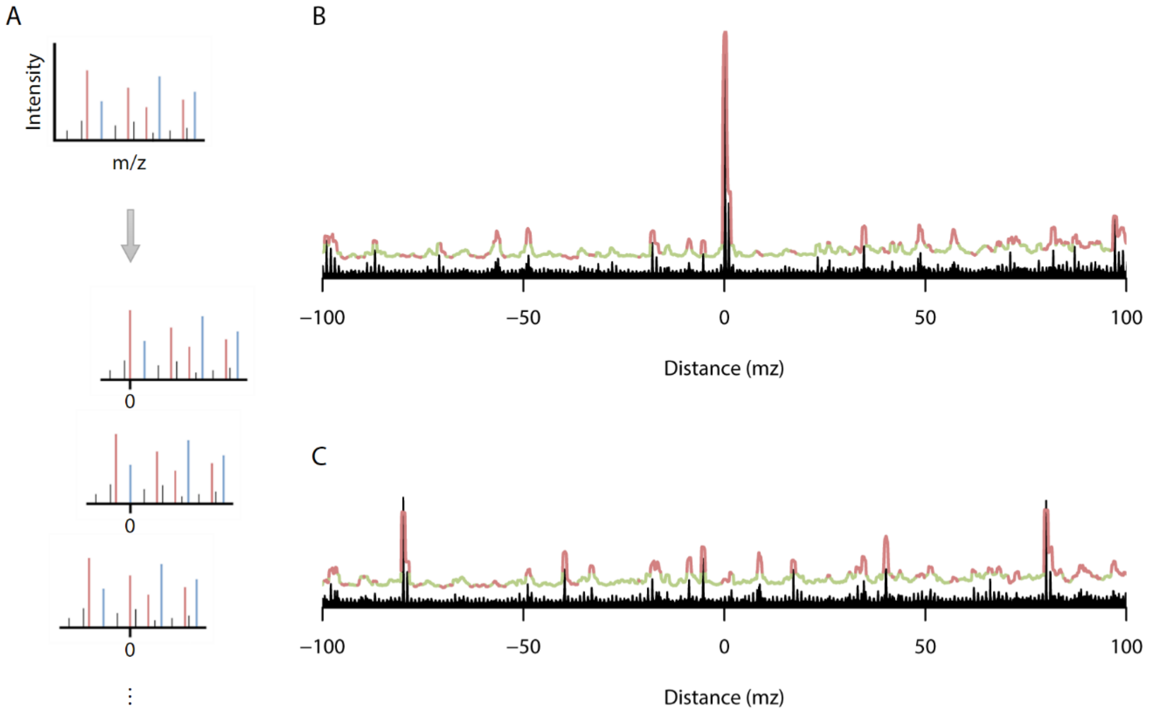


Figure 3. Determination of applicable fragment ion offsets. (A) Site determining ions for each confident PSMs are identified and for each of those a new spectrum is created *in silico* by subtracting the fragment's m/z value from all peaks. Obtained peak lists are combined for each fragmentation technique individually and signal abundances are summed up using a sliding window that corresponds to the respective fragment ion mass tolerance. Performing this procedure with either (B) true or (C) false site determining gives rise to the respective m/z distance spectra, shown here exemplarily for CID with ±0.5 m/z tolerance. Applicable (green) and non-applicable (red) offsets can subsequently be determined by comparing both m/z distance spectra. Additionally, summed up abundances are shown using a 100-fold narrower window (black).

