Middle-down approaches for mass spectrometry-based protein identification and characterization

THÈSE N^{O} 6918 (2016)

PRÉSENTÉE LE 12 FÉVRIER 2016
À LA FACULTÉ DES SCIENCES DE BASE
LABORATOIRE DE SPECTROMÉTRIE DE MASSE BIOMOLÉCULES
PROGRAMME DOCTORAL EN CHIMIE ET GÉNIE CHIMIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Kristina SRZENTIĆ

acceptée sur proposition du jury:

Prof. P. J. Dyson, président du jury
Dr M. Kussmann, Dr Y. Tsybin, directeurs de thèse
Dr C. Masselon, rapporteur
Prof. P. Picotti, rapporteuse
Prof. H. H. Girault, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2016

‘Happy is he who gets to know the reasons for things’

Virgil

To my mother and the loving memory of my grandmother

Your endless love and support made me the person I am today, and for that I am forever grateful

Mojoj majci i uspomeni na moju voljenu baku

Vaša beskrajna ljubav i podrška su me učinile osobom koja sam danas, i na tome sam vam vječno zahvalna

Table of contents.

Abstract. 12

Riassunto in italiano 14

List of papers 16

List of abbreviations. 20

Chapter 1. Introduction. 23

1.1. Studying biology with analytical techniques, or coupling uncertainties. 25

1.2. From genome to proteome - Era of proteomics 26

1.3. Mass spectrometry (MS): a magnifying glass into proteome complexity? 29

1.3.1. Proteome complexity: from gene to proteoform. 30

1.3.2. The protein abundance dynamic range in cells and tissues. 30

1.4. Classical MS approaches: bottom-up *vs* top-down 32

1.5. Middle-down (MD) MS approach 34

1.6. Selected application: Analysis of Immunoglobulins G (IgGs) 35

1.7. Aim of the thesis 40

1.8. Overview of Chapters 41

Chapter 2. Experimental methods 43

2.1. MS fundamentals: what to understand and how to apply it in proteomics analysis 45

2.1.1. Mass accuracy and mass resolution 45

2.1.2. Isotopes and their role in mass spectrometry of polypeptides 46

2.1.3. Signal to noise ratio, spectral dynamic range 49

2.2. Tandem mass spectrometry MS/MS 52

2.3. Ion activation and dissociation: Energy threshold-based activation 54

2.4. Ion activation and dissociation: Electron-based activation methods 57

2.5. Where to cleave a protein? 59

2.6. High-throughput *vs* targeted analysis in proteomics 61

2.7. Data analysis 63

Chapter 3. Instrumentation 67

3.1. Electrospray ionization 69

3.2. Hybrid mass spectrometers 71

3.3. Fourier transform mass spectrometry (FTMS) 74

3.3.1. Orbitrap-based mass spectrometer - design and working principle 76

3.4. Limitations of MS experiment in an Orbitrap FT MS 78

Chapter 4. Towards a 'stand-alone' MDP pipeline 81

4.1. Liquid chromatography: adopted parameters and reflections on mass spectrometric analysis of large peptides. 85

4.2. Evaluation of MS/MS parameters for MDP 90

4.3. Paper I: Proteome digestion specificity analysis for rational design of extended bottom-up and middle-down proteomics experiments 93

4.4. Paper II: Practical consideration for improving the productivity of mass spectrometry-based proteomics 115

Chapter 5. Enabling MDP proteolysis: a quest for a novel protease 123

5.1. Paper III: Extended bottom-up proteomics with secreted aspartic protease Sap9 127

5.2. Paper IV: Advantages of extended bottom-up proteomics using Sap9 for analysis of monoclonal antibodies 149

Chapter 6. An alternative to enzymatic digestion: Chemoselective MD approach 171

6.1. Paper V: Chemical-mediated digestion: an alternative realm for middle-down mass spectrometry 175

6.2. MD with Cysteine-based chemical digestion: targeted analysis of Immunoglobulins G 213

Chapter 7. Towards top-down MS *via* MD 219

7.1. Paper VI: Middle-down analysis of IgG mixtures using electron transfer dissociation allows light and heavy chain pairing characterization 223

7.2. Paper VII: Revealing chain connectivity in monoclonal IgG1 using GingisKHAN proteolysis and top-down electron transfer dissociation Orbitrap FTMS 251

Chapter 8. Conclusions 285

References 295

Acknowledgments 307

Curriculum Vitae 311

Abstract.

Mass spectrometry (MS) has emerged over the last two decades as the analytical technique of choice in systems-level protein studies, known as proteomics. Two are the MS-based approaches generally applied to proteomics: bottom-up (BU), which relies on the proteolytic digestion of proteins into short (~10 amino acids) peptides, and top-down (TD), where proteolysis is omitted, intact proteins are detected and fragmented in gas phase. Both methods present advantages as well as drawbacks. Here, we sought to establish a complete platform to put forward a third MS-based proteomic approach; middle-down (MD). It implies protein digestion as in BU, but aims to generate large peptides which size approaches the one of small intact proteins that are readily analyzed in TD. This novel domain aims to account for the shortcoming of both classical approaches. Until now, the main reasons behind the limited use of MD proteomics have been the lack of easy-to-use cleaving agents capable of producing peptides in the desired 3-15 kDa mass range with high specificity, limitations in MS and allied instrumentation, and the absence of dedicated bioinformatics tools for processing of acquired data. The latter greatly impedes the next milestone in MD proteomics – large-scale analysis. MD can potentially combine the analysis of large portions of proteins carrying set of biologically-relevant modifications – allowing exploring proteins of molecular weight or complexity incompatible with current TD capabilities – with the high-throughput hallmark of BU proteomics. Here, we first *in silico* evaluated the potential target amino acid residues to produce peptides in the MD mass range within the proteomes of different organisms. This bioinformatics work was followed by an experimental study based on synthetic MD-sized peptides, aimed at determining the optimal MS and tandem MS parameters for large peptide characterization. Next, we pursued two distinct ways of performing MD proteolysis: i) the use of an enzyme and ii) the use of a chemical reagent. We selected the protease Sap9 as a target enzyme for MD, which we fully characterized and successfully applied to the study of a mixture of monoclonal antibodies, where it showed an advantage over traditional BU in terms of reduced introduction of artifacts to the sample, allowing the post-translational modification investigation and unambiguous antibody identification. The chemical cleavage way we addressed *via* judicious protocol optimization for hydrolysis at the N-terminal side of cysteine with NTCB reagent. We also advanced MD protocols by generation of large (~50 kDa) subunits of monoclonal antibodies through the use of papain and another more specific novel protease, GingisKHAN, combined with new MS signal processing and data analysis capabilities. The developed workflow improved mapping of the connectivity of cysteines involved in inter- and intra-molecular disulfide bridges in antibodies. To summarize, we demonstrated that MD approach to mass spectrometry and proteomics is a powerful, yet underdeveloped, complement to BU and TD. This work has benchmarked MD for targeted protein analysis. In the near future, with advancements of the field, we envision its growing use for large-scale complex proteome analysis.

Keywords: mass spectrometry, MS; tandem mass spectrometry, MS/MS; middle-down, MD; Fourier transform mass spectrometry, FTMS; Orbitrap; protein; immunoglobulin G, IgG; chemical digestion; proteolysis.

Riassunto in italiano

La spettrometria di massa (MS) è assurta nelle ultime due decadi al ruolo di tecnologia d'eccellenza in studi di proteine a livelli sistemici, noti come proteomica. Due sono gli approcci generalmente impiegati in proteomica: bottom-up (BU), che si basa sulla digestione delle proteine in peptidi più corti (~10 amminoacidi), e top-down (TD), in cui la proteolisi viene omessa, e le proteine sono analizzate e poi frammentate ancora intere in fase gassosa. Entrambi i metodi presentano sia vantaggi che limiti. Attraverso il lavoro di questa Tesi abbiamo cercato di definire in modo completo una piattaforma per avanzzare l'applicazione di una terzo approccio di proteomica basata su spettrometria di massa: il middle-down (MD). MD può concettualmente essere pensato come un ibrido delle due metodologie tradizionali, BU e TD. Esso, infatti, implica la digestione delle proteine come nel caso del BU, ma allo scopo di generare peptidi lunghi, la cui massa rasenti quella di piccole proteine intatte che sono quelle piu' facilmente analizzabili nel TD. Questa nuova strategia ambisce a rispondere dei limiti di entrambi gli approcci classici. Le ragioni principali alla base della sinora ridotta diffusione della proteomica MD sono da ricercarsi nella mancanza di un agente per la digestione proteolitica di facile utilizzo e che fosse in grado di produrre con alevata specificità peptidi nell'intervallo di massa desiderato, tra 3 e 15 kDa, oltre a limiti nella strumentazione, tanto di MS che di tecnologie complementari, ed infine nell'assenza di strumenti bioinformatici dedicati all'analisi di dati MD. Quest'ultimo ostacolo tutt'ora impedisce in particolare il raggiungimento del prossimo traguardo nella proteomica MD, ossia la realizzazione di analisi su larga scala.L'MD può potenzialmente combinare l'analisi di ampie porzioni di proteina corredate da specifici gruppi di modifiche (genetiche o chimiche) di rilevanza biologica – acconsentendo dunque di esplorare proteine di peso molecolare o complessita' incompatibili con l'attuale livello raggiunto dal TD – con le caratteristiche high-throughput del bottom-up. Pertanto, abbiamo dapprima effettuato una ricerca *in silico* sui potenziali target amminoacidici utili a produrre peptidi nell'intervallo di massa tipico dell'MD, considerando i proteomi di diversi organismi. Poi, a questa ricerca bioinformatica è seguito

uno studio sperimentale basato su peptide sintetici di dimensioni compatibili con l'MD, mirato a determinare i parametri ottimali di MS ed MS/MS da impiegare nella caratterizzazione di peptidi di grandi dimensioni. Successivamente abbiamo percorso due vie distinte per ottenere una proteolisi per l'MD: i) attraverso l'uso di un enzima, e ii) con un reagente chimico. Abbiamo scelto come enzima per l'MD la proteasi Sap9, che abbiamo caratterizzato in modo completo e utilizzato con successo nello studio di una miscela di anticorpi monoclonali, applicazione nella quale ha dimostrato un vantaggio rispetto al tradizionale BU in termini di ridotta introduzione di artefatti nel campione, acconsentendo contemporaneamente alla mappatura di modifiche post-traduzionali e alla identificazione inequivocabile dei singoli anticorpi.Parallelamente, abbiamo implementato la modalità di digestione chimica attraverso una attenta ottimizzazione del protocollo di idrolisi del legame peptidico al lato N-terminale della cisteina attraverso il reagente NTCB. Sono stati infine migliorati anche i protocolli per MD destinati alla produzione di grandi subunità (~50 kDa) di anticorpi monoclonali attraverso l'uso di papaina e di un'altra proteasi ancor più specifica, combinati con nuove evoluzioni nel *signal processing* e nella analisi dei dati derivati dall'MS. In conclusione, abbiamo dimostrato che l'approccio MD alla spettrometria di massa ed alla proteomica è un strumento potente, benchè non ancora sviluppato a pieno, complementare tanto al BU quanto al TD. Questo lavoro ha valutato l'MD nell'analisi mirata di singole proteine. In futuro, grazie agli sviluppi del settore, prevediamo invece la sua applicazione estensiva ad studi di proteomica su larga scala.

Parole chiave: spettrometria di massa, MS; spettrometria di massa tandem, MS/MS; middle-down, MD; spettrometria di massa a trasformata di Fourier, FTMS; Orbitrap; proteina; immunoglobulina G, IgG; digestione chimica; proteolisi.

List of papers

The present Thesis is based on the following research articles:

I. Laskay Ü. A., Lobas A. A., **Srzentić K.**, Gorshkov M. V., Tsybin Y.O. Proteome digestion specificity analysis for rational design of extended bottom-up and middle-down proteomics experiments. *Journal of Proteome Research* (2013), 12 (12), 5558–5569

II. Laskay Ü. A., **Srzentić K.**, Fornelli L., Upir O., Kozhinov A. N., Monod M., Tsybin Y. O. Practical considerations for improving the productivity of mass spectrometry-based proteomics. *Chimia* (2013), 67 (4) 1–6

III. Laskay Ü. A., **Srzentić K.**, Tsybin Y.O. Extended bottom-up proteomics with secreted aspartic protease Sap9. *Journal of Proteomics* (2014), 110, 20-31

IV. **Srzentić K.,** Fornelli L., Laskay Ü. A., Monod M., Beck A., Ayoub D., Tsybin Y.O. Advantages of extended bottom-up proteomics using Sap9 for analysis of monoclonal antibodies. *Analytical Chemistry* (2014), 86 (19), 9945–9953

V. **Srzentić K.**, Zhurov K. O., Lobas A. A., Nikitin G., Gorshkov M. V. and Tsybin Y. O. Chemical-mediated digestion: an alternative realm for middle-down mass spectrometry. *Manuscript in preparation*

VI. Ayoub D., **Srzentić K**., Fornelli L., Tsybin Y. O. Middle-down electron transfer dissociation mass spectrometry of IgG mixtures allows light and heavy chain pairing characterization. *Manuscript in preparation*

VII. **Srzentić K.**, Nagornov K. O., Lobas A. A., Ayoub D., Fornelli L., Gorshkov M. V. and Tsybin Y. O. Revealing chain connectivity in monoclonal IgG1 using GingisKHAN proteolysis and top-down electron transfer dissociation Orbitrap FTMS. *Manuscript in preparation*

My personal contribution to these papers is the following:

I have performed the entire FTMS experiments for papers III - VI and partially for paper VII. Data acquisition for the paper I was performed with the assistance of Dr. Unige A. Laskay. I personally carried out all the sample preparation and most of data analysis for all the papers, except for sample preparation of papers I and VII and bioinformatics of papers II and VI for which we received important support from collaborators (listed as co-authors). Finally, I was involved in the discussion of all the research projects and actively contributed to the idea generating, writing and correction of all manuscripts.

Papers (research articles and reviews) not included in the discussion (* indicates equal contribution):

1. Gasilova N., **Srzentić K.**, Qiao L., Tsybin Y. O., Girault H. H. On-chip mesoporous functionalized magnetic microspheres for extended bottom-up proteomics, *Analytical Chemistry* (2015), *in print*
2. **Srzentić K.***, Schumann K.*, Tsybin Y.O. and Shevchenko A. Quantification of the membrane lipidome turnover by metabolic 15N labeling and ultra-high resolution Orbitrap FTMS. *Manuscript in preparation*

List of presentations.

Talks:

1. **Srzentić K.**, Zhurov K. O., Kussmann M., Tsybin Y.O. Chemoselective digestion: an alternative realm for middle-down mass spectrometry. Presented at the 33rd meeting of the Swiss Group for Mass Spectrometry, Beatenberg, Switzerland, October 29-30th 2015
2. **Srzentić K.**, Gasilova N., Zhurov K. O., Kussmann M. and Tsybin Y. O. Chemoselective digestion for middle-down proteomics and structural analysis of monoclonal antibodies. Presented at 9th Mass Spectrometry in Biotechnology and Medicine (MSBM) Summer School, July 5-11th 2015, Dubrovnik, Croatia
3. **Srzentić K.**, Laskay Ü. A., Tsybin Y.O. Utility of chemical agents for high-throughput protein characterization. Fall meeting of the Swiss Chemical Society, Lausanne, Switzerland, September 6 2013

Posters:

1. **Srzentić K**., Zhurov K., Nikitin G., Cindric M., Kussmann M., Tsybin Y. O. Chemoselective digestion for middle-down proteomics and structural analysis of monoclonal antibodies. Presented at 63rd ASMS Conference on Mass Spectrometry and Allied Topics. St. Louis, MI, US, May 31st – June 3rd 2014
2. Kozhinov A. N.; Wuehr M.; Corthésy J.; Nagornov K. O.; **Srzentić K**.; Dayon L.; Kussmann M.; Gygi S. P.; Tsybin Y. O. Super-resolution signal processing leverages multiplexed quantitative proteomics. Presented at 63rd ASMS Conference on Mass Spectrometry and Allied Topics. St. Louis, MI, US, May 31st – June 3rd 2014
3. Schuhmann K.; **Srzentić K**.; Nagornov K. O.; Gutmann T.; Coskun Ü.; Tsybin Y. O.; Shevchenko A. Quantification of the membrane lipidome turnover by metabolic 15N labeling and ultra-high resolution Orbitrap FTMS. Presented at 63rd ASMS Conference on Mass Spectrometry and Allied Topics. St. Louis, MI, US, May 31st –June 3rd 2014

4. Gasilova N., **Srzentić K**., Qiao L., Tsybin Y. O., Girault H. On-chip mesoporous functionalized magnetic microspheres for extended bottom-up proteomics. Presented at 63rd ASMS Conference on Mass Spectrometry and Allied Topics. St. Louis, MI, US, May 31st – June 3rd 2014
5. Tsybin Y. O., Nagornov K. O., **Srzentić K**, Kozhinov A. N. Advanced time-domain data analysis for improved FTMS-based proteomics. Presented at Proteomics Forum, Berlin, Germany, 22-25th March 2015
6. **Srzentić K**, Fornelli L, Zhurov K. O., and Tsybin Y. O. Chemical hydrolysis-based middle-down proteomics. Presented at 20th ISMS Conference Geneva, Switzerland, August 2014
7. Levitsky L. I., Ivanov M. V., Lobas A. A., Pridatchenko M. L., Tarasova I. A., **Srzentić K.,** Tsybin Y. O., Mitulović G., Gorshkov M. V. IdentiPy, an open-source MS/MS data search platform for shotgun proteomics. Presented at 13th Human Proteome Organization World Congress. Madrid, Spain, October 5-8th 2014
8. **Srzentić K.,** Karateev G.; Fornelli L., Levitsky L.I, Lobas A. A., Dubikovskaya E.; Gorskhov M. V., Laskay U. A., Ayoub D.; Tsybin Y.O. Chemical hydrolysis-based middle-down proteomics. Presented at 62nd ASMS Conference on Mass Spectrometry and Allied Topics. Baltimore, MA, US, June 15-19th 2014
9. Ayoub D.; Fornelli L., **Srzentić K.,** Laskay U. A., Beck A., Tsybin Y.O. Middle-down and extended bottom-up mass spectrometry for in-depth and rapid characterization of immunoglobulins and their mixtures. Presented at 62nd ASMS Conference on Mass Spectrometry and Allied Topics. Baltimore, MA, US, June 15-19th 2014
10. **Srzentić K**., Fornelli L., Ayoub D., Laskay Ü. A., Beck A., Tsybin Y.O. Characterization of therapeutic antibodies by middle-down MS strategies. Presented at European Proteomics Association Scientific Meeting. Saint Malo, France, October 14-17th 2013

List of abbreviations.

AC	Alternate current
AGC	Automatic Gain Control
BUP	Bottom-up proteomics
CDR	Complementarity determining regions
CID	Collision-induced dissociation
Da	Dalton
DC	Direct current
ECD	Electron capture dissociation
eFT	Enhanced Fourier transform
ETD	Electron transfer dissociation
F(ab)	Fragment antigen-binding
Fc	Fragment crystallizable region
FDR	False Discovery Rate
FTMS	Fourier transform mass spectrometry
FWHM	Full width at half maximum
HCD	Higher-energy collision-induced dissociation
Ig	Immunoglobulin
IgG	Immunoglobulin G
IRMPD	Infrared multiphoton dissociation
LC	Liquid chromatography
LC-MS/MS	Liquid chromatography-tandem mass spectrometry
LIT	Linear ion trap
LTQ	Linear trap quadrupole
m/z	Mass to charge ratio
mAb	Monoclonal antibody
MD	Middle-down
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
NCE	Normalized collision energy
ppm	Parts per million
PTM	Post-translational modification
QIT	Quadrupole ion trap
QQQ	Triple quadrupole
RF	Radio frequency
RP	Reversed phase
Sap9	Secreted aspartic protease 9
SNR	Signal-to-noise ratio
TD	Top-down
Th	Thomson
TOF	Time-of-flight

UVPD	Ultraviolet photo dissociation
DIA	Data-independent aquisition
DDA	Data-dependent aquisition

Chapter 1. Introduction

1.1. Studying biology with analytical techniques, or coupling uncertainties.

The concept of uncertainty inevitably occupies a pivotal role in epistemology, the branch of philosophy discussing the limits of knowledge. Naturally, also scientific research, as a study of the nature of things, is widely interconnected with this subject, and some of the most famous and celebrated scientific discoveries and achievements are fully centered on this topic. A notable example is given by 1927 Heisenberg's *uncertainty principle*, which postulates how there is a fundamental limit to the precision with which certain pairs of physical properties of a particle, such as position and momentum, can be known simultaneously [1]. This famous principle is valid in physics, specifically in quantum mechanics, but similar examples can be found in other scientific disciplines. In mathematical logic, for instance, Gödel's incompleteness theorems, formulated in 1931, describe the intrinsic limitations of formal mathematical systems that, simply speaking, cannot be at the same time consistent and complete. The *trait d'union* of Heisenberg's and Gödel's works seems to be found, at least at a first glimpse, in the ontological impossibility of a fully complete scientific knowledge.

It is needless to say that the epistemological consequences of such formulations are extremely sophisticated and would require a separate dissertation. This Thesis, however, will focus on the *study of biological systems and entities through the application of analytical techniques*: therefore, it is first important to acknowledge the existence of uncertainty principles also within life sciences. To this regard, a recent example is given by the work of Strippoli *et al* [2], demonstrating that the sequence of the genome of a living cell can be determined only with a level of uncertainty that is always greater than zero, and is determined as a function of mutation rate and size of the investigated genome.

An in-depth analysis and extension of this last work can lead us to more effectively describe biological uncertainty as a combination of: (i) inherent characteristics of living systems, which are subject to variations (in the mentioned case, genetic mutation of DNA base pairs), and (ii) limitations of the analytical techniques used for their study. This scheme, despite its extreme simplicity, can be related to the distinction between the notions of *aleatoric* (or statistical) and *epistemic* (or systematic) uncertainty, as defined by the field of uncertainty quantification.

Applying this conclusion as guiding light, the following paragraphs of this Introduction and subsequent Chapters will introduce the object of my scientific investigation, *the proteome*, describing the biological reasons of its deep complexity, and later will present some key features of the analytical tool used for its study, known as *mass spectrometry*, and the conceptual differences – and limitations in terms of achievable information – of the different approaches through which such technique can be used for the proteome analysis.

1.2. From genome to proteome - Era of proteomics

The sequencing of the DNA content within human cells accomplished by the Human Genome project revealed presence of around 25’000 genes [3]. A large number, which nevertheless does not fully explain the complexity of human beings [4]. This is indeed illustrated by gene effectors, or proteins. From the initial «*one gene-one enzyme*» hypothesis by Beadle and Tatum (1941), in the last decades science obtained more realistic picture of the landscape of expressed proteins. Unlike the finite genome (i.e. the complete complement of genetic information in a cell), the proteome is constantly changing in response to internal and external stimuli. Number of proteins certainly exceed the number of coding genes because of genetic (i.e. alternative splicing) and chemical modifications (endogenous or introduced, referred to as post-translational modifications, PTMs), which can influence structure and, subsequently, biological activity of proteins. It has been estimated how all these processes contribute to yield more than one million unique protein forms [5].

Proteomics, as counterpart of genomics, offers highly complementary information to genomics, deeper insight into complex cell communication and reveals biologically important processes at protein level carried out through their interactions. It is the proteome (i.e. the entire set of proteins expressed at a particular time in a cell) that ultimately determines cellular function and phenotype. While DNA and RNA sequencing data is used to predict gene products (e.g., protein sequences, Figure 1.1) and expression levels in cells, protein sequences and abundances can only be inferred based on genome and transcriptome data, and this is one of the major driving factors for the development of methods to directly analyze proteins.

Amino Acid	**DNA codons**
Isoleucine	ATT, ATC, ATA
Leucine	CTT, CTC, CTA, CTG, TTA, TTG
Valine	GTT, GTC, GTA, GTG
Phenylalanine	TTT, TTC
Methionine	ATG
Cysteine	TGT, TGC
Alanine	GCT, GCC, GCA, GCG
Glycine	GGT, GGC, GGA, GGG
Proline	CCT, CCC, CCA, CCG
Threonine	ACT, ACC, ACA, ACG
Serine	TCT, TCC, TCA, TCG, AGT, AGC
Tyrosine	TAT, TAC
Tryptophan	TGG
Glutamine	CAA, CAG
Asparagine	AAT, AAC
Histidine	CAT, CAC
Glutamic acid	GAA, GAG
Aspartic acid	GAT, GAC
Lysine	AAA, AAG
Arginine	CGT, CGC, CGA, CGG, AGA, AGG
Stop codons	TAA, TAG, TGA

Figure 1.1. List of twenty proteinogenic amino acids with their corresponding DNA codons representing each amino acid. All 64 possible 3-letter combinations of the DNA coding units T, C, A and G are used either to encode one of these amino acids or as one of the three stop codons that signals the end of a sequence. While DNA can be decoded unambiguously, it is not possible to predict a DNA sequence from its protein sequence. Because most amino acids have multiple codons, a number of possible DNA sequences might represent the same protein sequence. Adapted from http://www.hgvs.org/

This large-scale qualitative and quantitative study of proteins, particularly their structure and function is one of the cornerstone applications of mass spectrometry (MS) [6]. The applications of MS-based proteomics range from proteome profiling between healthy and diseased systems to recognizing point mutations and deletions in the protein sequence as well as identifying and localizing post-translational modifications.

However, state-of-the-art instrumentation available nowadays in MS along with the development and improvement of methodology still does not enable routine, e.g., clinical, applications which are still restricted by the time-consuming sample preparation as well as by sample complexity, e.g., the large dynamic range of proteins present [7].

1.3. Mass spectrometry (MS): a magnifying glass into proteome complexity?

Mass spectrometry (MS) is defined as an analytical technique for identifying and quantifying compounds by measuring their physical properties: the mass-to-charge ratios (*m/z*) and abundances of charged particles (ions) in the gas phase. The *m/z* is expressed in Thomson unit (Th) [8] fundamentally defined as:

$$1\ \mathrm{Th} = 1\ \mathrm{Da}/e = 1.036426 * 10^{-8}\ \mathrm{kg}\ \mathrm{C}^{-1}$$

Even though the first mass spectrometer was constructed by J. J. Thomson more than a hundred years ago (called parabola spectrograph [9], followed by Thomson's discovery of first stable isotopes 20 and 22 of neon (Ne)), one of the first commercial instruments, called *calutron*, was designed forty years later by Ernest O. Lawrence during the Manhattan Project. The instrument was designed to separate the isotopes of uranium (U-235). Since then, MS underwent countless improvements, all up to the advent of entirely new generation of instruments. Nowadays, emphasis in development is put to meet the needs of 'omics' (proteomics, metabolomics, transcriptomics, *etc.*). Hence, the field of proteomics is discretely becoming synonymous with high-throughput MS-based characterization of proteomes. As a result, mass spectrometry nowadays is established as one of the most versatile tools in protein structural analysis [10-12] and has become the method of choice for protein identification and quantification [13, 14], protein post-translational modifications (PTMs) mapping [15, 16] and the elucidation of protein-protein interactions [17].

The use of a sophisticated technique such as mass spectrometry is justified by the extreme complexity of proteomes, particularly of eukariota. Two distinct phenomena contribute to such complexity: (i) *proteoform variability* and (ii) *protein dynamic range.*

1.3.1. Proteome complexity: from gene to proteoform.

The term proteoform, introduced recently by Smith, Kelleher, and Top-Down Proteomics Consortium [18], represents the attempt of filling the gap separating the realm of genomics, which focuses on a relatively stable entity, the genome, comprising a defined set of protein-encoding genes, and proteomics, which studies gene products. A gene product in its biologically active form is the result of potential modifications, occurring both at the genetic level (e.g., alternative splicing, polymorphisms like snips (SNPs), *etc.*) as well as the chemical level (with the so-called post-translational modifications, or PTMs). A proteoform is, therefore, a protein which includes a *specific set* of such modifications.

To better understand the level of variability of the actual pool of proteoforms forming a proteome, compared to that of the set of genes originally responsible for protein transcription, we can consider that an eukariotic organism has a genome composed of a few thousands of protein-encoding genes (from the ~6'600 of *Saccharomyces cerevisiae* [19] to the ~25'000 of *Homo sapiens* [20]). Currently, the number of known PTMs exceeds 200 [21] and it was estimated that ~80% of protein-encoding genes undergoes alternative splicing [22]. A final number of one million proteoforms seems, therefore, a realistic estimate for the human proteome. Notably, this «natural» variability can be further increased in an artificial fashion, as demonstrated by Zubarev and co-workers [23], due to the procedures required for proteoform extraction/purification and for eventually preparing them for proteomic experiments.

1.3.2. The protein abundance dynamic range in cells and tissues.

The different biological functions carried out by proteins require a differentiation in terms of their *copy-per-cell* number (considering only cellular proteins) or, more in general, of their expression level (if we include also secreted proteins). Proteomes of the most different organisms seem to all follow one rule, which probably reflects some indispensable requirements for life to be possible: a limited number of proteins is highly abundant, then a large portion of the proteome belongs to what can be considered a medium abundant class and, finally, another limited pool of proteins is expressed in an extremely low number of copies (Figure 1.2.). The difference in the copy-per-cell numbers between the most and least abundant proteins, also called *protein dynamic range*, is dramatically pronounced: referring to the previously mentioned examples of yeast and human, a combination of transcriptomic and proteomic techniques allowed to estimate that for these two organisms the range is of 6 [24] and 7 [25] orders of magnitude, respectively. Furthermore, specific tissues or fluids, oftentimes of common value for biological or clinical research, can reach even higher values: for instance, the dynamic range of human plasma spans up to 11-12 orders of magnitude, with the consequence that even accessing 6 orders of magnitude of this range (operation that is complicated by the extreme abundance of a restricted number of proteins, namely immunoglobulins and albumin) would still completely exclude from the proteomic analysis most of the biologically relevant proteins such as interleukins and signalling-cascade activation factors [26, 27].

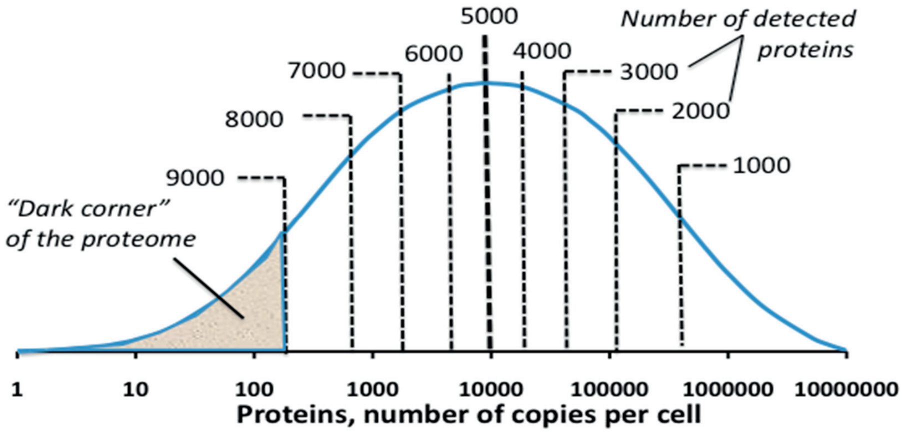


Figure 1.2. The distribution of protein abundances in a typical mammalian organism, Mus musculus, shows a bell-shaped distribution. Note, that the related mRNA distribution would have a similar shape, although shifted towards the left on the x-axis (data not shown). Figure adapted from references [28] and [29].

1.4. Classical MS approaches: bottom-up *vs* top-down

Proteomic studies employ two main approaches, «bottom-up» and «top-down», for characterization of complex protein mixtures, e.g., proteomes, using mass spectrometry, Figure 1.3. Bottom-up proteomics (BUP) is currently the method of choice for large-scale identification and characterization of proteins present in complex samples, such as cell lysates, body fluids or tissues. This conventional approach relies on protein digestion of whole or sub-proteome into short (6-30 amino acid) peptides with enzyme of choice (usually trypsin) before on-line chromatographic separation and tandem mass spectrometry (MS/MS) analysis. These peptides can be efficiently fragmented and identified, but their sequences bring limited specificity at the protein level and they are oftentimes discarded as ambiguous hits. In a typical shotgun proteomic experiment ~100,000 peptides may be present in the sample and a large number of low-abundance peptides are never analyzed [30]. Thus, among the major drawbacks of BUP approach in case of using tryptic digestion is the large number of peptides present in the sample. As a result, most of the precursor ion isolation windows (typically 2 to 3 m/z units) contain co-eluted peptides. The fragmentation patterns of these peptides may overlap and, depending on the level of precursor ion fraction, a search engine may return wrong sequence candidates. Additionally, an inherent limitation of BUP is the so-called *protein inference problem* [31]: several of the proteolytic peptides analyzed during an MS-based shotgun experiment can be associated with multiple gene products (thus leading to ambiguous protein identifications) or, when related to a unique gene, are shared among different proteoforms, that cannot be individually defined. In other words, bottom-up proteomics can yield the identification of *protein families or groups*, not of specific proteoforms.

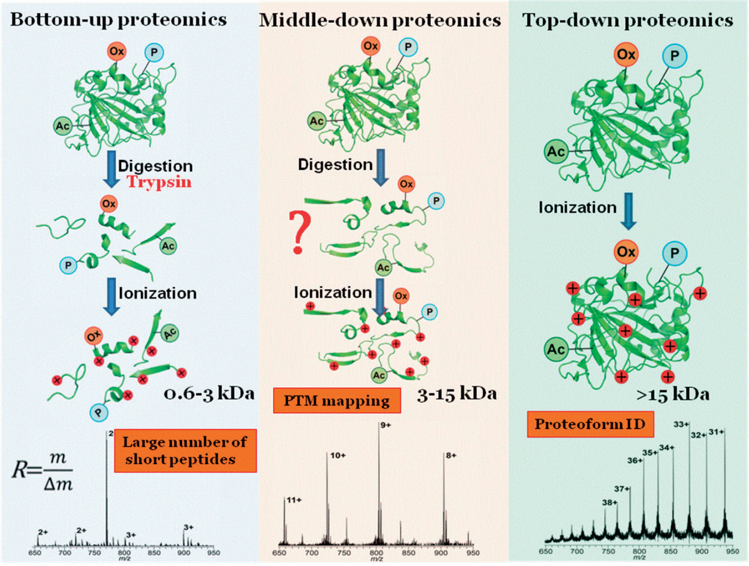


Figure 1.3. MS-based proteomic approaches.

Top-down MS is at the other end of available strategies for proteome coverage. This approach omits proteolysis and intact proteins (primarily in 15-50 kDa range) are fragmented in the gas phase [32, 33]. The major advantage of top-down proteomics (TDP) is the access to the entire protein sequence and information about possible genetic or post-translational modifications present, overcoming the protein inference problem [32]. However, technical challenges in MS methods, for what concerns both the detection of intact species as well as their fragmentation in the gas phase require the use of highly sophisticated instrumentation for TDP, which is currently efficient in the identification and characterization only of small proteins (typically below 30 kDa). Additionally, the separation of intact proteins using adsorption chromatography (*vide infra*) is also limited to small size proteins with low level of diversity. Because of the above reasons, the implementation of TDP approach has been so far circumscribed to low complexity mixtures of proteins of relatively low molecular weights, although a top-down approach has been used for the targeted characterization of proteins and protein complexes up to 150 kDa and above [34-36].

1.5. Middle-down (MD) MS approach

Middle-down proteomics (MDP) is an approach that aims to combine the benefits of bottom-up and top-down approaches, while minimizing their above-mentioned limitations. Here, similarly to bottom-up, proteins are digested, however, a restricted (less frequent) proteolysis is employed to increase the average size distribution of the resulting peptides. The target peptide distribution here is 30-150 residues or 3–15 kDa. The complexity of a mixture is reduced, allowing high resolution mass analysis on liquid chromatography (LC) separation timescale. In addition, the increased peptide length typically results in a larger number of charges per precursor ion thus increasing the efficiency of MS event [37].

Specifically, efficient separation of these long peptides can be readily performed on commercial chromatographic columns, and the elution profile and LC peak capacities are comparable to those of the bottom-up approach. Although a longer acquisition time is necessary for recording of high resolution MS/MS spectra, this is achievable with modern instrumentation, such as the hybrid instruments (*vide infra*, Chapter 3). In addition, the long amino acid series enhance the uniqueness of the sequence and increase the probability for localization of covalent modifications such as PTMs and single-point mutations. Therefore, we consider this paradigm shift towards analysis of longer peptides to be the key for achieving increased dynamic range of protein concentrations and high-throughput identification of targeted proteoforms.

Proof-of-principle preliminary experiments that attest the viability and importance of the MDP method have been conducted previously by prominent US-based research groups in the field of proteomics and mass spectrometry, namely undersigned by distinguished research investigators such as Catherine Fenselau (University of Maryland) and Scott McLuckey (Purdue University). Previous research efforts in MDP have been using proteolytic enzymes with well-established cleavage sites and optimized proteolysis conditions that cleave selectively at a single residue (AspN, LysC, and GluC) [38-42] as well as chemical cleavage such as microwave-assisted acid hydrolysis [37].

1.6. Selected application: Analysis of Immunoglobulins G (IgGs)

Immunoglobulins currently represent the fastest growing class of biotherapeutics in the pharmaceutical industry. Produced by B cells, these proteins recognize a target molecule (antigen) with both high specificity and selectivity (e.g., binding constant in the nano/picomolar range). Immunoglobulins are tetrameric glycoprotein complexes; their characteristic quaternary structure is composed of two identical light and two identical heavy chains, whose molecular weight (MW) is ~25 and 50 kDa each, respectively, for a total MW of ~150 kDa, Figure 1.4.

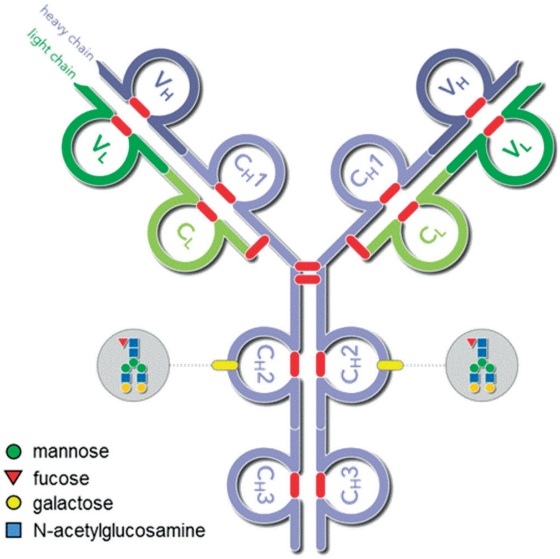


Figure 1.4. Schematic representation of an immunoglobulin class G1 (isotype IgG1). Heavy chains are colored in blue, light chains in green. Glycosylation sites are highlighted in grey. The typical G0F glycan is indicated in circles. CDRs are located on variable domains, indicated with V_L and V_H for light and heavy chain, respectively.

According to the type of heavy chain (indicated by Greek letters α, δ, ε, γ and μ), in mammals we distinguish five isotypes of immunoglobulins: IgA, IgD, IgE, IgG and IgM. Three of them are monomeric (IgD, IgE and IgG), whereas IgA are dimeric and IgMs are present in pentameric form. The most important isotype is that of IgG, which is divided into 4 subclasses (numbered from 1 to 4). These subclasses differ essentially for one of the most important structural features of immunoglobulins, which is the disulfide bond connectivity. Both the chains of IgGs are divided into globular domains, each of which includes an intramolecular disulfide bridge. Intermolecular disulfide bridges connect then each light chain with one heavy chain, and the heavy chains are connected together by a number of S-S bonds, variable according to the subclass, in the so-called *hinge region*.

Most of therapeutical immunoglobulins are IgG of the subclass 1, IgG1, which is also the most abundant in humans. These are characterized by two intermolecular disulfide bridges at the hinge region as depicted in Figure 1.4. Immunoglobulins belonging to the same isotype share most of the sequence of their heavy chains (specifically, the central and C-terminal portion). Similarly, this is the case also for the two possible kinds of light chains, λ and κ. The highly conserved regions of immunoglobulins form the *constant* domains, whereas the N-terminal part of each chain is referred to as the *variable* domain. The latter includes, on both light and heavy chains, three regions exposed to the surface, and loops, called *complementarity determining regions* (CDRs). CDRs are responsible for the binding of a specific region (epitope) of the target antigen. Antigen binding is mediated by the variable domains, mainly by three loops connecting individual β-strands in each domain (CDR). Like natural IgGs, all recombinant antibodies contain an -Asn-X-Ser/Thr-Y- consensus sequence for N-glycosylation in their heavy chain C_H2 constant domain where X and Y are amino acids different from proline.

- PTMs in IgG

Among the post-translational modifications found on IgGs, the most important is surely represented by the N-glycosylation (+> 1000 Da) of the heavy chain (occurring at Asn_{297}, following Kabat numbering (see Johnson [43] and references therein). The glycosylation profile of recombinant IgGs can change according to the host system. Further common PTMs include pyroglutamic acid formation, methionine oxidation, deamidation of glutamine and asparagine. Deamidation is a spontaneous phenomenon believed to target proteins for degradation [44], but it can also be enzymatically induced.

Deamidation converts Gln and Asn in glutamate and aspartate, respectively, with a mass shift of +0.984 Da. This implies that deamidation replaces a polar amino acid with a charged one, with potential structural consequences on the involved protein. Importantly, deamidation occurs via the formation of a cyclic intermediate, so that the final product can be a structural isomer (enantiomer) of Glu and Asp, like γ-glutamic acid or β-aspartic acid[45] (also known as iso-Asp), respectively (Figure 1.5).

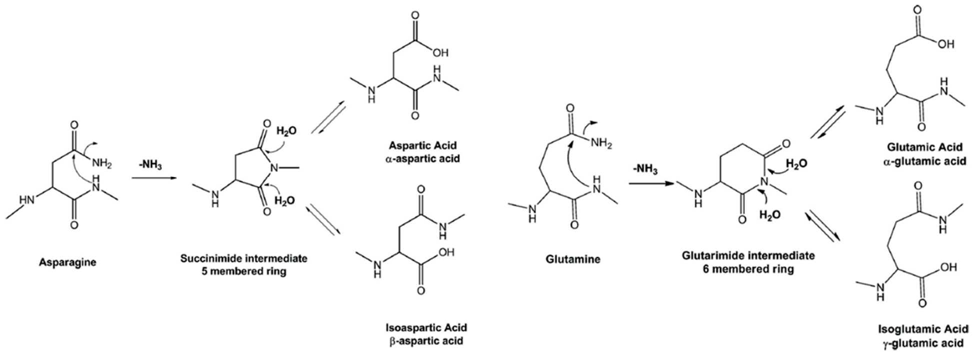


Figure 1.5. Pathways leading to the formation of structural isomers of Asp and Glu as a consequence of deamidation of Asn and Gln, respectively. Left panel shows the formation of a succinimide intermediate with possible production of either α-Asp or β-Asp. Right panel illustrates the formation of α-Glu or γ-Glu from a glutarimide intermediate. Schemes derived from references [46] and [47].

Mass spectrometry was shown to be a powerful tool in identification as well as quantitation of this PTM. Although enantiomer formation was first detected in MS-based studies applying collision-induced ion activation, electron-based activation methods rose as methods of choice for investigation of modifications involving structural isomers rearrangements. Here the distinction between amino acid isomers is achieved through the observation of specific reporter product ions that can be identified in the ECD/ETD tandem mass spectra only in presence of non-*α* amino acids. ETD was applied to the differentiation of α-Asp from β-Asp [48]. Deamidation is of particular importance in biotherapeutics quality control, as it is known how the deamidation of Asn in the paratope region (CDR2 of IgG1) can lead to unsuccessful antigen binding and affect its activity, hence, advancement of MS-based methods for its elucidation. Chapter 4 provides results relevant to this problematic by successful adaptation of novel extended bottom-up proteomics (eBUP) pipeline. Conceptual relevance of the obtained results is presented and summarized in the research article within Chapter 4 (Paper 4).

- MS-based structural analysis of IgGs and allied challenges

IgGs offer vast variety of problematics, making it both a suitable molecule and at the same time a challenging test bed for method development. The role of mass spectrometry in the characterization of immunoglobulins is fundamental for different reasons. First, due to the non-widely studied PTMs occurrence, whose masses are always very small compared to the overall size of the intact protein. Furthermore, the size of these proteins complicates the analysis with other traditionally employed techniques such as gel electrophoresis. Last but not the least, given the high sequence homology between IgG classes, as well as within the class (IgG1), it is important to confirm the IgG sequence obtained by genomic data, particularly for the CDR domains, given the high variability of these sequences. Traditionally, detailed characterization of IgGs has been carried out primarily by bottom-up MS [49]. The combination of data derived from the digestion of the antibody by different proteases can effectively result in high sequence coverage (up to 100%) and facilitate the identification of both large (e.g., glycosylation) and small (e.g., deamidation) PTMs [50].

Nevertheless, up to date, there is no universal MS approach that would suffice to yield successful analysis of all the underlined problematics relative to IgGs. However, in past decade, various approaches tackling different aspects of this protein were adopted, as depicted in Figure 1.6.

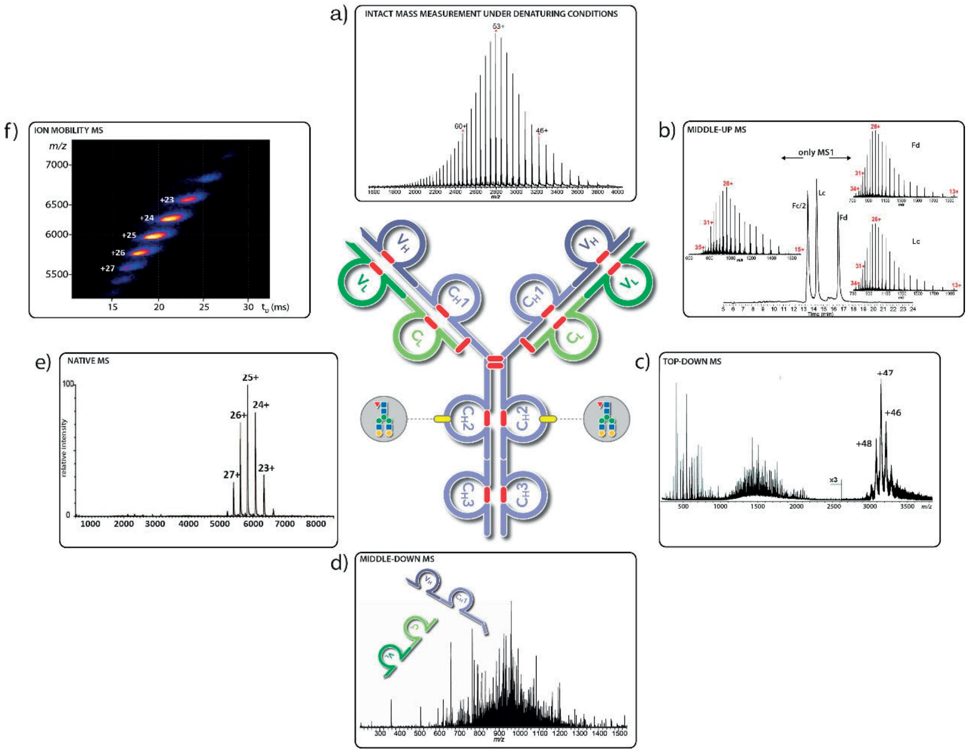


Figure 1.6. MS-based mainstream avenues adopted in structural analysis of IgGs. Figure adapted from references [35, 36, 51, 52]

Within this thesis, particularly in Chapters 5-7 we will introduce alternative and complementary avenues developed for structural analysis of IgGs.

1.7. Aim of the thesis

Comprehensive qualitative and quantitative description of biological systems requires further improvement of molecular structure analysis approaches. Current large-scale and targeted protein analysis based on bottom-up and top-down mass spectrometry cannot provide the required level of analytical characteristics. The common trait of the hereinafter presented research can be thus defined as **advancing structure analysis of proteins constituting complex biological systems, e.g., proteomes and protein complexes, by developing and applying middle-down approaches on the *state-of-the-art* mass spectrometry technology.** The term *state-of-the-art* is not referred only to hardware or software features, but rather to the application of advanced instruments to extreme cases, finalized to obtain a proof-of-concept and to push the limits of technology by recognizing and addressing its current limitations. Project objectives to be achieved were defined by work in the following interdisciplinary research directions: (i) bioinformatics for rational design of proteome digestion strategies and tailored processing of middle-down proteomics data; (ii) search for specific reagents, e.g., proteases, for middle-down proteomics; (iii) tandem mass spectrometry method and technique development, including optimization of radical chemistry-based ion activation and dissociation strategies; (iv) improved conditions for solution-phase protein fractionation and separation; (v) validation and comparison of the method performances using sets of known proteins with equimolar and variable concentrations and (vi) targeted applications of middle-down proteomics, particularly including structural analysis of monoclonal antibodies.

1.8. Overview of Chapters

In order to provide a fundamental understanding of herein applied technology and comprehensive explanation of results, this Thesis is structured as follows: Chapter 2 is aimed at clarifying in detail the basic mass spectrometry concepts and introducing the problematic of biological systems we aim to investigate, finally underlying the current limitations of the chosen approach, but at the same time the rationale supporting their choice. Chapter 3 is dedicated to in-depth description of the instrumentation and techniques used in this Thesis. Chapters 4 to 7 report introductions and summaries of results achieved including pertaining research articles: for the elucidation of the effects of enzymatically-induced modifications on a peptide. Finally, Chapter 8 summarizes the obtained results and outlines future research directions.

Chapter 2. Experimental methods

2.1. MS fundamentals: what to understand and how to apply it in proteomics analysis

2.1.1. Mass accuracy and mass resolution

The capability of determining the nature of an analyte observed in a mass spectrum depends heavily on two parameters, *mass resolution* and *mass accuracy*. This statement becomes more important with the increased complexity of the sample under analysis and the size of the pool of potential candidates to match with the experimental data – which is a situation well represented by proteomic studies.

Starting with mass resolution, this is a measure of the capability of the mass spectrometer of distinguishing two analytes represented by close signals in a mass spectrum (see Marshall and Hendricks [53] and references therein). Using a more precise formalism, mass resolution is defined as the minimum difference in mass, $\Delta m=m_2-m_1$ (with m_1 and m_2 being the masses of the two analytes, the first smaller than the second), between the signals of the two analytes such that the valley between the mentioned signals corresponds to a define percentage of the height of the the smaller peak. Traditionally, this percentage corresponds to 50%, and the relative mass resolution is indicated with $\Delta m_{50\%}$. The resolving power is defined as $m/\Delta m$, and typically the resolution that is considered is $\Delta m_{50\%}$, therefore we usually talk of resolving power as full width at half maximum, or FWHM.

On the other hand, mass accuracy corresponds to:

$$Mass\ Acc = \frac{m/z_{exp} - m/z_{theor}}{m/z_{theor}} \qquad (2.1.)$$

where m/z_{exp} and m/z_{theor} are the experimentally determined and theoretical mass-to-charge values of a given analyte, respectively. Mass accuracy is usually expressed in *parts-per-million* (ppm).

Back to the introduction of this paragraph, it will be now apparent how mass resolution and accuracy are related to each other, as, for instance, higher mass resolution allows for reduced interference of neighboring peak in determining the final peak shape of the signal of a given analyte, thus improving the correct positioning of the peak apex and, finally, mass accuracy. If it is true that the mass calibration of an MS instrument is obtained by fitting the observed *m/z* values of calibrants of known molecular formula with their exact masses (i.e., the mass obtained by summing the isotopes of each element included in its molecular formula, *vide infra*), performing experiments on unknown analytes (as in any proteomic project) we try to match an observed *m/z* value (and, therefore, mass) with a possible chemical formula. Increasing the mass resolution by using more sophisticated instrumentation (see Chapter 3) has dramatic beneficial effects in the reduction of the potential candidates: for instance, it has been estimated that for a tryptic peptide of *Saccharomyces cerevisiae* with a mass of ~2.36 kDa passing from a mass accuracy of 2 Da to 0.5 Da (or from ~850 to ~ 200 ppm) reduces four-fold the number of potential candidates [54]. Furthermore, a more recent *in silico* simulation [55] showed how that if MS measurements are performed at 1 ppm mass accuracy on tryptic peptides, it is possible to exclude 99% of peptides having the same nominal mass (i.e., the mass calculated using the integer mass of the most abundant isotope of each element in the chemical formula) but different chemical formula and, hence, amino acid compositions.

2.1.2. Isotopes and their role in mass spectrometry of polypeptides

In nature, each chemical species can be associated to a chemical formula, which defines the type and number of atoms present in that given neutral molecule or ion. In the case of this Thesis, the object of our analytical investigation are polypeptides, which are built of a limited number of amino acids (see Chapter 1). These building blocks for peptides and proteins are in turn composed of a restricted group of elements: hydrogen, carbon, oxygen, nitrogen, and sulfur (in order of relative abundance in proteins). Considering that each of this elements include different isotopes, it is apparent that, when analyzing a polypeptide we will be in the presence of not only a single species (a single ion, in the case of MS analysis), but of a group of *isotopologues*, or chemical species with the same chemical formula differing by the isotopic composition [56]. Although it is technically possible to isolate a single isotopologue, in practice an MS measurement usually detects isotopic envelopes (also known as isotopic distributions or isotopic clusters). In mass spectrometry the isotopologue ions of a single chemical species can allow the determination of the charge of the ion cluster itself – if the elemental composition is known, as it is the case for polypeptides. Specifically, we can approximate the mass difference between consecutive isotopologues in any envelope generated by peptide or protein ions as equal to the addition of a ^{13}C: going from left to right in the isotopic cluster we are increasing the mass of the isotopologues of m_{13c}-m_{12c}~1.003 u (essentially, the mass of a neutron). This approximation can be applied as carbon is one of the most abundant elements in polypeptides and at the same time the one with the highest relative abundance of heavy isotopes (^{13}C represents about 1% of the total carbon, whereas for instance ^{2}H is only ~0.01%). Therefore, provided that sufficient mass resolution is used, so that the exact *m/z* values can be assigned to each isotopologue, the spacing between consecutive isotopologues will be equal to 1.003/*z*, where *z* indicates the charge of the ion cluster. Obviously, knowing the charge *z* and the corresponding *m/z* value of an ion in a mass spectrum means to be in the position of determining its mass. To precisely determine the mass of the neutral molecule, also the ionization technique and/or the charge carrier has to be known (see Chapter 3.1).

The presence of isotopes is also the reason of the existence of multiple possible definitions of mass for each polypeptide (or, more in general, for any molecule): in particular, the *monoisotopic mass* corresponds to the mass value calculated by considering only the most abundant isotopes of each element, which in the case of a polypeptide correspond also to the lightest isotopes. Conversely, the *average mass* is obtained by summing the average atomic masses of all of the elements present in a molecule. Notably, in the mass spectrum representing a polypeptide, the monoisotopic peak, representing an ionized monoisotopic molecule, corresponds always to the isotopologue ion positioned at the extreme left of the isotopic cluster. Differently, the average mass is a pure mathematical estimate and does not physically correspond to any isotopologue ion in a cluster. Nevertheless, by increasing the size of the polypeptide the isotopic distribution assumes a progressively more pronounced Gaussian-like shape (Figure 2.1.), and the average mass becomes generally closer to the most abundant mass, which is the mass of the most abundant isotopologue. Indeed, the average mass can be also defined as the centroid of the isotopic distribution.

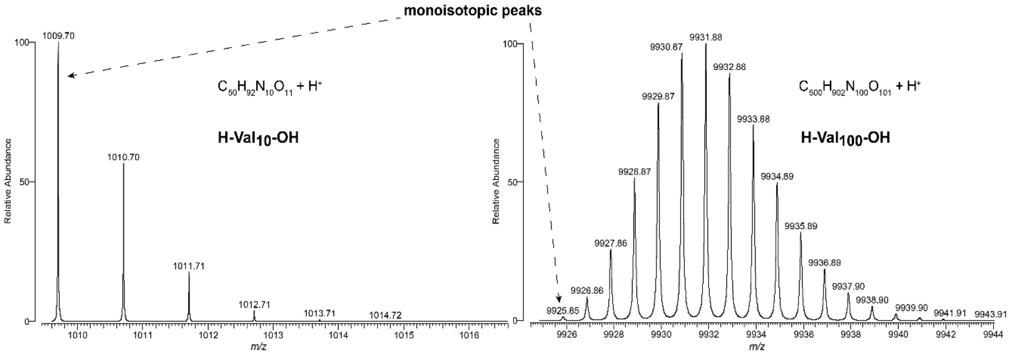


Figure 2.1. Isotopic distributions of the singly-charged ions of model homopeptides H-Val$_{10}$*-OH and H-Val*$_{100}$*-OH. The two examples show how the monoisotopic peak is also the most abundant isotopologue in the cluster for the shorter peptide, but becomes a relatively low-abundant isotopologue peak in the case of a larger peptide.*

Importantly, when the spectral mass resolution is not sufficient to distinguish the isotopologues present in an isotopic cluster (as, for example, in the case of particularly highly-charged ions), the charge state – and, therefore, the mass – of an ion can be still inferred if the analyte ionized producing several ion clusters of different charge state, as it is often the case for polypeptides. The sum of all the charge states referring to a single polypeptide is generally called *charge state envelope* or *distribution*.

Considering two consecutive charge states in the envelope, whose respective m/z values are indicated as X and Y (with X higher than Y), we will have that:

$$X = \frac{m}{z_1} = (m + z_1 * m_c)/z_1 \qquad (2.2.)$$

and also:

$$Y = \frac{m}{z_2} = \frac{m + z_2 * m_c}{z_2} = \frac{m + (z_1 + 1) * m_c}{z_1 + 1} \qquad (2.3.)$$

Where z_1 and z_2 are the charge states of X and Y, respectively, and m is the average mass of the analyte. In the simple and common case that the charge carrier is a proton, for the sake of simplicity we can approximate its mass to 1 u, and therefore the charge z_1 can be calculated as:

$$z_1 = \frac{Y - 1}{X - Y} \qquad (2.4.)$$

Finally, the average mass of the polypeptide will be approximately equal to:

$$m = (X * z_1) - z_1 \qquad (2.5.)$$

2.1.3. Signal to noise ratio, spectral dynamic range

In every analytical measurement, including mass spectrometric ones, it is possible to distinguish two main components: the *signal* of the desired analyte, and background *noise*. In general, the noise can cause distortions in the analyte signals, especially for low-abundant ones, and is detrimental for achieving high mass accuracy and resolution. For most mass spectrometers, including Fourier transform-based instruments discussed in the following Chapter of this Thesis, we can differentiate between two components in the spectral noise: *chemical noise*, which is the complex of signals originated from chemical components in the sample matrix other than the target analyte, and *thermal noise*, generated by the electronic apparatus used for ion detection. Maximizing the so-called *signal-to-noise ratio* (S/N or SNR) improves both the *limit of detection* (LOD) as well as the *limit of quantification* (LOQ) for analyzed molecules. Different methods have been applied to estimate the noise level or intensity and, thus, SNR. Commonly, the “N sigma” rule is applied: considered the stochastic nature of noise, this will have an certain distribution, whose width can be described through its standard deviation, σ; a signal level equal to Nσ (with N positive number) is then defined as the threshold below which the signal of the analyte is not clearly distinguishable from noise and cannot be considered for confident assignment [57]. Effectively, this method aims at defining a “noise baseline” to discriminate between true positive and false positive signals. This concept is then extended to several softwares used in proteomics: for instance, the algorithm THRASH calculates this baseline by assuming that noise-rich areas of the spectrum are characterized by the highest point (i.e., recorded signal) density, and subsequently allows for “true positive signals”-filtering using a user-defined value of SNR calculated on the base of such noise baseline [58]. Finally, among the strategies aimed at improving SNR there is the averaging of mass spectra or, in the case of Fourier transform MS (*vide infra*), time-domain signal. In the latter case, this time-consuming method produces an increase of SNR equal to $\sqrt{N}$, where N is the number of averaged time-domain signals (this depends on the fact that the signal amplitude increases linearly with N, whereas the noise amplitude with √) [59]. Closely related to the concept of SNR, the spectral *dynamic range* defines the ratio between the amplitude of the most over the least intense signal within a single mass spectrum. This parameter is essentially a function of the used hardware, more precisely of the mass analyzer (see Chapter 3). We can anticipate that in the case of the Orbitrap mass analyzer (*vide infra*), used for all the works presented in this Thesis, the spectral dynamic range has been estimated in 4 orders of magnitude in the best cases [60]. The effective dynamic range value that allows for accurate mass measurements, though, is reduced to about 5000. Under suboptimal experimental conditions, like in the case of particularly complex mass spectra (where the signal is spread through a multitude of different ions), this value can further decrease. This last scenario is very common in proteomic experiments, were multiple polypeptide ions can be simultaneously analyzed. As described in the previous paragraph, polypeptides ionize forming isotopic clusters, which are composed of a number of isotopologue ions that increases with the length of the polypeptide chain. Therefore, mass spectra of MD and TD experiments are generally characterized by a smaller dynamic range compared to those of BUP (where shorter peptides are investigated), and the identification of the monoisotopic peak in further complicated by its relatively-low abundance within the isotopic cluster. Further considerations on SNR and its improvement in MD experiments are discussed in research articles, Paper VI and VII, enclosed in the Chapter 7.

2.2. Tandem mass spectrometry MS/MS

In the previous section of this chapter the importance of mass accuracy and resolution is discussed, as a prerequisite for intact mass determination when elemental composition of the investigated protein/peptide is known. In such case accurate mass determination from the first mass measurement (referred to as MS^1 or the *survey scan*) is sufficient for the identification of analyte of interest. However, when two peptides share the same intact mass, but differ in the order of amino acid residues they are composed of (e.g. isomers or isobaric peptides), additional information is needed. For this reason, as well as the need to elucidate the structure of unknown species, or to obtain quantitative measurements, in 1960s a two stage mass analysis experiments known as *tandem mass spectrometry* (MS/MS or MS^2) were introduced [61, 62]. Conceptually MS/MS consists of three processes: i) *m/z* selection of the target ion, ii) activation/fragmentation of the selected ion and iii) analysis of the resulting product ions. By convention, the targeted ion is referred to as the precursor, and the fragment one as the product ion, earlier called parent and daughter ion, respectively. Depending on the instrument employed, and the way steps are carried out, tandem mass spectrometry can be performed either *in space* or *in time* [63]. In the former, the precursor selection, fragmentation and fragment detection occur in different mass analyzers (*e.g.,* as in triple quadrupole instruments, QQQ). On the other hand, when all the steps occur in the same mass analyzer in temporal sequence (e.g., linear ion traps, LIT) the MS/MS process is defined as in time. Ion activation and dissociation step ultimately allows elucidation of the peptide/protein *via* the cleavage of the polypeptide chain along the backbone, resulting in N- and/or C-terminal-containing product ions. Generally there are six main types of MS/MS product ions, depending on which of the three backbone bonds (N-C_α, C_α-C and C-N) is cleaved. Oftentimes other types of cleavages might occur, that lead to formation of internal fragments, or those that result in neutral losses (most commonly loss of H_20, CO_2, NH_3, *etc.*) or even losses of an entire side-chain. The latter might aid in localizing introduced or endogenous modification on a primary sequence; however accessing this information comes at the price of broad array of progeny ions present in the tandem mass spectra that are then challenging for deconvolution and characterization. The nomenclature for the polypeptide fragmentation pathways was originally proposed by Roepstorff and Fohlmann [64]. According to the currently accepted convention fragment ions have alphabetical assignments that denote type of the bond cleaved. Numerical denotation indicates the position of the cleavage site in the polypeptide chain. A graphical illustration of the above mentioned fragments nomenclature, as well as the activation methods that lead to their formation (which will be described in the following subsection) is shown in Figure 2.2.

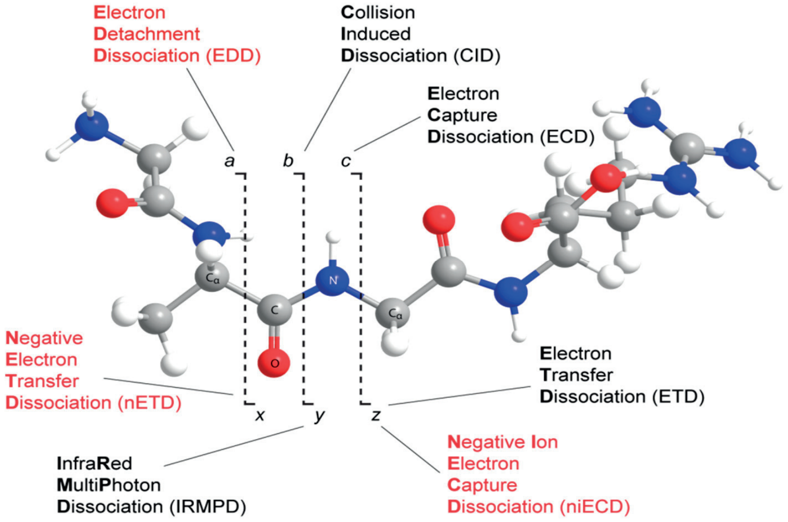


Figure 2.2. Scheme of MS/MS product ions and selection of ion activation methods that lead to their formation. Product ions that retain N-terminus are reffered to as a, b and c- ions, whereas those retaining C-terminus are x, y and z- ions. Complementary ion pairs a/x, b/y and c/z are yielded by the cleavage of C_a-C , C-N and N-C_a, respectively. Adapted from Zhurov et al.[65]

Fragmentation of a polypeptide chain is accomplished by employing different activation/fragmentation methods some of which are introduced in Figure 2.2. and roughly categorized by the type of the product ions they yield. However, generation and entity of product ions is circumspect by different applied mechanisms under which the backbone cleavages occur. Thus, ion activation methods are classified into two main groups that can be defined as: i) energy threshold-based activation and ii) radical-driven activation. We will further discuss their main concepts and list most widely used methods of each group. For the scope of this dissertation, only those employed for the research will be further described in subsections of this Chapter.

2.3. Ion activation and dissociation: Energy threshold-based activation

Tandem MS methods belonging to this group are considered as ‘ergodic’ process, which means that energy is impinged and randomized throughout all vibration modes before dissociation. Deposition of energy depends either on collisions with the neutral gas molecules; here we distinguish between *collision-induced dissociation* (CID) and *higher-energy collision dissociation* (HCD); or on ion-photon interactions. The latter can be further divided in interaction with low energy infrared photons used for *infrared multiphoton dissociation* (IRMPD) [66, 67] and high energy ultraviolet photons used for *ultraviolet photodissociation* (UVPD) [68, 69]. CID and IRMPD are ‘slow-heating’ methods because they require multiple energy accumulation events, while HCD and UVPD are more energetic and require fewer collisions/absorptions. Dissociation occurs when energetic barrier of a chemical bond is surpassed, leading to the cleavage of the weakest bond in polypeptide. In addition to (and separate from) the backbone C-N bond, these bonds are also present at the side chains of amino acid residues.

- **Collision Induced Dissociation**

CID or collision activated dissociation (CAD) [70-73] was the first ion activation employed and has an undeniable importance in overall tandem MS development. It is widely utilized and readily implemented in most of the mass spectrometers, still remaining the activation method of choice to which all other methods are measured up to for fragmentation of positively charged peptides [74]. Strictly speaking it is a two-step process which involves heterolytic cleavage of the sigma (σ) C-N bond ultimately resulting in dissociation of the precursor ion. In the case of a singly charged precursor dissociation will lead to the formation of a charged fragment ion and a neutral portion, while for a higher charge state precursor (2+ and above) two charged fragment ions could be produced. Historically, collision-induced ion excitation was developed in two variants: “soft” CID, performed through resonant excitation, and beam-type CID [75]. Hereinafter when using the acronym CID we will be referring uniquely to the former, whereas the latter will be identified through the name of one of its commercial implementations, HCD.

In collision induced ion activation the dampening of the kinetic energy of the precursor ions is achieved by their multiple inelastic collisions with the molecules of an inert bath-gas (such as He or N). As a result, each collision deposits an increment of internal energy whose excess is then translated into the vibrational energy of the bonds in precursor ion, until the dissociation threshold is exceeded which in turn results in a bond cleavage. Traditionally, in low-energy CID yielded collisions are in range of 1-100 eV [75] whereas high-energy collisions are considered to be in order of several keV. In CID, multiple, low energy collisions are required for reaching the bond fragmentation threshold, whilst the more energetic beam-type activation requires fewer collision events.

HCD is a beam-type collision induced activation (as the one implemented in QQQ) specific to Orbitrap-based instrument. Differently from CID, HCD does not suffer from a low mass cutoff from resonant excitation, so for example reporter ions from isobaric tagged peptides can be easily observed. Even though its name implies that dissociation occurs under high energy, this activation method is still in the regime of low energy (less than 100 eV) [76], however, the final applied energy is somewhat higher compared to the CID one, commonly used in bottom-up regime (30-35 eV). This derives from the fact that HCD is actually a charge dependent method for peptide fragmentation and the redistributed normalized collisional energy (NCE) is recalculated with respect to the precursor charge state and its m/z, for every precursor of the dependent scan in part, as given in the formula below:

$$HCD = \frac{NCE \times {}^{m}/_{z}}{500 \times CF} \qquad (2.6.)$$

Where m/z refers to the precursor in question, 500 is reference m/z at charge state 1, and CF is a HCD correction factor for each charge state as follows (correction factor (charge state)): 1 (1.0); 0.90 (2); 0.85 (3); 0.80 (4) and 0.75 (≥5).

Similarly to CID, HCD is thought to produce complementary *b*- and *y*- ion series. Given the very energetic activation and thus often disruptive for the more unstable *b*- ions which then break into internal fragments and smaller *b*-ions, the returned HCD product ions are mostly *y*-series. The reduced yield of the *b*-ions could be considered one of the shortcomings of this method for data interpretation of MD range (and larger) peptides, and will be discussed as such in the following Chapters.

2.4. Ion activation and dissociation: Electron-based activation methods

Differently from the previous group, electron-based methods could be classified as 'non-ergodic' process, given that dissociation is suggested to precede the energy randomization. These methods are based on interactions of a polypeptide cation or anion with an electron in the gas phase. As a result, a radical species within the peptide backbone is formed leading to the bond rupture. Two most commonly employed methods entailing attainment of an electron by a polyprotonated species are *electron capture dissociation* (ECD) and *electron transfer dissociation* (ETD). Due to the common traits between these methods (*vide infra*), and for the sake of simplicity hereinafter those will be also referred to as ExD (see Zhurov *et al.* [65] and references therein). For the past almost two decades, since the introduction of ECD in 1998 [77, 78] there is an ongoing debate about the underlying fragmentation mechanism, and various mechanistic approaches were proposed [79-82]. According to the '*Cornell mechanism*' ExD proceeds *via* electron capture/transfer at the site of ionizing proton (typically protonated nitrogen of the amine group at the N-terminus or a side chain of basic amino acid residues Arg, Lys, His) inducing a backbone rupture through migration of hydrogen and formation of a radical aminoketyl group [77, 83]. Regardless of the differences on where electron is attained, and how fragmentation proceeds, mechanisms generally agree on non-ergodic dissociation which results in the random homolytic cleavage of N-Cα bonds along the backbone (see Tureček *et al* [84] and references therein). However, recently, heterolytic cleavage was also put into consideration as an alternative pathway [85-87]

- **Electron Transfer Dissociation**

ETD [88] is a radical-driven fragmentation technique based on an electron transfer from an electron-donor molecule, a radical anion, and a multiply-charged cation. The characteristic ion-ion reaction upon ETD can be denoted as:

$$[M+nH]^{n+} + A^{-\cdot} \rightarrow [M+nH]^{(n-1)+\cdot} + A \rightarrow fragments \qquad (2.7.)$$

where "A" represents the electron donor molecule. The bond cleavage gives rise to the series of even-electron c'-ions and radical odd-electron $z^{\bullet}$-ions. ETD dependence on amino acid composition/sequence of the peptide/protein is generally weak. However, it is to be noted how the cleavage at the N-terminal side of proline cannot lead to product ion formation due to the structure of this amino acid. Due to the charge reduction proceeding the electron transfer it is obvious how only multiply charged ($z>1$) precursors can be subjected for a successful fragmentation outcome. However, it has been shown how ETD methods are mostly futile in dissociating doubly charged peptide cations [88, 89]. Consequently, supplemental activations (such as IR photo activation or low energy collision activation) have been implemented for ExD to allow for more effective product ion generation [90-92]. For better understanding of ETD, one of important parameters to consider is *charge density* which indicates the total number of charges distributed in a peptide/protein (here we refer to the number of residues or simply mass). After electron transfer-induced backbone cleavage, fragment ions are still bound in the $[c' + z^{\bullet}]^{(n-1)+\bullet}$ complex by non-covalent interactions. Their separation depends upon the Coulombic repulsions and the nature of the amino acids around particular cleavage site. Ions with greater charge density (lower m/z) yield repulsions sufficient to overcome the interactions and separate now detectable individual fragment ions. Conversely, low charge density induces more compact cationic structure, which in turn prevents separation and subsequent fragments detection. Instead, only m/z values of charge-reduced species are detected. This phenomenon of partitioning from a direct dissociation is denoted as electron transfer without dissociation (ETnoD) [93]. As a consequence, ETD percent fragmentation (a number of observed *c*- and *z*- type ions over the theoretical number of product ions for a considered peptide sequence; *e.g.* a 15 residue long primary peptide sequence has 14 cleavable backbone N-Cα bonds that could give rise to a total of 28 *c*- and *z*- product ions) is linearly decreasing with the increase of the precursor m/z On the other hand, this dependence is much less pronounced in sequence coverage of peptides (defined as a number of backbone bonds cleaved versus the theoretical number of all backbone bonds of the same types) as shown by Good *et al* [92].

Two peculiar features distinguishing ExD from energy threshold-based ion activation methods are: i) the capability of retaining labile PTMs such as phosphorylation, methylation, acetylation and glycosylation [94-102] (this is true for systems with smaller molecular weight such as peptides, see Chapter 4) and elucidating the enantiomere structure by generating diagnostic ions (*e.g.* deamidation of Asn to Asp/isoAsp)[46, 48, 103, 104] ii) the announced preference of cleaving disulfide bridges [79].

Practical aspects of above described fragmentation methods in respect to the instrument employed for this Thesis will be further considered in the Instrumentation Chapter.

2.5. Where to cleave a protein?

The definition of the MDP pipeline cannot disregard the key question of the ideal cleavage site on polypeptide chains. A protein is built using 20 different proteinogenic amino acids, differentiated by their side chains (Figure 2.3).

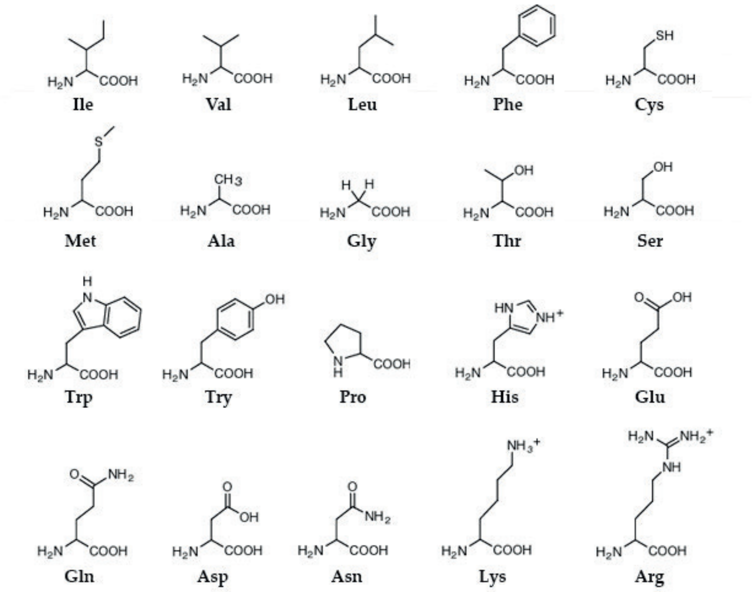


Figure 2.3. The twenty proteinogenic amino acids ordered by the increased hydrophobicity (top left to bottom right) according to the Kyte-Doolittle scale [105]. Amino acids are indicated with the three letter code.

Our purpose is to generate proteolytic peptides of a length of about 30-100 residues. Therefore, it will be important to consider two main aspects when selecting the best *theoretical targets* for protein cleavage: first, the *frequency* of occurrence of each of the 20 amino acids in the proteome of interest (and, eventually, of their combinations, in the case of protein digestion strategies based either on two consecutive cleavages or recognition of pairs of targets); second, the *position of specific amino acid residues*, such as the basic ones that can be protonated, within the sequence of the obtained peptide. The latter aspect, in combination with the average size of obtained polypeptides, is important as it might potentially affect the peptide fragmentation, potentially defining which ion activation technique is more suitable for this category of proteolytic peptide.

