Veronika Suni

Computational Methods and Tools for Protein Phosphorylation Analysis

Turku Centre *for* Computer Science

TUCS Dissertations
No 237, April 2019

Computational Methods and Tools for Protein Phosphorylation Analysis

Veronika Suni

To be presented, with the permission of the Faculty of Science and Engineering of the University of Turku, for public criticism in the Auditorium PharmaCity on April 5th, 2019 at 12:00.

University of Turku
Department of Future Technologies
Vesilinnantie 5, 20500 Turun yliopisto
Turku, Finland
2019

Supervised by

Professor Garry Corthals
Faculty of Science
Van ’t Hoff Institute for Molecular Sciences
Amsterdam, The Netherlands

Adj. Professor Laura Elo
Turku Centre for Biotechnology
University of Turku and Åbo Akademi University
Turku, Finland

Professor Tapio Salakoski
Faculty of Science and Engineering
University of Turku
Turku, Finland

Reviewed by

Assoc. Professor Fredrik Levander
Department of Immunotechnology
Lund University
Lund, Sweden

Assoc. Professor Liam McDonnell
Laboratory of Proteomics and Metabolomics
Fondazione Pisana per la Scienza
Pisa, Italy

Opponent

Professor Lukas Käll
Science for Life Laboratory
KTH Royal Institute of Technology
Stockholm, Sweden

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

ISBN 978-952-12-3795-9
ISSN 1239-1883
Painosalama Oy – Turku 2019

to Martti

Abstract

Signaling pathways represent a central regulatory mechanism of biological systems where a key event in their correct functioning is the reversible phosphorylation of proteins. Protein phosphorylation affects at least one-third of all proteins and is the most widely studied posttranslational modification. Phosphorylation analysis is still perceived, in general, as difficult or cumbersome and not readily attempted by many, despite the high value of such information. Specifically, determining the exact location of a phosphorylation site is currently considered a major hurdle, thus reliable approaches are necessary for the detection and localization of protein phosphorylation. The goal of this PhD thesis was to develop computation methods and tools for mass spectrometry-based protein phosphorylation analysis, particularly validation of phosphorylation sites. In the first two studies, we developed methods for improved identification of phosphorylation sites in MALDI-MS. In the first study it was achieved through the automatic combination of spectra from multiple matrices, while in the second study, an optimized protocol for sample loading and washing conditions was suggested. In the third study, we proposed and evaluated the hypothesis that in ESI-MS, tandem CID and HCD spectra of phosphopeptides can be accurately predicted and used in spectral library searching. This novel strategy for phosphosite validation and identification offered accuracy that outperformed the other currently existing popular methods and proved applicable to complex biological samples. And finally, we significantly improved the performance of our command-line prototype tool, added graphical user interface, and options for customizable simulation parameters and filtering of selected spectra, peptides or proteins. The new software, SimPhospho, is open-source and can be easily integrated in a phosphoproteomics data analysis workflow. Together, these bioinformatics methods and tools enable confident phosphosite assignment and improve reliable phosphoproteome identification and reporting.

Keywords: mass spectrometry, protein phosphorylation, spectral library, phosphosite validation, simulation of spectra

Tiivistelmä (Finnish summary)

Veronika Suni

Laskennalliset menetelmät ja työkalut proteiinien fosforylaatioanalyysiin

Signalointireitit ovat tärkeä biologisten järjestelmien säätelymekanismi, jossa keskeinen tapahtuma niiden oikeassa toiminnassa on proteiinien palautuva fosforylaatio. Proteiinien fosforylaatio vaikuttaa ainakin kolmasosaan kaikista proteiineista, ja on eräs laajimmin tutkittuja translaation jälkeisiä modifikaatioita. Fosforylaatioanalyysiä pidetään edelleen yleisesti ottaen vaikeana ja työläänä toteuttaa, eikä sitä siksi suoriteta tyypillisesti sen tarjoaman informaation hyödyllisyydestä huolimatta. Erityisesti eksaktin fosforylaatiosijainnin määrittäminen ja havainnointi nähdään keskeisenä haasteena, jonka ratkaisemiseksi tarvitaan luotettavia lähestymistapoja. Tämän väitöskirjan tavoite on kehittää ohjelmistollisia työkaluja massaspektrometriaperusteiseen fosforylaatioanalyysiin, erityisesti fosforylaatiosijainti validointiin. Kahdessa ensimmäisessä tutkimuksessa kehitimme metodeja fosforylaatiosijaintien identifioimiseen MALDI-tyyppisessä massaspektrometriadatassa. Näistä ensimmäisessä hyödynsimme useista matriiseista peräisin olevien spektrien automaatttista kombinaatiota, kun taas jälkimmäisessä esittelimme optimoidun prosessin näytteiden valmisteluun ja puhdistusolosuhteisiin. Kolmannessa tutkimuksessa esittelimme ja arvioimme hypoteesia, jossa fosfopeptidien ESI-, tandem CID- ja HCD –tyyppisten massaspektrometrien tuottamaa spektridataa voidaan ennustaa tarkasti ja näin hyödyntää spektrikirjastopohjaisessa haussa. Tämä uusi fosforylaatiosijaintien validoimis- ja identifikaatiostrategia tarjoaa nykyisin yleisesti käytössä olevia metodeja tarkempia tuloksia ja osoittautui soveltuvaksi kompleksisten biologisten näytteiden analysointiin. Lopuksi, toteutimme merkittäviä käytettävyys- ja tehokkuusparannuksia aiemmin esittelemäämme komentorivipohjaiseen prototyyppisovellukseen optimoimalla lähdekoodia ja lisäämällä siihen graafisen käyttöliittymän sekä joukon asetuksia simulaatioparametrien muokkaamiseen ja valittujen spektrien, peptidien tai proteiinien suodatukseen. Uusi sovellus nimeltä SimPhospho on lähdekoodiltaan avoin ja helposti integroitavissa fosfoproteomiikan data-analyysiprosesseihin. Yhdessä nämä bioinformatiikan metodit ja sovellustyökalut mahdollistavat luotettavan fosforylaatiosijaintien määrityksen ja parantavat fosfoproteomien identifikaation luotettavuutta.

Acknowledgements

First of all, I would like to express my deepest gratitude to my supervisors, Professor Garry Corthals, Professor Tapio Salakoski and Dr. Laura Elo for their support. I would like to thank Garry, my primary supervisor, for his fundamental role in my development as a researcher, for sharing his ideas and for his inspiring guidance.

I am grateful to Dr. Liam McDonnell and Dr. Fredrik Levander for reviewing my dissertation and giving valuable feedback and comments. I would like to thank Professor Lukas Käll for being my opponent at the public defence.

I wish to thank all the collaborators and co-authors that have made this research possible: Professor Rudolph Aebersold, Dr. Susumu Imanishi, Dr. Päivi Koskinen, Dr. Petri Kouvonen, Dr. Alessio Maiolica, Dr. Eeva Rainio, Tomi Suomi, and Tomoya Tsubosaka. I appreciate very much your important contribution in the projects for this thesis. I especially would like to thank Susumu for his ceaseless encouragement and for sharing his invaluable expertise.

Throughout the years it has been a pleasure working alongside so many great people and excellent researchers in Turku Centre for Biotechnology. I would like to thank my colleagues, past and present members of Proteomics groups for the enjoyable time, especially Anne, Arttu, Pekka, Anni, Dorota, Susanne, Fanni, Olli, Tanya and Aschwin. I would also like to thank all my colleagues from Bioinformatics group, especially Asta, Bishwa, Mehrad and Esko for interesting discussions and helping me so generously. A warm thank you to Sohrab, Johannes, Satu, Maria, Mikko, Ye, Deep, Thomas, Xu, Arfa, Narges, and Vladislav for being wonderful workmates.

Finally, I am very grateful to my friends and family for their support, love and continuous encouragement.

I would like to acknowledge the financial support provided by Turku Centre for Computer Science (TUCS), University of Turku, Academy of Finland, Turun Yliopistosäätiö, and Lounaissuomalaiset Syöpärahastot.

Turku, 15th of February *Veronika Suni*

List of original publications

P1. Kouvonen P, Rainio EM, Suni V, Koskinen P, Corthals GL. *Data combination from multiple matrix-assisted laser desorption/ionization (MALDI) matrices: opportunities and limitations for MALDI analysis.* Rapid Commun Mass Spectrom. 2010 Dec 15;24(23):3493-5.

P2. Kouvonen P, Rainio EM, Suni V, Koskinen P, Corthals GL. *Enrichment and sequencing of phosphopeptides on indium tin oxide coated glass slides.* Mol Biosyst. 2011 Jun;7(6):1828-37.

P3. Suni V*, Imanishi SY*, Maiolica A, Aebersold R, Corthals GL. *Confident site localization using a simulated phosphopeptide spectral library.* J Proteome Res. 2015 May 1;14(5):2348-59. (* equal contribution)

P4. Suni V, Suomi T, Tsubosaka T, Imanishi SY, Elo LL, Corthals GL. *SimPhospho: a software tool enabling confident phosphosite assignment.* Bioinformatics. 2018 Aug 1;34(15):2690-2692.

Contents

INTRODUCTION

Background and motivation

Signaling pathways represent a central regulatory mechanism of biological systems where a key event in their correct functioning is the reversible phosphorylation of proteins. Protein phosphorylation affects at least one-third of all proteins and is the most widely studied posttranslational modification (Hunter, 1995) because of its important role in cellular behaviour. Protein phosphorylation can be examined in several ways, but the most versatile non-radioactive methods that currently exist use mass spectrometry.

While protein identification by mass spectrometry is a standard technique, there are still challenges related to protein phosphorylation analysis, which can be roughly categorized as either biological or technical. The biological challenges that make phosphorylation difficult to detect are low stoichiometry, heterogeneity, and low abundance of protein phosphorylation. State of the art mass spectrometry, separation and phospho-enrichment techniques have been effectively used to tackle these problems. The technical challenges, which are addressed using bioinformatics tools, concern spectra interpretation, in particular finding the exact position in a protein sequence where a phosphorylation event is taking place. Importantly, unambiguous identification of phosphorylation sites can help avoid costly and tedious downstream biological characterization experiments misguided by incorrect site assignments (Gunawardena *et al.*, 2011; Vaga *et al.*, 2014).

Bioinformatics has now become a critical component of any successful proteomics experiment and it is also usually the most time-consuming step. The vitally important role of bioinformatics in proteome study has been recognized, which resulted in the considerable advances in the field of proteomics informatics

in the past decade, driven mainly by free and open-source software tools (Deutsch, Lam and Aebersold, 2008). The rapidly emerging field of bioinformatics has introduced the means to handle heterogeneous data sets and improved the knowledge discovery process (Blueggel, Chamrad and Meyer, 2004). As a result, mass spectrometry combined with enrichment strategies for phosphorylated proteins and advanced data processing has been employed to identify thousands of high-confidence phosphorylation sites (Giansanti *et al.*, 2015; Batth *et al.*, 2018). However, unambiguous identification of the phosphorylated residues is still considered a major hurdle.

At the time when this PhD project began, an interesting study was conducted by Proteome Informatics Research Group (Rudnick *et al.*, 2010). Their aim was to evaluate the consistency of reporting of phosphopeptide identifications and phopshosite localization across 22 laboratories in both academic and industry proteomics communities. Participants were given a common dataset and were allowed to use the bioinformatics methods and tools of their choice. The agreement between the groups on phosphosite localization was as low as ~38%, and what made the results more alarming was that the use of the same data analysis tools led to different, even conflicting results. Such inconsistency in phospho-proteomics informatics suggested that false-positive and false-negative rate in high-throughput phosphoproteomic data sets could be substantial and that the best practice is still to be defined (Lee, Jones and Hubbard, 2015).

There have been significant developments in the past years to improve reliable phosphoproteome identification. The most important ones include (1) the latest data format for identification results that support scores associated with localization of modifications, including phosphorylation, which should accelerate the development of phosphoproteomics data analysis pipelines (Vizcaíno *et al.*, 2017); (2) additional information about ambiguity of phosphosites in the phosphorylation databases (Gnad, Gunawardena and Mann, 2011; Hornbeck *et al.*, 2015); as well as (3) the appearance of repositories that include experimentally observed and validated mass spectra (Hummel *et al.*, 2007; Bodenmiller *et al.*, 2008; Farrah *et al.*, 2013). Collections of high quality tandem mass spectra, also called spectral libraries, enable a new promising validation method for phosphorylation analysis through spectral library matching which offers more precise similarity scores. However, the difficulty of generating enough data for such a library to be complete and useful for phosphosite validation appeared to be a major limitation to explore the potential of spectral library matching approach.

The aim of the work described in this PhD thesis was to develop new computational methods and tools to assist the analysis of protein phosphorylation, in particular identification and validation of phosphorylation sites.

Research questions

The thesis attempts to answer the following research questions:

1. *Can phosphorylation sites be successfully determined in MALDI-MS experiments through combination of spectra from multiple matrices?*
2. *What are the optimal sample loading and washing conditions for ITO-coated glass slides for phosphopeptide purification in MALDI-MS?*
3. *Can the spectra of phosphopeptides be accurately predicted using spectra of dephosphorylated peptides in ESI-MS experiments? Is the resemblance strong enough to be used for predictive purposes? Can this resemblance be used to confidently identify the correct phosphopeptide isoforms in CID and HCD tandem mass spectra?*
4. *What are the optimal parameters for generating simulated spectra of phosphopeptides?*

Hence, the first two research questions deal with methods for phosphopeptide analysis using MALDI-MS, while the other two research questions focus on ESI-MS. The goal of the first research question was to develop an approach for site-specific phosphorylation analysis that uses a panel of matrix-assisted laser desorption ionization (MALDI) matrices, and to compare their performance to each other and the commonly used workflows that employ electrospray ionization (ESI) (**P1**). The second research question concerns testing various sample loading and washing conditions for the indium tin oxide (ITO) coating glass slides used in MALDI-MS to maximize the phosphopeptide purification effect (**P2**). The third question is about the development of a software based phosphosite validation method for ESI-MS experiments using predicted spectra, and its integration into a phosphoproteomics experimental and data analysis workflow. It also includes benchmarking against several popular methodologies (**P3**). The fourth question concerns the optimization of the parameters for simulation of spectra as well as enhancing the functionality of the software (**P4**).

Organization of the Thesis

This thesis is structured as follows. In the first three chapters, I present a literature review introducing topics relevant to this thesis work. **Chapter 1** focuses on current mass-spectrometry based methods for protein phosphorylation analysis. After the introduction to protein phosphorylation and phosphoproteomics, there is a brief overview of the essential sample separation techniques and finally, types of instruments that were used in this dissertation. In **Chapter 2**, the main three categories of computational methods for spectral interpretation and peptide identification are presented. First, the manual spectral interpretation, known as *de novo* sequencing, is discussed. Next, the principles of protein sequence database and spectral library search engines are described. **Chapter 3** discusses localization of phosphorylation sites and their validation. Methods based on delta scores are presented, as well as probability-based tools. The chapter is concluded with a review of studies describing the most recent strategies for phosphosite validation.

In **Chapter 4**, the summary of this Thesis work is presented, each paper described in a separate section. The first paper (**P1**), that discusses a novel MALDI-MS method, describes the implementation of spectral data combination for comparison of different matrices in site-specific phosphorylation studies. The second paper (**P2**) systematically evaluates characteristics of indium tin oxide coated slides for phosphopeptide analysis using MALDI-MS. The third paper (**P3**) describes the method that was developed for ESI-MS, and its comparative analysis with popular phosphosite validation methods using real-life and synthetic datasets. The fourth article (**P4**) presents an improved version of the prototype program, SimPhospho, developed in the previous study P3. Improvements include a graphical user interface, better performance, additional features and enhanced predictive capability.

1 MASS-SPECTROMETRY BASED METHODS FOR PROTEIN PHOSPHORYLATION ANALYSIS

1.1 Protein phosphorylation

Proteins are a fundamental component of all living cells executing the majority of their functions. Proteins are large molecules made up of twenty different smaller molecules, amino acids. Each type of protein has a unique sequence of amino acids. The collection of proteins produced by organisms is termed the proteome and it is studied by proteomics in analogy with the complement of genes, genome, studied by genomics. In 2014 the first two drafts of the human proteome were presented in *Nature* journal (Kim *et al.*, 2014; Wilhelm *et al.*, 2014), showing protein evidence for 18,000 genes. Protein diversity is further increased through posttranslational modifications, which are chemical modifications of a protein after its translation from a gene.

Among the hundreds of types of protein modifications (Krishna and Wold, 1993), protein phosphorylation is among a few that have been proven to be of regulatory importance in biological processes (Hunter, 1995). The regulation of protein phosphorylation is so extensive that the majority of intracellular proteins are thought to be phosphorylated at any given time, many with more than one phosphate. The aims of protein phosphorylation studies include obtaining phosphorylation site information by detecting the amino acid residues that are phosphorylated in a particular protein, determining the number of phosphorylation sites and therefore, establishing protein phosphorylation stoichiometry, identifying the kinase(s) responsible for the phosphorylation event, and analyzing

the functional impact of the observed phosphorylation (Patterson, Aebersold, & Goodlett, 2001). Hence, phosphoproteomics, a sub-discipline of proteomics, is focused on deriving a comprehensive view of the extent and dynamics of protein phosphorylation, and its ultimate goal is the rapid analysis of entire phosphorylation-based signaling networks (Mumby and Brekken, 2005; Olsen *et al.*, 2006).

While phosphorylation can occur on nine out of twenty amino acids, the most common and studied phosphorylation targets are Serine, Threonine and Tyrosine. It has been estimated that 90% of phosphorylation events occur on Serine, 10% on Threonine and less than 0.1% on Tyrosine (Hunter, 2000). Lately, there has been more focus on Histidine phosphorylation (Fuhs *et al.*, 2015).

Current methods for analysis of the phosphoproteome rely heavily on mass spectrometry (MS), an analytical technique for the determination of the composition of a sample by identifying the chemical structure of peptides. Most MS-based strategies for identifying phosphorylation sites in proteins include the following three stages: sample preparation involving protein purification and enzymatic digestion; isolation and separation of the phosphopeptides from non-phosphorylated peptides through enrichment and concentration procedures; and finally, identification and structural characterization of the phosphopeptides by MS (Corthals, Aebersold and Goodlett, 2005).

1.2 Separation and enrichment

Prior to analysis in a mass spectrometer, peptides are often separated in a liquid phase column based on their hydrophobicity using e.g. high performance liquid chromatography (HPLC). HPLC can be directly coupled to the mass spectrometer or separation can be done offline.