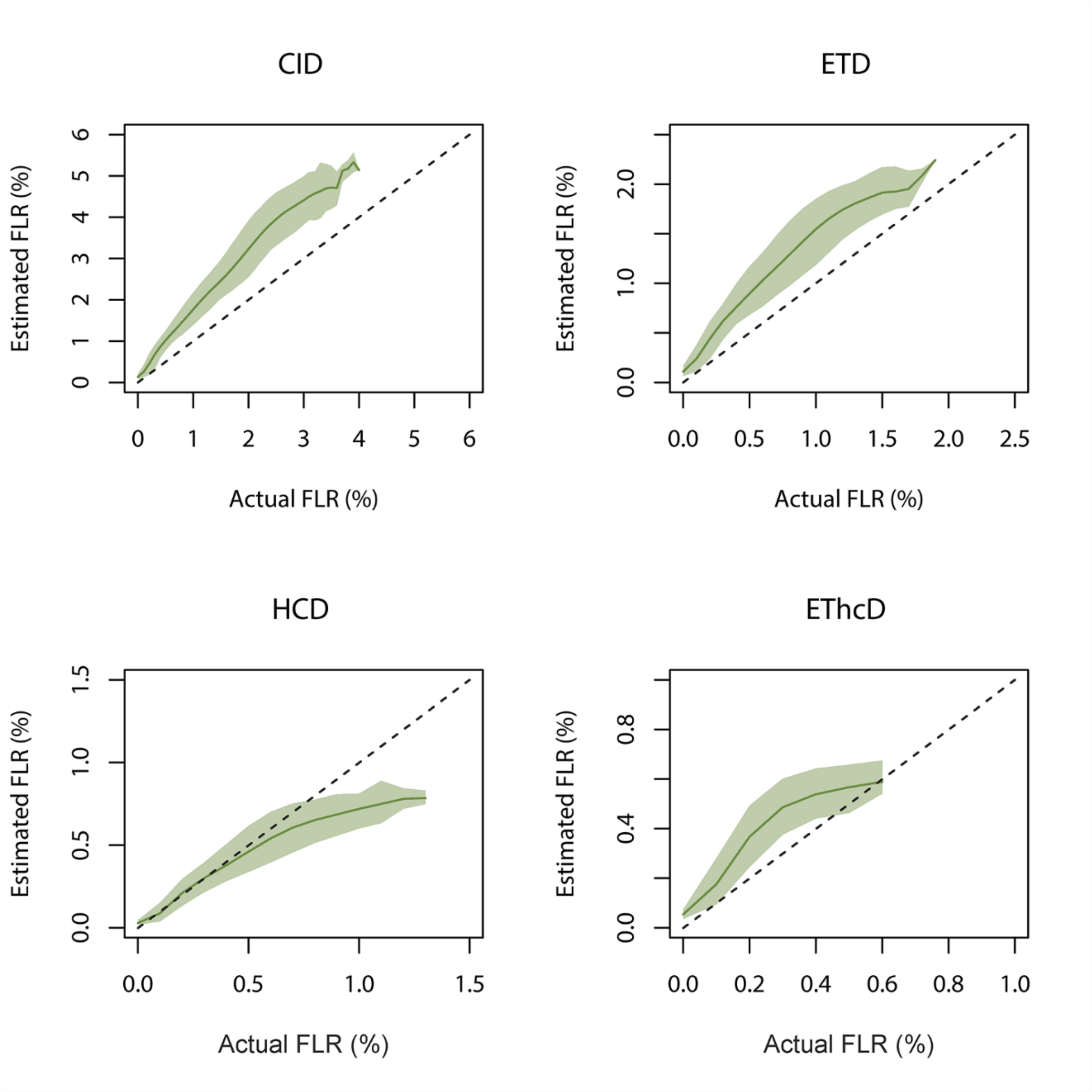


Figure 4. Accuracy of FLR estimation. Actual FLR, determined on the basis of known sites within synthetic phospho-peptides, is compared to the estimated one for (A) CID, (B) ETD, (C) HCD and (D) EThcD. Mean values (dark green) and 95% confidence intervals (light green), both approximated by bootstrapping, are illustrated.