This Thesis will present two distinct studies aimed at elucidating both of the above mentioned aspects: Paper I focuses on bioinformatics studies to define the set of residues to target for performing middle-down experiments in different model organism, from bacteria to human. Paper II, instead, discusses the effect of specific positioning of basic residues on the fragmentation of large (>20 amino acids) peptides.

Cela va sans dire, the practical fulfillment of an MDP experiment means to identify a cleaving agent – either enzymatic or chemical – that can actually yield the proteolytic cleavage suggested by the observations and calculations illustrated in Paper I. The field of proteomics has already defined a large set of proteases capable of cleaving proteins with different *specificities*. Figure 2.4. compares commercially available proteases, generally employed in bottom-up experiments (with the sole exclusion of Glu-C, which can be used for MDP of restricted groups of proteins, such as histones, characterized by a peculiar amino acidic composition), with Sap9, a secreted aspartic protease at the center of several pilot studies presented in this Thesis (see Papers III and IV on Chapter 5).

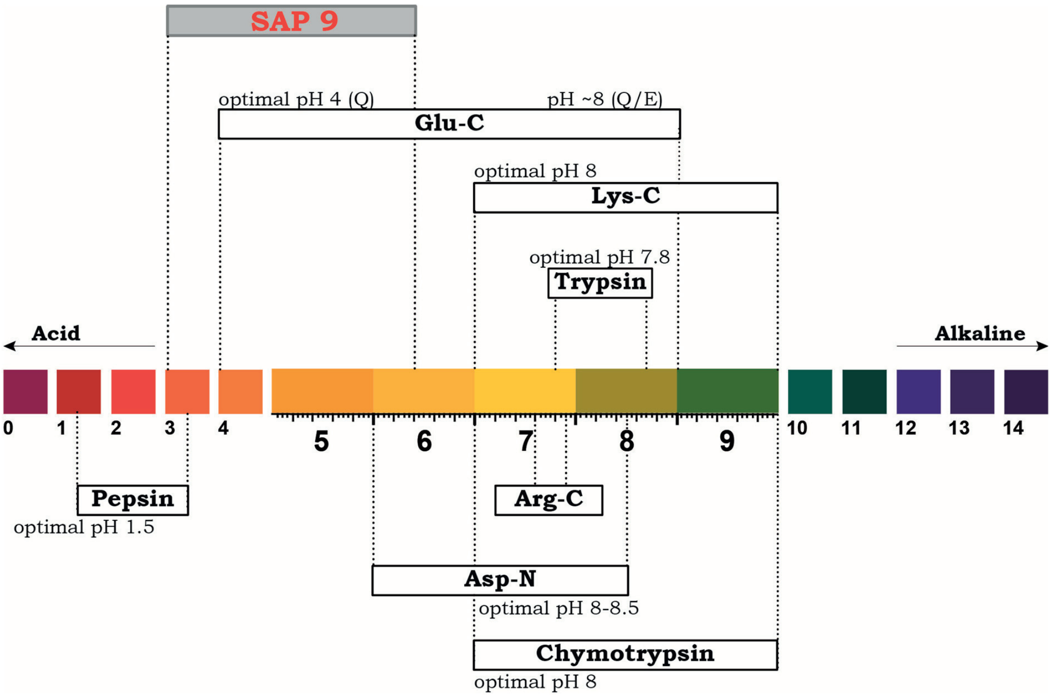


Figure 2.4. Scale of pH and relative range of activity of most common proteases used in proteomics. Sap9, which is objet of this research work, is highlighted in red.

2.6. High-throughput *vs* targeted analysis in proteomics

Going from general to particular according to the deductive method, it is apparent that the development of new ways of studying complex objects requires a *targeted* investigation. Large scale studies would indeed introduce issues in the interpretation of results. Specifically, in this thesis method development and validation of proposed pipelines was initially focused on a smaller sample pool of proteins whose complexity arises from the large size, the presence of multiple modifications whose biological relevance, and/or connectivity between such cannot be readily addressed by a currently adopted single methodology, either because proteins exceed mass for TD analysis (e.g., covalent bonds such as disulfide bridges in IgGs), or modifications are distant in such way that upon BUP proteolysis those end up in two different peptides and connectivity between them is lost.

Notably, once a new cleaving agent has been identified, *in silico* studies aimed at calculating the theoretical proteome coverage, and experiments based on model substrates, are fundamental to characterize the *main* features and *potential* use of the cleaving agent itself (*vide infra*, Chapter 2.8). Nevertheless, these studies cannot fully describe its effective utility in real proteomic applications. A protease recently described in scientific literature, Lys-N, which cleaves at the N-terminal side of lysine, [106] is a perfect example of the need for in-depth studies to determine the best application fields for novel peptide cleaving agents. In large-scale studies, for instance on a complex sample such as mouse heart, the use of Lys-N leads to the identification of a total number of proteins simply comparable to that achievable by digestion with trypsin or Lys-C [107]. This goes in agreement with our MS^2 studies on synthetic model peptides, described in Paper II, which suggest that the positioning of basic amino acid residues at either the N- or C-terminus of a peptide can influence the relative abundance of N- or C-terminal-containing product ions but does not significantly change the overall peptide sequence coverage. Conversely, studies on specific sub-proteomes, like the human phosphoproteome, show large differences in the phosphorylation sites mapped after digestion by Lys-N or trypsin, demonstrating that the two proteases digested the same pool of proteins in a complementary fashion, and the use of both proteases dramatically increases the number of observable phosphorylation sites [108]. Considering that large scale shotgun studies showed comparable numbers of IDs between Lys-C/trypsin and Lys-N, the discrepancies observed in the case of the phosphoproteome are likely to be linked to different *selectivity* of phosphoproteins towards the two groups of enzymes. This example underlines that the substrate selectivity might play a role as important as enzymatic specificity in the final outcome of a proteomic experiment.

2.7. Data analysis

High throughput proteomic analyses generate vast amounts of data for which manual analysis is close to impossible, as it would be extremely time-consuming, hence, an automated interpretation of large proteomic datasets is indispensable. However, even with automated analysis, manual validation of results is often advised. In shotgun proteomics, the most common way for the processing of MS/MS data is through comparison to theoretical fragmentation patterns of peptide sequences generated *in silico* from protein sequence databases following to the specific cleavage rule of the used protease or chemical cleaving agent, as depicted in figure 2.5.

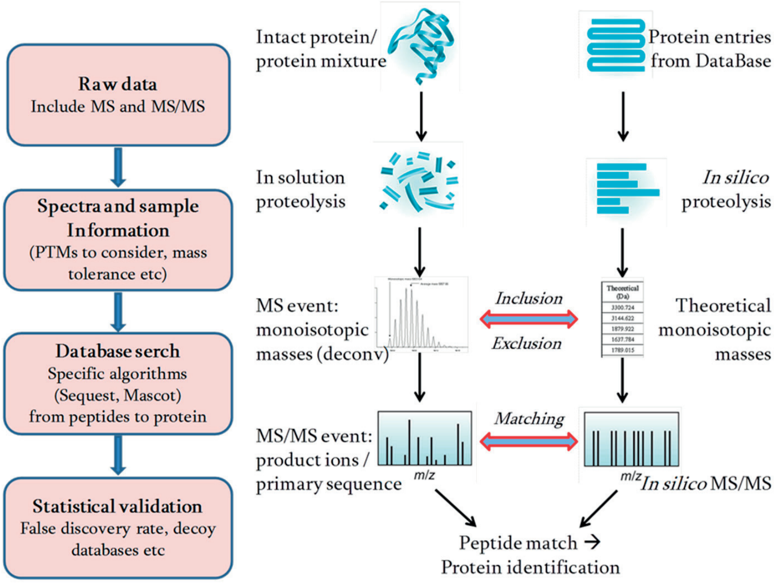


Figure 2.5. Schematic representation of an automated database search workflow

After the mass of the precursor is determined through deconvolution of the survey scan, and subsequently used to restrict the „search space“ to a sub-group of potential peptides (according to mass accuracy tolerances set by the experimenter), the tandem mass spectrum is matched against all the theoretical ones generated from the candidates (using specific mass tolerances for product ions), finally leading to the identification of a candidate peptide sequence. Various statistical methods are used to validate the candidate peptide sequences and corresponding protein matches (according to the *'two peptide per protein'* rule introduced in BUP).

One of the most utilized statistical methods for filtering the data is *false discovery rate* (FDR, see Nesvizhskii [109] and references therein). Simply speaking it is a rate of false-positives (arbitrary value by convention set to 1-2 %) which is estimated based on results obtained by searching against a *decoy* database (containing „false" protein sequences generated by scrambling or reversing the original ones), or in terms of MS, property of MS/MS spectra that defines expected proportion of incorrect assignements. The majority of database search algorithms initially developed for processing of BUP data, such as SEQUEST [110], Mascot [111], X!Tandem [112], OMSSA [113] and Andromeda, or proteomic computational platforms like MaxQuant [114] and the Trans-Proteomic Pipeline, readily calculate the FDR. However, it is to be stated how the FDR concept (as well as the entire database search pipeline) works under the assumption that peptides are generated with a residue-specific protease and hence that the set of cleavage rule is known. However, proteases so far described for MDP are generally partially-specific or even non-specific [37, 115-117]. In such case, it is almost impossible to create a reversed database, and the theoretical peptide candidates can be as short as one amino acid residue. With the exponential increase in the number of candidates in decoy database, the traditional FDR procedure applied to BUP would now remove an insufficient number of false-positives, and still render final results ambiguous.To date, there is no dedicated database search algorythm for MDP. Most of the aforementioned BUP softwares fall short in interpreting these data, primarily because the search space is generated strictly using a well-defined cleavage rule, with the consequence that many peptides generated by unspecific cleavage, or that carry unexpected or not-annotated PTMs produced by non-enzymatic cleavage (as in the case of certain chemical cleaving agents, that can modify peptides both through the main cleavage mechanism or also through side reactions) are excluded by the search process and therefore cannot be identified.

On the other hand, TD softwares can be adapted to the analysis of MDP experiments. This fact might be explained considering that large polypeptides like those typical of MD share several characteristics and MS-related features with small intact proteins analyzed in TD proteomic experiments: for instance, their isotopic distribution resembles that of a protein, with the monoisotopic peak being low abundant and located far from the apex of the isotopic distribution, differently from that of a short tryptic peptide, which is generally the most abundant of the isotopomers. Furthermore, the product ions generated by the activation of large MDP peptide ions are often multiply charged, similarly to those of proteins, and tandem mass spectra can be particularly convoluted, with overlapping ion species (a situation rarely encountered in BUP experiments). Hence, if we define three operations required for TD data analysis, and specifically: (i) *peak picking*, (ii) *spectral deconvolution* (for both MS and MS/MS) and (iii) *product ion assignment*, it is clear that the algorithms used for TD mass spectrometry can be successfully applied also to MDP. Particularly, steps (i) and (ii) rely in TD on the *estimation* of the protein monoisotopic mass applying an isotopic fitting model (for instance based on *averagine* [118], an amino acid of mass equal to the weight average of all amino acids present in proteomes, molecular formula C4.9384 H7.7583 N1.3577 O1.4773 S0.0417 and average mass of 111.1254 Da), rather than the identification of the monoisotopic peak in the *m/z* space as in BUP. Moreover, similar concepts are used also for the deconvolution of tandem mass spectra, which are acquired exclusively using high resolution MS (whereas in BUP they are generally acquired in low resolution). Finally, it is important to note that methods for the estimation of the isotopic distribution of large polypetide precursor or fragment ions would often not be accurate enough for small peptides [119]. In other words, TD data analysis methods would not perform perfectly for BUP analysis.

In the works presented in this Thesis, data analysis was therefore carried out using mostly a *combination* of the BUP-dedicated algorithm Sequest (Papers III and IV), and softwares tailored for processing of TD data, such as MS Align+ [120] (for data interpretation where the detection of unexpected or not-annotated PTMs was crucial, Paper V)), ProSight PC, ProSight Lite (for the generation of graphical fragmentation maps)[121] and, finally, the recently introduced MASH Suite [122] (applied to the MD analysis of immunoglobulins, Papers VI and VII).

Chapter 3. Instrumentation

3.1. Electrospray ionization

Coupling of electrospray ionization (ESI) to (biomolecular) mass spectrometry was introduced in the 1980's by J. B. Fenn. In 2002 he was awarded the Nobel Prize in chemistry for his work on ESI MS. Briefly, ESI is a *soft ionization* technique which entails a non-destructive analyte charging phenomenon, resulting in formation of multiply-charged ions [123] (*vide infra*). ESI revolutionized the field of MS by enabling the analysis of large biomolecules. The *m/z* ratios of biomolecular ions were shifted couple of hundreds, even thousands Th down the *m/z* axis, thus allowing for ion detection within the nominal mass limits of mass analyzers such as quadrupole based ones. Further success of this method was due to the implementation of micro and nano-electrospray (known as μESI with flow rates in the μl/min range and nESI with flow rates in the nl/min range) ion sources described by Wilm and Mann [124, 125] that facilitated on-line coupling of ESI with orthogonal front-end to MS separation techniques (such as reversed-phase liquid chromatography (RP-LC) and capillary electrophoresis (CE)) that require a constant flow of the liquid. Underlying concept of ESI stands on the basic principles of electrochemistry. Electrospray ion source can be considered as a controlled current electrolytic cell where high voltage (kV) is applied to a liquid and the process of protonation/deprotonation occurs. In terms of electrochemistry, if generation of protonated species $[M+zH]^{z+}$ takes place, the tip of the capillary acts as an anode and MS source inlet as its counterpart electrode (cathode) or inversely, cathode and anode, respectively, when deprotonated species $[M-zH]^{z-}$ are formed.

The ionization process depicted in Figure 3.1. starts with the accumulation of the charge at the liquid surface under the influence of the electric field (which is established between the tip of the capillary and the MS). In addition to the applied potential difference, liquid charging is aided by the presence of low concentration of acids (such as 0.1 -1 % formic acid) that act as proton donors. Ramping the voltage (typically between 1.5- 2 kV and 2-6 kV for nESI and μESI, respectively) causes the supposed spherical droplets at the tip of the probe to elongate into the characteristic shape called the *'Taylor cone'*.

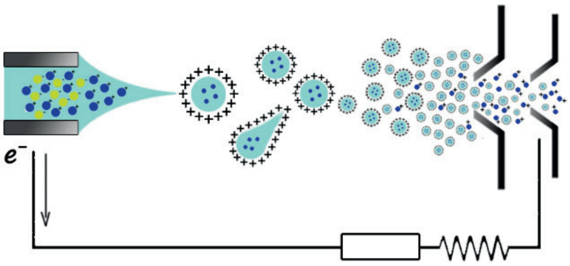


Figure 3.1. Graphic representation of an electrospray ion source interfaced to a mass spectrometer and the electrochemical process of ESI.

At an 'onset voltage' pressure is higher than the surface tension; droplets are released and the spray is formed. Evaporation of the excess of solvent contained in the charged parent droplets is aided by the vicinity of the heated inlet capillary (between 0.2-2 cm) and additionally by usage of the countercurrent gas (such as nitrogen). In practice, organic compounds such as methanol or acetonitrile are used as solvents in ESI; their surface tension is lower than that of water hence, those are easier to evaporate. Desolvation process causes the charged droplets to progressively shrink, thus increasing their charge per unit ratio. Once the droplets reach the *'Rayleigh limit'* [126] defined as the maximum amount of charge a liquid droplet could carry they further break down in a process of *Coulomb fission* [127]. Moiety of the offspring droplets (charged analyte ions) that has either net positive or negative charge (depending on the analyte of interest and the setup applied) enters the MS. The resulting ESI mass spectrum is composed of signals (peaks) corresponding to the different charge states of the same analyte ion generating the *charge state envelope*. It is noteworthy to mention how the charges ESI generates on the ions are merely a consequence of the described charge accumulation in the droplets and redox processes occurring at the probe tip (*vide supra*), and do not exactly reflect the charge state of the analyte in solution, as showed by *Kelly et al* on protein myoglobin [128].

Finally, it is important to mention the ESI current sensitivity dependence on the concentration rather than the total volume amount of the sample. This dependence gave rise to the further development of the nESI technique which coupled with MS instruments nowadays allows detection limit (or lower limit of detection, LOD) in ranges of femtomoles [129] and even attomole detection for certain analytes has been reported [130].

3.2. Hybrid mass spectrometers

Mass spectrometers use a combination of electric and magnetic field generated within electrostatic lenses and other *ion optics* elements to confine and manipulate analyte ions. As mentioned in the previous Chapter, a proteomic-oriented mass spectrometer has to allow both the detection of intact ions and also their controlled fragmentation. To maximize the flexibility in performing all these different operations, MS instruments have been equipped with multiple mass analyzers in series. Such instruments are known as *hybrid mass spectrometers* [131]. Hereinafter, the two mass analyzers present in the instrument used for this Thesis, a linear ion trap Orbitrap Fourier transform mass spectrometer (LTQ Orbitrap Elite, Thermo Scientific), will be described in the general terms of their respective working principles.

- Linear ion trap

Stand-alone linear quadrupole ion trap (LIT) mass spectrometers were introduced by Thermo Finnigan in 2002, as a high capacity alternative to three dimensional quadrupole ion traps (QIT) [132]. The claimed advantages of LITs over QITs include increased ion storage volume (reduction in space charge effect), enhanced sensitivity and higher trapping efficiency, as previously described by Syka and Fies [133]. A LIT as implemented by Thermo Finnigan consists of a quadrupole with rods split into three sections and front and backlenses (not shown), Figure 3.2.

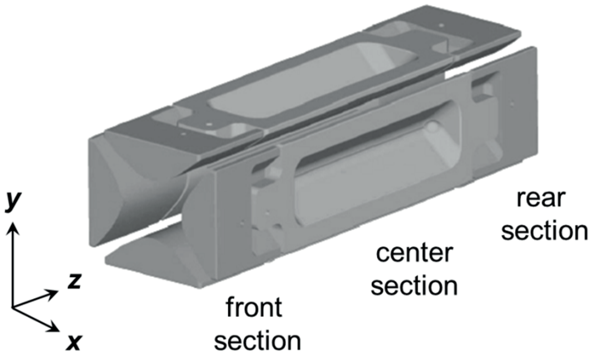


Figure 3.2. A perspective view of a linear ion trap as it is implemented within the LTQ Orbitrap Elite system. The two x-rods are provided with slits to enable radial resonance ion ejection. The hyperbolic rods are divided into inner and outer sections to add a constant voltage in z-direction that causes ion trapping. Adapted from Schwartz and Senko [134].

Radiofrequency (RF) potentials with opposite phases are applied to the two pairs of quadrupole rods to confine ions in the radial direction Direct current (DC) potentials are applied to the separate sections of the rods confine ions in the axial direction (Figure 3.3.). As with QITs, a bath gas provides collisional cooling of the ion cloud.

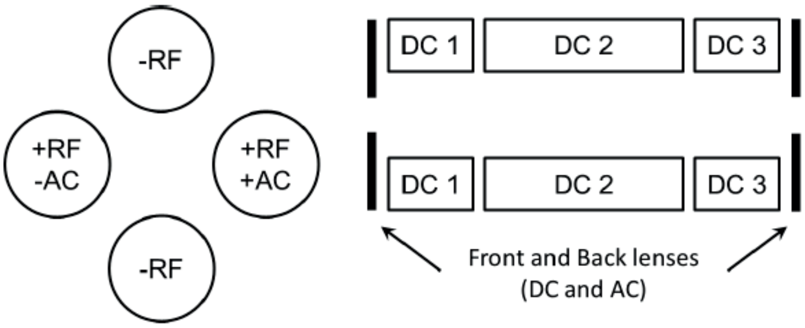


Figure 3.3. Schematic showing the application of RF, DC and AC potentials needed for operation of a linear quadrupole ion trap. Adapted from Schwartz and Senko [134]

Auxiliary alternating current (AC) waveforms with opposite phases are added to one pair of opposite rods for ion isolation and ion activation. Mass analysis is accomplished by ejection of ions (mass selective instability) through slots in the center section of the quadrupole rods. DC potentials on the front and rear lenses gate the flow of ions into and out of the ion trap. To enable ion/ion reactions, AC waveforms are also added to the front and back lenses of the LIT. This provides an axial trapping field for simultaneous storage of cations and anions.

Stability of the ions within the linear ion trap is described by Mathieu equations that operate with two dimensionless parameters, a and q:

$$a = \frac{4QU}{mr_0^2 2\omega^2} \qquad (3.1.)$$

$$q = \frac{2QV}{mr_0^2 \omega^2} \qquad (3.2.)$$

where Q is charge, U is DC voltage, V is RF voltage and ω is oscillation frequency. Plotting a as a function of q gives a stability diagram (Figure 3.4.)

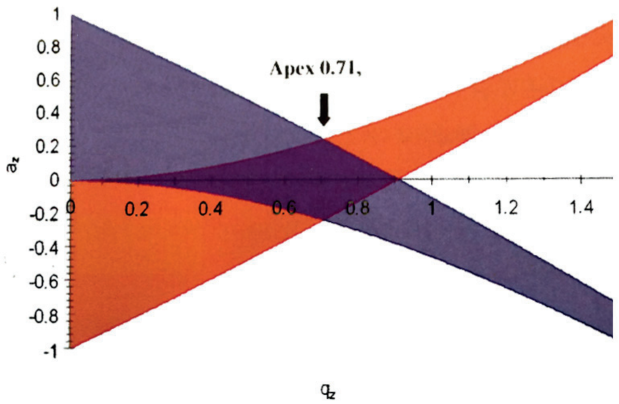


Figure 3.4 Mathieu stability diagram for linear ion trap. Area indicated in orange shows combinations of a and q which provide a stable trajectory in x- direction, whereas purple area indicates combinations of parameters under which trajectories are stable in y-direction. Figure adapted from www.planetorbitrap.com.

This diagram is a graphical representation of all solutions to Mathieu's equation for linear ion trap. Overlap of regions in Figure 3.4 indicates combinations of a and q under which ions are stable inside the trap. By ramping the AC voltage (known as resonance ejection voltage) ions are passing the q = 0.908 instability barrier and are axially ejected.

3.3. Fourier transform mass spectrometry (FTMS)

High mass resolution and accuracy measurements in hybrid mass spectrometers are obtained thanks to specific mass analyzers positioned downstream of the low resolution mass analyzer (with respect to the ion inlet). Typically, two types of mass analyzers are capable of achieving resolution >20'000: time-of-flight (TOF) and Fourier transform (FT)-based ones.

Conceptually, the working principle of TOF mass analyzer is relatively simple [135, 136]. Ions with identical starting position and velocity are accelerated to a final kinetic energy of zeV through a field-free space, such that their time-of-flight when they will reach the ion detector can be described as:

$$t = d\sqrt{\left(\frac{1}{2zeV}\right)}\sqrt{\frac{m}{z}} \qquad (3.1.)$$

where d is the length of the flight tube, z is ion's charge state, e is an elementary charge, V is an acceleration potential, and m is the mass of the ion of interest. TOF mass analyzers are based on a "single-ion counting" principle, as every single ion may generate a signal by reaching the detector, if it has enough energy.

Conversely, FT-based mass analyzers work by measuring frequency of motion of ion packets that are injected into the mass analyzers, where they are trapped, excited and, for a certain time, follow a specific periodic motion dictated by the use of magnetic or electric fields [137]. We can distinguish between two FT-based mass analyzers: i) the magnetic field-based ion cyclotron resonance (ICR) mass analyzer; and ii) the electrostatic field-based

Orbitrap mass analyzer. In FT-ICR MS, ions move immerged in a static, spatially uniform magnetic field B, and their cyclotron frequency is defined as:

$$\nu_c = ezB/2\pi m \qquad (3.2.)$$

with e being the elementary charge and m the ion's mass. In reality, due to the presence of electric fields in ICR mass analyzers, for example trapping and space charge fields, the measured frequency is the reduced cyclotron frequency, which is shifted by the magnetron frequency compared to the cyclotron frequency. Importantly, frequency of ion motion is inversely proportional to the m/z ratio of ions. As a consequence, resolution of FT-ICR MS drops following the same dependence with an increase in m/z.

In the Orbitrap, ions are confined using a quadro-logarithmic electrostatic potential. The frequency of ion oscillation along the central electrode (*vide infra*), also known as axial oscillation frequency, is determined as:

$$\nu_z = \sqrt{\frac{ezk}{m}}\frac{1}{2\pi} \qquad (3.3.)$$

where k is a constant describing the field curvature. Contrary to FT-ICR MS, ion frequency is inversely proportional to $\sqrt{\frac{m}{z}}$ and thus resolution in Orbitrap FTMS reduces substantially slower than in FT-ICR MS as a function of m/z.

For both mass analyzers, the signal of ion packets is recorded as a current induced by the passage of ions themselves close to pairs of electrodes. Such signal is then digitized and stored as a *time-domain signal*, commonly referred to as “transient signal” due to its characteristic decay in amplitude over time. Time-domain signals are Fourier transformed to obtain frequency domain spectra, which can be easily converted into m/z spectra using calibrants of known m/z and frequency as well as dependences presented in equations 3.2 and 3.3. In the following subsection we focus on one of the FTMS instruments used in this Thesis.

3.3.1. Orbitrap-based mass spectrometer - design and working principle

Our lab is equipped with a state-of-the-art hybrid LTQ Orbitrap Elite mass spectrometer from Thermo Scientific (Bremen, Germany), particularly suited for the selected research direction. The hybrid architecture of this instrument includes two different mass analyzers, a dual-pressure linear ion trap (LTQ) and the compact, high-field Orbitrap, arranged in series, Figure 2.

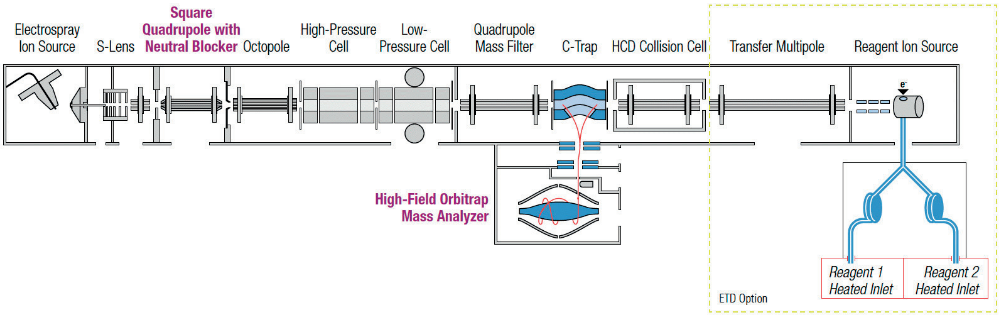
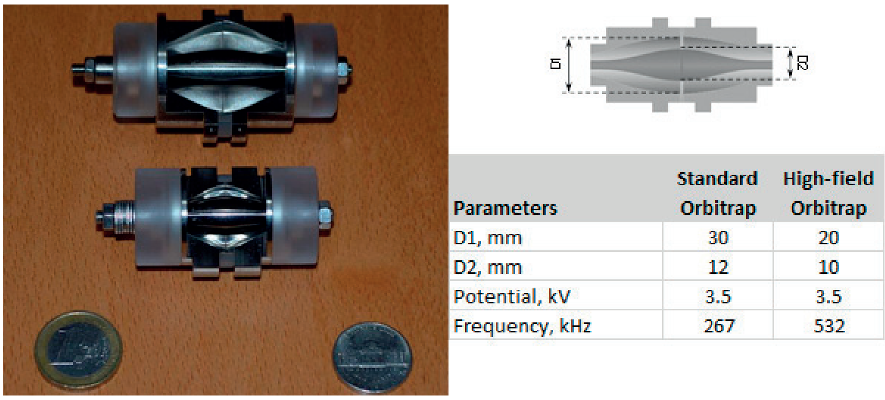


Figure 3.5. Schematics of the hybrid LTQ Orbitrap Elite mass spectrometer.

The Orbitrap technology, developed in 2000 by Makarov [138] represents the latest major achievement in mass spectrometry in general and in Fourier transform mass spectrometry (FTMS) in particular. This technology is based on the recording of the induced current produced by ion clouds trapped into the electrostatic Orbitrap cell. As in every FTMS-based mass analyzer, the final resolution is proportional to the length of the recorded transient signal.



Parameters	Standard Orbitrap	High-field Orbitrap
D1, mm	30	20
D2, mm	12	10
Potential, kV	3.5	3.5
Frequency, kHz	267	532

Figure 3.6. Comparison between standard and high-field Orbitrap mass analyzers. Left panel, picture of both the mass analyzers, indicating the size reduction occurred passing from the old to the new generation of mass analyzers. Right panel, some dimensions and other features distinguishing the two Orbitrap generations.

Among the figures of merit of the Orbitrap Elite series there are: the advanced signal processing algorithm (enhanced FT, or eFT) for absorption mode-type FT providing, together with increased frequency of ion axial oscillations, improved resolution up to 480’000 at m/z 400 (transient length 1536 ms), high sensitivity of the front-end ion optics equipped with the S-lens (stacked ring ion guide) [139].

Furthermore, the LTQ Orbitrap Elite allows different activation methods for tandem mass spectrometry (described in the Chapter 2). CID [73] is performed in the high pressure region of the LTQ (gas used: helium), as well as electron transfer dissociation (ETD) [88]. In the LTQ Orbitrap Elite the injection of the ETD reagent radical anions, fluoranthene, is performed from the back of the instrument. The amount of fluoranthene injected is controlled by the automatic gain control function (AGC) in the LTQ, exactly like for the target value of precursor ions for MS/MS or ions analyzed in the survey scan.

Finally, this mass spectrometer can perform also HCD[76]. Originally developed in the C-trap, this fragmentation method is now performed in a specific multipole trap positioned after the C-trap, see Figure 3.1. Importantly, product ions generated by HCD are analyzed uniquely in the Orbitrap mass analyzer. Nevertheless, HCD activation is fast, so that it is still possible to apply it to bottom-up proteomics as demonstrated by Michalski *et al.* As an example, by using a top-15 routine (which consists in the fragmentation of the 15 most intense precursors detected in the survey scan), the final duty cycle is of 3.3 s when the survey scan is performed at 240’000 resolution (at 400 *m/z*) and product ion detection following HCD is performed at 15’000 resolution (at 400 *m/z*). In addition to being a collision cell, HCD cell can be used for trapping of the large ions, to thermalize them *via* mild collisions with the bath gas, prior to squeezing of the ion package in the C-trap, thus improving their transfer efficiency when injected into the Orbitrap analyzer.

3.4. Limitations of MS experiment in an Orbitrap FT MS

FT-based instruments in general and Orbitrap-based instruments in particular are current state-of-the-art instruments used in proteomics research and could be considered *condicio sine qua non* tools for structural analysis of compounds for which high resolution is required. However, these instruments do meet their practical limitations. The mass analyzers described in this Chapter work by spatially confining ions within a limited space (gaining the generic name of *ion traps*). Although the motion of the ion inside the mass analyzer can be in a first approximation considered as dictated by a specific electric or magnetic field (for instance, the magnetic field produced by superconductive magnets used in ICR FTMS), ions also interact with each other. The observable ions behavior produced by these interactions, and particularly by Coulombic forces, take collectively the name of *space charging*.

The most obvious space charge effect affecting high-resolution FTMS mass analyzers, ICR and Orbitrap cells, is known as ion coalescence. Originally discovered and studied in ion cyclotron resonance FTMS [140], coalescence has been recently reported also in Orbitrap-based mass spectrometers [141]. This effect, that for ICR seems caused by phase-locking of ion populations with very similar motion frequencies (and, thus, originally characterized by very close *m/z* position in the mass spectrum), produces distortions in the peak positioning, as the peaks of neighboring analytes with very similar masses starts getting closer in the *m/z* space until they eventually become completely superimposed and, therefore, undistinguishable. Importantly, not only the two species will seem like one in the case of complete coalescence, but even when the phenomenon is only partial it still has detrimental effects on mass accuracy.

Although a recent study would suggest that this phenomenon should not be a problem for general bottom-up studies [142], it has to be considered that ion coalescence, as any space charge effect, depends on the number of ions introduced in the mass analyzer. Therefore, selected applications differing from shotgun BUP experiments might suffer problems if a large population of ions within a small *m/z* window has to be used [143], for example in the attempt of increasing the SNR. For example, if too many ions of a single charge state of a highly-charged protein are simultaneously introduced in an Orbitrap (or ICR), ion coalescence might affect the relative positioning of the isotopologues composing the isotopic distribution of a single charge state (which are already extremely close in the *m/z* axis), with the ultimate effect of preventing the correct charge assignment of the ion cluster and thus a possible mistake in the calculation of the protein mass.

Another application where it is of fundamental importance to keep ion coalescence under control is lipidomics, particularly during experiments aimed at distinguishing between lipidic species separated in mass by only a few milliDaltons.

Chapter 4. Towards a ‘stand-alone’ MDP pipeline

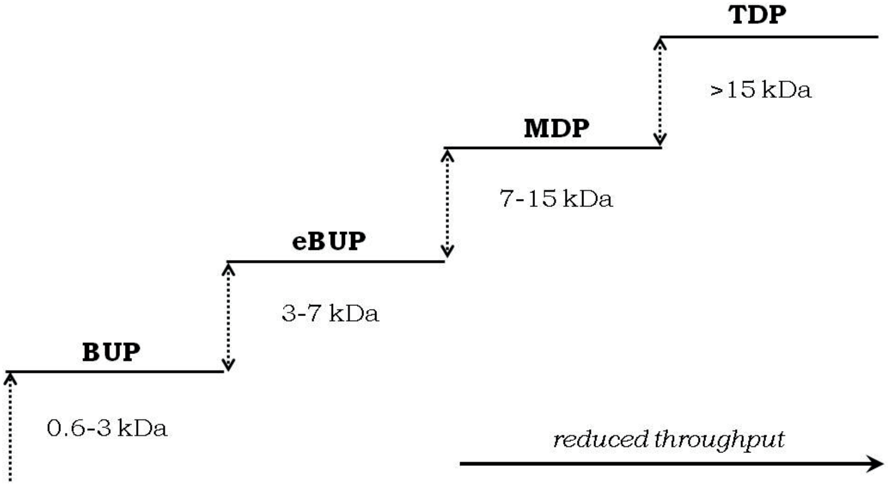
This Chapter is dedicated to the development and implementation of a rationale enabling a new subdomain, middle-down proteomics (MDP, as introduced in Chapter 1), in a well-defined field of mass spectrometry-based proteomics. To do so, we explored various *middle-down* (MD) *approaches* that will be presented through research articles in following Chapters. Throughout this Chapter we will refer to MDP when talking about the high-throughput shotgun identification of proteome(s), whereas MD will refer to distinct approaches within MDP. Herein, we addressed the initial requirements that would allow achieving the following goals set for MDP pipeline that targets 3-15 kDa peptides:

i) Identify the most optimal target backbone cleavage site(s) in proteins for optimizing the desired peptide length and amino acid distribution,
ii) Identify and characterize a suitable protease or other cleaving agent,
iii) Optimize front-end separation(s) for peptides in a targeted mass bin,
iv) Determine MS instrument parameters set to analyze resulting peptides,
v) Optimize data analysis workflows to improve analytical characteristics of MD mass spectrometry, such as sensitivity and spectral dynamic range

The final goal of the study was to better characterize complex biological systems by offering insights that currently employed approaches cannot provide. Hereinafter reported considerations are the initial premises later translated into research articles enclosed at the end of this Chapter:

- Proteome Digestion Specificity Analysis for Rational Design of Extended Bottom-up and Middle-down Proteomics Experiments (*Paper I*)
- Practical Considerations for Improving the Productivity of Mass Spectrometry-based Proteomics (*Paper II*)

Tailoring the proteolysis site to yield longer average peptides is the driving force for the development of MDP. Frequency of amino acids differs from proteome to preoteome, often even within sub compartments of a proteome. Hence, this fact alone implicates *a priori a* potential need for various proteome–specific MD approaches to allow for a MDP on a given organism. At the same time, sample preparation, peptide separation, ionization conditions, fragmentation parameters, data acquisition and data analysis must be adjusted to analysis of long peptides. Figure 4.1 shows peptide ranges classified on the basis of the type and figures of merit of the mass spectrometers that can be employed, chromatography considerations, and available database search algorithms for data analysis. Note how here we categorized the proteomics approaches in four mass bins according to the molecular weight of proteolytic pool. The fourth category, extended bottom-up proteomics (eBUP) derived as a sub-category of MDP after initial considerations of available pipeline in terms of the instrument performance and the choice of front-end chromatography (reversed-phase LC (RP-LC) *vide infra*). The former meets its limitation in current software feature that does not allow change of settings (such as collision energy, isolation window, target value for MS/MS, number of microscans, *etc.*) 'on a fly'. Hence, oftentimes analysis in MDP range has to be carried out in minimum of two repetitions, to accommodate the optimal settings for all peptides throughout the mass range. RP-LC on the other hand, emphasizes the hydrophobicity-based elution, while size-exclusion comes as a secondary, not strongly pronounced effect. This in turn means how shorter (30-50 residues) peptides could elute in close retention time window with longer (50-100 kDa) ones, if their hydrophobicity index is close. In respect to those two aspects, it seemed favorable to further split broad range of proteolytic pool of MDP (3-15 kDa) into lower molecular weight bin (3-7 kDa) and use that one as initial testing bed for MDP pipeline optimization.



LC separation	C18, SCX, WAX, HILIC	C18, C8, Zorbax C3, SCX, WAX, HILIC	C8, Zorbax C3, C4, monolithic	Zorbax C3, C4, monolithic
MS analyzer	Ion trap, QQQ, QTOF, FTMS	QTOF, FTMS	QTOF, FTMS	QTOF, FTMS
Activ. method	CID	CID, HCD, ExD	HCD, ExD	HCD, ExD, EThcD
Data analysis	Sequest, Mascot, X!Tandem	Mascot, X!Tandem, Sequest, MS-Align+	MS-Align+, ProSight 3.0	MS-Align+, ProSight 3.0, ProSightLite

Figure 4.1. Classification of mass spectrometry-based proteomic approaches based on the molecular size of the analytes. Adapted from Laskay et al. [144].

4.1. Liquid chromatography: adopted parameters and reflections on mass spectrometric analysis of large peptides.

In general, the application of separation techniques to complex proteomic samples is essential for successful mass spectrometric analysis. Front-end in-solution separation enhances the dynamic range of detection, minimizes ion suppression effect during the electrospray ionization process and greatly increases the depth of proteome analysis.

The most commonly employed separation technology in MS-based proteomics is liquid chromatography (LC), and specifically reversed-phase LC (RPLC). As any chromatographic technique, LC is performed by distributing the analytes between two phases: the mobile phase, in this case a liquid which carries the analytes, and the stationary phase, which is fixed inside a hollow column through which the mobile phase is forced. The *distribution coefficient* of each analyte between the two phases, which determines whether the analytes will be preferentially retained by one or the other phase, can be changed, primarily, by varying the composition of the mobile phase. In RPLC the analytes, in our case peptides, are characterized by certain degree of hydrophobicity. Therefore, they are initially loaded onto the chromatographic column using a very low percentage of organic solvent in the mobile phase, so that they will preferentially interact with the stationary phase, which is composed of hydrophobic material. The percentage of organic component in the mobile phase is then raised over time, progressively moving the distribution equilibrium of peptides from the stationary towards the mobile phase.

Traditionally, columns are packed with particles conjugated to specific functional groups. For reversed-phase LC, silica-based microparticles are decorated with linear alkanes of different length, the most common being C18, C8 and C4 (where 4, 8, or 18 denotes the number of carbon atoms in the chain). Longer alkane chains are more hydrophobic, being ideal for polypeptides of reduced size (e.g., tryptic peptides), whereas shorter chains are typically used for longer polypeptides and proteins. The choice of the stationary phase is important, along with other parameters (*vide infra*), to determine the final *chromatographic resolution*, defined as capability of separating the elution peaks of two different analyte molecules. In proteomics, high chromatographic resolution is of fundamental importance as it allows to minimize the number of peptides that are simultaneously directed towards the mass spectrometer, greatly reducing problems of signal suppression and overlapping and ultimately leading to the detection of a higher number of species with high specificity and selectivity.

Although a comprehensive discussion about liquid chromatography exceeds the scope of this Thesis, to better understand the choices of columns/stationary phases selected for the MDP platform (see Figure 4.1, bottom), and also discussing potential limitations of the currently available commercial columns, it might be important to briefly introduce the Van Deemter equation [145-147], which describes how the *height equivalent to a theoretical plate* (HEPT) is related to the linear velocity of mobile phase (v) in relation to three parameters (A, B and C):

$$HEPT = A + \frac{B}{v} + Cv$$

The theoretical plate is the minimal stage or portion of the column in which the two phases establish an equilibrium. Increasing the number of theoretical plates enhances the chromatographic performance characteristics, including the resolution. Therefore, it is important to minimize HEPT as, for a column of given length l, the number of theoretical plates N is given by

$$N = \frac{l}{HEPT}$$

In the Van Deemter equation, the three parameters A, B and C refer to physico-chemical properties of the column. Specifically, A is the Eddy-diffusion parameter, which is related to the quality of column packing by measuring possible variations in the analyte flow path due to inhomogeneities in the stationary phase; B represents the longitudinal diffusion coefficient, which is contrasted by increasing the mobile phase linear velocity, effectively concentrating in a tighter pack the analyte molecules; finally, C indicates the mass transfer coefficient, originated by the porous nature of the column packing material, and that is linearly proportional to the linear velocity of the mobile phase. Note, that increasing l to obtain a higher number of theoretical plates is a strategy that is widely pursued, but found its limit in the backpressure generated by the stationary phase, which is directly proportional to the column length.

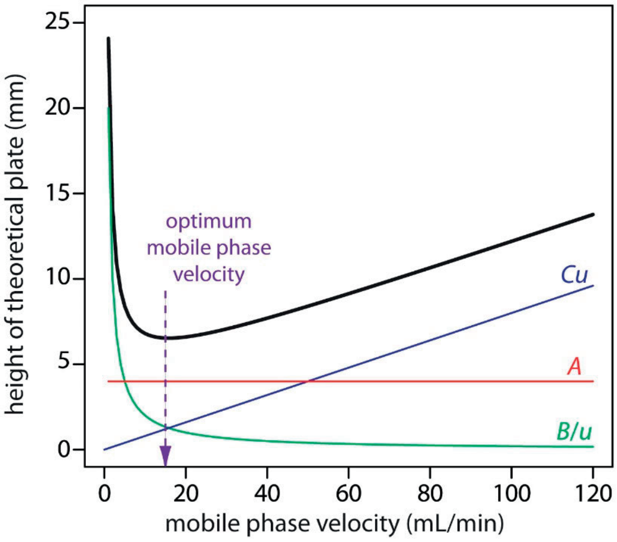


Figure 4.2. Plot illustrating the Van Deemter equation. The three components A, B/v and Cv are indicated in three different colors, while the final HEPT function is displayed in black. The dotted line indicates the optimal linear velocity of the mobile phase, which corresponds to the minimum in the HEPT curve.

For our MDP platform, we apply on-line LC nanospray (nESI) MS (i.e., with the column outlet directly linked to the nESI source) using a Dionex Ultimate 3000 (Thermo Scientific). This LC system is equipped with a nano pump (allowing for flow rates between 0.05 and 1 µl/min) and a micro pump (operating at >1 µl/min). The former is used for peptide separation on analytical nano columns by reversed-phase LC, whereas in our setup the latter is used to wash peptides trapped on a guard column right after injection and prior to analytical separation. The choice of columns for MDP was done taking into account the characteristics of middle-down peptides, which are larger and more hydrophobic than typical bottom-up ones. Hence, highly hydrophobic stationary phases such as C18 were excluded, to prevent irreversible binding of peptides to the column stationary phase. From a mass spectrometry point of view, slightly broader LC elution peaks resulting from the weaker hydrophobic interactions might be advantageous on the MS/MS working timescale for large peptides. In fact, the longer the peptide, the more charges it normally carries, and the higher the chance that its tandem mass spectrometry generates partially overlapping multiply charged product ions; therefore, high resolution is required for both MS and MS/MS in MDP experiments. This is translated into the need of using the high resolution mass analyzer for both survey scans and product ion detection, reducing the throughput. Furthermore, convoluted tandem mass spectra might require the implementation of strategies aimed at improving the spectral signal-to-noise ratio (SNR) that are generally detrimental for the speed of analysis (*vide infra*, paragraph 4.2).

With all these considerations in mind, a silica particle-based nano-column with C8 stationary phase was used in combination with a C8 trap column for most of MDP applications. The adopted flow rate was 0.8 µl/min, which was required by the large pore size (300 Å) to maintain the linear velocity of the mobile phase sufficiently close to optimal values, avoiding detrimental effects on chromatographic resolution, without leaving the nano-flow rate regime, which is favorable over the micro-flow rate regime in terms of ionization efficiency and sample consumption.

For selected applications, finally, a monolithic column was used [148]. Monolithic columns are produced by polymerizing the stationary phase directly inside the column. Differently from packed column, the stationary phase is in this case non-porous, and therefore the mass transfer is highly limited, with obvious benefits for the achievable resolution. Moreover, the monolithic stationary phase of these columns is characterized by high physical-chemical stability and allows to use columns of superior length due to the reduced back pressure. The monolithic column used in the here described MDP pipeline is a commercial PepSwift from Dionex, consisting of a cross-linked co-polymeric stationary phase based on divinylbenzene and polystyrene. Due to the relatively large internal diameter (100 um), the column was used at a flow rate of 1 µl/min, producing a slightly reduced ionization efficiency in comparison to that of its C8 counterpart.

4.2. Evaluation of MS/MS parameters for MDP

CID remains the fragmentation technique of choice for BUP, due to its speed, efficiency for short peptides and the possibility to use the LTQ for low-resolution product ion detection with high scan rate. For MDP applications, however, as the average size of analyzed peptides requires fragmentation methods leading to extensive sequence coverage and necessarily product ion detection performed in the Orbitrap, different techniques such as HCD and particularly ETD, could be considered as more suitable ion fragmentation methods.

Differently from CID, ETD produces extended sequence coverage even for extremely large peptides (or proteins), and the sequence information that it produces is generally not limited to the N- and C-termini. Furthermore, as mentioned in Chapter 2, ETD typically preserves labile PTMs. This feature is of a particular importance considering that one of the envisioned advantages of middle-down over the classical bottom-up approach is the possibility of determining with higher precision possible proteoforms, whose classification is essentially based on PTMs localization. It is to be noted how the loss of labile PTMs with CID is true for small systems such as peptides. However, if we recall how energy is randomized through 3N-6 or 3N-5 vibrational modes in a non-linear or linear molecule, respectively (where N indicates the number of nuclei present in the molecule), [63] it is obvious how for large biomolecules the number of available vibrational modes is significantly higher. This in turn means how the randomization of energy does not lead to the cleavage of the weakest bond present in a side chain, thus allowing preservation of PTMs. Given that percent fragmentation in ETD is dependent on the charge state of the selected cation, the higher average charge state of peptides in the mass range of MD experiments should be favorable for this activation method and provide better results than shorter and less charged peptides typical of BU. The ion-ion interaction time is decreasing with the increase in the precursor charge state, and this ensures that the duration of MS/MS event is substantially reduced (8-10-fold; 10 ms used in MD approach [149] *vs* 80-100 ms in BU approach [92]) by passing from BU to MD ETD LC-MS/MS). The overall fragmentation efficiency is lower than in energy threshold-based activation techniques (CID, HCD, IRMPD), which means that strategies aimed at improving the spectral signal-to-noise ratios are particularly important to achieve a high-quality ETD MS/MS spectrum. In hybrid FTMS instruments like the one used for this Thesis, it is possible to average several time-domain transients (the so-called *micro scans*) prior to their Fourier transformation (FT). This procedure is detrimental to the overall analysis throughput, as it requires a full cycle of ion injection, fragmentation and product ion detection to obtain the additional time-domain signals (micro scans). Furthermore, this strategy increases spectral SNR only with the square root of the number of averaged micro scans.

With all this considered, despite the reduced complexity of the proteolytic peptide landscape in MD compared to BU, which should help the peptide separation, the time constraints imposed by the LC-MS/MS setup usually employed in any proteomics study render the actual implementation of ETD for MDP challenging. On the other hand HCD, despite being a low-energy fragmentation method, should still be favorable over CID for MD peptides, due to the fact that applied energy is calculated with consideration to the precursor mass to charge ratio and charge (as discussed in the Chapter 3) and therefore should be more sensitive and adaptable to the broad span of charge states and peptide masses that is intrinsic property of MD proteolytic pool, and in turn result in more successful overall fragmentation per LC run. Additionally, HCD is on average faster than ETD (0.1-1 ms *vs* 0.1 – 100 ms per ion-neutral and ion-ion interaction event, respectively, not accounting for the radical anion transfer and accumulation time in the LTQ) and due to the position of the HCD collision cell it has shorter path of the fragments to the orbitrap mass analyzer, which is of importance for the successful transmission of larger product ions, especially the (unstable) radical ones. Both the short duration of ion-ion interaction in HCD, as well as the shorter path for transmission of product ions to the analyzer cell render the overall duty cycle shorter and in turn allow more MS/MS scans and/or micro scans per chromatographic elution peak. Shortcoming of HCD however lies in the fact that it does not preserve the labile PTMs which is detrimental to the PTM assessment goal of MD, as well as the fact that this activation method is known for prevalence of the *y*-ion series, whereas the *b*-series one is oftentimes missing.

It remains to be stated how, both ETD and HCD individually, or combined into a hybrid technique (namely EThcD [150, 151]), yield better results in analysis of large biomolecules, in both MD and TD fashion, than other available ion activation methods, despite the here underlined current practical difficulties which are mainly a result of instrument software and limitations of the separation methods preceding the MS. Additionally, certain considerations listed here are made with particular regards to the hybrid instrument used for this Thesis (i.e., Orbitrap Elite mass spectrometer), and do not apply to some of the new FT-based hybrid instruments, such as those based on the Quadrupole Mass Filter-Orbitrap-Linear Ion Trap platform. The latest iteration of this line of Orbitrap-based mass spectrometers (Orbitrap Fusion Lumos, Thermo Scientific), for instance, is not only natively equipped for EThcD, but can also perform the so-called "high-capacity ETD", which consists in ETD performed in the LTQ as in all other LTQ-Orbitrap instruments, but with precursor cations trapped in the central section of the high-pressure chamber, which can store a higher number of ions (three times higher, according to estimations) compared with the back section where traditionally precursor ions are stored [152]. As a final result, high-capacity ETD shows about 3-fold improvement in spectral signal-to-noise ratio of product ions when compared with the standard counterparts for the analysis of intact proteins. The importance of this achievement is apparent if we consider that, performing standard ETD, the averaging of nine micro scans would be required to match such signal-to-noise (S/N) gain. With continuous advancements in the state-of-the-art instrumentation like the one described above, it is easily expected that charge-dependent ion activations will become benchmark methods for MDP.

4.3. Paper I: Proteome digestion specificity analysis for rational design of extended bottom-up and middle-down proteomics experiments

Journal of proteome research

Article

pubs.acs.org/jpr

Proteome Digestion Specificity Analysis for Rational Design of Extended Bottom-up and Middle-down Proteomics Experiments

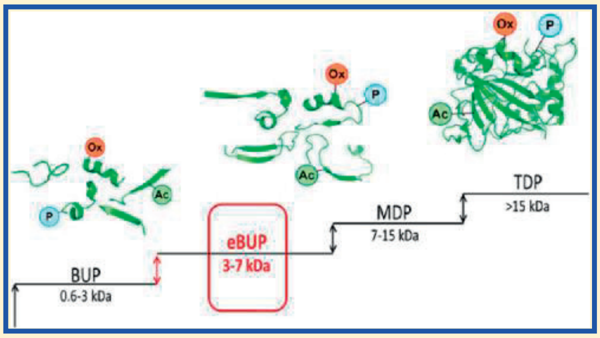
Ünige A. Laskay,[†] Anna A. Lobas,[‡,§] Kristina Srzentić,[†] Mikhail V. Gorshkov,[‡,§] and Yury O. Tsybin*[,†]

[†]Biomolecular Mass Spectrometry Laboratory, Ecole Polytechnique Fédérale de Lausanne, 2 av. Forel, 1015 Lausanne, Switzerland

[‡]Institute for Energy Problems of Chemical Physics, Russian Academy of Sciences, Leninskii Prospect 38, Bldg. 2,119334 Moscow, Russia

[§]Moscow Institute of Physics and Technology (State University), 9 Institutskiy per., 141707 Dolgoprudny, Moscow Region, Russia

Supporting Information



ABSTRACT: Mass spectrometry (MS)-based bottom-up proteomics (BUP) is currently the method of choice for large-scale identification and characterization of proteins present in complex samples, such as cell lysates, body fluids, or tissues. Technically, BUP relies on MS analysis of complex mixtures of small, <3 kDa, peptides resulting from whole proteome digestion. Because of the extremely high sample complexity, further developments of detection methods and sample preparation techniques are necessary. In recent years, a number of alternative approaches such as middle-down proteomics (MDP, addressing up to 15 kDa peptides) and top-down proteomics (TDP, addressing proteins exceeding 15 kDa) have been gaining particular interest. Here we report on the bioinformatics study of both common and less frequently employed digestion procedures for complex protein mixtures specifically targeting the MDP approach. The aim of this study was to maximize the yield of protein structure information from MS data by optimizing peptide size distribution and sequence specificity. We classified peptides into four categories based on molecular weight: 0.6−3 (classical BUP), 3−7 (extended BUP), 7−15 kDa (MDP), and >15 kDa (TDP). Because of instrumentation-related considerations, we first advocate for the extended BUP approach as the potential near-future improvement of BUP. Therefore, we chose to optimize the number of unique peptides in the 3−7 kDa range while maximizing the number of represented proteins. The present study considers human, yeast, and bacterial proteomes. Results of the study can be further used for designing extended BUP or MDP experimental workflows.

KEYWORDS: *mass spectrometry, MS, proteomics, middle-down proteomics, top-down proteomics, bottom-up proteomics*

■ (INTRODUCTION)

Bottom-up proteomics (BUP) is the current approach for high-throughput identification and quantitation of the proteins present in a biological sample.[1] This method entails digestion of proteins into short (6−30 amino acid residue) peptides that can be separated by liquid chromatography (LC) and analyzed by tandem mass spectrometry (MS/MS). The robustness and high throughput of the BUP approach combined with the state-of-the-art LC−MS/MS technology allow identification and quantitation of thousands of proteins with and without post-translational modifications in a single proteomic experiment.[1,2] Modern high-resolution MS instruments, such as Orbitrap Fourier transform mass spectrometer (FTMS), enable the identification of up to 2500 proteins from a human sample in a 90 min LC−MS/MS experiment.[3] Significant efforts on optimization of the ionization and subsequent fragmentation techniques have resulted in increased sensitivity and speed of MS instruments in use.[4] Despite these efforts, up to 85% of MS/MS spectra acquired in a typical LC-MS/MS experiment remain unidentified or result in false identifications, thus reducing the sensitivity of the analysis.[5] One of the shortcomings is related to the properties of enzymatically derived peptides, for example, their size and location of basic or acidic residues. For example, the most widely used trypsin digestion produces a large number of short (0.6 to 1 kDa) peptides. These can be efficiently fragmented and identified, but their sequences bring little specificity at the protein level and they are usually discarded as ambiguous hits. Peptides with more than 30 residues are typically multiply charged when electrospray ionization (ESI) source is used. The presence of long peptides is detrimental in BUP because all MS operation parameters such as the scanning rate, resolution, and fragmentation parameters[6] as well as the LC parameters (column type, dimensions, and flow rate)[7] are optimized for small, <3 kDa peptides.

Received: June 4, 2013
Published: September 27, 2013

© 2013 American Chemical Society

It has also been reported that several protein classes, such as membrane proteins and highly ordered and compacted globular proteins, are underrepresented in the shotgun proteomic experiment due to the inability of the proteases to access the cleavage sites embedded in the structure.[8] As a solution, a two-step digestion with two different proteases, for example, LysC followed by trypsin, has been found to offer more comprehensive sequence coverage of these proteins.[9] Peptide fragmentation efficiency as a function of peptide length and amino acid composition has also been addressed by the use of complementary fragmentation techniques, for example, collision-induced dissociation (CID) and electron-transfer dissociation (ETD).[10]

Among the other major drawbacks of the bottom-up approach in the case of using tryptic digestion is the large number of peptides present in the sample. As a result, most of the precursor ion isolation windows (typically 3 Da) contain coeluting peptides. The fragmentation patterns of these peptides may overlap, and, depending on the abundance of the product ions, a search engine may return wrong sequence candidates. High-resolution MS may not help to solve the issue with coeluting abundant peptides of close masses due to, for example, the coalescence phenomenon associated with ion trap mass analyzers.[11] Furthermore, peptide coisolation reduces the accuracy of isotopic labeling-based protein quantitation strategies, for example, of tandem mass tag (TMT)-based and iTRAQ approaches.[12] The problem associated with low precursor ion fraction in the isolation windows or chimera MS/MS spectra has also been recently recognized.[5a,13] This creates a major limitation for both specificity and sensitivity of the proteomic experiment. Several approaches addressing the above shortcomings are currently under evaluation at both the instrument development level by increasing the sensitivity and the scan rate of high-performance MS as well as by improved sample preparation, for example, equalizing the protein abundances using ProteoMiner technology[14] or improving the accessibility of the digestion site.[15]

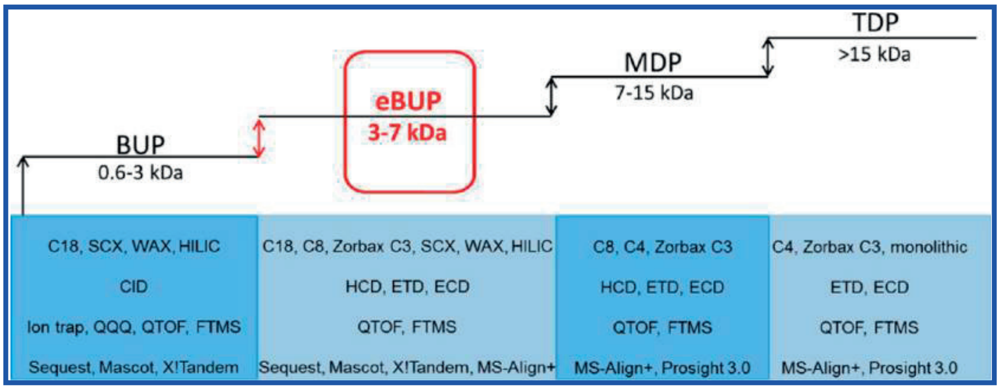
Top-down MS is an alternative strategy for proteome analysis. In this approach, the intact proteins or large protein fragments (primarily in 15−50 kDa range) are analyzed without the need for proteolysis.[16] Protein sequence information is typically obtained by fragmentation of the protein ions in the gas phase using electron capture dissociation (ECD) in Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometers[17] or ETD in hybrid FT-ICR MS, Orbitrap FTMS, or time-of-flight (qTOF) MS instruments.[18] The major advantage of top-down proteomics (TDP) is the access to the entire protein sequence and information about possible post-translational modifications (PTMs) present.[19] Therefore, TDP aims at proteoform-level analysis of heterogeneous protein mixtures or proteomes.[20] The term "proteoform" encompasses protein molecular forms produced from a single gene that are structurally different as a result of genetic variations at the DNA level, alternative splicing at the RNA level, and PTMs at the protein level. However, the analysis is greatly hindered, especially for large proteins, due to the inefficiency of the MS/MS techniques, including ECD and ETD, to provide extensive sequence coverage for proteins >50 kDa in a time-constraint experiment. In addition, the high charge states of the precursor ions, overlapping charge-state distributions of different proteins, and structural diversity of the intact proteins render the isolation of a single protein, and especially proteoform,[20] technically unfeasible. The resulting product ion mass spectra are convoluted and may contain product ions of multiple charge states. The separation of intact proteins using adsorption chromatography is also limited to small-size proteins with a low level of diversity. Because of the above reasons, the implementation of TDP approach has been limited to low-complexity mixtures of proteins of relatively low molecular weights (usually below 30−35 kDa). Therefore, TDP is routinely employed in only a handful of research laboratories. Nevertheless, recent developments in online capillary electrophoresis (CE)−MS/MS technologies, ETD and higher energy collision-induced dissociation (HCD) methods, sensitivity, and resolving power of mass spectrometry (MS), this strategy is gaining a widespread interest in the field of applied proteomic research.[21]

A third emerging technique toward protein identification is middle-down proteomics (MDP).[22] This approach benefits from the undeniable advantages of both bottom-up and top-down strategies and minimizes their above-mentioned shortcomings. In MDP, proteins of interest are also subjected to proteolysis, however, the resulting peptides are significantly larger (up to ~150 residues in a sequence and up to ~15 kDa molecular weight range). The complexity of a mixture is reduced compared with BUP, allowing high-resolution mass analysis on LC separation time scale of a larger fraction of peptides. In addition, the increased peptide length typically results in a larger number of charges per precursor ion, thus increasing the ETD/ECD efficiency.[23] With this approach, increased protein sequence coverage can be obtained due to the longer average size of peptides. In addition, the probability for localization of covalent modifications such as PTMs and single-point mutations arising from splicing variants increases for longer peptides.[24] However, PTMs may lead to increased combinatorial complexity in MDP because a higher number of longer peptides containing multiple PTMs may be required to represent all proteoforms. The information on PTM connectivity provided by MDP on these multiply modified peptides would be lost in BUP.

With recent advances in genome sequencing technologies, the protein databases are rapidly populated with a multitude of proteins that have unknown functions. Currently, one of the most accepted approaches toward assigning protein function and localization uses a similarity search by homology algorithms, such as BLAST.[25] Other approaches include a search for structural motifs[26] and prediction of secondary structures.[27] Bioinformatics studies using a database containing both prokaryotic and eukaryotic species revealed that proteins possess different amino acid compositions depending on their function and vice versa; the composition can hint toward protein function.[28] According to Cedano et al.,[28] besides the expected high frequency of hydrophobic residues in membrane proteins, the commonly targeted digestion site lysine (K) represents 6 to 8% of all amino acids in extracellular, intracellular, and nuclear proteins, whereas in membrane proteins it is less common with the occurrence frequency of ~4.4%. The mean value for arginine (R) is 4.2 to 5% for all protein classes, except those present in the nucleus, which had much higher frequency of 8.7% in R content. Interestingly, histidine (H) was found to be uniformly present in all protein classes at 2.1% frequency.

Importantly, the global amino acid composition of different kingdoms of life also varies significantly. Bogatyreva et al. found that the frequency of K is 5% in bacteria and 6.5% in archaea and eukaryotes, cysteine (C) and serine (S) are more frequent in eukaryotes than in other kingdoms, and tryptophan (W), methionine (M), R, phenylalanine (F), and aspartic acid (D) are uniformly represented in all kingdoms.[29]

Scheme 1. Classification of Mass Spectrometry-Based Proteomic Approaches Based on the Molecular Size of the Analytes[a]



C18, SCX, WAX, HILIC	C18, C8, Zorbax C3, SCX, WAX, HILIC	C8, C4, Zorbax C3	C4, Zorbax C3, monolithic
CID	HCD, ETD, ECD	HCD, ETD, ECD	ETD, ECD
Ion trap, QQQ, QTOF, FTMS	QTOF, FTMS	QTOF, FTMS	QTOF, FTMS
Sequest, Mascot, X!Tandem	Sequest, Mascot, X!Tandem, MS-Align+	MS-Align+, Prosight 3.0	MS-Align+, Prosight 3.0

[a](1) Type of LC column, (2) activation method, (3) mass analyzer, and (4) database search engine.

In summary, the currently employed "one enzyme for all proteomes" approach is not always suitable for blind analysis of complex protein mixtures, regardless of the targeted peptide size range. In this study, we use the peptide size distribution after single and two sequential cleavages as metrics for optimization of the "shotgun" proteomics workflow. This optimization aims at maximizing the number of proteins identified, thus improving both specificity and sensitivity of the proteome analysis and potentially increasing the probability of PTMs localizations.

In this bioinformatics assessment, we considered the possible cleavages at each of the 20 common amino acids for human, yeast, and bacteria proteomes. Figures of merit of commonly used proteolytic and chemical cleavage methods are presented herein. The comprehensive results containing all cleavage sites are presented in the Supporting Information (Figures S2−S3). Venn diagrams (not to scale) depicting the percentage of human proteins identified by individual and combined approaches are contained in the main text of the manuscript. Similar figures for the yeast and bacterial databases and the numerical representation of all Figures can be found in the Supporting Information (Figures S4−S12).

■ EXPERIMENTAL SECTION

In silico digestion of the *Homo sapiens* (human), *Saccharomyces cerevisiae* (yeast), and *Escherichia coli* (bacteria) protein databases (UniProt, release-2012_07) and calculation of peptide masses were performed using the tools of in-house built open-source Python library "Pyteomics".[30] We chose these three species as representative data sets for the mammalian, fungal, and bacterial kingdoms. The nonredundant databases contained 20 103 human, 6566 yeast, and 4243 bacteria proteins. Proteoforms were not included in the calculations. In presenting the data, we use the term "unique peptides" as accepted in the proteomics literature. In brief, the term "unique" refers to peptides representative of a single protein in a given proteome. In this case, peptides that had a shared sequence between multiple proteins were excluded from the statistics. For example, we considered unique peptides for calculating the number of proteins that could be potentially identified with a given pool of peptides.

In the light of currently existing proteomic approaches, the peptide size range identified by MDP and even the terminology for the analysis of long peptides is not consistent. Fenselau and others target the analysis of 3−10 kDa peptides and term the analysis middle-down or middle-out proteomics,[23,31] whereas Kelleher and coworkers, based on their extensive top-down experience, target 5−15 kDa peptides and also term their analysis MDP.[32] Wu and coworkers used the terminology of extended-range proteome analysis (ERPA) for the analysis of long peptides.[33] On the basis of the working regime (scan speed and mass resolution) of instruments employed in reported work as well as our own experience, we suggest differentiating between the analyses of 3−7 and 7−15 kDa peptides and use different terminology for their distinction. These peptide ranges were chosen on the basis of the type and figures of merit of the mass spectrometers that can be employed, chromatography considerations, and available database search algorithms for data analysis. Scheme 1 shows our proposed classification for distinction of the proteomics workflows. BUP has been chosen to be delimited from eBUP based on experimental considerations, such as chromatographic separation conditions and mass spectrometer operation parameters (resolution setting, ion accumulation time, etc.) For example, it has been found that 97% of all tryptic peptides of the yeast proteome identified by BUP have 7−35 residues, with an average peptide length of 10 amino acids.[34] Separation of these short peptides can be performed on C18 reverse-phase columns, and collisional dissociation methods can be used for MS/MS analysis. Because of the short average length of the precursor ions, it is likely that the average charge state is also low. Therefore, in most cases, there is no need for time-consuming high mass resolution analysis of the product ions. In contrast, to avoid strong and irreversible binding of longer peptides on the hydrophobic chromatographic column, a C8 column is more appropriate for separation of those. The broader separation peaks resulting from the weaker hydrophobic interactions are advantageous in the MS/MS working time scale, considering that multiple microscans are to be averaged for enhanced signal-to-noise ratio (S/N) of the highly charged product ions. Several intact protein chromatographic and electrophoretic separation technologies have been implemented in top-down proteomics.[35] In addition to the chromatographic considerations, the MS working parameters also greatly affect the data quality. In principle, the longer the peptide, the more charges it carries; therefore, high-resolution MS and MS/MS scans are required for eBUP, MDP, and TDP. As a result, practical issues such as the ion optics potentials or operating pressure also need to be considered.

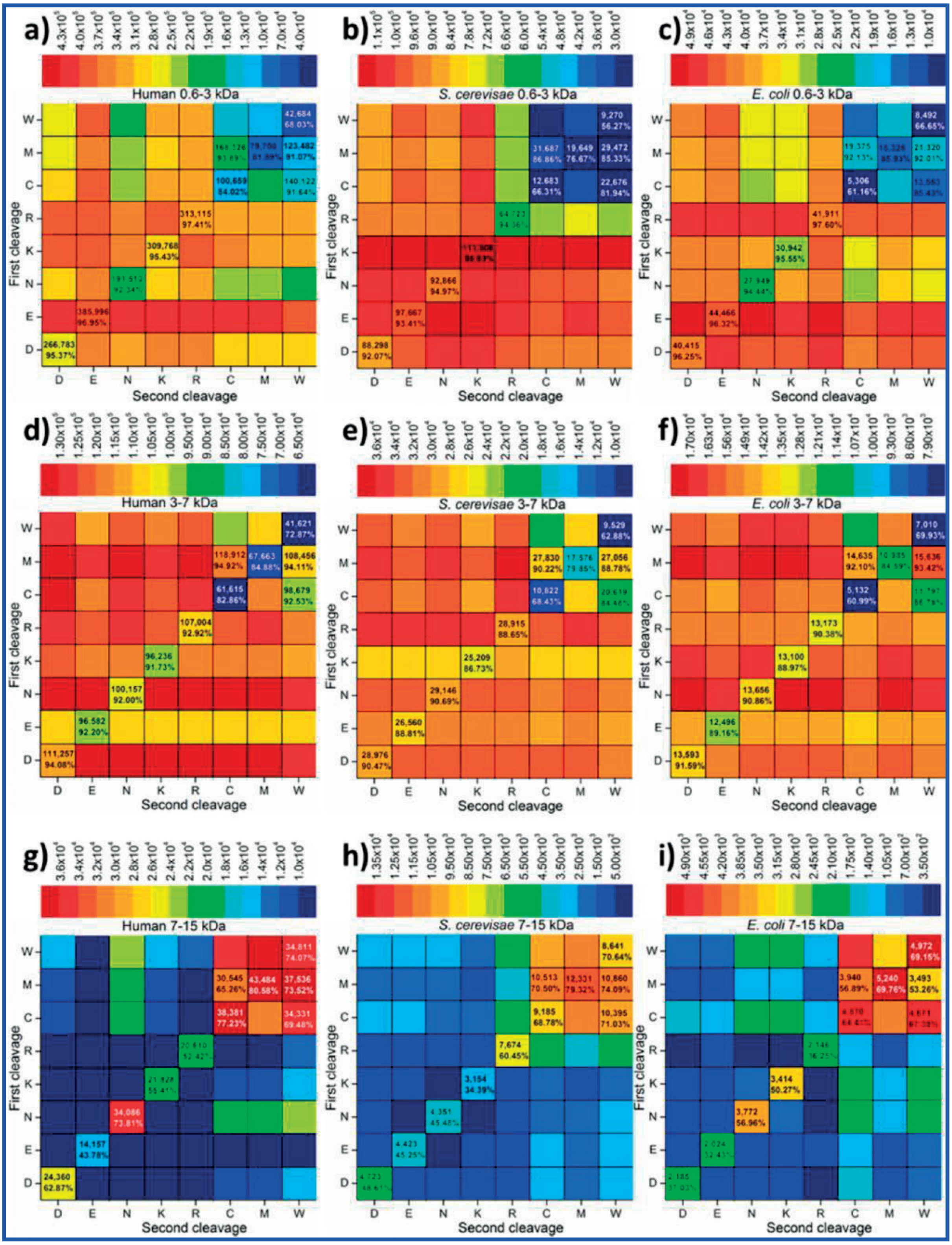


Figure 1. Theoretical distribution of the total number of peptides and percentage of the proteins identifiable with unique peptides in human (left column), yeast (middle column), and bacteria (right column) proteomes after two-step consecutive cleavage with 7 kDa MW cutoff after the first cleavage within (top row) 0.6–3, (middle row) 3–7, and (bottom row) 7–15 kDa ranges.

■ RESULTS AND DISCUSSION

Previous research efforts in MDP have been using proteolytic enzymes with well-established cleavage sites and optimized proteolysis conditions that cleave selectively at a single residue or at infrequently occurring dibasic sites. Herein, we investigated the feasibility of these and other proteases for eBUP and MDP using bioinformatics approach. Figure 1 shows the theoretical number of peptides and their size distributions as well as the percentage of proteome that can be identified in each mass range from the complete human, *S. cerevisae* and *E. coli* proteomes. The amino acid on the vertical axis represents the position of the first cleavage, whereas the position of second cleavage is shown on the horizontal axis. The color scale shows the number of peptides obtained for each cleavage. Cleavage combinations at those amino acid residues that can be targeted with currently known enzymatic or chemical methods are included in Figure 1, whereas Figure S2 (Supporting Information) contains information on all cleavages, regardless of the practical feasibility of the digestion. The amino acids are presented in order of decreasing hydrophobicity. For these calculations, we performed the first in silico digestion, removed the peptides with masses below 7 kDa, performed a second theoretical cleavage on the remaining peptides, and summed the resulting peptides after both digestion steps. The bottom-left to top-right diagonal positions therefore represent the peptide size distributions after single amino acid residue cleavage. For statistical purposes, and to illustrate sample complexity, we have included herein all peptides. Those shared between multiple proteins were counted once. The resulting peptides have been classified into the four (BUP, eBUP, MDP, and TDP) size ranges based on their lengths, Scheme 1 and Figure 1. As expected, in all species studied the number of peptides in the 3−7 kDa region is largest when both of the two consecutive cleavages occur at rare amino acids. Interestingly, the relative frequency of amino acids is different for the three species studied (Figure S1, Supporting Information); therefore, the optimal cleavage site combinations are species-dependent. In the following, we will consider selected amino acids and their pairs as candidates for enzymatic cleavage sites.

Lysine, Lys (K), and Arginine, Arg (R)

Targeting basic digestion sites is the most common approach in BUP, and trypsin is the most commonly used protease in proteomics applications. The Venn diagram in Figure 2 top panel shows the number of human proteins represented by unique peptides in the BUP, eBUP, and MDP size regions following tryptic digestion. Similar figures for the yeast and bacteria are included in the Supporting Information (Figures S3 and S4). Because of the high combined frequency of K and R in all kingdoms, the number of peptides in higher mass regions and the number of identified proteins in eBUP and MDP regimes are greatly reduced, and 15.7% of human proteins can be identified only by the BUP approach. Because Figure 1 reports on the two-step consecutive cleavage with 7 kDa cutoff filter after the first digestion step, the order of amino acids (K followed by R or R followed by K) influences the number of resulting peptides in all mass regimes. In contrast, the tryptic digestion data shown in Figure 2 consider all possible cleavages at K and R residues without an additional 7 kDa molecular weight filtering.

If the side chain of K residues is chemically derivatized prior to digestion, trypsin can be used to target only R residues. One application of this approach used propionic anhydride to block the K residues in histones, leading to increased average peptide size and allowing relative quantitation of modified histones.[36] If this approach was applied to analysis of the entire yeast, bacterial, or human proteome, the number of peptides in BUP would be about three times greater than the eBUP peptides and an order of magnitude larger than the MDP peptides (Figure 1). The Venn diagram in Figure 2, middle panel, illustrates that for the human proteome analysis of eBUP peptides after R cleavage may yield a comparable number of protein identifications and requires a significantly smaller number of peptide MS/MS spectra.

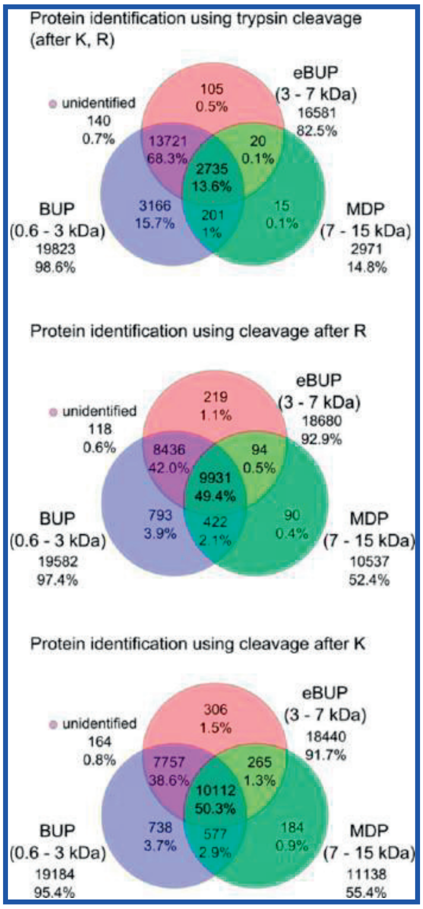


Figure 2. Venn diagrams (not to scale) of the number and percent of human proteins identifiable by unique peptides in BUP, eBUP, and MDP approaches using (top) tryptic proteolysis (both K- and R-specific cleavages), (middle) only K-specific cleavage, and (bottom) only R-specific cleavage.