Low stoichiometry, heterogeneity, and low abundance of protein phosphorylation make it difficult to detect. Thereby, isolation and separation of the phosphopeptides from nonphosphorylated peptides through enrichment is crucial for phosphorylation analysis by MS. Among the separation techniques available, two-dimensional phosphopeptide mapping (2DPP), two-dimensional gel electrophoresis (2DE) and immobilized metal affinity chromatography (IMAC) have all been successfully used for the separation of phosphopeptides (Corthals,

Aebersold and Goodlett, 2005; Larsen and Robinson, 2008). The initial peptide loading conditions of titanium dioxide (TiO_2) phosphopeptide enrichment (Pinkse *et al.*, 2004) were further optimized (Larsen *et al.*, 2005) and the approach has gained in popularity and is now widely used in phosphoproteomics studies. TiO_2 chromatography offers high purification efficiency and specificity, isolating phosphopeptides from nonphosphorylated peptides.

1.3 Mass-spectrometry

MS instruments, mass spectrometers, consist of three main parts: an ion source, a mass analyzer, and a detector. Before the components of the sample can be measured, the ion source has to ionize them and convert them into gaseous ions, i.e. electrically charged molecules. Next, these analyte ions are transferred to the mass analyzer, which separates them according to their mass-to-charge ratio (m/z). After separation, the components reach the detector, which records the output as the ion intensity at different m/z values. This output can be visualized by a plot with m/z on the X-axis and ion intensity on the Y-axis, or a mass spectrum. Generally, every MS instrument is classified by the ionization method and the type of analyzer(s) it uses.

1.3.1 Ionization methods

The most commonly used ionization methods in proteomics are electrospray ionization (ESI) and matrix-assisted laser desorption and ionization (MALDI). Both ionization techniques were recognized by the Nobel Prize in Chemistry in 2002. One of the fundamental differences between these methods is that MALDI is employed on samples in a solid state, whereas ESI is employed on samples in a liquid state.

For MALDI (Karas and Hillenkamp, 1988; Tanaka *et al.*, 1988), the analyte is first dissolved with a large amount of a chemical matrix. The mixture of sample and matrix is then spotted onto a plate and left to dry. The evaporation of residual water or other solvent from the sample allows the formation of a crystal lattice into which the peptide sample is integrated cocrystallizing analyte and matrix. MALDI creates ions by the laser energy striking the crystalline matrix, which has a specific absorption wavelength that is close to the laser wavelength, and consequently causing rapid excitation of matrix and subsequent ejection of matrix

and analyte ions into the gas-phase. Singly protonated analyte ions are formed, which are then guided to the mass analyzer of choice by electrical potentials.

For ESI (Whitehouse *et al.*, 1985; Fenn *et al.*, 1989), a liquid-chromatography (LC) column that contains the analyte and solvent molecules is held at a high electrical potential, which creates gas-phase ions. The sample enters the source through a flow stream and passes through a stainless-steel cone or needle held at high voltage. As the flow stream exits the needle, it sprays out in a fine spray of droplets. The droplets contain peptide ions as well as components of the LC mobile phase (water, acetonitrile, acetic acid, etc.). Next, the source separates the peptide ions from the solvent components through the process known as desolvation, and transfers the ions into the mass analyzer. Solvent is removed as the droplets enter the MS by heat or some other form of energy such as energetic collisions with an inert gas (Matthiesen, 2007).

In contrast to MALDI, ESI yields multiply charged ions, which means ESI spectra are considerably more complex than MALDI spectra, with a collection of peaks for each charged state. In addition, by producing multiply charged ions, ESI makes larger proteins accessible to analysis than MALDI does. ESI spectra always require mass spectral deconvolution, that is, extraction of the molecular mass from the distribution of multiply charged ions of the molecule of interest, while for MALDI deconvolution is needed not nearly as often: in the cases of overlapping isotope patterns in very complex peptide samples (Xu *et al.*, 2018). Recently the influence of the ion source on peptide detection in large-scale proteomics was investigated (Nadler *et al.*, 2017). Significant differences were observed with respect to amino acid composition, charge-related parameters, hydrophobicity, and modifications of the detected peptides. Also, it has been shown that MALDI can complement ESI ionization for phosphoproteomics, particularly in detection of acidic and phosphotyrosine containing peptides (Ruprecht *et al.*, 2016). ESI shows better overall performance and it is currently the dominant ionization process.

1.3.2 Analyzers

All mass spectrometers have at least one analyzer. Instruments constructed with two or more analyzers that are coupled together are known as tandem MS or hybrid instruments. There are three most widely used groups of analyzers: time of flight, quadrupole, and ion trap analyzers. Ion trap and quadrupole analyzers are normally coupled to ESI ion sources, whereas time of flight analyzers are usually employed with MALDI ion sources.

Time of flight (TOF) mass analyzer measures the time it takes for the ions to fly from one end of the analyzer to the other and to strike the detector (Weickhardt, Moritz and Grotemeyer, 1996). The speed with which the ions fly down the analyzer tube is determined by their kinetic energy and inversely proportional to the root of the mass. Using a constant accelerating the field, the speed is inversely proportional to the square root of the m/z ratio. Quadrupoles (Q) consist of four parallel poles or rods. The electric field between the rods deflects the ions in complex trajectories, and as a result ions with the selected m/z ratio pass through the analyzer to be collected at the detector while other ions with unstable trajectories will eventually collide with the rods (Leary and Schmidt, 1996). Quadrupoles essentially act as mass-filters. Ion traps (IT) are using electromagnetic fields to retain, or "trap", ions inside, and depending on the setup the trapped ions can either be detected or used for further fragmentation. There are three major types of IT analyzers: radiofrequency ion traps (2D or 3D), Fourier transform ion cyclotron resonance (FT-ICR), and Orbitrap (Nolting, Malek and Makarov, 2019). Radiofrequency ion traps resemble quadrupoles in design, where in a 3D ion trap two parallel rods are replaced with two hyperbolic metal electrodes, and in 2D, also called linear ion traps, electrodes are coupled with quadrupole rods. Because of using direct or alternating current and radio frequency, ion traps can be used both to collect and inject pulses of ions coming from the ion source. FT-ICR and Orbitrap detect ions on the basis of their oscillating frequencies, requiring a Fourier transform algorithm for the signal processing, and are known for their high resolution and mass accuracy. FT-ICR (Comisarow and Marshall, 1974) detects ions through excitation of their circular motion in a strong magnetic field (the ion's cyclotron motion). The frequency of the cyclotron motion is inversely proportional to the ion's m/z ratio. Orbitraps (Makarov, 2000) use a spindle-shaped electric field to define ion motion such that radially the ions rotate around the spindle, and axially the ions oscillate at a frequency inversely proportional to the ion's m/z ratio.

In this thesis, three different mass spectrometers were used. One instrument with a MALDI source (Ultraflex II by Bruker Daltonics), and two ESI instruments (LTQ Orbitrap Velos and Q Exactive, both by Thermo Fisher Scientific). The schematics are presented in Figures 1, 2.

Ultraflex II TOF/TOF (Figure 1), introduced in 2003, enables high-throughput protein identification by MALDI-TOF peptide mass fingerprinting, immediately followed by more detailed protein characterization using MALDI-TOF/TOF tandem mass spectrometry on the same prepared sample. The instrument features a linear and reflectron TOF analyzer, PAN ("panoramic") technology for high MS mass resolution over a broad mass range (Suckau *et al.*, 2003).

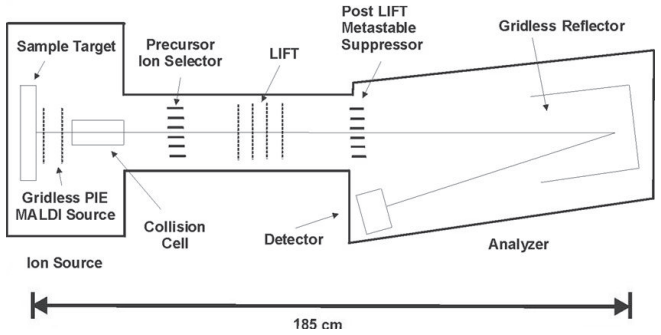

Figure 1 Schematic of Ultraflex II TOF/TOF (Image reproduced with permission of the rights holder, Bruker Daltonics)

Orbitrap LTQ Velos, launched in 2009, with the ETD module is a hybrid ion trap-Orbitrap instrument (Olsen *et al.*, 2009). It features a dual pressure linear ion trap, which is an independent MS detector that can store, isolate, and fragment ions and then send them either to the Orbitrap analyzer for further analysis or to a Secondary Electron Multiplier (SEM) detector.

Q Exactive (Figure 2) was launched in 2011 and it is one of the top Orbitrap-based mass spectrometers. It is a hybrid quadrupole-Orbitrap instrument that combines high-performance quadrupole precursor selection with high-resolution accurate mass Orbitrap detection (Michalski *et al.*, 2011). It features an S-lens ion source for increased sensitivity, a hyperbolic quadrupole mass filter for selection of precursor ions and ion transmission, a C-trap, Orbitrap mass analyzer for high

mass resolution and spectrum quality, and an HCD collision cell for high fragmentation efficiency for MS/MS spectra.

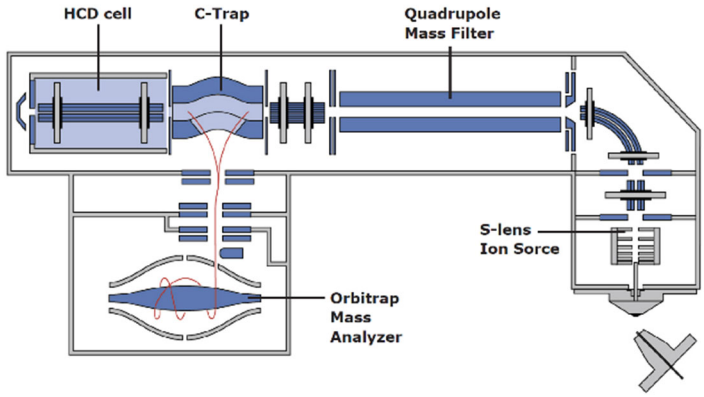

Figure 2 Schematic of Q Exactive (Image reproduced with permission of the rights holder, ThermoFisher Scientific)

A wide variety of MS configurations exists depending on the combination of the types of mass analyzers used, for instance tandem TOF (TOF/TOF), Q-TOF, triple quadrupole (TQ). Although these mass analyzers differ in the details of how they work, they all perform the same type of basic mass analysis. From a mixture of peptide ions generated by an ion source, the tandem MS analyzers select a single m/z species. This ion is then subjected to e.g. collision-induced dissociation (CID), which induces fragmentation of the peptide into fragment ions and neutral fragments. The fragment ions are then analyzed on the basis of their m/z to produce a product ion spectrum. The information contained in this tandem or MS/MS spectrum permits the sequence of the peptide to be deduced.

1.4 Peptide and phosphopeptide fragmentation

The characteristics of tandem mass spectra generated from peptide fragmentation depend on the peptide sequence, but also the mass spectrometer used, in particular the type of ion source and mass analyzer. In addition, the fragmentation techniques, or activation types, determine how the peptide ions are activated for fragmentation.

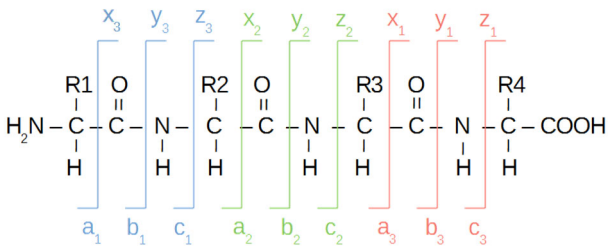

Figure 3 Generalised product ion series generated from a peptide fragmentation in the mass spectrometer

The general schematic of a peptide consisting of four amino acids is presented in Figure 3. Each amino acid has a central carbon atom (–C) attached to a carboxyl group (–COOH), an amino group (–H2N), a hydrogen atom (–H), and a side group (–R). Only the side group differs from one amino acid to another. Here, a chain of amino acids is formed through covalent peptide bonds, where one amino acid loses a hydrogen and oxygen from its carboxyl group (COOH) and the other loses a hydrogen from its amino group (NH2). This reaction produces a molecule of water (H_2O) and two amino acids join by a peptide bond (-CO-NH-). The two ends of a polypeptide chain are chemically different: the end carrying the free amino group is the amino terminus, or N-terminus, and the one carrying the free carboxyl group is a carboxyl terminus or C-terminus.

Fragments will only be detected if they carry at least one charge. If this charge is retained on the N terminal fragment, the ion is called either *a, b* or *c*; if the charge is retained on the C terminal, the ion type is either *x, y* or *z* (Roepstorff and Fohlman, 1984; Johnson *et al.*, 1987). A subscript indicates the number of residues in the fragment. When a *b/y*-type fragmentation occurs twice in the same molecule, it generates so-called internal fragment ions. An internal fragment with just a single side chain formed by a combination of *a*-type and *y*-type cleavage is called an immonium ion.

Three popular types of fragmentation were used in this thesis: CID, HCD and ETD. The most common peptide fragmentation method is collision-induced dissociation (CID), also known as collisionally activated dissociation (CAD) (Hayes and Gross, 1990; Morris *et al.*, 1996). CID fragmentation results in ions that are formed through cleavage of the weaker bonds within the peptide, resulting in *b*- and *y*-series of ions (Figure 3). In case of phosphopeptide fragmentation, prominent ions are produced after the loss of a phosphoryl group from

phosphorylated serine- and threonine- containing peptides, which can be either a direct loss of H_3PO_4 from the phosphorylated residue or the combined losses of HPO_3 and H_2O from the phosphorylation site and from an additional site within the peptide. The characteristics of CID tandem mass spectra of peptides were reviewed in detail (Papayannopoulos, 1995).

Another type of fragmentation method is called beam-type CID or higher energy collision dissociation (HCD). HCD is a variation of CID that requires increased radiofrequency voltage. Similar fragmentation patterns (*b/y*-type) are observed in CID and HCD, however, unlike CID, HCD spectra contain ions in low-mass regions, including a_2, b_2, y_1, y_2 and immonium ions (Olsen *et al.*, 2007). Phosphotyrosines typically lead to a unique immonium ion with m/z 216.0426, whereas HCD spectra of histidine-phosphorylated peptides have a diagnostic immonium ion with m/z 190.037 (Potel *et al.*, 2018). Neutral losses from phosphorylated serine and threonine are present in HCD spectra too, although the combined loss pathway was found to be less dominant under ion activation conditions associated with HCD-MS/MS than with CID-MS/MS (Cui *et al.*, 2014). Several studies compared HCD and CID fragmentation and investigated their performance for phosphopeptide analysis (Olsen *et al.*, 2009; Jedrychowski *et al.*, 2011; Michalski *et al.*, 2012).

Electron transfer dissociation (ETD) was introduced in 2004 (Syka *et al.*, 2004) and it is currently the most prominent alternative to CID. It was developed as a low-cost, more widely accessible ECD-like (Zubarev, Kelleher and McLafferty, 1998) dissociation method. ETD induces fragmentation of the peptide backbone along the pathways that are analogous to those observed in ECD. ETD generates *c*- and *z*-type fragment ions through cleavage of the bond between the amino group and carbon atom. The information content of an ETD spectrum is dramatically different form *b*- and *y*-type of ions observed in CID. The other important difference is that in case of phosphopeptide fragmentation, ions produced by loss of phosphoric acid are absent.

Another fragmentation type that combines ETD and HCD is called EThcD (Frese *et al.*, 2011). In principle, after an initial electron-transfer dissociation ETD step, all ions are subjected to CID, which yields both *b/y*- and *c/z*-type fragment ions in a single spectrum. This rich spectrum provides higher peptide sequence coverage and more confident localization of phosphorylation sites (Frese *et al.*, 2013).

2 IDENTIFICATION OF PEPTIDES AND PHOSPHOPEPTIDES

Peptides and proteins can be identified from MS1 level spectra using peptide mass fingerprinting (PMF) technique that was developed in 1993 by several groups (Henzel *et al.*, 1993; James *et al.*, 1993; Pappin, Hojrup and Bleasby, 1993; Yates *et al.*, 1993). In PMF, an unknown protein is digested with a protease with high bond specificity (e.g. trypsin) to yield constituent peptides. Molecular masses of these peptides are measured by MS1 analysis and the resulting list of masses is compared to a calculated peptide peak list obtained from the *in silico* digestion of each protein in a protein sequence database according to the rules defined by a set of user-defined parameters. The fingerprint of a protein is therefore the unique set of peptide masses generated by the cleavage.

The comparison of measured masses to calculated ones is done automatically by programs called search engines that calculate a score used for ranking the proteins. The most widely used PMF search engines are Mascot (Perkins *et al.*, 1999), ProFound (Zhang and Chait, 2000), and MS-Fit (Clauser, Baker and Burlingame, 1999). Among the parameters that can be specified in PMF search, the most critical one is the choice of protein sequence database. The other parameters include the enzyme used in the analysis, missed cleavages, mass tolerance and possible amino acid modifications.

Currently, the standard mass spectrometer for protein fingerprinting is MALDI-TOF type, but PMF can also be performed on spectra generated by ESI instruments. Before PMF can be performed, the acquired mass spectra that consist of signals of both sample and noise need to be preprocessed, to extract peaks lists.

Denoising and peak extraction is frequently done using the software that comes with a mass spectrometer, but there are also separate automated analysis tools, i.e. pipelines for high-throughput peptide mass fingerprinting (Samuelsson *et al.*, 2004). An additional step for ESI is deconvolution of multiply charged ions to singly charged ions. Post-processing would involve deisotoping and filtering out matrix (MALDI) or solvent (ESI) peaks, as well as contaminant peaks, such as keratins or enzyme autolysis peaks.

PMF was designed for protein identification, and it accomplishes it very well both for non-phosphorylated and phosphorylated proteins. However, it can also act as a peptide-identification technique in certain situations. The main limitation for peptide identification using PMF is that a number of different peptides may share the same molecular mass, if search is done e.g. against the entire SwissProt database. But if the original sample is known to contain just a few proteins, then the number of expected peptides is dramatically reduced, and those peptides that do not have in their sequences amino acids with identical molecular mass can be identified uniquely. For phosphopeptides, confident identification using PMF is also possible provided that the peptide has only one amino acid in its sequence that can be phosphorylated (S, T, or Y).

In most cases, however, peptides have more than one possible phosphorylation site and additional fragmentation of the peptide by e.g. CID is required. Methods used for the characterization of the fragments, which represent tandem mass spectra, are discussed in the following sections.

2.1 Peptide and phosphopeptide identification methods

In general, there are two main classes of methods for identifying peptides from MS/MS spectra. The first one is *de novo* sequencing (discussed in Section 2.1.1), where the peptide sequence is reconstructed from the spectrum based on the rules of the peptide fragmentation manually or automatically. The second class is by using specialized software tools that perform either protein database searching (Section 2.1.2) or spectral library matching (Section 2.1.3).