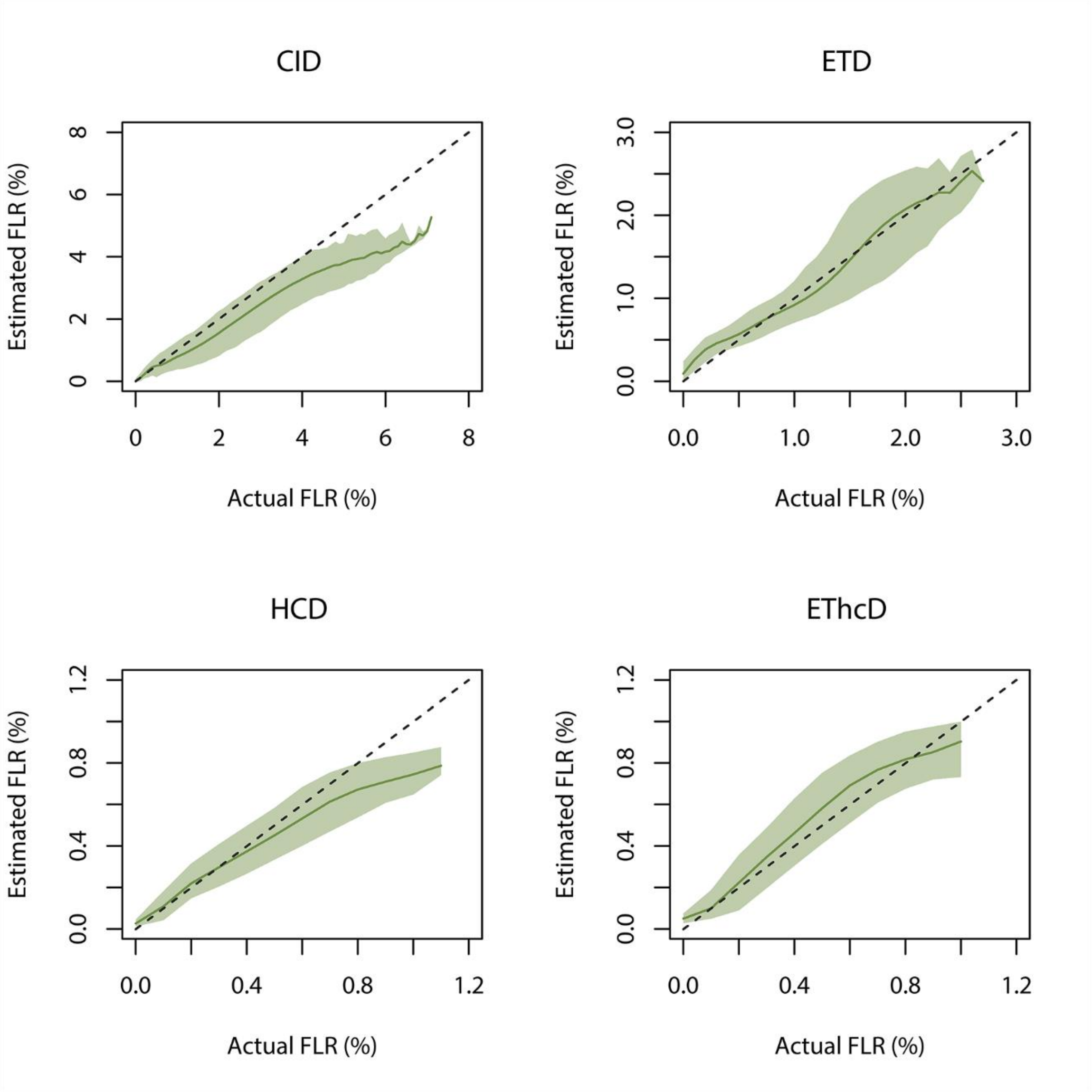


Figure 5. Effect of low spectral quality on accuracy of FLR estimation. After decreasing dramatically AGC target values for linear ion trap and orbitrap acquisition, actual and estimated FLR are compared for the different fragmentation techniques, (A) CID, (B) ETD, (C) HCD and (D) EThcD. Mean values (dark green) and 95% confidence intervals (light green), both approximated by bootstrapping, are illustrated.

REFERENCES

1. Gajadhar, A. S. & White, F. M. System level dynamics of post-translational modifications. *Curr. Opin. Biotechnol.* **28C,** 83–87 (2014).

2. Seet, B. T., Dikic, I., Zhou, M.-M. & Pawson, T. Reading protein modifications with interaction domains. *Nat. Rev. Mol. Cell Biol.* **7,** 473–83 (2006).

3. Yang, X.-J. Multisite protein modification and intramolecular signaling. *Oncogene* **24,** 1653–62 (2005).

4. Cohen, P. Protein kinases--the major drug targets of the twenty-first century? *Nat. Rev. Drug Discov.* **1,** 309–15 (2002).

5. Harsha, H. C. & Pandey, A. Phosphoproteomics in cancer. *Mol. Oncol.* **4,** 482–95 (2010).

6. Han, X., Aslanian, A. & Yates, J. R. Mass spectrometry for proteomics. *Curr. Opin. Chem. Biol.* **12,** 483–90 (2008).

7. Nilsson, T. *et al.* Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat. Methods* **7,** 681–5 (2010).

8. Zhao, Y. & Jensen, O. N. Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques. *Proteomics* **9,** 4632–41 (2009).

9. Eyrich, B., Sickmann, A. & Zahedi, R. P. Catch me if you can: mass spectrometry-based phosphoproteomics and quantification strategies. *Proteomics* **11,** 554–70 (2011).

10. Fenn, J. B., Mann, M., Meng, C. K. A. I., Wong, S. F. & Whitehouse, C. M. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science (80-.).* **246,** 64–71 (1989).

11. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–89 (1994).

12. Stein, S. E. & Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **5,** 859–66 (1994).

13. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–67 (1999).

14. Craig, R. & Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20,** 1466–7 (2004).

15. Shilov, I. V *et al.* The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Mol. Cell. Proteomics* **6,** 1638–55 (2007).

16. Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10,** 1794–805 (2011).

17. Mueller, L. N. *et al.* SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* **7,** 3470–80 (2007).

18. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–72 (2008).

19. Breitwieser, F. P. *et al.* General statistical modeling of data from protein relative expression isobaric tags. *J. Proteome Res.* **10,** 2758–66 (2011).

20. Choi, H. *et al.* SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nat. Methods* **8,** 70–3 (2011).

21. Beausoleil, S. a, Villén, J., Gerber, S. a, Rush, J. & Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **24,** 1285–92 (2006).