The frequency of occurrence of K is different than that of R in both yeast and bacterial databases; therefore, targeting this cleavage site showed a different peptide distribution than digestion at R only. Digestion after K residues yielded more BUP and less eBUP and MDP peptides for yeast (Figure 1b,e,h). Considering the 4.4 times larger number of peptides in BUP than in eBUP, the 95.9 and 86.6% proteins identified, respectively, suggest that analysis of peptides in eBUP region could be more advantageous than that in the BUP approach. In contrast with human, for bacteria the number of BUP peptides was greatly reduced, whereas a 1.5-fold increase in MDP peptides was obtained. Bacterial proteome digestion yielded 30 942 BUP peptides, only 2.3 times more than in eBUP; therefore, in the case of this organism, BUP still has the potential to provide satisfactory protein identification. Nonetheless, if the goal of the study requires the detection of higher protein sequence coverage, then eBUP and MDP have the potential to be more feasible than BUP.

There are other proteomics-grade enzymes that cleave around basic amino acids. For example, LysC has been used for characterizing protein glycosylation sites.[24] The benchmarked metallo-endoproteinase, *Grifola frondosa* or LysN, has been successfully characterized and implemented in the digestion of a standard yeast cell lysate;[37] peptides were analyzed in BUP regime using CID. Because with LysN digestion the K is present on the N terminus, this will sequester a proton and yield informative *c*-ion series in ETD, potentially facilitating de novo sequencing of peptides.[38] As seen in the bottom panel in Figure 2, analyses of BUP and eBUP peptides obtained with LysC have the potential to exclusively identify 3.7 and 1.5% of human proteins, respectively.

Dibasic Cleavage Sites

One of the recently proposed cleavage sites for MDP is the targeting of dibasic residues, that is, the positions where two consecutive basic amino acids such as K or R are present.[32,39] Several dibasic-site specific proteases have been characterized to date. Kex2 is a commercially available protein construct, which is described to be specific to KR and RR sites.[40] In silico digestion of the human database targeting these dibasic residues yielded a total of ∼175 000 peptides with average length of 123 amino acids (Supporting Information, Figure S4). OmpT is an outer membrane protease that has less specific dibasic-site preference, and, similar to Sap9, another aspartic protease described in the literature, this protease is supposed to target sites where two consecutive basic amino acids such as K and R are present.[32,39,41] In practice, neither Sap9 nor OmpT is an exclusively dibasic-site-specific protease.

When all dibasic site cleavages were allowed, the number of potential peptides was increased to >240 000, with average length of 75 amino acids (Figure S4 in the Supporting Information). As summarized in Table S1 (Supporting Information), by employing dibasic site cleavage for the human proteome, the number of peptides increases for the high mass range regions, except for the case when all dibasic site cleavages are allowed. In this latter instance, the number of BUP and eBUP peptides and their respective information value (number of proteins represented by unique peptides) were similar.

Contrary to expectations, targeting the dibasic sites does not offer advantages regarding protein identifications by analysis of long peptides, as 34.9% of proteins did not yield >15 kDa peptides. Also, as shown in the Venn diagram in Figure 3, top panel, when targeting all dibasic cleavage sites, all three peptide mass ranges carry unique protein information. Therefore, when this digestion strategy is chosen, it would be beneficial to analyze peptides in at least two different size ranges. Although dibasic-site-specific cleavage can be employed in all proteomics regimes, BUP could be slightly more beneficial than either eBUP or MDP, considering that analysis of 48 600 short peptides can lead to the identification of 76% of all human proteins. eBUP, MDP, and even the >15 kDa region contain a very large number of peptides, requiring higher instrument effort than analysis of short peptides.

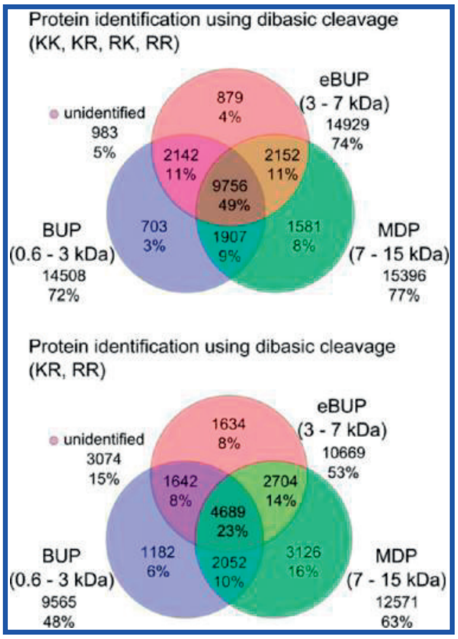


Figure 3. Venn diagrams (not to scale) of the number of unique peptides and percent of human proteins identifiable by BUP, eBUP, and MDP approaches using dibasic site-specific proteolysis with cleavages (top) at all combinations, KK, RR, KR, and RK, and (bottom) only at KR and RR sites.

On the basis of the information in Table S1 in the Supporting Information, top-down analysis of the 28 761 >15 kDa peptides can identify 79.5% of proteins. As expected, the less frequent digestion occurrence also yielded fewer peptides in all other peptide regions. Moreover, combined analysis of all three peptide regions left as much as 15% of all human proteins unidentified, suggesting that Kex2 digestion is most suitable when top-down-type analysis (>15 kDa peptide analysis) is sought, Figure 3, bottom panel.

Aspartic Acid, Asp (D)

One of the first digestion approaches targeted for MDP was microwave-assisted hydrolysis C-terminal to aspartic acid (D) using formic acid or acetic acid and short (order of minutes) microwave irradiation times.[42] An estimated 3.8% of human ribosomal amino acids are D, whereas tryptic sites (K and R) constitute 21.65%; therefore, trypsin cannot be used for comprehensive sequence analysis of these proteins. Swatkoski et al. obtained identification of 58 ribosomal proteins following a 20 min microwave-assisted acid hydrolysis that yielded peptides in the 500−5000 Da mass range.[42c] Although this technique was found to be beneficial for analysis of this select class of proteins, the strong cleavage conditions may lead to loss of phosphate groups and therefore cannot be used for unambiguous PTM localization.[42c] AspN and AspC are metallo-endoproteases that selectively cleave the peptide bond N and Cterminals to D residues, respectively. Therefore a similar peptide size distribution may be obtained with them as with acid hydrolysis without the detrimental loss of phosphate groups.[34] According to the peptide size distribution after digestion of the human proteome targeting D amino acid (Figure 1a,d,g, diagonal line, DD), cleavage yielded 266 783 theoretical peptides in the BUP region, 111 257 in eBUP, and 24 360 in MDP.

Statistical analysis of the yeast and bacterial proteomes offers a similar picture on the utility of targeting D for proteomic analyses. In silico digestion of yeast yielded 88 298 peptides in the BUP region, 28 976 in the eBUP region, and only 4658 in the MDP region (Figure 1). The number of long, >15 kDa peptides was 301. BUP can therefore identify 92.1% of the yeast proteome, whereas eBUP can identify 90.5% assuming that identification of a single unique peptide is sufficient for confident protein identification. Similarly, 40 415 BUP, 13 593 eBUP, 2185 MDP, and 140 >15 kDa peptides were found in bacterial proteome, corresponding to identification yield of 96.3 and 91.6% in BUP and eBUP, respectively. The number of eBUP peptides was three times lower than those in the 0.6−3 kDa region; therefore, the sample complexity can be greatly reduced. Considering that a similar number of proteins may be identified in BUP and eBUP, it is apparent that targeting D amino acid in all kingdoms is beneficial, and, as illustrated in Figure 4, top panel, analysis of peptides only in the 3−7 kDa region can provide similar information as BUP analysis.

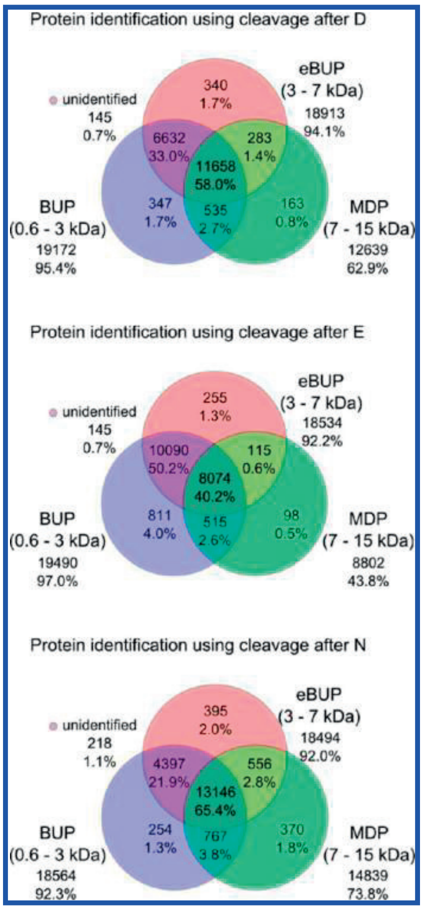


Figure 4. Venn diagrams (not to scale) of the number and percent of human proteins identifiable by unique peptides in BUP, eBUP, and MDP approaches using cleavage at (top) D, (middle) E, and (bottom) select N−X.

Glutamic acid, Glu (E)

Certain protein families, for example, histones, are highly rich in basic residues K and R; therefore, proteases targeting these cleavage sites yield very small peptides with uninformative sequence. GluC (*Staphylococcus aureus* Protease V8) is a serine protease that has been shown to cleave selectively at C terminal to E and has been previously used for mapping histone H3 PTMs.[43]

As shown in Figure 1b,e, in silico digestion of the yeast database targeting cleavage after E yielded 97 667 peptides in the BUP region compared with 26 560 with 3−7 kDa sizes. Presuming that MS/MS analysis and identification of all of these peptides is feasible, BUP can identify 93.4% and eBUP can identify 88.8% of the yeast proteome. Similarly, cleavage of the *E. coli* and human databases at E residues yielded about four times fewer peptides in the eBUP range than in BUP range, and these peptides represent ∼90% of the bacterial and human proteomes, respectively.

GluC digestion to with BUP can therefore be feasible for protein identification in less complex proteomes, such as bacteria, where the separation and analysis of the maximum number of peptides is not technically challenging. For more complex proteomes, eBUP may be more advantageous given that a comparable number of proteins may be identified with a smaller number of required MS/MS spectra (Figure 4, middle panel).

Asparagine, Asn (N)

Currently, no enzymatic or chemical cleavage method exists that is exclusively selective to N, regardless of the neighboring residues. N is a more frequently occurring amino acid in yeast (6.12%) than in human (3.6%) and bacteria (3.95%). Cleavage at all N residues yielded 92 866 peptides in BUP, 29 146 in eBUP, and 4351 peptides in the MDP range (Figure 1). In *E. coli*, 27 949 BUP and 13 656 eBUP peptides were obtained due to the rarer occurrence of this amino acid in this species.

Hydroxylamine has been shown to selectively cleave the N−G peptide bond; however, by using reducing agents and longer interaction times, the cleavage of N−L, N−M, and N−A was also observed.[44] These select cleavages can be beneficial for analysis of long peptides; as shown in Figure 4 bottom panel, eBUP may identify 62.6% of human proteins, whereas MDP analysis could identify 73.2% of the human proteome. The combination of the two techniques may lead to identification of 85.4% of human proteins.

Tryptophan, Trp (W)

2-(2′-Nitrophenylsulfonyl)-3-methyl-3-bromoindolenine (BNPS)-skatole is a brominating reagent that cleaves peptide bonds C terminal to W.[45] Another well-established protocol uses *o*-iodosobenzoic acid; the yield is 80% with few side reactions. W is one of the rarest amino acids in all three species studied; however, the number of peptides in the BUP and eBUP regions were comparable to those obtained in MDP ranges for all three species (Figure 1). Yeast yielded 9270 BUP and 9529 eBUP peptides (Figure 1b,e), whereas MDP and TDP ranges contained 8641 and 6535, respectively (Figure 1h,k). As expected, the number of proteins identified by analysis of unique small and midrange peptides was only 56.3 and 62.9%, respectively.

Unexpectedly, targeting W in the *E. coli* and human proteomes resulted in a greater number of BUP peptides than in eBUP. This could be indicating the preferential positioning of W residues toward protein termini; cleavage at this residue resulting in a short (6−30 amino acids) peptide and the complementary remainder protein fragment. In *E. coli*, digestion at W residues can enable identification of 66.6% and 70.0% of the proteins by BUP and eBUP, respectively (as shown in the Venn diagram in Figure S10, Supporting Information). A similar trend was observed for the human database. Contrary to the expected benefit of MDP, combined BUP and eBUP analysis provided unique peptides suitable for the identification of 85% of the human proteins (Figure 5 top panel). This is a clear illustration of how prior knowledge of the amino acid distribution of the proteome can aid in decision on the working regime.

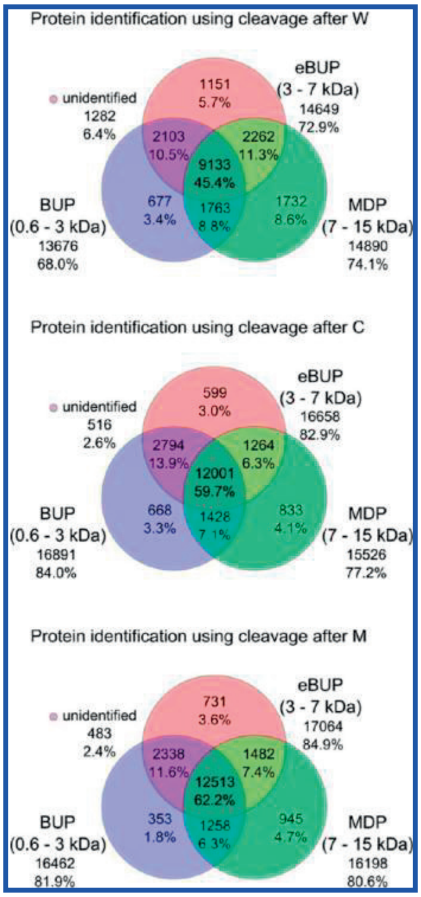


Figure 5. Venn diagrams (not to scale) of the number and % of human proteins identifiable by unique peptides in BUP, eBUP, and MDP approaches using cleavage at rare amino acids: (top) Trp (W), (middle) Cys (C), and (bottom) Met (M).

Cysteine, Cys (C)

2-Nitro-5-thiocyanobenzoic acid (NTCB) is a reagent that yields peptides with N-terminal C. The resulting peptides will be modified by an iminothiazolidine-carboxyl group.[46] C is a rarely occurring amino acid; therefore, a large number of long peptides (7−15 kDa) would be expected as a result of C-site specific digestion. However, as seen in Figure 1b,e,h, in the yeast proteome very similar numbers of BUP, eBUP, and MDP peptides are to be expected, and analysis in either regime would identify ~70% of the proteome. Surprisingly, digestion of the bacterial proteome yielded four to five times more peptides in the BUP region than in eBUP or MDP, and a similar trend was observed for the human proteome. It would therefore be unusually beneficial to perform the combined analysis of the BUP and MDP peptides to achieve identification of 94.6% of human proteins, as shown in Figure 5, middle panel.

Methionine, Met (M)

CNBr is a toxic reagent that has been used for decades for attaining cleavages N-terminal to M.[47] The three kingdoms studied again presented different trends of peptide size distribution when cleaved at M. Yeast showed comparable number of peptides in the three mass ranges, allowing identification of ~80% of the proteome in either regime. A dramatic difference in the number of BUP peptides was obtained for bacteria, yielding 15 325 BUP peptides that represent as much as 95.5% of the proteome. Analysis of either MDP or eBUP peptides would be less beneficial in this case. Contrary to the expected large number of long peptides, the human proteome showed a uniform distribution of short, medium, and long peptides. Combined analysis of any two peptide regimes would therefore allow identification of ~95% of the human proteome, as indicated in Figure 5, bottom panel.

Pairwise Cleavages

We investigated in detail those pairwise cleavages that would maximize the number of proteins identified in the 3−7 kDa eBUP range, with the least number of peptides analyzed. In brief, the peptide size distribution was investigated by in silico cleavage at the first amino acid, followed by a cleavage of peptides >7 kDa at the second amino acid. Therefore, the most logical choice was the coupling of two rare amino acid cleavage sites, while keeping in mind the practical feasibility of the experiment. For example, methionine oxidation can be induced by various experimental steps; therefore, it would be beneficial to target this amino acid as first cleavage site.

C followed by W. The peptide size distribution after two sequential cleavages was investigated by in silico cleavage at C, followed by a second cleavage of peptides >7 kDa at W. As a result, the number of short and medium-length peptides increased when compared with the number of peptides obtainable by single cleavages. For the yeast and bacterial databases the number of peptides in the 7−15 kDa (MDP) region also slightly increased when compared with the individual C or W cleavages (Figure 1). In contrast, for the human database, the number of MDP peptides remained very similar, or slightly decreased when compared with individual W and C cleavages, respectively. However, a second digestion at W of the 38 381 MDP peptides obtained after C cleavage resulted in 39 463 more BUP peptides, and 37 064 more eBUP peptides than C cleavage alone. Because 34 331 MDP peptides still remained, it is apparent that 4050 very long (>15 kDa) peptides exist in the human database that are highly rich in W. One source of such peptides is NADH dehydrogenase (UniProt accession O95167), a 9.3 kDa protein with no C residues and two W amino acids in close proximity of both termini. This sequential cleavage increases the number of proteins identified for all databases in BUP and eBUP regimes, and the most significant improvement was estimated for the human database. As shown in Figure 6, top panel, the overall number of proteins that can be identified in the human database using pairwise C and W cleavage increased in all mass regimes, and a combined analysis of BUP and eBUP peptides left only 2.6% of proteins unidentified versus 15% for individual W cleavage and 6.7% for C, respectively.

M followed by W

Targeting M followed by W yielded significant improvement in the number of proteins represented in BUP and eBUP regions for all databases when compared with individual W and M cleavages (Figure 1). For the *E. coli* and human databases, 92 and 91.1% of proteins were represented in the BUP region, and 93.4 and 94.1% of proteins were represented in the eBUP region, respectively. In contrast, this cleavage combination was not as beneficial for the yeast database, with a modest 85.3% of proteins represented in BUP and 88.8% in the eBUP region. As shown in Figure 6, middle panel, 99.3% of human proteins can be identified by a combined analysis of the three mass regions, with the eBUP range offering the most identifiable proteins individually (94.1%).

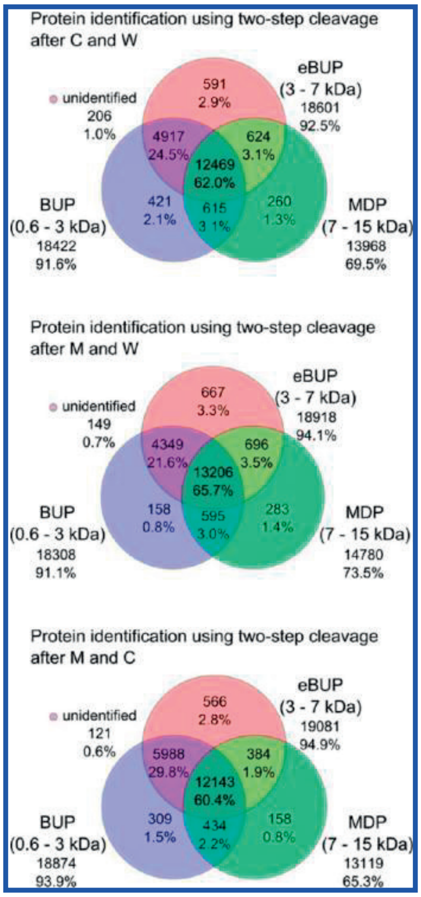

Protein identification using two-step cleavage after C and W
unidentified 206 1.0%
eBUP (3 - 7 kDa) 18601 92.5%
591 2.9%
4917 24.5%
624 3.1%
12469 62.0%
421 2.1%
615 3.1%
260 1.3%
BUP (0.6 - 3 kDa) 18422 91.6%
MDP (7 - 15 kDa) 13968 69.5%
Protein identification using two-step cleavage after M and W
unidentified 149 0.7%
eBUP (3 - 7 kDa) 18918 94.1%
667 3.3%
4349 21.6%
696 3.5%
13206 65.7%
158 0.8%
595 3.0%
283 1.4%
BUP (0.6 - 3 kDa) 18308 91.1%
MDP (7 - 15 kDa) 14780 73.5%
Protein identification using two-step cleavage after M and C
unidentified 121 0.6%
eBUP (3 - 7 kDa) 19081 94.9%
566 2.8%
5988 29.8%
384 1.9%
12143 60.4%
309 1.5%
434 2.2%
158 0.8%
BUP (0.6 - 3 kDa) 18874 93.9%
MDP (7 - 15 kDa) 13119 65.3%

Figure 6. Venn diagrams (not to scale) of the number and percent of human proteins identifiable by unique peptides in BUP, eBUP, and MDP approaches using two-step cleavages at amino acid pairs: (top) Cys (C) followed by Trp (W), (middle) Met (M) followed by Trp (W), and (bottom) Met (M) followed by Cys (C).

M followed by C

The last pairwise cleavage exemplified herein yielded peptides representing a higher number of yeast and human proteins in the short and midrange peptide regions when compared with the two previous examples. This improvement comes with the added benefit that in the case of yeast the number of peptides required to identify more proteins in the eBUP region is decreased. For the human proteome, this cleavage pair yields slightly more proteins identified than the C−W and M−W cleavages (Figure 6, bottom panel), leaving only 0.6% of proteins not represented in the three mass regions. This slight improvement, however, comes at the price of a significant increase in the number of peptides generated.

Comparison of Proteomics Approaches

Figure 7 shows a general overview of the results previously presented and summarizes the potential of the three techniques (MDP, eBUP, and BUP) for the in-depth characterization of the human proteome. The total number of peptides produced by each pairwise cleavage presented in Figure 7 was plotted against the percentage of proteins identifiable by unique peptides in the three mass ranges. Figure 7a shows the proteome coverage range between 10 and 100%, whereas the inset in Figure 7a,b shows expanded regions of the proteome coverage axis, between 90.5 and 94% and 95−98.5%, respectively. It becomes apparent that identification of tryptic peptides in the classical BUP mass range (Figure 7a) offers the highest proteome coverage (98.61%). This, however, can be achieved with the identification of ∼574 000 peptides. In practice, even the most modern, state-of-the-art MS instruments are not capable of achieving this performance. Michalski et al. was able to detect in excess of 100 000 peptides in MS mode in a 90 min gradient; however, only ∼16 000 of these yielded identifiable MS/MS spectra with the search parameters utilized.[5a] Increasing the gradient length does not scale proportionally with the number of peptides characterized; in a 4 h gradient Wisniewski et al. identified ∼40 000 peptides.[15] Utilizing an LC-ion mobility separation approach, as it is available on the Waters Synapt G2 instruments, allows for a three-fold increase in the protein identification rate when compared with a data-dependent LC-MS/MS approach.[48] This improved duty cycle is possible due to the high speed of the time-of-flight mass analyzer coupled to the data-independent peptide fragmentation. However, despite recent improvements in nano-LC separation technologies and MS instrumentation, the speed of peptide fragmentation and analysis does not allow the timely identification of all peptides, and implicitly of proteins, from a highly complex proteome. It is also apparent from Figure 7b that a multitude of cleavage sites can be targeted for BUP analysis. For example, a combination of consecutive cleavages at E and R (or vice versa) would decrease the number of peptides by ∼100 000 in BUP with a loss of only 0.5% of identifiable proteins.

In contrast, none of the MDP approaches described herein allows for identification of more than 80% of the human proteome (Figure 7a, blue dots). The most optimal cleavage site was found to be Cys, offering the possibility for identification of 77.2% of the human proteome, albeit with only ∼38 000 peptides in the 7−15 kDa range. Digestion at all possible dibasic sites or at Trp would be slightly less optimal, permitting identification of 73.5 or 76.6% of the proteome, respectively, with ∼37 000 peptides. The LC separation efficiency of these peptides may be reduced because of their length. In addition, current MS instruments require significantly longer acquisition times for the MS/MS spectra of these larger species. This greatly reduces the duty cycle and number of peptides that can be identified in a single LC−MS/MS experiment.

The inset in Figure 7a,b shows the performance of several eBUP approaches (red dots). It becomes apparent that several digestion protocols are suitable for the identification of >90% of the proteome with a single-step, well-established protocol. Digestion with LysC or GluC and analysis of the 3−7 kDa peptides are therefore the simplest means for reducing sample complexity while achieving high proteome coverage. It is not surprising that combination of a cleavage at a rare amino acid site (M or W) followed by a second digestion of the long peptides at a more frequent amino acid (D, K, R) or vice versa, optimizes the number of peptides in the 3−7 kDa size range. With such two-step cleavages, up to 97% of the human proteome could be identified, with as few as ∼125 000 peptides. The separation of these midsize peptides is less challenging, and the MS instrumentation parameters (fragmentation, ion accumulation, transfer, and detection) are closer to those required by a classical BUP experiment, as previously discussed.

■ CONCLUSIONS

On the basis of our findings, we propose that the analysis of peptides in the 3−7 kDa range is more optimal than targeting bottom-up (0.6−3 kDa) or middle-down (7−15 kDa) peptides

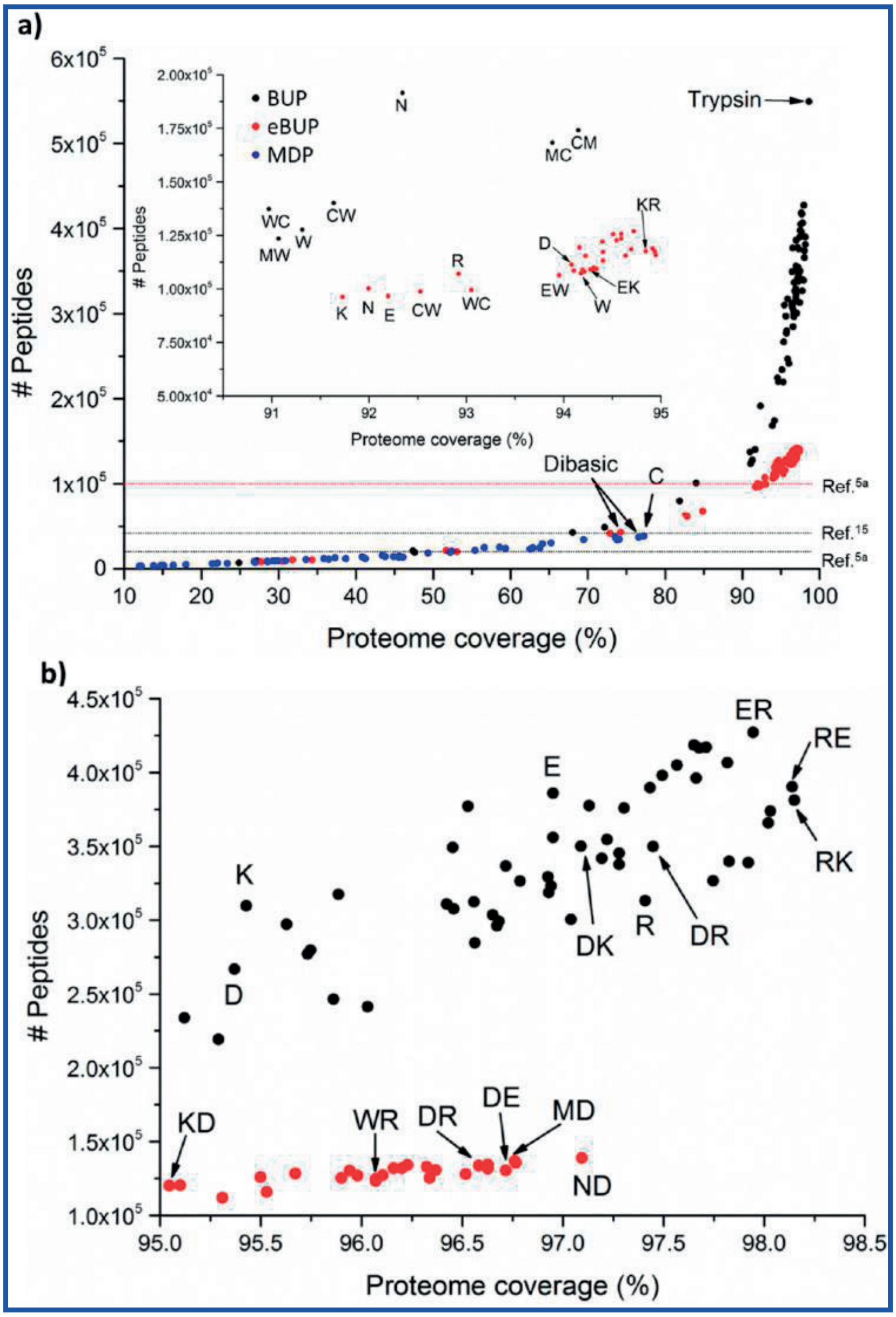


Figure 7. Comparison of the performance of the BUP, eBUP, and MDP approaches for identification of the human proteome shown as a dependence on the number of peptides on the percent of proteins represented by unique peptides. The total number of peptides shown on the *y* axis corresponds to Figure 1 and describes peptides produced in a single digestion or in a two-step process with a cleavage at the first amino acid, followed by a cleavage of peptides >7 kDa at the second amino acid. Also shown are the data corresponding to the cleavages by trypsin and dibasic enzymes. (See Table S1 in the Supporting Information.) The dotted black lines represent examples of detectable and identifiable peptides in a 90 min and 4 h LC−MS/MS experiment, respectively. The dotted red line shows the total number of peptides detected from a 90 min LC gradient reported in the corresponding reference.

due to several reasons. First, the number of theoretical peptides is, in some cases, an order of magnitude lower than that in the bottom-up regime. This may lead to the detection of a larger fraction of peptides on a chromatographic time scale. In addition, the identification of longer peptides inevitably offers better sequence coverage than short peptides, which can improve protein identification and may enhance PTMs localization efficiency. It was also demonstrated that it is desirable to maximize the number of proteins represented while minimizing the number of unique 3−7 kDa peptides. Theoretically, two-step cleavages targeting two distinct rare amino acids (combination of W, C, or M) are more appropriate for this purpose than targeting a single cleavage site. In contrast, the analysis of long (7−15 kDa) peptides alone did not seem to provide the expected advantages foreseen by promoters of the middle-down approach. None of the individual cleavage sites or a two-step cleavage seem to be yielding a complete protein identification for whole protein databases. It is also apparent that because of the variability in the amino acid frequencies of different species a different protease, chemical, or combination of these might be required for optimal protein identification. Therefore, to increase the number of proteins represented by these very long peptides, the digestion approach should be carefully tailored toward the subset of targeted databases or protein families studied. In the Supporting Information, a comprehensive analysis of all pairwise cleavages has been included for the yeast, bacterial, and human databases. This information can be further mined for the judicious selection of the cleavage agents for assessment of these species.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Number of peptides in the human protein database for the dibasic site cleavages in the different peptide mass ranges. Relative frequency of amino acids for the three species studied. Theoretical distribution of LC unique peptides and proteins that remain unidentified in the human, yeast, and bacteria proteome. Peptide length distribution of unique peptides from dibasic-site specific digestion of the human proteome. Venn diagrams of the number and % of yeast proteins and bacterial proteins identified by unique peptides in BUP, eBUP, and MDP approaches. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +41 21 693 97 51. Fax: +41 21 693 98 95. E-mail: yury.tsybin@epfl.ch.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The work was supported by the Swiss National Science Foundation (Projects 200021-125147 and 128357) and the European Research Council (ERC Starting grant 280271 to Y.O.T.). A.A.L. and M.V.G. thank EU FP7 Program support (Project Prot-HiSPRA, grant agreement 282506).

■ REFERENCES

(1) Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M. C.; Yates, J. R., 3rd. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chem. Rev.* **2013**, *113* (4), 2343−94.

(2) Mann, M.; Kulak, N. A.; Nagaraj, N.; Cox, J. The coming age of complete, accurate, and ubiquitous proteomes. *Mol. Cell* **2013**, *49* (4), 583−90.

(3) Michalski, A.; Damoc, E.; Hauschild, J. P.; Lange, O.; Wieghaus, A.; Makarov, A.; Nagaraj, N.; Cox, J.; Mann, M.; Horning, S. Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics* **2011**, *10* (9), M111 011015.

(4) Cox, J.; Mann, M. Quantitative, high-resolution proteomics for data-driven systems biology. *Annu. Rev. Biochem.* **2011**, *80*, 273−99.

(5) (a) Michalski, A.; Cox, J.; Mann, M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J. Proteome Res.* **2011**, *10* (4), 1785−93. (b) Frank, A. M.; Monroe, M. E.; Shah, A. R.; Carver, J. J.; Bandeira, N.; Moore, R. J.; Anderson, G. A.; Smith, R. D.; Pevzner, P. A. Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nat. Methods* **2011**, *8* (7), 587−91.

(6) Jedrychowski, M. P.; Huttlin, E. L.; Haas, W.; Sowa, M. E.; Rad, R.; Gygi, S. P. Evaluation of HCD- and CID-type fragmentation within their respective detection platforms for murine phosphoproteomics. *Mol. Cell. Proteomics* **2011**, *10* (12), M111 009910.

(7) (a) Kocher, T.; Swart, R.; Mechtler, K. Ultra-high-pressure RPLC hyphenated to an LTQ-Orbitrap Velos reveals a linear relation between peak capacity and number of identified peptides. *Anal. Chem.* **2011**, *83* (7), 2699−704. (b) McQueen, P.; Krokhin, O. Optimal selection of 2D reversed-phase-reversed-phase HPLC separation techniques in bottom-up proteomics. *Expert Rev. Proteomics* **2012**, *9* (2), 125−8.

(8) Vuckovic, D.; Dagley, L. F.; Purcell, A. W.; Emili, A. Membrane proteomics by high performance liquid chromatography-tandem mass spectrometry: Analytical approaches and challenges. *Proteomics* **2013**, *13* (3−4), 404−23.

(9) Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **2001**, *19* (3), 242−7.

(10) (a) Altelaar, A. F.; Heck, A. J. Trends in ultrasensitive proteomics. *Curr. Opin. Chem. Biol.* **2012**, *16* (1−2), 206−13. (b) Swaney, D. L.; McAlister, G. C.; Coon, J. J. Decision tree-driven tandem mass spectrometry for shotgun proteomics. *Nat. Methods* **2008**, *5* (11), 959−64.

(11) Gorshkov, M. V.; Fornelli, L.; Tsybin, Y. O. Observation of ion coalescence in Orbitrap Fourier transform mass spectrometry. *Rapid Commun. Mass Spectrom.* **2012**, *26* (15), 1711−1717.

(12) Wuhr, M.; Haas, W.; McAlister, G. C.; Peshkin, L.; Rad, R.; Kirschner, M. W.; Gygi, S. P. Accurate multiplexed proteomics at the MS2 level using the complement reporter ion cluster. *Anal. Chem.* **2012**, *84* (21), 9214−21.

(13) Houel, S.; Abernathy, R.; Renganathan, K.; Meyer-Arendt, K.; Ahn, N. G.; Old, W. M. Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *J. Proteome Res.* **2010**, *9* (8), 4152−60.

(14) Fonslow, B. R.; Carvalho, P. C.; Academia, K.; Freeby, S.; Xu, T.; Nakorchevsky, A.; Paulus, A.; Yates, J. R., 3rd. Improvements in proteomic metrics of low abundance proteins through proteome equalization using ProteoMiner prior to MudPIT. *J. Proteome Res.* **2011**, *10* (8), 3690−700.

(15) Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **2009**, *6* (5), 359−62.

(16) Tran, J. C.; Zamdborg, L.; Ahlf, D. R.; Lee, J. E.; Catherman, A. D.; Durbin, K. R.; Tipton, J. D.; Vellaichamy, A.; Kellie, J. F.; Li, M.; Wu, C.; Sweet, S. M.; Early, B. P.; Siuti, N.; LeDuc, R. D.; Compton, P. D.; Thomas, P. M.; Kelleher, N. L. Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **2011**, *480* (7376), 254−8.

(17) (a) Pan, J.; Borchers, C. H. Top-down structural analysis of posttranslationally modified proteins by Fourier transform ion cyclotron resonance-MS with hydrogen/deuterium exchange and electron capture dissociation. *Proteomics* **2013**, *13* (6), 974−81. (b) Mao, Y.; Valeja, S. G.; Rouse, J. C.; Hendrickson, C. L.; Marshall, A. G. Top-down structural analysis of an intact monoclonal antibody by electron capture dissociation-fourier transform ion cyclotron resonance-mass spectrometry. *Anal. Chem.* **2013**, *85* (9), 4239−46.

(18) (a) Fornelli, L.; Damoc, E.; Thomas, P. M.; Kelleher, N. L.; Aizikov, K.; Denisov, E.; Makarov, A.; Tsybin, Y. O. Analysis of intact monoclonal antibody IgG1 by electron transfer dissociation Orbitrap FTMS. *Mol. Cell. Proteomics* **2012**, *11* (12), 1758−67. (b) Tsybin, Y. O.; Fornelli, L.; Stoermer, C.; Luebeck, M.; Parra, J.; Nallet, S.; Wurm, F. M.; Hartmer, R. Structural analysis of intact monoclonal antibodies by electron transfer dissociation mass spectrometry. *Anal. Chem.* **2011**, *83* (23), 8919−27.

(19) Chamot-Rooke, J.; Mikaty, G.; Malosse, C.; Soyer, M.; Dumont, A.; Gault, J.; Imhaus, A. F.; Martin, P.; Trellet, M.; Clary, G.; Chafey, P.; Camoin, L.; Nilges, M.; Nassif, X.; Dumenil, G. Posttranslational

modification of pili upon cell contact triggers N. meningitidis dissemination. *Science* **2011**, *331* (6018), 778−82.

(20) Smith, L. M.; Kelleher, N. L. Consortium for Top Down Proteomics, Proteoform: a single term describing protein complexity. *Nat. Methods* **2013**, *10* (3), 186−7.

(21) (a) Ahlf, D. R.; Compton, P. D.; Tran, J. C.; Early, B. P.; Thomas, P. M.; Kelleher, N. L. Evaluation of the compact high-field orbitrap for top-down proteomics of human cells. *J. Proteome Res.* **2012**, *11* (8), 4308−14. (b) Tian, Z.; Tolic, N.; Zhao, R.; Moore, R. J.; Hengel, S. M.; Robinson, E. W.; Stenoien, D. L.; Wu, S.; Smith, R. D.; Pasa-Tolic, L. Enhanced top-down characterization of histone post-translational modifications. *Genome Biol.* **2012**, *13* (10), R86.

(22) Garcia, B. A.; Siuti, N.; Thomas, C. E.; Mizzen, C. A.; Kelleher, N. L. Characterization of neurohistone variants and post-translational modifications by electron capture dissociation mass spectrometry. *Int. J. Mass Spectrom.* **2007**, *259* (1−3), 184−196.

(23) Cannon, J.; Lohnes, K.; Wynne, C.; Wang, Y.; Edwards, N.; Fenselau, C. High-throughput middle-down analysis using an orbitrap. *J. Proteome Res.* **2010**, *9* (8), 3886−90.

(24) Wu, S. L.; Huhmer, A. F.; Hao, Z.; Karger, B. L. On-line LC-MS approach combining collision-induced dissociation (CID), electron-transfer dissociation (ETD), and CID of an isolated charge-reduced species for the trace-level characterization of proteins with post-translational modifications. *J. Proteome Res.* **2007**, *6* (11), 4230−44.

(25) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215* (3), 403−10.

(26) (a) Henikoff, S.; Henikoff, J. G. Automated assembly of protein blocks for database searching. *Nucleic Acids Res.* **1991**, *19* (23), 6565−72. (b) Bairoch, A. PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res.* **1992**, *20* (Suppl), 2013−8.

(27) Rost, B.; Sander, C. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **1993**, *232* (2), 584−99.

(28) Cedano, J.; Aloy, P.; Perez-Pons, J. A.; Querol, E. Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol.* **1997**, *266* (3), 594−600.

(29) Bogatyreva, N. S.; Finkelstein, A. V.; Galzitskaya, O. V. Trend of amino acid composition of proteins of different taxa. *J. Bioinf. Comput. Biol.* **2006**, *4* (2), 597−608.

(30) Goloborodko, A. A.; Levitsky, L. I.; Ivanov, M. V.; Gorshkov, M. V. Pyteomics–a Python framework for exploratory data analysis and rapid software prototyping in proteomics. *J. Am. Soc. Mass Spectrom.* **2013**, *24* (2), 301−4.

(31) Cannon, J.; Nakasone, M.; Fushman, D.; Fenselau, C. Proteomic identification and analysis of K63-linked ubiquitin conjugates. *Anal. Chem.* **2012**, *84* (22), 10121−8.

(32) Wu, C.; Tran, J. C.; Zamdborg, L.; Durbin, K. R.; Li, M.; Ahlf, D. R.; Early, B. P.; Thomas, P. M.; Sweedler, J. V.; Kelleher, N. L. A protease for 'middle-down' proteomics. *Nat. Methods* **2012**, *9* (8), 822−4.

(33) (a) Wu, S. L.; Kim, J.; Hancock, W. S.; Karger, B. Extended Range Proteomic Analysis (ERPA): a new and sensitive LC-MS platform for high sequence coverage of complex proteins with extensive post-translational modifications-comprehensive analysis of beta-casein and epidermal growth factor receptor (EGFR). *J. Proteome Res.* **2005**, *4* (4), 1155−70. (b) Wu, S. L.; Kim, J.; Bandle, R. W.; Liotta, L.; Petricoin, E.; Karger, B. L. Dynamic profiling of the post-translational modifications and interaction partners of epidermal growth factor receptor signaling after stimulation by epidermal growth factor using Extended Range Proteomic Analysis (ERPA). *Mol. Cell. Proteomics* **2006**, *5* (9), 1610−27.

(34) Swaney, D. L.; Wenger, C. D.; Coon, J. J. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J. Proteome Res.* **2010**, *9* (3), 1323−9.

(35) Capriotti, A. L.; Cavaliere, C.; Foglia, P.; Samperi, R.; Lagana, A. Intact protein separation by chromatographic and/or electrophoretic techniques for top-down proteomics. *J. Chromatogr., A* **2011**, *1218* (49), 8760−76.

(36) Garcia, B. A.; Mollah, S.; Ueberheide, B. M.; Busby, S. A.; Muratore, T. L.; Shabanowitz, J.; Hunt, D. F. Chemical derivatization of histones for facilitated analysis by mass spectrometry. *Nat. Protoc.* **2007**, *2* (4), 933−8.

(37) Hohmann, L.; Sherwood, C.; Eastham, A.; Peterson, A.; Eng, J. K.; Eddes, J. S.; Shteynberg, D.; Martin, D. B. Proteomic analyses using Grifola frondosa metalloendoprotease Lys-N. *J. Proteome Res.* **2009**, *8* (3), 1415−22.

(38) Taouatas, N.; Drugan, M. M.; Heck, A. J.; Mohammed, S. Straightforward ladder sequencing of peptides using a Lys-N metalloendopeptidase. *Nat. Methods* **2008**, *5* (5), 405−7.

(39) Laskay, Ü. A.; Srzentić, K.; Fornelli, L.; Upir, O.; Kozhinov, A. N.; Monod, M.; Tsybin, Y. O. Practical considerations for improving the productivity of mass spectrometry-based proteomics. *Chimia* **2013**, *67* (4), 244−249.

(40) Mizuno, K.; Nakamura, T.; Ohshima, T.; Tanaka, S.; Matsuo, H. Characterization of KEX2-encoded endopeptidase from yeast Saccharomyces cerevisiae. *Biochem. Biophys. Res. Commun.* **1989**, *159* (1), 305−11.

(41) Albrecht, A.; Felk, A.; Pichova, I.; Naglik, J. R.; Schaller, M.; de Groot, P.; Maccallum, D.; Odds, F. C.; Schafer, W.; Klis, F.; Monod, M.; Hube, B. Glycosylphosphatidylinositol-anchored proteases of Candida albicans target proteins necessary for both cellular processes and host-pathogen interactions. *J. Biol. Chem.* **2006**, *281* (2), 688−94.

(42) (a) Hua, L.; Low, T. Y.; Sze, S. K. Microwave-assisted specific chemical digestion for rapid protein identification. *Proteomics* **2006**, *6* (2), 586−91. (b) Hauser, N. J.; Han, H.; McLuckey, S. A.; Basile, F. Electron transfer dissociation of peptides generated by microwave D-cleavage digestion of proteins. *J. Proteome Res.* **2008**, *7* (5), 1867−72. (c) Swatkoski, S.; Gutierrez, P.; Wynne, C.; Petrov, A.; Dinman, J. D.; Edwards, N.; Fenselau, C. Evaluation of microwave-accelerated residue-specific acid cleavage for proteomic applications. *J. Proteome Res.* **2008**, *7* (2), 579−86.

(43) Taverna, S. D.; Ueberheide, B. M.; Liu, Y.; Tackett, A. J.; Diaz, R. L.; Shabanowitz, J.; Chait, B. T.; Hunt, D. F.; Allis, C. D. Long-distance combinatorial linkage between methylation and acetylation on histone H3 N termini. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (7), 2086−91.

(44) Bornstein, P.; Balian, G. Cleavage at Asn-Gly bonds with hydroxylamine. *Methods Enzymol.* **1977**, *47*, 132−45.

(45) Sass, E.; Blachinsky, E.; Karniely, S.; Pines, O. Mitochondrial and cytosolic isoforms of yeast fumarase are derivatives of a single translation product and have identical amino termini. *J. Biol. Chem.* **2001**, *276* (49), 46111−7.

(46) Stark, G. R. Cleavage at cysteine after cyanylation. *Methods Enzymol.* **1977**, *47*, 129−32.

(47) Andreev, Y. A.; Kozlov, S. A.; Vassilevski, A. A.; Grishin, E. V. Cyanogen bromide cleavage of proteins in salt and buffer solutions. *Anal. Biochem.* **2010**, *407* (1), 144−6.

(48) Rodriguez-Suarez, E.; Hughes, C.; Gethings, L.; Giles, K.; Wildgoose, J.; Stapels, M.; Fadgen, K. E.; Geromanos, S. J.; Vissers, J. P. C.; Elortza, F.; Langridge, J. I. An Ion Mobility Assisted Data Independent LC-MS Strategy for the Analysis of Complex Biological Samples. *Curr. Anal. Chem* **2013**, *9* (2), 199−211.

Supporting Information.

Proteome digestion specificity analysis for rational design of extended bottom-up and middle-down proteomics experiments

Ünige A. Laskay, Anna A. Lobas, Kristina Srzentić, Mikhail V. Gorshkov, and Yury O. Tsybin*

Table S1. Number of peptides in the human protein database for the dibasic site cleavages in the different peptide mass ranges. Also included are the percentage values of proteins that can be identified with at least one peptide by each cleavage, if peptides from a single mass range are analyzed.

	0.6-3 kDa		**3-7 kDa**		**7-15 kDa**		**>15 kDa**	
	#peptides	**Unrepr. proteins**	**#peptides**	**Unrepr. proteins**	**#peptides**	**Unrepr. proteins**	**#peptides**	**Unrepr. proteins**
KK	9,389	72.82%	10,287	68.19%	14,298	55.13%	28,232	14.39%
KR	6,212	78.04%	7,936	72.28%	11,848	58.78%	28,105	13.29%
RR	9,471	70.28%	1,0372	65.67%	13,530	53.93%	27,852	15.04%
RK	7,112	75.13%	8,776	69.55%	13,541	54.47%	28,899	14.07%
KK&RK	20,996	52.64%	21,364	48.29%	24,674	36.21%	28,244	20.88%
RR&KR	19,174	52.42%	20,062	46.93%	22,924	37.47%	28,761	20.48%
KK&RK & RR&KR	48,600	27.83%	43,117	25.74%	36,985	23.41%	22,825	34.85%

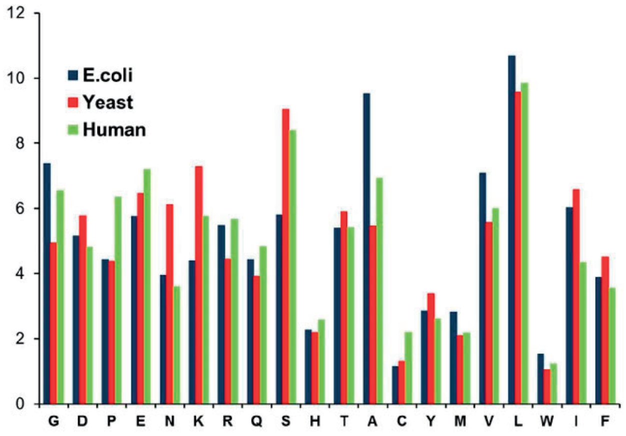


Figure S1. The relative frequency of amino acids for the three species studied.

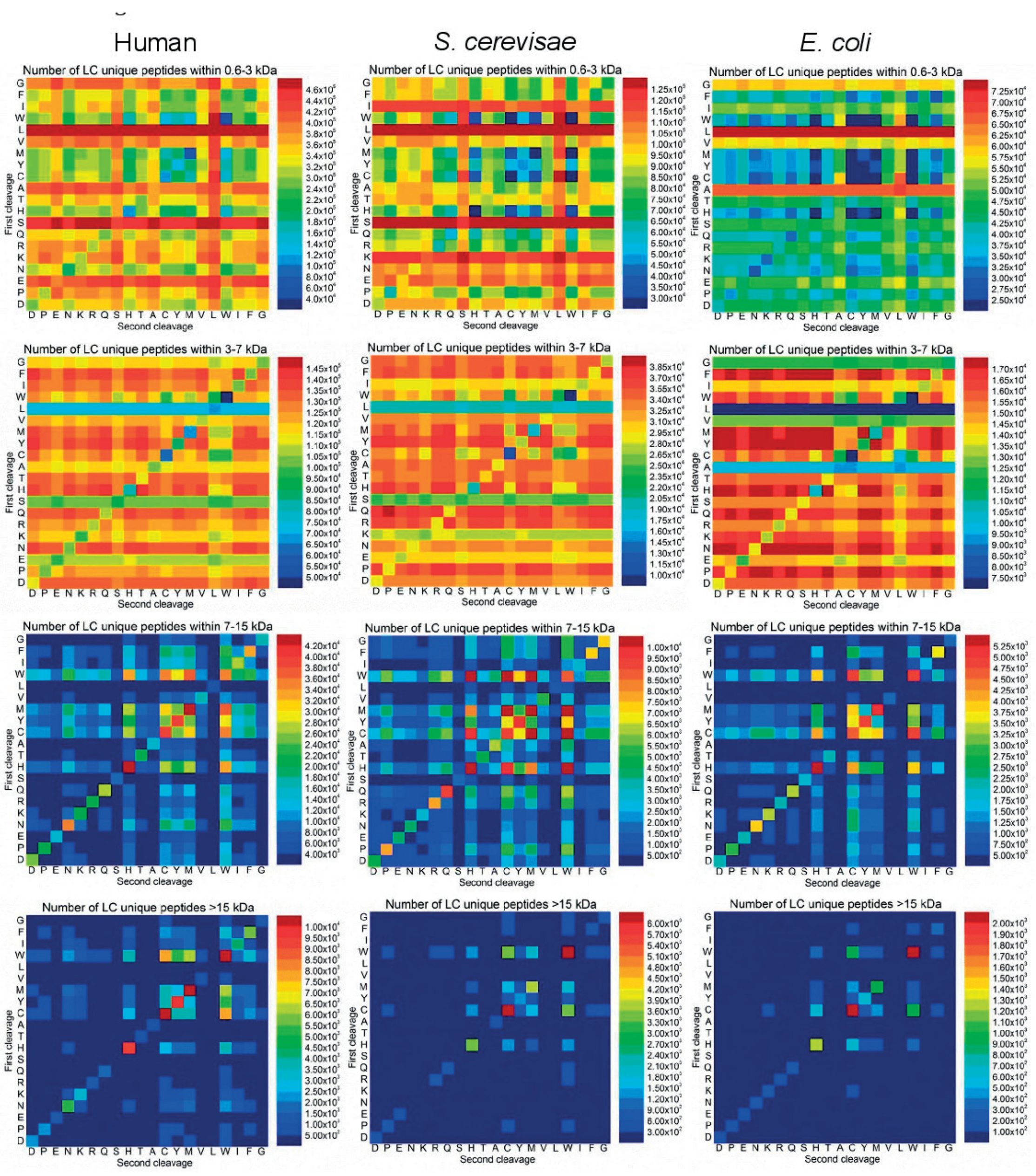


Figure S2. Theoretical distribution of LC unique peptides in the human (left), yeast (middle), and bacteria (right) proteome after two-step cleavage within a) 0.6-3 kDa, b) 3-7 kDa, c) 7-15 kDa, and e) >15 kDa ranges.

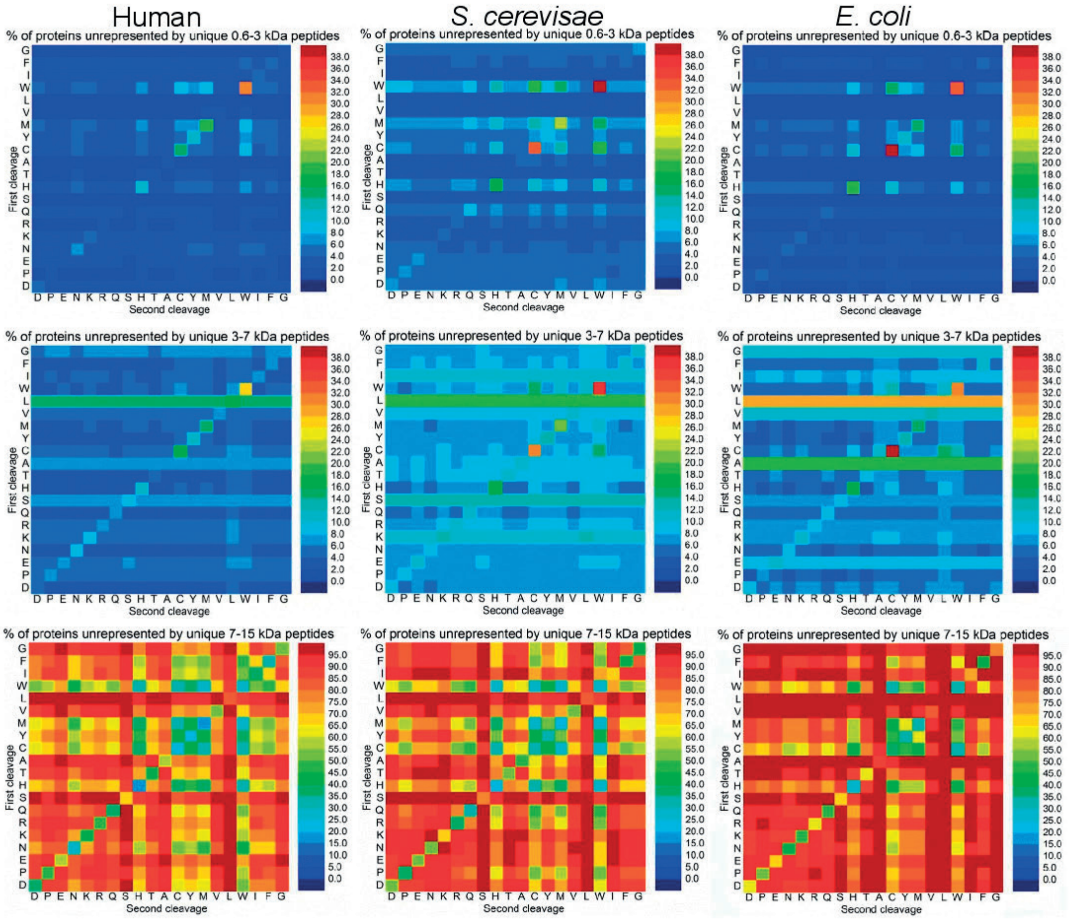


Figure S3. Theoretical distribution of proteins that remain unidentified in the human (left), yeast (middle), and bacteria (right) proteome after analysis of strictly unique peptides within a) 0.6-3 kDa, b) 3-7 kDa, c) 7-15 kDa, and e) >15 kDa ranges obtained after a two-step cleavage.

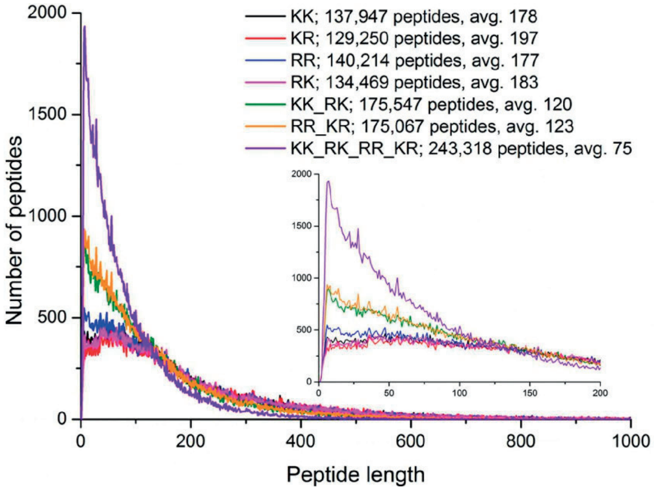


Figure S4. Peptide length distribution of unique peptides from dibasic-site specific digestion of the human proteome.

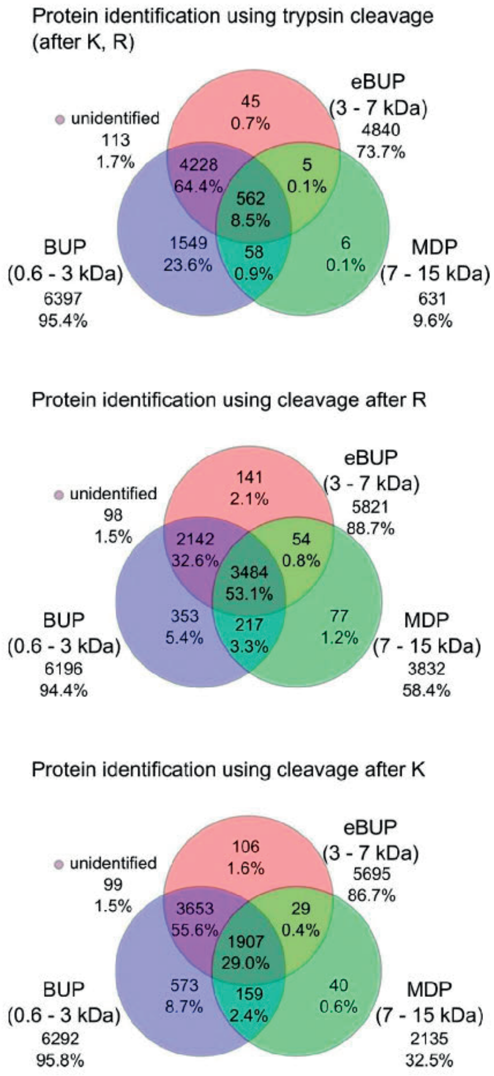


Figure S5. Venn diagrams of the number and % of yeast proteins identified by unique peptides in BUP, eBUP, and MDP approaches using (top) tryptic proteolysis (both K and R-specific cleavages); (middle) only K-specific cleavage; and (bottom) only R-specific cleavage.

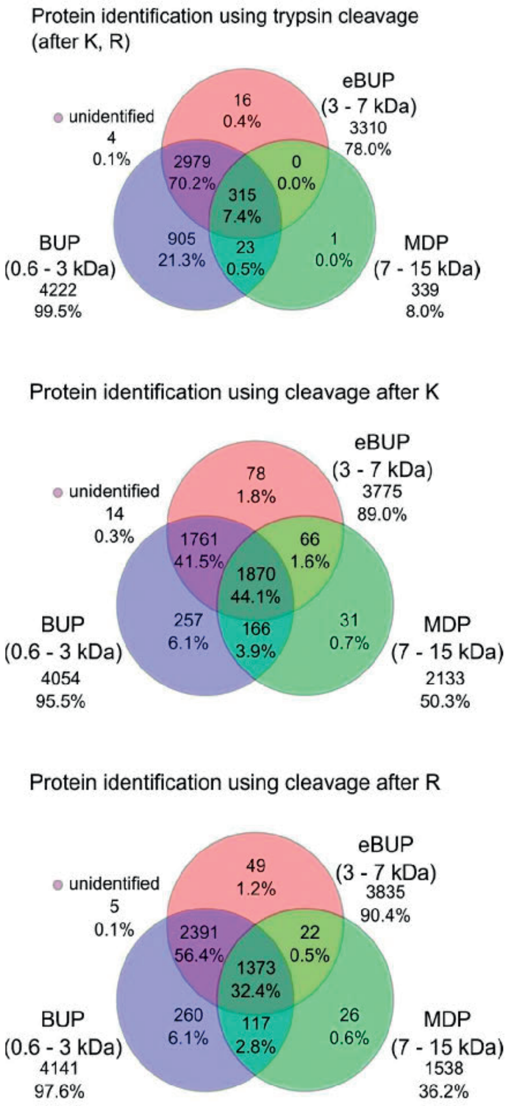
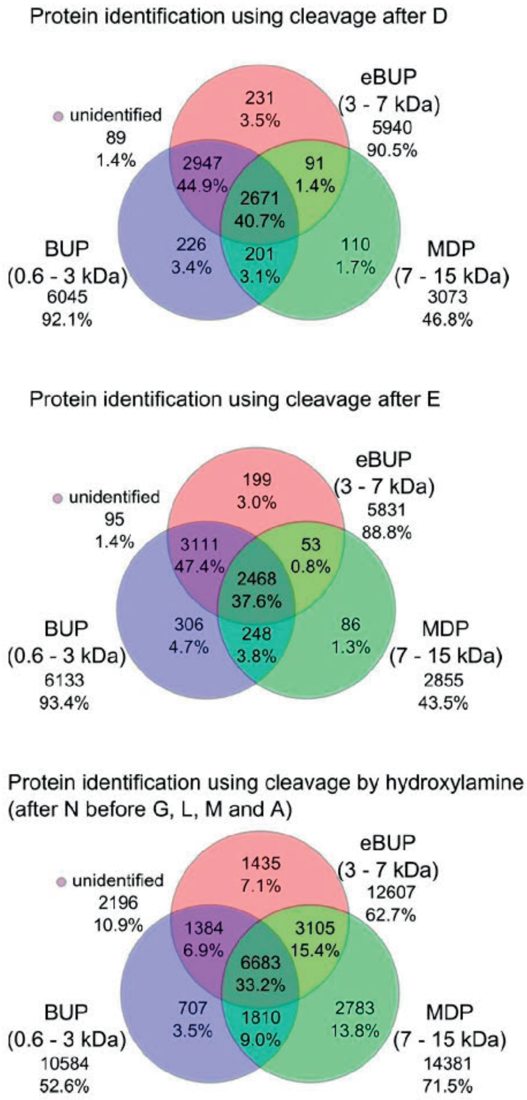


Figure S6. Venn diagrams of the number and % of bacterial proteins identified by unique peptides in BUP, eBUP, and MDP approaches using (top) tryptic proteolysis (both K and R-specific cleavage); (middle) only K-specific cleavage; and (bottom) only R-specific cleavage.

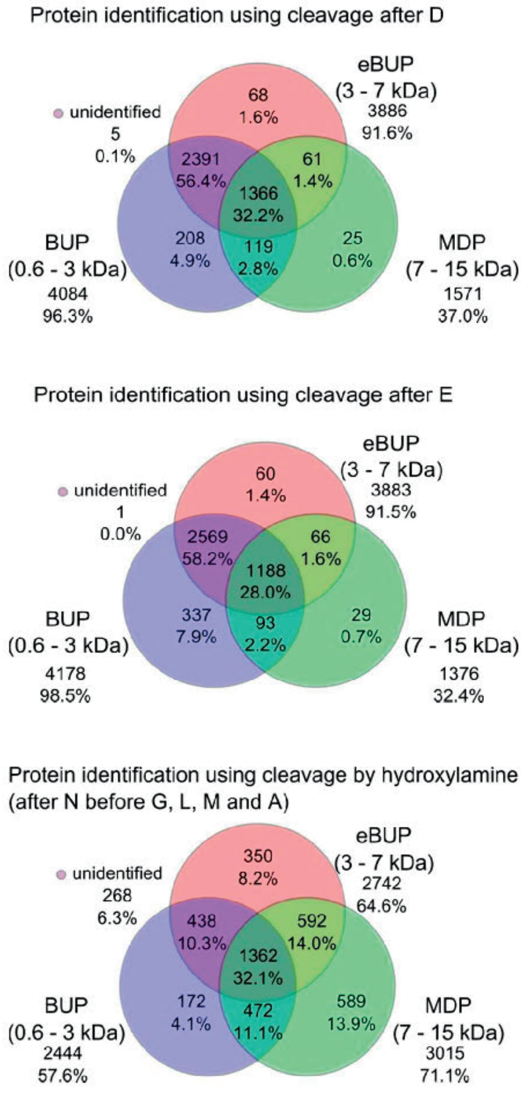


Protein identification using cleavage after D

Protein identification using cleavage after E

Protein identification using cleavage by hydroxylamine (after N before G, L, M and A)

Figure S7. Venn diagrams of the number and % of yeast proteins identified by unique peptides in BUP, eBUP, and MDP approaches using cleavage at (top)D; (middle) E; and (bottom) select N-X.



Protein identification using cleavage after D

Protein identification using cleavage after E

Protein identification using cleavage by hydroxylamine (after N before G, L, M and A)

Figure S8. Venn diagrams of the number and % of bacterial proteins identified by unique peptides in BUP, eBUP, and MDP approaches using cleavage at (top) D; (middle) E; and (bottom) select N-X.

.

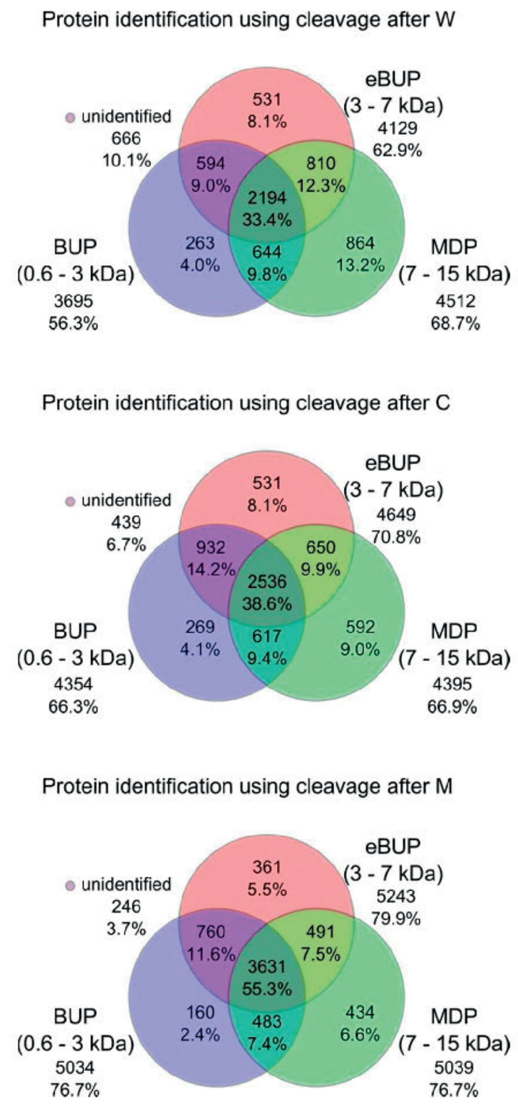


Figure S9. Venn diagrams of the number and % of yeast proteins identified by unique peptides in BUP, eBUP, and MDP approaches using cleavage at rare amino-acids: (top) Trp (W); (middle) Cys (C); and (bottom) Met (M).

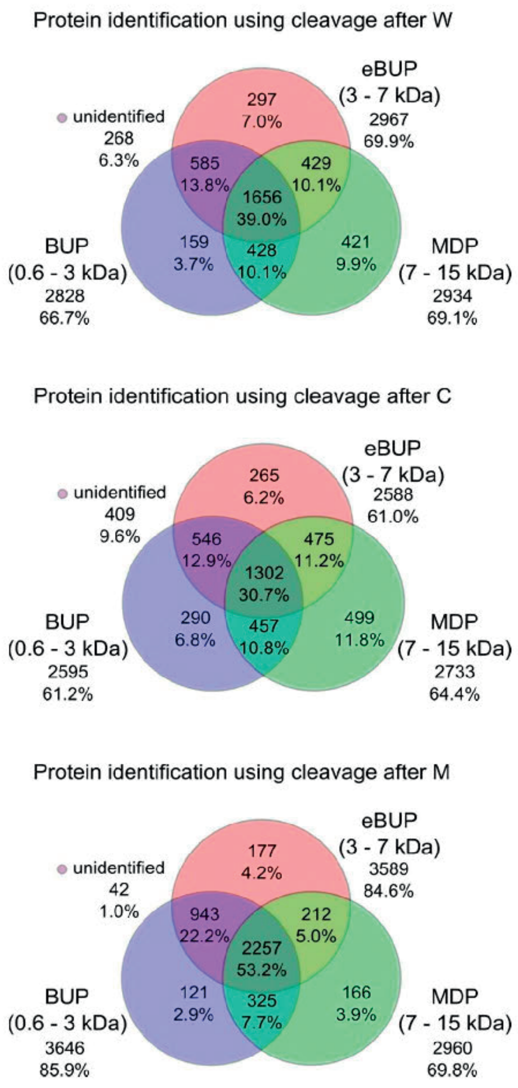


Figure S10. Venn diagrams of the number and % of bacterial proteins identified by unique peptides in BUP, eBUP, and MDP approaches using cleavage at rare amino-acids: (top) Trp (W); (middle) Cys (C); and (bottom) Met

(M).

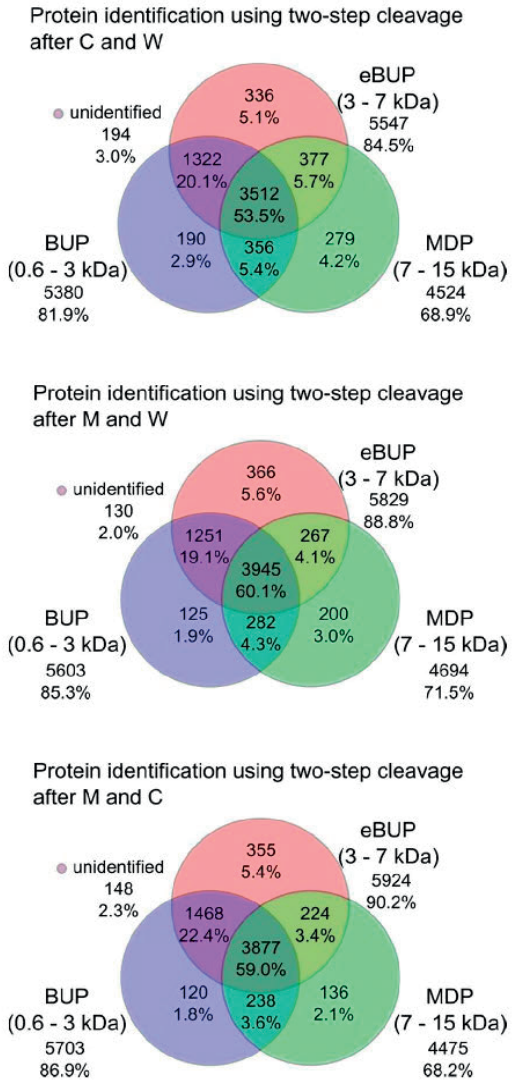


Figure S11. Venn diagrams of the number and % of yeast proteins identified by nique peptides in BUP, eBUP, and MDP approaches using two-step cleavages at amino-acid pairs: (top) Cys(C) followed by Trp (W); (middle) Met (M) followed by Trp (W); and (bottom) Met (M) followed by Cys (C).

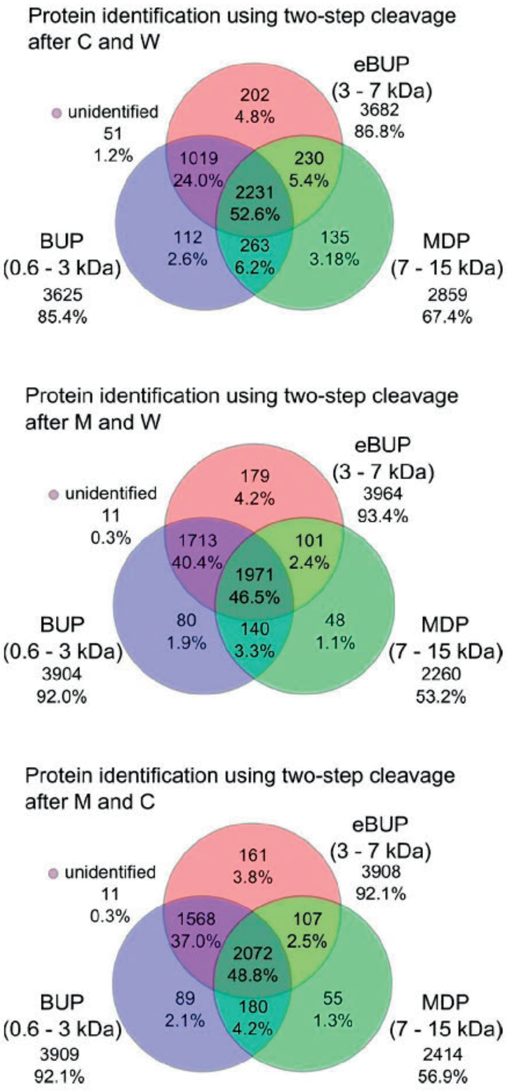


Figure S12. Venn diagrams of the number and % of bacterial proteins identified by unique peptides in BUP, eBUP, and MDP approaches using two-step cleavages at amino-acid pairs: (top) Cys(C) followed by Trp (W); (middle) Met (M) followed by Trp (W); and (bottom) Met (M) followed by Cys (C).

4.4. Paper II: Practical consideration for improving the productivity of mass spectrometry-based proteomics

244 CHIMIA **2013**, *67*, Nr. 4 LAUREATES: AWARDS AND HONORS, SCS FALL MEETING 2012

doi:10.2533/chimia.2013.244 Chimia 67 (2013) 244–249 © Schweizerische Chemische Gesellschaft

Practical Considerations for Improving the Productivity of Mass Spectrometry-based Proteomics

Ünige A. Laskay[§a], Kristina Srzentić[a], Luca Fornelli[a], Oxana Upir[a], Anton N. Kozhinov[a], Michel Monod[b], and Yury O. Tsybin[a*]

[§]SCS-Metrohm Foundation Award for best oral presentation

Abstract: Mass spectrometry (MS) is currently the most sensitive and selective analytical technique for routine peptide and protein structure analysis. Top-down proteomics is based on tandem mass spectrometry (MS/MS) of intact proteins, where multiply charged precursor ions are fragmented in the gas phase, typically by electron transfer or electron capture dissociation, to yield sequence-specific fragment ions. This approach is primarily used for the study of protein isoforms, including localization of post-translational modifications and identification of splice variants. Bottom-up proteomics is utilized for routine high-throughput protein identification and quantitation from complex biological samples. The proteins are first enzymatically digested into small (usually less than *ca.* 3 kDa) peptides, these are identified by MS or MS/MS, usually employing collisional activation techniques. To overcome the limitations of these approaches while combining their benefits, middle-down proteomics has recently emerged. Here, the proteins are digested into long (3–15 kDa) peptides *via* restricted proteolysis followed by the MS/MS analysis of the obtained digest. With advancements of high-resolution MS and allied techniques, routine implementation of the middle-down approach has been made possible. Herein, we present the liquid chromatography (LC)-MS/MS-based experimental design of our middle-down proteomic workflow coupled with post-LC supercharging.