2.1.1 De novo sequencing

The derivation of a peptide sequence from its MS/MS spectrum alone, without help of a protein database, is called *de novo* sequencing. *De novo* sequencing methods allow reliable proteoform recognition and identification of previously unknown peptide sequences. More importantly, for the unknown or not yet sequenced genomes, this method is the only option to get sequence identifications.

There is a set of rules that are generally applied to *de novo* sequencing, concerning amino acid composition, loss of ammonia and water, isobaric masses, and spectral intensity rules. For example, the rules describe how to determine if the tryptic peptide ends with a K or an R (diagnostic y_1 ion at 147 is observed for K and 175 for R), or which fragments to expect to lose ammonia (R, K, Q, and N), and water (S, T, and E), when to expect double cleavage (at P or H residue), or y- and b-ions swapping intensities (when a P, H, K, or R is encountered in the sequence).

Some of the popular tools for *de novo* are PEAKS (Ma *et al.*, 2003; Zhang *et al.*, 2012), pNovo+ (Chi *et al.*, 2013), pepNovo+ (Frank and Pevzner, 2005; Frank *et al.*, 2007; Frank, 2009b, 2009a) and Novor (Ma, 2015). Their performance was compared in the recent reviews (Gorshkov *et al.*, 2016; Muth and Renard, 2018).

Data analysis using software tools for *de novo* sequencing can be divided into four main steps: preprocessing, candidate computation, refined scoring, and confidence scoring. Below, these four steps are described using the example of PEAKS. The first step, preprocessing of the raw MS/MS data, typically includes noise filtering, peak centering, and deconvolution of the doubly and triply charged species to singly charged ions. Preprocessing was found critically important for successful *de novo* sequencing, as preprocessing done by vendors' software differs a lot and might confuse spectra interpretation. Second, for candidate computation, 10,000 best sequences of all possible combinations of amino acids for a given precursor ion mass are computed. The algorithm considers the main ions to explain observed fragment masses using a score that rewards peaks close to calculated ions and co-existing $-H_2O$ and $-NH_3$ ions, taking into account the intensities of the peaks, while the missing peaks are penalized. The idea is to find a sequence with a maximum total reward score. Refined scoring is performed on all candidates or on a lower number specified by the user. Mass error tolerance used for refined score calculation is stricter, and additionally immonium ions and internal cleavage ions are used for reward and penalty calculations. Finally, the

confidence score is calculated for each of the top-scoring peptide sequences that includes the confidence score of each residue in the top-scoring sequences.

2.1.2 Protein sequence database search

Tools for interpreting tandem mass spectra using protein databases are called search engines just as tools for PMF. Most of the protein database search engines operate in the same manner. Each acquired tandem mass spectrum of a peptide is compared against theoretical tandem mass spectra of peptides generated by *in silico* digestion of the specified protein sequence database. In other words, in addition to PMF type of search, *in silico* fragmentation is performed on peptides with matching masses, and finally the resulting fragment ion peaks are compared to acquired tandem mass spectra. The general process of sequence database searching is illustrated in Figure 4.

A sequence database search cannot be performed without the actual protein sequence database (e.g. Swiss-Prot or TrEMBL). The other parameters that are provided by the user when submitting the search usually contain information about (1) the characteristics of mass spectrometer that was used for acquisition of the spectra, which will define the types of fragment ions; (2) the protease used to digest the proteins in the sample, which will limit the number of candidate peptides; (3) number of missed cleavages (internal K or R residues in case of trypsin digestion) that algorithm for *in silico* digestion should allow; (4) method of calculation of the peptide mass, i.e. monoisotopic or average; (5) parent ion mass tolerance that determines how close the mass of the measured peptide and calculated mass of the candidate peptides from the sequence database should be; and (6) fragment ion mass tolerance. Generally all search engines consider posttranslational modifications, including phosphorylation. If phosphorylation is specified in the initial query of the search, the search time will be prolonged considerably, since phosphorylation is a variable modification and a search engine will be generating candidate peptides both with and without modified residues that can be phosphorylated.

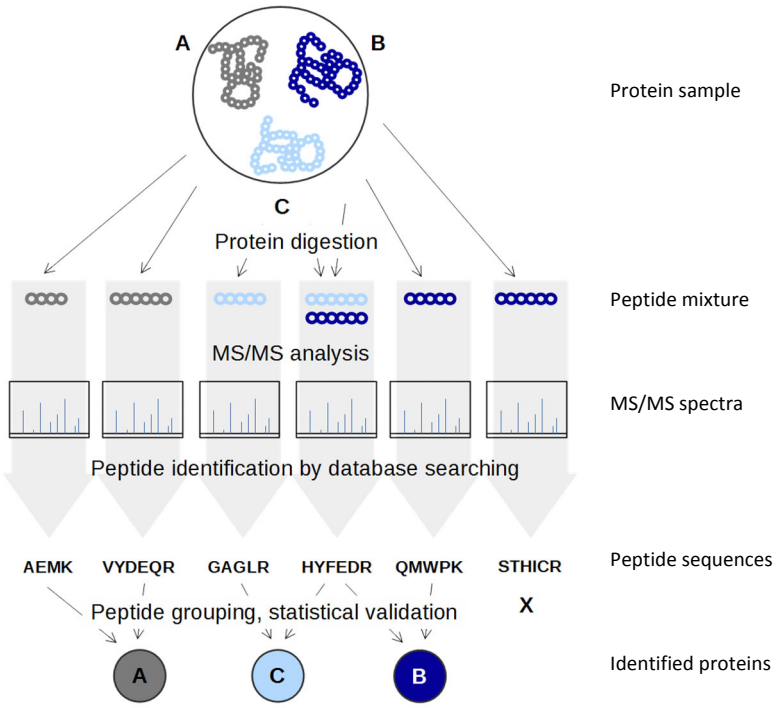

Figure 4 General view of the experimental steps and flow of the data in shotgun proteomics analysis (adapted from Nesvizhskii 2007)

Some of the most commonly used sequence database search engines, SEQUEST, Mascot and X!Tandem, are discussed below. Other popular tools that are not covered here are OMSSA (Open Mass Spectrometry Search Algorithm) (Geer *et al.*, 2004), Andromeda (Cox *et al.*, 2011), MS-GF+ (Kim and Pevzner, 2014) and Crux (Park *et al.*, 2008). MSFragger (Kong *et al.*, 2017) is the newest and reportedly the fastest to date database search tool designed specifically for PTM identification.

The scoring methods vary somewhat between engines, but comparisons reveal that they can be fairly similar in their performance, each finding a roughly similar number of peptides, although identifying different spectra and peptides. It has been shown that combining results of multiple search engines improves coverage of the analysis (Searle, Turner and Nesvizhskii, 2008; Shteynberg *et al.*, 2013).

This strategy has gained popularity and several proteomics data analysis pipelines support "simultaneous" analysis of the same raw data using more than one engine.

SEQUEST (Eng, McCormack and Yates, 1994) was the first database search engine that became commercially available. First, experimental spectra are pre-processed, the intensities of the peaks are normalized and low-intensity peaks are removed. Spectra of candidate peptides from the database are reconstructed using simplified fragmentation rules. SEQUEST then calculates the cross correlation score (Xcorr) between each experimental spectrum and its candidate peptides' spectra, which reflects the number of fragment ions that are common between each pair. Hence, the longer the peptide, the higher score is to be expected. Finally, the highest scoring peptide-spectrum match is reported, along with the Xcorr score and DeltaCn.

Mascot (Perkins *et al.*, 1999) is another widely used commercial database search engine. The main difference between Mascot and SEQUEST is in the way the similarity score is being calculated. It is also based on a number of matching fragment ion peaks, but it is not a simple count as in SEQUEST. Mascot estimates the probability of that number of matches to occur by chance, considering the numbers in the experimental and theoretical (predicted) spectra. The final score is called ion score.

X!Tandem (Craig and Beavis, 2003, 2004), originally called TANDEM, was the first open-source sequence database search engine. It relies on the assumption that for each identifiable protein, there should be at least one identifiable tryptic peptide. The workflow is therefore slightly different from the other tools. First, X!Tandem performs a quick database search; and from detected peptides, proteins are inferred. Then, it creates a database with those proteins, and finally searches the same tandem mass spectra again, this time extensively, including modified peptides, but against that database of reduced number of proteins. X!Tandem can be used through the interface of Trans-Proteomic Pipeline (Keller *et al.*, 2005; Deutsch *et al.*, 2010). The equivalent to SEQUEST's XCorr score is X!Tandems's hyperscore.

2.1.3 Spectral library search

Although sequence database searching is currently a standard method in proteomics, it has a number of shortcomings: it is time-consuming, error-prone and it can also be excessive, when within an experiment mostly the same spectra are searched multiple times and spectral information obtained in previous experiments remains unused. A decade ago a new type of search algorithm, called spectral library matching, has become available.

Unlike sequence database search engines, that compare experimental spectra to theoretical spectra generated based on protein sequence database, spectral library methods match experimental spectra to a library of spectra derived from previous identifications. By comparing two experimental spectra, this new approach takes advantage of comparison of specific spectral features (often ignored in database searches), including actual peak intensities, neutral losses from fragments, and various uncommon or even unknown fragments, to determine the best match. Therefore, the similarity scores of spectral searching algorithms are more precise and are well suited for identification of phosphopeptides, where tandem mass spectra are predominantly complex and rich in neutral loss ions. The relevance and potential of the use of spectral libraries in large-scale phosphoproteomics was highlighted in several studies (Bodenmiller *et al.*, 2008; Alcolea, Kleiner and Cutillas, 2009). The improved sensitivity of spectral library searching was demonstrated in a comparison study of spectral library search engine with several sequence database search engines using "equalized" search space, where the peptide content of the spectral libraries and the databases was the same (Zhang *et al.*, 2011). The results illustrated an additional advantage of spectral searching in identifying spectra of low-quality or spectra containing a lot of multiply charged fragments.

The first tools developed for spectral library searching were Bibliospec (Frewen *et al.*, 2006), X!Hunter (Craig *et al.*, 2006) and SpectraST (Lam *et al.*, 2007). All of them are free and open source. Spectrum Library Central, hosted at PeptideAtlas (Desiere *et al.*, 2006), provides access to the spectral libraries built specifically for spectrum library searching of tandem mass spectrometry data. Spectral libraries can be either downloaded from public resources or built in-house using the tools mentioned above.

A well prepared spectral library contains high quality spectra and reliable identifications with a single representative spectrum for every peptide, which can be constructed using several strategies. One approach is to select the highest scoring replicate spectrum (obtained in sequence database search) and discard the others, while the other is to combine replicate spectra into a consensus spectrum keeping those peaks that are present across replicates. Finally, depending on the tool, a different number of most intense peaks per spectrum will be kept in the library.

The typical search parameters used in spectral library searching are not nearly as many as in protein database searching and will usually include, in addition to the information about the spectral library to be searched, precursor mass or m/z tolerance and sometimes charge states. If one is interested in modified peptides, then those need to be present in the spectral library, otherwise they cannot be identified. Ideally, reference spectra for the library have been acquired on the same instrument or in the least the same type of instrument where the sample is analysed, which insures that fragmentation pattern will indeed be similar and spectra will be successfully and accurately matched to the peptides.

The latest reviews of the spectral library matching for proteomics (Griss, 2016; Shao and Lam, 2017) are listing more than 15 different tools. The most common scoring function used to estimate a pairwise spectrum comparison is the dot-product, also called the "spectral contrast angle". Its value ranges from 0 to 1, where 1 is given to a pair of identical tandem mass spectra.

Spectral library searching and protein sequence database searching can be considered to be complementary techniques. It has been shown that it is possible to increase the number of confidently identified tandem mass spectra by subsequent library search after the database search (Ahrné *et al.*, 2009). When searching the spectral library, constructed using database search results, it was possible to recover additional spectra that were noisy or represented modified and missed cleaved peptides. Interestingly, the majority of these spectra were also identified by database search, but were filtered out due to their low confidence scores. A similar hybrid approach was evaluated in another study (Cannon *et al.*, 2011).

2.2 Data formats and standards

Numerous commercial and open-source software tools have been developed for the analysis of proteomics data. That has become possible because of the adopted common data formats both by instrument and software vendors. There are several types of MS data, such as raw data, peak lists, peptide-level identifications, protein-level identification, output from quantification software and others. Currently, more and more software tools generate files directly in widely supported data formats, but converters to create compatible data files are still very popular.

Proteomics Standards Initiative (PSI) was established in 2002 by Human Proteome Organization (HUPO) with a goal to define and promote common standardized data formats and software tools for proteomics data (Orchard, Hermjakob and Apweiler, 2003). New standards are developed and old ones are maintained through PSI working groups by the members of the scientific community, and software and hardware vendors. Among the existing standards there are the MIAPE (Minimum Information About a Proteomics Experiment), mzML, mzIdentML, mzQuantML, and mzTab. The progress of the initiative over the years and planned future work were recently reviewed (Deutsch *et al.*, 2017). While some formats may be officially retired ("obsolete"), their use may still be compulsory as they may be the only supported input to certain software tools, such as SpectaST.

Trans-Proteomics Pipeline (TPP) (Keller *et al.*, 2005; Deutsch *et al.*, 2010) was the first proteomics data analysis platform that utilized open file formats, which were mainly XML-based (extensible markup language). TPP enabled a uniform analysis of MS/MS spectra using a variety of open source tools for sequence database and spectral library searching, validation of peptide and protein assignments as well as tools for peptide and protein quantitation. Other projects that provide a single environment for proteomics data analysis workflows are The OpenMS Proteomics Pipeline (TOPP) (Kohlbacher *et al.*, 2007) and MaxQuant (Cox and Mann, 2008).

One of the main advantages of using XML syntax is that it allows storing in a single document numerous structured components, i.e. elements that may include other elements. The examples of such elements are scans for files with spectral data and peptide-spectrum matches (PSM) for files with identifications. XML

formats used in proteomics have a well-defined structure, or schema. Invalid files, i.e. files that contain elements or their attributes that are not allowed, cannot be used for downstream analysis. XML data formats are both human readable and machine readable.

2.2.1 Spectra

The native output file formats from each mass spectrometer vendor that contain the raw measurement data (spectra) are different. For example, Thermo Fisher Scientific instruments generate .raw files, while .wiff files are typical to some of the ABI/Sciex mass spectrometers. mzXML data format (Pedrioli *et al.*, 2004) was developed at the Seattle Proteome Center (SPC) in the Institute for Systems Biology (ISB) as a common open data format for representation of MS data for data analysis within TPP and beyond. Even though .mzXML is a widely accepted format in the proteomics community, it is not a standard.

.mzML on the other hand is a standard that was designed as a joint project by HUPO-PSI, SPC-ISB and other members of proteomics research and industry (Deutsch, 2010). .mzML is a more complex format than .mzXML due to its integrated controlled vocabularies and in addition there exists a separate semantic validator.

2.2.2 Peptide identification

Data files generated by the tools for database or library search are called peptide identification files. Until recently, every search engine generated output in its own data format, which led to the development of parsers, or converters, that would allow data extraction from results of different search engines and automation of the downstream analysis, such as MascotDatfile (Helsens *et al.*, 2007), OMSSA parser (Barsnes *et al.*, 2009), and X!Tandem parser (Muth *et al.*, 2010). In addition, several converters are integrated within TPP, where the uniform format for storing and handling identification data has been .pepXML (.pep.xml). PepXML, however, similarly to .mzXML is an open format, but not a PSI standard.

In 2012, the first version of an exchange standard for peptide and protein identification data, mzIdentML, was published (Jones *et al.*, 2012). It was designed by HUPO-PSI to act as a single data format for identification data. In

2017, the latest version 1.2 of mzIdentML data standard was described (Vizcaíno *et al.*, 2017). One very important improvement included the implementation of features supporting scores associated with localization of PTMs on peptides, which were missing from the earlier versions.

3 VALIDATION METHODS OF PEPTIDE AND PHOSPHOPEPTIDE IDENTIFICATION

Methods for assessing the reliability of assignments of MS/MS spectra to peptides (peptide-spectrum matches, PSM) and for estimating error rates in the datasets have become very important in proteomics research and even more so in large-scale studies. Search engines return a list of the best matching peptides that were found in the database for almost every spectrum. However, the top peptide matches that are reported are not necessarily correct. In fact, in some cases the proportion of correct peptide sequence assignments is rather low.

A number of the likely reasons for poor results of sequence database searching have been summarized (Nesvizhskii, 2007). In addition to deficiencies of the search engines' scoring schemes that frequently use simplified rules for peptide ion fragmentation, poor quality of the spectra due to low signal-to-noise ratios, contamination or incomplete fragmentation will cause incorrect interpretation by the programs. Furthermore, when complex peptide mixtures are analysed, several peptide ions with similar m/z might fragment simultaneously, producing complicated MS/MS spectra that search engines would fail to assign correctly. And finally, charge state might be incorrectly determined, or search parameters specified by the user may be too strict, limiting the search space.

Manual validation of peptide assignments becomes impossible when dealing with thousands of spectra. Instead, historically peptide identification results would be ordered by the database search scores and simple score threshold would be applied, which is problematic when one wants to know the error rate or to compare or correlate results obtained using a different threshold or a search engine. An

improved approach is to convert a database search score into an expectation value (E-value), which reflects the expected number of peptides with scores equal to or better than the database search score by random chance. The lower the E-value, the less likely a PSM is to be random and therefore more likely to be correct. This type of probability-based scoring is implemented in Mascot.

3.1 False discovery rate

The most commonly used statistical measures for proteomics datasets are local and global false-discovery rates (FDR). The local FDR (also referred to as posterior error probability) is used to measure the statistical significance of individual PSMs, as in the case of Mascot's E-values. The global FDR, on the other hand, reflects the error rate for the set of PSMs and is very important for large datasets.

Local FDR estimation can be done using mixture model-based approach, where for every MS/MS spectrum, the frequencies of different scores from all the theoretical spectra in the database are fitted to a model distribution. This distribution is typically a mixture of two underlying distributions, representing correct and incorrect PSMs (Figure 5a). The two best known computational tools that implement model-based error rate analysis are PeptideProphet (Keller *et al.*, 2002; Choi and Nesvizhskii, 2008) and Percolator (Käll *et al.*, 2007; Brosch *et al.*, 2009). They also support global FDR estimations.

Global FDR can be estimated using target-decoy databases, statistical modelling, or a combination of both. Target-decoy strategy was first introduced in proteomics for evaluation of SEQUEST results (Moore, Young and Lee, 2002). The strategy is simple to implement and it is applicable to data generated by any search engine (Elias and Gygi, 2007). Tandem mass spectra are searched against the concatenated standard protein database (target) and database with clearly labelled reversed sequences (decoy). The proportion of decoy and target hits in the ordered list of highest scoring PSMs is used to guide the selection of the filtering criteria, such as discriminant score cut off, based on the desired statistic threshold (e.g. 1% FDR). Here, FDR is calculated as a number of false positives divided by the total number of PSMs above score thresholds. Decoy hits represent obvious false positives, but because there can be as many hidden false positives among the target hits, the convention has been to use the doubled number of decoy hits as false positive.