22. Olsen, J. V *et al.* Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127,** 635–48 (2006).

23. Savitski, M. M. *et al.* Confident phosphorylation site localization using the Mascot Delta Score. *Mol. Cell. Proteomics* **10,** M110.003830 (2011).

24. Taus, T. *et al.* Universal and confident phosphorylation site localization using phosphoRS. *J. Proteome Res.* **10,** 5354–62 (2011).

25. Ow, S. Y. *et al.* iTRAQ underestimation in simple and complex mixtures: “the good, the bad and the ugly”. *J. Proteome Res.* **8,** 5347–55 (2009).

26. Houel, S. *et al.* Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *J. Proteome Res.* **9,** 4152–60 (2010).

27. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol. Biol.* **604,** 55–71 (2010).

28. Cargile, B. J., Bundy, J. L. & Stephenson, J. L. Potential for false positive identifications from large databases through tandem mass spectrometry. *J. Proteome Res.* **3,** 1082–5 (2004).

29. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 9440–5 (2003).

30. Nesvizhskii, A. I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **4,** 787–97 (2007).

31. Moore, R. E., Young, M. K. & Lee, T. D. Qscore: an algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass Spectrom.* **13,** 378–86 (2002).

32. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J. & Gygi, S. P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2,** 43–50 (2003).

33. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4,** 207–14 (2007).

34. Baker, P. R., Trinidad, J. C. & Chalkley, R. J. Modification site localization scoring integrated into a search engine. *Mol. Cell. Proteomics* **10,** M111.008078 (2011).

35. Fermin, D., Walmsley, S. J., Gingras, A.-C., Choi, H. & Nesvizhskii, A. I. LuciPHOr: algorithm for phosphorylation site localization with false localization rate estimation using modified target-decoy approach. *Mol. Cell. Proteomics* **12,** 3409–19 (2013).

36. Wiesmüller, K. H. *et al.* Peptide and cyclopeptide libraries: Automated synthesis, analysis and receptor binding assays. *Comb. Pept. Non-peptide Libr. - A Handb. Search Lead Struct.* 203–246 (1996).

37. Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J. & Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 9528–33 (2004).

38. Olsen, J. V *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **4,** 709–12 (2007).

39. Frese, C. K. *et al.* Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. *Anal. Chem.* **84,** 9668–73 (2012).

40. Frese, C. K. *et al.* Unambiguous phosphosite localization using electron-transfer/higher-energy collision dissociation (EThcD). *J. Proteome Res.* **12,** 1520–5 (2013).

Chapter 6

Conclusion and Perspective

Protein phosphorylation, a widespread PTM, plays a key role in numerous important cellular processes and can be related to disseases, such as cancer. The ability to characterize phosphorylated peptides and proteins, especially in terms of modification assignments, is of utmost importance for the biological questions addressed. A variety of software solutions have been introduced that enable large scale phospho-site assignment based on large-scale tandem mass spectrometry data. Confident and sensitive site localization requires sufficient sequence coverage of measured fragment ions. In this regard, phosphoRS, one of the most popular phospho-site asignment tools, has been extended for the analysis of EThcD spectra that allow enhanced peptide sequence analysis (see section 5.1). The results suggest that applying EThcD more high confident site localizationes are obtained, when compared to existing fragmentation techniques. Further, it is reasoned that this novel dissociation method might also improve assignment of other PTMs.

Site localization tools are applied broadly to complex biological samples, although they were validated with a limited set of chemically synthesized peptides. Until today, there exists no generally accepted approach for accurate estimation of false localization rates, even though it is essential in order to guarantee quality of reported results, which form the basis for continuative biological studies. A novel approach is presented, which is based on fragment ion offsetting, that enables accurate FLR estimation (see section 5.2). The method has been validated on the basis of more than 700,000 high confident PSMs originating from roughly 60,000 distinct chemically synthesized phospho-peptides. The dataset provides MS/MS spectra acquired with commonly used fragmentation techniques, such as CID, ETD, HCD and EThcD, applying different instrumental settings. Owing to the size and diversity of this PSM collection, it can serve as a valuable resource for a variety of versatile applications, including the analysis of fragmentation patterns, retention time prediction and development of novel identification and characterization algorithms for phospho-peptides.

It can be envisioned that fragment ion offsetting could also be applicable for assignment of other PTMs but phosphorylation, although re-evaluation especially regarding applicable m/z offsets might be required, since they hold to a large extend PTM-specific information. Furthermore, shifting fragment ions that distinguish between individual peptide sequences might also be deployable for peptide and protein identification. It is therefore concluded that advancements and developments presented in this work will enhance the methodological and computational toolbox of proteomics and will form the basis for subsequent improvements of the proteomic technology in general.