Keywords: Electron transfer dissociation (ETD) · Higher-energy collisional dissociation (HCD) · Limited proteolysis · Middle-down proteomics · Post-column supercharging

Today, protein analysis using mass spectrometry (MS) is routine in numerous academic, commercial and clinical laboratories around the world. Depending on the goal of the study, among the most commonly employed approaches are bottom-up, top-down, and the newly emerging middle-down proteomics.[1] While bottom-up proteomics can be performed on fast, economical but low-resolution instruments, top-down and middle-down proteomics can only be performed on more expensive, high-resolution platforms, such as Fourier transform ion cyclotron resonance (FT-ICR) and Orbitrap, or state-of-the-art time-of-flight (TOF) instruments.[2]

Bottom-up proteomics is regularly used for high-throughput protein identification, quantitation, and targeted identification of post-translational modifications (PTMs). Here, proteins are cleaved into small (less than *ca.* 3 kDa) peptides with an enzyme, usually trypsin. The peptides are separated on a chromatographic column and analyzed individually, typically using data-dependent tandem mass spectrometry (MS/MS). Standardized protocols for both in-solution and in-gel digestion of proteins are well tested, separation of the small peptides is routinely achievable using both micro- and nanoflow rate liquid chromatography (LC), and high-throughput identification of thousands of proteins is possible from a single chromatographic run (*e.g.* 2,500 proteins/90 min.).[3] The MS/MS activation method can also be tailored to maximize the quality of the mass spectra. Collision-induced dissociation (CID) or higher-energy collisional dissociation (HCD) of short tryptic peptides is usually very efficient, and the interpretation of the mass spectra is straightforward. However, these collisional activation methods lead to preferential cleavage of the weak covalent bonds, therefore localization of PTMs is cumbersome. Electron capture and electron transfer dissociation (ECD and ETD) are techniques that can be used for fragmentation of multiply charged precursor ions. The main advantage of ETD and ECD is that the labile PTMs are preserved, however, fragmentation of peptides carrying less than three charges is inefficient and the MS/MS spectra of these are often uninformative.

Although a vast number of research efforts are employing bottom-up proteomics, this conventional approach carries several limitations arising from the high sample complexity and limited instrumental performance.[4] Specifically, biological samples typically contain thousands of proteins in a wide concentration range, simultaneous digestion of these leads to tens of thousands of peptides, which greatly increases sample complexity. Due to the limited separation capabilities of liquid chromatography and the time allotted for each MS/MS scan, only the most abundant

---

Correspondence: Prof. Dr. Y. O. Tsybin[a]
Tel.: +41 21 693 97 51
E-mail: yury.tsybin@epfl.ch
[a]Biomolecular Mass Spectrometry Laboratory
Ecole Polytechnique Fédérale de Lausanne
EPFL LSMB BCH 4307
CH-1015 Lausanne
[b]Department of Dermatology
Centre Hospitalier Universitaire Vaudois
CH-1011 Lausanne

of the co-eluting peptides are analyzed in data-dependent MS/MS.[5] This can be partially overcome by data independent MS/MS (termed MS^E MS/MS^{ALL} or SWATH), where all peptides present in a given *m/z* window are simultaneously fragmented, however, in such approach, the fragment ions and the precursors must be precisely related on the basis of their elution profile.[6,7] An additional shortcoming of bottom-up proteomics is that many proteins are present in multiple isoforms (PTMs, splice variants, *etc.*).[8] If a protein is present in only two isoforms (for instance, single phosphorylation and the non-phosphorylated variety), the only prerequisite for identification of both isoforms is the identification of the modified and non-modified peptide. However, without prior knowledge of the number of isoforms, or if multiple isoforms exist, the relationship between the modified peptide and the originating protein sequence is lost.[9]

In contrast, top-down proteomics is the MS-based method for the analysis of intact proteins. High-throughput identification of isoforms for select proteins has been reported in recent years. On-line 2D separation of purified histones on weak cation-exchange and hydrophilic interaction chromatography (WCX-HILIC) columns followed by ETD of the intact proteins allowed for the identification of 708 isoforms present in HeLa cells.[10] High-throughput PTM localization is also possible using top-down proteomics, Tran and coworkers identified 3000 variants of 1043 human proteins.[11] However, to achieve such a result, extensive four-dimensional fractionation and multiple LC-MS/MS runs were required, which is a serious shortcoming when limited sample quantities are available. The protein fractionation is necessary due to several reasons. The efficient LC separation of proteins is technically more difficult to achieve than it is for peptides. Since proteins are present in multiple charge states in the mass spectrum, co-eluting proteins might have overlapping signals that hinder the isolation of the individual protein signals. In addition, protein fragmentation is more cumbersome than peptide MS/MS analysis, and the fragmentation mass spectra often contain intersecting multiple charge state product ion signals. Moreover, a much higher number of MS/MS scans per LC peak must be accumulated for achieving sufficient signal to noise (S/N) ratio of the fragment ions. Another consideration is the extremely diverse protein sizes present in a complex mixture. Although analysis of small proteins (<20 kDa) can be performed without major instrument modifications, MS/MS analysis of larger proteins is not trivial. Recently, up to 32% sequence coverage of ~150 kDa intact monoclonal antibodies was obtained using electron transfer dissociation (ETD) on an Orbitrap FTMS and time-of-flight MS, as well as by electron capture dissociation (ECD) on FT-ICR MS.[12] However, for these accomplishments, hundreds of ETD/ECD mass spectra had to be averaged, which is not possible for all proteins on the timescale of LC separation. In general, application of ECD/ETD to proteins yields larger sequence coverage than collisional activation-based MS/MS. Nevertheless, if the protein is highly folded or the structure protected by disulfide bonds, fragmentation of the internal backbone bonds is not efficient by either MS/MS method.

Middle-down proteomics is an approach that aims to combine the benefits of bottom-up and top-down approaches, while minimizing their above-mentioned limitations. Here, similarly to bottom-up, proteins are digested, however, a restricted (less frequent) proteolysis is employed to increase the average size of the resulting peptides (3–15 kDa), as detailed below. Due to the lower number of resulting peptides, the sample is less complex than in the bottom-up approach, but MS/MS analysis can still be performed in a high-throughput manner. Specifically, efficient separation of these long peptides can be readily performed on commercial chromatographic columns, and the elution profile and LC peak capacities are comparable to those of the bottom-up approach. Moreover, due to the decreased sample complexity, the number of co-eluting peptides is also reduced, as detailed further below. Although a longer acquisition time is necessary for recording of high resolution MS/MS spectra, this is achievable with modern instrumentation, such as the Orbitrap FTMS. In addition, the long amino acid series enhances the uniqueness of the sequence and increases the chance for identification of peptides that carry a modification.

Therefore, we consider this paradigm shift towards analysis of longer peptides to be the key for achieving increased dynamic range of protein concentrations and high-throughput identification of targeted protein isoforms. However, sample preparation, peptide separation, ionization conditions, fragmentation parameters, data acquisition and data analysis must be appropriately tailored to analysis of long peptides. Herein, we first identify a suitable protease for the middle-down approach that ensures a fast digestion into long peptides and results in high protein sequence coverage. In addition, we present post-column supercharging employed to increase the average charge state of the long peptides, and, consequently, the fragmentation efficiency; and finally consider the practical aspects of such setup for high-throughput protein analysis.

Theoretical Considerations

Although trypsin is a well-characterized protease that is routinely used in bottom-up proteomics, it is not ideal, since it produces tens of thousands of short (4–10 residue) peptides. This, in turn, is detrimental due to ineffective use of the LC column when the valuable binding sites are saturated by excessive numbers of short, uninformative peptides. In addition, an MS/MS scan is performed on all multiply charged precursor ions and, if several species elute in a narrow elution window, fragmentation of short peptides may be performed to the detriment of co-eluting longer ones. Moreover, these fragmentation mass spectra are often not useful, since the probability that the sequence is unique to a particular protein decreases with decreasing peptide length. Finally, short peptides (<1 kDa) may be removed prior to analysis using molecular weight cut-off (MWCO) filters, however, as any additional handling operation, this may also lead to sample loss. It is therefore desirable to perform the proteolysis at less frequent amino acid sites or specific amino acid patterns with a required repetition rate.

Our survey of the yeast proteome (as well as that of human and bacteria, data not shown here) suggests that cleavage at dibasic residues greatly decreases the number of peptides yielded by proteolysis. Fig. 1 shows the theoretical peptide size distribution of proteolytic peptides obtained *via in silico* digestion of the *E. coli* proteins, with zero missed cleavages allowed, with trypsin (top panel) and with a protease Sap9 (bottom panel) that cleaves after two adjacent basic amino acids, arginine and lysine, *vide infra*. Trypsin produced a total number of ~285,000 peptides with unique sequences, 280,000 of these were within the length range of 2–50 amino acid residues and had an average length of 15.2 residues. In contrast, the dibasic site cleavage yielded *ca.* 100,000 peptides, with 65,000 within the length range of 2–100 residues and had an average length of 58.2 residues. Because the peptides obtained with dibasic residue cleavage are more uniformly distributed across the 20–100 amino acid size range, one may expect that the chromatographic separation of these will be more efficient. Particularly, the elution will be spread along a wider gradient region. As a result, a better MS sampling of the eluting peptides may be obtained. This is important when considering that long, highly charged peptides yield multiply charged product ions, requiring high resolution mass analysis and, consequently, longer acquisition time per MS/MS spectrum compared to analysis of short peptides in a trap instrument.

Tailoring the proteolysis site to ensure

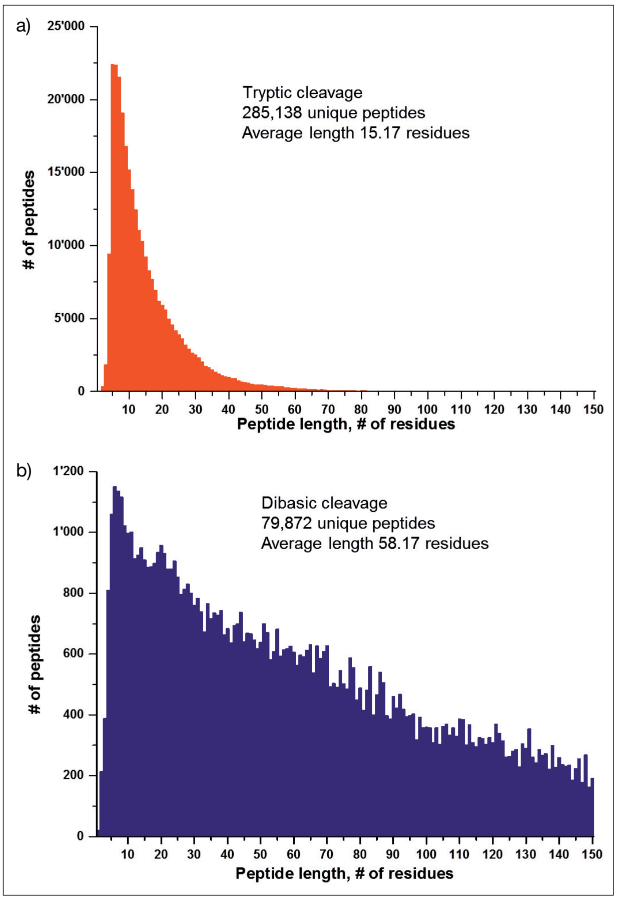


Fig. 1. Size distribution of proteolytic peptides in the 1–100 residue range obtained *via in silico* digestion of the *E. coli* proteins database with trypsin (panel a) and a dibasic protease Sap9 (panel b). Zero missed cleavages were allowed in the calculations.

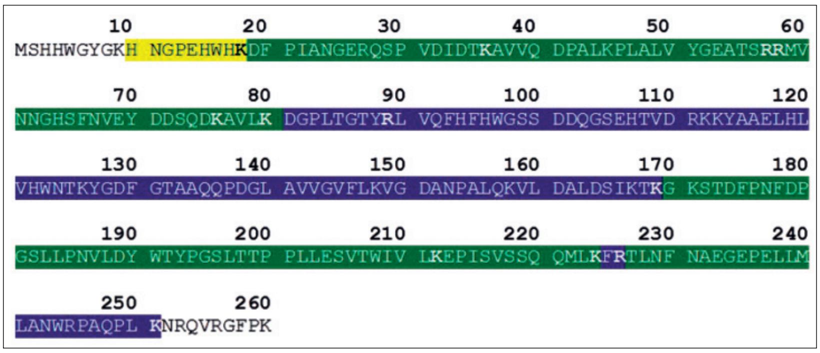
longer average peptides is the driving force for the development of middle-down proteomics. Although several proteases, such as AspN, LysC and GluC, as well as microwave-assisted acid proteolysis have been utilized for obtaining long peptides,[13–17] the occurrence of these amino acid sites (data not shown) is more frequent than the occurrence of two adjacent basic residues. As a result, targeting the dibasic residues for cleavage offers more desirable peptide size range. Kex2[18] and OmpT[19] have been previously described to have dibasic site specificity. However, recombinant Kex2 is specific only to KR and RR, and not KK and RK sites therefore decreasing the cleavage possibilities and increasing the average peptide size beyond the 10 kDa mass range. The working regime for very long peptides approaches the top-down approach, where both the LC and the mass spectrometer operating parameters must be specifically tailored for efficient separation and timely fragmentation. OmpT is a protein construct which appears to be cumbersome to produce, required re-folding prior utilization, and it has been found to extensively cleave at other amino acids, as well.

Sap9, a Novel Protease for Middle-down Proteomics

In our approach, we sought to establish the enzymatic activity of the *Candida albicans* aspartic protease Sap9 overexpressed in *Pichia pastoris*. The protease production is highly efficient (in the order of g/L), it is excreted in the extracellular medium, and can be effortlessly purified from the supernatant using the His-tag approach, as described by Albrecht and coworkers.[20] For the cleavage specificity study, commercial carbonic anhydrase (29 kDa) obtained from Protea (Morgantown, WV) was digested with Sap9 at pH 4.5, enzyme:protein ratio 1:2.5, 37 °C. We performed extensive enzyme activity studies and found these conditions to be the most appropriate for efficient proteolysis. Aliquots were removed hourly for eight hours. Peptides were desalted with C18 ZipTip (Millipore, Billerica, MA) and separated on a C8 column (Thermo Scientific, 15 cm, 2 µm, 100 Å) using a 45 min. H_2O/ACN:MeOH:TFE (2:5:1) gradient. Eluted peptides were detected using LTQ Orbitrap Elite FTMS equipped with a high-field Orbitrap mass analyzer and provided with the eFT signal processing algorithm. MS scan was performed at 60,000 resolution (at 400 *m/z*) while HCD mass spectra were acquired at 15,000 resolution setting. Protein database search was performed using Sequest and Mascot against the *Bos taurus* database. We performed a fully tryptic search allowing for nine missed cleavages, as well as a no-enzyme search to avoid biasing of the results towards preferred cleavage sites.

Fig. 2 shows the sequence coverage obtained by Sequest (yellow) and Mascot after four hours of proteolysis. The regions identified by both algorithms are noted in green (tryptic peptides) and blue (nonspecific cleavage). Amino acids denoted in bold letters indicate the C terminal positions of the trypsin-like cleavages.

After four hours of digestion, using a fully tryptic database search with nine missed cleavages, Mascot identified 58.8% of the carbonic anhydrase sequence, the average peptide mass was 3.1 kDa. Sequest performed slightly better, yielding 62.3% sequence coverage, average peptide mass 2.7 kDa. The longest peptide identified was 6106.08 Da and the sequence (170–224) is unique to bovine carbonic anhydrase. When performing a no-enzyme search, we obtained 95% sequence coverage



MSHHWGYGKH NGPEHWH**K**DF PIANGERQSP VDIDT**K**AVVQ DPALKPLALV YGEATS**RR**MV
NNGHSFNVEY DDSQD**K**AVL**K** DGPLTGTY**R**L VQFHFHWGSS DDQGSEHTVD RKKYAAELHL
VHWNTKYGDF GTAAQQPDGL AVVGVFLKVG DANPALQKVL DALDSIKT**K**G KSTDFPNFDP
GSLLPNVLDY WTYPGSLTTP PLLESVTWIV L**K**EPISVSSQ QML**K**F**R**TLNF NAEGEPELLM
LANWRPAQPL **K**NRQVRGFPK

Fig. 2. Sequence coverage of bovine carbonic anhydrase digested for 4 h with Sap9 at pH 4.5, 37 °C, protein:enzyme ratio 1:2.5 (w/w). In yellow is the tryptic peptide identified by Sequest, in green are indicated the regions identified by both Sequest and Mascot. Blue indicates the protein region identified using a no enzyme search.

with both engines. Of the 81 total cleavages observed, only 6 occurred at dibasic sites, 13 were tryptic and 32 half-tryptic. In addition, we observed 26 peptides with cleavage C-terminal to hydrophobic residues. Only 5 peptides were detected with hydrophilic C terminus other than K or R. Shorter digestion times (<2 h) yielded a combination of short (<2 kDa) and long (>10 kDa) peptides, while longer interaction times (5–8 h) yielded increasingly shorter peptides.

The obtained results indicate that at the conditions where Sap9 is the most active, its site specificity is not ideal. Under the experimental conditions presented herein we observed a secondary selectivity towards hydrophobic residues in addition to the basic sites. We are currently investigating the optimal proteolysis conditions where cleavage specificity is tailored toward more predictable dibasic sites whereas the protease activity is still sufficiently high.

Peptide Fragmentation Study

We sought to establish the efficiency and quality of the mass spectra obtained using the two most commonly used fragmentation techniques, HCD and ETD, applied to the analysis of middle-down range peptides. Specifically, we investigated whether the exact location of the dibasic site cleavage, *i.e.* proteolysis after or between the two adjacent basic residues, has an effect on the quality of the MS/MS mass spectra. We performed a series of experiments where the HCD activation energy and ETD reaction time was gradually varied for all charge states of the 34 residue synthetic peptides mimicking cytochrome C sequence (Peptide Synthesis Facility, University of Lausanne, Switzerland) TGQAPGFSYTDANKNKGITWGEETLMEYLENPKK and KTGQAPGFSYTDANKNKGITWGEETLMEYLENPK. Peptides were dissolved in 50:50 ACN:H_2O solvent mixture containing 0.1% of formic acid to the final concentrations of *ca.* 10 µM. Ions were generated using a nano-electrospray ionization (nESI) ion source (Triversa Nanomate, Advion Biosciences, Ithaca, NY, USA) at a flow rate of *ca.* 300 nL/min. HCD normalized collision energy (NCE) was varied between 0 and 35 in increments of 5, while ETD interaction time was varied from 0.1 to 100 ms in increments of 5 ms. Fig. 3 shows the representative MS/MS spectra for the 5+ charge state of the two peptides obtained at NCE 20 and 25 ms ETD reaction time, respectively. The most notable difference between the HCD mass spectra of the two peptides is the absence of the b_3-b_5 ion series when both peptide termini are K. This is likely due to the sequestration of the proton by the K side-chain does not allow charge-directed fragmentation, as predicted by the mobile proton model.[21] As indicated by the absence of products from the innermost positions, the distal positioning of the two fixed charges is detrimental for the cleavage of peptide bonds in the middle of the sequence, even at the collision energy where virtually no precursor ion remains. In contrast, the c_2-c_6 ion series is completely absent from the ETD mass spectrum of the peptide with adjacent basic residues, indicating that, for these product ions, the N-terminal charge is neutralized upon ETD.[22] Nonetheless, ETD fragmentation of both peptides yielded informative mass spectra and almost complete (93%) sequence coverage. We performed similar studies on numerous peptides with varying lengths. We have found that, as expected, sequence coverage is improved for both

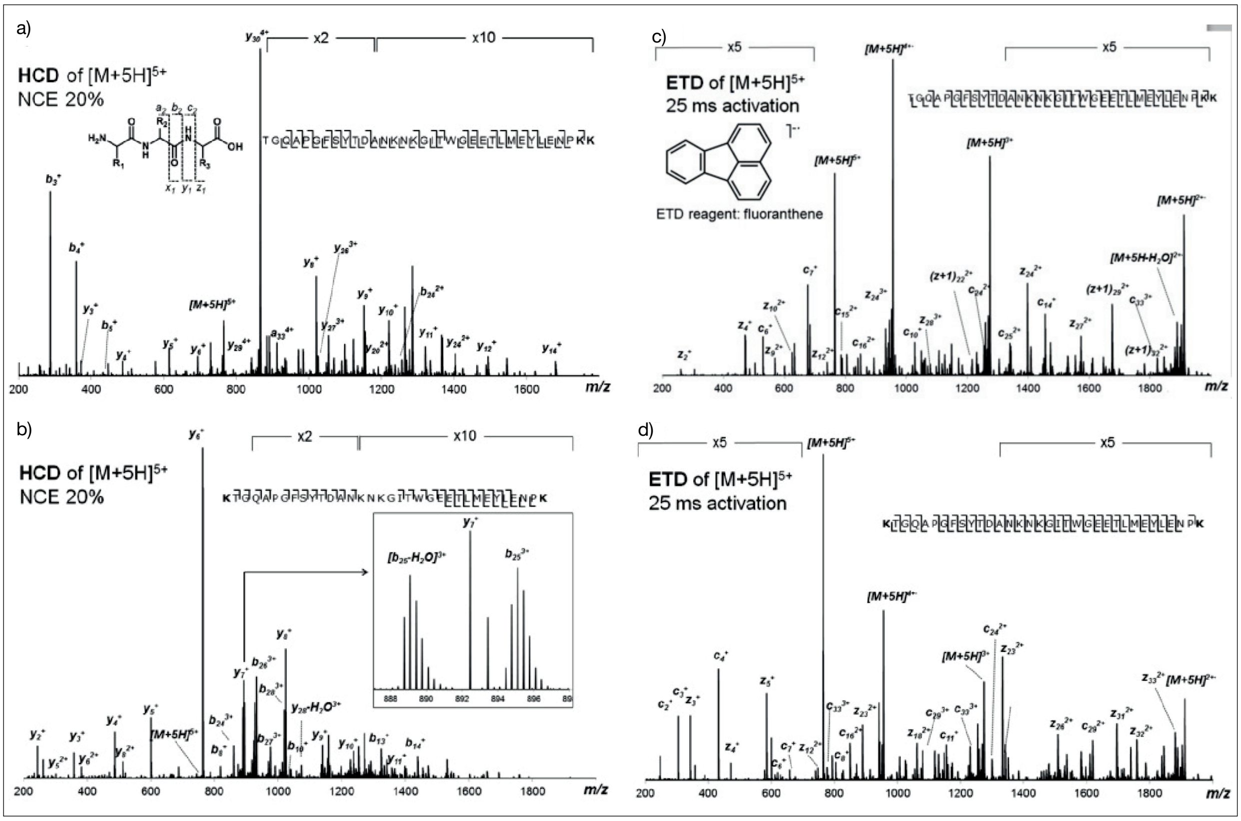


Fig. 3. Comparison of the fragmentation mass spectra of the 5+ precursor ions of the peptides TGQAPGFSYTDANKNKGITWGEETLMEYLENPKK and KTGQAPGFSYTDANKNKGITWGEETLMEYLENPK obtained with HCD (panels *a* and *b*) and ETD (panels *c* and *d*). Ion activation conditions were: NCE 20 for HCD, and 25 ms of ion activation time for ETD. The inset in panel b demonstrates the utility of high resolution in the identification and assignment of product ions.

HCD and ETD fragmentation with increasing precursor ion charge state of each peptide, regardless of the positioning of the terminal basic residues.

Post-column Supercharging

To ensure that the peptides carry the maximum number of charges and to increase the signal to noise ratio,[23] while minimizing the total number of charge states, we utilized post-column supercharging. For this, a mixture of five proteins was digested overnight with LysC, the peptides eluting from the chromatographic column were continuously reacted with 0.5% *m*-NBA (*m*-nitrobenzyl alcohol) in 50% ACN (both from Sigma Aldrich). The reagent was introduced *via* a zero-dead-volume Y junction (Idex, Oak Harbor, WA), as illustrated in Fig. 4. The supercharging reagent flow was set to match the flow from the column (0.3 µL/min) to minimize turbulence. The inset in Fig. 4 shows the minimal effect of post-column supercharging on the chromatographic peak shape. Representative mass spectra obtained with and without supercharging are shown in Fig. 5. All other experimental conditions (LC flow rate, spray voltage, number of precursor ions, maximum ion injection time, mass resolution setting) were the same for both experiments. To ensure that the change in charge state distribution is the effect of *m*-NBA, and not a solvent-effect, for the non-supercharged experiment we infused 50% ACN.

The mass spectra in Fig. 5 contain several species, two of which are highlighted in red and blue. Under normal conditions (no supercharging, top panel) the 'red' peptide was present under four different charge states (+4, +5, +6, and +7), while the 'blue' peptide was detected with +3 and +4 charges. The most abundant charge state for the 'red' and 'blue' peptides were +6, and +4, respectively. When the supercharging reagent was added, the peaks corresponding to the 'red' +4 and +5 charge states were greatly diminished, while the +7 peak now had the highest S/N. Similarly, the ratio of the +4 and +3 precursors of the 'blue' peptide increased from 2:1 to 12:1. Fig. 6 shows the total (not unique) number of precursor ions with different charge states obtained with and without supercharging. As indicated by the ratio of the red and blue columns, the number of +2 precursor ions decreased slightly, while the number of MS/MS spectra that had precursor ions of >2 charges increased. This shows that the overall S/N of high charge state species increases and more precursor ions are selected for MS/MS in the data dependent scanning event. This is an important aspect for middle-down proteomics, since higher S/N precursor ion yields better quality MS/MS spectra and requires shorter scanning times.

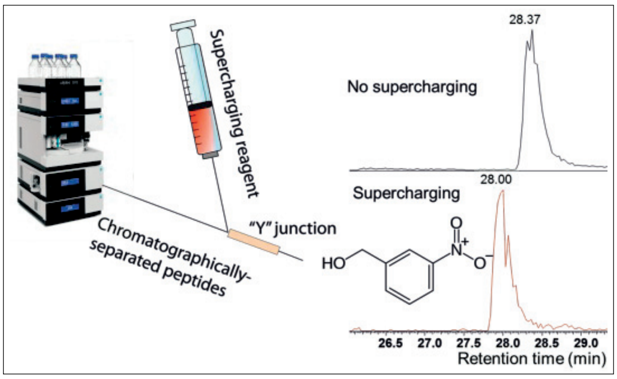


Fig. 4. Schematic representation for the introduction of the supercharging reagent after chromatographic separation, using a zero-dead-volume Y connector. The inset shows that the chromatographic peak shape is not significantly affected with introduction of the reagent.

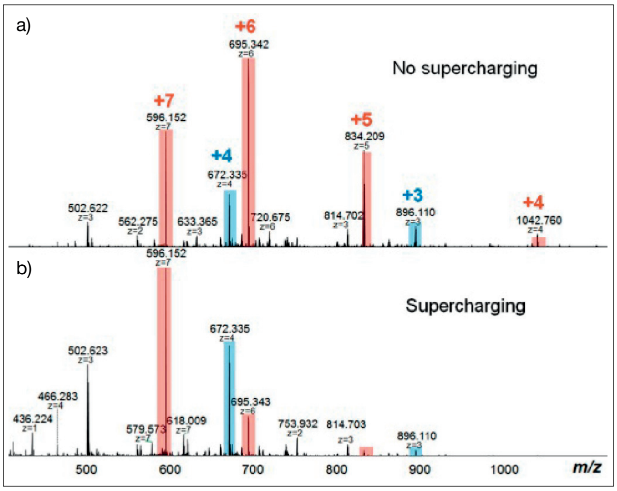


Fig. 5. Experimental mass spectra of peptides eluted from the chromatographic column a) without b) with addition of 0.5% *m*-NBA supercharging reagent. The S/N of high charge states for peptides indicated in red and blue increases by supercharging, while the signal of low charge states are significantly diminished.

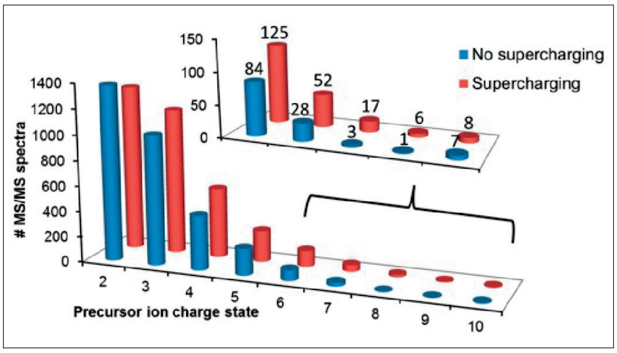


Fig. 6. Total precursor ion charge state distribution with (red) and without (blue) post-column supercharging for a mixture of five standard proteins digested overnight with LysC.

Conclusions

Sap9 is a promising protease for middle-down proteomics and yields close to complete sequence coverage in as short as four hours of digestion. This, in turn, improves the probability for detection and localization of PTMs and the identification of splice variants. It can also be important when the goal of the study is to distinguish between proteins with very similar sequences, such as in targeted species identification based on a reference protein.[24] To improve the MS and MS/MS data quality, the S/N of highly charged peptides can be increased by introduction of the supercharging reagent post column. This has the advantage of not interacting with the stationary phase, therefore eliminating contamination, and preventing changes in chromatographic separation. These improvements of the proteomic workflow can be implemented for the high throughput analysis of complex protein mixtures using state-of-the-art high resolution instrumentation. In addition, for the peptide size range <7 kDa, protein database search engines well established for bottom-up proteomics (Sequest and Mascot) can be successfully employed, without the need for customized in-house built algorithms.

Acknowledgements

The work was supported by the Swiss National Science Foundation (Projects 200021-125147/1 and 128357) and the European Research Council (ERC Starting Grant 280271). We thank Konstantin Zhurov for the helpful discussions.

Received: March 4, 2013

---

[1] V. Marx, *Nat. Methods* **2013**, *10*, 201.

[2] Y. O. Tsybin, L. Fornelli, A. N. Kozhinov, A. Vorobyev, S. M. Miladinovic, *Chimia* **2011**, *65*, 641.

[3] A. Michalski, E. Damoc, J. Hauschild, O. Lange, A. Wieghaus, A. Makarov, N. Nagaraj, J. Cox, M. Mann, S. Horning, *Mol. Cell. Proteom.* **2011**, *10*, M111.011015-M111.011015.

[4] M. W. Duncan, R. Aebersold, R. M. Caprioli, *Nat. Biotechnol.* **2010**, *28*, 659.

[5] A. Michalski, J. Cox, M. Mann, *J. Proteome Res.* **2011**, *10*, 1785.

[6] T. Geiger, J. Cox, M. Mann, *Mol. Cell. Proteom.* **2010**, *9*, 2252.

[7] L. C. Gillet, P. Navarro, S. Tate, H. Roest, N. Selevsek, L. Reiter, R. Bonner, R. Aebersold, *Mol. Cell. Proteomics* **2012**, *18*, O111.016717.

[8] H. Schluter, R. Apweiler, H. Holzhutter, P. Jungblut, *Chem. Cent. J.* **2009**, *3*, 11.

[9] L. M. Smith, N. L. Kelleher, M. Linial, D. Goodlett, P. Langridge-Smith, Y. Ah Goo, G. Stafford, L. Bonilla, G. Kruppa, R. Zubarev, J. Rontree, J. Chamot-Rooke, J. Garavelli, A. Heck, J. Loo, D. Penque, M. Hornshaw, C. Hendrickson, L. Pasa-Tolic, C. Borchers, D. Chan, N. Young, J. Agar, C. Masselon, M. Gross, F. McLafferty, Y. Tsybin, Y. Ge, I. Sanders, J. Langridge, J. Whitelegge, A. Marshall, *Nat. Methods* **2013**, *10*, 186.

[10] Z. Tian, N. Tolic, R. Zhao, R. Moore, S. Hengel, E. Robinson, D. Stenoien, S. Wu, R. Smith, L. Pasa-Tolic, *Genome Biol.* **2012**, *13*, R86.

[11] J. C. Tran, L. Zamdborg, D. R. Ahlf, J. E. Lee, A. D. Catherman, K. R. Durbin, J. D. Tipton, A. Vellaichamy, J. F. Kellie, M. Li, C. Wu, S. M. Sweet, B. P. Early, N. Siuti, R. D. LeDuc, P. D. Compton, P. M. Thomas, N. L. Kelleher, *Nature* **2011**, *480*, 254.

[12] L. Fornelli, E. Damoc, P. M. Thomas, N. L. Kelleher, K. Aizikov, E. Denisov, A. Makarov, Y. O. Tsybin, *Mol. Cell. Proteom.* **2012**, *11*, 1758.

[13] S. D. Taverna, B. M. Ueberheide, Y. Liu, A. J. Tackett, R. L. Diaz, J. Shabanowitz, B. T. Chait, D. F. Hunt, C. D. Allis, *Proc. Natl. Acad. Sci. U.S. A.* **2007**, *104*, 2086.

[14] S. Wu, A. F. R. Hühmer, Z. Hao, B. L. Karger, *J. Proteome Res.* **2007**, *6*, 4230.

[15] L. Hohmann, C. Sherwood, A. Eastham, A. Peterson, J. K. Eng, J. S. Eddes, D. Shteynberg, D. B. Martin, *J. Proteome Res.* **2009**, *8*, 1415.

[16] N. J. Hauser, H. Han, S. A. McLuckey, F. Basile, *J. Proteome Res.* **2008**, *7*, 1867.

[17] S. Swatkoski, P. Gutierrez, C. Wynne, A. Petrov, J. D. Dinman, N. Edwards, C. Fenselau, *J. Proteome Res.* **2008**, *7*, 579.

[18] M. Rholam, N. Brakch, D. Germain, D. Y. Thomas, C. Fahy, H. Boussetta, G. Boileau, P. Cohen, *Eur. J. Biochem.* **2008**, *227*, 707.

[19] C. Wu, J. C. Tran, L. Zamdborg, K. R. Durbin, M. Li, D. R. Ahlf, B. P. Early, M. Thomas, J. V. Sweedler, N. L. Kelleher, *Nat. Methods* **2012**, *9*, 822.

[20] J. Sarfati, M. Monod, P. Recco, A. Sulahian, C. Pinel, E. Candolfi, T. Fontaine, J. P. Debeaupuis, M. Tabouret, J. P. Latgé, *Diagn. Microbiol. Infect. Dis.* **2006**, *55*, 279.

[21] A. R. Dongre, J. L. Jones, A. Somogyi, V. H. Wysocki, *J. Am. Chem Soc.* **1996**, *118*, 8365.

[22] K. O. Zhurov, L. Fornelli, M. D. Wodrich, Ü. A. Laskay, Y. O. Tsybin, *Chem. Soc. Rev.* **2013**, DOI: 10.1039/c3cs35477f.

[23] S. M. Miladinović, L. Fornelli, Y. Lu, K. M. Piech, H. H. Girault, Y. O. Tsybin, *Anal. Chem.* **2012**, *84*, 4647.

[24] Ü. A. Laskay, J. Burg, E. J. Kaleta, I. -. E. Vilcins, S. R. Telford III, A. G. Barbour, V. H. Wysocki, *Biol. Chem.* **2012**, *393*, 195.

Chapter 5. Enabling MDP proteolysis: a quest for a novel protease

This Chapter is dedicated to the characterization of a novel protease for middle-down (MD) approach. Particularly, we addressed prerogative physical/chemical properties of suggested protease and investigated its efficiency in generating peptides in required mass bin with MS. Finally, we developed and applied the novel pipeline to a case study of immunoglobulins structural analysis, tackling one of the major PTM (deamidation) for this class of proteins.

Hereinafter reported considerations and results are translated into research articles enclosed at the end of this Chapter:

- Extended bottom-up proteomics with secreted aspartic protease Sap9 (Paper III)
- Advantages of extended bottom-up proteomics using Sap9 for analysis of monoclonal antibodies (Paper IV)

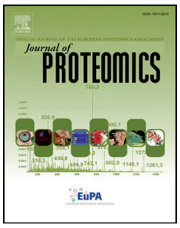
5.1. Paper III: Extended bottom-up proteomics with secreted aspartic protease Sap9

JOURNAL OF PROTEOMICS 110 (2014) 20–31

Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot



Extended bottom-up proteomics with secreted aspartic protease Sap9

Ünige A. Laskay[a], Kristina Srzentić[a], Michel Monod[b], Yury O. Tsybin[a,]*

[a]Biomolecular Mass Spectrometry Laboratory, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
[b]Department of Dermatology, Centre Hospitalier Universitaire Vaudois, 1011 Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 15 April 2014
Accepted 18 July 2014

Keywords:
Bottom-up proteomics
Extended bottom-up proteomics
Middle-down proteomics
Enzymatic digestion
High-resolution mass spectrometry

ABSTRACT

We investigate the benefits and experimental feasibility of approaches enabling the shift from short (1.7 kDa on average) peptides in bottom-up proteomics to about twice longer (~3.2 kDa on average) peptides in the so-called extended bottom-up proteomics. *Candida albicans* secreted aspartic protease Sap9 has been selected for evaluation as an extended bottom-up proteomic-grade enzyme due to its suggested dibasic cleavage specificity and ease of production. We report the extensive characterization of Sap9 specificity and selectivity revealing that protein cleavage by Sap9 most often occurs in the vicinity of proximal basic amino acids, and in select cases also at basic and hydrophobic residues. Sap9 is found to cleave a large variety of proteins in a relatively short, ~1 h, period of time and it is efficient in a broad pH range, including slightly acidic, e. g., pH 5.5, conditions. Importantly, the resulting peptide mixtures contain representative peptides primarily in the target 3–7 kDa range. The utility and advantages of this enzyme in routine analysis of protein mixtures are demonstrated and the limitations are discussed. Overall, Sap9 has a potential to become an enzyme of choice in an extended bottom-up proteomics, which is technically ready to complement the traditional bottom-up proteomics for improved targeted protein structural analysis and expanded proteome coverage.

Biological significance

Advances in biological applications of mass spectrometry-based bottom-up proteomics are oftentimes limited by the extreme complexity of biological samples, e.g., proteomes or protein complexes. One of the reasons for it is in the complexity of the mixtures of enzymatically (most often using trypsin) produced short (<3 kDa) peptides, which may exceed the analytical capabilities of liquid chromatography and mass spectrometry. Information on localization of protein modifications may also be affected by the small size of typically produced peptides. On the other hand, advances in high-resolution mass spectrometry and liquid chromatography have created an intriguing opportunity of improving proteome analysis by gradually increasing the size of enzymatically-derived peptides in MS-based bottom-up proteomics. Bioinformatics has already confirmed the

---

Abbreviations: AGC, automatic gain control; ETD, electron transfer dissociation; FTMS, Fourier transform mass spectrometer/mass spectrometry; HCD, higher energy collision dissociation; IAA, iodoacetamide; MeOH, methanol; PTM, post-translational modification; RFU, relative fluorescence units; Sap9, secreted aspartic protease 9; TFA, trifluoroacetic acid; TFE, 2,2,2-trifluoroethanol.

* *Corresponding author at:* EPFL ISIC LSMB, BCH 4307, 1015 Lausanne, Switzerland. Tel.: +41 21 693 97 51; fax: +41 21 693 98 95.
E-mail address: yury.tsybin@epfl.ch (Y.O. Tsybin).

http://dx.doi.org/10.1016/j.jprot.2014.07.035
1874-3919/© 2014 Elsevier B.V. All rights reserved.

envisioned advantages of such approach. The remaining bottle-neck is an enzyme that could produce longer peptides. Here, we report on the characterization of a possible candidate enzyme, Sap9, which may be considered for producing longer, e.g., 3–7 kDa, peptides and lead to a development of extended bottom-up proteomics.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

With the development of increasingly faster and more sensitive high-resolution mass spectrometers, qualitative and quantitative protein analysis is becoming routine in various research, clinical, and industrial laboratories [1–6]. The two main avenues for protein identification with mass spectrometry (MS) are the "bottom-up" and the "top-down" approaches. In bottom-up proteomics, the mixtures of proteins are digested into short (0.6–3 kDa) peptides, which are then analyzed individually with a high-throughput liquid chromatography–tandem mass spectrometry (LC–MS/MS) [1,2,5]. The advantage of this approach lies in the ease of separation, fragmentation and detection of these short peptides. The specificity toward basic residues of trypsin, the most widely employed enzyme in bottom-up proteomics, ensures that more than 95% of proteins from diverse proteomes can be, theoretically, represented by unique peptides [7]. Naturally, this is the most commonly used approach in routine proteomics [8]. However, due to the large sample complexity and wide dynamic concentration range of the proteins oftentimes present in the sample [9,10], it may fail to characterize the entire proteome without rigorous pre-fractionation and multiple technical replicate experiments [10,11]. Orthogonal or restricted cleavage specificities provided by other enzymes show certain advantages for increased proteome coverage with bottom-up proteomics [8]. Nevertheless, the number of enzymatically-produced peptides remains extremely high and the information about the presence of multiple proteoforms (isoforms, point mutations, and post-translational modifications (PTMs)) is often lost [12].

In contrast, top-down proteomics involves fragmentation and analysis of proteins in their intact form directly in the gas phase [13–16]. Since the primary structure is preserved, identification of proteins on the proteoform level is, in principle, possible [17]. However, this technique is far from being routinely implemented for complete proteome analysis due to the difficulties arising from the chromatographic separation of the proteins on the timescale of the experiment. Small (<40 kDa) proteins can be readily separated on C4 reverse phase columns, whereas separation of larger (up to 100 kDa) proteins requires multidimensional separation involving polymeric reverse phase stationary phase (PLRP-S) columns [15] or capillary electrophoresis [18]. Although these emerging separation techniques enable on-line MS/MS analysis of intact proteins, extensive pre-fractionation of the mixture is required prior to analysis [19]. In addition, the MS/MS spectra obtained are highly convoluted due to the large number of product ions present with different charge states. Due to the presence of PTMs and their combinations, proteoform-level characterization requires specialized mass spectra deconvolution tools and database search algorithms [13,20], which require further development.

A third, newly emerging direction is analysis of larger, >3 kDa, peptides, dubbed "middle-down proteomics" [21–24]. This approach entails the chemical or enzymatic digestion of the proteins, much as in the case of bottom-up proteomics; however, the resulting targeted peptides should be in the 3–15 kDa range. We previously proposed the division of this wide mass range due to practical considerations into the 3–7 kDa (extended bottom-up) and 7–15 kDa (middle-down) mass ranges [7].

Recent comprehensive bioinformatics study of the human, yeast and bacterial proteomes revealed that there is no unique cleavage site that allows for whole proteome complete identification based on unique peptides in the 3–7 kDa mass range [7]. In-silico digestion using cleavage rules of currently utilized proteases such as LysC, GluC, AspN, and acid hydrolysis [25] yielded prevalently small (0.6–3 kDa) peptides. Other proteases, such as Kex2 [26] or OmpT [21] have been previously described as dibasic-site specific enzymes capable of yielding large (>3 kDa) peptides. However, bioinformatics studies also showed that ~25% of human proteins remain unidentified with unique peptides with this cleavage rule. Also, the true dibasic-site specificity has not yet been confirmed experimentally for any enzyme. Moreover, proteins such as serum albumin were not identified after digestion with OmpT, likely to the large size (>20 kDa) of the peptides generated [21].

Limiting the enzymatic reaction time for known proteases, e.g., trypsin, to ensure missed cleavage sites has been also previously investigated and is finding renewed interest [27]. A recent study on the trypsin digestion kinetics by Lowenthal and co-workers has determined that tryptic peptides have complex formation kinetics depending on their primary sequence, relative position in the protein tertiary structure and the presence of missed cleavage sites [28]. Long peptides containing missed cleavages were found to be formed slower than short peptides. Other proteases, such as chymotrypsin, pepsin and thermolysin, could potentially find an application for extended bottom-up proteomics. However, digestion reproducibility and substrate specificity of these enzymes at short reaction times remain unclear. Therefore, an optimum protease for either extended bottom-up or middle-down proteomics is yet to be found.

Candida albicans Sap9 is an aspartic protease from the yapsin family and fulfills a major role in maintaining cell wall integrity in fungi. Figure S1 (Supplementary Information) details Sap9 primary structure information [29]. Sap9 has been proposed to cleave peptide backbone primarily after Lys–Arg and Arg–Arg tandem sequences. This specificity has been reported for Sap9 processing of a small number of synthetic, 8–9 amino acid long peptides containing dibasic residues [29], but not whole proteins. In a separate study, Aoki and coworkers used a FRETS-25Xaa library containing 475 peptides, and concluded that Sap9 cleaves preferentially peptides containing basic residues, albeit also cleaves at leucine [30]. More recently, it has been shown that this protease has broad substrate specificity and cleaves several cell wall proteins [31]. Cleavage at select Lys or Arg residues and the importance of neighboring residues on the cleavage position was also noted. Overall, the enzymatic activity of Sap9 has been previously investigated at different pH values but only at 37 °C. Sap9 was found to be active in acidic sodium citrate buffer at pH 3.5–6.0 [31]. Despite these encouraging preliminary results, the proteomic-grade utility of this enzyme has not been established. The effect of pH, temperature, digestion time, and enzyme:protein (E:P) ratio on the protein cleavage specificity (i.e., the amino acid sites targeted under different conditions) has not yet been studied. To summarize, although the general ability of Sap9 to cleave proteins on the cell surface of fungi has been revealed, only the limited cleavage specificity study was carried out using short synthetic model peptides.

Herein we evaluate the applicability of Sap9 to cleave proteins into mid-range peptides, suitable for extended bottom-up proteomics. We first search for the optimal proteolytic conditions of Sap9 using fluorescence-based enzyme activity assay and proteomics experiment-derived enzyme specificity analysis under different experimental conditions (pH, temperature, E:P ratio, and digestion time). In the following, we apply the optimized digestion conditions to evaluate the performance of Sap9-based proteomics for analysis of model protein mixtures with up to 48 proteins.

2. Experimental procedures

2.1. Production of recombinant proteases

Recombinant *C. albicans* His_6-tagged Sap9 was produced using *Pichia pastoris* as an expression system. The sense and antisense primers (5′-ATGCTCGAGAAAAGAGCTAAGGCACCTTTCAAAATC-3′ and 5′-GAATCTAGATTAATGGTGATGGTGATGGTGAGCACCAATGACTTCAATCGA-3′) were used to generate a PCR fragment encoding His_6-tagged Sap9 with genomic DNA of a *C. albicans* clinical isolate as a target. The PCR product was digested with *Xho*I and *Xba*I, and subsequently inserted into the *P. pastoris–Escherichia coli* shuttle vector pPICZαA digested with the same restriction enzymes to generate the expression plasmid pSap9_H-6. *P. pastoris* KM71 (Invitrogen, Carlsbad, CA, USA) was transformed by electroporation with SacI linearized plasmid DNA, and transformants were selected on YPD agar medium (2% (w/v) Difco Bacto Peptone (Difco laboratories, Detroit, MI, USA), 1% (w/v) Difco Bacto yeast extract, 2% (w/v) dextrose, 2% agar) containing 100 μg/mL zeocin.

For enzyme production, *P. pastoris* transformants were grown to near saturation (OD600 = 10) at 30 °C in 1 L of glycerol-based yeast media (0.1 M potassium phosphate buffer at pH 6.0, containing 10 g/L yeast extract, 20 g/L peptone, 13 g/L yeast nitrogen base without amino acids (Becton-Dickinson, Sparks, MD), 10 mL/L glycerol and 40 mg/L biotin). Cells were harvested and resuspended in 200 mL of the same medium with 5 mL/L methanol instead of glycerol and incubated for 48 h. Then, the culture supernatant was harvested after centrifugation (3000 ×*g*, 4 °C, 5 min).

2.2. Purification of heterologously produced Sap9

The secreted proteins from 200 mL of *P. pastoris* culture supernatant were concentrated by ultrafiltration to 6 mL using a Centricon Plus-70 (30 kDa cut-off) (Millipore, Volketswil, Switzerland). The His_6-tagged target protein was extracted with a Ni-NTA resin (Qiagen, Hilden, Germany) column with histidine elution buffer (50 mM histidine in PBS 1×) as previously described [29]. Active fractions were pooled and concentrated by ultrafiltration using Amicon Ultra (Millipore 30 kDa cut-off). Protein concentrations were measured with Nanodrop (Thermo Scientific, Wilmington, DE, USA). The image of the 1D SDS-PAGE gel of the supernatant before and after His_6-tag purification is shown in Figure S2 (Supplementary Information).

2.3. Enzymatic activity assay

The effect of pH, temperature, and E:P ratio on proteolytic activity of *C. albicans* Sap9 was determined using fluorescence-based kinetic assays. The pH-insensitive green-fluorescent BODIPY FL dye-labeled casein was used as a substrate (EnzChek Protease Assay Kit, Molecular Probes, Eugene, OR, USA). Activity assays were performed at 25, 37 and 45 °C, 3.5–6 pH with 0.5 pH unit increment and E:P (w/w) ratio from 1:2.5 to 1:100. Lyophilized casein was reconstituted with 50 mM sodium citrate buffer (pH 6) to a concentration of 1 μg/μL, and further diluted to final concentration of 0.01 μg/μL in assay buffers (pH 3.5–6). 100 μL of substrate solution was placed in each well of 96 well microplate (8 × 12 size Wallac black/clear bottom). Enzyme dilutions were prepared using sodium citrate buffer at different pH, and a solution containing substrate in absence of protease was monitored as reference control set. Measurements initiated immediately upon the addition of reaction buffer in all wells (~10 min from initializing reaction in first well) and were carried out using a Victor X3 multilabel plate reader (Perkin-Elmer, Waltham, MA, USA). Fluorescence was monitored using a 485 nm wavelength excitation, 535 nm emission, and cut-off filter at 580 nm. The sensitivity setting was varied as required, and the fluorescence data was acquired utilizing bottom reading. Each assay was measured with 300 plate reading repeats (up to 12 h, 60 s delay between repeats, shaking for 5 s prior to each reading) in triplicate. Baseline casein fluorescence was found to be pH dependent, therefore all experimental data presented were normalized to the fluorescence recorded without addition of protease at each working condition by subtracting the baseline fluorescence values.

2.4. MS-based enzyme specificity analysis

The substrates employed included a single protein (bovine carbonic anhydrase 2 from Protea Biosciences, Morgantown, WV), a 7 protein mixture (yeast enolase 1 and 2, bovine apotransferrin, serum albumin, pancreatic ribonuclease A, chicken egg white lysozyme (all from Sigma Aldrich, St. Louis, MO, USA)) and bovine carbonic anhydrase 2, the standard equimolar 48 protein mixture (UPS-1, Sigma Aldrich), and the proteomics dynamic range standard set of the same 48 proteins (UPS-2, Sigma Aldrich). Proteins were resuspended in 6.8 M urea, 100 mM ammonium bicarbonate buffer, reduced with 3 molar equivalents of DTT at 50 °C for 1 h, and alkylated with 100 mM iodoacetamide for 30 min at room temperature in dark. The protein mixture was then diluted 20× in 50 mM sodium citrate buffer pH 5.5, and digested with Sap9 in varying E:P ratios with a reaction time between 30 min and 8 h at room temperature. Peptides were desalted using C4 and C18 ZipTip (Millipore, Billerica, MA), the eluents from the two desalting columns were pooled. 5–8 pmol peptide mixture was loaded on a 75 μm ID precolumn (C8, 2 cm long, 100 Å pore size, 5 μm particles) for 10 min at a flow rate of 8 μL/min with 0.1% TFA, and separated using a Dionex Ultimate 3000 nanoLC system fitted with 150 mm C8 Acclaim PepMap300 column with 300 Å pore size, 5 μm size particles, 75 μm ID (Thermo Scientific, Bremen, Germany) at a flow rate of 0.8 μL/min. Compositions of the eluents were A: 0.1% formic acid, B: 50% methanol, 20% acetonitrile, 10% 2,2,2-trifluoroethanol (Alfa Aesar GmbH & Co KG, Karlsruhe, Germany), 0.1% formic acid. This solvent composition has been found by Mitulović and co-workers to minimize carryover and extend the column lifetime without detrimental effect on the separation and analysis of peptides in the LC system employed herein [32]. The percentage of the organic phase was increased from 5 to 60%, the length of the gradient varied between 10 and 60 min, depending on the sample complexity. Eluted peptides were nanoelectrosprayed with 2.4 kV needle potential and analyzed using an Orbitrap Elite ETD FTMS (Thermo Scientific, Bremen, Germany). MS survey scans were acquired at 60,000 resolution at *m/z* 400. Isolation window was set to 3.0 Th, monoisotopic peak selection was disabled. Precursor ions with charge states 1, 2 or 1, 2, and 3 were excluded for MS/MS. Fragmentation of the top 5 peaks was carried out using higher energy collision dissociation (HCD) with normalized collision energy 27%. Alternatively, ion trap collision induced dissociation (CID) of the top 10 peaks was performed using 35 V collision energy. Three microscans were acquired and averaged for each MS/MS spectrum at resolution 15,000 at *m/z* 400 in the Orbitrap for both activation methods. The maximum injection time was set to 200 ms for both MS and MS/MS, and the automatic gain control (AGC) was set to 10^6 charges for MS scan and 5* 10^4 charges for MS/MS, respectively.

In a reference study, the UPS-1 standard was digested overnight with sequencing grade modified trypsin (Promega, Madison, WI, USA) in E:P ratio 1:25 (w/w) following the manufacturer's protocol. Peptides were desalted using C18 ZipTip and separated on a Dionex Ultimate 3000 nanoLC system fitted with 150 mm C18 Acclaim PepMap100 column with 100 Å pore size, 3 μm particles, 75 μm ID (Thermo Scientific, Bremen, Germany) at a flow rate of 0.3 μL/min using the same gradient elution as described above. MS survey scans were performed at 60,000 resolution setting (at *m/z* 400). The top 10 peaks were fragmented using either CID or the top 5 peaks with HCD. A single microscan was recorded for each MS/MS spectrum in Orbitrap FTMS at resolution setting 15,000 (at *m/z* 400).

Peak lists were generated using Proteome Discoverer 1.4, and the MS/MS spectra were analyzed using Sequest. The precursor ion mass tolerance was set to 10 ppm, product ion tolerance to 0.02 Da. For the study of carbonic anhydrase and the 7 protein mixture searches were performed against a database containing 756 proteins, including reviewed sequence homologs of the 7 proteins from all species in the UniProt knowledgebase.

The UPS-1 digestion data was searched against a database containing the non-redundant UniProt human proteins (ver. 2013_02), the 48 UPS-1 proteins and their shuffled sequences (in total, 40,527 entries). Cys carbamidomethylation was set as fixed modification, N-acetylation and oxidation of Met were allowed as dynamic modifications. The cleavage specificity was set to trypsin with two missed cleavages for the trypsin experiments and to no enzyme for Sap9 peptides. To determine the occurrence of Sap9 autolysis, data was additionally searched against a *C. albicans* database containing 1218 reviewed entries. Peptide and protein false discovery rates (FDRs) were determined using Scaffold (Proteome Software, Inc., Portland, OR).

The performance of other search engines such as Mascot has also been investigated. After manual validation, peptide score threshold was set to 15 for HCD and to 20 for CID mass spectra.

3. Results

3.1. Enzyme activity characterization

The effect of the temperature on Sap9 enzyme activity was studied performing the digestion at constant pH and different E:P ratios. Fig. 1 shows the activity curves at pH 4.5 recorded at 25, 37, and 45 °C, at E:P ratios 1:2.5, 1:10, and 1:25. At all E:P ratios the curves show high enzyme activity within the first 2 h of digestion, and a considerable enzyme activity was observed even at the very first time point recorded (~10 min). At 2 h the slope of all curves decreased, but fluorescence response did not reach plateau even at 12 h under the experimental conditions employed (data not shown). The highest activity was obtained at E:P ratio 1:2.5 at 37 °C (red line), the fluorescence recorded at 2 h was ~4 fold higher than at 1:10 ratio, and a ten-fold higher compared to 1:25 ratio. This ratio was optimal at all temperatures, although activity at 2 h and 45 °C was 2.5 times lower than at 37 °C. Another two-fold decrease in activity was recorded at 25 °C. Based on the enzyme activity assay, we concluded that Sap9 is most aggressive at pH 4.5, 37 °C.

The effect of pH and E:P ratio on Sap9 enzyme activity was studied performing the digestion at selected reaction temperatures and time points. Fig. 2 shows the Sap9 enzyme activity

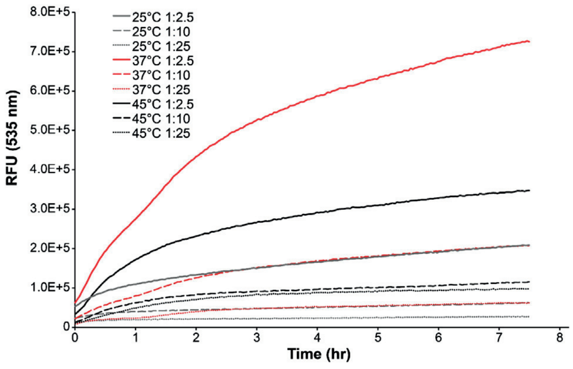


Fig. 1 – Effect of temperature on enzymatic activity of Sap9 protease at pH 4.5, E:P ratios 1:2.5, 1:10, and 1:25. The substrate is green fluorescent β-casein.

within one standard deviation obtained at 25, 37, and 45 °C at different pH values and E:P ratios achieved at 4 h reaction time with the green fluorescent casein substrate. The complete activity curves recorded demonstrate that at this time point the slope of all curves decreased, indicating that the enzyme is saturated with substrate (Figures S3, S4, and S5 in Supplementary Information). As expected, enzyme activity increased with increasing the relative enzyme quantity. At optimal pH conditions (~4.5) the protease maintained its activity even at very low (1:100) E:P ratios. At all temperatures studied, the protease maintained its activity in the entire pH range studied with the highest fluorescence values recorded at pH values between 4 and 5.

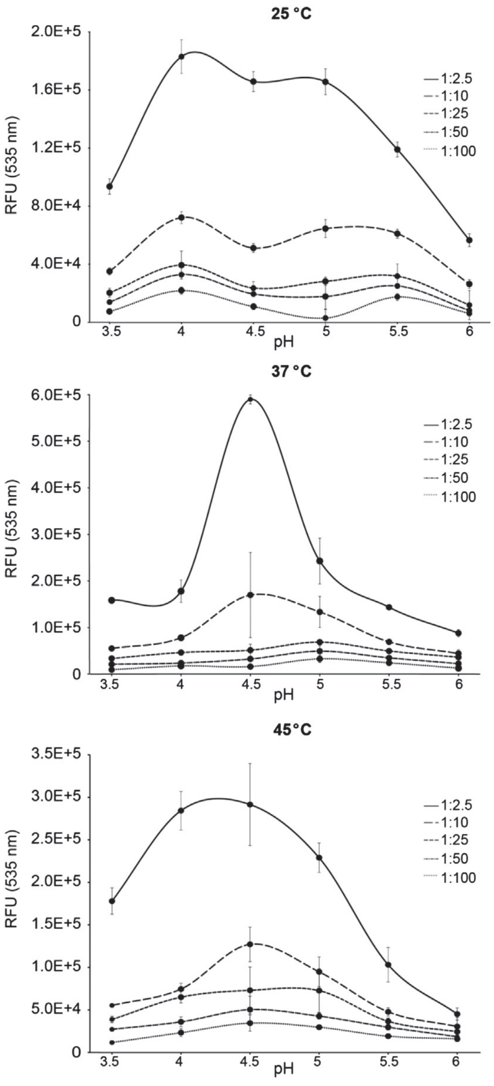


Fig. 2 – Sap9 enzymatic activity recorded at 4 h time point at 25, 37, and 45 °C and pH values between 3.5 and 6. Error bars indicate one standard deviation of the mean from the triplicate measurement. The substrate is green fluorescent β-casein.

3.2. Sap9-proteolysis of carbonic anhydrase

To determine which temperature yields optimal number of peptides for extended bottom-up proteomics, we tested the dependence between size and charge distribution of peptides obtained when proteolysis of a single protein is performed at different temperatures. Carbonic anhydrase 2 (~29 kDa) was selected as a first substrate because it contains an RR and an RKK sites positioned within the first 120 amino acid residues from the N terminus, yielding two peptides in the targeted 3–7 kDa mass range. The E:P ratio 1:2.5 and pH 4.5 were chosen as starting conditions because they provide the highest proteolytic activity, Figs. 1 and 2. LC–MS/MS experiments on carbonic anhydrase digests showed that at this E:P ratio temperature does not noticeably affect average peptide lengths at a given pH. At 1 h, the average size of proteolytically-derived peptides obtained for this protein was 2.8–3.0 kDa at 15, 25, and 45 °C. In contrast, at 37 °C the average size of peptides was 2.4 kDa, indicating a more aggressive enzymatic activity (Table S1, Supplementary Information). At the database search filtering parameters employed, Sequest was more suitable for identification of longer peptides than Mascot.

100% sequence coverage was obtained at the shortest time point sampled (1 h) at all temperatures, except at 37 °C, and even after 8 h digestion at temperatures above or below 37 °C. Presumably, at around 37 °C Sap9 digested the protein into short peptides that were not retained on the C8 LC column. Furthermore, considering the increasing number of short peptides detected at longer sampling times, it is apparent that the protease continues to further digest the long peptides. Therefore, a short digestion time is, perhaps, more suitable for Sap9 proteolysis-based extended bottom-up MS. We have found Sap9 autolysis products at 37 and 45 °C, E:P ratio 1:2.5, but not with lower relative amount of enzyme and/or lower temperatures.

3.3. Sap9-proteolysis of a seven protein mixture

We proceeded with characterization of Sap9 for the digestion of a mixture containing seven proteins of different sizes and tertiary structures (see Experimental section). Table S2 (Supporting Information) shows the resulting size and charge state information of Sap9-derived proteolytic peptides identified, as well as the average sequence coverage of the considered proteins. At pH 4.5 Sap9 is very active at all temperatures, and comparable average sequence coverage was obtained at all time-points between 1 and 4 h, regardless of the experimental conditions. Based on the short average peptide sizes obtained as soon as 1 h after digestion was started, we concluded that, naturally, either the E:P ratio of 1:2.5 was too high, or the digestion time was too long. Therefore, to establish the effect of the pH, the reaction time was limited to 30 min, and a lower E:P ratio of 1:10 was employed. Temperature of 25 °C was chosen for following experiments as the practically most convenient for experimental setup.

Table 1 shows the effect of pH on the size of peptides obtained by 30 minute Sap9 digestion of the seven protein mixture at 25 °C in 1:10 E:P ratio. Based on this study, we established that at these experimental conditions the protease was active at all pH values and average protein sequence coverage of 40–56% was obtained. pH 5–5.5 was found to be the most optimal for extended bottom-up approach, since the highest average peptide length and charge state were obtained at these conditions. In addition, similarly to a single protein study, we have found that Sequest was, in general, more suitable for identification of long peptides than Mascot, as the latter failed to identify peptides with charge states >7+.

An important consideration in establishing the utility of Sap9 for MS-based proteomics was the study of the digestion reproducibility by performing parallel digestions and comparing the extracted ion chromatograms of the LC–MS experiments. The resulting peptides were identical in three replicates performed, albeit, in some cases, with slightly varying relative intensities, Figure S6 (Supplementary Information). Whether the change in the relative abundance of some peptides is a result of parallel or competing digestion reactions or an artifact arising from spraying conditions remains to be determined.

3.4. *Sap9-proteolysis of a 48 protein mixture*

Digestion conditions where the Sap9 enzyme is less aggressive (25 °C, pH 5.5, 1:10 E:P ratio, 1 hour digestion) were applied to the equimolar mixture of the commercial 48 protein standard UPS-1. These conditions were determined to be those at which longer than tryptic peptides were observed for the 7 protein mixture.

The extracted ion chromatogram of the UPS-1 peptides produced after 1 hour digestion with Sap9 is shown to demonstrate the performance of the method in Fig. 3. As expected, shorter, less hydrophobic peptides were eluted at the first minutes of the gradient, whereas with increasing the organic component, longer, more highly charged peptides were separated. This aspect of peptide separation, although predictable, is particularly important for scheduling data dependent analysis of peptides with broad charge state distribution using an Orbitrap mass analyzer. To optimize the number of peptides identified throughout LC separation, two or more activation methods could be employed, monoisotopic peak selection can be enabled/disabled, and different minimum signal threshold, number of microscan events, or injection time could be set for the different stages of the gradient, depending on the length and charge state of eluting peptide ions [33,34].

Table 1 – Effect of pH on Sap9 activity: size and charge distribution of proteolytic peptides from seven protein mixture after 30 minute Sap9 digestion obtained by Sequest and Mascot database search engines.

pH	Average		Maximum		Coverage (%)	Total # peptides
	MW, kDa	Charge	MW, kDa	Charge		
Sequest						
4	2.6	4	6.6	8	55	156
4.5	2.5	3.7	5.5	8	52.1	193
5	2.7	3.9	7.1	9	56.5	196
5.5	2.8	4	7.1	9	56.3	203
6	2.9	4.3	7.1	10	40.6	149
Mascot						
4	2.3	3.3	5.1	6	52.1	165
4.5	2.2	3.3	4.2	7	47.1	155
5	2.2	3.3	4.2	7	46.2	145
5.5	2.4	3.4	5	7	48.9	145
6	2.2	3.3	4.2	5	31.4	97

As exemplified in Fig. 4, HCD MS/MS produced extensive sequence coverage of large proteolytic peptides on the LC time scale. As shown in the insets of Figs. 3 and 4, high resolution is necessary for the accurate charge state determination of the precursor ions, as well as that of the resulting product ions in MS/MS. CID yielded similar fragmentation mass spectra, although with more prevalence of the *b* ion series, as expected [35].

A comparison between the number of proteins and peptides identified from the 1 hour Sap9 digestion of the protein mixture using different data dependent MS/MS strategies is presented in Table 2. Precursor ion activation with CID and HCD was performed including or excluding peptide precursor ions of charge state 3+. Precursor ion charge states 1+ and 2+ were excluded in all Sap9 experiments. Also included is the summary of a typical bottom-up proteomics experiment on an overnight tryptic digest, fragmented using data dependent top 10 CID in LTQ excluding charge state 1+, followed by ion detection in Orbitrap FTMS. Using these conditions, the average peptide mass identified from Sap9 was between 2.7 and 3.3 kDa, depending on the activation method, whereas with trypsin we identified peptides with average mass of 1.9 kDa.

The peptides generated by 1 hour Sap9 digestion and fragmented using CID MS/MS (charge states >3+) resulted in identification of 41 proteins with at least 2 peptides (and of 46 proteins with 1 peptide) at FDR level 1%. Of these, 39 proteins (43 with 1 peptide, respectively) were from the listed UPS-1 standard, as detailed in Table 2. In addition, tetranectin (E9PHK0) and malate dehydrogenase (P40926) were also identified. With the same filtering criteria, HCD MS/MS yielded the identification of 42 proteins with >2 peptides, and 50 proteins with at least 1 peptide, of which 46 were UPS-1 standard proteins. When peptides with 3+ charge state were included in the data dependent analysis, CID MS/MS led to identification of 46 proteins (2 peptides rule), including 44 UPS-1 proteins, malate dehydrogenase, and tetranectin. Inclusion of the 3+ charge state peptides increased the sequence coverage of the proteins identified previously, and the average sequence coverage of the UPS-1 proteins was 60% using CID and 59% with HCD activation, with an average of ~12 peptides/protein.

The only protein from the UPS-1 protein mixture formulation that was not identified after digestion with Sap9 was C-reactive protein P02741. This protein did not produce peptides in the targeted mass and charge range, most likely due to the presence of 5 dibasic sites in its sequence. Furthermore, we have identified keratin in the Sap9-digested UPS-1 mixture, with a single peptide of 3+ charge state fragmented with HCD. This observation indicates that perhaps this protein family is very quickly degraded by Sap9 and is in line with the physiological activity of this protease, Sap9 is used by the fungus for host cell adhesion and causes epithelial cell damage [29,30].

In comparison, at protein FDR setting of 1%, 64 proteins containing 45 UPS-1 proteins were identified with at least 2 peptides by LC–MS/MS of UPS-1 sample subjected to a standard overnight trypsin proteolysis, Table 2. In addition to the ambiguous assignment to hemoglobin delta chain instead

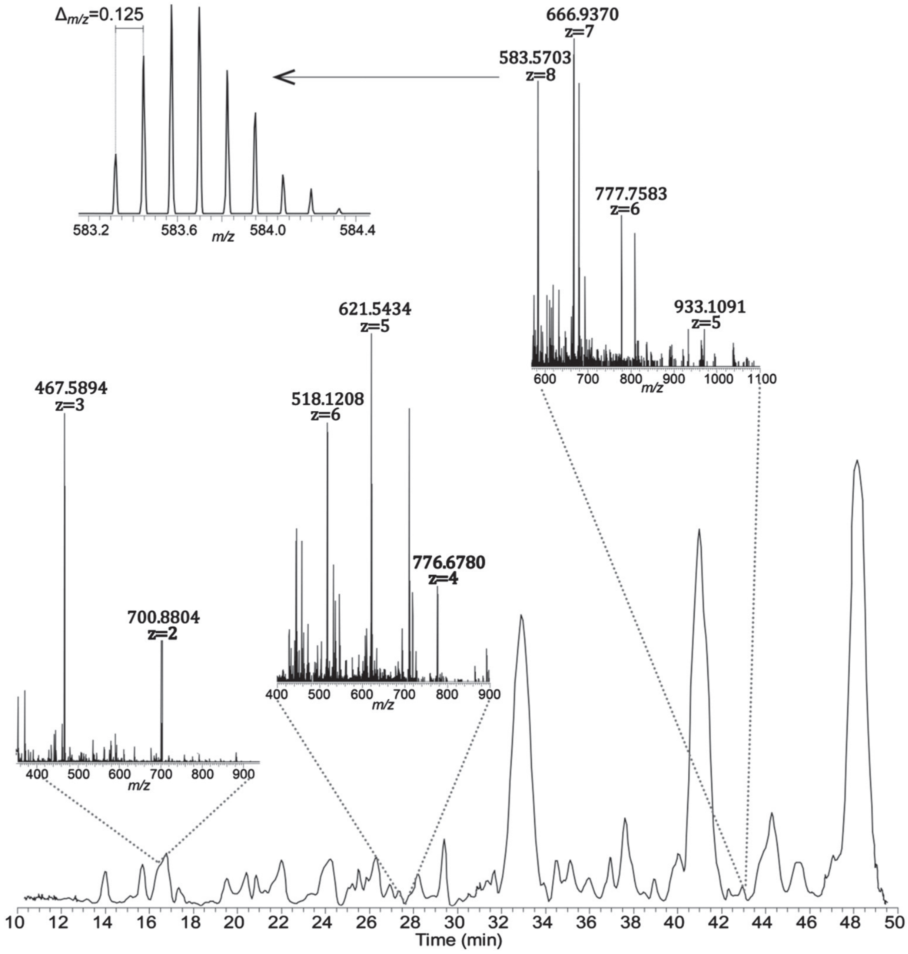


Fig. 3 – Extracted ion chromatogram of the UPS-1 peptides obtained after 1 h digestion with Sap9. The insets show examples of the *m/z* distribution of peptides of varying size in the LC gradient. 15 cm C8 column, 300 Å, 5 μm, 70 min gradient MeOH:ACN: TFE:0.1%FA (5:3:1:1), solvent A 0.1% FA (8 pmol on column).

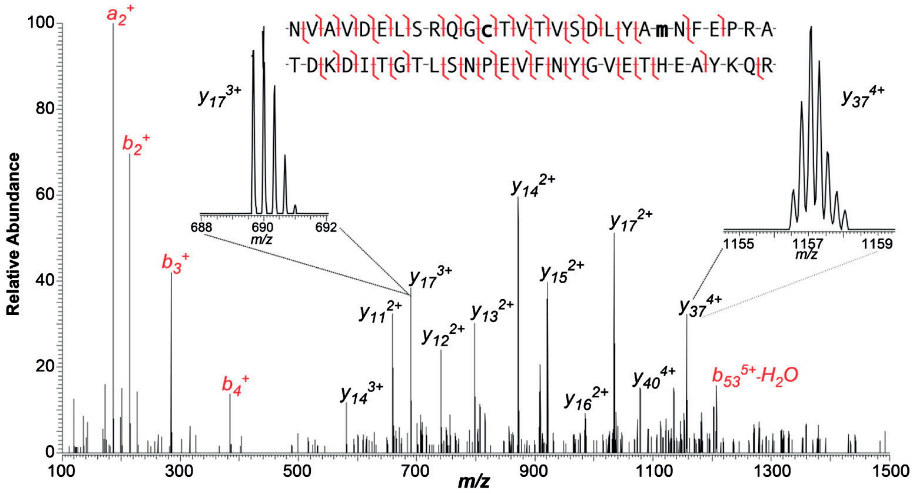


Fig. 4 – Orbitrap FTMS HCD mass spectrum of a 7+ precursor peptide ion from UPS-1 component ribosyldihydronicotinamide dehydrogenase acquired on the LC timescale.

Table 2 – Sequence coverage of UPS-1 proteins after 1 h Sap9 and overnight trypsin digestion obtained by Sequest at protein FDR 1%, peptide XCorr score thresholds 2.5 (2+), 3.5 (3+), and 3.8 (≥4+). Data dependent MS/MS fragmentation was triggered at signal threshold 15,000. Peptides were fragmented using CID and HCD MS/MS and precursor ions of charge state 3+ were either included (>2+) or excluded (>3+) from the data dependent ion selection. Searches were performed against a database containing human proteins, UPS formulation proteins and their shuffled sequences. Sap9 data was searched using no-enzyme setting and trypsin data was searched using trypsin with 2 missed cleavages.

Protein	Accession #	MW, kDa	Trypsin overnight		Sap9 1 h							
			CID		>3+ CID		>3+ HCD		>2+ CID		>2+ HCD	
			#pept.	Seq. cov.	#pept.	Seq. cov.	#pept.	Seq. cov.	#pept.	Seq. cov.	#pept.	Seq. cov.
Alpha-lactalbumin	P00709	16	5	38%	6	79%	5	42%	10	100%	8	63%
Annexin A 5	P08758	36	22	74%	2	13%	2	13%	4	26%	3	13%
Antithrombin-III	P01008	53	30	71%	6	28%	9	36%	16	42%	17	51%
BH3 interacting domain death agonist 1	P55957	22	15	70%	1	10%	2	17%	4	38%	6	55%
Beta-2-microglobulin	P61769	14	5	49%	4	77%	5	77%	7	77%	8	77%
C-reactive protein	P02741	25	7	27%								
Carbonic anhydrase 1	P00915	29	15	73%	2	35%	3	37%	3	41%	2	28%
Carbonic anhydrase 2	P00918	29	17	71%	3	11%	4	20%	6	26%	6	26%
Catalase OS = *Homo sapiens*	P04040	60	46	71%	14	38%	16	42%	18	48%	25	56%
Complement C5	P01031	188	1	18%	13	66%	4	76%	3	79%	7	86%
Creatine kinase M-type	P06732	43	25	67%	4	54%	14	44%	20	48%	22	54%
Cytochrome b_5	P00167	15	5	41%	13		1	16%			2	22%
Cytochrome c	P99999	12	7	46%	5	68%	10	100%	7	68%	12	99%
Fatty acid-binding protein, heart	P05413	15	7	51%	12	100%	12	100%	15	100%	19	100%
Gamma-synuclein	O76070	13	13	82%	12	99%	12	99%	15	99%	16	99%
Gelsolin	P06396	86	29	48%	17	57%	15	44%	21	49%	29	57%
Glutathione S-transferase A1	P08263	26	7	30%	6	23%	6	21%	11	43%	9	28%
Glutathione S-transferase P	P09211	23	13	71%			1	13%			2	22%
GTPase HRas [Chain 1189]	P01112	21	12	62%	6	71%	7	62%	10	62%	10	62%
Hemoglobin subunit alpha	P69905	15	9	84%	5	80%	7	92%	11	92%	9	88%
Hemoglobin subunit beta	P68871	16	11	84%	5	77%	5	77%	8	77%	9	85%
Histidyl-tRNA ligase, cytoplasmic	P12081	57	32	64%	6	20%	7	16%	18	35%	13	32%
Insulin-like growth factor II	P01343	20	4	76%	5	76%	7	28%	10	37%	11	100%
Interferon gamma	P01579	19	11	50%	9	87%	10	75%	12	87%	13	87%
Interleukin-8	P10145	8.4			7	74%	10	85%	13	96%	14	96%
Lactotransferrin	P02788	78	56	72%	21	36%	25	46%	36	58%	37	56%
Leptin	P41159	19	6	41%	5	53%	5	53%	10	65%	12	65%
Lysozyme C	P61626	17	4	45%	3	65%	2	25%	7	64%	4	27%
Microtubule-associated protein tau	P10636	79	23	60%	28	52%	34	52%	31	52%	37	52%
Myoglobin	P02144	17	18	92%	6	92%	8	71%	9	92%	11	92%
NAD(PH dehydrogenase) [quinone] 1	P15559	31	16	49%	4	34%	6	41%	6	29%	8	40%
NEDD8	Q15843	9	3	41%	7	54%	6	94%	9	94%	7	94%
Peptidyl-prolyl cis-trans isomerase A	P62937	18	12	77%	3	46%	3	46%	6	64%	6	46%
Peroxiredoxin 1	Q06830	22	16	68%	2	13%	3	13%	3	13%	4	18%
Platelet-derived growth factor subunit B	P01127	27			6	26%	6	69%	11	99%	18	100%
Ubiquitin	P62988	26	3	43%	1	49%	1	49%			1	49%
Pro-epidermal growth factor	P01133	134			1	87%	1	87%	4	83%	4	87%
Retinol-binding protein 4	P02753	23	9	72%	10	62%	10	61%	13	56%	14	65%
Ribosyldihydronicotinamide dehydrogenase [quinone]	P16083	26	14	73%	4	34%	9	50%	7	34%	9	50%
Serotransferrin	P02787	77	54	73,00%	32	62%	47	66%	56	73%	67	78%
Serum Albumin	P02768	69	32	66%	27	69%	29	75%	31	78%	34	74%
Small ubiquitin-related modifier 1	P63165	39	23	57%	5	32%	8	39%	14	53%	15	58%
SUMO-conjugating enzyme UBC9	P63279	18	7	58%	4	78%	4	60%	7	94%	10	94%
Superoxide dismutase [Cu-Zn]	P00441	16	8	66%	4	62%	4	40%	4	48%	6	56%
Thioredoxin	P10599	12	6	63%					3	40%	4	43%
Tumor necrosis factor [TNF-alpha]	P01375	26	8	54%	2	36%	2	24%	3	30%	3	16%
Ubiquitin-conjugating enzyme E2 C V = 1	O00762	20	13	81%			1	11%	6	39%	4	39%
Ubiquitin-conjugating enzyme E2 E1	P51965	21	3	16%	1	7%	2	17%	2	20%	3	29%
Total Number of peptides identified			681		326		390		520		590	
Average # peptides/protein			15.1		7.6		8.5		11.8		12.6	
Average peptide mass, kDa			1.9		3.3		3.0		2.9		2.7	
Average peptide charge			2.7+		4.6+		4.6+		4.0+		4.0+	
Average sequence coverage			60%		54%		51%		60%		59%	

(continued on next page)

Table 2 (*continued*)

Protein	Accession #	MW, kDa	Trypsin overnight		Sap9 1 h							
			CID		>3+ CID		>3+ HCD		>2+ CID		>2+ HCD	
			#pept.	Seq. cov.	#pept.	Seq. cov.	#pept.	Seq. cov.	#pept.	Seq. cov.	#pept.	Seq. cov.
% spectra identified			21%		34%		46%		42%		55%	
Total Number of UPS-1 proteins/total proteins with 2 peptides			44/61		39/41		40/42		44/46		46/48	
Total Number of UPS-1 proteins/total proteins with 1 peptide			45/64		43/46		46/50		44/46		47/54	

of hemoglobin beta, 6 keratins, 2 IgG chains, as well as tetranectin, apolipoprotein A-I, and malate dehydrogenase were also identified. Similarly to the work of Mann and others, interleukin-8 (P10145), platelet-derived growth factor B (P01127), and pro-epidermal growth factor (P01133) belonging to 48 protein standard were not identified with trypsin [36]. Interestingly, these proteins were detected using Sap9 in all data acquisition modes.