Numerous variations of the target-decoy method exist. In addition to simply reversing the sequences from the target database, decoys can also be generated by randomizing or shuffling. Importantly, for every method the size of the decoy database has to be the same as the target database. The decoy database search can be performed in tandem (concatenated database) or in sequence, where two separate searches are performed, one against the target and one against the decoy database (Fitzgibbon, Li and McIntosh, 2008). Alternatively, decoy fusion method, implemented in PEAKS (Zhang *et al.*, 2012), can be applied.

FDR estimation of peptide identifications using the target-decoy strategy was also evaluated for spectral library searching (Lam, Deutsch and Aebersold, 2010). An alternative FDR estimation method was later developed for SpectraST with a goal to enable decoy-free validation (Shao, Zhu and Lam, 2013). The similarity scoring function was modified to produce a score distribution that has a good discrimination power and in addition is more accurately fitted by PeptideProphet.

3.2 False localization rate

When database search results are scored, ranked and validated, they might be reported together with their FDR value and that is generally acceptable. However, in phosphoproteomic studies, when the results contain mainly phosphopeptides, that is not enough to ensure that the number of false identifications is minimum. The problem arises from the fact that the top-scoring alternative peptide identifications for each spectrum may have the exact same score, but only one of them is shown among results, or difference in the score is insignificant, suggesting that both are valid candidates. In case of spectra of modified peptides, these candidates are phosphopeptide isoforms (homologues), i.e. the same peptides, but with different phosphorylation sites. Consequently, identifications that were filtered for FDR may still contain incorrect assignments, because the reported hits are not necessarily correct. Below is a schematic from a paper by (Zubarev, Zubarev and Savitski, 2008) (Figure 5b) highlighting that a task of setting up a score threshold above which the assignments are considered reliable becomes complicated if not impossible in case of a dataset with homologous hits, such as a phosphoproteomics dataset.

The implications of phosphorylation site mis-assignments may not seem severe, because the peptide sequence is after all typically correct and it will contribute to identification of the correct protein. However, unambiguous identification of

phosphorylation sites can help to avoid costly and time-consuming downstream biological characterization experiments misguided by incorrect site assignments (Gunawardena *et al.*, 2011; Vaga *et al.*, 2014).

For a while, there has not been any uniform method for calculating and reporting ambiguity of phosphopeptide identifications, until a study conducted in 2010 by ABRF Proteome Informatics Research Group (iPRG). The focus of the study was to evaluate how bioinformatics analysis is done in proteomics laboratories across the world using a common dataset of phosphopeptides. It was shown to be generally challenging for all the participants to identify phosphopeptides and to localize the phosphorylation sites correctly. In the framework of this study, the term False Localization Rate (FLR) was introduced. FLR is calculated by dividing the number of incorrect site assignments by the total number of site assignment as a function of the score. True FLR can only be measured when the correct phosphosites are known, which is the case when synthetic phosphopeptides are used.

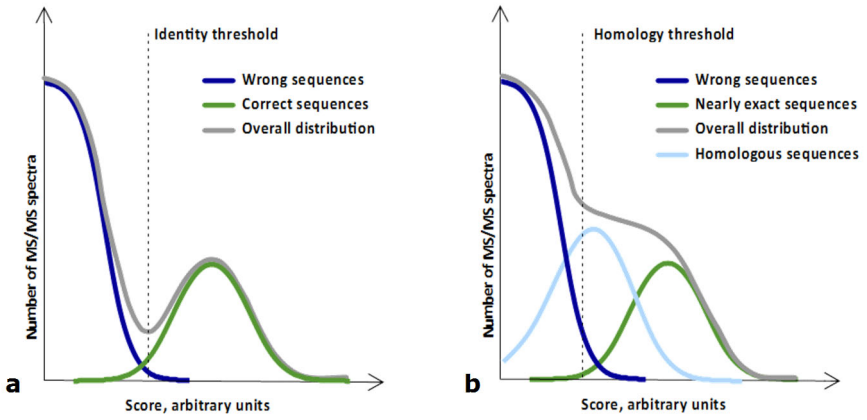

Figure 5 Theoretical shapes of the distributions of search-engine scores. (a) Idealized "hit or miss" situation; (b) more realistic situation taking into account the existence of homologs in MS/MS datasets (adapted from Zubarev et al.)

Various computational approaches have been developed for large-scale phosphoproteomics to localize phosphorylation sites in peptide sequences (Chalkley and Clauser, 2012; Wiese *et al.*, 2014; Ferries *et al.*, 2017). As a result, the quality of reported phosphopeptide identification data has improved significantly. Some of the representative methods, which can be divided roughly into two groups, are reviewed below in Section 3.3 (methods based on delta scores) and Section 3.4 (specialized phosphosite validation tools).

3.3 Delta scores

3.3.1 Mascot delta score

Database search engine Mascot outputs Mascot delta (MD-) score, which reflects the difference of the Mascot ion scores between the highest and the second highest ranking peptide spectrum match. A related score, normalized MD-score, was at some point considered and used for phosphosite validation. Specifically, normalized MD-score was defined as the difference between the top two Mascot ion scores of alternative phosphorylation sites in the same peptide sequence divided by the ion score of the top ranking site. However, after re-evaluation of the ability of MD-score without normalization to estimate the probability of correct phosphosite localization (Savitski *et al.*, 2011), it showed outstanding performance compared to normalized MD-score and popular phosphosite validation tool, Ascore. In that study, FLR of each method was calculated using a benchmarking data set of 180 synthetic phopshopeptides with known phosphorylation sites. The authors also found that phosphosite assignments are more reliable both by MD-score and Ascore if the potential phosphorylation sites are more than one amino acid apart when compared to sites that are adjacent. Afterwards, Mascot results in phosphoproteomics experiments would frequently be reported together with Mascot delta score information.

3.3.2 Sequest delta score

Delta Cn, ΔC_n, is the normalized difference between the best and the second-best scores reported by SEQUEST (Eng, McCormack and Yates, 1994). Similarly to MD- or normalized MD-score, this score helps determine the uniqueness of a match. If there are multiple matches, the reported ones would be the best and the

matches with the score within 5% of that best match. A commonly used cutoff for ΔC_n is 0.10. The value ΔC_n indicates the difference of the normalized correlation parameters between the first and second ranked sequences from the species-specific database search. These days SEQUEST results would be more commonly accompanied by the Ascore score, described below (3.4.1), rather than ΔC_n.

3.3.3 SpectraST delta score

By default, delta score calculated by spectral library search engine SpectraST reflects the difference between the dot-score of top and second-ranking non-homologous peptides. Therefore, it deliberately excludes phosphopeptide isoforms from the comparison. It is however possible to change that and compare the hits with the same peptide sequence but different phosphorylation sites. We called that difference score recalculated delta-score in P2. We have shown that a cutoff of 0.005 helped improve FLR in the experiment with synthetic peptides. Similarly to sequence database search results, there are situations (not as common though) where delta score is equal or near zero and such assignments can be eliminated as ambiguous. After the recalculated delta score is obtained, discriminant score (f-value) can be recalculated as well. F-value summarizes dot score, delta dot score, and a score that reflects the number of peaks that contributed to the match.

3.4 Phosphosite validation tools

3.4.1 Ascore

Ascore (Beausoleil *et al.*, 2006) was developed to enable validation of phosphopeptide identifications made by SEQUEST database search engine. Ascore measures the probability of correct phosphosite localization based on the presence and intensity of site-determining ions. To evaluate the Acore algorithm, several datasets with known phosphosites were used, including synthetic phosphopeptides and manually validated datasets. An Ascore value of at least 19 is recommended to achieve 1% FLR. The performance was compared with Mascot and SEQUEST, as well as specific Mascot and SEQUEST (shown further in brackets) scores, such as Delta Ions score (dCn) and Ions Score (XCorr).

MS/MS spectra are preprocessed to contain a certain number of most intense peaks per 100 m/z units. That number is called peak depth and it varies from 1 to 10. Water and phosphoric acid ion losses are removed, as well as all but one peak per isotopic cluster. The predicted spectra of b- and y- ions for every possible phosphopeptide isoform are compared separately with the preprocessed observed spectrum. Based on the matching peaks and peak depth, binomial probability is computed and only the best scoring phosphopeptide is reported. The actual Ascore is calculated using site determining ions.

Even though it is generally perceived that Ascore is only applicable to SEQUEST results, there is a study describing a workflow that included file conversions and the application of Scaffold software that allowed assigning Ascore values to Mascot results (Taus *et al.*, 2011). On a separate note, Ascore was developed further by another group into SLoMo (Bailey *et al.*, 2009) or turbo-SLoMo (Collins *et al.*, 2014). SLoMo supports files from SEQUEST, Mascot and OMSSA, low and high resolution CID, ETD and ECD data. SLoMo, like Ascore, reports the best scoring site(s) assignments, second best are not known.

3.4.2 PTM score

PTM score (Olsen *et al.*, 2006) is a phosphosite validation tool that is included in MaxQuant (Cox and Mann, 2008) and can be used after Andromeda (Cox *et al.*, 2011) or Mascot database search. PTM score is based on an algorithm previously published by the same group, that improved peptide identification by two consecutive stages of mass spectrometric fragmentation (MS^2 and MS^3) (Olsen and Mann, 2004). PTM score is using, similarly to Ascore, so-called peak depth, only it is pre-set to 4 most intense fragment ions per 100 m/z units in MS/MS. Unlike Ascore, all potential phosphorylation sites are reported together with their PTM scores, which represent localization probabilities. A cutoff value of 0.75 is used to separate peptide-spectrum matches with ambiguous phopshosite localization and more reliable ones that are called “class I” sites.

3.4.3 PhosphoRS

PhosphoRS (Taus *et al.*, 2011), is a phosphosite localization tool implemented in Java and integrated in Proteome Discoverer (Thermo Fisher Scientific) analysis software. PhosphoRS was developed to support phosphosite validation of spectra from several fragmentation techniques. The scoring parameters are optimized for MSA, HCD, ETD and ETD-high. Unlike other similar tools, the peak depth for peak extraction is not fixed in phosphoRS, but the optimal value is dynamically determined for each 100 m/z window (maximum of 8 peaks) by calculating the cumulative binomial probability for each phopshosite. Another difference compared to e.g. Ascore is that not only site-determining fragment ions are used to calculate probability of the site, but all theoretical fragment ions. According to the authors, phosphorylation site probability of 0.99 should correspond to FLR of 1%, while phosphoRS site probability of 0.75 should still lead to FLR of less than 2%.

In the original study phosphoRS was compared to Ascore and MD-score using synthetic phosphopeptide and biological data sets. PhosphoRS could localize phosphosites at 1% FLR in a higher number of PSMs and unique phosphopeptides than the other validation methods. Later, phosphoRS was renamed into ptmRS, since the application of the tool was expanded to assign probabilities to other modifications.

3.4.4 LuciPHOr

LuciPHOr (Fermin *et al.*, 2013) is a phosphorylation site localization algorithm with direct FLR estimation, compatible with input from Mascot, SEQUEST and X!Tandem search engines. Relying on a novel target-decoy-based approach, the algorithm uses both mass accuracy and peak intensities for site localization scoring and FLR estimation. For each identified peptide, permutations with modification on every amino acid are considered. Decoy permutations are generated by placing phosphorylation on amino acids that cannot be phosphorylated (non-S/T/Y), while non-decoy permutations are the ones with modification on S, T, or Y. All permutations are scored based on how well the observed spectrum matches the theoretically calculated fragment ions. Based on the score distributions of decoy and non-decoy permutations, LuciPHOr calculates the FLRs using the statistical method for estimating FDRs with the

empirical Bayes method. LuciPHOr's performance was evaluated using synthetic phosphopeptide data and it was compared with MD-score and Ascore.

Later, LuciPHOr2, a re-implementation of the original LuciPHOr, was developed (Fermin *et al.*, 2015). It has several improvements and novel features, such as operation system independence, reduced computation time and in addition to phosphorylation, support for site localization of generic PTMs. Another important improvement is that LuciPHOr2 reports the two best site localizations, and not just one, as previously, allowing identification of positional isomers, i.e. co-eluting species of the same peptide with different phosphorylation sites.

3.5 Other strategies

Several other tools for phosphosite validation should be mentioned, such as PhosSA (Saeed *et al.*, 2012, 2013) that supports Mascot and SEQUEST results compared to PhosphoRS in the original study, modification site localization score SLIP (Baker, Trinidad and Chalkley, 2011) integrated into Protein Prospector (Chalkley *et al.*, 2008), and finally the most recent method, P-bracket (phospho-bracket) (Xiao *et al.*, 2017), that relies on ion-pairs, which are distinctive to phosphorylation sites.

Besides delta scores and scores generated by designated validation tools, orthogonal phosphopeptide identification methods can be used for phosphosite validation. For example, it may be enough to analyse raw MS/MS data of phosphopeptides using different search engines, thereby enabling analysis by additional validation tools, which were not directly supported by the single search engine originally used. Alternatively, in addition to protein database search engines, spectral library searching can be used, provided a suitable spectral library is available. We have developed semi-simulated spectral libraries for this purpose (studies P3 and P4), and others have suggested to combine spectral libraries of high-confidence phosphopeptide spectra with predicted spectra to improve the coverage (Shao and Lam, 2017) of the proteome in the library. It has been shown that PTMs can be discovered by "blind", or "open" modification search, where the algorithm tries to detect mass-shifted spectral matches (Bandeira, 2007). The open modification search tools designed for spectral library searching are pMatch (Ye *et al.*, 2010) and QuickMod (Ahrné *et al.*, 2011). Open search is now also supported by SpectraST (Ma and Lam, 2014).

One of the main reasons for ambiguous assignments of phosphopeptides using tandem mass spectrometry and automated tools for spectra interpretation is the fact that sequence information is usually incomplete, i.e. not all the fragment ions are present in the spectra as predicted by the algorithms. The situation can be improved by repeating the analysis with complementary fragmentation or ionization techniques, using other protease than trypsin or performing digestion twice. That in turn would require more advanced bioinformatics tools that would combine the complementary data.

4 SUMMARY OF THE THESIS WORK

Contributions

This dissertation is composed of four original publications, all of which have been peer-reviewed and are now published. The topic for all studies remained the same, development of methods to improve accuracy of site-specific phosphorylation analysis by mass spectrometry. In **P1**, we have shown that combining spectra from different matrices when analyzing the same sample with MALDI ionization without prior phosphopeptide enrichment can generate higher number of confidently identified phosphopeptides than when using matrices separately or using ESI ionization with phosphopeptide enrichment. My contribution in this project included data analysis, development of scripts for spectra combination and for comparison of the results, and manuscript review and editing. In **P2**, various sample loading and washing conditions for the indium tin oxide (ITO) coating glass slides used in MALDI-MS were systematically tested and the optimal ones were determined for the most efficient phosphopeptide enrichment. My contribution was the development of scripts to support data analysis and integration, and manuscript review and editing. In **P3**, we have tested the hypothesis that spectra of phosphopeptides can be predicted based on the spectra of dephosphorylated peptides and then used for phosphosite validation using spectral library searching. In this project, I implemented and refined the phospho-peptide spectral simulation method, performed protein database and spectral library searches, compared the performance of software tools, and wrote the original draft of the manuscript. In **P4**, we significantly improved the software implementation of SimPhospho from P3 and optimized the simulation parameters. My contribution was the design of the graphical user interface, performance

optimizations, introduction of additional features, data analysis, and writing the original draft of the manuscript.

P1, **P2** and **P3** were published in specialized proteomics journals (*Rapid communications in mass spectrometry*, *Molecular Biosystems*, and *Journal of proteome research*), while a more readily accessible and established solution presented in **P4** was published in a more general journal, *Bioinformatics*.

4.1 Data combination from multiple MALDI matrices: opportunities and limitations for MALDI analysis

In study P1, we examined spectra generated from different MALDI matrices and found that they sometimes can be regarded as complementary. That led us to the development of a method that allowed identification of a larger number of phosphopeptides with MALDI-TOF/TOF using four different matrices and without prior phosphopeptide enrichment than when using ESI-qTOF after TiO_2-purification. In addition, the sample amount needed for the MALDI multi-matrix workflow was reduced greatly compared to ESI analysis.

The following workflow was used. NFATc1 protein sample was digested with trypsin. The resulting peptide sample was first analysed with four different MALDI matrices. Then, MS-data (peptide masses) were collected from all the matrices and combined for PMF. Identified phosphopeptides were subjected to MS/MS analysis with all four matrices regardless from which matrix they were originally detected. MS/MS spectra from the same precursor mass were merged and used for MS/MS Mascot database search. The illustration of the data flow is shown in Figure 6.

The initial data analysis was done using Biotools (Bruker Daltonics). Through the interface of this program one can perform both peptide mass fingerprinting (MS level) and Mascot database search (MS/MS level). We wanted to see if combined, or in other words, merged spectra used as an input would generate comparable results with methods that use phosphopeptide enrichment. There were no means to accomplish that, so my task was to develop a Microsoft Excel macro that would allow creating different but easily customizable combined peak lists. These

resulting merged peak lists needed to be in a certain format, to be compatible with downstream analysis. The created macro was run repeatedly in the sample preparation optimization phase of the project. The user had to select the correct corresponding input files and followed the progress of the data analysis. In principle, mass-intensity pairs acquired on different matrices are collected by the script in one list, and then sorted by their mass. When peaks are found in different matrices in close proximity to each other, the average m/z is calculated and the highest intensity value is assigned to that averaged mass. Excel offered a familiar interface to the end-user and an interactive environment for automated execution of a series of computational and data manipulation steps. Macros were implemented using Visual Basic for Applications (VBA).