References

Literature

1. Alberts, B. *et al. Molecular biology of the cell* 5th ed. (Garland Science, 2008).
2. Nelson, D. L. & Cox, M. M. *Principles of biochemistry* 5th ed. (W. H. Freeman and Company, 2008).
3. Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. Nucleotide sequence of bacteriophage lambda DNA. *J. Mol. Biol.* **162,** 729–773 (1982).
4. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274,** 546–67 (1995).
5. Blattner, F. R. *et al.* The complete genome sequence of escherichia coli K-12. *Science* **277,** 1453–1462 (1997).
6. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
7. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–51 (2001).
8. Church, D. M. *et al.* Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS biology* **7,** e1000112 (2009).
9. Nilsson, T. *et al.* Mass spectrometry in high-throughput proteomics: ready for the big time. *Nature methods* **7,** 681–5 (2010).
10. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246,** 64–71 (1989).

11. Zhao, Y. & Jensen, O. N. Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques. *Proteomics* **9,** 4632–41 (2009).

12. Eyrich, B., Sickmann, A. & Zahedi, R. P. Catch me if you can: mass spectrometry-based phosphoproteomics and quantification strategies. *Proteomics* **11,** 554–70 (2011).

13. Olsen, J. V. *et al.* Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127,** 635–48 (2006).

14. Bodenmiller, B. *et al.* Phosphoproteomic analysis reveals interconnected system-wide responses to perturbations of kinases and phosphatases in yeast. *Science signaling* **3,** rs4 (2010).

15. Huttlin, E. L. *et al.* A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143,** 1174–89 (2010).

16. Beausoleil, S. A., Villen, J., Gerber, S. A., Rush, J. & Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature biotechnology* **24,** 1285–92 (2006).

17. Taus, T. *et al.* Universal and confident phosphorylation site localization using phosphoRS. *Journal of proteome research* **10,** 5354–62 (2011).

18. Cohen, P. The origins of protein phosphorylation. *Nature cell biology* **4,** E127–30 (2002).

19. Cohen, P. Protein kinases–the major drug targets of the twenty-first century? **1,** 309–315 (2002).

20. Harsha, H. C. & Pandey, A. Phosphoproteomics in cancer. *Molecular oncology* **4,** 482–95 (2010).

21. Aivaliotis, M. *et al.* Ser/Thr/Tyr protein phosphorylation in the archaeon Halobacterium salinarum–a representative of the third domain of life. *PloS one* **4,** e4777 (2009).

22. Macek, B. *et al.* The serine/threonine/tyrosine phosphoproteome of the model bacterium Bacillus subtilis. *Molecular & cellular proteomics : MCP* **6,** 697–707 (2007).

23. Macek, B. *et al.* Phosphoproteome analysis of E. coli reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation. *Molecular & cellular proteomics : MCP* **7,** 299–307 (2008).

24. Gnad, F. *et al.* High-accuracy identification and bioinformatic analysis of in vivo protein phosphorylation sites in yeast. *Proteomics* **9,** 4642–52 (2009).

25. Hilger, M., Bonaldi, T., Gnad, F. & Mann, M. Systems-wide analysis of a phosphatase knock-down by quantitative proteomics and phosphoproteomics. *Molecular & cellular proteomics : MCP* **8,** 1908–20 (2009).

26. Pan, C., Gnad, F., Olsen, J. V. & Mann, M. Quantitative phosphoproteome analysis of a mouse liver cell line reveals specificity of phosphatase inhibitors. *Proteomics* **8,** 4534–46 (2008).

27. Matthews, H. R. Protein kinases and phosphatases that act on histidine, lysine, or arginine residues in eukaryotic proteins: a possible regulator of the mitogen-activated protein kinase cascade. *Pharmacol. Ther.* **67,** 323–50 (1995).

28. Fuhrmann, J. *et al.* McsB is a protein arginine kinase that phosphorylates and inhibits the heat-shock regulator CtsR. *Science* **324,** 1323–7 (2009).

29. Besant P. G., P. M. J. Attwood P. V. Focus on phosphoarginine and phospholysine. *Curr. Protein Pept. Sci.* **10,** 536–50 (2009).

30. Cieśla, J., Fraczyk, T. & Rode, W. Phosphorylation of basic amino acid residues in proteins: important but easily missed. *Acta biochimica Polonica* **58,** 137–48 (2011).

31. Schmidt, A. *et al.* Quantitative phosphoproteomics reveals the role of protein arginine phosphorylation in the bacterial stress response. *Molecular & cellular proteomics : MCP* **13,** 537–50 (2014).

32. Attwood, P. V., Piggott, M. J., Zu, X. L. & Besant, P. G. Focus on phosphohistidine. *Amino acids* **32,** 145–56 (2007).

33. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422,** 198–207 (2003).

34. Nesvizhskii, A. I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Molecular & cellular proteomics : MCP* **4,** 1419–40 (2005).

35. Hawkridge, A. M. *et al.* Quantitative mass spectral evidence for the absence of circulating brain natriuretic peptide (BNP-32) in severe human heart failure. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 17442–7 (2005).

36. Uttenweiler-Joseph, S. *et al.* Toward a full characterization of the human 20S proteasome subunits and their isoforms by a combination of proteomic approaches. *Methods in molecular biology (Clifton, N.J.)* **484,** 111–30 (2008).

37. Zabrouskov, V. *et al.* Stepwise deamidation of ribonuclease A at five sites determined by top down mass spectrometry. *Biochemistry* **45,** 987–92 (2006).

38. Siuti, N. & Kelleher, N. L. Decoding protein modifications using top-down mass spectrometry. *Nature methods* **4,** 817–21 (2007).

39. Pesavento, J. J., Mizzen, C. A. & Kelleher, N. L. Quantitative analysis of modified proteins and their positional isomers by tandem mass spectrometry: human histone H4. *Analytical chemistry* **78,** 4271–80 (2006).

40. Du, Y., Parks, B. A., Sohn, S., Kwast, K. E. & Kelleher, N. L. Top-down approaches for measuring expression ratios of intact yeast proteins using Fourier transform mass spectrometry. *Analytical chemistry* **78,** 686–94 (2006).

41. Waanders, L. F., Hanke, S. & Mann, M. Top-down quantitation and characterization of SILAC-labeled proteins. *Journal of the American Society for Mass Spectrometry* **18,** 2058–2064 (2007).

42. Macek, B., Mann, M. & Olsen, J. V. Global and site-specific quantitative phosphoproteomics: principles and applications. *Annual review of pharmacology and toxicology* **49,** 199–221 (2009).

43. Andersson, L. & Porath, J. Isolation of phosphoproteins by immobilized metal (Fe3+) affinity chromatography. *Analytical biochemistry* **154,** 250–4 (1986).

44. Posewitz, M. C. & Tempst, P. Immobilized Gallium (III) Affinity Chromatography of Phosphopeptides peptides , as a front end to mass spectrometric analysis , the use of an immobilized metal affinity chromatography. *Analytical chemistry* **71,** 2883–2892 (1999).

45. Ikeguchi, Y. & Nakamura, H. Determination of Organic Phosphates by Column-Switching High Performance Anion-Exchange Chromatography Using On-Line Preconcentration on Titania. *Analytical Sciences* **13,** 479–483 (1997).

46. Pinkse, M. W. H., Uitto, P. M., Hilhorst, M. J., Ooms, B. & Heck, A. J. R. Selective isolation at the femtomole level of phosphopeptides from proteolytic digests using 2D-NanoLC-ESI-MS/MS and titanium oxide precolumns. *Analytical chemistry* **76,** 3935–43 (2004).

47. Mazanek, M. *et al.* A new acid mix enhances phosphopeptide enrichment on titanium- and zirconium dioxide for mapping of phosphorylation sites on protein complexes. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences* **878,** 515–24 (2010).

48. Beausoleil, S. A. *et al.* Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 12130–5 (2004).

49. Gruhler, A. *et al.* Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Molecular & cellular proteomics : MCP* **4,** 310–27 (2005).

50. Villén, J., Beausoleil, S. A., Gerber, S. A. & Gygi, S. P. Large-scale phosphorylation analysis of mouse liver. *Proceedings of the National Academy of Sciences of the United States of America* **104,** 1488–93 (2007).

51. De Graaf, E. L., Giansanti, P., Altelaar, A. F. M. & Heck, A. J. R. Single step enrichment by Ti4+-IMAC and label free quantitation enables in-depth monitoring of phosphorylation dynamics with high reproducibility and temporal resolution. *Molecular & cellular proteomics : MCP,* 1–28 (2014).

52. Alpert, A. J. Electrostatic repulsion hydrophilic interaction chromatography for isocratic separation of charged solutes and selective isolation of phosphopeptides. *Analytical chemistry* **80,** 62–76 (2008).

53. Alpert, A. J. Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds. *Journal of chromatography* **499,** 177–96 (1990).

54. Anderson, N. L. The Human Plasma Proteome: History, Character, and Diagnostic Prospects. *Molecular & Cellular Proteomics* **1,** 845–867 (2002).

55. Giddings, J. C. Two-dimensional separations: concept and promise. *Analytical chemistry* **56,** 1258A–1270A (1984).

56. Link, A. J. *et al.* Direct analysis of protein complexes using mass spectrometry. *Nature biotechnology* **17,** 676–82 (1999).