An important aim of the present study was to establish the specificity and determine the preferential digestion sites of the Sap9 protease. The peptides identified from the 1 hour UPS-1 digest were aligned to include the four amino acids prior and the four amino acids following the digestion site, respectively. Duplicate sequences were removed, and the frequency of each amino acid in the resulting P4–P4′ positions was plotted using iceLogo against the non-redundant human UniProt database, Fig. 5 [37]. Here, the *x* axis represents the amino acid position between P4 and P4′ (with cleavage occurring between P1 and P1′), and the *y* axis shows the frequency of each amino acid expressed as percent difference between the experimental dataset and the UniProt database. As a result, amino acids present in the peptides identified more frequently than in the human proteome set have positive *y* values, whereas those underrepresented have negative *y* values. It is therefore apparent that Sap9 cleavage occurs more likely in the vicinity of basic residues, Lys and Arg, and it is more likely to occur before acidic amino acids, such as Asp and Glu. Sap9 digestion is not likely in the vicinity of Pro, and, although possible, cleavage at hydrophobic residues such as Ile, Val or Leu is less prominent than in the vicinity of basic residues.

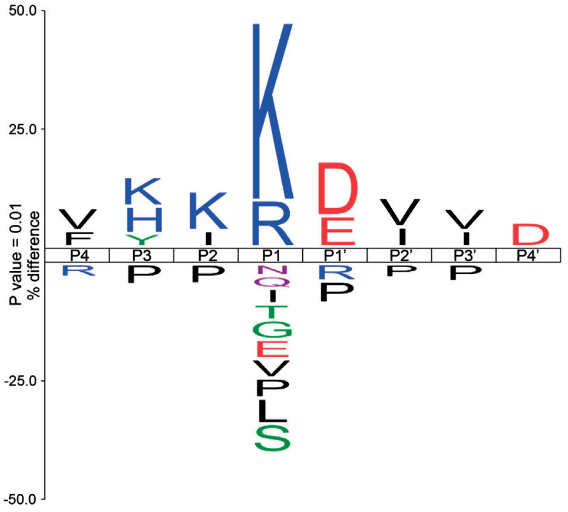


Fig. 5 – IceLogo representation of Sap9 cleavage specificity determined based on the peptides generated from the 48 proteins present in the UPS-1 standard mixture.

In agreement with a typical bottom-up proteomics experiment [11], only 21% of the 10,095 CID MS/MS mass spectra acquired for the tryptic digest resulted in peptide identification. In contrast, 42% of the 3848 CID MS/MS mass spectra of the 1 h Sap9 digest were assigned. This two-fold improvement in spectral identifications illustrates the utility of long peptides for unambiguous peptide and protein assignment, the decreased likelihood for hits against a shuffled, non-existing sequence, as well as the ability of the existing data acquisition and analysis software to handle peptides in this size and charge range. Importantly, when charge state 3+ was excluded, Sap9 identified a similar number of proteins with 2 times lower number of tandem mass spectra compared to trypsin (326 identified peptides versus 681 for Sap9 and trypsin, respectively).

Lowering the data dependent acquisition precursor ion intensity threshold from 15,000 to 5000 greatly increased the number of tandem mass spectra triggered, and enhanced the number of peptides identified, whereas the sequence coverage did not improve significantly (Table S3, Supplementary Information). Increasing the minimum signal threshold inherently increases the number of precursor ions isolated and enhances the quality of the tandem mass spectra [33]. This is especially important for long peptides with high charge states, as it is the case of those produced by Sap9 digestion.

The peptide size distributions obtained after 30 min or 1 hour Sap9 digestion in comparison to those observed with trypsin were extracted using RawMeat 2.1 (VAST Scientific, Cambridge, MA) and are shown in Fig. 6. Note, the y-axis values in the top and middle panels are the total number of peptides that were selected for fragmentation, not only those that were identified. As it can be seen from Fig. 6, top panel, the peptides generated with Sap9 after 30 minute digestion (average 4.1 kDa) were considerably longer than after 1 hour digestion (average 3.5 kDa). However, with the LC–MS/MS conditions applied not all proteins could be identified. Many of the peptides produced with 30 min Sap9 digestion were not reliably identified using the fragmentation methods, data acquisition strategies, and interpretation algorithms employed herein. Therefore, the results from this set of experiments were excluded from further discussion. In comparison with Sap9 results, tryptic peptides were, on average, considerably shorter (~1.9 kDa), Fig. 6 middle panel. Finally, Fig. 6 bottom panel shows the theoretical peptide size distributions for the 48

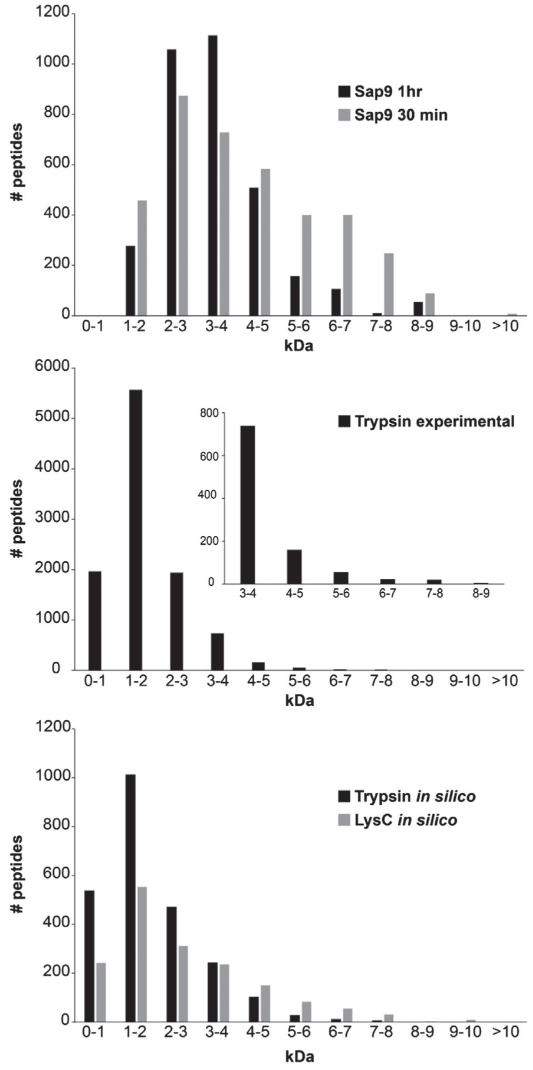


Fig. 6 – Peptide size distribution of the 48 protein mixture UPS-1 obtained with a) 30 minute or 1 hour Sap9 digestion, b) overnight trypsin digestion (the inset shows an expanded view of heavy peptide distribution), and c) theoretical size distribution of peptides obtained with trypsin (one missed cleavage allowed) and LysC.

proteins digested with trypsin (1 missed cleavage allowed) and LysC. Both enzymes would yield shorter peptides than Sap9 digestion.

Table S4 (Supplementary Information) shows the performance of Sap9 in the identification of proteins from the UPS-2 dynamic range standard. Both trypsin and Sap9 have identified proteins in the 50 fmol–50 pmol range (1 fmol–1 pmol on the LC column) with comparable average sequence coverage of 56% for trypsin and 54% for Sap9. Sap9 identified only 3 proteins in the lowest concentration, whereas trypsin identified 7, albeit trypsin also identified several contaminant proteins and some that were not in the UPS-2 formulation (such as ATP-binding cassette sub-family D member 2 Q9UBJ2).

4. Discussion

The characterization of Sap9 protease is a part of our effort for benchmarking novel digestion procedures for production of the mid-size (3–7 kDa) peptides for extended bottom-up proteomics. Sap9 has been found to be active in the pH range 3.5–6, temperature range 25–45 °C, and at E:P ratios (w/w) down to 1:100. As expected, enzyme activity increased with increasing the enzyme quantity, whereas at optimal pH conditions the protease maintained its activity even at very low E:P ratios. At 37 °C, the highest fluorescence was recorded at pH 4.5. However, one important observation to make is that optimal conditions for most aggressive enzyme activity (37 °C, pH 4.5, E:P 1:2.5) were not the most suitable for obtaining peptides in the 3–7 kDa range. Under experimental settings were the enzyme is most active, the peptides detected were on average shorter, and lower protein sequence coverage values were observed with the employed extended bottom-up proteomics conditions. In contrast, by performing the digestion at room temperature, at pH 5.5, and E:P ratio of 1:10 (w/w), the enzyme activity was reduced and longer peptides were detected, leading to increased sequence coverage. At room temperature and pH 5.5, a 30 minute incubation yielded peptides from approximately 50% of the proteins in the UPS-1 mixture. The substrate specificity and ability to digest a variety of proteins using shorter digestion times (less than 1 h) at pH 4.5, 37 °C remains to be investigated.

The MS/MS-based study of peptides generated from the 48 protein mixture revealed that Sap9 cleaves at adjacent basic residues or in their vicinity. It most likely cleaves when a basic or a hydrophobic residue is followed by an acidic amino acid. This restricted but not exclusive cleavage specificity is important for ensuring that a variety of proteins can be processed within a short incubation time, regardless of their primary sequence. As a result, proteins that are rich in basic residues and cannot be identified with trypsin or even LysC (such as a large number of ribosomal proteins and interleukin-8 (P10145), platelet-derived growth factor B (P01127), and pro-epidermal growth factor (P01133) investigated in this study) can be readily studied using Sap9 digestion. Therefore, Sap9 could be used as a complementary protease for identification of proteins that could not be detected using other benchmarked proteases due to their primary amino acid sequence.

When DTT was added in 1:3 molar ratio prior to digestion, no Sap9 autolysis products were observed. However, when higher concentration of DTT was used for reduction of disulfide bonds (i.e., 5 mM), several Sap9 peptides were observed, indicating autolysis. This is most likely caused by the reduction of the disulfide bond linking the two Sap9 subunits by the excess DTT remaining in the digestion solution.

Importantly, state-of-the art mass spectrometers, e.g., time-of-flight (TOF) MS or Orbitrap FTMS, are capable to efficiently analyze 3–7 kDa peptides without any hardware modification [18,38–40]. MS/MS analysis of a complex protein mixture digested with Sap9 shows that the size distribution of the peptides generated by Sap9 is suitable for separation on a chromatographic column. The single change in the workflow of currently employed bottom-up proteomics technique is the requirement for the C8 stationary phase, instead of the more commonly utilized C18. In addition, increased size and charge of peptides in extended bottom-up proteomics may be advantageous for MS/MS performance, specifically for electron transfer dissociation (ETD) [41,42]. In general, we identified slightly more proteins and peptides using HCD compared to CID. This is likely due to the fact that HCD activation energy is calculated for each precursor ion in part and it is based on the initial energy setting normalized to the mass and charge of the selected precursor ion. Including peptides with charge state 3+ as precursor ions has the benefit of increasing slightly the number of proteins identified, albeit it does not significantly improve the sequence coverage of proteins already identified.

Data analysis software already employed in conventional bottom-up proteomics can be employed for data interpretation, although the search algorithms Sequest and Mascot used herein could be further improved to take into account the high mass resolution and mass accuracy of the product ions [43]. Therefore, with minor adjustments to the data acquisition strategy and to the chromatographic separation, extended bottom-up proteomics is straightforward and can be readily implemented on high resolution MS/MS platforms.

The advantages of using the Sap9 protease versus the more commonly employed enzymes, e.g., trypsin, are the shorter digestion time (1 h vs. overnight) and higher protein sequence coverage obtainable from a single peptide analysis. This, in turn, may aid in proteoform-level protein identification by enabling localization of point mutations, deletions, insertions and/or consecutive PTMs from a single peptide. In addition to the peptides identified using the data acquisition and analysis workflow described herein, a large number of highly charged species (>10+) have been observed. These, in most cases, did not yield good quality MS/MS spectra and were therefore discarded. Optimization of the data dependent fragmentation and acquisition parameters, such as number of charges (AGC setting), maximum injection time, as well as of the physical instrument operation parameters, such as gas pressure in the transfer region between linear ion trap and Orbitrap, may further improve the identification of the long peptides generated by Sap9.

Another important advantage of Sap9 for proteome analysis arises from its high activity in acidic environment, under conditions where conventional proteases (i.e., trypsin) are inactive. In acidic pH thiol-disulfide exchange reaction is inhibited, therefore no disulfide bond rearrangements can occur [44]. Also, at basic pH values oxidation by ambient oxygen is more likely to occur than at acidic conditions [45]. This aspect is important, for example, in the proteomic analysis of therapeutic antibody samples [24,46].

The main disadvantage of using a non-specific protease for high throughput proteomics is that multiple peptides with partially overlapping sequences might be generated, increasing the complexity of the mixture. This phenomenon is increasingly severe at longer digestion times and at digestion conditions approaching the optimal enzyme activity conditions. However, in certain application areas, the presence of multiple peptides with varying lengths from the same sequence region might also be beneficial. For instance, we have successfully employed Sap9 for characterization of monoclonal antibodies, where the long peptides with partially overlapping sequences aid in determination of the variable regions and in chain identification [47]. Similarly, Sap9 could be utilized for the study of purified protein extracts, such as immunoprecipitated samples, where the aim of the study is proteoform-level characterization.

It can be envisaged that the specificity of Sap9 (or that of other non-specific enzymes) could be tailored by increasing the rigidity of the protein structure using protein engineering. A reduced specificity aimed towards exclusively dibasic and proximal basic residues would reduce the number of peptides generated. This, in turn, would further add to the utility of the enzyme in high-throughput proteomics.

Transparency document

The http://dx.doi.org/10.1016/j.jprot.2014.07.035 associated with this article can be found, in the online version.

Acknowledgments

We thank Luca Fornelli and Daniel Ayoub for the discussion of the data analysis, Mark Ivanov and Mikhail Gorshkov for assistance with the database search, and Karine Salamin for Sap9 production. We thank Johan Auwerx and Norman Moullan for the access to the fluorometer. The work was supported by the Swiss National Science Foundation (Project 200021-147006) and the European Research Council (ERC Starting grant 280271 to YOT).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jprot.2014.07.035.

REFERENCES

[1] Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature 2003;422:198–207.
[2] Han X, Aslanian A, Yates 3rd JR. Mass spectrometry for proteomics. Curr Opin Chem Biol 2008;12:483–90.
[3] Liu T, Belov ME, Jaitly N, Qian W-J, Smith RD. Accurate mass measurements in proteomics. Chem Rev 2007;107:3621–53.
[4] Mann M, Kelleher NL. Precision proteomics: the case for high resolution and high mass accuracy. Proc Natl Acad Sci 2008; 105:18132–8.
[5] Sabido E, Selevsek N, Aebersold R. Mass spectrometry-based proteomics for systems biology. Curr Opin Biotechnol 2012;23: 591–7.
[6] Zubarev RA, Makarov A. Orbitrap mass spectrometry. Anal Chem 2013;85:5288–96.
[7] Laskay ÜA, Lobas AA, Srzentic K, Gorshkov MV, Tsybin YO. Proteome digestion specificity analysis for rational design of extended bottom-up and middle-down proteomics experiments. J Proteome Res 2013;12:5558–69.
[8] Meyer JG, Kim S, Maltby DA, Ghassemian M, Bandeira N, Komives EA. Expanding proteome coverage with orthogonal-specificity α-lytic proteases. Mol Cell Proteomics 2014;13:823–35.
[9] Corthals GL, Wasinger VC, Hochstrasser DF, Sanchez JC. The dynamic range of protein expression: a challenge for proteomic research. Electrophoresis 2000;21:1104–15.

[10] Zubarev RA. The challenge of the proteome dynamic range and its implications for in-depth proteomics. Proteomics 2013;13:723–6.

[11] Michalski A, Cox J, Mann M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC–MS/MS. J Proteome Res 2011;10:1785–93.

[12] Smith LM, Kelleher NL, Top Down Proteomics C. Proteoform: a single term describing protein complexity. Nat Methods 2013;10:186–7.

[13] Ahlf DR, Thomas PM, Kelleher NL. Developing top down proteomics to maximize proteome and sequence coverage from cells and tissues. Curr Opin Chem Biol 2013;17:787–94.

[14] Tsybin YO, Fornelli L, Stoermer C, Luebeck M, Parra J, Nallet S, Wurm FM, Hartmer R. Structural analysis of intact monoclonal antibodies by electron transfer dissociation mass spectrometry. Anal Chem 2011;83:8919–27.

[15] Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, Durbin KR, Tipton JD, Vellaichamy A, Kellie JF, Li MX, Wu C, Sweet SMM, Early BP, Siuti N, LeDuc RD, Compton PD, Thomas PM, Kelleher NL. Mapping intact protein isoforms in discovery mode using top-down proteomics. Nature 2011;480:254–8.

[16] Liu X, Sirotkin Y, Shen Y, Anderson G, Tsai YS, Ting YS, Goodlett DR, Smith RD, Bafna V, Pevzner PA. Protein identification using top-down. Mol Cell Proteomics 2012; 11(M111):008524.

[17] Tian Z, Tolic N, Zhao R, Moore RJ, Hengel SM, Robinson EW, Stenoien DL, Wu S, Smith RD, Pasa-Tolic L. Enhanced top-down characterization of histone post-translational modifications. Genome Biol 2012;13:R86.

[18] Taichrib A, Pelzing M, Pellegrino C, Rossi M, Neususs C. High resolution TOF MS coupled to CE for the analysis of isotopically resolved intact proteins. J Proteomics 2011;74: 958–66.

[19] Doucette AA, Tran JC, Wall MJ, Fitzsimmons S. Intact proteome fractionation strategies compatible with mass spectrometry. Expert Rev Proteomics 2011;8:787–800.

[20] Compton PD, Zamdborg L, Thomas PM, Kelleher NL. On the scalability and requirements of whole protein mass spectrometry. Anal Chem 2011;83:6868–74.

[21] Wu C, Tran JC, Zamdborg L, Durbin KR, Li M, Ahlf DR, Early BP, Thomas PM, Sweedler JV, Kelleher NL. A protease for 'middle-down' proteomics. Nat Methods 2012;9:822–4.

[22] Laskay ÜA, Srzentić K, Fornelli L, Upir O, Kozhinov AN, Monod M, Tsybin YO. Practical considerations for improving the productivity of mass spectrometry-based proteomics. Chimia 2013;67:244–9.

[23] Cannon J, Lohnes K, Wynne C, Wang Y, Edwards N, Fenselau C. High-throughput middle-down analysis using an orbitrap. J Proteome Res 2010;9:3886–90.

[24] Fornelli L, Ayoub D, Aizikov K, Beck A, Tsybin YO. Middle-down analysis of monoclonal antibodies with electron transfer dissociation orbitrap fourier transform mass spectrometry. Anal Chem 2014;86:3005–12.

[25] Hua L, Low TY, Sze SK. Microwave-assisted specific chemical digestion for rapid protein identification. Proteomics 2006;6: 586–91.

[26] Mizuno K, Nakamura T, Ohshima T, Tanaka S, Matsuo H. Characterization of KEX2-encoded endopeptidase from yeast *Saccharomyces cerevisiae*. Biochem Biophys Res Commun 1989; 159:305–11.

[27] Yang HJ, Shin S, Kim J, Hong J, Lee S, Kim J. Vortex-assisted tryptic digestion. Rapid Commun Mass Spectrom 2011;25: 88–92.

[28] Lowenthal MS, Liang Y, Phinney KW, Stein SE. Quantitative bottom-up proteomics depends on digestion conditions. Anal Chem 2014;86:551–8.

[29] Albrecht A, Felk A, Pichova I, Naglik JR, Schaller M, de Groot P, Maccallum D, Odds FC, Schafer W, Klis F, Monod M, Hube B. Glycosylphosphatidylinositol-anchored proteases of *Candida albicans* target proteins necessary for both cellular processes and host-pathogen interactions. J Biol Chem 2006;281:688–94.

[30] Aoki W, Kitahara N, Miura N, Morisaka H, Yamamoto Y, Kuroda K, Ueda M. Comprehensive characterization of secreted aspartic proteases encoded by a virulence gene family in *Candida albicans*. J Biochem 2011;150:431–8.

[31] Schild L, Heyken A, de Groot PW, Hiller E, Mock M, de Koster C, Horn U, Rupp S, Hube B. Proteolytic cleavage of covalently linked cell wall proteins by *Candida albicans* Sap9 and Sap10. Eukaryot Cell 2011;10:98–109.

[32] Mitulovic G, Stingl C, Steinmacher I, Hudecz O, Hutchins JRA, Peters JM, Mechtler K. Preventing carryover of peptides and proteins in nano LC–MS separations. Anal Chem 2009;81: 5955–60.

[33] Kalli A, Smith GT, Sweredoski MJ, Hess S. Evaluation and optimization of mass spectrometric settings during data-dependent acquisition mode: focus on LTQ-Orbitrap mass analyzers. J Proteome Res 2013;12:3071–86.

[34] Wong CCL, Cociorva D, Venable JD, Xu T, Yates JR. Comparison of different signal thresholds on data dependent sampling in orbitrap and LTQ mass spectrometry for the identification of peptides and proteins in complex mixtures. J Am Soc Mass Spectrom 2009;20:1405–14.

[35] Michalski A, Neuhauser N, Cox J, Mann M. A systematic investigation into the nature of tryptic HCD spectra. J Proteome Res 2012;11:5479–91.

[36] Geiger T, Cox J, Mann M. Proteomics on an orbitrap benchtop mass spectrometer using all-ion fragmentation. Mol Cell Proteomics 2010;9:2252–61.

[37] Colaert N, Helsens K, Martens L, Vandekerckhove J, Gevaert K. Improved visualization of protein consensus sequences by iceLogo. Nat Methods 2009;6:786–7.

[38] Rodriguez-Suarez E, Hughes C, Gethings L, Giles K, Wildgoose J, Stapels M, Fadgen KE, Geromanos SG, Vissers JPC, Elortza F, Langridge J. An ion mobility assisted data independent LC–MS strategy for the analysis of complex biological samples. Curr Anal Chem 2013;9:199–211.

[39] Makarov A. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. Anal Chem 2000;72:1156–62.

[40] Scigelova M, Makarov A. Orbitrap mass analyzer—overview and applications in proteomics. Proteomics 2006;6(Suppl. 2): 16–21.

[41] Good DM, Wirtala M, McAlister GC, Coon JJ. Performance characteristics of electron transfer dissociation mass spectrometry. Mol Cell Proteomics 2007;6:1942–51.

[42] Zhurov KO, Fornelli L, Wodrich MD, Laskay UA, Tsybin YO. Principles of electron capture and transfer dissociation mass spectrometry applied to peptide and protein structure analysis. Chem Soc Rev 2013;42:5014–30.

[43] Wenger CD, Coon JJ. A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. J Proteome Res 2013;12:1377–86.

[44] Hansen RE, Ostergaard H, Norgaard P, Winther JR. Quantification of protein thiols and dithiols in the picomolar range using sodium borohydride and 4,4′-dithiodipyridine. Anal Biochem 2007;363:77–82.

[45] Creighton TE. Disulfide bond formation in proteins. Methods Enzymol 1984;107:305–29.

[46] Wang Y, Lu QZ, Wu SL, Karger BL, Hancock WS. Characterization and comparison of disulfide linkages and scrambling patterns in therapeutic monoclonal antibodies: using LC–MS with electron transfer dissociation. Anal Chem 2011;83:3133–40.

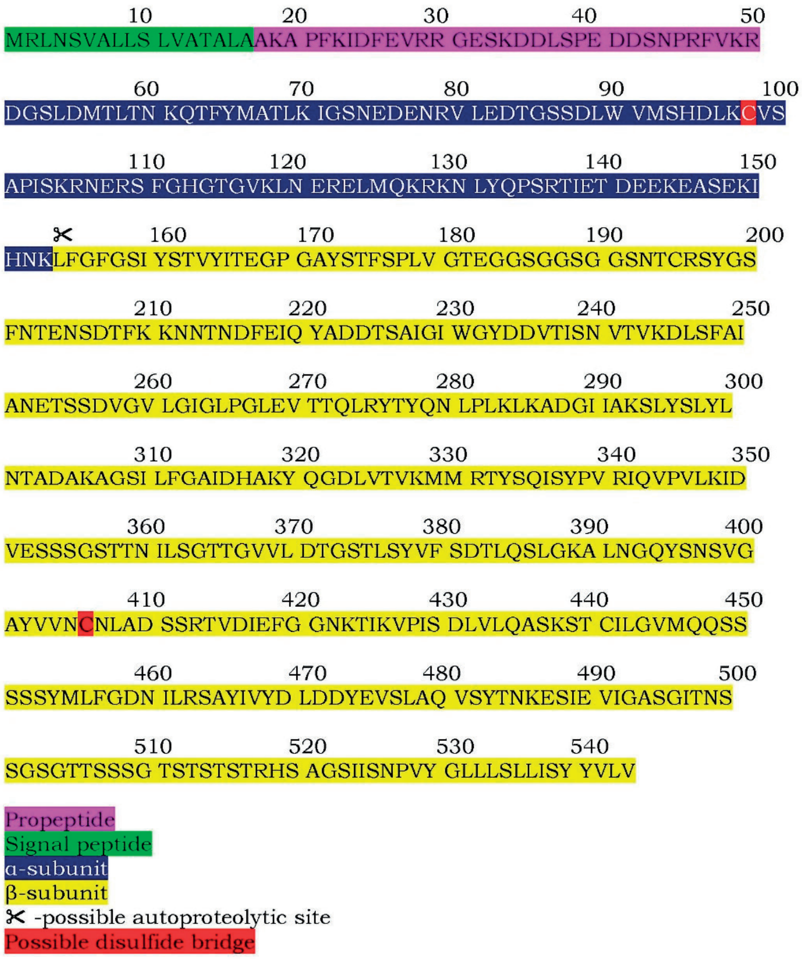
[47] Srzentić K, Fornelli L, Laskay UA, Beck A, Ayoub D, Tsybin YO. Advantages of extended bottom-up proteomics using Sap9 for analysis of monoclonal antibodies; 2014 [submitted].

Supplementary Information.

Extended Bottom-Up Proteomics with Secreted Aspartic Protease Sap9

Ünige A. Laskay, Kristina Srzentić, Michel Monod, and Yury O. Tsybin

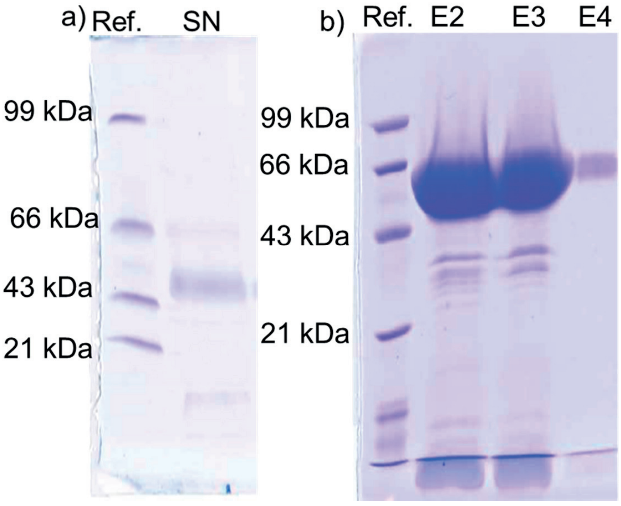
Figure S1. Amino acid sequence of C. albicans Sap9 from UniProt database (accession number O42779).



```
            10         20         30         40         50
MRLNSVALLS LVATALAAKA PFKIDFEVRR GESKDDLSPE DDSNPRFVKR

            60         70         80         90        100
DGSLDMTLTN KQTFYMATLK IGSNEDENRV LEDTGSSDLW VMSHDLKCVS

           110        120        130        140        150
APISKRNERS FGHGTGVKLN ERELMQKRKN LYQPSRTIET DEEKEASEKI

  ✂        160        170        180        190        200
HNKLFGFGSI YSTVYITEGP GAYSTFSPLV GTEGGSGGSG GSNTCRSYGS

           210        220        230        240        250
FNTENSDTFK KNNTNDFEIQ YADDTSAIGI WGYDDVTISN VTVKDLSFAI

           260        270        280        290        300
ANETSSDVGV LGIGLPGLEV TTQLRYTYQN LPLKLKADGI IAKSLYSLYL

           310        320        330        340        350
NTADAKAGSI LFGAIDHAKY QGDLVTVKMM RTYSQISYPV RIQVPVLKID

           360        370        380        390        400
VESSSGSTTN ILSGTTGVVL DTGSTLSYVF SDTLQSLGKA LNGQYSNSVG

           410        420        430        440        450
AYVVNCNLAD SSRTVDIEFG GNKTIKVPIS DLVLQASKST CILGVMQQSS

           460        470        480        490        500
SSSYMLFGDN ILRSAYIVYD LDDYEVSLAQ VSYTNKESIE VIGASGITNS

           510        520        530        540
SGSGTTSSSG TSTSTSTRHS AGSIISNPVY GLLLSLLISY YVLV
```

Propeptide
Signal peptide
α-subunit
β-subunit
✂ -possible autoproteolytic site
Possible disulfide bridge

Figure S2. 10% 1D SDS PAGE gel with a) 10 uL of *P. pastoris* supernatant (SN), b) 5 uL of the E2-E4 fractions eluted from the Ni column.



His6 affinity column allow the production of highly purified Sap9 where non-tagged proteins are not retained on the column. The bands below the 43 kDa marker on FigureS2 panel b arise from degradation products of the recombinant Sap9. This was confirmed by western blotting using an anti-Sap9 antiserum (data not shown). An advantage of the P. pastoris expression system is that this yeast secretes very low level of native proteins.

Figure S3. Fluorescence-based enzyme activity assay recorded at 25 °C, pH 3.5-6, enzyme:protein ratios a) 1:2.5-1:25, b) 1:50-1:100.

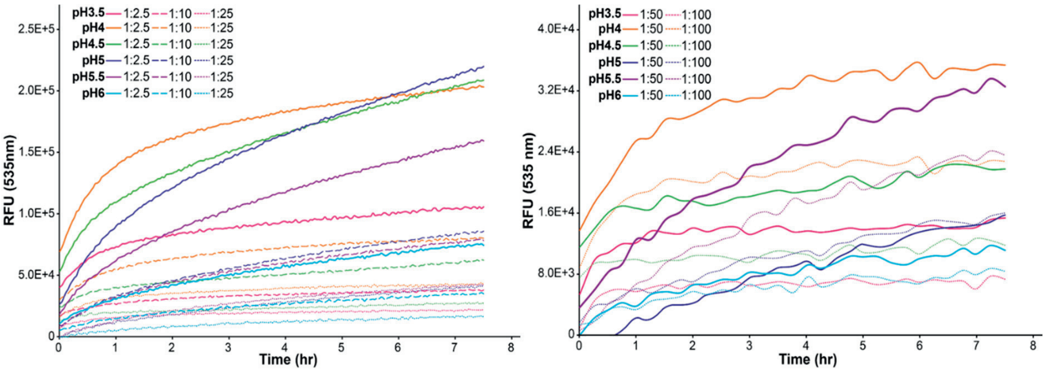


Figure S4. Fluorescence-based enzyme activity assay recorded at 37 °C, pH 3.5-6, enzyme:protein ratios a) 1:2.5-1:25, b) 1:50-1:100

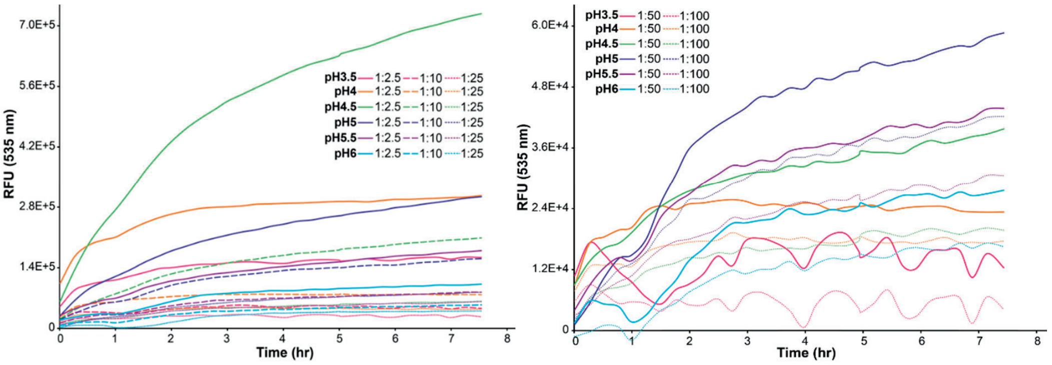


Figure S5. Fluorescence-based enzyme activity assay recorded at 45 °C, pH 3.5-6, enzyme:protein ratios a) 1:2.5-1:25, b) 1:50-1:100.

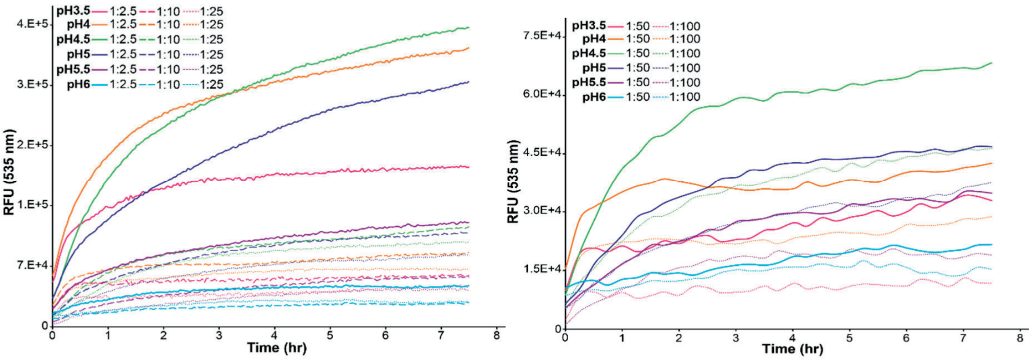


Figure S6. Extracted ion chromatograms of three technical replicates of LC-MS/MS of carbonic anhydrase digested with Sap9.

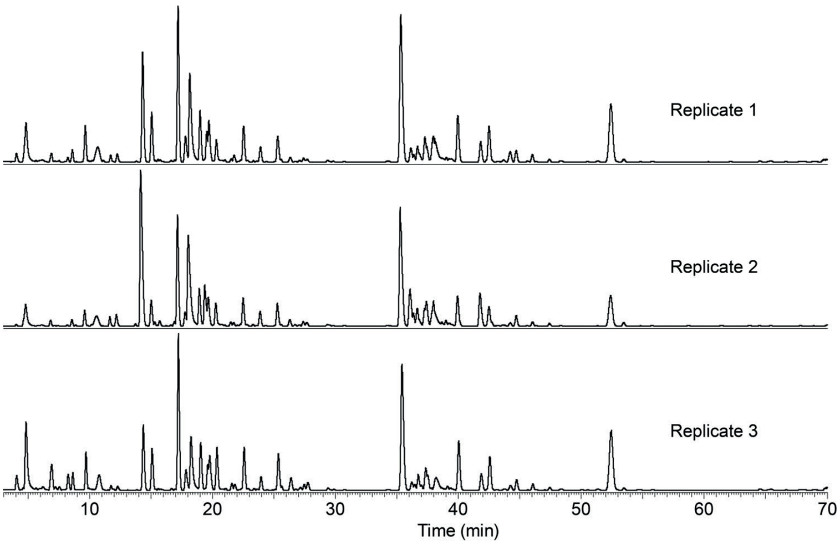
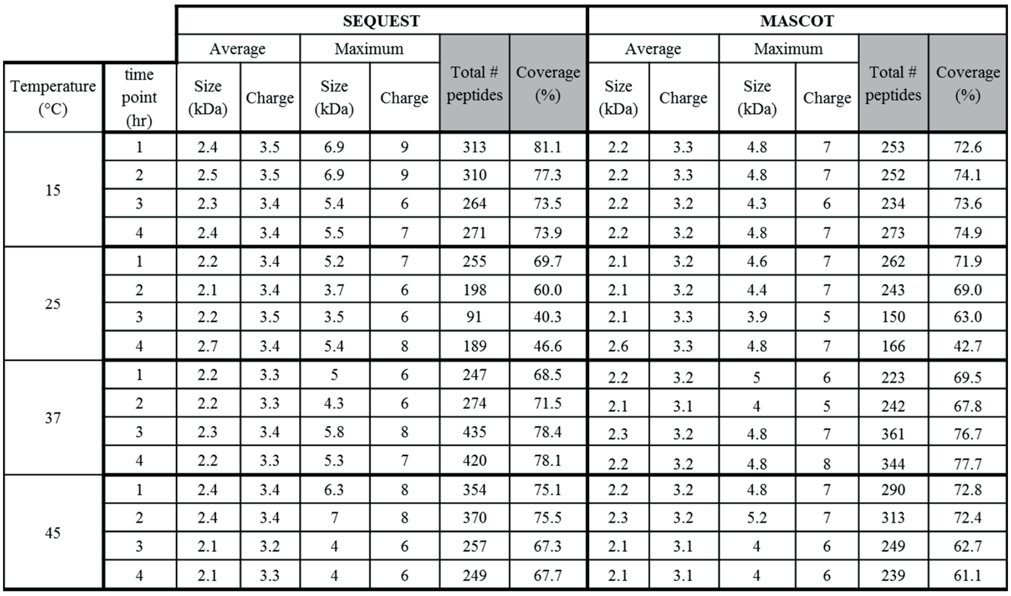


Table S1. Effect of temperature on digestion at E:P ratio 1:2.5 (w/w) at pH 4.5. Size and charge state distribution of proteolytic peptides from bovine carbonic anhydrase 2 after 1 to 8 hours Sap9 digestion obtained by Sequest and Mascot.

		SEQUEST					MASCOT				
		Average		Maximum			Average		Maximum		
Temperature (°C)	time point (hr)	Size (kDa)	Charge	Size (kDa)	Charge	Coverage (%)	Size (kDa)	Charge	Size (kDa)	Charge	Coverage (%)
15°C	1	3	4.3	6.5	10	100	2.4	3.5	4.8	6	96.92
	2	3	4.3	9.5	10	100	2.4	3.4	4.8	5	96.92
	3	3.1	4.3	8.8	12	100	2.4	3.4	4.8	5	100
	4	3	4.2	7.8	10	100	2.4	3.4	4.8	5	100
	5	2.8	4	6	8	100	2.4	3.3	8.4	7	96.92
	6	3	4.3	8.8	13	100	2.3	3.3	4.8	5	96.92
	7	2.8	3.9	6.5	9	100	2.4	3.4	4.8	6	96.92
	8	2.8	3.9	6.1	8	100	2.4	3.4	4.8	5	96.92
25°C	1	3	4.3	7.3	11	100	2.4	3.5	4.8	7	100
	2	3	4.3	9.5	11	100	2.3	3.4	4.8	5	100
	3	2.9	4.3	6	9	100	2.4	3.5	4.8	6	100
	4	2.7	4	6	9	100	2.4	3.4	8.4	7	96.92
	5	2.8	4	9.5	8	100	2.4	3.4	4.8	5	96.92
	6	2.8	4	9.5	8	100	2.4	3.3	4.8	5	96.92
	7	2.8	3.9	9.5	8	100	2.4	3.3	4.8	6	96.92
	8	2.7	3.9	9.5	9	100	2.3	3.3	4.8	5	96.92
37°C	1	2.4	3.8	4.1	7	58	2.2	3.5	3.7	5	58
	2	2.4	4	4.1	7	44	2.3	3.4	3.5	6	44
	3	2.5	3.9	4.1	6	38	2.1	3.4	3.6	6	38
	4	2.6	4	4.8	7	77	2.2	3.2	3.5	5	77
	5	2.8	4.6	4.1	7	41	2.1	3	3.1	5	41
	6	2.8	4.4	4.1	6	22	2.1	3.1	3.2	5	22
	7	2.9	4.1	4.8	6	53	2.1	3.1	3.2	6	53
	8	2.8	4.8	4.1	6	31	2	3.6	3.2	6	31
45°C	1	2.8	4	6	9	100	2.4	3.4	4.8	5	100
	2	2.5	3.6	5.8	7	100	2.3	3.4	4.8	5	100
	3	2.5	3.6	5.7	8	100	2.2	3.3	4.8	5	100
	4	2.7	3.8	7.8	9	100	2.3	3.3	4.8	5	100
	5	2.8	3.8	7.8	9	100	2.3	3.3	4.8	5	100
	6	2.8	4	8.8	13	100	2.3	3.3	4.8	5	100
	7	2.8	4	7.8	10	100	2.3	3.3	4.8	6	96.92
	8	2.8	3.9	6	9	100	2.3	3.3	4.8	5	96.92

Table S2. Effect of temperature on Sap9 cleavage. Size and charge state distribution of proteolytic peptides obtained from seven protein mixture after 1 to 4 hours Sap9 digestion at E:P ratio 1:2.5 (w/w) identified by Sequest and Mascot.



		SEQUEST						MASCOT					
		Average		Maximum				Average		Maximum			
Temperature (°C)	time point (hr)	Size (kDa)	Charge	Size (kDa)	Charge	Total # peptides	Coverage (%)	Size (kDa)	Charge	Size (kDa)	Charge	Total # peptides	Coverage (%)
15	1	2.4	3.5	6.9	9	313	81.1	2.2	3.3	4.8	7	253	72.6
	2	2.5	3.5	6.9	9	310	77.3	2.2	3.3	4.8	7	252	74.1
	3	2.3	3.4	5.4	6	264	73.5	2.2	3.2	4.3	6	234	73.6
	4	2.4	3.4	5.5	7	271	73.9	2.2	3.2	4.8	7	273	74.9
25	1	2.2	3.4	5.2	7	255	69.7	2.1	3.2	4.6	7	262	71.9
	2	2.1	3.4	3.7	6	198	60.0	2.1	3.2	4.4	7	243	69.0
	3	2.2	3.5	3.5	6	91	40.3	2.1	3.3	3.9	5	150	63.0
	4	2.7	3.4	5.4	8	189	46.6	2.6	3.3	4.8	7	166	42.7
37	1	2.2	3.3	5	6	247	68.5	2.2	3.2	5	6	223	69.5
	2	2.2	3.3	4.3	6	274	71.5	2.1	3.1	4	5	242	67.8
	3	2.3	3.4	5.8	8	435	78.4	2.3	3.2	4.8	7	361	76.7
	4	2.2	3.3	5.3	7	420	78.1	2.2	3.2	4.8	8	344	77.7
45	1	2.4	3.4	6.3	8	354	75.1	2.2	3.2	4.8	7	290	72.8
	2	2.4	3.4	7	8	370	75.5	2.3	3.2	5.2	7	313	72.4
	3	2.1	3.2	4	6	257	67.3	2.1	3.1	4	6	249	62.7
	4	2.1	3.3	4	6	249	67.7	2.1	3.1	4	6	239	61.1

Table S3. Sequence coverage of UPS-1 proteins after 1 hr Sap9 and overnight trypsin digestion obtained by Sequest at protein FDR 1%, peptide XCorr score thresholds 2.5 (2+), 3.5 (3+), and 3.8 (≥4+). Data dependent MS/MS fragmentation was triggered at signal threshold 5000. Database search was performed against a database containing human proteins, UPS-1 formulation proteins and their shuffled sequences. Sap9 data was searched using no-enzyme setting and trypsin data was searched using trypsin with 2 missed cleavages.

Protein	**Accesion #**	**MW**	**Trypsin overnight**		**Sap9 1 hr**							
			CID		**> 3+ CID**		**>3+ HCD**		**>2+ CID**		**>2+ HCD**	
			#pept.	Seq. Cov.	#pept.	Seq. Cov.	#pept.	Seq. Cov.	#pept.	Seq. Cov.	#pept.	Seq. Cov.
Alpha-lactalbumin	P00709	16 kDa	5	38%	3	25%	4	46%	7	54%	8	54%
Annexin A5	P08758	36 kDa	22	74%	2	13%	2	13%	3	13%	4	13%
Antithrombin-III	P01008	53 kDa	30	71%	5	27%	4	23%	12	41%	10	41%
BH3-interacting domain death agonist 1	P55957	22 kDa	15	70%			1	10%	4	38%	4	34%
Beta-2-microglobulin	P61769	14 kDa	5	49%	3	36%	3	65%	5	64%	6	77%
C-reactive protein	P02741	25 kDa	7	27%								
Carbonic anhydrase 1	P00915	29 kDa	14	73%	1	9%	3	22%	2	22%	3	28%
Carbonic anhydrase 2	P00918	29 kDa	17	71%	4	19%	4	20%	5	25%	6	26%
Catalase OS=Homo sapiens	P04040	60 kDa	46	71%	10	28%	15	39%	23	48%	24	52%
Complement C5 OS=Homo sapiens	P01031	188 kDa	1	18%	3	85%	3	85%	2	39%	4	85%
Creatine kinase M-type	P06732	43 kDa	25	67%	11	35%	8	24%	14	34%	17	31%
Cytochrome b5	P00167	15 kDa	5	41%	1	14%	1	14%	2	20%	2	20%
Cytochrome c	P99999	12 kDa	7	46%	4	52%	6	91%	8	68%	8	91%
Fatty acid-binding protein, heart	P05413	15 kDa	7	51%	9	86%	11	100%	17	100%	17	100%
Gamma-synuclein	O76070	13 kDa	13	82%	7	50%	9	86%	11	86%	10	72%
Gelsolin	P06396	86 kDa	29	48%	6	20%	8	20%	14	32%	13	31%
Glutathione S-transferase A1	P08263	26 kDa	7	30%	2	10%	3	18%	5	21%	5	21%
Glutathione S-transferase P	P09211	23 kDa	13	71%					1	6%		
GTPase HRas (Chain 1-189)	P01112	21 kDa	12	62%	5	51%	6	58%	9	68%	6	52%
Hemoglobin subunit alpha	P69905	15 kDa	9	84%	2	21%	5	88%	10	88%	8	64%
Hemoglobin subunit beta	P68871	16 kDa	11	84%	2	21%	3	51%	6	51%	5	46%
Histidine--tRNA ligase, cytoplasmic	P12081	57 kDa	32	64%	5	13%	4	10%	3	11%	6	14%
Insulin-like growth factor II	P01344	20 kDa	4	76%	2	25%	4	76%	8	100%	8	37%
Interferon gamma	P01579	19 kDa	11	50%	10	87%	10	87%	11	87%	10	87%
Lactotransferrin	P02788	78 kDa	56	72%	10	26%	10	26%	13	23%	19	31%
Leptin	P41159	19 kDa	6	41%	3	33%	5	37%	6	49%	6	49%
Lysozyme C	P61626	17 kDa	4	45%	1	12%	1	13%	1	9%	2	12%
Microtubule-associated protein tau	P10636-8	79 kDa	23	60%	25	40%	25	54%	34	54%	32	32%
Myoglobin	P02144	17 kDa	18	92%	4	46%	7	87%	6	87%	9	92%
NAD(PH dehydrogenase [quinone] 1	P15559	31 kDa	16	49%	3	34%	4	40%	4	40%	6	44%
NEDD8	Q15843	9 kDa	3	41%	3	33%	6	94%	8	94%	6	94%
Peptidyl-prolyl cis-trans isomerase A	P62937	18 kDa	12	77%	2	18%	3	35%	3	35%	4	35%
Peroxiredoxin-1	Q06830	22 kDa	16	68%	2	13%	2	13%	3	13%	3	13%
Platelet-derived growth factor subunit B	P01127	27 kDa										
Polyubiquitin-B	P62988	26 kDa	3	43%	1	28%	1	28%	1	28%	1	17%
Pro-epidermal growth factor	P01133	134 kDa							1	15%	1	15%
Retinol-binding protein 4	P02753	23 kDa	9	72%	3	24%	3	24%	5	49%	5	49%
Ribosyldihydronicotinamide dehydrogenase [quinone]	P16083	26 kDa	14	73%	4	34%	7	41%	5	39%	7	41%
Serotransferrin	P02787	77 kDa	54	73%	23	43%	22	46%	28	56%	37	64%
Serum albumin	P02768	69 kDa	32	66%	5	13%	6	18%	6	23%	7	21%
Small ubiquitin-related modifier 1	P63165	39 kDa	23	57%	3	12%	6	46%	12	38%	12	56%
SUMO-conjugating enzyme UBC9	P63279	18 kDa	7	58%	1	8%	1	9%	7	73%	5	53%
Superoxide dismutase [Cu-Zn]	P00441	16 kDa	8	66%	1	15%	2	35%	3	24%	5	48%
Thioredoxin	P10599	12 kDa	6	63%								
Tumor necrosis factor	P01375	26 kDa	8	54%	1	21%	4	45%	4	52%	5	59%
Ubiquitin-conjugating enzyme E2 C V=1	O00762	20 kDa	13	81%					4	28%	1	7%
Ubiquitin-conjugating enzyme E2 E1	P51965	21 kDa	3	16%	2	7%	1	7.0%	1	7%	1	7%
Total Number of peptides identified			681		194		233		337		358	
Average # peptides/protein			15.1		4.9		5.7		7.7		8.3	
Average sequence coverage			60%		30%		43%		44%		45%	
% spectra identified			21%		46%		42%		48%		54%	
Total Number of UPS-1 proteins/total proteins with 2 peptides			**44/61**		**32/33**		**35/36**		**39/41**		**39/41**	
Total Number of UPS-1 proteins/total proteins with 1 peptide			45/64		40/42		41/43		44/46		43/45	

Table S4. Sequence coverage of UPS-2 proteins after 1 hr Sap9 and overnight trypsin digestion obtained by Sequest at protein FDR 1%, peptide XCorr score thresholds 2.5 (2+), 3.5 (3+), and 3.8 (≥4+). Data dependent MS/MS fragmentation was triggered at signal threshold 5000. Database search was performed against a database containing human proteins, UPS-2 formulation proteins and their shuffled sequences. Sap9 data was searched using no-enzyme setting and trypsin data was searched using trypsin with 2 missed cleavages.

Protein		Trypsin overnight		Sap9 1 hour	
	fmol	# peptides	Seq. Coverage	# peptides	Seq. Coverage
Carbonic anhydrase 1	50000	53	96%	38	94%
Carbonic anhydrase 2	50000	43	79%	34	97%
Complement C5	50000	3	42%	18	85%
Hemoglobin subunit alpha	50000	31	93%	35	100%
Hemoglobin subunit beta	50000	39	90%	39	95%
Leptin	50000	38	92%	31	100%
Serum albumin	50000	102	58%	160	98%
Ubiquitin	50000	11	83%	4	53%
Catalase	5000	52	80%	41	67%
Cytochrome b5	5000	13	74%	2	20%
Myoglobin	5000	16	80%	17	92%
NAD(P)H dehydrogenase [quinone] 1	5000	24	62%	5	29%
Peptidyl-prolyl cis-trans isomerase A	5000	19	70%	9	43%
Peroxiredoxin 1	5000	16	71%	4	13%
Pro-Epidermal growth factor	5000	2	76%	6	83%
Small ubiquitin-related modifier 1	5000	24	73%	18	59%
Alpha-lactalbumin	500	5	33%	7	88%
Creatine kinase M-type	500	9	37%	5	17%
Histidyl-tRNA synthetase, cytoplasmic	500	19	50%		
Lysozyme C	500	3	40%	1	8.50%
NEDD8	500	5	35%		
Retinol-binding protein 4	500	7	68%	4	29%
Ribosyldihydronicotinamide dehydrogenase [quinone]	500	10	71%	3	34%
SUMO-conjugating enzyme UBC9	500	4	50%	2	13%
Antithrombin-III (Chain 33-464)	50	5	21%		
Beta-2-microglobulin	50	1	7.70%	1	12%
BH3-interacting domain death agonist	50	1	22%		
Gamma-synuclein	50	2	37%		
Glutathione S-transferase A1	50	1	5.90%	1	4.10%
Insulin-like growth factor II	50	1	36%	1	19%
Thioredoxin	50	1	10.00%		
Average sequence coverage		56%		54%	
Total number of peptides		560		486	

5.2. Paper IV: Advantages of extended bottom-up proteomics using Sap9 for analysis of monoclonal antibodies

analytical chemistry

Article

pubs.acs.org/ac

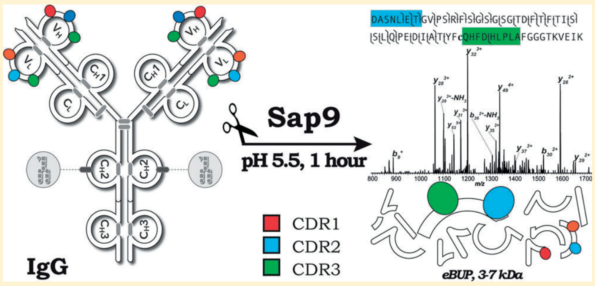
Advantages of Extended Bottom-Up Proteomics Using Sap9 for Analysis of Monoclonal Antibodies

Kristina Srzentić,[†] Luca Fornelli,[†] Ünige A. Laskay,[†] Michel Monod,[‡] Alain Beck,[§] Daniel Ayoub,[†] and Yury O. Tsybin*,[†]

[†]Biomolecular Mass Spectrometry Laboratory, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
[‡]Department of Dermatology, Centre Hospitalier Universitaire Vaudois, 1011 Lausanne, Switzerland
[§]Centre d'Immunologie Pierre Fabre, 74160 St. Julien-en-Genevois, France

Supporting Information

ABSTRACT: Despite the recent advances in structural analysis of monoclonal antibodies with bottom-up, middle-down, and top-down mass spectrometry (MS), further improvements in analysis accuracy, depth, and speed are needed. The remaining challenges include quantitatively accurate assignment of post-translational modifications, reduction of artifacts introduced during sample preparation, increased sequence coverage per liquid chromatography (LC) MS experiment, and ability to extend the detailed characterization to simple antibody cocktails and more complex antibody mixtures. Here, we evaluate the recently introduced extended bottom-up proteomics (eBUP) approach based on proteolysis with secreted aspartic protease 9, Sap9, for analysis of monoclonal antibodies. Key findings of the Sap9-based proteomics analysis of a single antibody include: (i) extensive antibody sequence coverage with up to 100% for the light chain and up to 99−100% for the heavy chain in a single LC-MS run; (ii) connectivity of complementarity-determining regions (CDRs) via Sap9-produced large proteolytic peptides (3.4 kDa on average) containing up to two CDRs per peptide; (iii) reduced artifact introduction (e. g., deamidation) during proteolysis with Sap9 compared to conventional bottom-up proteomics workflows. The analysis of a mixture of six antibodies via Sap9-based eBUP produced comparable results. Due to the reasons specified above, Sap9-produced proteolytic peptides improve the identification confidence of antibodies from the mixtures compared to conventional bottom-up proteomics dealing with shorter proteolytic peptides.



Immunoglobulin Gs (IgGs) are glycoproteins involved in the adaptive immune response in animals and represent 75% of human immunoglobulins circulating in serum. IgG structure is composed of two identical light chains (~25 kDa each) and two identical heavy chains (~50 kDa each), with a total molecular weight of approximately 150 kDa. Currently, monoclonal recombinant IgGs are the most popular biotherapeutics and are employed in the treatment of a variety of pathologies such as cancer and inflammatory diseases.[1] Mass spectrometry (MS) has emerged as the analytical technique of choice for the in-depth characterization of IgGs,[2] and the technique is particularly important for product quality control (QC) in biotechnology and in the pharmaceutical industry. During bioproduction, purification, and storage, monoclonal IgGs are subject to a variety of post-translational modifications (PTMs). Prior to approval for therapeutic use, thorough qualitative and quantitative characterization of IgGs is required by sanitary authorities. Furthermore, with the impending patent expiration of some originator IgGs and the arrival of biosimilar IgG products to the market, the need for high-throughput analytical methods allowing comparability studies between originator and biosimilar molecules became urgent. MS-based peptide mapping and PTM assignment of IgGs is primarily conducted following a bottom-up proteomics (BUP) approach, entailing proteolytic digestion of the antibody with one or a cocktail of the BUP benchmarked proteases of choice (most commonly trypsin, GluC, or LysC), which yields peptides with an average mass of ~2 kDa.[3−5] Recently, top-down (TD) mass spectrometry approaches have been applied for the study of intact IgGs.[6−8] TD MS offers certain advantages over BUP, including limited sample manipulation and, hence, reduced probability for introducing artifacts in the sample. Nevertheless, BUP is better supported by the current state-of-the-art in terms of both technical equipment and data analysis software, therefore remains the technique of choice for most MS laboratories and QC units. Furthermore, the use of proteolytic enzymes with different cleavage specificities combined with multiple liquid chromatography-tandem mass spectrometry (LC-MS/MS) experiments can result in up to 100% sequence coverage of IgGs, whereas TD MS is currently limited to ~30%.[6]

Received: July 24, 2014
Accepted: September 10, 2014
Published: September 10, 2014

© 2014 American Chemical Society

One of the major limitations in BUP studies of the IgGs is introduction of artifacts. These may ensue from the lengthy sample preparation and protein digestion protocols carried out under basic pH conditions, which are typical, for instance, in denaturation with urea prior to trypsin proteolysis. Specifically, these conditions can favor deamidation, which consists in the replacement of Asn and Gln amino acid residues with Asp and Glu, respectively, through the formation of a cyclic intermediate.[9,10] This phenomenon implies the introduction of charged residues in the polypeptidic chain. Moreover, the process can generate both stereo and structural isomers of Asp and Glu, potentially altering also the high-order structure of the protein. Deamidation is therefore responsible for the loss of biological activity of IgGs when located in the antigen-binding domains and specifically in one of the complementarity-determining regions (CDRs).[11,12] Harris et al. showed that ~17% of Trastuzumab, an important IgG1 biotherapeutic, contains deamidations at two different sites in its CDRs (one in the light chain and one in heavy chain). The deamidated IgGs were separated with ion exchange chromatography, and deamidation sites were assigned using Edman degradation and matrix-assisted laser desorption ionization time-of-flight (MALDI TOF) MS. It was found that the deamidated IgG variants present significantly reduced potency.

Tandem mass spectrometry (MS/MS) is an efficient tool for the assessment of protein deamidation, notably by radical-driven MS/MS.[13] Pioneering studies by O'Connor and co-workers demonstrated the capability of electron capture dissociation (ECD)[14] and electron transfer dissociation (ETD)[15] for distinguishing structural isomers derived by deamidation of Asn and Gln through the detection of diagnostic *c*- and *z*-type product ions.[10,16,17] Zubarev and co-workers extended the application of this technique to large-scale proteome investigations, with ECD performed under LC-MS/MS time constraints.[18]

Some of the above-mentioned limitations of BUP might be solved by the modification of the IgG digestion protocol. Middle-down (MD) MS, performed with limited proteolysis targeting the IgG hinge region using enzymes such as papain or IdeS, was proven to lead to sequence coverage of up to ~65% and has shown potential for the identification of post-translational modifications (PTMs) like Met oxidation without noticeable introduction of artifact PTMs, as shown by Fornelli et al.[19,20] Nevertheless, the analysis of ~25 kDa IgG subunits requires the availability of ion activation/fragmentation techniques efficient on large biomolecules and high-resolution mass analyzers. Similarly to TD MS, ETD is currently the MS/MS technique of choice for MD MS of IgGs. Novel MS/MS techniques, such as the recently introduced ultraviolet photodissociation (UVPD) at 193 nm by Shaw et al.,[21] are expected to improve the performance of both TD and MD MS. Finally, the impetus in modern drug discovery is to extend the MS approaches for the characterization of IgG mixtures, ranging from simple cocktails containing 3−10 IgGs, to complex mixtures of IgGs present in serum. However, the current performance level of TD MS and MD MS for analysis of complex (>5 proteins) IgG mixtures is limited, specifically by the inefficient separation and fractionation of solution-phase protein mixtures. The analysis of IgG mixtures is further complicated by the high proteoform-level complexity of each IgG and high structural similarity of IgG constant regions.

Following the path signed by MD MS studies, herein we describe a method for characterization of IgGs which is aimed at maintaining and strengthening the advantages inherent to BUP, particularly the high sequence coverage, while minimizing some of its characteristic drawbacks. The proposed approach, referred to as extended bottom-up proteomics (eBUP),[22] is based on IgG proteolysis by the secreted aspartic protease 9 (Sap9) from *C. albicans*, which oftentimes produces peptides substantially larger (~3.5 kDa on average) than in BUP.[23] The entire protocol centered on Sap9 digestion was designed for keeping the sample under slightly acidic conditions during all the preparation steps. This is important for reducing de facto Asn and Gln deamidation in respect to typical BUP workflows, as assessed here by LC-MS/MS for Asn deamidation. The advantages of the proposed eBUP approach for IgG structural analysis are demonstrated for an isolated IgG and a cocktail of IgGs, with envisioned method extension for analysis of complex IgG mixtures.

■ EXPERIMENTAL SECTION

Samples and Sample Preparation. Monoclonal antibodies, IgGs, of subclasses IgG1, Adalimumab (Humira, Abbott Laboratories), Bevacizumab (Avastin, Genentech/Roche), Rituximab (Rituxan, IDEC Pharmaceuticals/Genentech) and Trastuzumab (Herceptin, Genentech), IgG2, Panitumumab (Vectibix, Amgen), and IgG4, Natalizumab (Antegren, Biogen IDEC) were obtained in their formulation buffers and versions approved by European Medicines Agency. Ammonium bicarbonate, urea, iodoacetamide, *tris*(2-carboxyethyl)phosphine (TCEP), dithiothreitol (DTT), and sodium citrate were obtained from Sigma-Aldrich (Buchs, Switzerland). Protein-grade guanidinium-chloride (GdnCl) was purchased from Carl Roth GmbH (Karlsruhe, Germany). Acetonitrile (ACN), methanol (MeOH), formic acid (FA), and 2,2,2-trifluoroethanol (TFE) were obtained in LC-MS purity grade. Acetonitrile and methanol were purchased from Fluka Analytical (Buchs, Switzerland). Formic acid was obtained from Merck (Zug, Switzerland), and 2,2,2-trifluoroethanol from Alfa Aesar GmbH & Co KG (Karlsruhe, Germany). Proteases trypsin and Glu-C were purchased from Promega (Dübendorf, Switzerland). Secreted aspartic protease 9 (Sap9) from *C. albicans* was recombinantly expressed in *P. pastoris* and purified as previously described.[24]

Sap9 Proteolysis. Single IgG (Trastuzumab) or an equimolar mixture of six IgGs were subjected to two sample preparation procedures at different pH values before digestion at pH 5.5, as follows. Procedure I: Sample was diluted with 6.8 M urea in 100 mM ammonium bicarbonate buffer (pH 7.8), reduced by addition of DTT to a final concentration of 5 mM and incubated at 50 °C for 1 h, followed by 45 min alkylation at room temperature in the dark with 18 mM final iodoacetamide. To reduce urea concentration below 0.6 M and to obtain optimal digestion pH, samples were further diluted 20× with 50 mM sodium citrate buffer (pH 5.5). Procedure II: Samples were diluted with 6 M GdnCl in 50 mM sodium citrate buffer (pH 5.5) and reduced with TCEP (final concentration 5 mM) for 1 h at room temperature. Alkylation of reduced thiols was omitted due to acidic environment.

Reduced samples prepared either with Procedure I or II were digested in triplicate with Sap9 in enzyme to protein (E/P) ratio 1:10 (w/w) for 1 h at 25 °C and at pH 5.5.[23] Reactions were quenched by addition of TFA to 1% final concentration.

Trypsin and GluC Proteolysis. Single IgG was diluted with 6 M GndCl in 100 mM ammonium bicarbonate buffer (pH 7.8), followed by reduction and alkylation as described above. Sample was further diluted 20× in ammonium bicarbonate buffer (pH 7.8) and digested with trypsin, E/P ratio 1:20 (w/w), at 37 °C overnight.[25] The mixture sample was reduced and alkylated as described in Procedure I, diluted 20× in 100 mM ammonium bicarbonate buffer (pH 7.8), and subjected to overnight proteolysis with trypsin or GluC in E/P ratio 1:20 (w/w) at 37 °C.

LC-MS/MS Analysis. Proteolytic peptides obtained by digestion with Sap9, trypsin or GluC were desalted off-line using pooled C4 and C18 (described in Supporting Information) or just C18 ZipTip cartridges (Millipore, Billerica, MA), respectively, prior to LC separation. Approximately 8 pmol of peptide mixture was loaded onto C8 (2 cm, 100 Å, 5 μm) or C18 (2 cm, 100 Å, 3 μm) trap-columns for 10 min with 0.1% FA at a flow rate of 8 or 4 μL/min, respectively. Reversed-phase nano LC was performed using a Dionex Ultimate 3000 system (Thermo Scientific, Bremen, Germany) equipped with C8 column (i.d. 75 μm, 150 mm, 300 Å, 5 μm) for separation of Sap9 proteolytic peptides or C18 column (i.d. 75 μm, 250 mm, 100 Å, 3 μm) for peptides obtained from trypsin and GluC digestions. Solvent A was composed of 0.1% of FA in water and solvent B of 50% MeOH, 20% ACN, 10% TFE, and 0.1% FA. The percentage of the organic phase was increased from 5 to 60% over 60 min for all performed analyses.

The outlet of chromatographic column was coupled online with a nano electrospray ionization (ESI) source (Nanospray Flex ion source, Thermo Scientific) equipped with a metallic emitter at a 2.2 kV potential. Mass spectrometric analysis was performed on a hybrid high-field LTQ Orbitrap Elite FTMS (Thermo Scientific) equipped with ETD. Analysis of peptides was carried out using different ion activation/fragmentation techniques in data-dependent mode. In all LC-MS/MS runs, the survey scan was performed at 60 000 resolution (at 400 m/z) in the Orbitrap FTMS with automatic gain control (AGC) set at 1E6. Dynamic exclusion of the precursor ion masses was enabled with 60 s duration.

Isolated precursor ions of Sap9-produced peptides were subjected to higher-energy collision induced dissociation (HCD), collision-induced dissociation (CID), or ETD in separate LC-MS/MS runs. Singly and doubly charged precursor ions were excluded from triggering MS/MS event. The AGC (number of charges) target value for MS/MS events was always set to 5E4. HCD was performed in a top-5 mode with product ion detection in the Orbitrap FTMS operating at 15 000 resolution (at 400 m/z) with 3 microscans per each scan. Normalized collision energy (NCE) was set at 27% (default charge state: 3+).[26] CID was performed in a top-10 mode with product ion detection in the LTQ (normal scan speed) and NCE of 35%. Finally, ETD was performed (for IgG mixture only) in a top-5 mode (AGC target value for fluoranthene radical anions 5E5, max injection time 50 ms) with charge-dependent duration enabled (default value of 80 ms for charge state 2+). ETD product ion detection was carried out in the Orbitrap FTMS operating at 15 000 resolution (at 400 m/z) with 3 microscans. Signal-to-noise (S/N) threshold was set to 15 000 throughout all experiments (relative intensity units). Peptides obtained from trypsin and GluC proteolysis were analyzed by CID in a top-10 mode with the instrument operating with the same parameters indicated above, with the exception of rejecting only singly charged peptide ions for MS/MS triggering.

For differentiation of stereoisomers produced by deamidation, either single Trastuzumab or IgG mixture digests obtained with trypsin were analyzed in a LC-MS/MS fashion with the instrument working in targeted mode, using a predefined precursor ion list. Dynamic exclusion was disabled and ETD was performed with AGC target value for fluoranthene of 2E5. Product ion detection was performed in the Orbitrap FTMS at 15 000 resolution (at 400 m/z) averaging 7 microscans. This LC-MS/MS experiment was repeated for the Sap9 IgG digest (Sap9 sample preparation Procedure II).

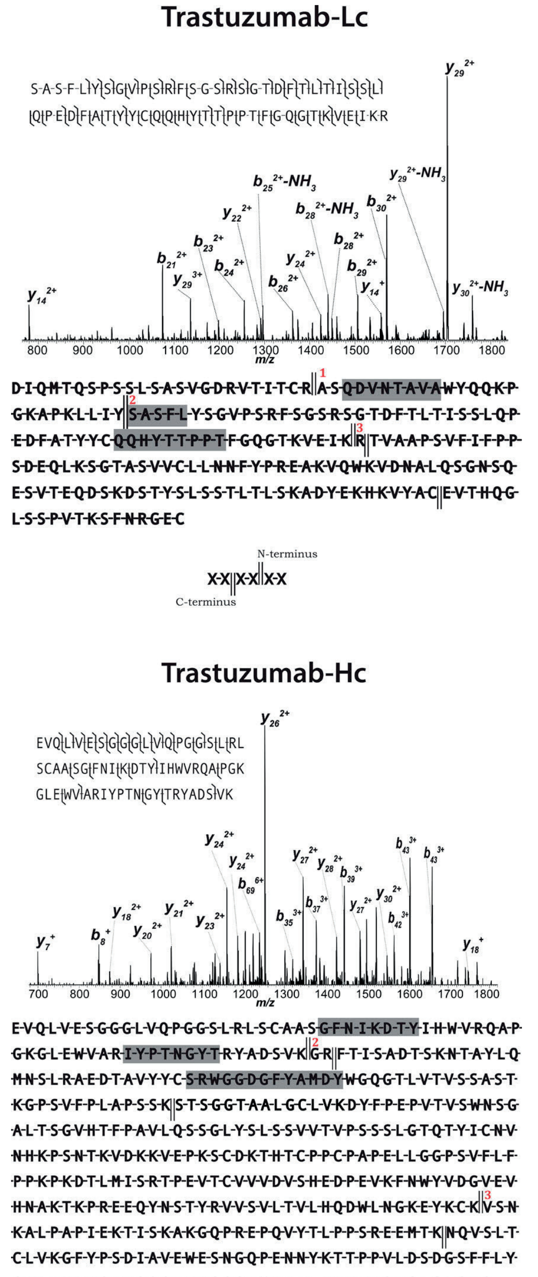


Figure 1. CID mass spectra and sequence coverage of (top) light and (bottom) heavy chains of therapeutic antibody IgG1 Trastuzumab obtained by analysis of Sap9-derived peptides from a single LC-MS/MS analysis. Sequenced regions are shown in bold letters, CDRs are highlighted in gray. Vertical double lines indicate Sap9 cleavage site; numbers 1 and 2 in light chain and 1 to 3 in heavy chain indicate peptides unique to Trastuzumab.

The MS/MS spectra were analyzed using Sequest (Proteome Discoverer 1.4, Thermo Scientific). The Sequest XCorr threshold values were empirically determined and set to 3.5 for CID and 3.0 for HCD. In case of ETD, the threshold values were 2.0, 2.5, and 2.8 for charge states 3, 4, and >5, respectively (arbitrary values determined by visual inspection of mass spectra). The precursor ion mass tolerance was set to 10 ppm,

Table 1. Comparison of Sequence Coverage, Average Peptide Length and Charge, Number of Identified Unique Peptides of Every IgG in Part, and Number of Total Peptides Obtained from a Single LC-MS/MS Analysis after Single Digestion with Sap9 (1 h), Trypsin (Overnight), and GluC (Overnight)

	Trypsin		GluC		Sap9					
	CID		CID		CID		HCD		ETD	
	sequence coverage (%)	# of unique peptides**	sequence coverage (%)	# of unique peptides**	sequence coverage (%)	# of unique peptides	sequence coverage (%)	# of unique peptides	sequence coverage (%)	# of unique peptides
Adalimumab HC	66	5/6	63	1/3	87	5	99	3	64	2
Adalimumab LC	83	2/4	62	1/1	96	4	83	4	61	3
Bevacizumab HC	68	5/6	53	1/1	93	3	97	5	57	0
Bevacizumab LC	72	1/2	50	0/0	72	1	78	1	44	0
Natalizumab HC	47	6/7	55	9/15	75	9	69	14	46	7
Natalizumab LC	63	2/3	56	1/1	91	3	91	4	74	3
Panitumumab HC*	60	8/12	39	2/7	92	13	84	14	30	4
Panitumumab LC	92	3/3	91	2/5	100	2	79	2	51	1
Rituximab HC	54	3/3	53	1/1	92	3	97	4	62	2
Rituximab LC	78	2/3	62	1/1	99	3	87	2	51	1
Trastuzumab HC	63	5/7	63	3/6	93	4	98	3	74	2
Trastuzumab LC	87	2/3	61	1/1	100	2	87	3	65	2
Average peptide size unique vs.total (kDa)	2.2/2.2/2.3		2.9/2.5/2.5		4.3/3.3		4.6/3.3		2.7/2.0	
Average peptide charge unique vs total (+)	2.2/2.5/2.4		3.4/3.1/3.1		4.5/4.1		5.4/4.5		3.9/3.9	

*Unique Panitumumab Hc peptide illustrated and discussed in Figure 3 was accounted for in total number of unique peptides for Sap9 CID and HCD results. **Number of unique peptides for trypsin and GluC is given without/with 2 missed cleavages allowed.[30]

product ion tolerance to 0.02 Da. IgG peptides were searched against the six IgG database and their shuffled sequences, allowing for Cys carbamidomethylation, oxidation of Met, N-terminal pyroglutamic acid formation from Gln and deamidation of Asn as dynamic modifications. The cleavage specificity was set to no-enzyme for Sap9, and to trypsin or GluC allowing for two missed cleavages for corresponding experiments.

■ RESULTS AND DISCUSSION

Trastuzumab Analysis with Sap9. Trastuzumab IgG was subjected to a 1 h Sap9 digestion (sample preparation Procedure I). The resultant peptide mixture was then analyzed using CID LC-MS/MS, Figure 1. Single antibody peptide mapping resulted in up to 100% sequence coverage for both chains of Trastuzumab from a single LC-MS/MS experiment. In comparison, currently employed digestion protocols for antibody analysis (digestion with trypsin and/or endoproteinases such as GluC or LysC)[2] reach these values by combining two or more proteases, long digestion hours, and/or multiple LC runs. Moreover, we identified peptides unique to Trastuzumab that ensure almost 50% sequence coverage of the light (peptides 1 and 2, Figure 1, top panel) and 30% of the heavy chains (peptides 1 and 2, Figure 1, bottom panel). Figure 1 also depicts CID mass spectra of peptides containing consecutive CDRs for both heavy and light chains with assigned CID product ion series (*b*- and *y*-ions), additionally confirming the correct peptide identification.