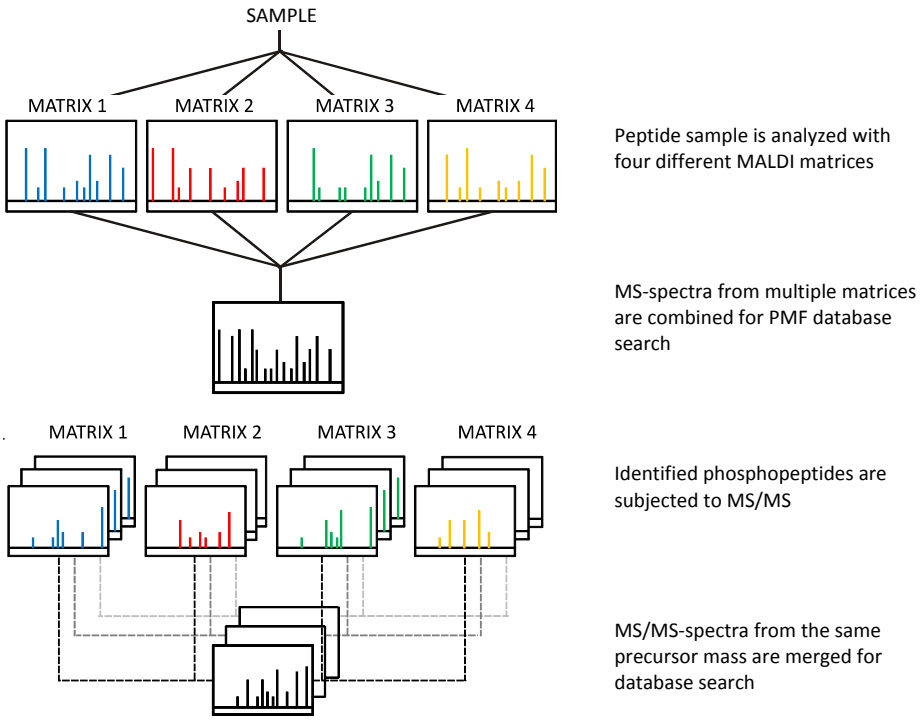

Figure 6 Schematic of multimatrix approach for phosphopeptide identification (adapted from Kouvonen 2010)

4.2 Enrichment and sequencing of phosphopeptides on indium tin oxide coated glass slides.

In paper P2, we evaluated the capability of commercially available indium tin oxide (ITO) coated glass slides used with MALDI ionization for phosphopeptide enrichment, earlier reported in (Imanishi *et al.*, 2009). Multiple sample loading and washing conditions were tested and finally an optimized protocol for extremely fast and sensitive phosphopeptide purification was suggested.

First, the effects of ammonia, washing solution and loading solutions were systematically tested using transcription factor NFATc1 phosphorylated with protein kinase A. We confirmed that the use of loading solution in combination with an ammonia (pH 11) wash prior to the matrix addition improved detection of phosphopeptides. Next, we tested the effect of TFA concentration (0.5%, 1%, 2%, 3%, 4%, 5%, and 6%) in the washing solution on phosphorylated and nonphosphorylated peptides. We found that already with 1% TFA phosphopeptide peak area was close to maximum and that concentration was chosen for the subsequent analysis.

To determine the sensitivity of ITO-purification and to compare our method with TiO_2 affinity chromatography, we analysed in triplicates samples with dilutions series containing 200, 100, 50, 25, 12 and 6 fmol of tryptic phosphopeptides. ITO-glass slide purification outperformed TiO_2 enrichment by allowing positive identifications starting at 12 fmol amount of sample, while TiO_2 required at least 100 fmol. In addition to determining the limit of identification described above, we compared ITO and TiO_2 in terms of limit of detection. Limit of detection for ITO-coated glass slides was 0.75 fmol and for TiO_2 is was 12 fmol.

The initial data analysis were performed similarly to the previous study. To allow a thorough comparison of many samples and conditions tested I developed VBA scripts for data analysis and integration of the results.

4.3 Confident site localization using a simulated phosphopeptide spectral library

In this paper, we described a novel method for detecting the sites of phosphorylation in phosphopeptides using ESI mass spectrometry. It enabled automated validation of phosphorylation sites using a reference-facilitated strategy (Imanishi *et al.*, 2007) on a large scale via spectral library searching of simulated spectra. The strategy is based on the observation that very similar fragmentation patterns are observed when spectra of phosphopeptides are compared with spectra of their dephosphorylated counterparts.

Four datasets including synthetic peptides and HeLa data were used in this study for method development and its benchmarking (Table 1).

Dataset	Fragmentation	Description
HeLa		
HeLa phosphorylated peptides	HCD, ETD, CID, MSA	Phosphopeptides were enriched from a HeLa tryptic digest
HeLa dephospho-rylated peptides	HCD, ETD, CID, MSA	Enriched HeLa phosphopeptides were dephosphorylated
Synthetic		
20 synthetic phosphopeptides	HCD, CID	20 singly phosphorylated peptides were selected and synthesized
Dataset from Marx *et al.*, 2013	HCD	>100,000 of singly phosphorylated peptides and nonphosphorylated counterparts were synthesized

Table 1 LC-MS/MS data sets used in this study (adapted from Suni *et al.*, 2015)

Figure 7 below shows the principle that was tested and evaluated using spectral library searching. Spectra of experimentally observed peptide, LFEDDDSNEK, with and without phosphorylation of Serine show a shift of 18 Da (denoted in blue) of ions containing a phosphosite. Simulated spectrum is shown on top and is used as a reference. In this study we have shown that it is possible to accurately simulate spectra of phosphopeptides by introducing a mass shift in the fragmentation pattern in a predicted way while preserving the intensity of the intact fragments.

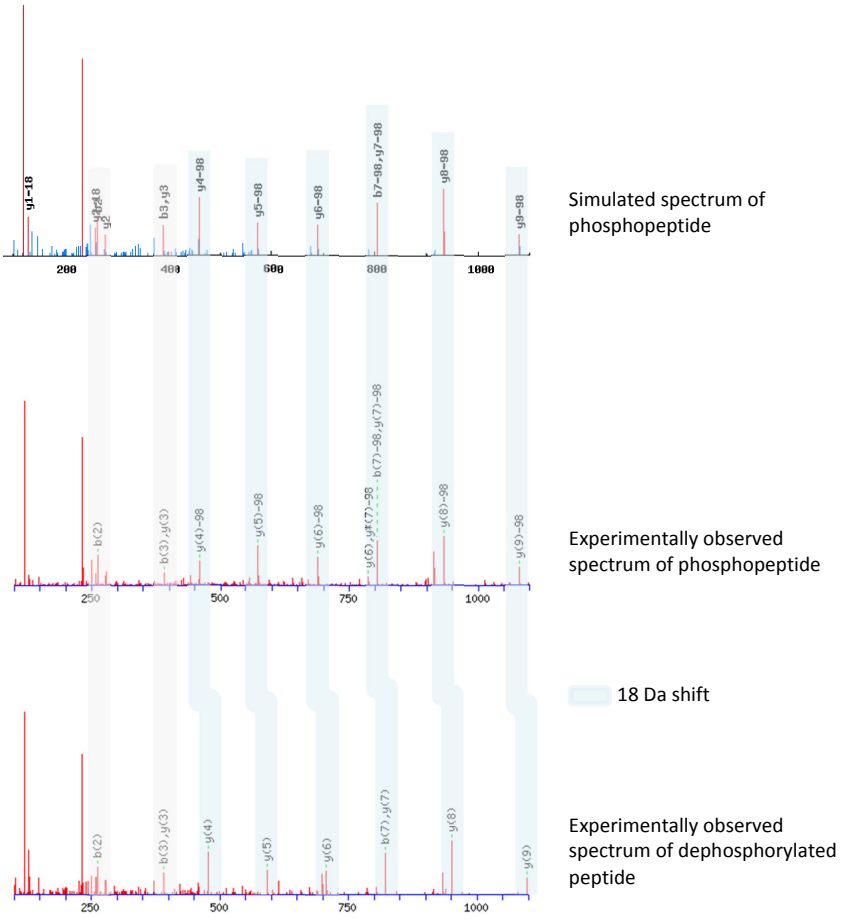

Figure 7 Representative HCD MS/MS spectra of phosphorylated and dephosphorylated peptides (LFEDDDS$_{167}$NEK)

The experimental workflow was as follows (Figure 8). Human HeLa cell protein sample was digested with trypsin and enriched for phosphopeptides using TiO_2 and then half of the sample was dephosphorylated. This sample preparation was done in Institute of Molecular Systems Biology, ETH Zurich. Then, in Proteomics facility in Turku Centre for Biotechnology, both samples were analyzed on the same instrument, LTQ Orbitrap Velos, by CID, MSA, ETD and HCD. We observed superior performance of HCD in terms of number of identifications, both in the analysis of phosphorylated and dephosphorylated peptides. Also, the most striking similarity between spectra of phosphorylated and dephosphorylated peptides was observed in HCD spectra. HCD spectra of the dephosphorylated

peptides were used to build a reference spectral library, for which spectra of all possible singly phosphorylated peptides were simulated. The simulated spectra were used to identify the phosphorylated peptides by spectral library searching.

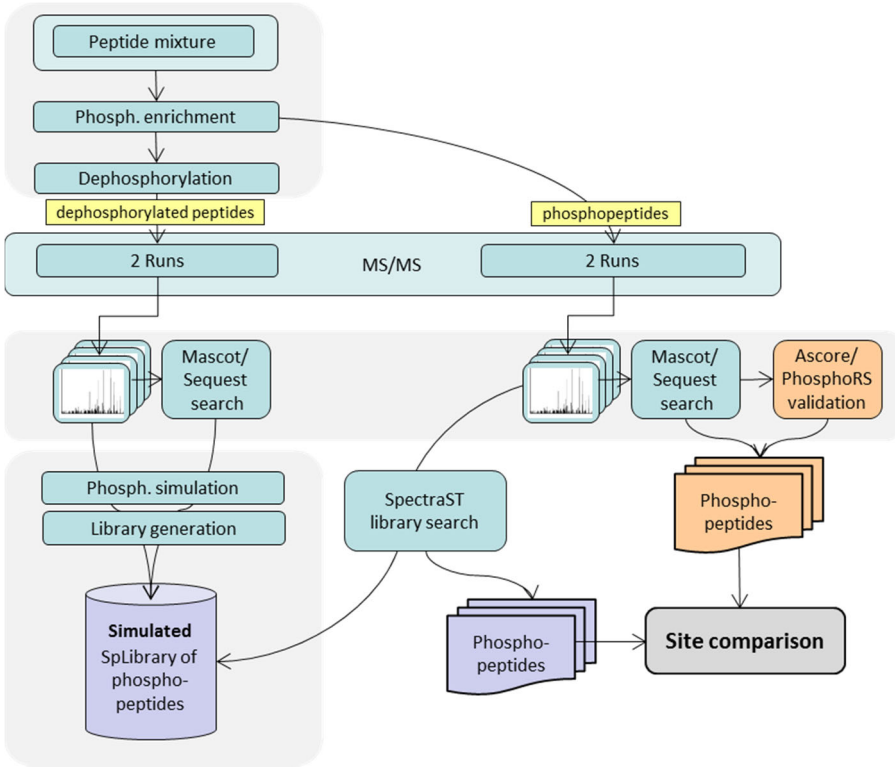

Figure 8 Experimental outline of reference-facilitated spectral library matching of phosphopeptides

We compared the performance of our simulated spectral library with publicly available ones, as well as combined our library to a larger background library. Datasets are organized below in a table (Table 2).

Spectral library	Num. of spectra	Description of source data
Experimental		
Synthetic phosphopeptide spectral library	31	HCD spectra of 20 synthetic phosphopeptides
Simulated		
SimHeLa library	23,126	Phosphopeptide HCD spectra simulated based on HeLA dephosphorylated peptides
SimMarx library (Marx *et al.*, 2013)	285,252	Phosphopeptide HCD spectra simulated based on synthetic nonphosphorylated peptides of Marx dataset
Combined		
SimHeLa-mouse-yeast library (Hu and Lam, 2013)	162,789	SimHeLa library merged with CID spectral libraries of mouse and yeast phosphopeptides
Hu&Lam library (Hu and Lam, 2013)	106,330	Library of experimentally obtained and simulated CID spectral of phosphopeptides

Table 2 Spectral libraries used in the study of simulated spectra of phosphopeptides (adapted from *Suni et al.*, 2015)

In addition to 20 synthetic phosphopeptides that we selected for synthesis and had analyzed in Turku, we used a published HCD dataset of 100,000 synthetic peptides. To simulate spectra for those, we used the spectra of their nonphosphorylated forms, generated by (Marx *et al.*, 2013). We wanted to evaluate the simulated spectral library using this large dataset that had pairs of phosphorylated and nonphosphorylated synthetic peptides. Majority of the spectral matches for that dataset of nonphosphorylated peptides had ambiguous identifications (30% Mascot delta score 0, and 60% Mascot delta score $\leq$ 10). For simulation we only used those matches that had Mascot identification higher than 1% FDR. Our results showed that our simulated libraries outperformed other simulated libraries and that they can be used as a promising alternative strategy for phosphosite validation.

4.4 SimPhospho: a software tool enabling confident phosphosite assignment

We were motivated to develop further the command line tool for simulation of phosphopeptide spectra implemented as part of the **P3** article. First of all, we wanted to investigate how the different intensity combinations would affect the accuracy of localization of phosphosites and to make improvements to the implementation, to make the simulation process faster, and to add more features. This new version of SimPhospho included a user-friendly Qt based GUI, executables for Linux, Mac and Windows operating systems, and the option to filter the initial MS/MS spectra using the scan number, specific peptide sequence or protein name. The default simulation parameters in the new program are set to the ones we found to produce the best results for accurate site-specific identifications. Interface of SimPhospho is shown in Figure 9.

As in the original algorithm, simulation of phosphopeptide fragment spectra (MS/MS) is performed using MS/MS of non-phosphorylated version of peptides. Ser/Thr phosphorylation and Tyr phosphorylation are simulated differently. For Ser/Thr phosphorylation we predict intact ions and neutral loss ions, but for Tyr it is only intact ions. In the previous implementation, pTyr containing fragment ions had intensity percentage that was automatically set equal to neutral loss ion intensity for pSer/pThr. In the new program this parameter is set separately. Also it is possible to select which ion types are going to be used for simulated spectra.

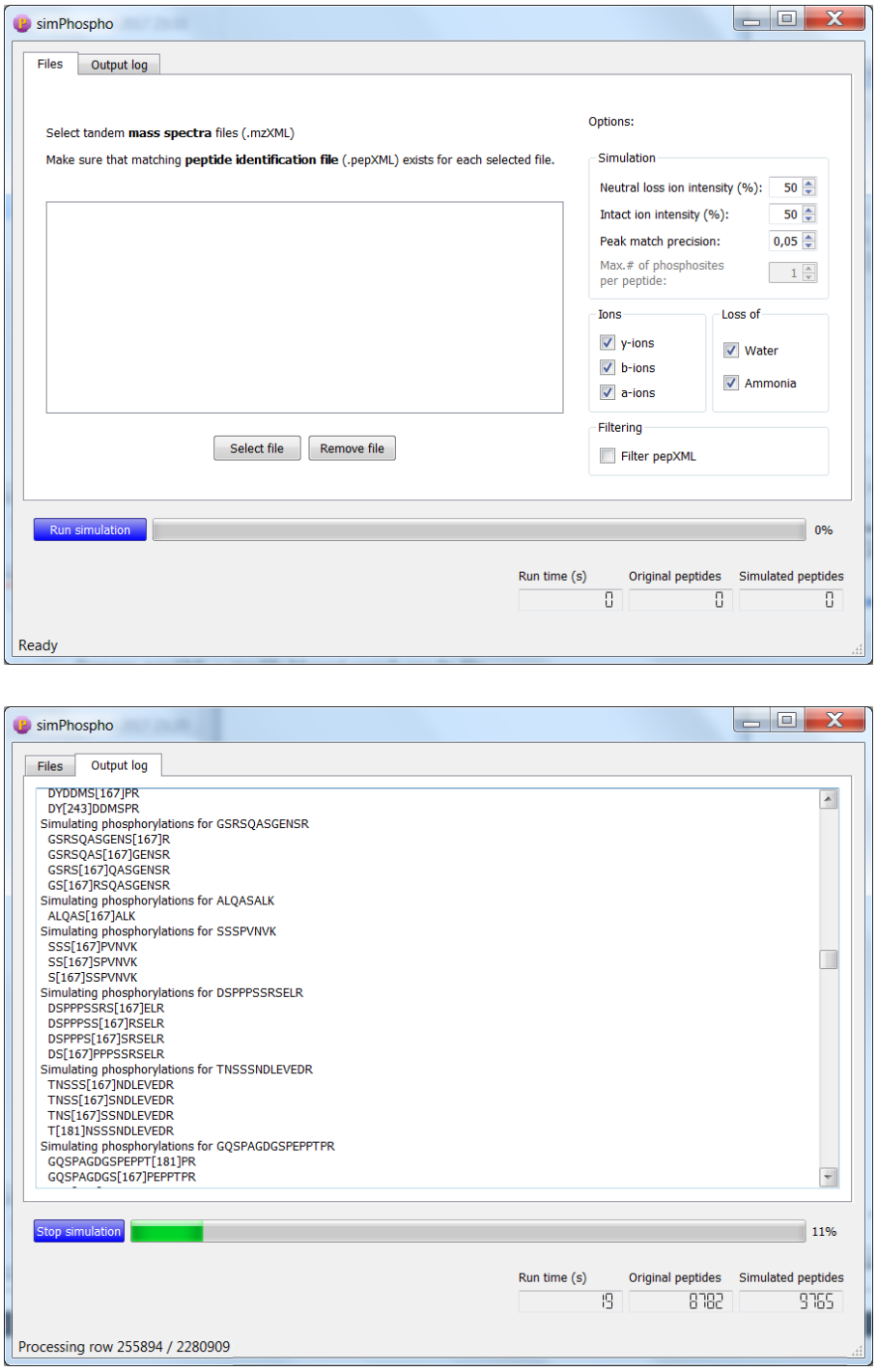

Figure 9 Interface of SimPhospho. Main view (top) and output log view (bottom)

We have studied a number of intensity combinations for simulated peaks. Originally, in **P3**, the values of these parameters were set empirically, by applying prior knowledge about the spectra of phosphopeptides. There, we chose to use 100% neutral loss, 10% intact ion intensity for pSer and pThr containing fragments, and 100% intact ion intensity for pTyr contacting fragments. However, in this study (**P4**) the best combination we found was (50%-50%ST,50%Y). Overall, the new tool provided faster data processing and improved phosphopeptides' MS/MS simulation compared to the older version implemented for **P3.** For example, on the same files provided as test data, the runtime improved from 48 min to 53 sec, the new version being ~50 times faster than the old one. This faster data processing has been achieved by the changes in the software implementation. We saw a clear increase in the number of correct spectra identified using SpectraST and simulated libraries. The increase at 1% FLR was 49%.

The new SimPhospho was significantly better than the prototype program, but at the same time it still relies on preparation of input data using TPP and conversion tools, just like the old version did. It is also biased towards singly phosphorylated peptides, in other words, peptides with multiply phosphorylated sites cannot be identified using a spectral library built using simulated spectra by SimPhospho.

DISCUSSION

In the past two decades, development of methods for phosphosite validation has been an active area of research. In the beginning of 2000s when phosphosite identifications were primarily reported without validation, many of those results inevitably contained error. Currently, the issue of ambiguity of site assignment by search engines is widely recognized. Consequently, the term “false localization rate” was adopted by the proteomics community, which helps researchers interpret the results with better understanding of their reliability.