57. Wolters, D. A., Washburn, M. P. & Yates, J. R. An automated multidimensional protein identification technology for shotgun proteomics. *Analytical chemistry* **73,** 5683–90 (2001).

58. Washburn, M. P., Wolters, D. & Yates, J. R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature biotechnology* **19,** 242–7 (2001).

59. Yates, J. R., Ruse, C. I. & Nakorchevsky, A. Proteomics by mass spectrometry: approaches, advances, and applications. *Annual review of biomedical engineering* **11,** 49–79 (2009).

60. Michalski, A., Cox, J. & Mann, M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *Journal of proteome research* **10,** 1785–93 (2011).

61. Taus, T. *phosphoRS: A novel probability-based algorithm for sensitive protein phosphorylation-site localization from high-throughput LC-MS/MS data* Bachelor thesis (Vienna University of Technology, 2011).

62. Karas, M. & Hillenkamp, F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical chemistry* **60,** 2299–301 (1988).

63. Chalkley, R. Instrumentation for LC-MS/MS in proteomics. *Methods in molecular biology (Clifton, N.J.)* **658,** 47–60 (2010).

64. Paul, W. & Steinwedel, H. Ein neues Massenspektrometer ohne Magnetfeld. *RZeitschrift für Naturforschung A* **8,** 448–450 (1953).

65. Schwartz, J. C., Senko, M. W. & Syka, J. E. P. A two-dimensional quadrupole ion trap mass spectrometer. *Journal of the American Society for Mass Spectrometry* **13,** 659–69 (2002).

66. Douglas, D. J., Frank, A. J. & Mao, D. Linear ion traps in mass spectrometry. *Mass spectrometry reviews* **24,** 1–29 (2005).

67. Comisarow, M. B. & Marshall, A. G. Fourier transform ion cyclotron resonance spectroscopy. *Chemical Physical Letters* **25,** 282–283 (1974).

68. Kingdon, K. H. A method for the neutralization of electron space charge by positive ionization at very low gas pressures. *Physical Review* **21,** 408 (1923).

69. Makarov, A. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Analytical chemistry* **72,** 1156–62 (2000).

70. Hunt, D. F., Buko, A. M., Ballard, J. M., Shabanowitz, J. & Giordani, A. B. Sequence analysis of polypeptides by collision activated dissociation on a triple quadrupole mass spectrometer. *Biomedical mass spectrometry* **8,** 397–408 (1981).

71. Morris, H. R. *et al.* High sensitivity collisionally-activated decomposition tandem mass spectrometry on a novel quadrupole/orthogonal-acceleration time-of-flight mass spectrometer. *Rapid communications in mass spectrometry : RCM* **10,** 889–96 (1996).

72. Makarov, A. *et al.* Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Analytical chemistry* **78,** 2113–20 (2006).

73. Roepstorff, P. & Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical mass spectrometry* **11,** 601 (1984).

74. Wysocki, V. H., Tsaprailis, G., Smith, L. L. & Breci, L. A. Mobile and localized protons: a framework for understanding peptide dissociation. *Journal of mass spectrometry : JMS* **35,** 1399–406 (2000).

75. Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J. & Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 9528–33 (2004).

76. Paizs, B. & Suhai, S. Fragmentation pathways of protonated peptides. *Mass spectrometry reviews* **24,** 508–48 (2005).

77. Schroeder, M. J., Shabanowitz, J., Schwartz, J. C., Hunt, D. F. & Coon, J. J. A neutral loss activation method for improved phosphopeptide sequence analysis by quadrupole ion trap mass spectrometry. *Analytical chemistry* **76,** 3590–8 (2004).

78. Olsen, J. V. *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nature methods* **4,** 709–12 (2007).

79. McAlister, G. C., Phanstiel, D. H., Brumbaugh, J., Westphall, M. S. & Coon, J. J. Higher-energy collision-activated dissociation without a dedicated collision cell. *Molecular & cellular proteomics : MCP* **10,** O111.009456 (2011).

80. Michalski, A., Neuhauser, N., Cox, J. & Mann, M. A systematic investigation into the nature of tryptic HCD spectra. *Journal of proteome research* **11,** 5479–91 (2012).

81. Zubarev, R. A. *et al.* Electron capture dissociation for structural characterization of multiply charged protein cations. *Analytical chemistry* **72,** 563–73 (2000).

82. Swaney, D. L. *et al.* Supplemental activation method for high-efficiency electron-transfer dissociation of doubly protonated peptide precursors. *Analytical chemistry* **79,** 477–85 (2007).

83. Frese, C. K. *et al.* Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. *Analytical chemistry* **84,** 9668–73 (2012).

84. Frese, C. K. *et al.* Unambiguous phosphosite localization using electron-transfer/higher-energy collision dissociation (EThcD). *Journal of proteome research* **12,** 1520–5 (2013).