CDR3 in a variable domain of a heavy chain (VH) is one of the hypervariable loops constituting the antigen-binding site of an antibody that displays high diversity both in sequence and in length. VH CDR3 can be up to 62 residues long, as recently reported by Larsen et al.,[27] and, interestingly, it is the precise length of VH CDR3 that was shown to be the crucial specificity determining factor in formation of an antigen-specific binding site.[28] Nowadays, CDR3 grafting is a widely employed technique for in vitro humanization of antibodies from mammalian cells. Ergo, unambiguous identification and characterization of complete CDR3 from the antibody of interest is important. Figure 1, bottom panel, shows a ~7 kDa long unique peptide, indicated with number 2, which contains the entire CDR3 of a heavy chain (Hc) of Trastuzumab. This, along with the extensive backbone cleavage assignment, offers complete information about sequence, length, and localization of this loop. This would not be the case with benchmarked proteases which would cleave in the middle, beginning, or end of the region, resulting in lost connectivity between different parts and, thus, making it less feasible to get length information and assignment to one antibody.

Peptide indicated by number 3 in light chain (Lc, which is not to confuse with LC being "liquid chromatography") of Trastuzumab (Figure 1 top panel) is another long peptide whose identification dramatically augments sequence coverage and, hence, IgG identification. Notably, peptide 3 of the Hc, aside from increasing sequence coverage, is long enough to be unique to a class of IgG despite its position deep inside the conserved region (Figure S1, Supporting Information). Hence, it further aids in distinguishing between IgG classes in a complex mixture. Complete sequence coverage for both chains (100%) was also reached when HCD fragmentation was applied (data not shown). Moreover, both CID and HCD results attained for a single IgG analysis are comparable with the ones obtained for Trastuzumab from the mixture sample, Table 1, taken as an example, vide infra.

The presence of these long peptides reduces the number of peptide identifications required per LC-MS/MS run to obtain full sequence coverage of investigated proteins. Theoretically, it should also reduce the yield of returned ambiguous hits from a database search. However, the latter remains to be validated on a larger data set, as the presence of substantially longer and highly charged peptides increases spectral complexity in terms of isotopically resolved precursor ions and renders product ion charge state assignments more challenging.

IgG Mixture Analysis with Sap9. We further assessed our method's efficiency of peptide mapping in an antibody mixture. An equimolar mixture of six monoclonal antibodies, IgGs, was subjected to a 1 h proteolytic digestion with Sap9 (sample preparation Procedure I). Proteolytic peptides were analyzed by LC-MS/MS. Sap9 proteolysis was repeated three times, exhibiting high digestion reproducibility as depicted in Figure 2, top panel. The applied eBUP workflow allowed the identification of all six

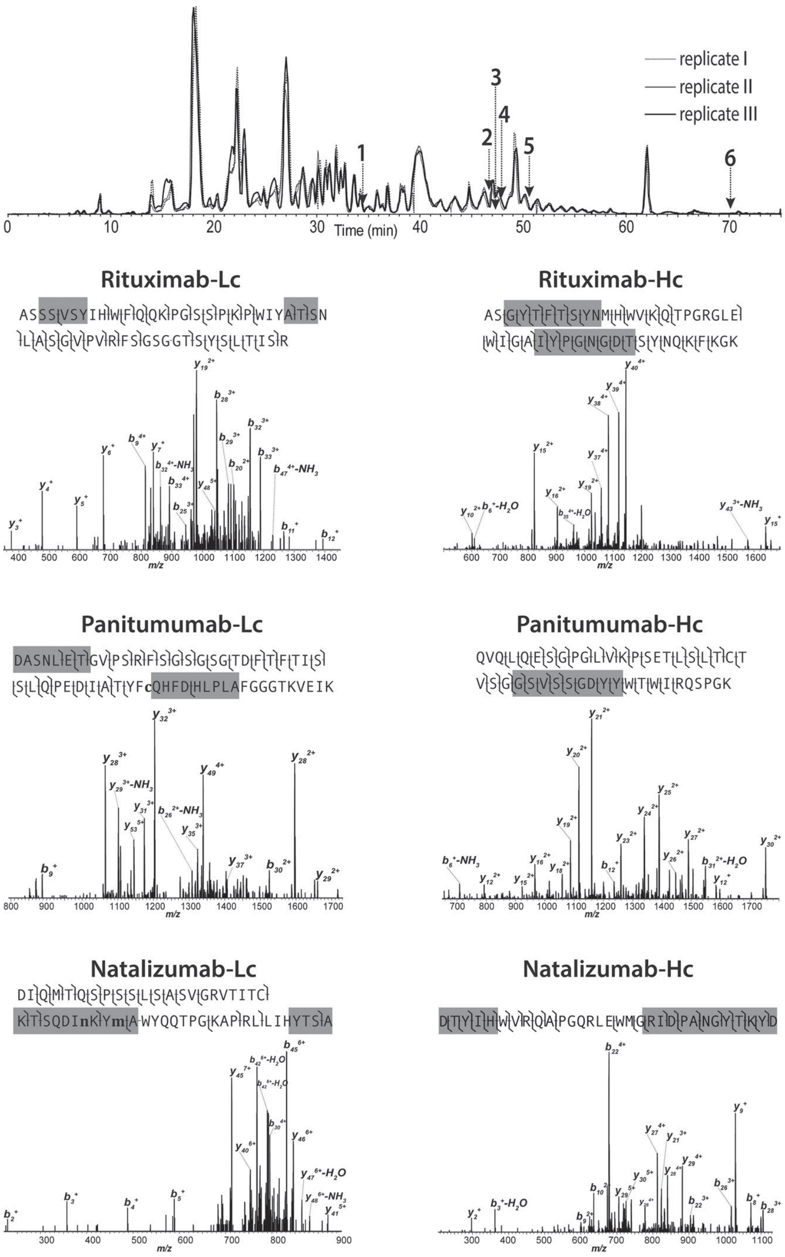


Figure 2. Extracted LC-MS/MS (CID) base peak chromatogram of IgG peptides obtained after triplicate 1 h digestion with Sap9 (sample preparation Procedure I). Panels 1−6 show examples of the unique peptides containing the CDRs for three selected IgGs separated by LC: (1) Natalizumab Hc (IgG4), (2) Rituximab Hc (IgG1), (3) Panitumumab Hc (IgG2), (4) Rituximab Lc, (5) Natalizumab Lc, (6) Panitumumab Lc.

IgGs in a single LC-MS/MS run using either of the ion activation methods applied (CID, HCD, and ETD). The most extensive protein sequence coverage and the highest number of unique backbone cleavages was obtained with CID and in part (sequence coverage) with HCD, whereas the numbers are lower for ETD, Table 1. Discrepancy between fragmentation efficiency in CID and HCD/ETD could be due to a broad span of charge states (3+ to 13+) present in the sample. Since HCD and ETD are charge-dependent activation methods, their successful application requires further optimization of fragmentation parameters such as collision energy in HCD, ion/ion reaction time and number of interacting ions in ETD, or potentially employing the hybrid method that combines the two, such as EThcD.[29]

Importantly, peptides containing CDR entities, required for discrimination between IgGs, were detected for each IgG. Complete and assigned fragmentation maps and companion sequence coverages (with the exception of ETD due to the yet insufficient fragmentation) for each IgG are given in Figure S2, Supporting Information. For comparison, the same IgG mixture was subjected to a digestion with benchmark proteases, trypsin, and GluC, commonly employed for quality control analysis of recombinant IgGs. As expected, both proteases identified all IgGs,

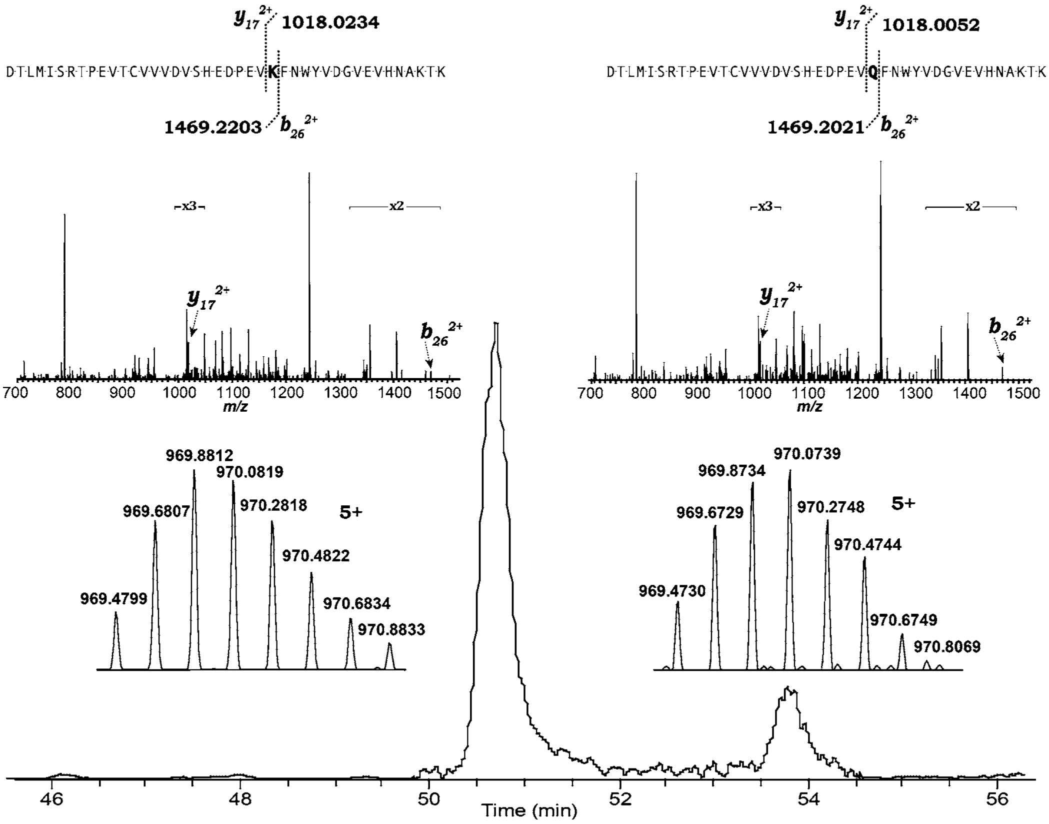


Figure 3. Extracted ion chromatogram of monoisotopic peaks at *m*/*z* 969.4727 and 969.4799 (5+) of the peptides DTLMISRTPEVTCVVVDVSHEDPEVQFNWYVDGVEVHNAKTK and DTLMISRTPEVTCVVVDVSHEDPEVKFNWYVDGVEVHNAKTK, corresponding to a peptide unique to Panitumumab Hc (Q containing), and to a peptide that is shared between IgG1 subclass analyzed herein (K containing), respectively.

with average sequence coverage of 69 and 59%, respectively. Importantly, Sap9-based eBUP analysis resulted in noticeably higher average sequence coverage (up to 91% for CID, Table 1). Notably, there is a 1.5-fold increase in average length of Sap9-produced peptides with respect to those obtained by classical BUP experiments (3.3 and 2.2 kDa, respectively), Table 1. This, in turn, increases the likelihood of the presence of two consecutive CDRs within one peptide produced by Sap9, which is generally not observed for shorter proteolytic peptides.

CDRs are parts of the IgG variable domains, which derive from different genes linked during a genetic recombination process. As a result, different IgGs might share one or more CDR regions, even when targeting different antigens. Specifically, in the mixture used for this study, the CDRs of the light chains of Panitumumab and Bevacizumab demonstrate a shared sequence: QDISNY. In addition, a CDR belonging to the Hc of Trastuzumab (sequence: GFNIKDTY) is identical to part of the variable domain of the Hc of Natalizumab (Figure S2, Supporting Information). Therefore, the presence of peptides containing two CDRs may lead to more reliable identification of each IgG present in a mixture, and it is pertinent to determine connectivity between these regions. Results from a single LC-MS/MS run of 1 h Sap9 digest indicate that highly charged long peptides containing consecutive CDRs for all IgGs in the mixture were successfully generated, separated, and identified. Figure 2 illustrates these representative peptides for each subclass of IgGs with the exception of Panitumumab Hc with their respective CID mass spectra with *b*- and *y*-product ions assignment. For this latter chain, the representative peptide was present in the chromatographic run, but due to its low abundance and resulting poor fragmentation, it did not pass the database search threshold and is not illustrated here. Nevertheless, Sap9 proteolysis of Panitumumab Hc yielded a long peptide with a single CDR, Figure 2.

Amino Acid Substitution Analysis. The genetic recombination processes involved in the characteristic hypervariability of CDRs can eventually lead to single point mutations (i.e., substitutions of single bases in the DNA), potentially resulting in replacements of one amino acid residue with another within the polypeptidic chain. The Lys-to-Gln substitution can be caused by this process because the codons encoding for the two amino acids differ by a single base (Figure S2, Supporting Information). Figure 3 highlights the ability of the applied LC-MS/MS-based method to identify and localize a single amino acid substitution on a long peptide, wherein a Lys residue, typical of IgGs1, is substituted for Gln in the analyzed IgG2, Panitumumab. This modification produces a net mass change of 36 mDa/*z*, where *z* is the charge state of an ion, which is translated into a difference of only 7 mTh (or 0.007 *m*/*z*) for a 5+ precursor ion. However, both the *b* and *y* product ions (in this case b_{26} and y_{17}) are required for the localization of this modification. The localization of the substitution relies on MS/MS, whereas high-resolution and mass accuracy are required for product ion detection, considering the very small mass shift. This falls within the lowest tolerance window for product ion assignment allowed by the search algorithm (0.02 Da), leading to potential ambiguous peptide identification on product ion level. Note that the

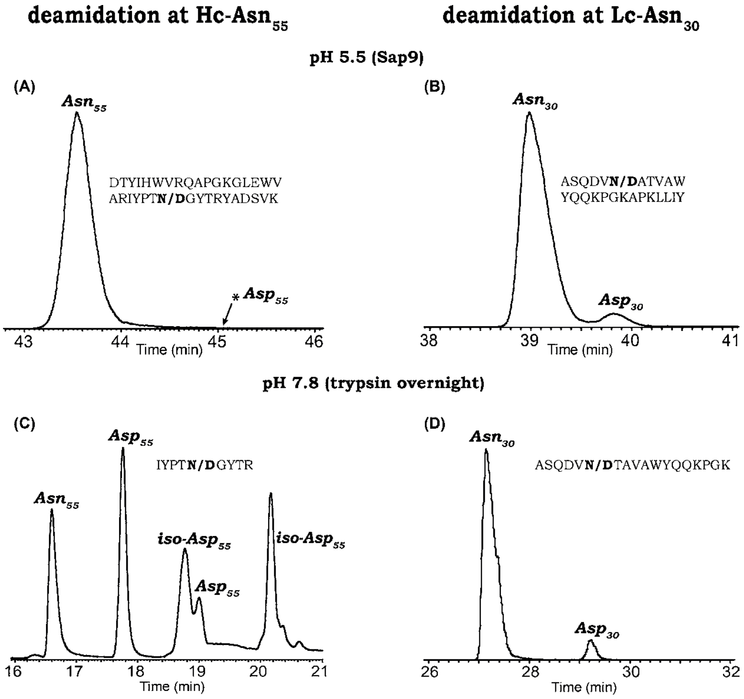


Figure 4. Assessment of Trastuzumab deamidation induced by different sample preparation/digestion protocols. Top panel: 1 h Sap9 at pH 5.5 following sample preparation Procedure II (A) Asn$_{55}$-Hc, (B) Asn$_{30}$–Lc; bottom panel: overnight trypsin at pH 7.8 (C) Asn$_{55}$-Hc, (D) Asn$_{30}$-Lc. * indicates deamidated peptide, corresponding to ~0.5% deamidation.

glutamine-modified peptide from Figure 3 was manually assigned and verified, as initial automatic assignment via the database search wrongly identified the lysine analogue. The manual assignment was possible not only due to the high-resolution detection of both *b*- and *y*-ion series, but also because the chromatographic elution of the two peptides is baseline separated under the applied conditions, and the 4:1 ratio between peak areas of the two corresponds to the presence of four IgGs1 sharing the Lys-containing peptide versus the single IgG2 with the Gln-peptide.

Deamidation Assessment of Trastuzumab. Trastuzumab contains two asparagine amino acid residues prone to deamidation in its CDRs. One is in the CDR1 of the Lc (Asn$_{30}$), and one in the CDR2 (Asn$_{55}$) of the Hc. Strong cation exchange chromatography (SCX) is generally employed to obtain a charge variant profile of IgGs. This profile generally contains a major peak corresponding to the most abundant proteoform of the IgG with peaks to its left corresponding to acidic proteoforms, whereas basic proteoforms elute after this main peak. Basic proteoforms can generally be attributed to the presence of the C-terminal lysine on one or both heavy chains and C-terminal proline amidation, whereas acidic charge proteoforms are generally attributed to Asn or Gln deamidation and N-terminal glutamine cyclization to form pyroglutamic acid. However, SCX alone does not allow the attribution and assignment of charged proteoforms in the IgG sequence. Peptide mapping with bottom-up LC-MS/MS is a powerful tool to sequence digested peptides and assess PTMs; however, common enzymes such as trypsin and GluC require basic pH and prolonged incubation times at 37 °C to achieve efficient digestion. This induces the deamidation of Asn rendering impossible the quantitatively accurate assignment of the endogenous deamidated asparagine amino acids. As shown in Figure 4, trypsin-digested Trastuzumab shows high deamidation rates with different deamidation products as confirmed by ETD MS/MS, vide infra. On the other hand, Sap9 digestion is performed at pH 5.5 over only 1 h (sample preparation Procedure II), which does not induce artifact deamidation, Figure 4, top panel. As an example, we can see that for the peptide-containing Asn$_{30}$ of Lc, we have two forms, a major nondeamidated form (92%) and a minor deamidated form of nearly 8%. The MS quantitation is in accordance with the results obtained by a more traditional technique (i.e., peak area calculation using the UV chromatographic trace) by Harris et al.[11]

As deamidation can lead to aspartic and isoaspartic acid formation, we employed ETD MS/MS to determine the isomer type of deamidated peptides, Figure 5. ETD MS/MS has been described to yield diagnostic ions differentiating aspartic and isoaspartic acids. These ions are a $z'-58$ Da and a $c^{\bullet}+58$ Da species for isoaspartic acid, as previously described by O'Connor and co-workers.[10] We found that with Sap9-based eBUP the deamidation at Asn$_{30}$ in the Lc resulted in formation of aspartic acid and at Asn$_{55}$ in the Hc in isoaspartic acid, whereas several deamidated forms for the tryptic peptides containing these deamidation sites were observed. Figure 5 shows an extracted ion chromatogram of the tryptic peptide containing Asn$_{55}$ of Hc. Although we have observed only one peak corresponding to the nondeamidated form, four different peaks corresponded to the deamidated forms. ETD-based LC-MS/MS showed that two of those corresponded to aspartic acid, and two others corresponded to isoaspartic acid. Generally, the deamidation of asparagine via a succinimide intermediate followed by partial hydrolysis of the succinimide ring leads to isoaspartate as the major product and aspartic acid as the minor product.[9,10] Alternative mechanisms by a nucleophilic attack without the succinimide intermediate leading to aspartic acid have also been

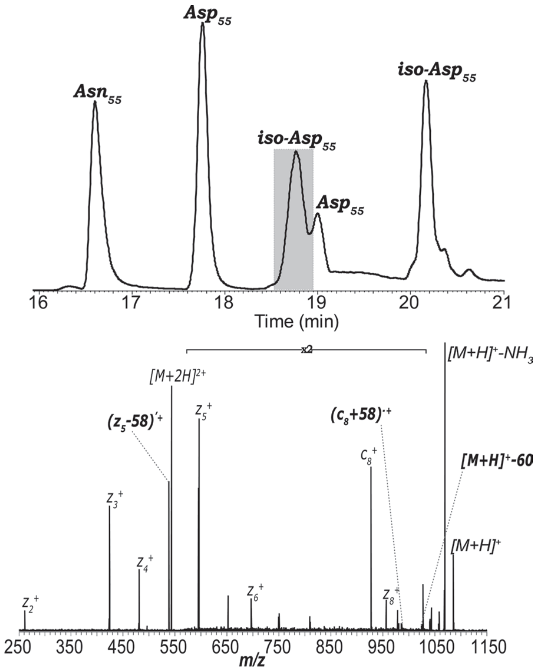


Figure 5. Differentiation between Asp and iso-Asp by ETD MS/MS. Top panel shows extracted base peak chromatogram of tryptic peptide IYPTNGYTR from Trastuzumab. Bottom panel illustrates ETD mass spectrum of doubly charged peptide containing iso-Asp from region highlighted in gray in chromatogram displayed above. Detected diagnostic c_8+58 radical and z_5−58 even-electron ions are indicated.

reported.[11,12] Here, for the tryptic peptide, we have observed two isomers eluting at different retention times. This is presumably due to L and D forms of the aspartate/isoaspartate residues. Finally, we also attempted to assess deamidation of Trastuzumab when it is present in IgG mixture, and we obtained comparable results (Figure S3, Supporting Information).

CONCLUSIONS

The primary consequence of the increased length of proteolytic peptides under eBUP conditions is that, with the application of high-resolution MS, 100% sequence coverage is obtained for both Hc and Lc in targeted analysis of a single IgG. The analysis of an IgG mixture (here composed of six IgGs from three subclasses) returned up to 99 and 100% sequence coverage for Hc and Lc, respectively. Importantly, these results were achieved in a single LC-MS/MS run preceded by a quick (1 h) Sap9 proteolysis at slightly acidic pH conditions. Results obtained using both sample preparation procedures (performed at pH 7.8 or pH 5.5) are comparable in terms of protein identification. Furthermore, it is noteworthy that Sap9 often generates peptides which include two consecutive CDRs, hence dramatically increasing the confidence in the IgG identification. This aspect is of particular importance for the analysis of IgG mixtures, such as a pool of polyclonal antibodies or a cocktail of monoclonal antibodies,[31] as the case of different IgGs sharing full sequence of one CDR domain is not rare, as exemplified here by CDR1 in light chains of Panitumumab and Bevacizumab. In summary, we optimized a quick and simple proteomics-grade pipeline that can be readily implemented in the current state-of-the-art proteomic setup. It enabled identification of modifications as small as near-isobaric single point amino acid variation and assessment of deamidation with results comparable to up to date figures of merit.

ASSOCIATED CONTENT

Supporting Information

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

AUTHOR INFORMATION

Corresponding Author

*E-mail: yury.tsybin@epfl.ch.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Authors thank Konstantin O. Zhurov for constructive discussions. We are grateful for financial support through the Swiss National Science Foundation (SNF project 200021_147006) and the European Research Council (ERC Starting Grant 280271 to Y.O.T.).

REFERENCES

(1) Walsh, G. *Nat. Biotechnol.* **2010**, *28*, 917−924.
(2) Beck, A.; Wagner-Rousset, E.; Ayoub, D.; Van Dorsselaer, A.; Sanglier-Cianferani, S. *Anal. Chem.* **2013**, *85*, 715−736.
(3) Diepold, K.; Bomans, K.; Wiedmann, M.; Zimmermann, B.; Petzold, A.; Schlothauer, T.; Mueller, R.; Moritz, B.; Stracke, J. O.; Molhoj, M.; Reusch, D.; Bulau, P. *PLoS One* **2012**, *7*, e30295.
(4) Kroon, D. J.; Baldwinferro, A.; Lalan, P. *Pharm. Res.* **1992**, *9*, 1386−1393.
(5) Harris, R. J.; Murnane, A. A.; Utter, S. L.; Wagner, K. L.; Cox, E. T.; Polastri, G. D.; Helder, J. C.; Sliwkowski, M. B. *Biotechnology* **1993**, *11*, 1293−1297.
(6) Fornelli, L.; Damoc, E.; Thomas, P. M.; Kelleher, N. L.; Aizikov, K.; Denisov, E.; Makarov, A.; Tsybin, Y. O. *Mol. Cell. Proteomics* **2012**, *11*, 1758−1767.
(7) Tsybin, Y. O.; Fornelli, L.; Stoermer, C.; Luebeck, M.; Parra, J.; Nallet, S.; Wurm, F. M.; Hartmer, R. *Anal. Chem.* **2011**, *83*, 8919−8927.
(8) Mao, Y.; Valeja, S. G.; Rouse, J. C.; Hendrickson, C. L.; Marshall, A. G. *Anal. Chem.* **2013**, *85*, 4239−4246.
(9) Yang, H.; Zubarev, R. A. *Electrophoresis* **2010**, *31*, 1764−1772.
(10) O'Connor, P. B.; Cournoyer, J. J.; Pitteri, S. J.; Chrisman, P. A.; McLuckey, S. A. *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 15−19.
(11) Harris, R. J.; Kabakoff, B.; Macchi, F. D.; Shen, F. J.; Kwong, M.; Andya, J. D.; Shire, S. J.; Bjork, N.; Totpal, K.; Chen, A. B. *J. Chromatogr. B: Biomed. Sci. Appl.* **2001**, *752*, 233−245.
(12) Yan, B. X.; Steen, S.; Hambly, D.; Valliere-Douglass, J.; Bos, T. V.; Smallwood, S.; Yates, Z.; Arroll, T.; Han, Y. H.; Gadgil, H.; Latypov, R. F.; Wallace, A.; Lim, A.; Kleemann, G. R.; Wang, W. C.; Balland, A. *J. Pharm. Sci.* **2009**, *98*, 3509−3521.
(13) Zhurov, K. O.; Fornelli, L.; Wodrich, M. D.; Laskay, U. A.; Tsybin, Y. O. *Chem. Soc. Rev.* **2013**, *42*, 5014−5030.
(14) Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W. *J. Am. Chem. Soc.* **1998**, *120*, 3265−3266.
(15) Syka, J. E. P.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 9528−9533.
(16) Cournoyer, J. J.; Pittman, J. L.; Ivleva, V. B.; Fallows, E.; Waskell, L.; Costello, C. E.; O'Connor, P. B. *Protein Sci.* **2005**, *14*, 452−463.
(17) Li, X.; Lin, C.; O'Connor, P. B. *Anal. Chem.* **2010**, *82*, 3606−3615.
(18) Yang, H.; Fung, E. Y.; Zubarev, A. R.; Zubarev, R. A. *J. Proteome Res.* **2009**, *8*, 4615−4621.
(19) Ayoub, D.; Jabs, W.; Resemann, A.; Evers, W.; Evans, C.; Main, L.; Baessmann, C.; Wagner-Rousset, E.; Suckau, D.; Beck, A. *mAbs* **2013**, *5*, 699−710.
(20) Fornelli, L.; Ayoub, D.; Aizikov, K.; Beck, A.; Tsybin, Y. O. *Anal. Chem.* **2014**, *86*, 3005−3012.

(21) Shaw, J. B.; Li, W.; Holden, D. D.; Zhang, Y.; Griep-Raming, J.; Fellers, R. T.; Early, B. P.; Thomas, P. M.; Kelleher, N. L.; Brodbelt, J. S. *J. Am. Chem. Soc.* **2013**, *135*, 12646−12651.

(22) Laskay, U. A.; Lobas, A. A.; Srzentic, K.; Gorshkov, M. V.; Tsybin, Y. O. *J. Proteome Res.* **2013**, *12*, 5558−5569.

(23) Laskay, U. A.; Srzentic, K.; Monod, M.; Tsybin, Y. O. *J. Proteomics* **2014**, *110*, 20−31.

(24) Albrecht, A.; Felk, A.; Pichova, I.; Naglik, J. R.; Schaller, M.; de Groot, P.; Maccallum, D.; Odds, F. C.; Schafer, W.; Klis, F.; Monod, M.; Hube, B. *J. Biol. Chem.* **2006**, *281*, 688−694.

(25) Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M. *Anal. Chem.* **1996**, *68*, 850−858.

(26) Laskay, U. A.; Srzentic, K.; Fornelli, L.; Upir, O.; Kozhinov, A. N.; Monod, M.; Tsybin, Y. O. *Chimia* **2013**, *67*, 244−249.

(27) Larsen, P. A.; Smith, T. P. *BMC Immunol.* **2012**, *13*, 52.

(28) Barrios, Y.; Jirholt, P.; Ohlin, M. *J. Mol. Recognit.* **2004**, *17*, 332−338.

(29) Frese, C. K.; Altelaar, A. F. M.; van den Toorn, H.; Nolting, D.; Griep-Raming, J.; Heck, A. J. R.; Mohammed, S. *Anal. Chem.* **2012**, *84*, 9668−9673.

(30) Since Sap9 does not exhibit prevalence for any particular residue surrounding the scissile bond number of missed cleavages is impossible to deduce, and hence, the number of unique peptides given in this table is a count of longest peptides, where peptides with overlapping sequence, regardless of their uniqueness, were not accounted for. Sequence coverage, on the other hand, is calculated as an average of all identified peptides in a chromatographic run, for all proteolytic experiments.

(31) Logtenberg, T. *Trends Biotechnol.* **2007**, *25*, 390−394.

Supporting Information.

Advantages of extended bottom-up proteomics using Sap9 for analysis of monoclonal antibodies

Kristina Srzentić[1], Luca Fornelli[1], Ünige A. Laskay[1], Michel Monod[2], Alain Beck[3], Daniel Ayoub[1], and Yury O. Tsybin[1*]

[1] Biomolecular Mass Spectrometry Laboratory, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

[2] Department of Dermatology, Centre Hospitalier Universitaire Vaudois, 1011 Lausanne, Switzerland

[3] Centre d'Immunologie Pierre Fabre, 74160 St. Julien-en-Genevois, France

Table of SI contents:

1. Experimental methods details: desalting protocol.
2. Figure S1. Sequence alignment of light and heavy chains of IgGs.
3. Figure S2. Sequence coverage of light and heavy chains of IgGs.
4. Figure S3. Comparison of Trastuzumab deamidation induced in a six IgG mixture by different sample preparation/digestion protocols.

Experimental methods details: pooled C4-C18 ZipTip desalting protocol.

Materials

- Resin conditioning solution: MeOH
- Wash solution: 0.1% TFA in H_20
- Peptide elution solution: 80% ACN/20% H_20/0.1% FA (v/v/v)
-

Procedure (recommended sample volume 10 µl)

Start with C4 ZipTip cartridge as follows:

1. Condition - Aspirate 10 µl of conditioning solution and discard to waste; repeat five times
   - Aspirate 10 µl of wash solution and discard to waste; repeat five times

2. Load - Slowly aspirate 10 µl of sample and expel the liquid back into the tube; aspirate-expel 15 times in the tube

3. Wash - Aspirate 10 µl of wash solution and expel back into the **sample tube**; repeat 5 times (*note*: this will increase the final volume in the sample tube, vortex prior to C18 procedure)
4. Elute - Aspirate 5 µl of wash solution and expel into the **new tube**; aspirate-expel 15 times in the tube

Repeat steps 1-4 with C18 ZipTip cartridge, with step 3 modified as follows:

3* Aspirate 10 µl of wash solution and **discard to waste**; repeat 5 times

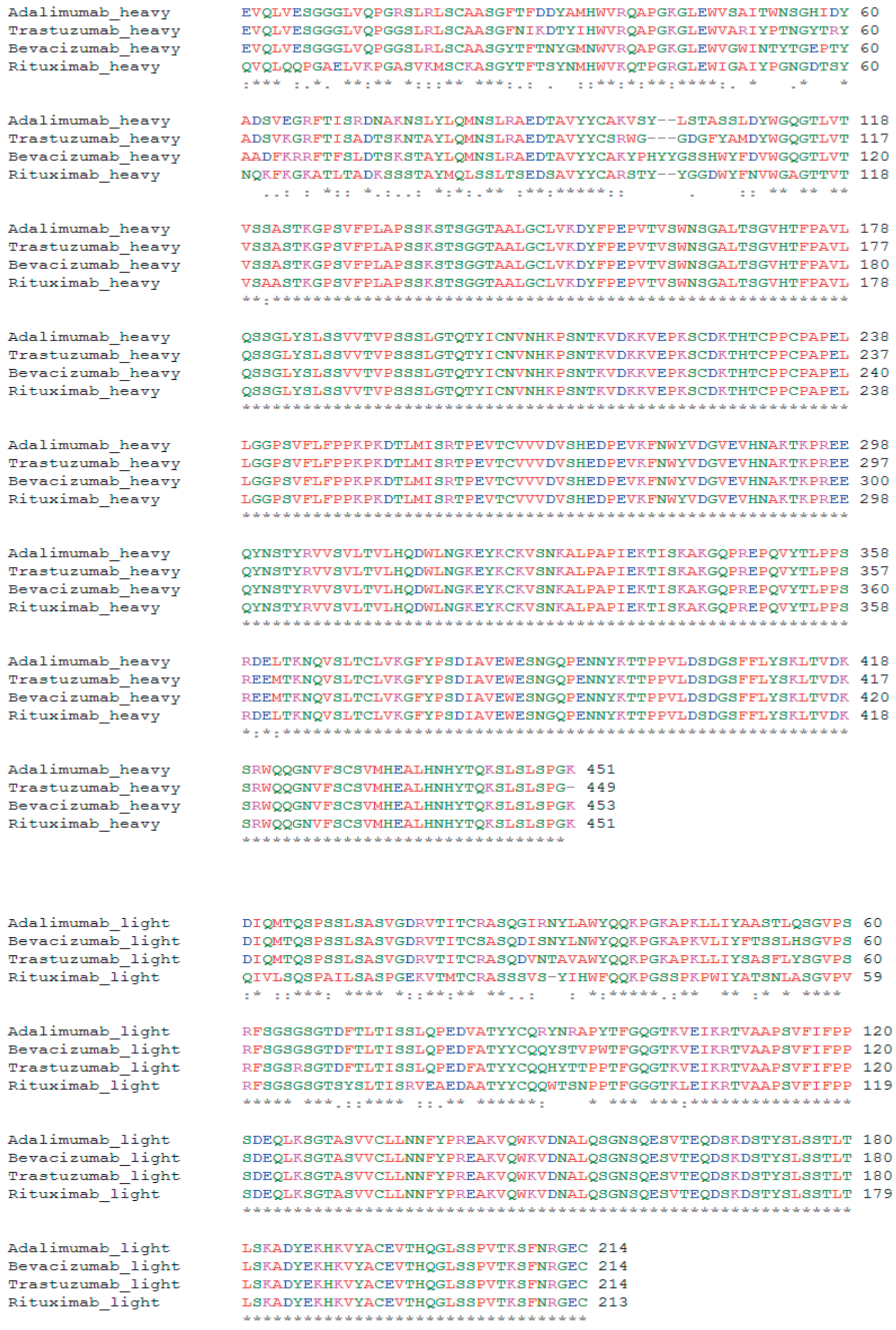
Pool fractions together and dilute with 0.1%FA to final 10% ACN

Alternatively: mix with 5% ACN/0.1% FA and inject directly on RP nano C8 column

Figure S1. Sequence alignment of light and heavy chains for a) all IgG1 present in the mixture, b) IgG1 (Adalimumab), IgG2 (Panitumumab) and IgG4 (Natalizumab). Residues are colored according to their physicochemical properties. "*" indicates positions which have a single, fully conserved residue, ":" indicates conservation between groups of strongly similar properties,

"." indicates conservation between groups of weakly similar properties.

a)



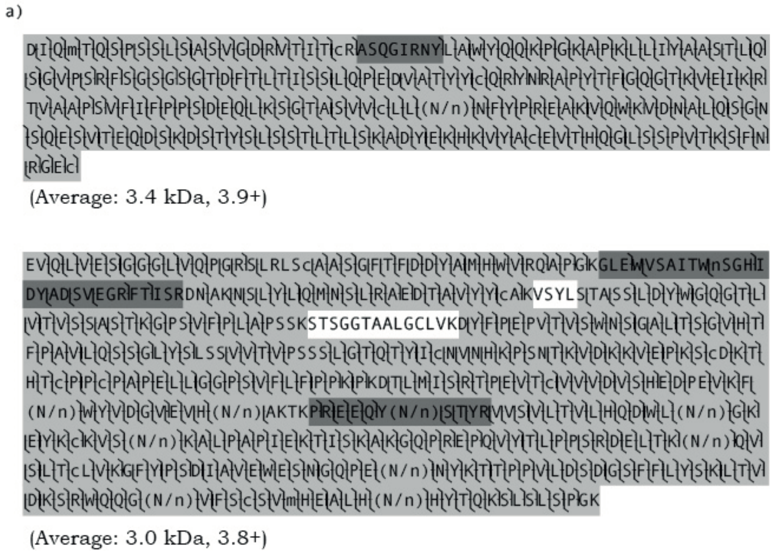
```
Adalimumab_heavy        EVQLVESGGGLVQPGRSLRLSCAASGFTFDDYAMHWVRQAPGKGLEWVSAITWNSGHIDY 60
Trastuzumab_heavy       EVQLVESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKGLEWVARIYPTNGYTRY 60
Bevacizumab_heavy       EVQLVESGGGLVQPGGSLRLSCAASGYTFTNYGMNWVRQAPGKGLEWVGWINTYTGEPTY 60
Rituximab_heavy         QVQLQQPGAELVKPGASVKMSCKASGYTFTSYNMHWVKQTPGRGLEWIGAIYPGNGDTSY 60
                        :*** :.*. **:** *:::** ***:.: .  ::**:*:**:****:. *   .*   *

Adalimumab_heavy        ADSVEGRFTISRDNAKNSLYLQMNSLRAEDTAVYYCAKVSY--LSTASSLDYWGQGTLVT 118
Trastuzumab_heavy       ADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYYCSRWG---GDGFYAMDYWGQGTLVT 117
Bevacizumab_heavy       AADFKRRFTFSLDTSKSTAYLQMNSLRAEDTAVYYCAKYPHYYGSSHWYFDVWGQGTLVT 120
Rituximab_heavy         NQKFKGKATLTADKSSSTAYMQLSSLTSEDSAVYYCARSTY--YGGDWYFNVWGAGTTVT 118
                          ..: : *:: *.:..: *:*:.** :**:*****::      .    :: ** ** **

Adalimumab_heavy        VSSASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVL 178
Trastuzumab_heavy       VSSASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVL 177
Bevacizumab_heavy       VSSASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVL 180
Rituximab_heavy         VSAASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVL 178
                        **:*********************************************************

Adalimumab_heavy        QSSGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKVDKKVEPKSCDKTHTCPPCPAPEL 238
Trastuzumab_heavy       QSSGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKVDKKVEPKSCDKTHTCPPCPAPEL 237
Bevacizumab_heavy       QSSGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKVDKKVEPKSCDKTHTCPPCPAPEL 240
Rituximab_heavy         QSSGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKVDKKVEPKSCDKTHTCPPCPAPEL 238
                        ************************************************************

Adalimumab_heavy        LGGPSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEVKFNWYVDGVEVHNAKTKPREE 298
Trastuzumab_heavy       LGGPSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEVKFNWYVDGVEVHNAKTKPREE 297
Bevacizumab_heavy       LGGPSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEVKFNWYVDGVEVHNAKTKPREE 300
Rituximab_heavy         LGGPSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEVKFNWYVDGVEVHNAKTKPREE 298
                        ************************************************************

Adalimumab_heavy        QYNSTYRVVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQPREPQVYTLPPS 358
Trastuzumab_heavy       QYNSTYRVVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQPREPQVYTLPPS 357
Bevacizumab_heavy       QYNSTYRVVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQPREPQVYTLPPS 360
Rituximab_heavy         QYNSTYRVVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQPREPQVYTLPPS 358
                        ************************************************************

Adalimumab_heavy        RDELTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYKTTPPVLDSDGSFFLYSKLTVDK 418
Trastuzumab_heavy       REEMTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYKTTPPVLDSDGSFFLYSKLTVDK 417
Bevacizumab_heavy       REEMTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYKTTPPVLDSDGSFFLYSKLTVDK 420
Rituximab_heavy         RDELTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYKTTPPVLDSDGSFFLYSKLTVDK 418
                        *:*:********************************************************

Adalimumab_heavy        SRWQQGNVFSCSVMHEALHNHYTQKSLSLSPGK 451
Trastuzumab_heavy       SRWQQGNVFSCSVMHEALHNHYTQKSLSLSPG- 449
Bevacizumab_heavy       SRWQQGNVFSCSVMHEALHNHYTQKSLSLSPGK 453
Rituximab_heavy         SRWQQGNVFSCSVMHEALHNHYTQKSLSLSPGK 451
                        ********************************


Adalimumab_light        DIQMTQSPSSLSASVGDRVTITCRASQGIRNYLAWYQQKPGKAPKLLIYAASTLQSGVPS 60
Bevacizumab_light       DIQMTQSPSSLSASVGDRVTITCSASQDISNYLNWYQQKPGKAPKVLIYFTSSLHSGVPS 60
Trastuzumab_light       DIQMTQSPSSLSASVGDRVTITCRASQDVNTAVAWYQQKPGKAPKLLIYSASFLYSGVPS 60
Rituximab_light         QIVLSQSPAILSASPGEKVTMTCRASSSVS-YIHWFQQKPGSSPKPWIYATSNLASGVPV 59
                        :* ::***: **** *::**:** **..:   : *:*****.:**  ** :* * ****

Adalimumab_light        RFSGSGSGTDFTLTISSLQPEDVATYYCQRYNRAPYTFGQGTKVEIKRTVAAPSVFIFPP 120
Bevacizumab_light       RFSGSGSGTDFTLTISSLQPEDFATYYCQQYSTVPWTFGQGTKVEIKRTVAAPSVFIFPP 120
Trastuzumab_light       RFSGSRSGTDFTLTISSLQPEDFATYYCQQHYTTPPTFGQGTKVEIKRTVAAPSVFIFPP 120
Rituximab_light         RFSGSGSGTSYSLTISRVEAEDAATYYCQQWTSNPPTFGGGTKLEIKRTVAAPSVFIFPP 119
                        ***** ***.::**** ::.** ******:    * *** ***:****************

Adalimumab_light        SDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLT 180
Bevacizumab_light       SDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLT 180
Trastuzumab_light       SDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLT 180
Rituximab_light         SDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLT 179
                        ************************************************************

Adalimumab_light        LSKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC 214
Bevacizumab_light       LSKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC 214
Trastuzumab_light       LSKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC 214
Rituximab_light         LSKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC 213
                        **********************************
```

b)

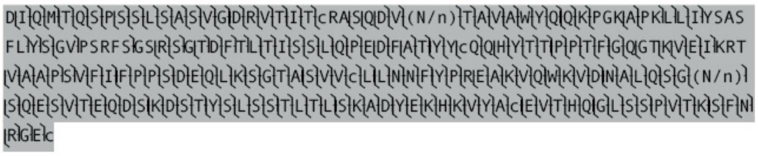
```
Adalimumab_heavy        EVQLVESGGGLVQPGRSLRLSCAASG--FTFDDYAMHWVRQAPGKGLEWVSAITWNSGHI 58
Panitumumab_heavy       QVQLQESGPGLVKPSETLSLTCTVSGGSVSSGDYYWTWIRQSPGKGLEWIGHIYY-SGNT 59
Natalizumab_heavy       QVQLVQSGAEVKKPGASVKVSCKASG--FNIKDTYIHWVRQAPGQRLEWMGRIDPANGYT 58
                        :*** :**  : :*. :: ::* .**  ..  *    *:**:**: ***:. *   .*

Adalimumab_heavy        DYADSVEGRFTISRDNAKNSLYLQMNSLRAEDTAVYYCAKVSYLS--TASSLDYWGQGTL 116
Panitumumab_heavy       NYNPSLKSRLTISIDTSKTQFSLKLSSVTAADTAIYYCVRDR-----VTGAFDIWGQGTM 114
Natalizumab_heavy       KYDPKFQGRVTITADTSASTAYMELSSLRSEDTAVYYCAREGYYGNYGVYAMDYWGQGTL 118
                        .*  ..:.*.**: *.: .   :::.*: : ***:***.:        . ::* *****:

Adalimumab_heavy        VTVSSASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPA 176
Panitumumab_heavy       VTVSSASTKGPSVFPLAPCSRSTSESTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPA 174
Natalizumab_heavy       VTVSSAKTTGPSVFPLAPCSRSTSESTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPA 178
                        ******.*.*********.*:*** .**********************************

Adalimumab_heavy        VLQSSGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKVDKKVEPKSCDKTHTCPPCPAP 236
Panitumumab_heavy       VLQSSGLYSLSSVVTVPSSNFGTQTYTCNVDHKPSNTKVDKTVERKCC---VECPPCPAP 231
Natalizumab_heavy       VLQSSGLYSLSSVVTVPSSSLGTKTYTCNVDHKPSNTKVDKRVESKYG---PPCPSCPAP 235
                        *******************.:**:** ***:********** ** *       **.****

Adalimumab_heavy        ELLGGPSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEVKFNWYVDGVEVHNAKTKPR 296
Panitumumab_heavy       PVAG-PSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEVQFNWYVDGVEVHNAKTKPR 290
Natalizumab_heavy       EFLGGPSVFLFPPKPKDTLMISRTPEVTCVVVDVSQEDPEVQFNWYVDGVEVHNAKTKPR 295
                         . * ******************************:*****:******************

Adalimumab_heavy        EEQYNSTYRVVSVLTVLHQDWLNGKEYKCKVSNKALPAPIEKTISKAKGQPREPQVYTLP 356
Panitumumab_heavy       EEQFNSTFRVVSVLTVVHQDWLNGKEYKCKVSNKGLPAPIEKTISKTKGQPREPQVYTLP 350
Natalizumab_heavy       EEQFNSTYRVVSVLTVLHQDWLNGKEYKCKVSNKGLPSSIEKTISKAKGQPREPQVYTLP 355
                        ***:***:********:*****************.**:.*******:*************

Adalimumab_heavy        PSRDELTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYKTTPPVLDSDGSFFLYSKLTV 416
Panitumumab_heavy       PSREEMTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYKTTPPMLDSDGSFFLYSKLTV 410
Natalizumab_heavy       PSQEEMTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYKTTPPVLDSDGSFFLYSRLTV 415
                        **::*:**************************************:***********:***

Adalimumab_heavy        DKSRWQQGNVFSCSVMHEALHNHYTQKSLSLSPGK 451
Panitumumab_heavy       DKSRWQQGNVFSCSVMHEALHNHYTQKSLSLSPGK 445
Natalizumab_heavy       DKSRWQEGNVFSCSVMHEALHNHYTQKSLSLSLGK 450
                        ******:************************* **
```

```
Adalimumab_light        DIQMTQSPSSLSASVGDRVTITCRASQGIRNYLAWYQQKPGKAPKLLIYAASTLQSGVPS 60
Panitumumab_light       DIQMTQSPSSLSASVGDRVTITCQASQDISNYLNWYQQKPGKAPKLLIYDASNLETGVPS 60
Natalizumab_light       DIQMTQSPSSLSASVG-RVTITCKTSQDINKYMAWYQQTPGKAPRLLIHYTSALQPGIPS 59
                        **************** ******::**.* :*: ****.*****:***: :* *:.*:**

Adalimumab_light        RFSGSGSGTDFTLTISSLQPEDVATYYCQRYNRAPYTFGQGTKVEIKRTVAAPSVFIFPP 120
Panitumumab_light       RFSGSGSGTDFTFTISSLQPEDIATYFCQHFDHLPLAFGGGTKVEIKRTVAAPSVFIFPP 120
Natalizumab_light       RFSGSGSGRDYTFTISSLQPEDIATYYCLQYDNL-WTFGQGTKVEIKRTVAAPSVFIFPP 118
                        ******** *:*:*********:***:* :::.   :** ********************

Adalimumab_light        SDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLT 180
Panitumumab_light       SDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLT 180
Natalizumab_light       SDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLT 178
                        ************************************************************

Adalimumab_light        LSKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC 214
Panitumumab_light       LSKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC 214
Natalizumab_light       LSKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC 212
                        **********************************
```

Figure S2. Sequence coverage of light (top) and heavy (bottom) chains of a) Adalimumab, b) Trastuzumab, c) Bevacizumab, d) Panitumumab, e) Natalizumab and f) Retuximab obtained from single LC-MS/MS analysis of IgGs mixture. N-terminal *b*-ions and C-terminal *y*-ions obtained by CID and HCD MS/MS are indicated on the fragmentation map. Fragments of all observed charge states were assigned based on the analysis of MS/MS spectra for all identified peptides that triggered data-dependent MS/MS acquisition. Identified peptides and sequenced regions are indicated in light gray (CID) and dark gray (only identified by HCD). Peptides carrying pyroglutamic acid formed from Glutamine (Q) PTMs are shown in red. N/n indicates sites where non-deamidated and deamidated asparagine were confirmed by manual analysis.

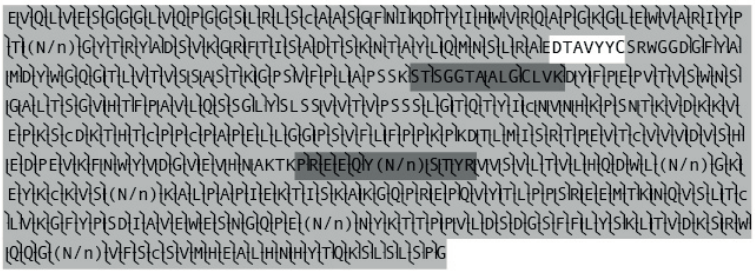

a)
DIQmTQSPSSLSASVGDRVTITcRASQGIRNYLAWYQQKPGKAPKLLIYAASTLQ
SGVPSRFSGSGSGTDFTLTISSLQPEDVATYYcQRYNRAPYTFGQGTKVEIKR
TVAAPSVFIFPPSDEQLKSGTASVVcLL(N/n)NFYPREAKVQWKVDNALQSGN
SQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYAcEVTHQGLSSPVTKSFN
RGEc
(Average: 3.4 kDa, 3.9+)
EVQLVESGGGLVQPGRSLRLScAASGFTFDDYAMHWVRQAPGKGLEWVSAITWnSGHI
DYADSVEGRFTISRDNAKNSLYLQMNSLRAEDTAVYYcAKVSYLSTASSLDYWGQGTL
VTVSSASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHT
FPAVLQSSGLYSLSSVVTVPSSSLGTQTYIcNVNHKPSNTKVDKKVEPKScDKT
HTcPPcPAPELLGGPSVFLFPPKPKDTLMISRTPEVTcVVVDVSHEDPEVKF
(N/n)WYVDGVEVH(N/n)AKTKPREEQY(N/n)STYRVVSVLTVLHQDWL(N/n)GK
EYKcKVS(N/n)KALPAPIEKTISKAKGQPREPQVYTLPPSRDELTK(N/n)QV
SLTcLVKGFYPSDIAVEWESNGQPE(N/n)NYKTTPPVLDSDGSFFLYSKLTV
DKSRWQQG(N/n)VFScSVmHEALH(N/n)HYTQKSLSLSPGK
(Average: 3.0 kDa, 3.8+)

b)



DIQMTQSPSSLSASVGDRVTITcRASQDV(N/n)TAVAWYQQKPGKAPKLLIYSAS
FLYSGVPSRFSGSRSGTDFTLTISSLQPEDFATYYcQQHYTTPPTFGQGTKVEIKRT
VAAPSVFIFPPSDEQLKSGTASVVcLLNNFYPREAKVQWKVDNALQSG(N/n)
SQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYAcEVTHQGLSSPVTKSFN
RGEc

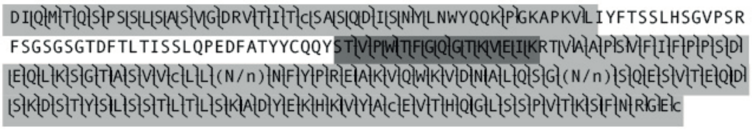
(Average: 3.5 kDa, 4.0+)



EVQLVESGGGLVQPGGSLRLScAASGFNIKDTYIHWVRQAPGKGLEWVARIYP
T(N/n)GYTRYADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYYCSRWGGDGFYA
MDYWGQGTLVTVSSASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNS
GALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYIcNVNHKPSNTKVDKKV
EPKScDKTHTcPPcPAPELLGGPSVFLFPPKPKDTLMISRTPEVTcVVVDVSH
EDPEVKFNWYVDGVEVHNAKTKPREEQY(N/n)STYRVVSVLTVLHQDWL(N/n)GK
EYKcKVS(N/n)KALPAPIEKTISKAKGQPREPQVYTLPPSREEMTKNQVSLTc
LVKGFYPSDIAVEWESNGQPE(N/n)NYKTTPPVLDSDGSFFLYSKLTVDKSRW
QQG(N/n)VFScSVMHEALHNHYTQKSLSLSPG

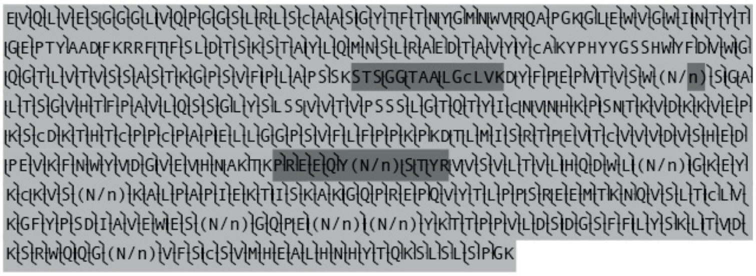
(Average: 3.3 kDa, 4.2+)

c)



DIQMTQSPSSLSASVGDRVTITcSASQDISNYLNWYQQKPGKAPKVLIYFTSSLHSGVPSR
FSGSGSGTDFTLTISSLQPEDFATYYCQQYSTVPWTFGQGTKVEIKRTVAAPSVFIFPPSD
EQLKSGTASVVcLL(N/n)NFYPREAKVQWKVDNALQSG(N/n)SQESVTEQD
SKDSTYSLSSTLTLSKADYEKHKVYAcEVTHQGLSSPVTKSFNRGEc

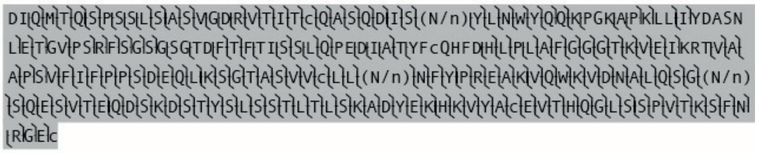
(Average: 3.4 kDa, 4.0+)



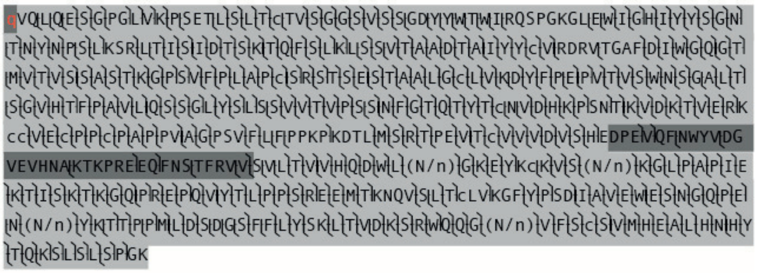
EVQLVESGGGLVQPGGSLRLScAASGYTFTNYGMNWVRQAPGKGLEWVGWINTYT
GEPTYAADFKRRFTFSLDTSKSTAYLQMNSLRAEDTAVYYcAKYPHYYGSSHWYFDVWG
QGTLVTVSSASTKGPSVFPLAPSSKSTSGGTAALGcLVKDYFPEPVTVSW(N/n)SGA
LTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYIcNVNHKPSNTKVDKKVEP
KScDKTHTcPPcPAPELLGGPSVFLFPPKPKDTLMISRTPEVTcVVVDVSHED
PEVKFNWYVDGVEVHNAKTKPREEQY(N/n)STYRVVSVLTVLHQDWL(N/n)GKEY
KcKVS(N/n)KALPAPIEKTISKAKGQPREPQVYTLPPSREEMTKNQVSLTcLV
KGFYPSDIAVEWES(N/n)GQPE(N/n)(N/n)YKTTPPVLDSDGSFFLYSKLTVD
KSRWQQG(N/n)VFScSVMHEALHNHYTQKSLSLSPGK

(Average: 3.3 kDa, 4.2+)

d)

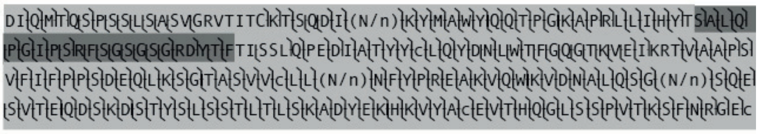

DIQMTQSPSSLSASVGDRVTITcQASQDIS(N/n)YLNWYQQKPGKAPKLLIYDASN
LETGVPSRFSGSGSGTDFTFTISSLQPEDIATYFcQHFDHLPLAFGGGTKVEIKRTVA
APSVFIFPPSDEQLKSGTASVVcLL(N/n)NFYPREAKVQWKVDNALQSG(N/n)
SQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYAcEVTHQGLSSPVTKSFN
RGEc

(Average: 3.5 kDa, 4.1+)

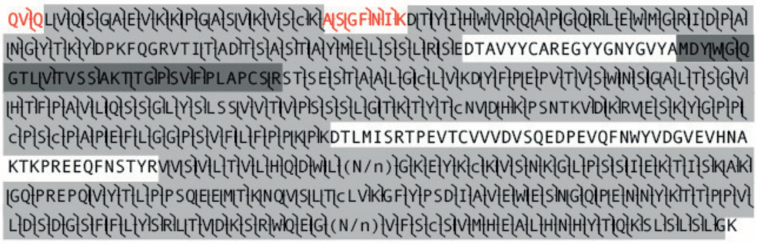

qVQLQESGPGLVKPSETLSLTcTVSGGSVSSGDYYWTWIRQSPGKGLEWIGHIYYSGN
TNYNPSLKSRLTISIDTSKTQFSLKLSSVTAADTAIYYcVRDRVTGAFDIWGQGT
MVTVSSASTKGPSVFPLAPcSRSTSESTAALGcLVKDYFPEPVTVSWNSGALT
SGVHTFPAVLQSSGLYSLSSVVTVPSSNFGTQTYTcNVDHKPSNTKVDKTVERK
ccVEcPPcPAPPVAGPSVFLFPPKPKDTLMSRTPEVTcVVVDVSHEDPEVQFNWYVDG
VEVHNAKTKPREEQFNSTFRVVSVLTVVHQDWL(N/n)GKEYKcKVS(N/n)KGLPAPIE
KTISKTKGQPREPQVYTLPPSREEMTKNQVSLTcLVKGFYPSDIAVEWESNGQPE
N(N/n)YKTTPPMLDSDGSFFLYSKLTVDKSRWQQG(N/n)VFScSVMHEALHNHY
TQKSLSLSPGK

(Average: 3.0 kDa, 4.0+)

e)

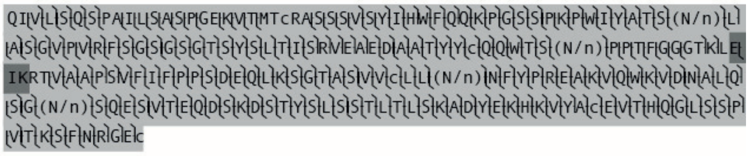

DIQMTQSPSSLSASVGRVTITCKTSQDI(N/n)KYMAWYQQTPGKAPRLLIHYTSALQ
PGIPSRFSGSGSGRDYTFTISSLQPEDIATYYcLQYDNLWTFGQGTKVEIKRTVAAPS
VFIFPPSDEQLKSGTASVVcLL(N/n)NFYPREAKVQWKVDNALQSG(N/n)SQE
SVTEQDSKDSTYSLSSTLTLSKADYEKHKVYAcEVTHQGLSSPVTKSFNRGEc

(Average: 3.3 kDa, 3.9+)

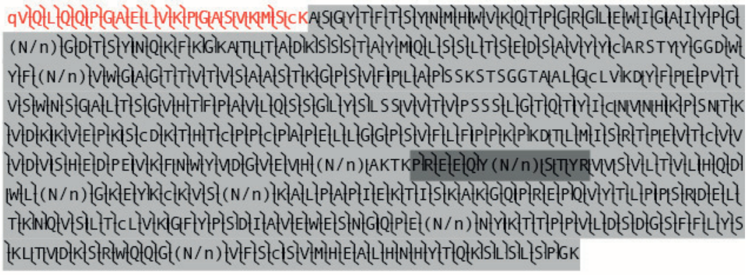

QVQLVQSGAEVKKPGASVKVScKASGFNIKDTYIHWVRQAPGQRLEWMGRIDPA
NGYTKYDPKFQGRVTITADTSASTAYMELSSLRSEDTAVYYCAREGYYGNYGVYAMDYWGQ
GTLVTVSSAKTTGPSVFPLAPCSRSTSESTAALGcLVKDYFPEPVTVSWNSGALTSGV
HTFPAVLQSSGLYSLSSVVTVPSSSLGTKTYTcNVDHKPSNTKVDKRVESKYGPP
cPScPAPEFLGGPSVFLFPPKPKDTLMISRTPEVTCVVVDVSQEDPEVQFNWYVDGVEVHNA
KTKPREEQFNSTYRVVSVLTVLHQDWL(N/n)GKEYKcKVSNKGLPSSIEKTISKAK
GQPREPQVYTLPPSQEEMTKNQVSLTcLVKGFYPSDIAVEWESNGQPENNYKTTPPV
LDSDGSFFLYSRLTVDKSRWQEG(N/n)VFScSVMHEALHNHYTQKSLSLSLGK

(Average: 2.7 kDa, 3.6+)

f)

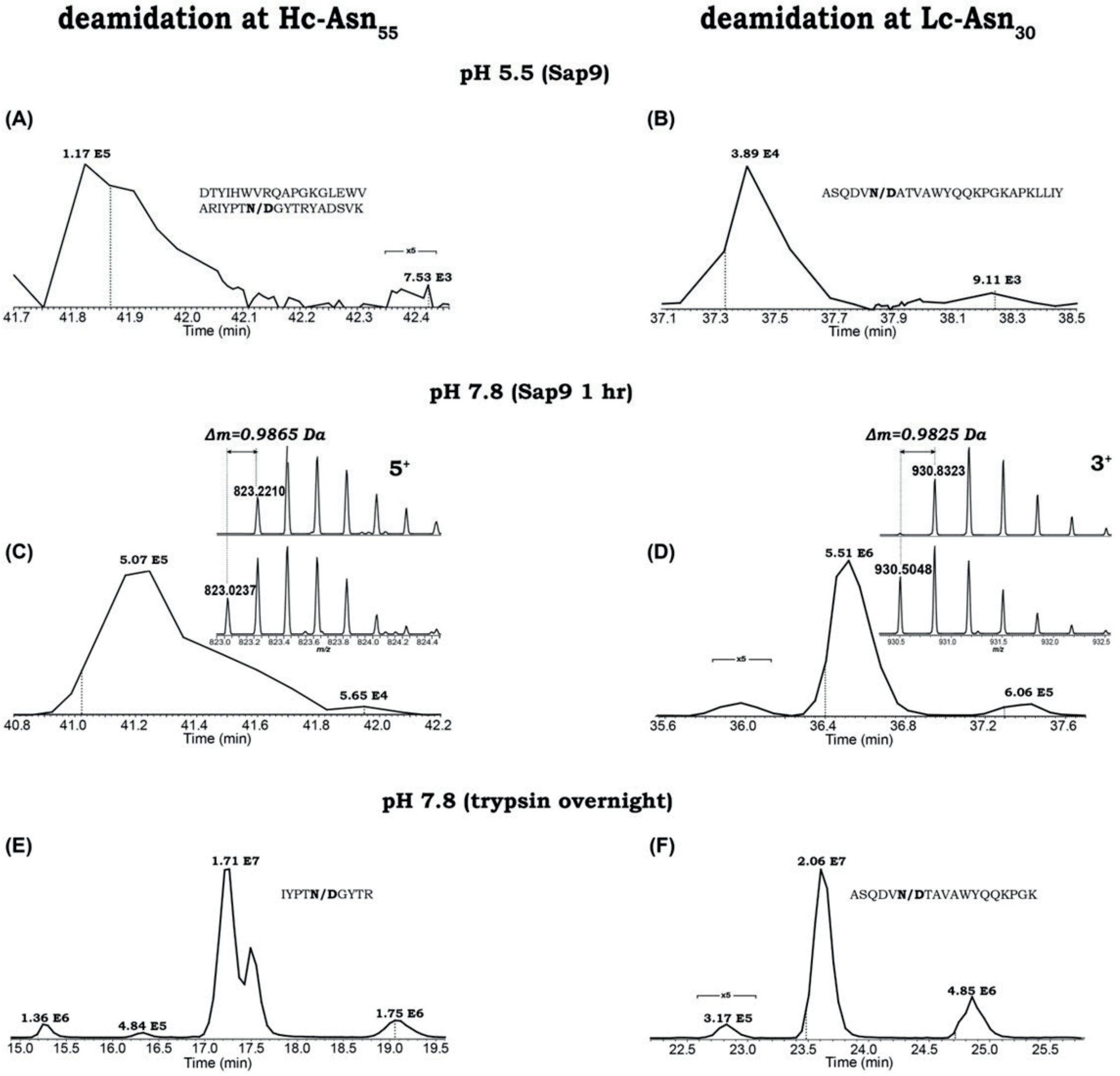


(Average: 3.5 kDa, 4.1+)



(Average: 3.1 kDa, 3.9+)

Figure S3. Comparison of Trastuzumab deamidation induced in a six IgG mixture by different sample preparation/digestion protocols. Top panel: 1 hr Sap9 at pH 5.5, sample preparation Procedure II (A) Asn_{55}-Hc, (B) Asn_{30}–Lc; middle panel: 1 hr Sap9 at pH 5.5 but sample preparation performed at pH 7.8 following Procedure I (C) Asn_{55}-Hc, (D) Asn_{30}-Lc; bottom panel: overnight trypsin at pH 7.8 (E) Asn_{55}-Hc, (F) Asn_{30}-Lc.



Chapter 6. An alternative to enzymatic digestion: Chemoselective MD approach

Extreme proteome complexity requires development of approaches complementary to bottom-up proteomics that analyzes mixtures of proteolytic (enzymatic) peptides lighter than 3 kDa and top down proteomics that targets intact proteins and protein fragments heavier than 15 kDa. Improvements in high resolution mass spectrometry (MS) are currently heading proteomics towards application of middle-down (MD) approaches and allow their implementation for proteome characterization using heavy, 3-15 kDa, proteolytic peptides. So far, MD has been limited by the absence of a highly specific protease generating large (>3 kDa) peptides. Chemical reagents may be powerful alternatives to enzymes for obtaining heavy (3-7, 7-15 kDa) peptides. We therefore considered the use of chemicals historically known in biochemistry for cleaving polypeptide chains with elevated amino acid specificity. Particularly, we chose to target single and less frequent amino acid residues (methionine, cysteine and tryptophane). We sought to optimize chemical-based MD proteomics platform at four levels: digestion protocols, by reducing duration and maximizing efficiency of proteolysis; peptide fractionation, for simplified MS analysis; development of liquid chromatography-MS strategies, and search algorithms tailored for large peptides with specific cleavage signatures.

Hereinafter reported considerations are the initial premises later translated into sorted results for a targeted application and a research article and enclosed at the end of this Chapter:

- Chemical-mediated digestion: an alternative realm for middle-down mass spectrometry (*Paper V*)

6.1. Paper V: Chemical-mediated digestion: an alternative realm for middle-down mass spectrometry

Chemical-mediated digestion: an alternative realm for middle-down mass spectrometry

Kristina Srzentić[1], Konstantin O. Zhurov[1], Anna A. Lobas[2,3], Gennady Nikitin[1] Mikhail Gorskhov[2,3] and Yury O. Tsybin[4*]

[1] Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
[2] Institute for Energy Problems of Chemical Physics, Russian Academy of Sciences, Leninskii Prospect 38, Bldg. 2,119334 Moscow, Russia
[3] Moscow Institute of Physics and Technology (State University), 9 Institutskiy per., 141707 Dolgoprudny, Moscow Region, Russia
[4] Spectroswiss Sarl, EPFL Innovation Park Build I, 1015 Lausanne, Switzerland

Correspondence should be addressed to Dr. Yury O. Tsybin, Spectroswiss Sarl, EPFL Innovation Park Build I, 1015 Lausanne, Switzerland. E-mail: tsybin@spectroswiss.ch

Running title: Chemical cleavage-based middle-down approach

Submitted to: *Anal.Chem*

Current manuscript date: 31 January, 2016

Keywords: middle-down (MD) mass spectrometry; Fourier transform mass spectrometry, FTMS; Orbitrap; chemicals; chemical cleavage; rare amino acids

Abstract

Recent advances in mass spectrometry (MS) instrumentation provide the necessary platform for the application of middle-down (MD) proteomics approach and investigation of increasingly longer (>3 kDa) peptides in a high-throughput manner. Generating longer peptides represents a paramount step in launching MD pipeline. In bottom-up proteomics (BUP) approach substantial protein degradation targeting lysine and arginine residues yields moderately evened distribution of peptides in the investigated mass bin (typically 6-30 residues-long resulting peptides) for all major kingdoms as well as their subproteome/sub compartment level. Whereas frequent digestion works well for BUP mass regime, MD proteolysis requires the exact opposite. Increasing the length of resulting peptides greatly depends upon the right selection of the targeting residue, which is further contingent on distribution of amino acid residues within analyte's primary sequence, and might drastically vary passing from one kingdom to another. Hence, MD presently meets its practical limitations in the selection of a cleaving agent that would provide the desired peptides of 3-15 kDa. To date, only a couple of attempts have been made to propose cleaving strategy for MDP, elucidating potential cleaving agent candidates. Our recent bioinformatics study has shown that cleavage at rare amino acid residues such as methionine, tryptophan or cysteine, and even a combination of these digestion sites would be beneficial for MD approach. Chemical cleavage allows targeting of rare amino- acids, for which no proteolytic enzymes exist. Herein we have investigated the performance of several chemical agents targeting primarily Met, Cys and Trp (CNBr, BNPS-Skatole and NTCB, respectively) as potential candidates for MD proteolysis. Figures of merit such as digestion reproducibility, peptide size distribution and presence of side reactions are discussed.

Introduction

Proteolysis is considered the main avenue for in-solution and in-gel digestion of complex protein mixtures.([1-5] One of the main challenges of both extended bottom-up (eBU) ([6, 7] and middle-down (MD) proteomics ([8, 9] is the generation of peptides in the defined mass range (3-7 or 7-15 kDa, respectively), possibly without creating at the same time peptides that are either below or above the targeted size. Recent reports described proteases such as LysC, GluC, Sap9 or OmpT [8] as a 'way to go' for generating large peptides in the mass bin suitable for MD. Claimed advantages of these proteases should be, aside from providing the targeted length of yielding peptides, their reproducible digestion and robustness. Due to these reasons, alongside with their non-toxic nature, proteases are typically favored over chemicals in proteomics. Although digestion with chemical agents essentially performs the same type of reaction, i.e., hydrolysis of peptide bonds in proteins, this alternative approach was exclusively applied in a targeted fashion, for cleavage of select proteins, mainly with the purpose of obtaining primary sequence information, for protein engineering, and to aid in elucidation of protein structure [10] for obtaining primary sequence information of selected proteins. With the exception of acid hydrolysis [9, 11-13], the utility of chemical digestion has not been yet investigated in MD proteomic studies. One of the main reasons for this delay might also lay in the fact that these small molecules target peptide bonds located on one of the termini of less frequent amino acid residues such as methionine, tryptophan or cysteine, and the larger peptides they generate could not have been investigated in a high-throughput manner until recently, due to technological limitations in mass spectrometry (MS). High resolution mass spectrometers, such as those equipped with orbitrap or time-of-flight mass analyzers, have been adapted for high-throughput characterization of large biomolecules just in the last decade. Additionally, until recently, a comprehensive information about the average size distribution of peptides obtainable by different chemical cleavages was missing. Cleavage of proteins with small molecules, in a difference to enzymatic cleavage, entails significant modification to the side chain of the targeted residue. This feature alone changes the input information on cleavage rules given to the search algorithm. Nowadays, adequate supporting bioinformatics platform, often developed for or also applied to top-down (TD) proteomics [14-16], can tackle analysis of this type of peptides by creating specific databases, properly dealing with complex tandem mass spectra obtained by the fragmentation of large polypeptides (similarly to what happens with TD MS), and also accounting for eventual unexpected modifications [17]. However, there were previous attempts to integrate cleavage with chemical agents into routine MS analysis of proteins. For instance, cleavage after methionine using cyanogen bromide (CNBr) is one of the more commonly used chemical cleavage methods that has limited side reactions and has a yield of 90-100% [18]. Vestling *et al.* reported on cleavage at C-terminal side of tryptophan residue using BNPS-skatole [19]. There were periodical attempts to utilize cysteine as target cleavage site solution for analysis of proteins with particular structure and/or structure adjacent problematics. Most of these studies however, were limited to the analysis of a single or a couple of proteins, investigation of particular modification introduced, such as dehydroalanine formation on cysteines in serum albumin [20] or alternatively to optimization of the single step in the reaction or entire reaction with the purpose of explaining the underlined mechanism in organic chemistry. To our knowledge, the herein described chemical cleavage procedures [21] have not been applied to large scale MD studies, nor have their cleavage efficiencies been evaluated in comparison with enzymatic procedures for the proteolysis of simpler protein mixtures analyzed by MS in terms of the effective capability of producing peptides in the desired mass range. Nonetheless, our group previously reported a bioinformatics study about the potential of all amino acid residues for generating MD-sized peptides [22], from where Met, Cys and Trp emerged as viable candidate targets for the development of a novel MD platform, with a potential for a future application to the whole proteome scale. Here, we first report protocol refinements for the chemical cleavage at the three aforementioned amino acid residues, followed by the experimental results of the analysis of digestion products of a simple seven-protein mixture digested with different chemical agents with a high-resolution Orbitrap FTMS, validate those that are suitable for future large scale middle-down proteomics applications.

Experimental methods

Sample preparation. An equimolar model mixture (100 µM) consisting of yeast enolase 1 and 2, bovine apo-transferrin, serum albumin, pancreatic ribonuclease A, chicken egg white lysozyme (all from Sigma Aldrich, St. Louis, MO, USA) and bovine carbonic anhydrase 2 (Protea Biosciences, Morgantown, WV) was prepared in 6.8 M urea in 100 mM ammonium bicarbonate buffer (pH 7.8). An aliquote of the mixture was removed for the digestion with NTCB, and the remainder was reduced with DTT (5 mM final concentration) at 50 °C for 1 hr, followed by 45 min alkylation at room temperature in dark with 18 mM final iodoacetamide. Sample was then split into aliquots (each containing 1 nmol of the protein mixture) dried to pellet in a SpeedVac concentrator (Eppendorf) and resuspended in the appropriate buffer for each of the chemical digestion procedures. For all procedures in part, digestion was carried out on three biological replicates.

Cleavage protocols. All cleavage procedures employed (N-terminal Cys cleavage with NTCB, C- terminal Met cleavage with CNBr, C- terminal Trp cleavage with o-iodosobenzoic acid or BNPS Skatole) were based on previously tailored protocols [23-24].

LC-MS/MS analysis. All peptides obtained through different chemical digestion procedures were subjected to the pooled C4-C18 stage tip cleanup with ZipTip cartridges (Millipore, Billerica, MA) as described previously [7] prior to LC separation. Approximately 8 pmol of peptide mixture was loaded onto C8 (2 cm, 100 Å, 5 um) trap-column for 10 minutes with 0.1% FA at a flow rate of 8 uL/min, respectively. Reversed-phase nano LC was performed using a Dionex Ultimate 3000 system (Thermo Scientific, Bremen, Germany) equipped with C8 column (150 mm, 300 Å, 5 um). Solvent A was composed of 0.1 % of FA in water and solvent B of 50 % MeOH, 20 % ACN, 10 % TFE, and 0.1 % FA. The percentage of the organic phase was increased from 5 to 60%

over 60 minutes for all performed analyses. For all chemical procedures employed, each digestion replicate was analyzed in three consecutive technical replicates.

The outlet of chromatographic column was coupled on-line with a nano electrospray ionization (ESI) source (Nanospray Flex ion source, Thermo Scientific) equipped with a metallic emitter to which a 2.2 kV potential was applied. Mass spectrometric analysis was performed on a hybrid high-field LTQ Orbitrap Elite FTMS (Thermo Scientific). Identification of peptides from Met, Cys or Trp digestions was carried out on a LC-time scale, with the mass spectrometer operating in data-dependent mode. In all the LC-MS/MS runs, the survey scan was performed at 60'000 resolution (at 400 *m/z*) in the Orbitrap FTMS with automatic gain control (AGC) set at 1 E6. Dynamic exclusion was enabled with 60 s duration.