Several phosphosite validation tools were developed, of which delta scores perhaps provide the easiest first measure of quality assessment of phosphopeptide identification. By applying simple arbitrary cutoffs – which have been accepted in the field – one separates and recognizes those phosphopeptide identifications that need to be taken with caution or require additional computational validation. Another promising approach involves the use of spectral libraries, which would be ideal for phosphopeptide identification because of the sensitivity of the similarity metrics used. Competing hits that represent the main challenge of spectra interpretation originate mainly from confusion between spectra of the phosphopeptide isoforms that frequently would have the majority of fragment ions overlapping. The matching scores used in spectral library searching can more accurately account for every peak and therefore are able to reflect the better hit more effectively compared to sequence database search engines. The limiting factor of the widespread use of spectral libraries in phosphoproteomics is that they are incomplete. Some phosphopeptide isoforms may not have yet been previously observed and therefore will be missed in the identification. Simulation of spectra to supplement the libraries was attempted before and worked well for nonphosphorylated peptides.

The novelty and the main contribution of this thesis work was the development of the simulation method that allowed simulation of all possible singly phosphorylated peptide isoforms if spectra of corresponding nonphosphorylated or dephosphorylated peptide were available. The major finding to emerge from this thesis work is that spectral library searching against simulated spectra of phosphopeptides achieves greater sensitivity in site-specific identification compared to the tools currently considered a golden standard.

A quantitative phosphoproteomic study (Kauko *et al.*, 2015) demonstrated the application of our developed approach and software, SimPhospho, where candidate phosphorylation sites obtained from a sequence database searches were further refined. Database search results were compared to the results of the spectral library search against a simulated library and based on the agreement and overlap, phosphorylation sites were either 1) confirmed, 2) flagged as ambiguous, or 3) complemented by phosphorylation sites identified only using a simulated library.

It is noteworthy that the use of spectral libraries has gained a lot of attention largely due to the popularity of targeted and data-independent acquisition proteomics, where spectral libraries play a central role in the data analysis. This thesis, however, applies spectral libraries in shotgun phosphoproteomics. Currently, spectral libraries can be integrated in a number of popular proteomics data analysis pipelines, including Trans-Proteomic Pipeline (TPP), Proteome Discoverer and OpenMS.

Data sharing in proteomics has been a great influence. Raw mass spectrometry data is deposited online, along with identification files and other needed resources to reproduce the results or to assess the quality of the data. Commonly, these data would be available not only for the reviewing procedure but also for the other researchers, that way supporting new method development and validation by the community. For instance, our published simulated spectral libraries from **P3** were used in a recent study that presented an automated pipeline called Epsilon-Q (Cho *et al.*, 2017). It implements the idea of combining the results from multiple simulated spectral libraries and DB searches. In addition to spectral library search and statistical estimation, results are combined and protein abundances are calculated within Epsilon-Q. Another example is an article describing a phosphosite validation method using phospho-brackets (Xiao *et al.*, 2017) that

used our HeLa phosphopeptide raw data and identifications that we made public in **P3** study.

Hopefully, this thesis work will raise awareness of the challenge of assigning phosphorylation sites accurately. We demonstrated that the use of simulated MS/MS spectra, i.e. (1) merged/combined spectra from different matrices in MALDI, and (2) simulated based on spectra of nonphosphorylated peptides in ESI, provides higher quality input data (MALDI, **P1**) or reference data for the library (ESI, **P3, P4**) and improves site-specific identification and validation of protein phosphorylation, reducing FLR of the analysis.

Future directions

The following areas for future research can be highlighted.

The most apparent future development areas include adapting the simulation method to multiply phosphorylation peptides, as well as spectra from another fragmentation type and other PTMs. In addition to CID and HCD spectra, we have observed the resemblance of fragmentation and intensity patterns of phosphorylated and dephosphorylated peptides in ETD spectra. However, we have not explored the simulation of that type of spectra yet. Further research might investigate if spectra of other PTMs would behave similarly to phosphorylation, and based on spectra of nonmodifed peptides, we could build spectral libraries of PTM of interest. An example of difficult to investigate but important PTM where simulated libraries could be tested includes glycosylation. Incidentally, our approach was recognized in the recent review article covering algorithms and design strategies towards automated glycoproteomics analysis (Hu, Khatri and Zaia, 2017).

All the methods developed in this thesis concern phosphorylation of serine, threonine and tyrosine residues, the three most common phosphorylation targets. We have considered the possibility to expand SimPhospho to simulate non-canonical phosphorylation e.g. of phosphohistidine. The detection of phosphohistidine has been shown to be very challenging (Kee and Muir, 2012). Based on the studies that observed spectra of phosphohistidines (Oslund *et al.*,

2014), we can anticipate that with some changes of our program, simulation of spectra of peptides containing phosphohistidine could be successful.

A significant improvement to our software would be the support of the latest HUPO-PSI standard formats. There are popular tools that employ these standards and their user-base is likely lost due to our software being incompatible with them. Workflow systems, such as KNIME (Berthold *et al.*, 2009; Fillbrunn *et al.*, 2017), allow integration of various old and new tools and creation of truly customizable data analysis pipelines. Proteome Discoverer offers similar interface based on nodes, but the choice of the nodes is limited, while tools, such as SearchGUI (Vaudel *et al.*, 2011; Barsnes and Vaudel, 2018) with PeptideShaker (Vaudel *et al.*, 2015) offer a user-friendly way to handle proteomics data.

And finally, having an option of fully automated phosphosite validation would be very beneficial. For that, considerably more work would need to be done. In particular, our suggested workflow would need to be optimized to be used on-the-fly, possibly together with the sequence database search engines, ultimately having all the data combined into a single result file. This would require the adaptation to the aforementioned standard formats and development of independent module that could be integrated with workflow systems.

Bibliography

Ahrné, E. *et al.* (2009) 'A simple workflow to increase MS2 identification rate by subsequent spectral library search.', *Proteomics*, 9(6), pp. 1731–6.

Ahrné, E. *et al.* (2011) 'QuickMod: A tool for open modification spectrum library searches', *Journal of proteome research*, 10(7), pp. 2913–2921.

Alcolea, M. P., Kleiner, O. and Cutillas, P. R. (2009) 'Increased confidence in large-scale phosphoproteomics data by complementary mass spectrometric techniques and matching of phosphopeptide data sets.', *Journal of proteome research*, 8(8), pp. 3808–15.

Bailey, C. M. *et al.* (2009) 'SLoMo: automated site localization of modifications from ETD/ECD mass spectra', *Journal of proteome research*, 8(4), pp. 1965–1971.

Baker, P. R., Trinidad, J. C. and Chalkley, R. J. (2011) 'Modification Site Localization Scoring Integrated into a Search Engine', *Molecular & Cellular Proteomics*, 10(7), p. M111.008078.

Bandeira, N. (2007) 'Spectral networks: a new approach to de novo discovery of protein sequences and posttranslational modifications', *BioTechniques*, 42(6), pp. 687–695.

Barsnes, H. *et al.* (2009) 'OMSSA Parser: An open-source library to parse and extract data from OMSSA MS/MS search results', *Proteomics*, 9(14), pp. 3772–3774.

Barsnes, H. and Vaudel, M. (2018) 'SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines', *Journal of proteome research*, 17(7), pp. 2552–2555.

Batth, T. S. *et al.* (2018) 'Large-Scale Phosphoproteomics Reveals Shp-2 Phosphatase-Dependent Regulators of Pdgf Receptor Signaling', *Cell Reports*, 22(10), pp. 2784–2796.

Beausoleil, S. A. *et al.* (2006) 'A probability-based approach for high-throughput protein phosphorylation analysis and site localization.', *Nat Biotechnol*, 24(10), pp. 1285–1292.

Berthold, M. R. *et al.* (2009) 'KNIME - the Konstanz information miner', *ACM SIGKDD Explorations Newsletter*, 11(1), p. 26.

Blueggel, M., Chamrad, D. and Meyer, H. E. (2004) 'Bioinformatics in proteomics.', *Current pharmaceutical biotechnology*, 5(1), pp. 79–88.

Bodenmiller, B. *et al.* (2008) 'PhosphoPep--a database of protein phosphorylation sites in model organisms', *Nat Biotechnol*, 26(12), pp. 1339–1340.

Brosch, M. *et al.* (2009) 'Accurate and Sensitive Peptide Identification with Mascot Percolator', *Journal of proteome research*, 8(6), pp. 3176–3181.

Cannon, W. R. *et al.* (2011) 'Large improvements in MS/MS-based peptide identification rates using a hybrid analysis.', *Journal of proteome research*, 10(5), pp. 2306–17.

Chalkley, R. J. *et al.* (2008) 'In-depth Analysis of Tandem Mass Spectrometry Data from Disparate Instrument Types', *Molecular & Cellular Proteomics*, 7(12), pp. 2386–2398.

Chalkley, R. J. and Clauser, K. R. (2012) 'Modification site localization scoring: strategies and performance', *Mol Cell Proteomics*, 11(5), pp. 3–14.

Chi, H. *et al.* (2013) 'pNovo+: de novo peptide sequencing using complementary HCD and ETD tandem mass spectra.', *Journal of proteome research*, 12(2), pp. 615–25.

Cho, J.-Y. *et al.* (2017) 'Epsilon-Q: An Automated Analyzer Interface for Mass Spectral Library Search and Label-Free Protein Quantification', *Journal of proteome research*, 16(12),

pp. 4435–4445.

Choi, H. and Nesvizhskii, A. I. (2008) 'Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics.', *Journal of proteome research*, 7(1), pp. 254–265.

Clauser, K. R., Baker, P. and Burlingame, A. L. (1999) 'Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching.', *Analytical chemistry*, 71(14), pp. 2871–82.

Collins, M. O. *et al.* (2014) 'Confident and sensitive phosphoproteomics using combinations of collision induced dissociation and electron transfer dissociation.', *Journal of proteomics*, 103, pp. 1–14.

Comisarow, M. B. and Marshall, A. G. (1974) 'Fourier transform ion cyclotron resonance spectroscopy', *Chemical physics letters*, 25(2), pp. 282–283.

Corthals, G. L., Aebersold, R. and Goodlett, D. R. (2005) 'Identification of Phosphorylation Sites Using Microimmobilized Metal Affinity Chromatography', in *Methods in enzymology*, pp. 66–81.

Cox, J. *et al.* (2011) 'Andromeda: a peptide search engine integrated into the MaxQuant environment', *Journal of proteome research*, 10(4), pp. 1794–1805.

Cox, J. and Mann, M. (2008) 'MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification', *Nat Biotechnol*, 26(12), pp. 1367–1372.

Craig, R. *et al.* (2006) 'Using annotated peptide mass spectrum libraries for protein identification', *Journal of proteome research*, 5(8), pp. 1843–1849.

Craig, R. and Beavis, R. C. (2003) 'A method for reducing the time required to match protein sequences with tandem mass spectra', *Rapid Communications in Mass Spectrometry*, 17(20), pp. 2310–2316.

Craig, R. and Beavis, R. C. (2004) 'TANDEM: matching proteins with tandem mass spectra', *Bioinformatics*, 20(9), pp. 1466–1467.

Cui, L. *et al.* (2014) 'Quantification of Competing H3PO4 Versus HPO3 + H2O Neutral Losses from Regioselective 18O-Labeled Phosphopeptides', *Journal of The American Society for Mass Spectrometry*, 25(1), pp. 141–148.

Desiere, F. *et al.* (2006) 'The PeptideAtlas project.', *Nucleic Acids Res*, 34(Database issue), pp. D655-8.

Deutsch, E. W. *et al.* (2010) 'A guided tour of the Trans-Proteomic Pipeline', *Proteomics*, 10(6), pp. 1150–1159.

Deutsch, E. W. (2010) 'Mass spectrometer output file format mzML.', *Methods in molecular biology (Clifton, N.J.)*. NIH Public Access, 604, pp. 319–31.

Deutsch, E. W. *et al.* (2017) 'Proteomics Standards Initiative: Fifteen Years of Progress and Future Work', *Journal of proteome research*, 16(12), pp. 4288–4298.

Deutsch, E. W., Lam, H. and Aebersold, R. (2008) 'Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics', *Physiol Genomics*, 33(1), pp. 18–25.

Elias, J. E. and Gygi, S. P. (2007) 'Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry', *Nature Methods*, 4(3), pp. 207–214.

Eng, J. K., McCormack, A. L. and Yates, J. R. (1994) 'An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database', *J Am Soc Mass Spectrom*, (5), pp. 976–989.

Farrah, T. *et al.* (2013) 'The State of the Human Proteome in 2012 as Viewed through PeptideAtlas', *Journal of proteome research*, 12(1), pp.

162–171.

Fenn, J. B. *et al.* (1989) 'Electrospray ionization for mass spectrometry of large biomolecules.', *Science*, 246(4926), pp. 64–71.

Fermin, D. *et al.* (2013) 'LuciPHOr: algorithm for phosphorylation site localization with false localization rate estimation using modified target-decoy approach', *Mol Cell Proteomics*, 12(11), pp. 3409–3419.

Fermin, D. *et al.* (2015) 'LuciPHOr2: site localization of generic post-translational modifications from tandem mass spectrometry data.', *Bioinformatics*, 31(7), pp. 1141–3.

Ferries, S. *et al.* (2017) 'Evaluation of Parameters for Confident Phosphorylation Site Localization Using an Orbitrap Fusion Tribrid Mass Spectrometer', *Journal of proteome research*, 16(9), pp. 3448–3459.

Fillbrunn, A. *et al.* (2017) 'KNIME for reproducible cross-domain analysis of life science data', *Journal of Biotechnology*, 261, pp. 149–156.

Fitzgibbon, M., Li, Q. and McIntosh, M. (2008) 'Modes of inference for evaluating the confidence of peptide identifications.', *Journal of proteome research*, 7(1), pp. 35–9.

Frank, A. M. *et al.* (2007) 'De Novo Peptide Sequencing and Identification with Precision Mass Spectrometry', *Journal of proteome research*, 6(1), pp. 114–123.

Frank, A. M. (2009a) 'A Ranking-Based Scoring Function for Peptide−Spectrum Matches', *Journal of proteome research*, 8(5), pp. 2241–2252.

Frank, A. M. (2009b) 'Predicting Intensity Ranks of Peptide Fragment Ions', *Journal of Proteome Research*, 8(5), pp. 2226–2240.

Frank, A. and Pevzner, P. (2005) 'PepNovo: de novo peptide sequencing via probabilistic network modeling.', *Analytical chemistry*, 77(4), pp. 964–73.

Frese, C. K. *et al.* (2011) 'Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-Orbitrap Velos.', *Journal of proteome research*, 10(5), pp. 2377–2388.

Frese, C. K. *et al.* (2013) 'Unambiguous Phosphosite Localization using Electron-Transfer/Higher-Energy Collision Dissociation (EThcD)', *Journal of proteome research*, 12(3), pp. 1520–1525.

Frewen, B. E. *et al.* (2006) 'Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries', *Anal Chem*, 78(16), pp. 5678–5684.

Fuhs, S. R. *et al.* (2015) 'Monoclonal 1- and 3-Phosphohistidine Antibodies: New Tools to Study Histidine Phosphorylation', *Cell*, 162(1), pp. 198–210.

Geer, L. Y. *et al.* (2004) 'Open Mass Spectrometry Search Algorithm', *Journal of proteome research*, 3(5), pp. 958–964.

Giansanti, P. *et al.* (2015) 'An Augmented Multiple-Protease-Based Human Phosphopeptide Atlas.', *Cell reports*, 11(11), pp. 1834–43.

Gnad, F., Gunawardena, J. and Mann, M. (2011) 'PHOSIDA 2011: the posttranslational modification database', *Nucleic Acids Research*, 39(Database), pp. D253–D260.

Gorshkov, V. *et al.* (2016) 'Peptide de novo sequencing of mixture tandem mass spectra', *Proteomics*, 16(18), pp. 2470–2479.

Griss, J. (2016) 'Spectral library searching in proteomics', *PROTEOMICS*, 16(5), pp. 729–740.

Gunawardena, H. P. *et al.* (2011) 'Unambiguous Characterization of Site-specific Phosphorylation of Leucine-rich Repeat Fli-I-interacting Protein 2 (LRRFIP2) in Toll-like Receptor 4 (TLR4)-mediated Signaling', *Journal of Biological Chemistry*, 286(13), pp. 10897–10910.

Hayes, R. N. and Gross, M. L. (1990) 'Collision-induced dissociation.', *Methods in enzymology*, 193, pp. 237–63.

Helsens, K. *et al.* (2007) 'MascotDatfile: an open-source library to fully parse and analyse MASCOT MS/MS search results.', *Proteomics*, 7(3), pp. 364–366.

Henzel, W. J. *et al.* (1993) 'Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases.', *Proceedings of the National Academy of Sciences of the United States of America*, 90(11), pp. 5011–5.

Hornbeck, P. V. *et al.* (2015) 'PhosphoSitePlus, 2014: mutations, PTMs and recalibrations', *Nucleic Acids Research*, 43(D1), pp. D512–D520.

Hu, H., Khatri, K. and Zaia, J. (2017) 'Algorithms and design strategies towards automated glycoproteomics analysis.', *Mass spectrometry reviews*, 36(4), pp. 475–498.

Hu, Y. and Lam, H. (2013) 'Expanding Tandem Mass Spectral Libraries of Phosphorylated Peptides: Advances and Applications', *Journal of Proteome Research*, 12(12), pp. 5971–5977.

Hummel, J. *et al.* (2007) 'ProMEX: a mass spectral reference database for proteins and protein phosphorylation sites', *BMC Bioinformatics*, 8, p. 216.

Hunter, T. (1995) 'Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling.', *Cell*, 80(2), pp. 225–36.

Hunter, T. (2000) 'Signaling--2000 and beyond.', *Cell*, 100(1), pp. 113–27.

Imanishi, S. Y. *et al.* (2007) 'Reference-facilitated phosphoproteomics: fast and reliable phosphopeptide validation by microLC-ESI-Q-TOF MS/MS', *Mol Cell Proteomics*, 6(8), pp. 1380–1391.

Imanishi, S. Y. *et al.* (2009) 'Phosphopeptide enrichment with stable spatial coordination on a titanium dioxide coated glass slide', *Rapid Communications in Mass Spectrometry*, 23(23), pp. 3661–3667.

James, P. *et al.* (1993) 'Protein Identification by Mass Profile Fingerprinting', *Biochemical and Biophysical Research Communications*, 195(1), pp. 58–64.

Jedrychowski, M. P. *et al.* (2011) 'Evaluation of HCD- and CID-type fragmentation within their respective detection platforms for murine phosphoproteomics.', *Molecular & cellular proteomics*, 10(12), p. M111.009910.

Johnson, R. S. *et al.* (1987) 'Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine', *Analytical Chemistry*, 59(21), pp. 2621–2625.