85. Jones, A. R. & Hubbard, S. J. An introduction to proteome bioinformatics. *Methods in molecular biology (Clifton, N.J.)* **604,** 1–5 (2010).

86. Hughes, C., Ma, B. & Lajoie, G. A. De novo sequencing methods in proteomics. *Methods in molecular biology (Clifton, N.J.)* **604,** 105–21 (2010).

87. Ma, B. *et al.* PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry : RCM* **17,** 2337–42 (2003).

88. Hubbard, S. J. Computational approaches to peptide identification via tandem MS. *Methods in molecular biology (Clifton, N.J.)* **604,** 23–42 (2010).

89. Kapp, E. & Schütz, F. Overview of tandem mass spectrometry (MS/MS) database search algorithms. *Current protocols in protein science* **25,** Unit25.2 (2007).

90. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **5,** 976–89 (1994).

91. Craig, R. & Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics (Oxford, England)* **20,** 1466–7 (2004).

92. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–67 (1999).

93. Colinge, J., Masselot, A., Giron, M., Dessingy, T. & Magnin, J. OLAV: towards high-throughput tandem mass spectrometry data identification. *Proteomics* **3,** 1454–63 (2003).

94. Geer, L. Y. *et al.* Open mass spectrometry search algorithm. *Journal of proteome research* **3,** 958–64 (2004).

95. Shilov, I. V. *et al.* The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Molecular & cellular proteomics : MCP* **6,** 1638–55 (2007).

96. Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of proteome research* **10,** 1794–805 (2011).

97. Dorfer, V. *et al.* MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *Journal of proteome research* (2014).

98. Stein, S. E. & Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry* **5,** 859–66 (1994).

99. Yates, J. R., Morgan, S. F., Gatlin, C. L., Griffin, P. R. & Eng, J. K. Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. *Analytical chemistry* **70,** 3557–65 (1998).

100. Craig, R., Cortens, J. C., Fenyo, D. & Beavis, R. C. Using annotated peptide mass spectrum libraries for protein identification. *Journal of proteome research* **5,** 1843–9 (2006).

101. Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S. & MacCoss, M. J. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Analytical chemistry* **78,** 5678–84 (2006).

102. Lam, H. *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7,** 655–67 (2007).

103. Savitski, M. M. *et al.* Confident phosphorylation site localization using the Mascot Delta Score. *Molecular & cellular proteomics : MCP* **10,** M110.003830 (2011).

104. Houel, S. *et al.* Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *Journal of proteome research* **9,** 4152–60 (2010).

105. Ow, S. Y. *et al.* iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly". *Journal of proteome research* **8,** 5347–55 (2009).

106. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods in molecular biology (Clifton, N.J.)* **604,** 55–71 (2010).

107. Cargile, B. J., Bundy, J. L. & Stephenson, J. L. Potential for false positive identifications from large databases through tandem mass spectrometry. *Journal of proteome research* **3,** 1082–5 (2004).

108. Nesvizhskii, A. I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature methods* **4,** 787–97 (2007).

109. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 9440–5 (2003).

110. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical chemistry* **74,** 5383–92 (2002).

111. Moore, R. E., Young, M. K. & Lee, T. D. Qscore: an algorithm for evaluating SEQUEST database search results. *Journal of the American Society for Mass Spectrometry* **13,** 378–86 (2002).

112. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J. & Gygi, S. P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *Journal of proteome research* **2,** 43–50 (2003).

113. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods* **4,** 207–14 (2007).

114. Lam, H., Deutsch, E. W. & Aebersold, R. Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. *Journal of proteome research* **9,** 605–10 (2010).

115. Gupta, N., Bandeira, N., Keich, U. & Pevzner, P. A. Target-decoy approach and false discovery rate: when things may go wrong. *Journal of the American Society for Mass Spectrometry* **22,** 1111–20 (2011).

116. Reiter, L. *et al.* Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Molecular & cellular proteomics : MCP* **8,** 2405–17 (2009).

117. Baker, P. R., Trinidad, J. C. & Chalkley, R. J. Modification site localization scoring integrated into a search engine. *Molecular & cellular proteomics : MCP* **10,** M111.008078 (2011).

118. Fermin, D., Walmsley, S. J., Gingras, A.-C., Choi, H. & Nesvizhskii, A. I. LuciPHOr: algorithm for phosphorylation site localization with false localization rate estimation using modified target-decoy approach. *Molecular & cellular proteomics : MCP* **12,** 3409–19 (2013).

Messbox zur Druckkontrolle

— Druckgröße kontrollieren! —

Breite = 100 mm
Höhe = 50 mm

— Diese Seite nach dem Druck entfernen! —