Isolated precursor ions were subjected to higher-energy collision induced dissociation (HCD), with singly- and doubly-charged precursor ions excluded from triggering MS/MS event. The AGC (number of charges) target value for MS/MS events was set to 5 E4. HCD was performed in a top-5 mode with product ion detection in the Orbitrap FTMS operating at 15'000 resolution (at 400 *m/z*) with 3 microscans per each scan. Normalized collision energy (NCE) was set at 27% (default charge state: 3+).[25]. Signal to noise (S/N) threshold was set to 15'000 throughout all experiments (relative intensity units).Additionally, intact mass measurement of individual proteins selected upon first round of data analysis was performed in order to investigate and confirm protein N terminal truncation.

Data processing. Theoretical distributions as well as *in-silico* digestions of the non-redundant protein databases of human, yeast, and bacteria were performed using an in-house Python-based interface based on pyteomics library. The peptide size distributions were determined for currently targeted amino-acid cleavage sites for MDP (dibasic, D, Q/E, W, M and C). Obtained .raw files for all analysis in part were peak picked using ReadW, centroided and converted to mzXML format for deconvolution with SNR threshold set to 3.

Database search. Data was searched in 'PTM discovery mode' by MS Align+ against custom database (containing primary sequences of 7 proteins used in the experiments). Precursor tolerance was set to 10 ppm, and product ion tolerance to 0.1 Da. In all cases except for the Cys-based protocol, carbamydometylation of cysteine residue was enabled as a fixed modification.

Data analysis. Manual validation was performed to confirm expected label-introduced modifications as well as to elucidate potential unexpected modifications on the side chains of residues. For that purpose, the search algorithm results output was ported to Excel, wherein the mass shifts of all the protein species were logged, along with purported mass shift localization data. The mass shifts were split into cleavage inducing and non-cleavage inducing groups (e.g. +25Da on N terminus, associated with formation of itz-peptide N-terminal to a cysteine vs. +16Da on methionine associated with oxidation of methionine). Next, the mass shifts were evaluated with respect to plausibility of occurrence under the expected chemistry for a given chemical method. Notably, a significant percentage of mass shifts were, in fact, associated with sums of multiple non-localized (due to lack of ion assignments) mass shifts within a region of a peptide (e.g. 41 Da = 25 Da + 16 Da, commonly occurring at the N-terminal end of a peptide with a proximate methionine). In such cases, manual ion assignment was carried out on .raw data, with aid of Protein Prospector in attempt to localize the individual mass shifts, most of the time successfully. Furthermore, the results output was subjected to additional treatment in cases where the PrSM with the highest E-score produced clearly mis-assigned output (i.e. ones containing short peptide sequences either side of a truncation with mass shifts of several thousand Da).

In certain instances, the correct net mass shift was identified by the software, but the contributing individual mass shifts were associated with a wrong residue, hence matching the peptide with primary sequence which has one residue more or less than the true peptide has. In all such cases, the .raw data was analyzed in *'de novo'* fashion to verify the proposed alternative assignments.

Results and Discussion.

Choice of cleaving agent in middle-down (MD) is a key step towards a successful pipeline. Several residues were suggested as target for MD so far, such as Q, D and dibasic (cleaving agents proposed: Lys-C, formic acid [12, 13], OmpT [8] Sap9 [6], respectively). Comparing obtainable peptide mass distribution from i*n silico* digestion of three proteomes (human, yeast and bacterial) shown in Figure S1 for each of these methods with less frequent residues (M, C and W), it stands to reason to conclude how latter ones could be more suitable or strongly complementary avenue for MD digestion. In terms of proteome coverage shown, theoretical survey shows how targeting all three residues yields substantial proteome coverage (Figure 1). However, throughout kingdoms prevalence of methionine as residue of choice is strongly implied; 99.85, 99.97 and 99.98 % for human, bacterial and yeast proteome respectively. (Figure S2). Reaction mechanism of small molecules employed here (Scheme 1) is somewhat different than those of proteases. As depicted in Scheme 2, this type of hydrolysis introduces a modification to the side chain of the targeted residue upon cleavage. Additionally, a different characteristic mass modification to the side chain of a residue is introduced even in the case the cleavage is omitted (e.g. -34 Da on Cys from dehydroalanine formation, -29.99 Da on Met from homoserine formation, +15.99 Da on Trp from oxindolyalanine formation). This structurally different way of cleaving represents a certain challenge for current data analysis software available for treatment of data in proteomics (*vide infra*). Table 1. shows figures of merit for employed chemical cleavage of seven protein mixture. The experimentally obtained average molecular weights for the peptides generated by each protocols (5.8, 7.8/8.8 and 8.3 kDa for cleavage at C, W and C, respectively) are in accordance with theoretical distribution (Figure SI1) and fall in the middle of the desired 3-15 kDa mass bin for middle-down. For the iodobenzoic acid protocol, the average sequence coverage was 51.4%, with cleavages at W detected in 79.4% (27/34) of cases. Notably, several of the identified peptides showed cleavage at the N-terminal, rather than C-terminal side to the tryptophan. In other cases, a previously unreported mass shift of 31.98 Da was observed (13.98+H_2O). Notably, for this particular protein mixture, the identified peptides, specifically for the larger proteins tended to be located at or very close to the termini of the proteins (e.g., in the case of serotransferrin, seroalbumin and enolase). As expected, ribonuclease, for lack of tryptophans, and due to its relatively small size (13.6 kDa) was detected as an entire protein in MD instrument setup. For the BNPS protocol similar results were obtained: overall sequence coverage (excluding ribonuclease) was 47.8% and cleavages were observed at 73.5% (25/34) of all tryptophan residues. Finally, some of the identified peptides contained cleavages around basic and acidic amino acids and not in vicinity of tryptophans. An in-depth analysis of the cleavage sites identified for the CNBr protocol revealed that, contrary to the chemoselectivity proposed for the reagents, cleavages were equally likely to occur at methionine and at tryptophan. In fact, disregarding N-terminal methionines, cleavages were identified at 72.5% (37/51) of methionines and 76.5 % (26/34) of tryptophans – the latter percentage being effectively equal to the two observed for the tryptophan-specific protocols, *vide supra*. A mechanism involving halogenation of tryptophan followed by HBr loss and hydrolysis has been proposed to explain cleavage at tryptophans, see Scheme 1 panel b. Finally, it should be noted that multiple instances of protein cleavage near aspartic and glutamic acid have also been observed, (likely a result of deprotonation of the side chain followed by nucleophilic substitution at the carbonyl group of the amide bond). Notably, cleavage at multiple amino acids gave rise to significantly higher sequence coverage than in either tryptophan cleaving protocols, with 78.8 %. Reproducibility of total ion chromatograms depicted in Figure S4 secludes NTCB-based method as the most reproducible throughout three digestion replicates. Figure 2. Illustrates the comparison of selected NTCB protocol with the benchmarked BUP trypsin protocol based on the the number of identifiable proteins as a function of peptide length given as a count of residues for human proteome. Considering the entire mass range of peptides, trypsin identifies slightly higher number of proteins compared to NTCB (19955 *vs* 19687, respectively). However, if we take into account only working regimes for both methods, number of peptides and therefore valuable information on proteins 'lost' is almost six times higher for trypsin (4061 is in the >30 residues long bin) than it is the case for NTCB (721 fall into 0-30 residues long bin). Even though this does not mean how protein identification is completely lost by the absence of these peptides (as those can be identified with another peptides, or couple of them, in appropriate mass bin), significant amount of information on the protein family is lost, lowering their unambiguous identification in BUP regime. For MD regime this could also mean a loss of PTM information, or connectivity between adjacent PTMs, however the ambiguity of identification will not be hindered as much, since the probability of having another unique peptide is much higher due to their 2-5-fold increased peptide length. Similar trend was found for yeast and bacterial kingdoms (Figure S2). In addition, b panels of Figures 2 and S2 show the number of unique peptides as a function of the peptide length for both methods in part. Surprisingly, in 0-30 residues bin, the number of unique peptides generated by targeting cysteine is comparable to the one obtainable by trypsin, but expectedly significantly higher in respective 30-n bin. Another striking information obtained is the number of unique peptides that trypsin generates, which only in BUP regime exceeded 500 000 for human proteome, while there is still substantial portion of peptides in the MD bin. Note how this number does not account for multiple instances of the same peptide (various modifications to the peptide primary sequence) and in actual experiment it is increased by a minimum of two-fold, overcrowding the LC run and therefore challenging their successful elution, ionization and fragmentation under LC time constraints. In case of NTCB this number is six times reduced (~ 100 000 unique peptides in MD bin). However, even with the reduction of proteolytic pool, intrinsic problem inherent to top-down (TD) – high heterogeneity of molecular weights and charges in the same LC run is coherent with MD regime as well (Figure 3), challenging their successful fragmentation and obtaining the complete sequence coverage, which in turn can bias the identification.

The NTCB protocol was found to be most chemoselective of the four protocols under consideration: only a single instance of a cleavage non-proximate to a cysteine was observed. Notably, this cleavage was observed in all four protocols and is thus deemed to be non-reagent specific and likely is a result of local structural effects that render hydrolysis under basic conditions particularly efficient. Five additional instances of cleavages within three amino acids of a cysteine were observed (Figure 5b). Notably, these cleavages are located around specific cysteines in a given protein and in four cases serine was present, whilst in three cases either lysine, threonine or/and aspartic acid were present – i.e. primary sequence and local secondary structure both likely played a role. Cleavages were observed at 62.5% (60/96) of cysteines with sequence coverage

above 80% (excluding carbonic anhydrase, which lacks cysteines, and the two enolases which contain one cysteine and would form fragments circa 18.9 kDa and 24.8 kDa in mass – unlikely to be detected with a MDP instrumental set-up). Importantly, as can be seen from Figure 5, the number of missed cleavages, as a result of DHA formation or label that did not lead to cleavage, was 40 % and 20 % respectively, relative to the number of itz-peptides (note that these numbers do not represent a weighted average, but simply the number of assigned instances overall thus representing species diversity than relative kinetics of the two reactions, i.e. if the most abundant species did not contain DHA, and one of the least abundant may contain two DHA's – this will go as equal number of counts for itz and DHA peptides). It is important to stress out how MD data analysis in general, and in this study in particular, is circumspect by the lack of appropriate tools for data interpretation. Currently available softwares for TD can be used for MD data, however manual validation is often required, due to occasional missed assignment of peptide N- or C- terminal residue. This might derive from the fact that typical TD software accounts for fragmentation pattern, rather than the cleavage rule, hence the crucial parameter on peptide's starting residue (in case of N-terminal cleavage) or ending residue (in case of C- terminal cleavage) is excluded from the defined search guidelines *a priori*. This in turn means how the *in-silico* database is not constructed (as it is the case in search engines designed for proteolytic-based data) and the identification is based on matched (or missing) fragmentation ladder and the mass of the investigated precursor. The lather can be biased by the various modifications occurring on the peptide and lead to an incorrect peptide assignment as exemplified for bovine serotransferrin (UniProt acc. number G3X6N3) in Figure S5. Here an additional amino acid (in respect to the expected primary sequence based on the cleaving agent specificity) was included at the N-terminus, identifying the peptide which starts with serine instead of cysteine. This is likely due to the high (>50Da) mass shift associated with expected N-terminal amino acid, and the lack of N-terminal fragment ions identified. First fragment ion assigned (b_{448}; numerical denotation indicates the position of the bond cleaved in respect to the entire protein sequence, as considered by the search algorithm employed) identified unknown mass shift which matched theoretical mass of a protein portion that corresponds to amino acid sequence starting with serine with a loss of -19.03 that is not in the list of known PTMs. Manual assignment of product ions from the raw spectra and mass matching to the theoretical list of product ions for the peptide starting with cysteine identified *b*- ion series consistent with cyanylation of cysteine (itz-peptide formation; C+ 24.99 Da) as well as *b*- and *y*-ions localizing carbamylation at lysine (K+43 Da), reported to occur in this protocol [26]. It should be noted that due to the basis set of mass shifts identified, one could not perform an *en masse* automated assignment because certain mass shift combinations effectively produced sum mass shifts of same nominal mass involving different number of cysteines (from zero to four). Importantly, of the ~3000 individual entries returned by the software, for 9 LC-MS runs for NTCB protocol, over 99 % of mass shifts were rationalized and found to be consistent with chemicals reactions occurring during the execution of the experimental protocol. For other cleaving methods employed ~ 10 -15 % of mass shifts (data not shown here) remained unclarified, possibly due to the unexplained underlying mechanisms and/or alternative pathways of the reactions, which were outside of the scope of this evaluation study.

Conclusions

We found NTCB protocol to be in line with desired characteristics for a MD protocol owing to: i) high selectivity towards a target residue, ii) high reproducibility, iii) generation of long peptides Thus, it can be proposed for a qualitative MD analysis and it shows a promise for development into a quantitatively accurate approach as well. Appearance of unconstrained mass modifications in the data analysis revealed that the strategy of changing the nature of the base towards reduced basicity with reasonable nucleophilicity and low steric hindrance enabled us to affect the branching ratio of competing pathways toward proteolysis, rather than formation of dihydroalanine. The characteristic mass shifts associated with both major reaction channels, coupled to the chemoselectivity of the reactions in question, enable for facile data interpretation and manual validation of assignments. Further work will include validation of this approach at the large scale, which is currently circumspect by the lack of automated search algorithm that accommodates requirements of such type of peptides.

Acknowledgments

Authors would like to thank Grigory Karateev and Dr. Elena Dubikovskaya for technical support, and Dr. Luca Fornelli for fruitful discussions. We are grateful for financial support through the Swiss National Science Foundation (SNF project 200021-125147/1) and the European Research Council (ERC Starting Grant 280271).

References

1. Aebersold, R. and M. Mann, *Mass spectrometry-based proteomics.* Nature, 2003. **422**(6928): p. 198-207.

2. Tsiatsiani, L. and A.J. Heck, *Proteomics beyond trypsin.* FEBS J, 2015. **282**(14): p. 2612-26.

3. Chait, B.T., *Mass spectrometry in the postgenomic era.* Annu Rev Biochem, 2011. **80**: p. 239-46.

4. Chait, B.T., *Mass Spectrometry: Bottom-Up or Top-Down?* Science, 2006. **314**(5796): p. 65-66.

5. Zhang, X., *Less is More: Membrane Protein Digestion Beyond Urea-Trypsin Solution for Next-level Proteomics.* Mol Cell Proteomics, 2015. **14**(9): p. 2441-53.

6. Laskay, U.A., et al., *Extended bottom-up proteomics with secreted aspartic protease Sap9.* J Proteomics, 2014. **110**: p. 20-31.

7. Srzentic, K., et al., *Advantages of extended bottom-up proteomics using Sap9 for analysis of monoclonal antibodies.* Anal Chem, 2014. **86**(19): p. 9945-53.

8. Wu, C., et al., *A protease for 'middle-down' proteomics.* Nat Meth, 2012. **9**(8): p. 822-824.

9. Cannon, J., et al., *High-throughput middle-down analysis using an orbitrap.* J Proteome Res, 2010. **9**(8): p. 3886-90.

10. Chapman, E., J.S. Thorson, and P.G. Schultz, *Mutational Analysis of Backbone Hydrogen Bonds in Staphylococcal Nuclease.* Journal of the American Chemical Society, 1997. **119**(30): p. 7151-7152.

11. Cannon, J.R., N.J. Edwards, and C. Fenselau, *Mass-biased partitioning to enhance middle down proteomics analysis.* J Mass Spectrom, 2013. **48**(3): p. 340-3.

12. Fenselau, C., O. Laine, and S. Swatkoski, *Microwave assisted acid cleavage for denaturation and proteolysis of intact human adenovirus.* Int J Mass Spectrom, 2011. **301**(1-3): p. 7-11.

13. Swatkoski, S., et al., *Evaluation of microwave-accelerated residue-specific acid cleavage for proteomic applications.* Journal of Proteome Research, 2008. **7**(2): p. 579-586.

14. Liu, X., et al., *Protein identification using top-down.* Mol Cell Proteomics, 2012. **11**(6): p. M111 008524.

15. Liu, X., et al., *Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach.* Mol Cell Proteomics, 2010. **9**(12): p. 2772-82.

16. D. LeDuc, R. and N. L. Kelleher, *Using ProSight PTM and Related Tools for Targeted Protein Identification and Characterization with High Mass Accuracy Tandem MS Data*, in *Current Protocols in Bioinformatics*. 2002, John Wiley & Sons, Inc.

17. Ansong, C., et al., *Top-down proteomics reveals a unique protein S-thiolation switch in Salmonella Typhimurium in response to infection-like conditions.* Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(25): p. 10153-10158.

18. Smith, B.J., *Basic Protein and Peptide Protocols*, in *Methods in Molecular Biology*. 1994, Humana Press, Totowa, NJ. p. 297-309.

19. Vestling, M.M., M.A. Kelly, and C. Fenselau, *Optimization by mass spectrometry of a tryptophan-specific protein cleavage reaction.* Rapid Commun Mass Spectrom, 1994. **8**(9): p. 786-90.

20. Bar-Or, R., L.T. Rael, and D. Bar-Or, *Dehydroalanine derived from cysteine is a common post-translational modification in human serum albumin.* Rapid Communications in Mass Spectrometry, 2008. **22**(5): p. 711-716.

21. Han, K.-K., C. Richard, and G. Biserte, *Current developments in chemical cleavage of proteins.* International Journal of Biochemistry, 1983. **15**(7): p. 875-884.

22. Laskay, U.A., et al., *Proteome digestion specificity analysis for rational design of extended bottom-up and middle-down proteomics experiments.* J Proteome Res, 2013. **12**(12): p. 5558-69.

23. Degani, Y. and A. Patchornik, *Cyanylation of sulfhydryl groups by 2-nitro-5-thiocyanobenzoic acid. High-yield modification and cleavage of peptides at cysteine residues.* Biochemistry, 1974. **13**(1): p. 1-11.

24. Crimmins, D.L., S.M. Mische, and N.D. Denslow, *Chemical cleavage of proteins in solution.* Curr Protoc Protein Sci, 2005. **Chapter 11**: p. Unit 11 4.

25. Laskay, U.A., et al., *Practical considerations for improving the productivity of mass spectrometry-based proteomics.* Chimia (Aarau), 2013. **67**(4): p. 244-9.

26. Tang, H.-Y. and D.W. Speicher, *Identification of alternative products and optimization of 2-nitro-5-thiocyanatobenzoic acid cyanylation and cleavage at cysteine residues.* Analytical Biochemistry, 2004. **334**(1): p. 48-61.

Figure captions.

Scheme 1. Reaction mechanism for chemical-mediated protein digestion: a) cleavage at X-Cys with NTCB, b) cleavage at: Met-X with CNBr (left panel), cleavage at Trp-X with CNBr (right panel) c) cleavage at Trp-X with BNPS skatole (left panel) and Trp oxidation (right panel).

Scheme 2. Characteristic mass shifts upon (a) Met-X, (b) X-Cys and (c): Trp-X cleavages. Top panel indicates mass shifts without miscleavage and bottom panels indicate the mas shift yielded by a miscleavage.

Figure 1. Venn diagram of human proteome coverage with herein proposed chemical methods targeting less frequent residues (M,C and W). Percentage of non-cleavable proteins is further separated into two categories by mass (MW >30 kDa indicated with pink circle; MW <30 kDa indicated with yellow circle).

Figure 2. Histogram representing a) the number of proteins as a function of peptide length given as a count of residues for NTCB *vs* trypsin based cleavage in human proteome. For both cleaving agents two mass bin ranges were considered (0-n and 30-n residues) Maximum attainable number of identifiable proteins in both bins is given. b) Number of unique peptides for both NTCB and Trypsin cleavage variant in the respective mas bin defined approach (BU (0-30 residues) and MD (30- 150 residues), respectively)

Figure 3. Evaluation of peptide mass and charge state variation over LC gradient obtained with LTQ Orbitrap Elite FTMS. Insets illustrate three examples of short (~ 20 residues), moderate (~50 residues) and long (> 60 residues) peptides.

Figure 4. Representative example of a single peptide sufficient for unambiguous identification of a single protein against Swiss-Prot database. Isolated 13+ precursor ion of serotransferrin (Bos Taurus) derived from N-terminal Cys digestion with NTCB is indicated in the right inset. Assignment

verification is confirmed by presence of expected chemical label (N-terminal cyanylation, Δ mass = 24.99 Da) with HCD MS/MS. Identified diagnostic ITZ-peptide *b*- ions are highlighted in red.

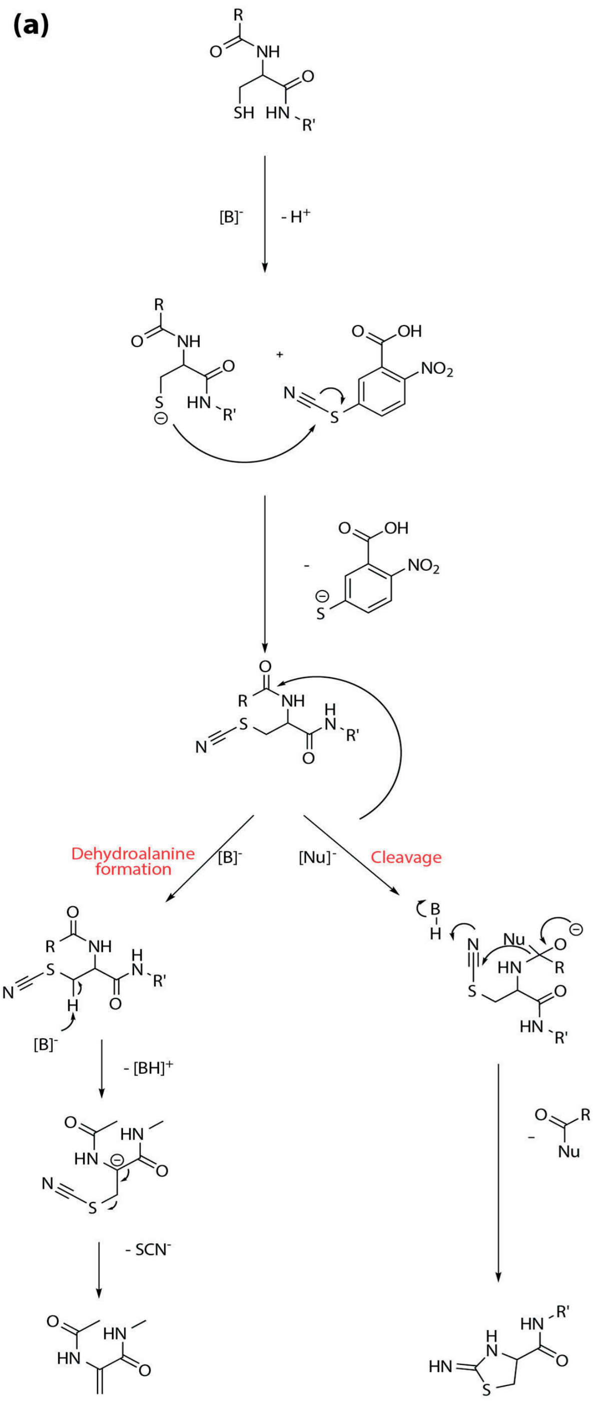
Figure 5. Statistical analysis of N-terminal Cys cleavage on a model seven protein mixture. a) Main panel depicts predominance of the cys cyanylation (ITZ peptide formation) reaction channel over the competing β elimination (DHA peptide formation). Right inset: Alternative channels and unknown modifications are presented as their observed mass shift and indicated with threshold (5%) given as the sum of the corresponding peptide entries divided by total number of entries returned by automated MS Align+ search, over nine replicas (three digestion replicates repeated in three technical replicates). Left inset: miscleavage rate- of 12% of identified peptides resulted in either DHA formation (uncleavable product) or peptides that carry a label but the cleavage was omitted due to the lack of the base.

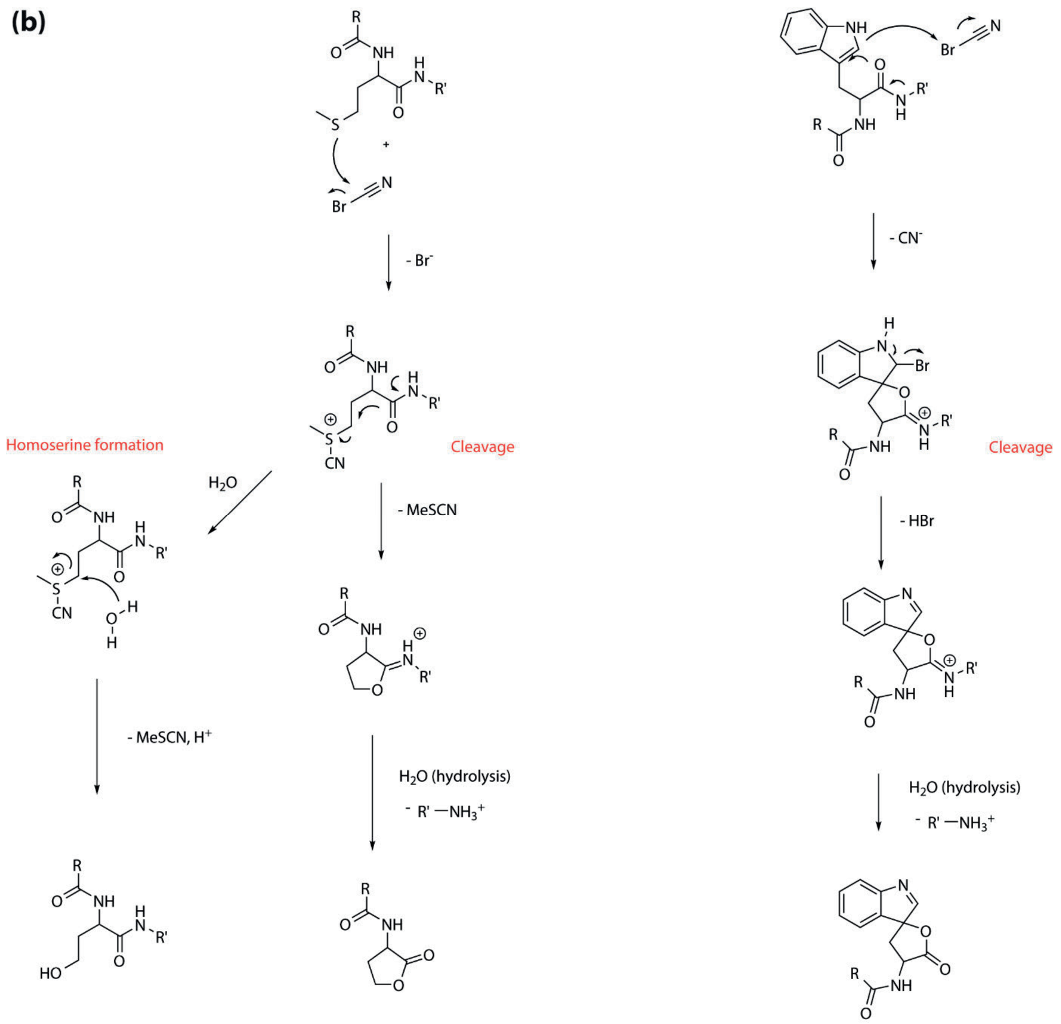
b) Cleavage specificity given as the count of cleavage sites at cysteine or at other than cysteine residue (averaged throughout nine total replicas), resulting from base hydrolysis. Other than cysteine cleavage sites are divided into cleavages in the vicinity of cysteine residues (within 3 amino acids) and those further away.

Table 1. Sequence coverage, number of identified peptides and average peptide length for 7 protein mixture digested with chemical methods employed. Number of expected *vs.* experimentally assigned cleavage sites and observed secondary cleavages are indicated for each chemical agent. *protein-specific secondary cleavage; “”protocol independent secondary cleavage, observed in all protocols.

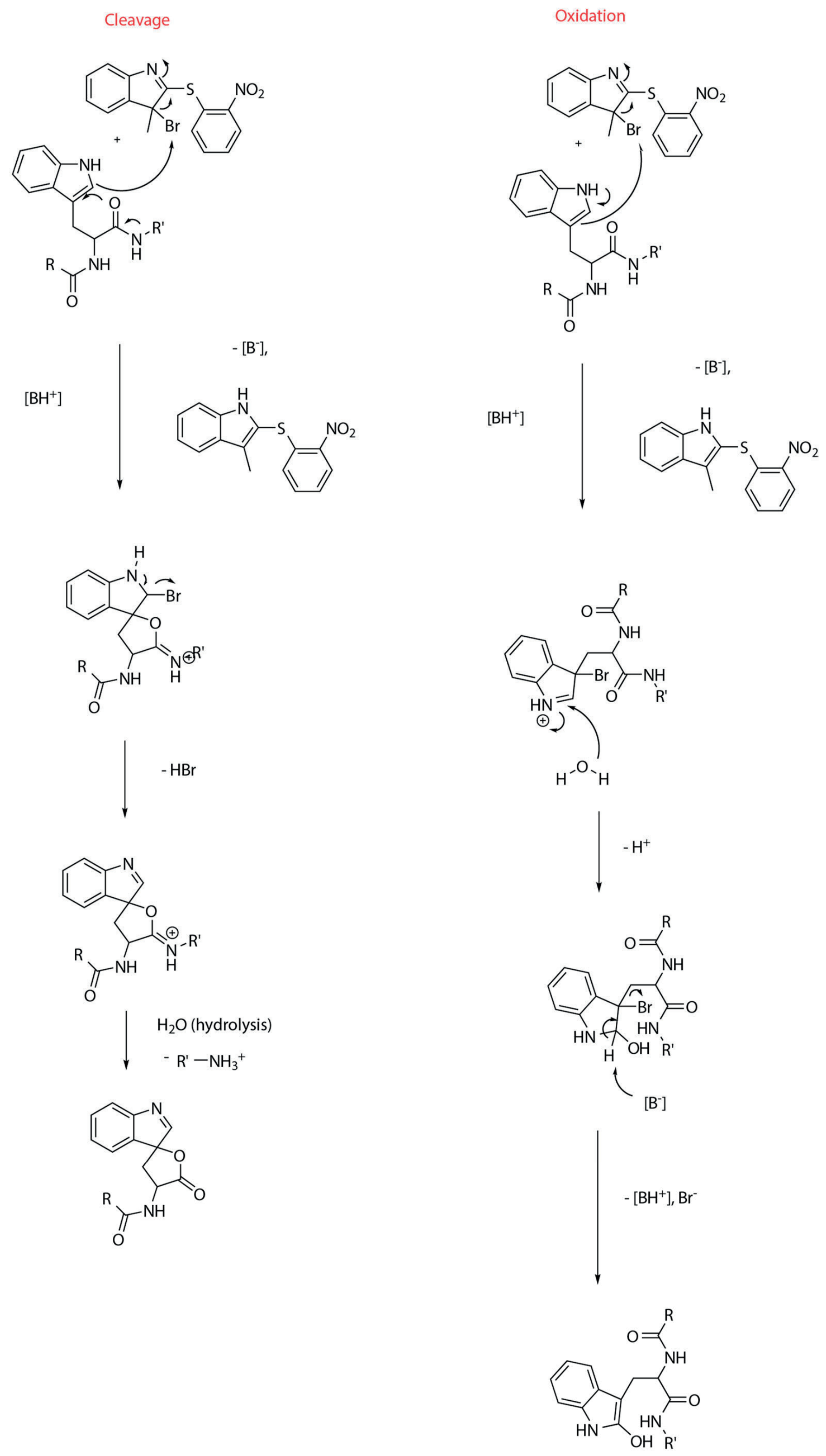
Targeted residue (cleaving agent)	# prot ID	Avg MW (kDa)	Avg charge	Observed *vs* theoretical site	Primary sequence mapped, %	Secondary cleavages
C (NTCB)	5	4.3	5.8	60/96	82.4	SCHTGL* DKKSCHT* CGDNTRK* SSNYCN* TKDRCK QSNSKD“”
W (BNPS skatole)	7*	6.1	7.8	25/34	47.8	D, N
W (iodosoben. acid)	7	6.5	8.8	27/34	51.4	D, N
M (CNBr)	7	5.7	7.3	37/51	78.8	26/34 of W, D

Scheme 1.




(b)
- Br-
Homoserine formation
H2O
Cleavage
- MeSCN
- MeSCN, H+
H2O (hydrolysis)
- R' —NH3+
- CN-
Cleavage
- HBr
H2O (hydrolysis)
- R' —NH3+

(c)


Cleavage
Oxidation
NO$_2$
Br
[BH$^+$]
- [B$^-$],
- HBr
H$_2$O (hydrolysis)
- R' —NH$_3$$^+$
- H$^+$
[B$^-$]
- [BH$^+$], Br$^-$

Scheme 2.

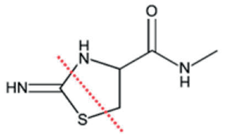
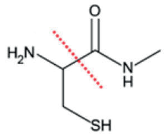
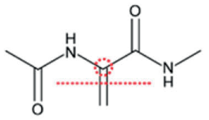
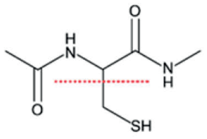
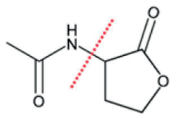
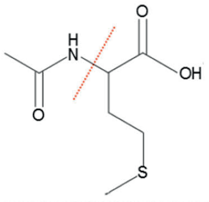
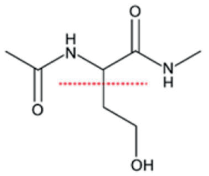
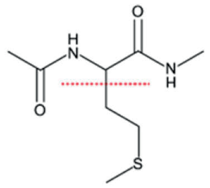
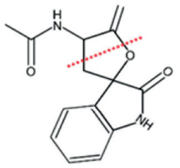
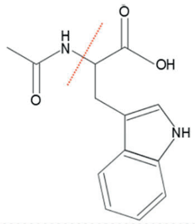
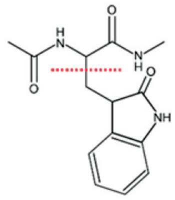
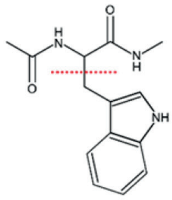
cleaving agent	Δ mass modification	modified residue	residue residue
NTCB	cleavage modification Δ[CN - H] = 24.99525 Da	 **ITZ peptide**	
	side chain modification (cleavage omitted) Δ[- SH_2] = -33.98772 Da	 **dehydroalanine (DHA)**	
CNBr	cleavage modification Δ[-SCH4] = -48.00337 Da	 **HEL peptide**	
	side chain modification (cleavage omitted) Δ[O - SCH_2] = -29.99281 Da	 **homoserine**	
BNPS skatole	cleavage modification Δ[O - H_2] = 13.97926 Da	 **DAS peptide**	
	side chain modification (cleavage omitted) Δ[O] = 15.99491 Da	 **oxindolyalanine**	

Figure 1.

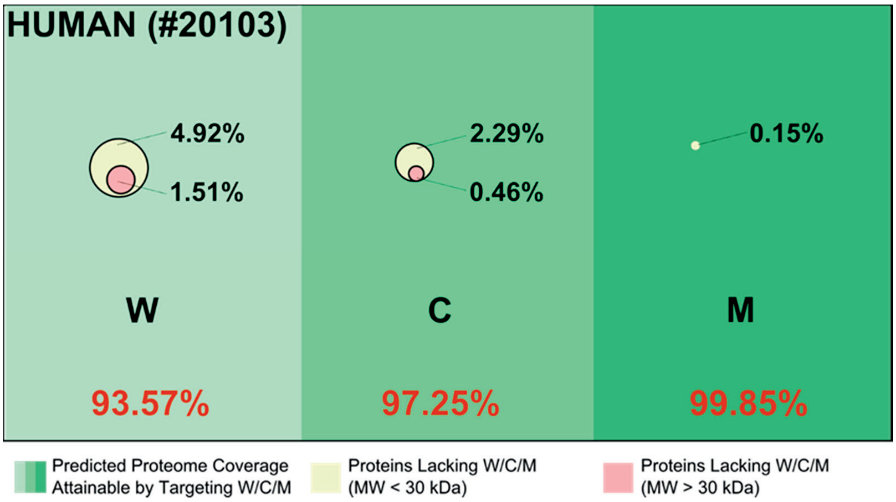
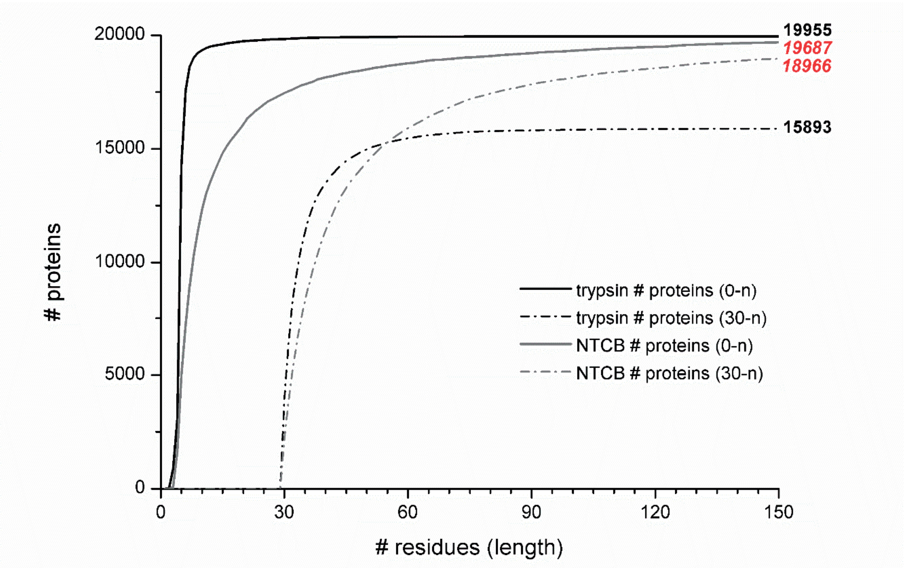

HUMAN (#20103)
4.92%
1.51%
2.29%
0.46%
0.15%
W
C
M
93.57%
97.25%
99.85%
Predicted Proteome Coverage Attainable by Targeting W/C/M
Proteins Lacking W/C/M (MW < 30 kDa)
Proteins Lacking W/C/M (MW > 30 kDa)

Figure 2.

a)



b)

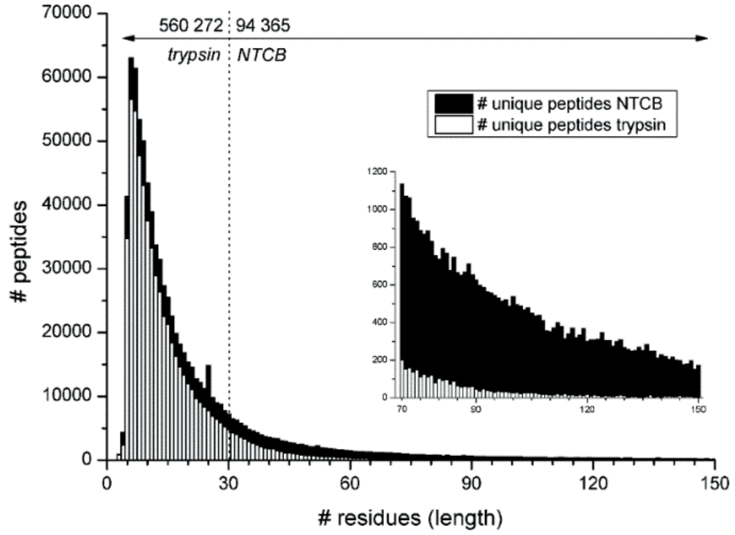

560 272
94 365
trypsin
NTCB
# unique peptides NTCB
# unique peptides trypsin
# peptides
# residues (length)
0
10000
20000
30000
40000
50000
60000
70000
0
30
60
90
120
150
0
200
400
600
800
1000
1200
70
90
120
150

Figure 3.

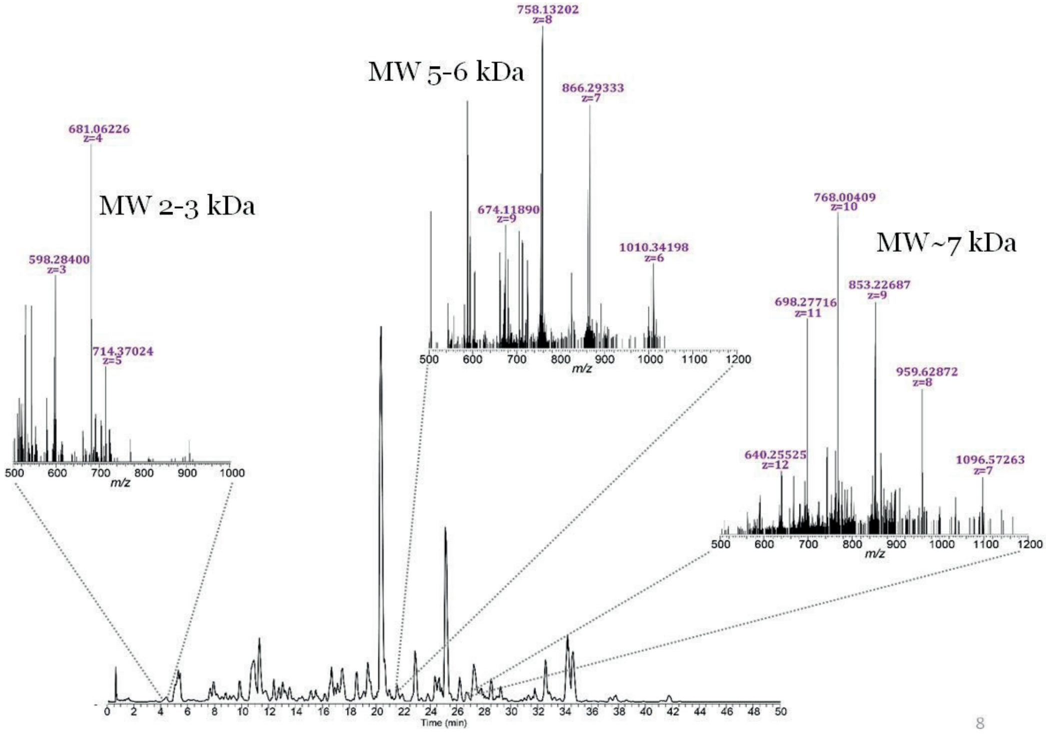

MW 2-3 kDa
681.06226
z=4
598.28400
z=3
714.37024
z=5
MW 5-6 kDa
758.13202
z=8
866.29333
z=7
674.11890
z=9
1010.34198
z=6
MW~7 kDa
768.00409
z=10
698.27716
z=11
853.22687
z=9
959.62872
z=8
640.25525
z=12
1096.57263
z=7
m/z
Time (min)
8

Figure 4.

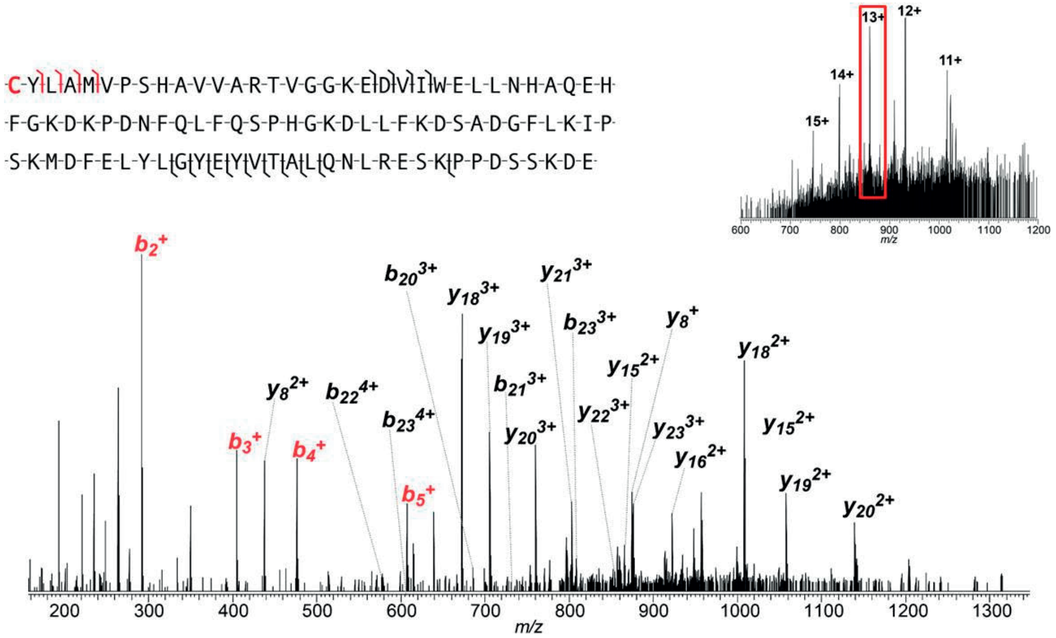
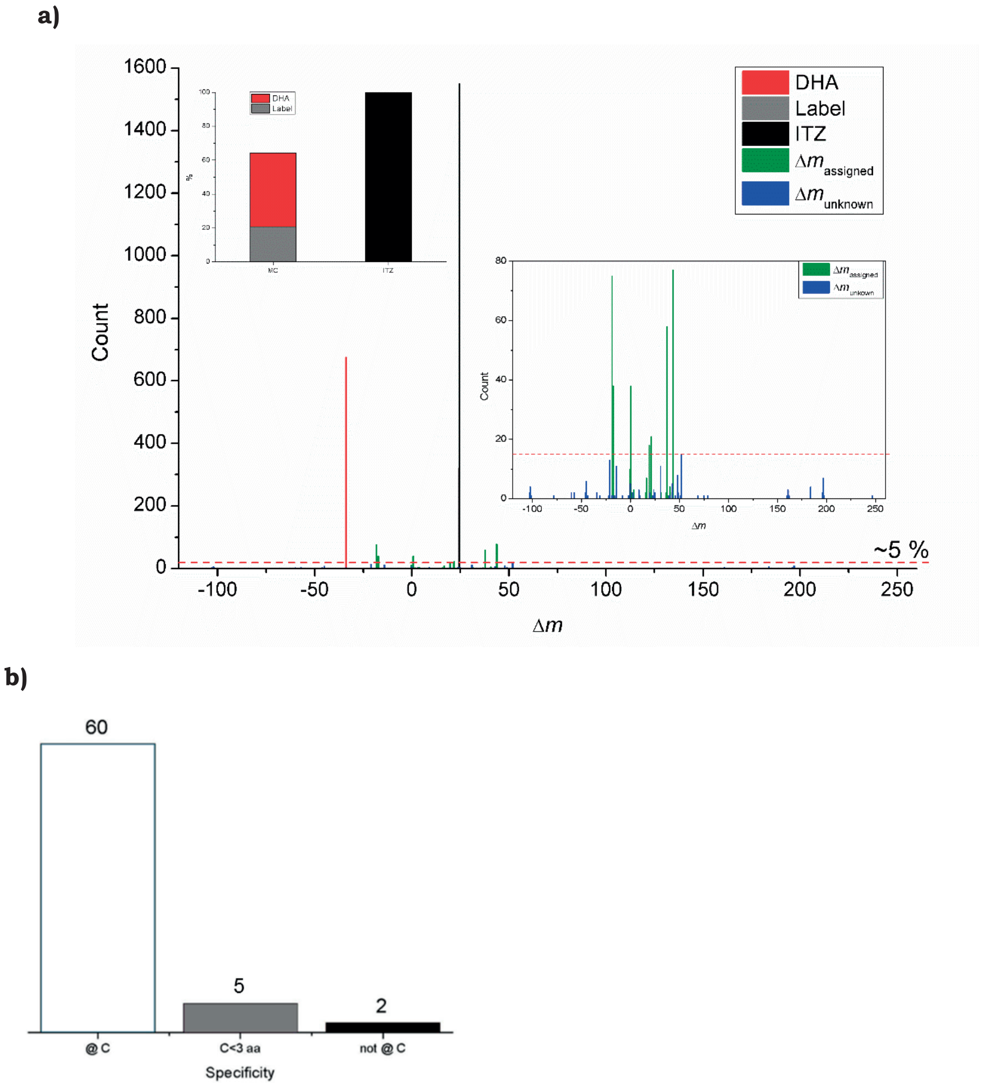

C-Y-L-A-M-V-P-S-H-A-V-V-A-R-T-V-G-G-K-E-D-V-I-I-W-E-L-L-N-H-A-Q-E-H
F-G-K-D-K-P-D-N-F-Q-L-F-Q-S-P-H-G-K-D-L-L-F-K-D-S-A-D-G-F-L-K-I-P
S-K-M-D-F-E-L-Y-L-G-Y-E-Y-V-T-A-L-Q-N-L-R-E-S-K-P-P-D-S-S-K-D-E
13+
12+
14+
11+
15+
600
700
800
900
1000
1100
1200
m/z
b_2^+
b_{20}^{3+}
y_{21}^{3+}
y_{18}^{3+}
b_{23}^{3+}
y_8^+
y_{19}^{3+}
y_{15}^{2+}
y_{18}^{2+}
y_8^{2+}
b_{22}^{4+}
b_{21}^{3+}
y_{22}^{3+}
b_{23}^{4+}
y_{20}^{3+}
y_{23}^{3+}
y_{15}^{2+}
b_3^+
b_4^+
y_{16}^{2+}
b_5^+
y_{19}^{2+}
y_{20}^{2+}
200
300
400
500
600
700
800
900
1000
1100
1200
1300
m/z

Figure 5.


a)
1600
1400
1200
1000
800
600
400
200
0
Count
-100
-50
0
50
100
150
200
250
Δm
~5 %
DHA
Label
ITZ
$\Delta m_{assigned}$
$\Delta m_{unknown}$
%
MC
ITZ
Count
Δm
b)
60
5
2
@C
C<3 aa
not @ C
Specificity

Supporting information.

Chemical-mediated digestion: an alternative realm for middle-down mass spectrometry

Kristina Srzentić, Konstantin O. Zhurov, Anna A. Lobas, Gennady Nikitin Mikhail Gorskhov and Yury O. Tsybin[4]

Figure 1S. Bioinformatics survey: *In silico* digestion of (a) human, (b) yeast and (c) bacterial proteome with recently introduced methods for MD MS (Lys-C, formic acid, OmpT/Sap9) and herein proposed chemical methods targeting less frequent residues (M,C and W)

Figure 2S. Venn diagram of (top panel) human, (middle panel) yeast and (bottom panel) bacterial proteome coverage with herein proposed chemical methods targeting less frequent residues (M, C and W). Percentage of non-cleavable proteins is further separated into two categories by mass (MW >30 kDa indicated with pink circle; MW <30 kDa indicated with yellow circle).

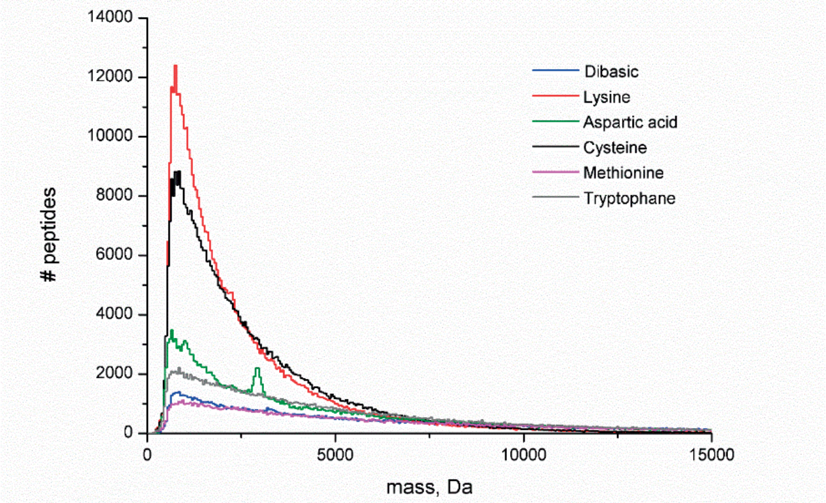
Figure 3S. Histograms for a) bacterial and b) yeast proteome. Top panel shows the number of proteins as a function of peptide length given as a count of residues for NTCB *vs* trypsin based cleavage (top panel). For both cleaving agents two mass bin ranges were considered (0-n and 30-n residues) Maximum attainable number of identifiable proteins in both bins is given. Bottom panel shows the number of unique peptides for both NTCB and Trypsin cleavage variant in the respective mas bin defined approach (BU (0-30 residues) and MD (30- 150 residues), respectively).

Figure 4S. Extracted base peak chromatogram of peptides obtained by triplicate chemical hydrolysis of seven protein mixture with (a) NTCB, (b) CNBr and (c) BNPS skatole. Experiments performed with LTQ Orbitrap Elite FTMS.

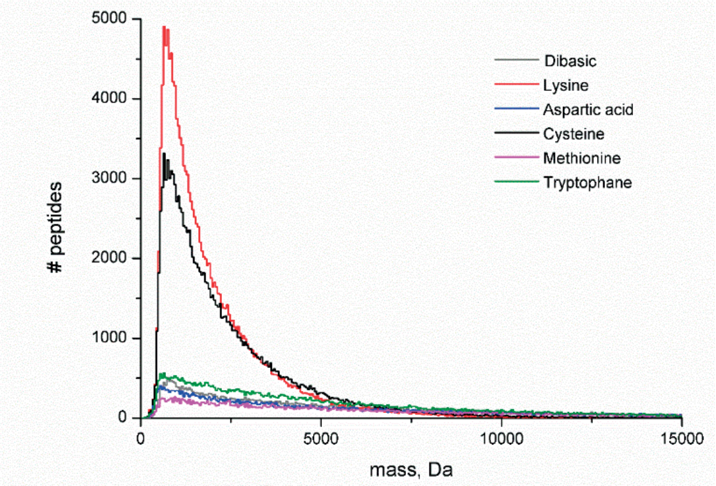
Figure 5S. Example of the incorrect assignment of the cleavage site by the search algorithm and correct manual assignment. Product ions that confirm modifications and their respective localization are indicated in red. Correct delta mass shifts are indicated above respective residue in blue.

Figure S1.

(a)



(b)


5000
4000
3000
2000
1000
0
# peptides
0
5000
10000
15000
mass, Da
Dibasic
Lysine
Aspartic acid
Cysteine
Methionine
Tryptophane

(c)

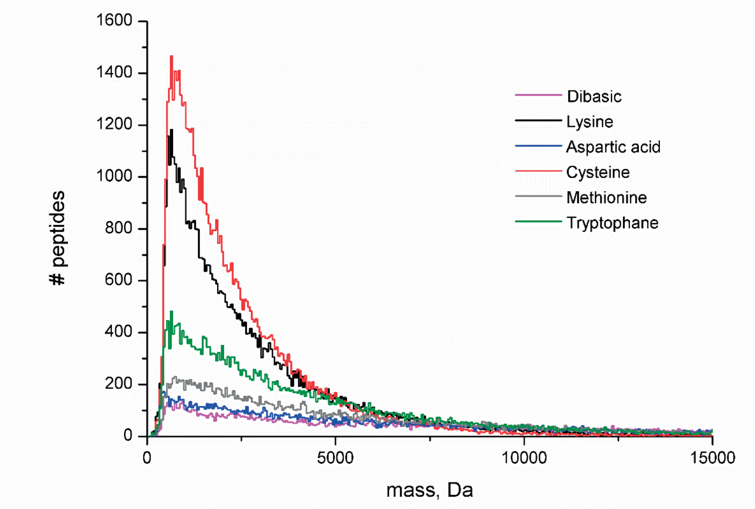

1600
1400
1200
1000
800
600
400
200
0
# peptides
0
5000
10000
15000
mass, Da
Dibasic
Lysine
Aspartic acid
Cysteine
Methionine
Tryptophane

Figure S2.

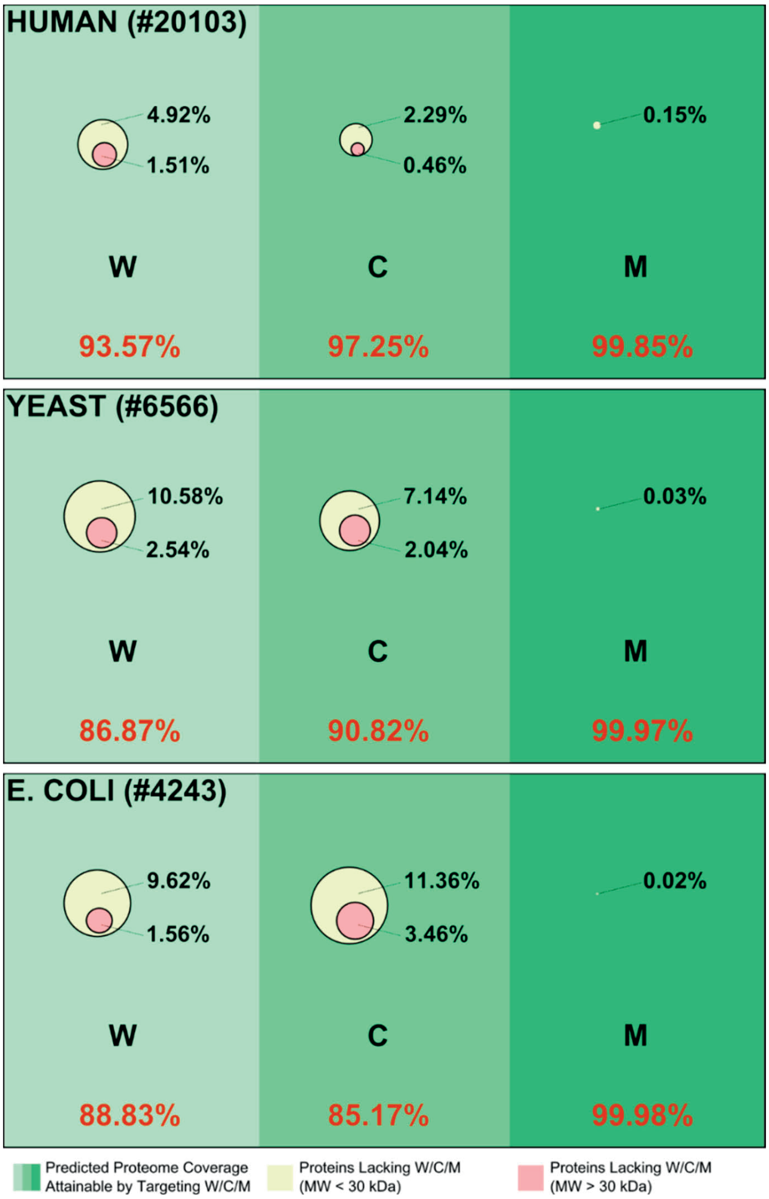
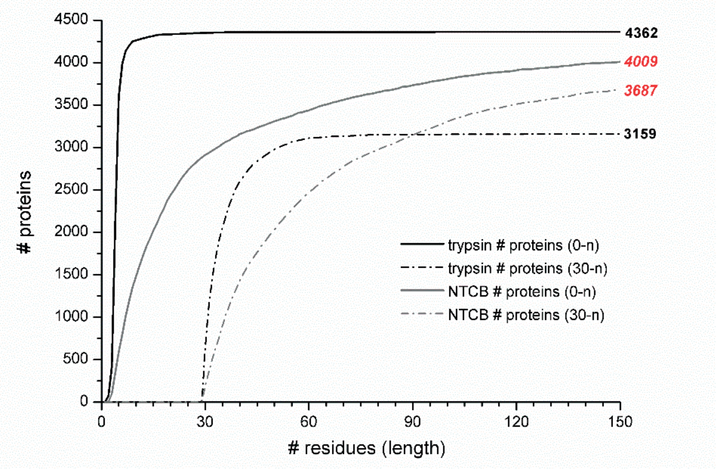
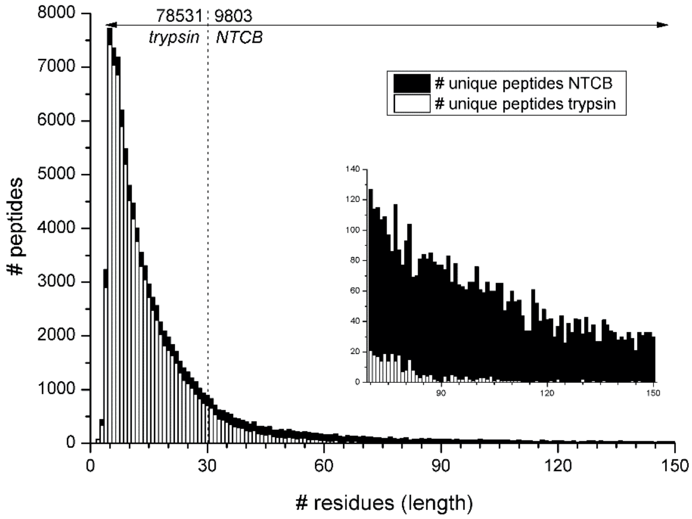

HUMAN (#20103)
4.92%
1.51%
2.29%
0.46%
0.15%
W
C
M
93.57%
97.25%
99.85%
YEAST (#6566)
10.58%
2.54%
7.14%
2.04%
0.03%
W
C
M
86.87%
90.82%
99.97%
E. COLI (#4243)
9.62%
1.56%
11.36%
3.46%
0.02%
W
C
M
88.83%
85.17%
99.98%
Predicted Proteome Coverage Attainable by Targeting W/C/M
Proteins Lacking W/C/M (MW < 30 kDa)
Proteins Lacking W/C/M (MW > 30 kDa)

Figure S3.

(a)


4500
4000
3500
3000
2500
2000
1500
1000
500
0
# proteins
0
30
60
90
120
150
# residues (length)
4362
4009
3687
3159
trypsin # proteins (0-n)
trypsin # proteins (30-n)
NTCB # proteins (0-n)
NTCB # proteins (30-n)


8000
7000
6000
5000
4000
3000
2000
1000
0
# peptides
0
30
60
90
120
150
# residues (length)
78531
9803
trypsin
NTCB
# unique peptides NTCB
# unique peptides trypsin
140
120
100
80
60
40
20
0
90
120
150

(b)

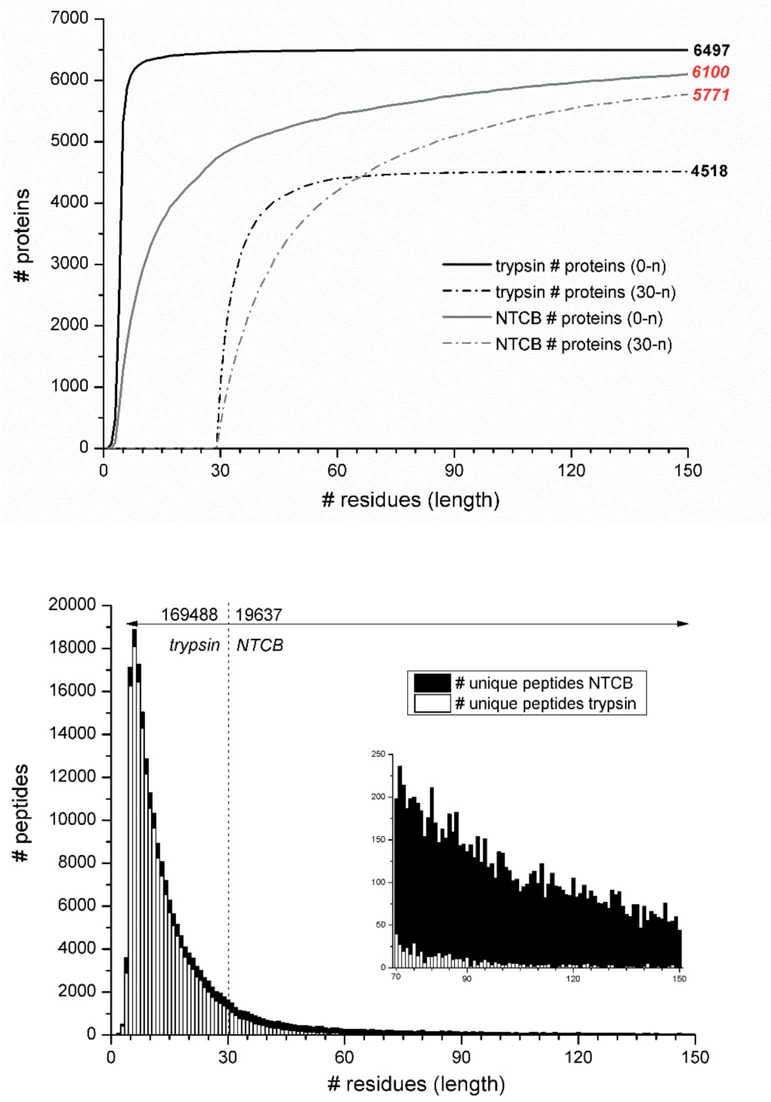

7000
6000
5000
4000
3000
2000
1000
0
# proteins
0
30
60
90
120
150
# residues (length)
6497
6100
5771
4518
trypsin # proteins (0-n)
trypsin # proteins (30-n)
NTCB # proteins (0-n)
NTCB # proteins (30-n)
20000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
# peptides
169488
19637
trypsin
NTCB
# unique peptides NTCB
# unique peptides trypsin
250
200
150
100
50
0
70
90
120
150
0
30
60
90
120
150
# residues (length)

Figure S4.

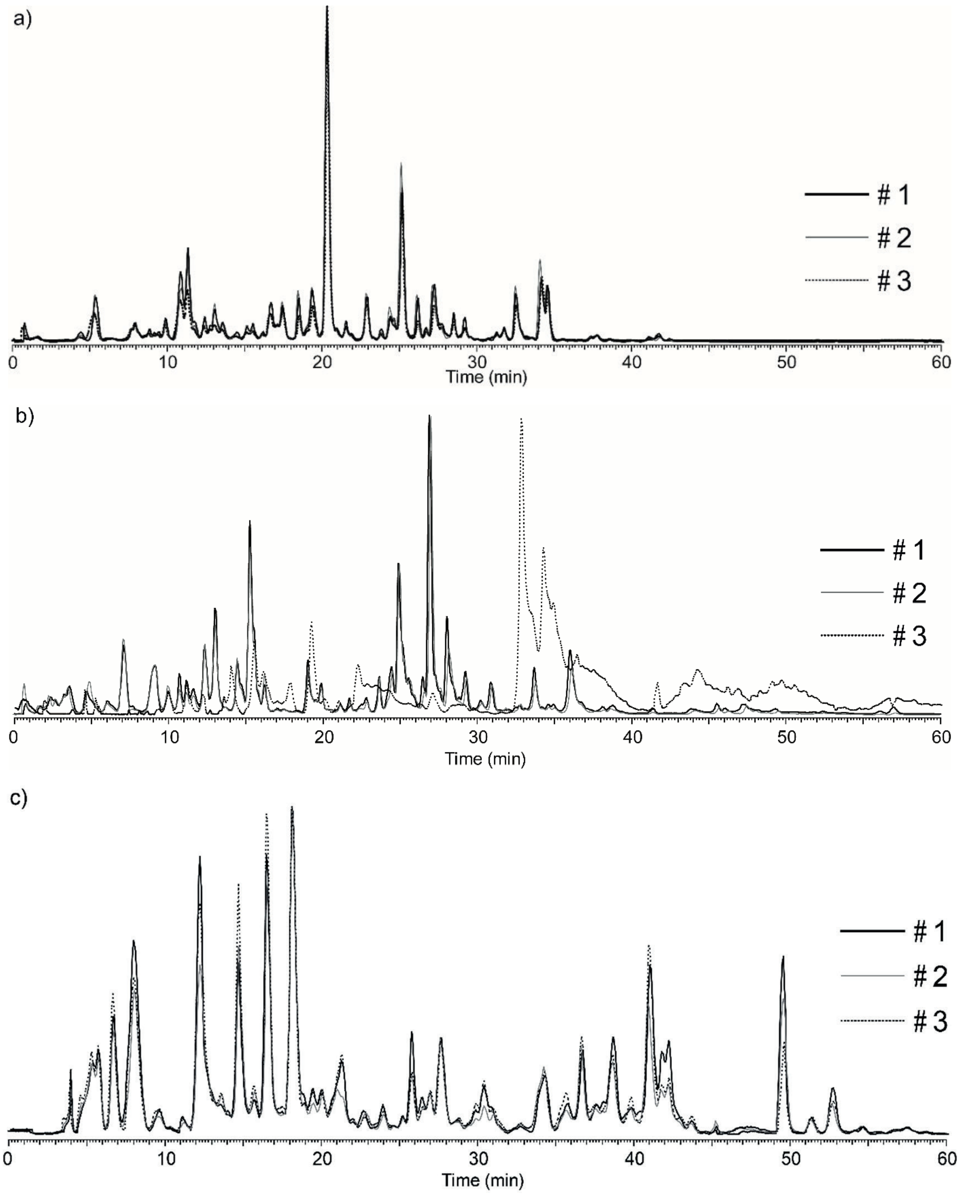
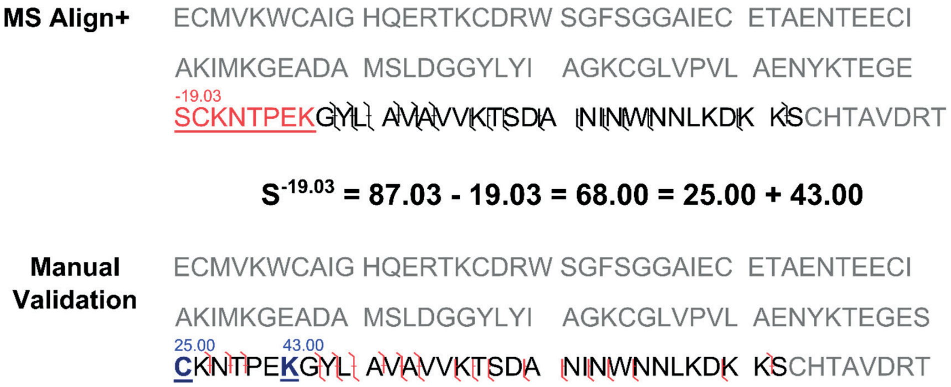

a)
# 1
# 2
# 3
0
10
20
30
40
50
60
Time (min)
b)
# 1
# 2
# 3
0
10
20
30
40
50
60
Time (min)
c)
# 1
# 2
# 3
0
10
20
30
40
50
60
Time (min)

Figure S5.


MS Align+
ECMVKWCAIG HQERTKCDRW SGFSGGAIEC ETAENTEECI
AKIMKGEADA MSLDGGYLYI AGKCGLVPVL AENYKTEGE
-19.03
SCKNTPEKGYL AVAVVKTSDA NINWNNLKDK KSCHTAVDRT
S-19.03 = 87.03 - 19.03 = 68.00 = 25.00 + 43.00
Manual Validation
ECMVKWCAIG HQERTKCDRW SGFSGGAIEC ETAENTEECI
AKIMKGEADA MSLDGGYLYI AGKCGLVPVL AENYKTEGES
25.00
43.00
CKNTPEKGYL AVAVVKTSDA NINWNNLKDK KSCHTAVDRT

6.2. MD with Cysteine-based chemical digestion: targeted analysis of Immunoglobulins G

Comprehensive structural analysis of antibodies, specifically of immunoglobulins G (IgGs), is indispensable due to their leading role as biotherapeutical drugs. The potential of electron transfer dissociation (ETD) Orbitrap Fourier transform (FT) mass spectrometry (MS) in characterizing IgGs is evaluated as a ‘way to go’ for structural analysis of these large biomolecules. Here, we proceed with deeper analysis of the IgG primary structure by further developing middle-down (MD) MS on a high-field Orbitrap Elite ETD FTMS.

IgGs became indispensable in treatment of variety of diseases; hence, the impetus in drug discovery is to extend the MS approaches for their characterization. Hypervariable domain is the fingerprint of an antibody thus, its correct assignment and sequence information is necessary for distinguishing targeted IgG between variety of sequence homologues, but with incomplete sequencing of Complementarity Determining Regions (CDRs), which are responsible for the IgG affinity and specificity. Currently employed bottom-up (BUP) avenues do not meet this requirement due to the frequent digestion, resulting in short peptides and oftentimes even a single CDR is split into two or more short peptides. Alternatively, analysis of intact antibodies (top-down, TD) is in principle the most adequate approach for structural characterization of IgGs. Relatively limited characterization of post-translational modifications (PTMs), important for modulating or impairing IgG biological activity, has been achieved. It is however limited to ~35% (REF Marshall, Luca TOF) sequence coverage, where most of the fragments derive from the Fc portion of the IgG, complementarity determining regions (CDRs) is obscured.

To address and overcome limitations inherent to BUP, and current technical restrictions of TD, we previously optimized extended bottom-up approach employing a non-selective protease (Sap9). Here, we introduce a chemoselective cysteine mediated digestion for IgG peptide mapping (Figure 6.1.).

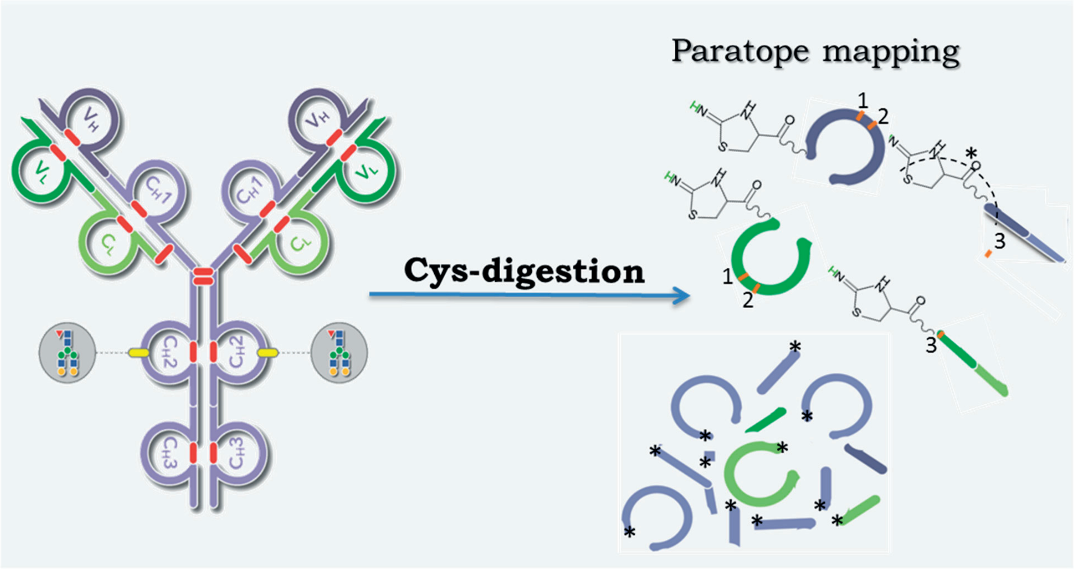


Figure 6.1. IgG characterization: proposed MD pipeline for paratope mapping by C-digestion.

To test viability of the named method in targeted structural analysis of antibodies, we digested therapeutical monoclonal antibody of class 1 (IgG1), Trastuzumab with NTCB agent as described in the Paper V in Chapter 5 and analysed via RP-LC-MS/MS (*vide supra*, Chapter 4). Obtained total ion chromatogram is shown in Figure 6.2. As expected, we successfully separated and identified long (6 kDa on average) peptides for both light and heavy chains of IgG.