Jones, A. R. *et al.* (2012) 'The mzIdentML Data Standard for Mass Spectrometry-Based Proteomics Results', *Molecular & Cellular Proteomics*, 11(7), p. M111.014381.

Käll, L. *et al.* (2007) 'Semi-supervised learning for peptide identification from shotgun proteomics datasets', *Nature methods*, 4(11), pp. 923–925.

Karas, M. and Hillenkamp, F. (1988) 'Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons.', *Analytical chemistry*, 60(20), pp. 2299–301.

Kauko, O. *et al.* (2015) 'Label-free quantitative phosphoproteomics with novel pairwise abundance normalization reveals synergistic RAS and CIP2A signaling', *Scientific Reports*, 5(1), p. 13099.

Kee, J.-M. and Muir, T. W. (2012) 'Chasing Phosphohistidine, an Elusive Sibling in the Phosphoamino Acid Family', *ACS Chemical Biology*, 7(1), pp. 44–51.

Keller, A. *et al.* (2002) 'Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database

search', *Anal Chem*, 74(20), pp. 5383–5392.

Keller, A. *et al.* (2005) 'A uniform proteomics MS/MS analysis platform utilizing open XML file formats', *Molecular Systems Biology*, 1(1), pp. E1–E8.

Kim, M.-S. *et al.* (2014) 'A draft map of the human proteome', *Nature*, 509(7502), pp. 575–581.

Kim, S. and Pevzner, P. A. (2014) 'MS-GF+ makes progress towards a universal database search tool for proteomics', *Nature Communications*, 5(1), p. 5277.

Kohlbacher, O. *et al.* (2007) 'TOPP--the OpenMS proteomics pipeline', *Bioinformatics*, 23(2), pp. e191–e197.

Kong, A. T. *et al.* (2017) 'MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics', *Nature Methods*, 14(5), pp. 513–520.

Krishna, R. G. and Wold, F. (1993) 'Post-translational modification of proteins.', *Advances in enzymology and related areas of molecular biology*, 67, pp. 265–98.

Lam, H. *et al.* (2007) 'Development and validation of a spectral library searching method for peptide identification from MS/MS', *Proteomics*, 7(5), pp. 655–667.

Lam, H., Deutsch, E. W. and Aebersold, R. (2010) 'Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics.', *Journal of proteome research*, 9(1), pp. 605–610.

Larsen, M. R. *et al.* (2005) 'Highly Selective Enrichment of Phosphorylated Peptides from Peptide Mixtures Using Titanium Dioxide Microcolumns', *Molecular & Cellular Proteomics*, 4(7), pp. 873–886.

Larsen, M. R. and Robinson, P. J. (2008) 'Chapter 12 Phosphoproteomics', *Comprehensive Analytical Chemistry*. Elsevier, 52, pp. 275–296.

Leary, J. J. and Schmidt, R. L. (1996) 'Quadrupole Mass Spectrometers: An Intuitive Look at the Math', *Journal of Chemical Education*, 73(12), p. 1142.

Lee, D. C. H., Jones, A. R. and Hubbard, S. J. (2015) 'Computational phosphoproteomics: From identification to localization', *Proteomics*, 15(5–6), pp. 950–963.

Ma, B. *et al.* (2003) 'PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry.', *Rapid communications in mass spectrometry*, 17(20), pp. 2337–42.

Ma, B. (2015) 'Novor: real-time peptide de novo sequencing software.', *Journal of the American Society for Mass Spectrometry*, 26(11), pp. 1885–94.

Ma, C. W. M. and Lam, H. (2014) 'Hunting for Unexpected Post-Translational Modifications by Spectral Library Searching with Tier-Wise Scoring', *Journal of Proteome Research*, 13(5), pp. 2262–2271.

Makarov, A. (2000) 'Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis', *Analytical chemistry*, 72(6), pp. 1156–62.

Marx, H. *et al.* (2013) 'A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics', *Nat Biotechnol*, 31(6), pp. 557–64.

Matthiesen, R. (2007) *Mass spectrometry data analysis in proteomics*. Humana Press.

Michalski, A. *et al.* (2011) 'Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer', *Molecular & Cellular Proteomics*, 10(9), p. M111.011015.

Michalski, A. *et al.* (2012) 'A Systematic Investigation into the Nature of Tryptic HCD Spectra', *Journal of Proteome Research*, 11(11), pp. 5479–5491.

Moore, R. E., Young, M. K. and Lee, T.

D. (2002) 'Qscore: An algorithm for evaluating SEQUEST database search results', *Journal of the American Society for Mass Spectrometry*, 13(4), pp. 378–386.

Morris, H. R. *et al.* (1996) 'High Sensitivity Collisionally-activated Decomposition Tandem Mass Spectrometry on a Novel Quadrupole/Orthogonal-acceleration Time-of-flight Mass Spectrometer', *Rapid Communications in Mass Spectrometry*, 10(8), pp. 889–896.

Mumby, M. and Brekken, D. (2005) 'Phosphoproteomics: new insights into cellular signaling', *Genome Biology*, 6(9), p. 230.

Muth, T. *et al.* (2010) 'XTandem Parser: An open-source library to parse and analyse X!Tandem MS/MS search results', *Proteomics*, 10(7), pp. 1522–1524.

Muth, T. and Renard, B. Y. (2018) 'Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification?', *Briefings in bioinformatics*, 19(5), pp. 954–970.

Nadler, W. M. *et al.* (2017) 'MALDI versus ESI: The Impact of the Ion Source on Peptide Identification', *Journal of proteome research*, 16(3), pp. 1207–1215.

Nesvizhskii, A. I. (2007) 'Protein Identification by Tandem Mass Spectrometry and Sequence Database Searching', in *Mass Spectrometry Data Analysis in Proteomics*. New Jersey: Humana Press, pp. 87–120.

Nolting, D., Malek, R. and Makarov, A. (2019) 'Ion traps in modern mass spectrometry', *Mass Spectrometry Reviews*, 38(2), pp. 150–168.

Olsen, J. V. *et al.* (2009) 'A Dual Pressure Linear Ion Trap Orbitrap Instrument with Very High Sequencing Speed', *Molecular & Cellular Proteomics*, 8(12), pp. 2759–2769.

Olsen, J. V *et al.* (2006) 'Global, in vivo, and site-specific phosphorylation dynamics in signaling networks', *Cell*, 127(3), pp. 635–648.

Olsen, J. V *et al.* (2007) 'Higher-energy C-trap dissociation for peptide modification analysis', *Nature Methods*, 4(9), pp. 709–712.

Olsen, J. V and Mann, M. (2004) 'Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation.', *Proceedings of the National Academy of Sciences of the United States of America*, 101(37), pp. 13417–22.

Orchard, S., Hermjakob, H. and Apweiler, R. (2003) 'The Proteomics Standards Initiative', *Proteomics*, 3(7), pp. 1374–1376.

Oslund, R. C. *et al.* (2014) 'A Phosphohistidine Proteomics Strategy Based on Elucidation of a Unique Gas-Phase Phosphopeptide Fragmentation Mechanism', *Journal of the American Chemical Society*, 136(37), pp. 12899–12911.

Papayannopoulos, I. A. (1995) 'The interpretation of collision-induced dissociation tandem mass spectra of peptides', *Mass Spectrometry Reviews*, 14(1), pp. 49–73.

Pappin, D. J., Hojrup, P. and Bleasby, A. J. (1993) 'Rapid identification of proteins by peptide-mass fingerprinting.', *Current biology*, 3(6), pp. 327–32.

Park, C. Y. *et al.* (2008) 'Rapid and Accurate Peptide Identification from Tandem Mass Spectra', *Journal of Proteome Research*. American Chemical Society, 7(7), pp. 3022–3027.

Pedrioli, P. G. A. *et al.* (2004) 'A common open representation of mass spectrometry data and its application to proteomics research.', *Nature biotechnology*, 22(11), pp. 1459–66.

Perkins, D. N. *et al.* (1999) 'Probability-based protein identification by searching sequence databases using mass spectrometry data.', *Electrophoresis*, 20(18), pp. 3551–67.

Pinkse, M. W. H. *et al.* (2004) 'Selective Isolation at the Femtomole Level of Phosphopeptides from Proteolytic Digests Using 2D-NanoLC-ESI-MS/MS and Titanium Oxide Precolumns', *Analytical Chemistry*, 76(14), pp. 3935–3943.

Potel, C. M. *et al.* (2018) 'Widespread bacterial protein histidine phosphorylation revealed by mass spectrometry-based proteomics', *Nature Methods*, 15(3), pp. 187–190.

Roepstorff, P. and Fohlman, J. (1984) 'Proposal for a common nomenclature for sequence ions in mass spectra of peptides.', *Biomed Mass Spectrom*, 11(11), p. 601.

Rudnick, P. A. *et al.* (2010) *Proteome Informatics Research Group 2010 Study*.

Ruprecht, B. *et al.* (2016) 'MALDI-TOF and nESI Orbitrap MS/MS identify orthogonal parts of the phosphoproteome.', *Proteomics*, 16(10), pp. 1447–56.

Saeed, F. *et al.* (2012) 'An Efficient Dynamic Programming Algorithm for Phosphorylation Site Assignment of Large-Scale Mass Spectrometry Data', *Proceedings (IEEE Int Conf Bioinformatics Biomed)*, pp. 618–625.

Saeed, F. *et al.* (2013) 'PhosSA: Fast and accurate phosphorylation site assignment algorithm for mass spectrometry data', *Proteome Sci*, 11(Suppl 1), p. S14.

Samuelsson, J. *et al.* (2004) 'Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting', *Bioinformatics*, 20(18), pp. 3628–3635.

Savitski, M. M. *et al.* (2011) 'Confident phosphorylation site localization using the Mascot Delta Score', *Mol Cell Proteomics*, 10(2), p. M110.003830.

Searle, B. C., Turner, M. and Nesvizhskii, A. I. (2008) 'Improving Sensitivity by Probabilistically Combining Results from Multiple MS/MS Search Methodologies', *Journal of Proteome Research*, 7(1), pp. 245–253.

Shao, W. and Lam, H. (2017) 'Tandem mass spectral libraries of peptides and their roles in proteomics research', *Mass Spectrometry Reviews*, 36(5), pp. 634–648.

Shao, W., Zhu, K. and Lam, H. (2013) 'Refining similarity scoring to enable decoy-free validation in spectral library searching', *Proteomics*, 13(22), pp. 3273–3283.

Shteynberg, D. *et al.* (2013) 'Combining Results of Multiple Search Engines in Proteomics', *Molecular & Cellular Proteomics*, 12(9), pp. 2383–2393.

Suckau, D. *et al.* (2003) 'A novel MALDI LIFT-TOF/TOF mass spectrometer for proteomics', *Analytical and Bioanalytical Chemistry*, 376(7), pp. 952–965.

Syka, J. E. P. *et al.* (2004) 'Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry', *Proceedings of the National Academy of Sciences*, 101(26), pp. 9528–9533.

Tanaka, K. *et al.* (1988) 'Protein and polymer analyses up tom/z 100 000 by laser ionization time-of-flight mass spectrometry', *Rapid Communications in Mass Spectrometry*, 2(8), pp. 151–153.

Taus, T. *et al.* (2011) 'Universal and confident phosphorylation site localization using phosphoRS.', *Journal of proteome research*, 10(12), pp. 5354–5362.

Vaga, S. *et al.* (2014) 'Phosphoproteomic analyses reveal novel cross-modulation mechanisms between two signaling pathways in yeast', *Mol Syst Biol*, 10, p. 767.

Vaudel, M. *et al.* (2011) 'SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches', *Proteomics*, 11(5), pp. 996–999.

Vaudel, M. *et al.* (2015) 'PeptideShaker enables reanalysis of MS-derived proteomics data sets', *Nature*

Biotechnology, 33(1), pp. 22–24.

Vizcaíno, J. A. *et al.* (2017) 'The mzIdentML Data Standard Version 1.2, Supporting Advances in Proteome Informatics.', *Molecular & cellular proteomics*, 16(7), pp. 1275–1285.

Weickhardt, C., Moritz, F. and Grotemeyer, J. (1996) 'Time-of-flight mass spectrometry: State-of the-art in chemical analysis and molecular science', *Mass Spectrometry Reviews*, 15(3), pp. 139–162.

Whitehouse, C. M. *et al.* (1985) 'Electrospray interface for liquid chromatographs and mass spectrometers.', *Analytical chemistry*, 57(3), pp. 675–9.

Wiese, H. *et al.* (2014) 'Comparison of alternative MS/MS and bioinformatics approaches for confident phosphorylation site localization', *Journal of proteome research*, 13(2), pp. 1128–1137.

Wilhelm, M. *et al.* (2014) 'Mass-spectrometry-based draft of the human proteome', *Nature*, 509(7502), pp. 582–587.

Xiao, K. *et al.* (2017) 'Accurate phosphorylation site localization using phospho-brackets', *Analytica Chimica Acta*, 996, pp. 38–47.

Xu, G. *et al.* (2018) 'Deconvolution in mass spectrometry based proteomics', *Rapid Communications in Mass Spectrometry*, 32(10), pp. 763–774.

Yates, J. R. *et al.* (1993) 'Peptide Mass Maps: A Highly Informative Approach to Protein Identification', *Analytical Biochemistry*, 214(2), pp. 397–408.

Ye, D. *et al.* (2010) 'Open MS/MS spectral library search to identify unanticipated post-translational modifications and increase spectral identification rate.', *Bioinformatics*, 26(12), pp. i399-406.

Zhang, J. *et al.* (2012) 'PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification.', *Molecular & cellular proteomics*, 11(4), p. M111.010587.

Zhang, W. and Chait, B. T. (2000) 'ProFound: an expert system for protein identification using mass spectrometric peptide mapping information.', *Analytical chemistry*, 72(11), pp. 2482–9.

Zhang, X. *et al.* (2011) 'Understanding the improved sensitivity of spectral library searching over sequence database searching in proteomics data analysis', *Proteomics*, 11(6), pp. 1075–1085.

Zubarev, R. A., Kelleher, N. L. and McLafferty, F. W. (1998) 'Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process', *Journal of the American Chemical Society*, 120(13), pp. 3265–3266.

Zubarev, R. A., Zubarev, A. R. and Savitski, M. M. (2008) 'Electron Capture/Transfer versus Collisionally Activated/Induced Dissociations: Solo or Duet?', *Journal of the American Society for Mass Spectrometry*, 19(6), pp. 753–761.

Turku Centre for Computer Science
TUCS Dissertations

1. **Marjo Lipponen**, On Primitive Solutions of the Post Correspondence Problem
2. **Timo Käkölä**, Dual Information Systems in Hyperknowledge Organizations
3. **Ville Leppänen**, Studies on the Realization of PRAM
4. **Cunsheng Ding**, Cryptographic Counter Generators
5. **Sami Viitanen**, Some New Global Optimization Algorithms
6. **Tapio Salakoski**, Representative Classification of Protein Structures
7. **Thomas Långbacka**, An Interactive Environment Supporting the Development of Formally Correct Programs
8. **Thomas Finne**, A Decision Support System for Improving Information Security
9. **Valeria Mihalache**, Cooperation, Communication, Control. Investigations on Grammar Systems.
10. **Marina Waldén**, Formal Reasoning About Distributed Algorithms
11. **Tero Laihonen**, Estimates on the Covering Radius When the Dual Distance is Known
12. **Lucian Ilie**, Decision Problems on Orders of Words
13. **Jukkapekka Hekanaho**, An Evolutionary Approach to Concept Learning
14. **Jouni Järvinen**, Knowledge Representation and Rough Sets
15. **Tomi Pasanen**, In-Place Algorithms for Sorting Problems
16. **Mika Johnsson**, Operational and Tactical Level Optimization in Printed Circuit Board Assembly
17. **Mats Aspnäs**, Multiprocessor Architecture and Programming: The Hathi-2 System
18. **Anna Mikhajlova**, Ensuring Correctness of Object and Component Systems
19. **Vesa Torvinen**, Construction and Evaluation of the Labour Game Method
20. **Jorma Boberg**, Cluster Analysis. A Mathematical Approach with Applications to Protein Structures
21. **Leonid Mikhajlov**, Software Reuse Mechanisms and Techniques: Safety Versus Flexibility
22. **Timo Kaukoranta**, Iterative and Hierarchical Methods for Codebook Generation in Vector Quantization
23. **Gábor Magyar**, On Solution Approaches for Some Industrially Motivated Combinatorial Optimization Problems
24. **Linas Laibinis**, Mechanised Formal Reasoning About Modular Programs
25. **Shuhua Liu**, Improving Executive Support in Strategic Scanning with Software Agent Systems
26. **Jaakko Järvi**, New Techniques in Generic Programming – C++ is more Intentional than Intended
27. **Jan-Christian Lehtinen**, Reproducing Kernel Splines in the Analysis of Medical Data
28. **Martin Büchi**, Safe Language Mechanisms for Modularization and Concurrency
29. **Elena Troubitsyna**, Stepwise Development of Dependable Systems
30. **Janne Näppi**, Computer-Assisted Diagnosis of Breast Calcifications
31. **Jianming Liang**, Dynamic Chest Images Analysis
32. **Tiberiu Seceleanu**, Systematic Design of Synchronous Digital Circuits
33. **Tero Aittokallio**, Characterization and Modelling of the Cardiorespiratory System in Sleep-Disordered Breathing
34. **Ivan Porres**, Modeling and Analyzing Software Behavior in UML
35. **Mauno Rönkkö**, Stepwise Development of Hybrid Systems
36. **Jouni Smed**, Production Planning in Printed Circuit Board Assembly
37. **Vesa Halava**, The Post Correspondence Problem for Market Morphisms
38. **Ion Petre**, Commutation Problems on Sets of Words and Formal Power Series
39. **Vladimir Kvassov**, Information Technology and the Productivity of Managerial Work
40. **Frank Tétard**, Managers, Fragmentation of Working Time, and Information Systems

41. **Jan Manuch**, Defect Theorems and Infinite Words
42. **Kalle Ranto**, Z_4-Goethals Codes, Decoding and Designs
43. **Arto Lepistö**, On Relations Between Local and Global Periodicity
44. **Mika Hirvensalo**, Studies on Boolean Functions Related to Quantum Computing
45. **Pentti Virtanen**, Measuring and Improving Component-Based Software Development
46. **Adekunle Okunoye**, Knowledge Management and Global Diversity – A Framework to Support Organisations in Developing Countries
47. **Antonina Kloptchenko**, Text Mining Based on the Prototype Matching Method
48. **Juha Kivijärvi**, Optimization Methods for Clustering
49. **Rimvydas Rukšėnas**, Formal Development of Concurrent Components
50. **Dirk Nowotka**, Periodicity and Unbordered Factors of Words
51. **Attila Gyenesei**, Discovering Frequent Fuzzy Patterns in Relations of Quantitative Attributes
52. **Petteri Kaitovaara**, Packaging of IT Services – Conceptual and Empirical Studies
53. **Petri Rosendahl**, Niho Type Cross-Correlation Functions and Related Equations
54. **Péter Majlender**, A Normative Approach to Possibility Theory and Soft Decision Support
55. **Seppo Virtanen**, A Framework for Rapid Design and Evaluation of Protocol Processors
56. **Tomas Eklund**, The Self-Organizing Map in Financial Benchmarking
57. **Mikael Collan**, Giga-Investments: Modelling the Valuation of Very Large Industrial Real Investments
58. **Dag Björklund**, A Kernel Language for Unified Code Synthesis
59. **Shengnan Han**, Understanding User Adoption of Mobile Technology: Focusing on Physicians in Finland
60. **Irina Georgescu**, Rational Choice and Revealed Preference: A Fuzzy Approach
61. **Ping Yan**, Limit Cycles for Generalized Liénard-Type and Lotka-Volterra Systems
62. **Joonas Lehtinen**, Coding of Wavelet-Transformed Images
63. **Tommi Meskanen,** On the NTRU Cryptosystem
64. **Saeed Salehi**, Varieties of Tree Languages
65. **Jukka Arvo**, Efficient Algorithms for Hardware-Accelerated Shadow Computation
66. **Mika Hirvikorpi**, On the Tactical Level Production Planning in Flexible Manufacturing Systems
67. **Adrian Costea**, Computational Intelligence Methods for Quantitative Data Mining
68. **Cristina Seceleanu**, A Methodology for Constructing Correct Reactive Systems
69. **Luigia Petre**, Modeling with Action Systems
70. **Lu Yan**, Systematic Design of Ubiquitous Systems
71. **Mehran Gomari**, On the Generalization Ability of Bayesian Neural Networks
72. **Ville Harkke**, Knowledge Freedom for Medical Professionals – An Evaluation Study of a Mobile Information System for Physicians in Finland
73. **Marius Cosmin Codrea**, Pattern Analysis of Chlorophyll Fluorescence Signals
74. **Aiying Rong**, Cogeneration Planning Under the Deregulated Power Market and Emissions Trading Scheme
75. **Chihab BenMoussa**, Supporting the Sales Force through Mobile Information and Communication Technologies: Focusing on the Pharmaceutical Sales Force
76. **Jussi Salmi**, Improving Data Analysis in Proteomics
77. **Orieta Celiku**, Mechanized Reasoning for Dually-Nondeterministic and Probabilistic Programs
78. **Kaj-Mikael Björk**, Supply Chain Efficiency with Some Forest Industry Improvements
79. **Viorel Preoteasa**, Program Variables – The Core of Mechanical Reasoning about Imperative Programs
80. **Jonne Poikonen**, Absolute Value Extraction and Order Statistic Filtering for a Mixed-Mode Array Image Processor
81. **Luka Milovanov**, Agile Software Development in an Academic Environment
82. **Francisco Augusto Alcaraz Garcia**, Real Options, Default Risk and Soft Applications
83. **Kai K. Kimppa**, Problems with the Justification of Intellectual Property Rights in Relation to Software and Other Digitally Distributable Media
84. **Dragoş Truşcan**, Model Driven Development of Programmable Architectures
85. **Eugen Czeizler**, The Inverse Neighborhood Problem and Applications of Welch Sets in Automata Theory

86. **Sanna Ranto**, Identifying and Locating-Dominating Codes in Binary Hamming Spaces
87. **Tuomas Hakkarainen**, On the Computation of the Class Numbers of Real Abelian Fields
88. **Elena Czeizler**, Intricacies of Word Equations
89. **Marcus Alanen**, A Metamodeling Framework for Software Engineering
90. **Filip Ginter**, Towards Information Extraction in the Biomedical Domain: Methods and Resources
91. **Jarkko Paavola**, Signature Ensembles and Receiver Structures for Oversaturated Synchronous DS-CDMA Systems
92. **Arho Virkki**, The Human Respiratory System: Modelling, Analysis and Control
93. **Olli Luoma**, Efficient Methods for Storing and Querying XML Data with Relational Databases
94. **Dubravka Ilić**, Formal Reasoning about Dependability in Model-Driven Development
95. **Kim Solin**, Abstract Algebra of Program Refinement
96. **Tomi Westerlund**, Time Aware Modelling and Analysis of Systems-on-Chip
97. **Kalle Saari**, On the Frequency and Periodicity of Infinite Words
98. **Tomi Kärki**, Similarity Relations on Words: Relational Codes and Periods
99. **Markus M. Mäkelä**, Essays on Software Product Development: A Strategic Management Viewpoint
100. **Roope Vehkalahti**, Class Field Theoretic Methods in the Design of Lattice Signal Constellations
101. **Anne-Maria Ernvall-Hytönen**, On Short Exponential Sums Involving Fourier Coefficients of Holomorphic Cusp Forms
102. **Chang Li**, Parallelism and Complexity in Gene Assembly
103. **Tapio Pahikkala**, New Kernel Functions and Learning Methods for Text and Data Mining
104. **Denis Shestakov**, Search Interfaces on the Web: Querying and Characterizing
105. **Sampo Pyysalo**, A Dependency Parsing Approach to Biomedical Text Mining
106. **Anna Sell**, Mobile Digital Calendars in Knowledge Work
107. **Dorina Marghescu**, Evaluating Multidimensional Visualization Techniques in Data Mining Tasks
108. **Tero Säntti**, A Co-Processor Approach for Efficient Java Execution in Embedded Systems
109. **Kari Salonen**, Setup Optimization in High-Mix Surface Mount PCB Assembly
110. **Pontus Boström**, Formal Design and Verification of Systems Using Domain-Specific Languages
111. **Camilla J. Hollanti**, Order-Theoretic Mehtods for Space-Time Coding: Symmetric and Asymmetric Designs
112. **Heidi Himmanen**, On Transmission System Design for Wireless Broadcasting
113. **Sébastien Lafond**, Simulation of Embedded Systems for Energy Consumption Estimation
114. **Evgeni Tsivtsivadze**, Learning Preferences with Kernel-Based Methods
115. **Petri Salmela**, On Commutation and Conjugacy of Rational Languages and the Fixed Point Method
116. **Siamak Taati**, Conservation Laws in Cellular Automata
117. **Vladimir Rogojin**, Gene Assembly in Stichotrichous Ciliates: Elementary Operations, Parallelism and Computation
118. **Alexey Dudkov**, Chip and Signature Interleaving in DS CDMA Systems
119. **Janne Savela**, Role of Selected Spectral Attributes in the Perception of Synthetic Vowels
120. **Kristian Nybom**, Low-Density Parity-Check Codes for Wireless Datacast Networks
121. **Johanna Tuominen**, Formal Power Analysis of Systems-on-Chip
122. **Teijo Lehtonen**, On Fault Tolerance Methods for Networks-on-Chip
123. **Eeva Suvitie**, On Inner Products Involving Holomorphic Cusp Forms and Maass Forms
124. **Linda Mannila**, Teaching Mathematics and Programming – New Approaches with Empirical Evaluation
125. **Hanna Suominen**, Machine Learning and Clinical Text: Supporting Health Information Flow
126. **Tuomo Saarni**, Segmental Durations of Speech
127. **Johannes Eriksson**, Tool-Supported Invariant-Based Programming

128. **Tero Jokela**, Design and Analysis of Forward Error Control Coding and Signaling for Guaranteeing QoS in Wireless Broadcast Systems
129. **Ville Lukkarila**, On Undecidable Dynamical Properties of Reversible One-Dimensional Cellular Automata
130. **Qaisar Ahmad Malik**, Combining Model-Based Testing and Stepwise Formal Development
131. **Mikko-Jussi Laakso**, Promoting Programming Learning: Engagement, Automatic Assessment with Immediate Feedback in Visualizations
132. **Riikka Vuokko**, A Practice Perspective on Organizational Implementation of Information Technology
133. **Jeanette Heidenberg**, Towards Increased Productivity and Quality in Software Development Using Agile, Lean and Collaborative Approaches
134. **Yong Liu**, Solving the Puzzle of Mobile Learning Adoption
135. **Stina Ojala**, Towards an Integrative Information Society: Studies on Individuality in Speech and Sign
136. **Matteo Brunelli**, Some Advances in Mathematical Models for Preference Relations
137. **Ville Junnila**, On Identifying and Locating-Dominating Codes
138. **Andrzej Mizera**, Methods for Construction and Analysis of Computational Models in Systems Biology. Applications to the Modelling of the Heat Shock Response and the Self-Assembly of Intermediate Filaments.
139. **Csaba Ráduly-Baka**, Algorithmic Solutions for Combinatorial Problems in Resource Management of Manufacturing Environments
140. **Jari Kyngäs**, Solving Challenging Real-World Scheduling Problems
141. **Arho Suominen**, Notes on Emerging Technologies
142. **József Mezei**, A Quantitative View on Fuzzy Numbers
143. **Marta Olszewska**, On the Impact of Rigorous Approaches on the Quality of Development
144. **Antti Airola**, Kernel-Based Ranking: Methods for Learning and Performace Estimation
145. **Aleksi Saarela**, Word Equations and Related Topics: Independence, Decidability and Characterizations
146. **Lasse Bergroth**, Kahden merkkijonon pisimmän yhteisen alijonon ongelma ja sen ratkaiseminen
147. **Thomas Canhao Xu**, Hardware/Software Co-Design for Multicore Architectures
148. **Tuomas Mäkilä**, Software Development Process Modeling – Developers Perspective to Contemporary Modeling Techniques
149. **Shahrokh Nikou**, Opening the Black-Box of IT Artifacts: Looking into Mobile Service Characteristics and Individual Perception
150. **Alessandro Buoni**, Fraud Detection in the Banking Sector: A Multi-Agent Approach
151. **Mats Neovius**, Trustworthy Context Dependency in Ubiquitous Systems
152. **Fredrik Degerlund**, Scheduling of Guarded Command Based Models
153. **Amir-Mohammad Rahmani-Sane**, Exploration and Design of Power-Efficient Networked Many-Core Systems
154. **Ville Rantala**, On Dynamic Monitoring Methods for Networks-on-Chip
155. **Mikko Pelto**, On Identifying and Locating-Dominating Codes in the Infinite King Grid
156. **Anton Tarasyuk**, Formal Development and Quantitative Verification of Dependable Systems
157. **Muhammad Mohsin Saleemi**, Towards Combining Interactive Mobile TV and Smart Spaces: Architectures, Tools and Application Development
158. **Tommi J. M. Lehtinen**, Numbers and Languages
159. **Peter Sarlin**, Mapping Financial Stability
160. **Alexander Wei Yin**, On Energy Efficient Computing Platforms
161. **Mikołaj Olszewski**, Scaling Up Stepwise Feature Introduction to Construction of Large Software Systems
162. **Maryam Kamali**, Reusable Formal Architectures for Networked Systems
163. **Zhiyuan Yao**, Visual Customer Segmentation and Behavior Analysis – A SOM-Based Approach
164. **Timo Jolivet**, Combinatorics of Pisot Substitutions
165. **Rajeev Kumar Kanth**, Analysis and Life Cycle Assessment of Printed Antennas for Sustainable Wireless Systems
166. **Khalid Latif**, Design Space Exploration for MPSoC Architectures

167. **Bo Yang**, Towards Optimal Application Mapping for Energy-Efficient Many-Core Platforms
168. **Ali Hanzala Khan**, Consistency of UML Based Designs Using Ontology Reasoners
169. **Sonja Leskinen**, m-Equine: IS Support for the Horse Industry
170. **Fareed Ahmed Jokhio**, Video Transcoding in a Distributed Cloud Computing Environment
171. **Moazzam Fareed Niazi**, A Model-Based Development and Verification Framework for Distributed System-on-Chip Architecture
172. **Mari Huova**, Combinatorics on Words: New Aspects on Avoidability, Defect Effect, Equations and Palindromes
173. **Ville Timonen**, Scalable Algorithms for Height Field Illumination
174. **Henri Korvela**, Virtual Communities – A Virtual Treasure Trove for End-User Developers
175. **Kameswar Rao Vaddina**, Thermal-Aware Networked Many-Core Systems
176. **Janne Lahtiranta**, New and Emerging Challenges of the ICT-Mediated Health and Well-Being Services
177. **Irum Rauf**, Design and Validation of Stateful Composite RESTful Web Services
178. **Jari Björne**, Biomedical Event Extraction with Machine Learning
179. **Katri Haverinen**, Natural Language Processing Resources for Finnish: Corpus Development in the General and Clinical Domains
180. **Ville Salo**, Subshifts with Simple Cellular Automata
181. **Johan Ersfolk**, Scheduling Dynamic Dataflow Graphs
182. **Hongyan Liu**, On Advancing Business Intelligence in the Electricity Retail Market
183. **Adnan Ashraf**, Cost-Efficient Virtual Machine Management: Provisioning, Admission Control, and Consolidation
184. **Muhammad Nazrul Islam**, Design and Evaluation of Web Interface Signs to Improve Web Usability: A Semiotic Framework
185. **Johannes Tuikkala**, Algorithmic Techniques in Gene Expression Processing: From Imputation to Visualization
186. **Natalia Díaz Rodríguez**, Semantic and Fuzzy Modelling for Human Behaviour Recognition in Smart Spaces. A Case Study on Ambient Assisted Living
187. **Mikko Pänkäälä**, Potential and Challenges of Analog Reconfigurable Computation in Modern and Future CMOS
188. **Sami Hyrynsalmi**, Letters from the War of Ecosystems – An Analysis of Independent Software Vendors in Mobile Application Marketplaces
189. **Seppo Pulkkinen**, Efficient Optimization Algorithms for Nonlinear Data Analysis
190. **Sami Pyöttiälä**, Optimization and Measuring Techniques for Collect-and-Place Machines in Printed Circuit Board Industry
191. **Syed Mohammad Asad Hassan Jafri**, Virtual Runtime Application Partitions for Resource Management in Massively Parallel Architectures
192. **Toni Ernvall**, On Distributed Storage Codes
193. **Yuliya Prokhorova**, Rigorous Development of Safety-Critical Systems
194. **Olli Lahdenoja**, Local Binary Patterns in Focal-Plane Processing – Analysis and Applications
195. **Annika H. Holmbom**, Visual Analytics for Behavioral and Niche Market Segmentation
196. **Sergey Ostroumov**, Agent-Based Management System for Many-Core Platforms: Rigorous Design and Efficient Implementation
197. **Espen Suenson**, How Computer Programmers Work – Understanding Software Development in Practise
198. **Tuomas Poikela**, Readout Architectures for Hybrid Pixel Detector Readout Chips
199. **Bogdan Iancu**, Quantitative Refinement of Reaction-Based Biomodels
200. **Ilkka Törmä**, Structural and Computational Existence Results for Multidimensional Subshifts
201. **Sebastian Okser**, Scalable Feature Selection Applications for Genome-Wide Association Studies of Complex Diseases
202. **Fredrik Abbors**, Model-Based Testing of Software Systems: Functionality and Performance
203. **Inna Pereverzeva**, Formal Development of Resilient Distributed Systems
204. **Mikhail Barash**, Defining Contexts in Context-Free Grammars
205. **Sepinoud Azimi**, Computational Models for and from Biology: Simple Gene Assembly and Reaction Systems
206. **Petter Sandvik**, Formal Modelling for Digital Media Distribution

207. **Jongyun Moon**, Hydrogen Sensor Application of Anodic Titanium Oxide Nanostructures
208. **Simon Holmbacka**, Energy Aware Software for Many-Core Systems
209. **Charalampos Zinoviadis**, Hierarchy and Expansiveness in Two-Dimensional Subshifts of Finite Type
210. **Mika Murtojärvi**, Efficient Algorithms for Coastal Geographic Problems
211. **Sami Mäkelä**, Cohesion Metrics for Improving Software Quality
212. **Eyal Eshet**, Examining Human-Centered Design Practice in the Mobile Apps Era
213. **Jetro Vesti**, Rich Words and Balanced Words
214. **Jarkko Peltomäki**, Privileged Words and Sturmian Words
215. **Fahimeh Farahnakian**, Energy and Performance Management of Virtual Machines: Provisioning, Placement and Consolidation
216. **Diana-Elena Gratie**, Refinement of Biomodels Using Petri Nets
217. **Harri Merisaari**, Algorithmic Analysis Techniques for Molecular Imaging
218. **Stefan Grönroos**, Efficient and Low-Cost Software Defined Radio on Commodity Hardware
219. **Noora Nieminen**, Garbling Schemes and Applications
220. **Ville Taajamaa**, O-CDIO: Engineering Education Framework with Embedded Design Thinking Methods
221. **Johannes Holvitie**, Technical Debt in Software Development – Examining Premises and Overcoming Implementation for Efficient Management
222. **Tewodros Deneke**, Proactive Management of Video Transcoding Services
223. **Kashif Javed**, Model-Driven Development and Verification of Fault Tolerant Systems
224. **Pekka Naula**, Sparse Predictive Modeling – A Cost-Effective Perspective
225. **Antti Hakkala**, On Security and Privacy for Networked Information Society – Observations and Solutions for Security Engineering and Trust Building in Advanced Societal Processes
226. **Anne-Maarit Majanoja**, Selective Outsourcing in Global IT Services – Operational Level Challenges and Opportunities
227. **Samuel Rönnqvist**, Knowledge-Lean Text Mining
228. **Mohammad-Hashem Hahgbayan**, Energy-Efficient and Reliable Computing in Dark Silicon Era
229. **Charmi Panchal**, Qualitative Methods for Modeling Biochemical Systems and Datasets: The Logicome and the Reaction Systems Approaches
230. **Erkki Kaila**, Utilizing Educational Technology in Computer Science and Programming Courses: Theory and Practice
231. **Fredrik Robertsén**, The Lattice Boltzmann Method, a Petaflop and Beyond
232. **Jonne Pohjankukka**, Machine Learning Approaches for Natural Resource Data
233. **Paavo Nevalainen**, Geometric Data Understanding: Deriving Case-Specific Features
234. **Michal Szabados**, An Algebraic Approach to Nivat's Conjecture
235. **Tuan Nguyen Gia**, Design for Energy-Efficient and Reliable Fog-Assisted Healthcare IoT Systems
236. **Anil Kanduri**, Adaptive Knobs for Resource Efficient Computing
237. **Veronika Suni**, Computational Methods and Tools for Protein Phosphorylation Analysis

Turku Centre *for* Computer Science

http://www. tucs.fi

tucs@abo.fi

University of Turku
Faculty of Science and Engineering
- Department of Future Technologies
- Department of Mathematics and Statistics

Turku School of Economics
- Institute of Information Systems Science

Åbo Akademi University
Faculty of Science and Engineering
- Computer Engineering
- Computer Science

Faculty of Social Sciences, Business and Economics
- Information Systems

ISBN 978-952-12-3796-6
ISSN 1239-1883