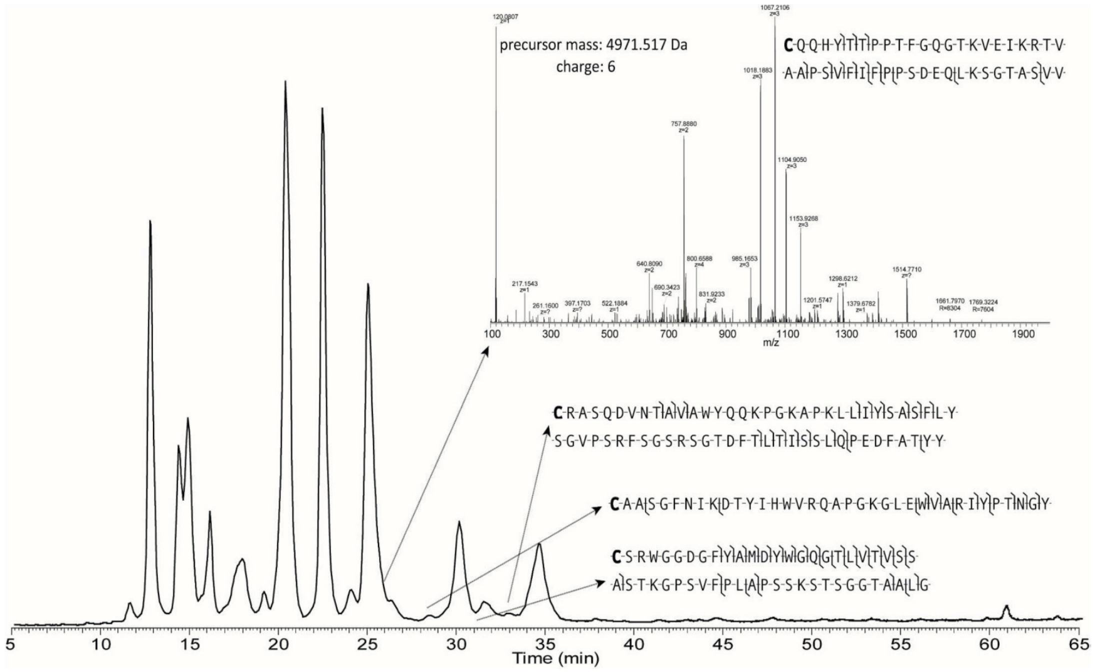


Figure 6.2. Extracted base peak chromatogram of Trastuzumab (IgG1) digested with NTCB. Inset: HCD mass spectrum of a 6+ precursor of the peptide identifying CDR3 of the light chain. Arrows indicate elution of the peptides containing adjacent CDR1 and 2 of the light (light orange) and the heavy chain (light green) as well as the peptide containing CDR3 of the heavy chain (light green) with their respective backbone fragmentation.

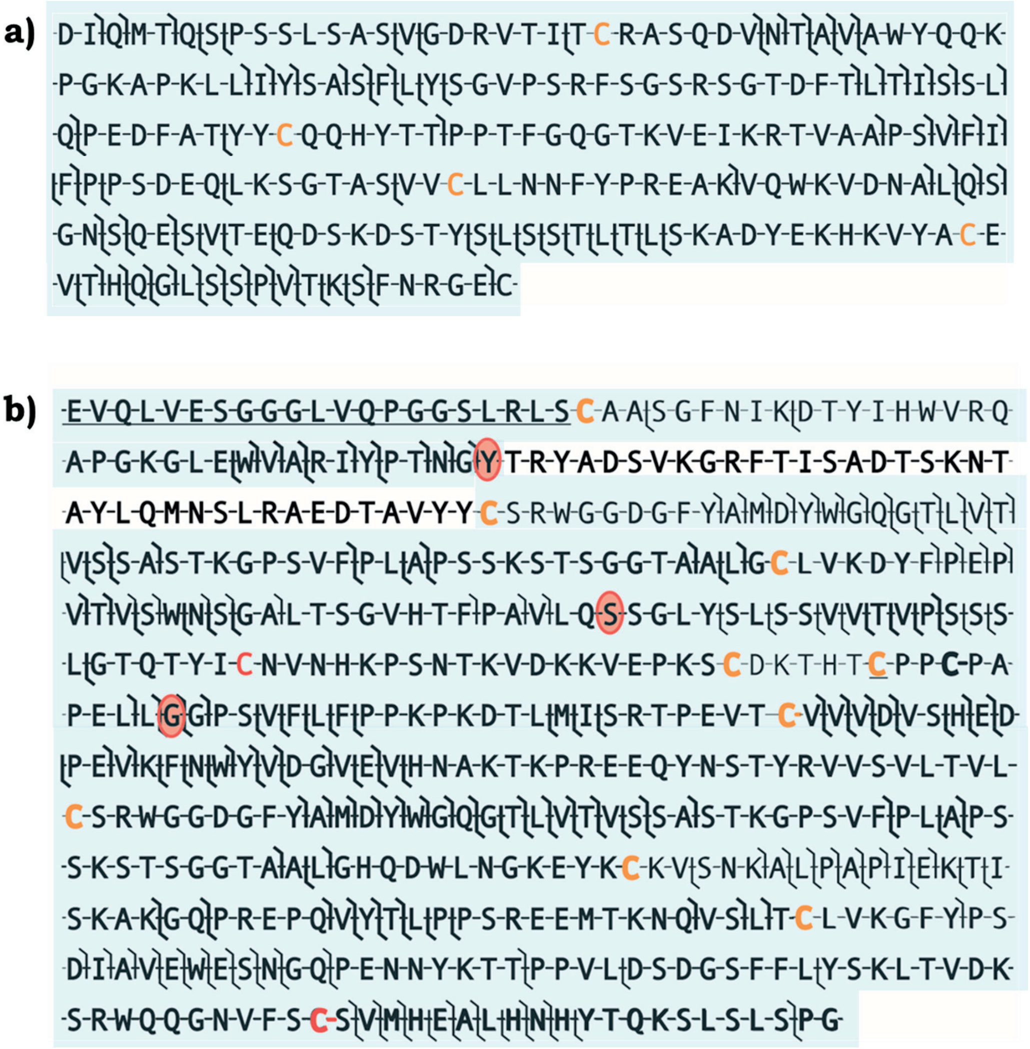

a)
D-I-Q-M-T-Q-S-P-S-S-L-S-A-S-V-G-D-R-V-T-I-T-C-R-A-S-Q-D-V-N-T-A-V-A-W-Y-Q-Q-K
P-G-K-A-P-K-L-L-I-Y-S-A-S-F-L-Y-S-G-V-P-S-R-F-S-G-S-R-S-G-T-D-F-T-L-T-I-S-S-L
Q-P-E-D-F-A-T-Y-Y-C-Q-Q-H-Y-T-T-P-P-T-F-G-Q-G-T-K-V-E-I-K-R-T-V-A-A-P-S-V-F-I
F-P-P-S-D-E-Q-L-K-S-G-T-A-S-V-V-C-L-L-N-N-F-Y-P-R-E-A-K-V-Q-W-K-V-D-N-A-L-Q-S
G-N-S-Q-E-S-V-T-E-Q-D-S-K-D-S-T-Y-S-L-S-S-T-L-T-L-S-K-A-D-Y-E-K-H-K-V-Y-A-C-E
V-T-H-Q-G-L-S-S-P-V-T-K-S-F-N-R-G-E-C
b)
E-V-Q-L-V-E-S-G-G-G-L-V-Q-P-G-G-S-L-R-L-S-C-A-A-S-G-F-N-I-K-D-T-Y-I-H-W-V-R-Q
A-P-G-K-G-L-E-W-V-A-R-I-Y-P-T-N-G-Y-T-R-Y-A-D-S-V-K-G-R-F-T-I-S-A-D-T-S-K-N-T
A-Y-L-Q-M-N-S-L-R-A-E-D-T-A-V-Y-Y-C-S-R-W-G-G-D-G-F-Y-A-M-D-Y-W-G-Q-G-T-L-V-T
V-S-S-A-S-T-K-G-P-S-V-F-P-L-A-P-S-S-K-S-T-S-G-G-T-A-A-L-G-C-L-V-K-D-Y-F-P-E-P
V-T-V-S-W-N-S-G-A-L-T-S-G-V-H-T-F-P-A-V-L-Q-S-S-G-L-Y-S-L-S-S-V-V-T-V-P-S-S-S
L-G-T-Q-T-Y-I-C-N-V-N-H-K-P-S-N-T-K-V-D-K-K-V-E-P-K-S-C-D-K-T-H-T-C-P-P-C-P-A
P-E-L-L-G-G-P-S-V-F-L-F-P-P-K-P-K-D-T-L-M-I-S-R-T-P-E-V-T-C-V-V-V-D-V-S-H-E-D
P-E-V-K-F-N-W-Y-V-D-G-V-E-V-H-N-A-K-T-K-P-R-E-E-Q-Y-N-S-T-Y-R-V-V-S-V-L-T-V-L
C-S-R-W-G-G-D-G-F-Y-A-M-D-Y-W-G-Q-G-T-L-V-T-V-S-S-A-S-T-K-G-P-S-V-F-P-L-A-P-S
S-K-S-T-S-G-G-T-A-A-L-G-H-Q-D-W-L-N-G-K-E-Y-K-C-K-V-S-N-K-A-L-P-A-P-I-E-K-T-I
S-K-A-K-G-Q-P-R-E-P-Q-V-Y-T-L-P-P-S-R-E-E-M-T-K-N-Q-V-S-L-T-C-L-V-K-G-F-Y-P-S
D-I-A-V-E-W-E-S-N-G-Q-P-E-N-N-Y-K-T-T-P-P-V-L-D-S-D-G-S-F-F-L-Y-S-K-L-T-V-D-K
S-R-W-Q-Q-G-N-V-F-S-C-S-V-M-H-E-A-L-H-N-H-Y-T-Q-K-S-L-S-L-S-P-G

Figure 6.3. Sequence coverage of light a) and heavy b) chains of Trastuzumab. (b, y) from a light chain and heavy chain identified peptides obtained from a NTCB digest and sequenced regions are indicated in light blue. Cysteines of ITZ peptide are indicated in orange, cysteines with DHA formation are shown in red. Circled residues indicate position of a non-specific cleavage introduced by base hydrolysis.

Due to the position of cysteines within particular IgG structure, generated peptides contained adjacent CDRs, revealing the connectivity information. Results obtained in this study indicate how employing NTCB-based cleavage could in future facilitate paratope discrimination of a single IgG from a mixture of IgGs, while reducing the number of required peptide IDs, and in turn, increase the obtainable sequence coverage per LC run.

Chapter 7. Towards top-down MS *via* MD

Hereinafter reported considerations are the initial premises later translated into research articles enclosed at the end of this Chapter:

- Revealing chain connectivity in monoclonal IgG1 by electron transfer dissociation Orbitrap FTMS on F(ab) subunits following KGP proteolysis (*Paper VI*)
- Middle-down analysis of IgG mixtures using electron transfer dissociation allows light and heavy chain pairing characterization (*Paper VII*)

7.1. Paper VI: Middle-down analysis of IgG mixtures using electron transfer dissociation allows light and heavy chain pairing characterization

Middle-Down Electron Transfer Dissociation of IgGs Enables Light and Heavy Chain Pairing Characterization

Daniel Ayoub[1#], Kristina Srzentić[1], Luca Fornelli[1$], Anna A. Lobas[2,3], Konstantin Aizikov[4], Alain Beck[5], Mikhail V. Gorskhov[2,3], and Yury O. Tsybin[6*]

[1] Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

[2] Institute for Energy Problems of Chemical Physics, Russian Academy of Sciences, Leninskii Prospect 38, Bldg. 2,119334 Moscow, Russia

[3] Moscow Institute of Physics and Technology (State University), 9 Institutskiy per., 141707 Dolgoprudny, Moscow Region, Russia

[4] Thermo Scientific, Bremen, Germany

[5] Centre d'Immunologie Pierre Fabre (CIPF), Saint-Julien-en-Genevois, France

[6] Spectroswiss Inc., EPFL Innovation Park, 1015 Lausanne, Switzerland

[*] Correspondence should be addressed to Dr. Yury O. Tsybin, Spectroswiss Inc., EPFL Innovation Park, Building I, 1015 Lausanne, Switzerland. E-mail: tsybin@spectroswiss.ch

[#]Current affiliation: Luxembourg Clinical Proteomics Center (LCP), Luxembourg Institute of Health, Strassen, Luxembourg

[$]Current affiliation: Proteomics Center of Excellence, Northwestern University, Evanston, IL, USA

Running title: Middle-Down ETD for Chain Pairing in IgG

Abstract

Identification of antigen-specific immunoglobulins G, IgGs, from antibody repertoire in the serum of immunized animals requires information on pairing of light and heavy chains in IgGs. Previously, bottom-up proteomics has been employed to guide the IgG identification in these experiments. However, digestion of IgGs to small peptides following the bottom-up approach scrambles information on the chain pairing. Here, we put forward a middle-down approach that should provide the required IgG chain pairing information with high selectivity and specificity. Briefly, here the IgGs are quickly digested above the hinge region with papain into large 50 kDa subunits, Fab and Fc. Thus obtained Fab subunits are disulfide bond-bound complexes of a complete light chain and the extended N-terminal part, Fd, of a heavy chain. Intact Fab subunits are then submitted to gas-phase fragmentation using electron transfer dissociation which results in formation of structure-specific product ions. Some of these product ions contain parts of both chains still bound by an intact S-S bond. Importantly, cleavage sites for some of these "pairing-specific" product ions are located in the variable domains of IgGs, increasing the confidence of pairing characterization. We validate the proposed approach for analysis of a single IgG1, single IgG4, a mixture of three IgG1 proteins which provide baseline separation of Fab subunits by liquid chromatography, and a mixture of two IgG1 proteins which provide Fab subunits co-eluting from liquid chromatography. The reported proof-of-principle experiments constitute a first step toward the use of the described middle-down approach for analysis of significantly more complex antibody mixtures with the final goal of improving IgG drug discovery.

Keywords: top-down mass spectrometry; immunoglobulin G1, IgG1; chain pairing; Fourier transform mass spectrometry, FTMS; Orbitrap; transient averaging;

Introduction

The development of monoclonal antibody (mAb)-based therapeutics requires their thorough and deep characterization due to mAb complexity as bio-products [1-3]. MAbs and their derivatives, unlike small molecule therapeutics are very heterogeneous and present in a large number of variants [1]. For safety requirements, these variants need to be thoroughly identified and characterized to comply with sanitary authorities guidelines. Mass spectrometry (MS)-based techniques are implemented throughout all stages of mAbs production and development. It provides valuable information about mAb structure, modifications and heterogeneity, ranging from high order structure and conformation to sequence and co-/post-translational modifications (PTM) mapping.

Monoclonal antibodies used as therapeutics are from the immunoglobulin G (IgG) class. IgGs are tetrameric glycoprotein complexes each composed of two ~50 kDa heavy chains and two ~25 kDa light chains. Each light chain is linked by a disulfide bond to a heavy chain and the heavy chains are linked together by two to four disulfide bonds depending on the IgG isotype. Four IgG isotypes exist and are defined by their heavy chain amino acid sequence: IgG1, IgG2, IgG3 and IgG4. IgG3 are not used as therapeutics due to a rapid clearance (7 vs 21 days compared to IgG1 in some cases). Disulfide bridges (16 for IgG1 and IgG4; 18 for IgG2) and non-covalent interactions maintain their three-dimensional structure (H2L2 homoheterodimers). The heavy and light chains are linked by one disulfide bond and the heavy chains by two (for IgG1 and IgG4) or three (for IgG2) disulfide bonds located in a short hinge domain (Hi). The other 12 cysteine bridges are intramolecular and delimit six different globular domains: one variable (VL) and one constant for the light chains (CL) and one variable (VH) and three constant for the heavy chains (CH1, CH2, and CH3). Antigen binding is mediated by the variable domains, mainly by three loops connecting individual β-strands in each domain (CDR). The mass measurement of the intact molecule is a basic and fast way to assess a mass profile of the whole antibody. It allows the confirmation of the elemental amino-acid composition by comparing expected and measured masses. The major glycoforms are also resolved. The resolution and the mass accuracy of the mass spectrometer along with the quality of the sample have a direct effect on the quality of the mass measurement. The molecular weight measurement of intact IgGs gives an overall profile picture of the protein. However, it does not provide structural resolution. Reducing the complexity of intact IgGs by chopping them into smaller subunits before mass analysis provides more resolution for structural heterogeneities. By analogy to bottom-up MS, Zhang *et al*. introduced the term middle-up MS to designate the analysis of large subunits of IgGs [4]. This is not to be confused with middle-down MS which refers to the MS/MS sequencing and analysis of these large subunits. These smaller subunits can be easily obtained by reduction of the disulfide bonds of IgGs dissociating them into two 25 kDa LCs and the two 50 kDa HCs. These are subsequently analyzed by LC-MS. Over the time, the mass analysis of separated light and heavy chains of IgGs has become common in biopharmaceutical laboratories. If the reduction is performed in denaturing condition, i.e. in the presence of chaotroping agents such as urea or guanidine-HCl, all disulfide bonds would be reduced [5-7]. The inter-chain disulfide bonds can also be selectively reduced in the absence of denaturing conditions leaving the intramolecular bonds intact [8]. Middle-up MS analysis can also involve limited proteolysis of the heavy chain, in non-denaturing conditions, yielding both two Fab fragments (50 kDa each) and an Fc (50 kDa) fragment (in the case of upper hinge cleavage), or a Fab'2 (100 kDa) fragment with two half Fc (25 kDa) fragments (in the case of cleavage under the hinge). These can be further disulfide bond reduced to obtain three ~25 kDa subunits (light chains, half Fc and Fd). Several proteases have been used for these middle-up approaches with the most common ones being papain [9], pepsin [10], ficin [11] and endoprotease LysC [12]. Papain cleaves IgGs just above the hinge region generating from each IgG molecule two fragment antigen binding subunits (Fab) and an Fc subunit. The two generated Fabs are identical except for the case of bispecific antibodies. Every Fab fragment is constituted of a light chain and the N-terminal half of a heavy chain called the Fd domain linked together by a disulfide bond. Each of the light chain and the Fd contains two intra-chain disulfide bonds. Papain and ficin generate Fab fragments in the presence of cysteine, whereas Fab'2 fragments can also be generated in the absence of cysteine. These two proteases have the advantage of cleaving IgGs from a broad range of species. In addition, ficin cleaves the Fc into multiple small fragments. Pepsin on the other hand generates Fab'2 fragments at acidic pH, it does not require cofactors and thiols but suffers from low yield. LysC generates generally Fab fragments at pH 8, however, it does not cleave all IgG isotypes. IgG2s for example resist digestion under native conditions. Limited proteolysis by these enzymes followed by reduction frees the light chain and the Fd part from the glycosylation heterogeneity (except in the case of Fab glycosylation [13]) and the half Fc fragments bear only one glycosylation site. This, conjugated to their smaller size when compared to intact IgGs (and even intact heavy chains) renders LC-MS analysis easier and potentially with higher mass accuracy giving access to a more straightforward characterization of micro-heterogeneities. Recently, a new bacterial cysteine protease, IdeS (Immunoglobulin-degrading enzyme of Streptococcus pyogenes), is becoming more popular due to its high specificity and yield. It has the advantage of being rapid (30 mins for complete cleavage), not requiring any cofactors and for cleaving at pHs of formulation buffers therefore limiting artifact introduction [13-15].

Here, we describe a middle-down tandem mass spectrometry approach enabling a structural analysis of intact 50 kDa Fab subunits generated by papain digestion above the hinge region of single IgG1 and single IgG4 proteins, as well as simple mixtures of two to three IgG1 proteins. We put a particular focus on ETD-derived product ions that contain parts of both light and heavy chains of IgG proteins. Thus obtained "chain pairing" information is of a potential importance in drug discovery where antigen-specific IgGs need to be identified from complex mixtures of antibodies present in serum of immunized animals, e.g., rabbits [16].

Experimental methods

Reagents. Water, acetonitrile (ACN), trifluoroacetic acid (TFA) and isopropanol (IPA) were purchased in LC-MS purity grade. Water and ACN were obtained from Fluka Analytical (Buchs, Switzerland), formic acid (FA) from Merck (Zug, Switzerland), IPA from Thermo Fisher Scientific (Switzerland), and guanidinium chloride (GdnCl) from Carl Roth (Germany). Tris-HCl, EDTA, papain and L-cysteine-HCl were purchased from Sigma-Aldrich. Therapeutic monoclonal antibodies of the IgG1 class, adalimumab (Humira, Abbot Laboratories), trastuzumab (Herceptin, Genentech), palivizumab (Synagis, MedImmune), and rituximab (Rituxan, Roche), and IgG4 class, natalizumab (Tysabri, Biogen Idec) were the European Medicines Agency approved versions and formulations, available commercially to the general public.

Sample preparation. Samples containing single or mixtures of antibodies are prepared using 100 µg of each IgG. Antibodies are first diluted to 1.3 mg/mL in digestion buffer: Tris-HCl 100 mM, EDTA 4 mM, L-cysteine-HCl 5.5 mM, pH 7.6. Papain digestion performed using an enzyme to substrate ratio of 1:100 at 37° C for two hours. The mixture is then buffer-exchanged to ammonium acetate 50 mM using Zeba 0.5 mL desalting spin columns (Pierce, Thermo Fisher Scientific). The solution is then acidified to pH 2-4 using TFA and analyzed by LC-MS.

Liquid chromatography – mass spectrometry. The chromatographic separation of IgG proteolytic fragments was performed using an Ultimate 3000 LC system (Thermo Scientific, Amsterdam, The Netherlands) under UPLC conditions. A combination of reversed phase C4 trap-column (Acquity UPLC PrST C4 VanGuard pre-column, 2.1x5 mm, particle size 1.7 µm, pore size 300 Å, Waters) and C4 column (Acquity UPLC PrST C4, 1x150 mm, particle size 1.7 µm, pore size 300 Å, Waters) was employed to ensure on-line IgG fragment desalting and separation. For each injection, 1 µg of digestion product was loaded on the column, heated at 65° C. After initial loading at 5% solution B (organic phase), a gradient of solution B from 15 to 45% in 15 minutes was used at a flow rate of 100 µl/min. Solution A consisted of 0.1% of FA in water, whereas solution B was composed of 39.9% IPA, 60% ACN, and 0.1% FA. The LC column outlet was on-line coupled with the electrospray (ESI) source of the mass spectrometer. MS experiments were performed on an ETD-enabled hybrid linear ion trap high-field Orbitrap FT mass spectrometer (LTQ Orbitrap Elite FTMS, Thermo Scientific, Bremen, Germany). Separate LC-MS experiments were dedicated to record broadband mass spectra and ETD tandem mass spectra. Instrumental parameters were set as follows: S-lens RF level was set to 70%, the temperature of heated transfer capillary was 350° C, microESI source (IonMax source, Thermo Scientific) was used with a 3.7 kV potential, and sheath gas was set to 20 and auxiliary gas to 10 arbitrary units. All the mass spectra were acquired using ion detection in the Orbitrap FTMS, in the *m/z* range 200-2000. For broadband and tandem mass spectrometry, we both reduced the gas (N2) in the HCD cell to provide "delta pressure" in the Orbitrap detector region of about 0.1E-10 torr, and applied "HCD trapping", which is a trapping and a temporary ion storage in the HCD cell before ion transmission to the Orbitrap mass analyzer through the C-trap [17]. Broadband mass spectra were recorded with either 15'000 or 120'000 resolution at 400 *m/z*, with a target value for the automatic gain control (AGC) of 1 million charges in either MS or MS/MS modes. For ETD experiments, precursor ions were isolated in the high pressure chamber of the LTQ and subsequently subjected to ETD MS/MS. The AGC target value for fluoranthene radical anions was set to 7-8E5 charges, with anion maximum injection time of 50 ms. ETD duration (i.e., ion-ion interaction time) was progressively increased from 3 ms to 9 ms in consecutive experiments. Product ion detection in the Orbitrap mass analyzer was performed with 120'000 resolution at 400 *m/z* (enhanced FT, eFT, enabled). All Orbitrap FTMS scans were recorded averaging 10 microscans to improve the signal-to-noise ratio (SNR) via on-board time-domain (transient) averaging prior to eFT signal processing. Isolation windows for ETD of IgG fragments included one charge state per precursor ion (isolation width: 15 Th) in the case of bevacizumab and trastuzumab, or multiple charge states for adalimumab (isolation width: 100 Th and wider).

Data processing and tandem MS analysis. Data were analyzed both as single LC-MS/MS runs, and after additional data processing aimed at improving SNR of tandem mass spectra. In the latter case, time-domain (transient) signals recorded in separate LC-MS/MS experiments were processed as previously described for top-down LC-MS/MS of mAbs [18]. Briefly, Orbitrap FTMS transient signals were first recorded in MIDAS *.dat format using advanced user interface installed on Orbitrap FTMS under a license from manufacturer [19]. Thus obtained transients were grouped according to the IgG fragment type and duration of ETD MS/MS, averaged, and finally subjected to time-to-frequency conversion with the eFT procedure using proprietary manufacturer's tools. The resulting standard Thermo .RAW files could be then opened and processed with commercial XCalibur software (Thermo Scientific), and were thus ready for the data analysis. For each ETD duration, an averaged mass spectrum for each IgG fragment was obtained. In addition, a total tandem mass spectrum was built by averaging all the transients (i.e., transients derived from different ETD duration experiments) available for a single IgG fragment. Data analysis was performed using Xtract and ProSightPC 3.0 (Thermo Scientific) [20]. First, Xtract was used for tandem mass spectra deconvolution, peak centroiding, and peak picking. Then, cleavage sites were assigned with ProSightPC using 15 ppm mass tolerance. For disulfide-bridged ETD product ions, the searches were performed using an in-house developed algorithm.

Results and discussion

LC-MS analysis of papain digested IgG mixture. In this experiment a mixture of three therapeutic monoclonal antibodies was used as a model sample. The mixture consisted of three IgG1 class monoclonal antibodies: adalimumab, trastuzumab, and palivizumab. After digestion using papain, the IgG mixture was directly analyzed by LC-MS. **Figure 1** shows the total ion chromatogram of the mixture. Expectedly, the Fc subunits of all three IgGs co-eluted as they share nearly the same sequences. The Fc subunits contain glycosylation and other information that are important for effector functions, interaction with Fc receptors and stability, but they provide no information about Lc and Hc pairing as they originate solely from the heavy chains. In the context of Fab analysis, Fc subunits can be removed from the sample using protein A or protein G capture to simplify the analysis. The three Fab subunits originating from the three IgGs were near baseline separated by liquid chromatography. Maintaining the LC column at a constant temperature (65 °C) ensured the high elution reproducibility necessary for subsequent LC-time scheduled ETD fragmentation. The mass spectra of the three Fabs subunits show charge state distributions stretching mainly from *m/z* 1000 to 2500 and centered around charge states 33-35+, Figure 1 insets. The measured molecular masses (47’637 Da, 47’681 Da, and 47’528.5 Da) are consistent with the theoretical masses calculated from the sequences of trastuzumab, adalimumab and palivizumab, respectively. For palivizumab, the heavy chain’s N-terminal glutamine is cyclized into a pyroglutamic acid, a common post-translational modification in IgGs with N-terminal glutamines.

ETD LC-MS/MS analysis of papain generated IgG1 Fab subunits mixture. The ETD settings were first optimized to maximize sequence coverage. Namely, like previously reported by Fornelli *et al.* for 25 kDa IgG subunits [15], high *m/z* product ion transmission was improved by reducing the nitrogen gas pressure in the HCD cell to obtain a “delta pressure” in the Orbitrap detector of 0.1x10-10 Torr and by applying HCD trapping. HCD trapping consists of temporarily storing the ions in the HCD cell before their transmission to the Orbitrap mass analyzer via the C-trap. Several ETD reaction times ranging from 3 to 25 ms were tested. The 10 and 15 ms reaction times we found to be the most effective to maximize sequence coverage and provide complementary product ions. The high reproducibility of the Fab subunits chromatographic elution allowed for setting the instrument to perform ETD MS/MS in a time-scheduled fashion, i.e. only during selected time windows corresponding to the elution times of the different Fab IgG subunits. Isolation windows of 200 Th to include ~5 charge states of precursor ions were used to increase ETD efficiency, Figure 1 insets. The transients from 15 LC-MS/MS runs using 10 ms reaction time and 15 LC-MS/MS runs using 15 ms reaction time were summed together for each Fab elution peak. The resulting ETD mass spectra were searched using ProSight PC to assign the peaks. The searches were performed independently for the light chain and the Fd subunit considering that the disulfide bond linking the two chains is cleaved upon ETD fragmentation.

As expected, sequence coverages were between 21 and 30 % for the different chains of the different Fabs subunits. This is comparable to results obtained on intact IgGs which, like Fab subunits, retain highly structured areas mainly in correspondence to the immunoglobulin domain and disulfide bond protected areas [18, 21]. Higher sequence coverage can be obtained if the disulfide bonds were reduced as shown in a previous paper using IdeS digestion followed by DTT reduction to produce 25 kDa IgG subunits [15]. Unlike in bottom-up approaches where the sequence coverage is calculated based on the identified peptides without regards to the cleavage sites assigned upon MS/MS, in top-down and middle-down experiments the sequence coverage is calculated as the ratio of assigned cleavage sites to the total number of possible cleavage sites, **Figure 2**. The fragmentation map of adalimumab is presented in Figure 2 top panel. The CDR 3 of the light and heavy chains are covered with cleavage sites corresponding to the complementary *c* and *z* ions. CDR 3 is the part of the sequence that is clone-specific to the IgG and is therefore crucial for clone identity determination. CDR 1 and 2 of both chains are also covered by at least one fragment.

Disulfide bonds are known to be cleaved using ETD as demonstrated by the 23.9 % and 25.7 % sequence coverage obtained for adalimumab's light chain and half heavy chain respectively. However, the propensity of ETD to cleave disulfide bonds in IgG is not known and, presumably, is less than 100%. Therefore, in the case of two independent polypeptide chains linked together by a disulfide bond, ETD may generate product ions from each chain that are linked together by a preserved disulfide bond. We believe this is the case for IgGs and their Fab subunits. According to our hypothesis, these product ions can provide light and heavy chain pairing information in the case of IgG mixtures. Nevertheless, these product ions bound by disulfide bridges are not accounted for by classical top-down MS or proteomics search algorithm. Furthermore, they would have a high mass and present in high charge states which would further complicate the challenge of assigning them. Therefore, we in-house developed a dedicated algorithm that would i) calculate the masses of all possible disulfide bond-bridged product ions based on the sequences of the light and heavy chains; and ii) compare them to the product ions from the experimental ETD MS/MS spectra within a certain mass error. The assignments of the product ions are then manually controlled and the ambiguous matches are discarded, namely masses that would match several possible product ions. **Figure 3** shows an example of such ions with expanded views on simple ions and disulfide bond-bridged ions. It is important here to highlight the high resolution of the employed mass spectrometer that allows isotopic resolution of these large product ions and the benefits of the improved sensitivity brought by transient averaging over multiple LC-MS/MS runs. The fragmentation map obtained is presented in Figure 2 bottom panel. When considering only product ions that consist of a light chain fragment and an Fd fragment linked together by a disulfide bond, the sequence coverage is nearly 16 % for both chains. All ions containing inter-chain disulfide bonds are logically *z*-type ions since in IgG1s, the two chains, Lc and Fd are linked by their C-termini: namely the C-terminal cysteine of the Lc and the cysteine of the Fd which is the fifth C-terminal residue. Many of the ETD cleavage sites identified using these disulfide bond-bridged product ions have already been observed in the first search using ProSight PC, most of them are complementary *z*-type ions. The fact that the most simple ions identified are *c*-type ions and the number of complementary *z* type ions identified contains an intra-chain disulfide bond might indicate that in many cases, upon ETD backbone cleavage, the inter-chain disulfide bond is conserved. When accounting for the new cleavage sites corresponding to these newly assigned ions, the sequence coverage reaches 26.8 % for the Lc and 31 % of the Fd. While this increase in sequence coverage is not high, identifying several ions sharing the same ETD cleavage sites does strengthen the confidence in the assignments. More interestingly, the CDR 3 domains of the Lc and the Fd are also well covered by several cleavage sites assigned from simple and disulfide bond-bridged product ions. This would confirm the IgG clone identity and, with the intact masses measured allows access to the Lc and Hc pairing information.

ETD LC-MS/MS analysis of papain generated IgG4 Fab subunit. After achieving the proof of concept for IgG1, the compatibility of the approach for other IgG subclasses was evaluated. The disulfide bond linkage is one of the main differences between the four IgG subclasses. IgG1s and IgG4s contain two disulfide bridges in the hinge region linking the two heavy chains. More related to Fab subunits is the linkage between the light chain and the heavy chain. In IgG1s, the disulfide bond linking the Lc to the Hc is between the fifth and last cysteine of the light chain and the fifth cysteine of the Fd, whereas in IgG2 and IgG4 the link is between the fifth cysteine of the light chain and the third cysteine of the heavy chain, **Figure 4**. Natalizumab was used here as an IgG4 benchmark to test the method. ETD was performed exactly the same way as for the IgG1 Fab mixture. **Figure 5** presents the fragmentation map obtained with the Prosight PC search, i.e. accounting only for product ions with the disulfide bond between the two chains cleaved. The sequence coverage achieved is near 30 % for both chains with both CDR 3 domains well covered. When running the search for product ions with an inter-chain disulfide bond, the sequence coverage increases to nearly 35 % and 36 % for the Lc and Fd respectively. Note that for IgG4s like natalizumab, unlike for

IgG1s, not all the disulfide linked ions are *z* ions: *z* ions from the light chain can be linked with *z* ions from the Fd for cleavage before cysteine 96 (form the N-terminus) or with *c* ions for cleavages to after cysteine 96. These results demonstrate that the approach can work for all isotypes as IgG2, IgG3 and IgG4 share the same type of Fab disulfide linkage.

ETD LC-MS/MS of co-eluting papain generated Fab subunit. In the first papain digested IgG mixture, all Fab subunits where separated by chromatography. This allows for the use of broad isolation windows to maximize ETD efficiency and sequence coverage as the generated product ions will correspond to one Fab subunit specie. However, separating different Fab subunits in a mixture using chromatography is not always trivial and co-elution of different Fab molecules is very common. For highly complex mixtures, other separation and fractionation techniques such as ion exchange chromatography or off-gel isoelectrofocusing can be used prior to reverse phase LC-MS analysis. Nevertheless, co-elution can still occur for molecules sharing close pI and hydrophobicity factors. In that case, one can use narrow isolation windows for one charge state from each co-eluting IgG and obtain MS/MS spectra that correspond to only one Fab subunit. However, this will translate in less product ions and reduced signal to noise ratio. Another alternative is to isolate for each molecule several charge states and fragment them independently and then sum the transients for each Fab all together. This on an LC-MS peak elution profile is very challenging as only a few transients can be acquired which renders increasing the number of runs necessary. A mixture of trastuzumab and rituximab, two IgGs that provide co-eluting Fab subunits in the LC-MS setting used were analyzed using a broad isolation window of 230 Th to test if the quality of the Fab sequencing would be affected and if it would still be possible to identify ions connected by an inter-chain disulfide bond, **Figure 6**.

The sequence coverage for trastuzumab was 23 and 26 % for the Lc and Fd respectively when not considering intra-chain disulfide bond linked ions.

For rituximab 22 and 24 % sequence coverage for the Lc and the Fd were obtained, respectively. When considering only inter-chain disulfide bond linked *z*-ions, about 17 % sequence coverage was obtained for both Lc and Fd for both rituximab and trastuzumab. In all cases CDR-3 domains were covered with several product ions. Rituximab presents cyclized glutamine to pyroglutamic acid on both chain. These results show that this co-isolation of several charge states from both IgGs did not affect the outcome of the sequence coverage in either search (with and without disulfide linked ions). This proves that the sequencing and light chain and heavy chain pairing determination is possible even in the case of co-elution proving the validity of the middle-down approach at the Fab level for such IgG mixtures.

Conclusions

Middle-down ETD-MS/MS at the Fab level constitutes an interesting approach for the characterization of monoclonal antibodies. Importantly, Fab subunit conserves the pairing information of light and heavy chain, which is otherwise lost in analysis of reduced or further digested IgG. The method achieves sequence coverages similar to those reported in intact IgG top-down characterization studies. Additionally, the reduced, 50 kDa, Fab subunit size compared to the 150 kDa intact IgG is more readily analyzed with high-resolution MS and MS/MS. The complementarity determining regions (CDR) 3 of both the light and the heavy chains are well sequenced. Furthermore, searching for disulfide bond-bridged ions allows for increasing sequence coverage and along with the intact Fab molecular mass and CDR-3 identity can constitute a valuable tool to determine light and heavy chain pairing in mixtures of antibodies. The latter information is practically unachievable with top-down MS of intact IgGs, and even more so from the mixtures of intact IgGs. The results show a strong proof of principle and that the approach can be applied to all IgG subclasses and even in the case of co-eluted and co-isolated IgGs. The long-term goal here is to apply this approach to complex mixtures of serum antibodies and search the mass spectra against a reference B cell genome-derived database [16].

Acknowledgements

We express our gratitude to Thermo Fisher Scientific Inc. for providing access under license to LTQ Orbitrap Elite FTMS transient signals and related functionality enabled by the developers' kit. This work was supported by the Swiss National Science Foundation (Projects 200021-125147 and 128357 to Y.O.T.) and the European Research Council (ERC Starting Grant 280271 to Y.O.T.).

References

[1] Beck, A.; Wagner-Rousset, E.; Ayoub, D.; Van Dorsselaer, A.; Sanglier-Cianférani, S.: Characterization of Therapeutic Antibodies and Related Products. *Analytical Chemistry* **85,** 715-736 (2013)

[2] Beck, A.; Sanglier-Cianférani, S.; Van Dorsselaer, A.: Biosimilar, Biobetter, and Next Generation Antibody Characterization by Mass Spectrometry. *Analytical Chemistry* **84,** 4637-4646 (2012)

[3] Beck, A.; Wurch, T.; Bailly, C.; Corvaia, N.: Strategies and challenges for the next generation of therapeutic antibodies. *Nat Rev Immunol* **10,** 345-352 (2010)

[4] Zhang, Z.; Pan, H.; Chen, X.: Mass spectrometry for structural characterization of therapeutic antibodies. *Mass Spectrometry Reviews* **28,** 147-176 (2009)

[5] Wang, L.; Amphlett, G.; Lambert, J. M.; Blättler, W.; Zhang, W.: Structural Characterization of a Recombinant Monoclonal Antibody by Electrospray Time-of-Flight Mass Spectrometry. *Pharm Res* **22,** 1338-1349 (2005)

[6] Rehder, D. S.; Dillon, T. M.; Pipes, G. D.; Bondarenko, P. V.: Reversed-phase liquid chromatography/mass spectrometry analysis of reduced monoclonal antibodies in pharmaceutics. *Journal of Chromatography A* **1102,** 164-175 (2006)

[7] Chelius, D.; Jing, K.; Lueras, A., *et al.*: Formation of Pyroglutamic Acid from N-Terminal Glutamic Acid in Immunoglobulin Gamma Antibodies. *Analytical Chemistry* **78,** 2370-2376 (2006)

[8] Yu, L.; Remmele, R. L.; He, B.: Identification of N-terminal modification for recombinant monoclonal antibody light chain using partial

reduction and quadrupole time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry* **20,** 3674-3680 (2006)

[9] Yan, B.; Valliere-Douglass, J.; Brady, L., *et al.*: Analysis of post-translational modifications in recombinant monoclonal antibody IgG1 by reversed-phase liquid chromatography/mass spectrometry. *Journal of Chromatography A* **1164,** 153-161 (2007)

[10] Gadgil, H. S.; Bondarenko, P. V.; Pipes, G.; Rehder, D.; McAuley, A.; Perico, N.; Dillon, T.; Ricci, M.; Treuheit, M.: The LC/MS analysis of glycation of IgG molecules in sucrose containing formulations. *Journal of Pharmaceutical Sciences* **96,** 2607-2621 (2007)

[11] Mariant, M.; Camagna, M.; Tarditi, L.; Seccamani, E.: A new enzymatic method to obtain high-yield F(ab)2 suitable for clinical use from mouse IgGl. *Molecular Immunology* **28,** 69-77 (1991)

[12] Gadgil, H. S.; Bondarenko, P. V.; Pipes, G. D.; Dillon, T. M.; Banks, D.; Abel, J.; Kleemann, G. R.; Treuheit, M. J.: Identification of cysteinylation of a free cysteine in the Fab region of a recombinant monoclonal IgG1 antibody using Lys-C limited proteolysis coupled with LC/MS analysis. *Analytical Biochemistry* **355,** 165-174 (2006)

[13] Ayoub, D.; Jabs, W.; Resemann, A., *et al.*: Correct primary structure assessment and extensive glyco-profiling of cetuximab by a combination of intact, middle-up, middle-down and bottom-up ESI and MALDI mass spectrometry techniques. *mAbs* **5,** 699-710 (2013)

[14] Chevreux, G.; Tilly, N.; Bihoreau, N.: Fast analysis of recombinant monoclonal antibodies using IdeS proteolytic digestion and electrospray mass spectrometry. *Analytical Biochemistry* **415,** 212-214 (2011)

[15] Fornelli, L.; Ayoub, D.; Aizikov, K.; Beck, A.; Tsybin, Y. O.: Middle-Down Analysis of Monoclonal Antibodies with Electron Transfer Dissociation

Orbitrap Fourier Transform Mass Spectrometry. *Analytical Chemistry* **86,** 3005-3012 (2014)

[16] Cheung, W. C.; Beausoleil, S. A.; Zhang, X., *et al.*: A proteomics approach for the identification and cloning of monoclonal antibodies from serum. *Nat Biotech* **30,** 447-452 (2012)

[17] Rosati, S.; Rose, R. J.; Thompson, N. J.; van Duijn, E.; Damoc, E.; Denisov, E.; Makarov, A.; Heck, A. J. R.: Exploring an Orbitrap Analyzer for the Characterization of Intact Antibodies by Native Mass Spectrometry. *Angewandte Chemie International Edition* **51,** 12992-12996 (2012)

[18] Fornelli, L.; Damoc, E.; Thomas, P. M.; Kelleher, N. L.; Aizikov, K.; Denisov, E.; Makarov, A.; Tsybin, Y. O.: Analysis of Intact Monoclonal Antibody IgG1 by Electron Transfer Dissociation Orbitrap FTMS. *Molecular & Cellular Proteomics* **11,** 1758-1767 (2012)

[19] Senko, M. W.; Canterbury, J. D.; Guan, S.; Marshall, A. G.: A High-performance Modular Data System for Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Rapid Communications in Mass Spectrometry* **10,** 1839-1844 (1996)

[20] Zamdborg, L.; LeDuc, R. D.; Glowacz, K. J., *et al.*: ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Research* **35,** W701-W706 (2007)

[21] Mao, Y.; Valeja, S. G.; Rouse, J. C.; Hendrickson, C. L.; Marshall, A. G.: Top-Down Structural Analysis of an Intact Monoclonal Antibody by Electron Capture Dissociation-Fourier Transform Ion Cyclotron Resonance-Mass Spectrometry. *Analytical Chemistry* **85,** 4239-4246 (2013)

Figure captions.

Figure 1. Extracted total ion chromatogram of separated 50 kDa Fab subunits of three monoclonal IgG1s from a mixture digested with papain. Numbered panels show broadband FTMS mass spectra of the charge state envelopes of 1) trastuzumab, 2) adalimumab and 3) palivizumab. The red rectangle shows the isolation windows (200 Th and wider) used for subsequent ETD fragmentation, centered around 34-36+, 33-35+ and 35-37+ charge states for trastuzumab, adalimumab and palivizumab, respectively.

Figure 2. Fragmentation map of Fab subunit of adalimumab. Top panel: ETD product ions (*c*- and *z*- type) assigned using ProSight PC, excluding those ions linked by intermolecular disulfide bond. Bottom panel: linked ions (*z*+*z*) assigned using an in-house developed tool. CDRs of each chain are highlighted in orange. Obtained sequence coverage is indicated below each chain in each panel.

Figure 3. Expanded view of ETD mass spectrum (m/z 1500-1640 range) of Fab subunit of trastuzumab. Indicated are *c*- ions and disulfide bond linked *z*+*z* ions. Insets show expanded views of isotopically resolved product ions: linked ion interchain disulfide bond z_{104}-z_{118} (13+) (left) and the c_{103} (7+) classical ion (right).

Figure 4. Cartoon representation of linkage of Fab subunits by intermolecular disulfide bond for IgG1s (left) and IgG2s, IgG3s and IgG4s (right). Both inter- and intramolecular disulfide bonds are shown in red.

Figure 5. Fragmentation map of Fab subunit of natalizumab. Top panel: ETD product ions (*c*- and *z*- type) assigned using ProSight PC, excluding those ions linked by intermolecular disulfide bond. Bottom panel: linked ions (*z*+*z*) assigned using an in-house developed tool. CDRs of each chain are highlighted in orange. Obtained sequence coverage is indicated below each chain in each panel.

Figure 6. Broadband FTMS mass spectrum (top) and ETD tandem mass spectrum (bottom) of co-eluting trastuzumab and rituximab Fab subunits. Respective charge state envelopes are color coded. The red rectangle shows the isolation window of 230 Th used for ETD fragmentation. The isolation window contains five to six charge states of each Fab subunit.

Figure 1.

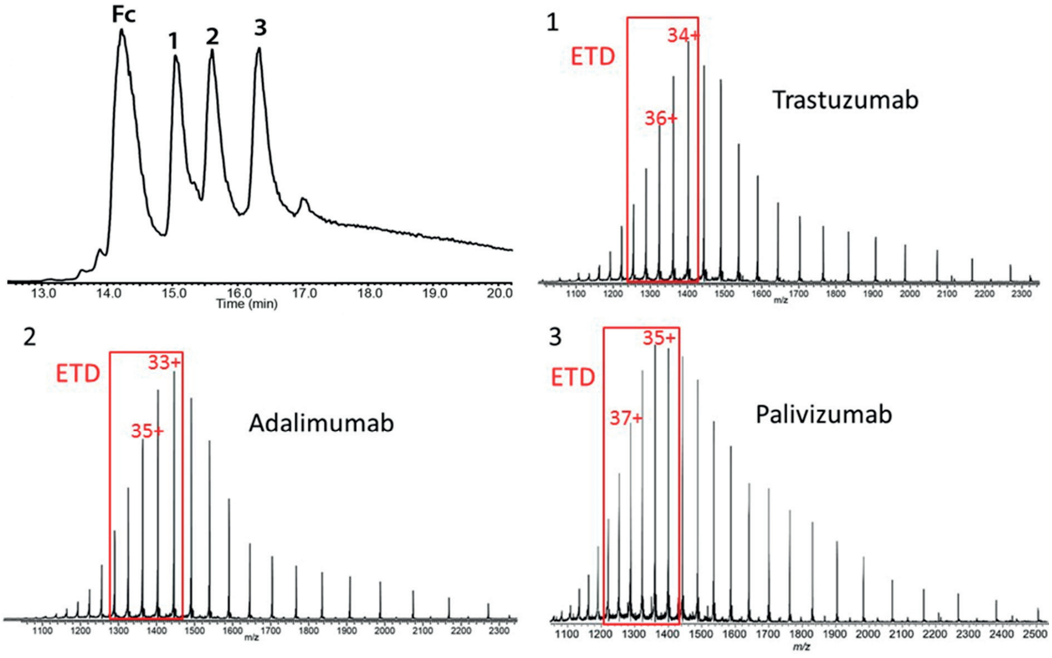

Fc
1
2
3
13.0
14.0
15.0
16.0
17.0
18.0
19.0
20.0
Time (min)
1
ETD
34+
36+
Trastuzumab
1100
1200
1300
1400
1500
1600
m/z
1700
1800
1900
2000
2100
2200
2300
2
ETD
33+
35+
Adalimumab
1100
1200
1300
1400
1500
1600
m/z
1700
1800
1900
2000
2100
2200
2300
3
ETD
35+
37+
Palivizumab
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
2200
2300
2400
2500
m/z

Figure 2.

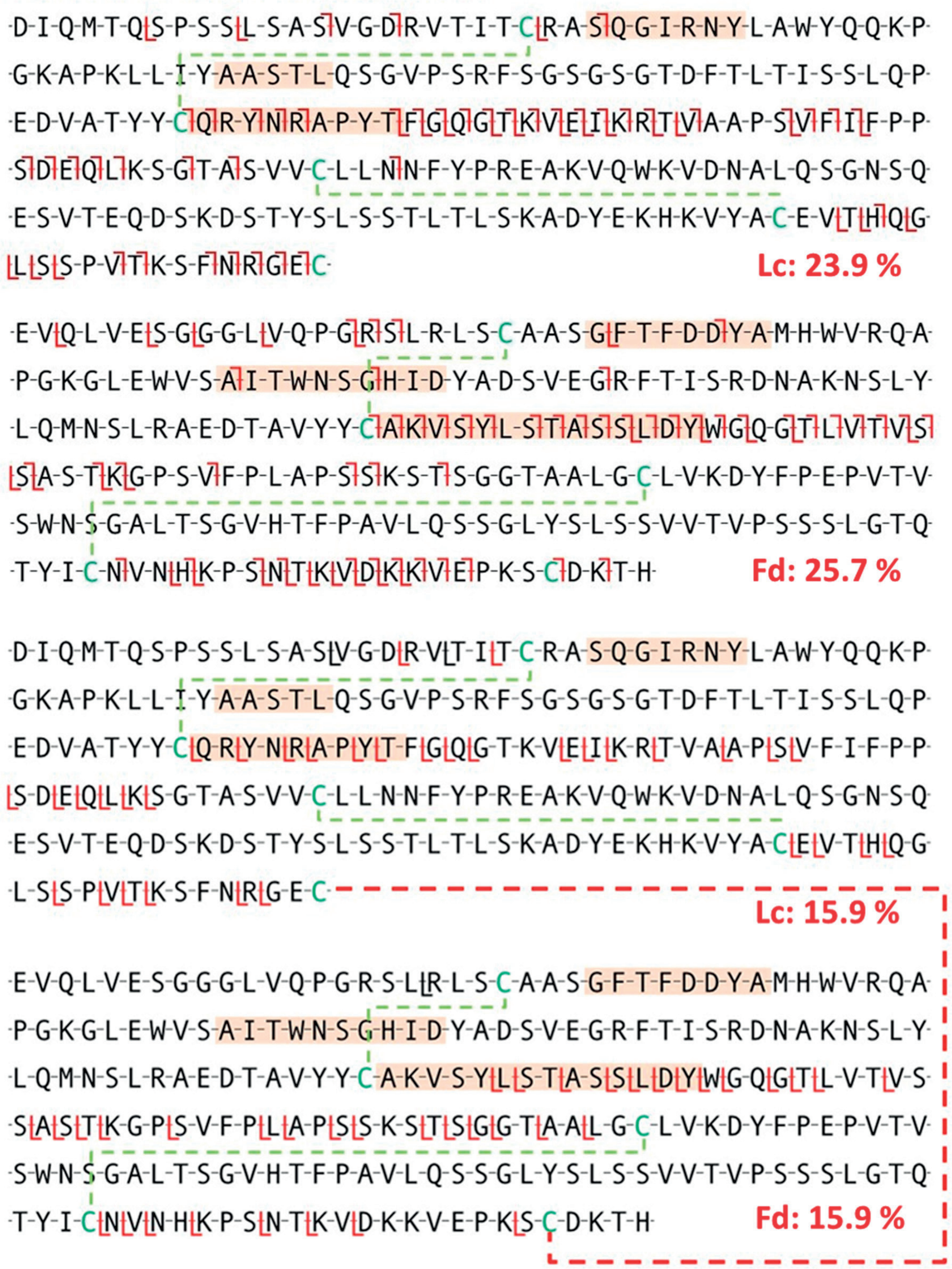

D-I-Q-M-T-Q-S-P-S-S-L-S-A-S-V-G-D-R-V-T-I-T-C-R-A-S-Q-G-I-R-N-Y-L-A-W-Y-Q-Q-K-P
G-K-A-P-K-L-L-I-Y-A-A-S-T-L-Q-S-G-V-P-S-R-F-S-G-S-G-S-G-T-D-F-T-L-T-I-S-S-L-Q-P
E-D-V-A-T-Y-Y-C-Q-R-Y-N-R-A-P-Y-T-F-G-Q-G-T-K-V-E-I-K-R-T-V-A-A-P-S-V-F-I-F-P-P
S-D-E-Q-L-K-S-G-T-A-S-V-V-C-L-L-N-N-F-Y-P-R-E-A-K-V-Q-W-K-V-D-N-A-L-Q-S-G-N-S-Q
E-S-V-T-E-Q-D-S-K-D-S-T-Y-S-L-S-S-T-L-T-L-S-K-A-D-Y-E-K-H-K-V-Y-A-C-E-V-T-H-Q-G
L-S-S-P-V-T-K-S-F-N-R-G-E-C
Lc: 23.9 %
E-V-Q-L-V-E-S-G-G-G-L-V-Q-P-G-R-S-L-R-L-S-C-A-A-S-G-F-T-F-D-D-Y-A-M-H-W-V-R-Q-A
P-G-K-G-L-E-W-V-S-A-I-T-W-N-S-G-H-I-D-Y-A-D-S-V-E-G-R-F-T-I-S-R-D-N-A-K-N-S-L-Y
L-Q-M-N-S-L-R-A-E-D-T-A-V-Y-Y-C-A-K-V-S-Y-L-S-T-A-S-S-L-D-Y-W-G-Q-G-T-L-V-T-V-S
S-A-S-T-K-G-P-S-V-F-P-L-A-P-S-S-K-S-T-S-G-G-T-A-A-L-G-C-L-V-K-D-Y-F-P-E-P-V-T-V
S-W-N-S-G-A-L-T-S-G-V-H-T-F-P-A-V-L-Q-S-S-G-L-Y-S-L-S-S-V-V-T-V-P-S-S-S-L-G-T-Q
T-Y-I-C-N-V-N-H-K-P-S-N-T-K-V-D-K-K-V-E-P-K-S-C-D-K-T-H
Fd: 25.7 %
D-I-Q-M-T-Q-S-P-S-S-L-S-A-S-V-G-D-R-V-T-I-T-C-R-A-S-Q-G-I-R-N-Y-L-A-W-Y-Q-Q-K-P
G-K-A-P-K-L-L-I-Y-A-A-S-T-L-Q-S-G-V-P-S-R-F-S-G-S-G-S-G-T-D-F-T-L-T-I-S-S-L-Q-P
E-D-V-A-T-Y-Y-C-Q-R-Y-N-R-A-P-Y-T-F-G-Q-G-T-K-V-E-I-K-R-T-V-A-A-P-S-V-F-I-F-P-P
S-D-E-Q-L-K-S-G-T-A-S-V-V-C-L-L-N-N-F-Y-P-R-E-A-K-V-Q-W-K-V-D-N-A-L-Q-S-G-N-S-Q
E-S-V-T-E-Q-D-S-K-D-S-T-Y-S-L-S-S-T-L-T-L-S-K-A-D-Y-E-K-H-K-V-Y-A-C-E-V-T-H-Q-G
L-S-S-P-V-T-K-S-F-N-R-G-E-C
Lc: 15.9 %
E-V-Q-L-V-E-S-G-G-G-L-V-Q-P-G-R-S-L-R-L-S-C-A-A-S-G-F-T-F-D-D-Y-A-M-H-W-V-R-Q-A
P-G-K-G-L-E-W-V-S-A-I-T-W-N-S-G-H-I-D-Y-A-D-S-V-E-G-R-F-T-I-S-R-D-N-A-K-N-S-L-Y
L-Q-M-N-S-L-R-A-E-D-T-A-V-Y-Y-C-A-K-V-S-Y-L-S-T-A-S-S-L-D-Y-W-G-Q-G-T-L-V-T-V-S
S-A-S-T-K-G-P-S-V-F-P-L-A-P-S-S-K-S-T-S-G-G-T-A-A-L-G-C-L-V-K-D-Y-F-P-E-P-V-T-V
S-W-N-S-G-A-L-T-S-G-V-H-T-F-P-A-V-L-Q-S-S-G-L-Y-S-L-S-S-V-V-T-V-P-S-S-S-L-G-T-Q
T-Y-I-C-N-V-N-H-K-P-S-N-T-K-V-D-K-K-V-E-P-K-S-C-D-K-T-H
Fd: 15.9 %

Figure 3.

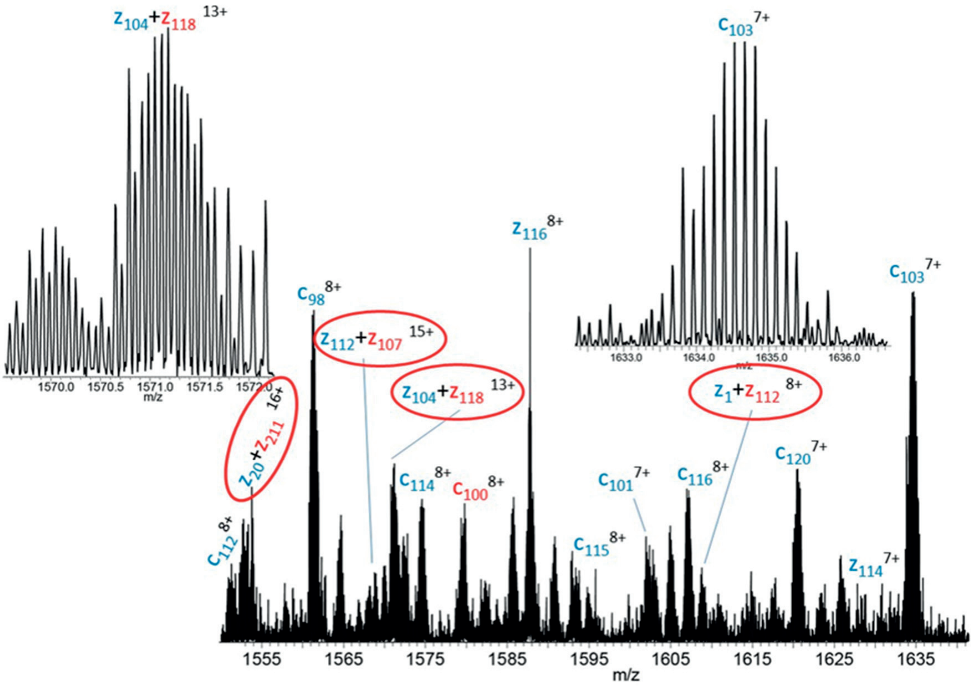
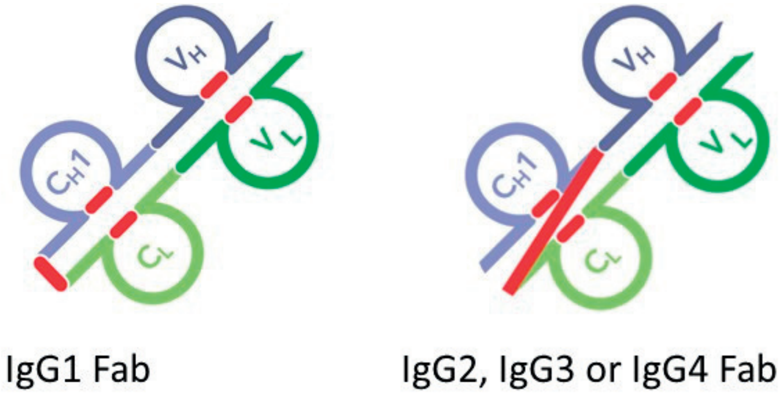


Figure 4.



IgG1 Fab

IgG2, IgG3 or IgG4 Fab

Figure 5.

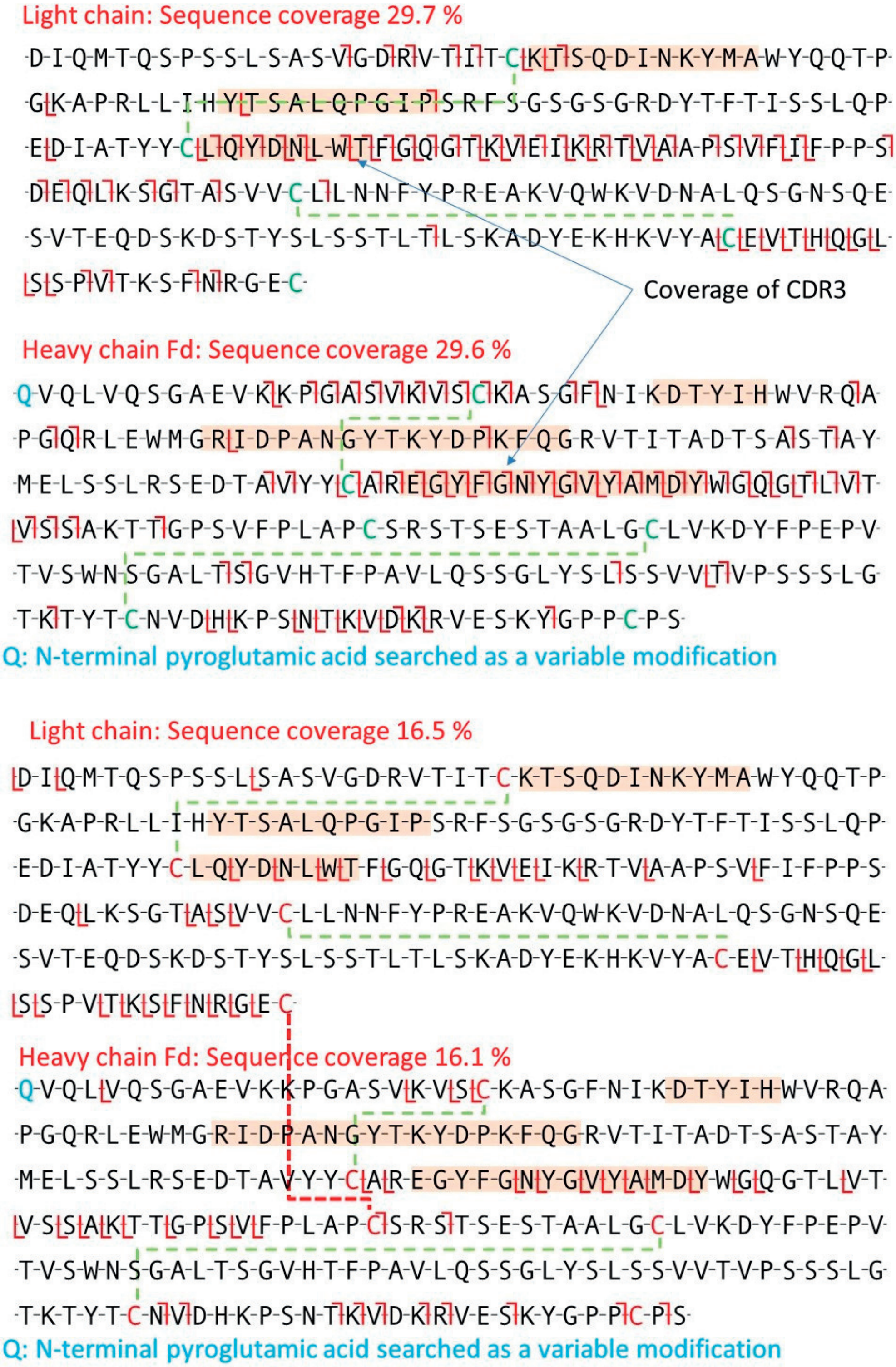
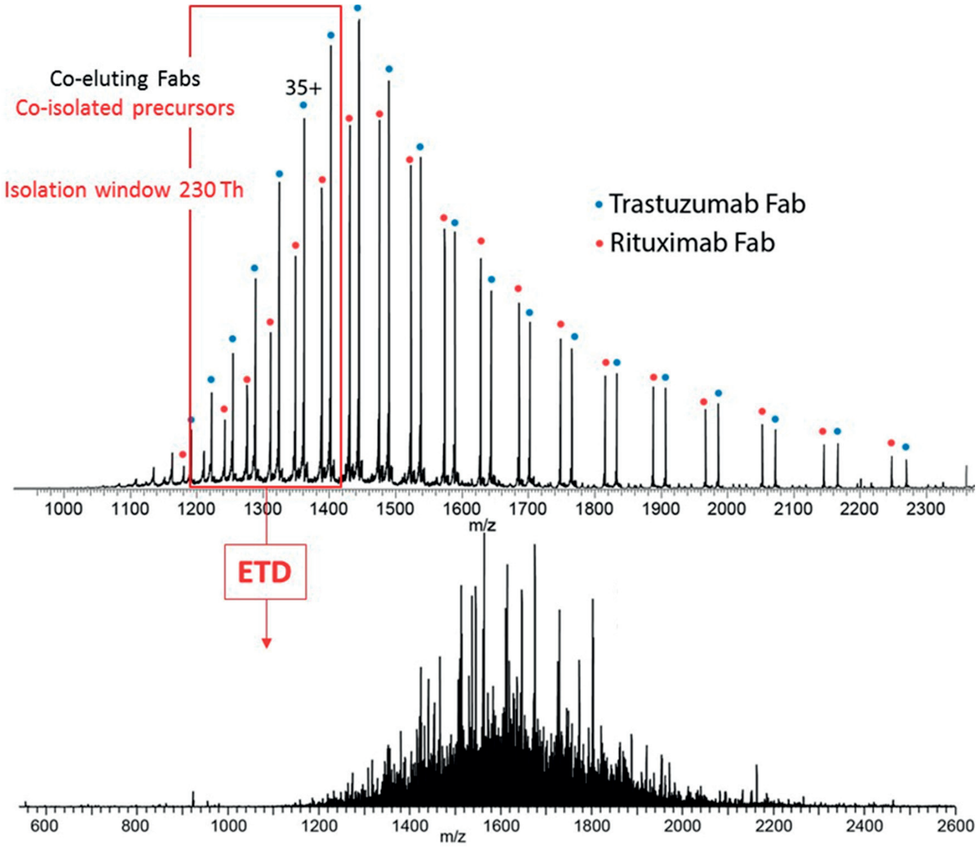

Light chain: Sequence coverage 29.7 %
D-I-Q-M-T-Q-S-P-S-S-L-S-A-S-V-G-D-R-V-T-I-T-C-K-T-S-Q-D-I-N-K-Y-M-A-W-Y-Q-Q-T-P
G-K-A-P-R-L-L-I-H-Y-T-S-A-L-Q-P-G-I-P-S-R-F-S-G-S-G-S-G-R-D-Y-T-F-T-I-S-S-L-Q-P
E-D-I-A-T-Y-Y-C-L-Q-Y-D-N-L-W-T-F-G-Q-G-T-K-V-E-I-K-R-T-V-A-A-P-S-V-F-I-F-P-P-S
D-E-Q-L-K-S-G-T-A-S-V-V-C-L-L-N-N-F-Y-P-R-E-A-K-V-Q-W-K-V-D-N-A-L-Q-S-G-N-S-Q-E
S-V-T-E-Q-D-S-K-D-S-T-Y-S-L-S-S-T-L-T-L-S-K-A-D-Y-E-K-H-K-V-Y-A-C-E-V-T-H-Q-G-L
S-S-P-V-T-K-S-F-N-R-G-E-C
Coverage of CDR3
Heavy chain Fd: Sequence coverage 29.6 %
Q-V-Q-L-V-Q-S-G-A-E-V-K-K-P-G-A-S-V-K-V-S-C-K-A-S-G-F-N-I-K-D-T-Y-I-H-W-V-R-Q-A
P-G-Q-R-L-E-W-M-G-R-I-D-P-A-N-G-Y-T-K-Y-D-P-K-F-Q-G-R-V-T-I-T-A-D-T-S-A-S-T-A-Y
M-E-L-S-S-L-R-S-E-D-T-A-V-Y-Y-C-A-R-E-G-Y-F-G-N-Y-G-V-Y-A-M-D-Y-W-G-Q-G-T-L-V-T
V-S-S-A-K-T-T-G-P-S-V-F-P-L-A-P-C-S-R-S-T-S-E-S-T-A-A-L-G-C-L-V-K-D-Y-F-P-E-P-V
T-V-S-W-N-S-G-A-L-T-S-G-V-H-T-F-P-A-V-L-Q-S-S-G-L-Y-S-L-S-S-V-V-T-V-P-S-S-S-L-G
T-K-T-Y-T-C-N-V-D-H-K-P-S-N-T-K-V-D-K-R-V-E-S-K-Y-G-P-P-C-P-S
Q: N-terminal pyroglutamic acid searched as a variable modification
Light chain: Sequence coverage 16.5 %
D-I-Q-M-T-Q-S-P-S-S-L-S-A-S-V-G-D-R-V-T-I-T-C-K-T-S-Q-D-I-N-K-Y-M-A-W-Y-Q-Q-T-P
G-K-A-P-R-L-L-I-H-Y-T-S-A-L-Q-P-G-I-P-S-R-F-S-G-S-G-S-G-R-D-Y-T-F-T-I-S-S-L-Q-P
E-D-I-A-T-Y-Y-C-L-Q-Y-D-N-L-W-T-F-G-Q-G-T-K-V-E-I-K-R-T-V-A-A-P-S-V-F-I-F-P-P-S
D-E-Q-L-K-S-G-T-A-S-V-V-C-L-L-N-N-F-Y-P-R-E-A-K-V-Q-W-K-V-D-N-A-L-Q-S-G-N-S-Q-E
S-V-T-E-Q-D-S-K-D-S-T-Y-S-L-S-S-T-L-T-L-S-K-A-D-Y-E-K-H-K-V-Y-A-C-E-V-T-H-Q-G-L
S-S-P-V-T-K-S-F-N-R-G-E-C
Heavy chain Fd: Sequence coverage 16.1 %
Q-V-Q-L-V-Q-S-G-A-E-V-K-K-P-G-A-S-V-K-V-S-C-K-A-S-G-F-N-I-K-D-T-Y-I-H-W-V-R-Q-A
P-G-Q-R-L-E-W-M-G-R-I-D-P-A-N-G-Y-T-K-Y-D-P-K-F-Q-G-R-V-T-I-T-A-D-T-S-A-S-T-A-Y
M-E-L-S-S-L-R-S-E-D-T-A-V-Y-Y-C-A-R-E-G-Y-F-G-N-Y-G-V-Y-A-M-D-Y-W-G-Q-G-T-L-V-T
V-S-S-A-K-T-T-G-P-S-V-F-P-L-A-P-C-S-R-S-T-S-E-S-T-A-A-L-G-C-L-V-K-D-Y-F-P-E-P-V
T-V-S-W-N-S-G-A-L-T-S-G-V-H-T-F-P-A-V-L-Q-S-S-G-L-Y-S-L-S-S-V-V-T-V-P-S-S-S-L-G
T-K-T-Y-T-C-N-V-D-H-K-P-S-N-T-K-V-D-K-R-V-E-S-K-Y-G-P-P-C-P-S
Q: N-terminal pyroglutamic acid searched as a variable modification

Figure 6.


Co-eluting Fabs
Co-isolated precursors
Isolation window 230 Th
35+
• Trastuzumab Fab
• Rituximab Fab
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
2200
2300
m/z
ETD
600
800
1000
1200
1400
1600
1800
2000
2200
2400
2600
m/z

7.2. Paper VII: Revealing chain connectivity in monoclonal IgG1 using GingisKHAN proteolysis and top-down electron transfer dissociation Orbitrap FTMS

Revealing chain connectivity in monoclonal IgG1 using GingisKHAN proteolysis and top-down electron transfer dissociation Orbitrap FTMS

Kristina Srzentić[1], Konstantin O. Nagornov[2], Anna A. Lobas[3,4], Daniel Ayoub[1], Luca Fornelli[1], Mikhail V. Gorskhov[3,4], Konstantin Ayzikov[5] and Yury O. Tsybin[2]*

[1] Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

[2] Spectroswiss Inc., EPFL Innovation Park, 1015 Lausanne, Switzerland

[3] Institute for Energy Problems of Chemical Physics, Russian Academy of Sciences, Leninskii Prospect 38, Bldg. 2,119334 Moscow, Russia

[4] Moscow Institute of Physics and Technology (State University), 9 Institutskiy per., 141707 Dolgoprudny, Moscow Region, Russia

[5] Thermo Scientific, Bremen, Germany

* Correspondence should be addressed to Dr. Yury O. Tsybin, Spectroswiss Inc., EPFL Innovation Park, Building I, 1015 Lausanne, Switzerland. E-mail: tsybin@spectroswiss.ch

Running title: Chain connectivity in IgG1 *via* ETD Orbitrap FTMS

Abstract

Pairing light and heavy chains in monoclonal antibodies using top-down mass spectrometry may complement chain sequence information provided by high-throughput genomics sequencing for rational selection of drug candidates. The 50 kDa F(ab) subunit of monoclonal antibodies is the smallest structural unit that contains the required information on pairing and can be enzymatically produced with high specificity. Here we develop and characterize the associated top-down workflow comprising the following steps: i) enzymatic digestion with single cleavage site specificity using GingisKHAN protease for production of F(ab) subunits; ii) multiple liquid chromatography (LC) – Orbitrap Fourier transform mass spectrometry (FTMS) runs with intact F(ab) fragmentation in the gas phase using electron transfer dissociation tandem mass spectrometry (MS/MS); iii) spectral averaging of tandem mass spectra across multiple LC-MS/MS runs acquired in reduced or full profile mode; iv) transient averaging across multiple LC-MS/MS runs followed by enhanced FT or absorption mode FT signal processing; and v) comprehensive automated and manual data analysis using ProSight Light and MASH Suite top-down software. We first benchmark the described workflow using myoglobin as a target protein and then apply thus validated method for the analysis of F(ab) subunit of trastuzumab. Obtained results confirm the envisioned benefits in terms of increased sensitivity from averaging of multiple LC-MS/MS runs for top-down protein analysis for both spectral and transient averaging, both of which are now accessible for general users.

Keywords: top-down mass spectrometry; immunoglobulin G1, IgG1; chain pairing; Fourier transform mass spectrometry, FTMS; Orbitrap; transient averaging; spectral averaging; absorption mode FT

Introduction

Monoclonal antibodies in general and immunoglobulins G1 (IgG1) in particular constitute an important class of biotherapeutics with demonstrated success for treatment of life-threatening diseases, including [1-3]. For example, trastuzumab IgG1, commercially known as herceptin, is widely used to treat breast cancer [4]. Therefore, the development of improved analytical methods and techniques addressing in-depth structural analysis of IgGs is important not only for accompanying targeted IgG development and production, but also for the discovery of new IgG-based drugs. Regarding the latter, bottom-up mass spectrometry (MS), in combination with high-throughput genomic sequencing, has been used for the identification of organism-produced IgGs present as complex biomolecular mixtures in body tissues or fluids like blood. Genomics-derived data enables the creation of databases containing sequence information on separately light and heavy chains of IgGs present in these mixtures [5]. Mass spectrometry data thus needs to pair the two antibody chains. Thorough protein degradation resulting from frequent digestion in bottom-up proteomics hinges the chain-pairing information and therefore the list of potential IgG lead candidates produced with bottom-up proteomics is not specific and contains false positive suggestions.

To overcome the limitations of bottom-up proteomics in providing pairing information on IgGs, we previously reported on the use of top-down (TD) mass spectrometry applied to the characterization of ~50 kDa F(ab) subunits of monoclonal IgGs. To obtain F(ab) fragments, we employed papain for IgG cleavage in the hinge region, whereas to increase the sensitivity of the TD approach we performed transient averaging followed by enhanced Fourier transform (eFT) signal processing (a commercial solution to improve resolution in Orbitrap Fourier transform mass spectrometry (FTMS) returning absorption mode [6]) at the manufacturer's site (Thermo Scientific, Bremen, Germany). The reported data validated such method for the analysis of up to three different IgGs simultaneously present in a mixture. The limitations of the developed approach are both in sample preparation, as the relatively low specificity of papain can potentially result in multiple digestion sites, and at the MS data processing level, given the restricted access to eFT for the general users.

In regard to the latter aspect, some consideration can be made with specific reference to top-down mass spectrometry. In general, sensitivity is a major bottleneck of TD MS. While in survey mass spectra the signal of an electrosprayed protein is detected as a complex charge state envelope, reducing the signal-to-noise ratio (SNR) of each single charge state, in tandem MS (MS/MS) selected precursor ions are dissociated in the gas phase, with the incoming total ion charge being split between many – hundreds or even thousands – of product ion channels. Therefore, signal amplitude of product ions in each single tandem mass spectrum (or microscan, in Orbitrap FTMS) can be very low. As a result, to construct a sufficiently accurate and abundant isotopic envelope of a product ion for unambiguous product ion assignment, averaging of a large number of microscans is required. Effectively, according to the fundamentals of signal processing in FTMS [7], the most sensitive and accurate approach to analysis data from a set of microscans is to first perform averaging of the time-domain data (transients) and then Fourier transform the final averaged transient to yield frequency spectrum, which can be further calibrated into a mass spectrum. The expected increase in signal-to-noise ratio (SNR) for ions in this case would scale as $\sqrt{N}$, where N is the number of averaged microscans. However, the number of available microscans is limited by the total time allocated for selected precursor ion analysis. In the case of LC-MS experiments, this analysis time is determined by the elution time of a precursor protein from the LC column and by the complexity of the sample – dictating the need to perform MS/MS on different co-eluting precursors, as in the case of proteomics studies.

For targeted TD experiment a total of 10-40 microscans are acquired per precursor protein ion elution peak within a single LC-MS/MS run [8-10]. However, to reach the spectral SNR level required to identify low-abundant product ions in convoluted MS spectra, the total number of acquired microscans is to be significantly, preferably more than 10 fold, increased. This result can be achieved either using off-line protein fractionation and subsequent MS/MS experiment from direct infusion of purified proteins, or performing multiple consecutive on-line LC-MS/MS experiments. The latter approach has the benefits of a better separation of co-eluting proteins, improved ionization efficiency, and a wider range of applications, including analysis of simple protein mixtures. Similarly to the above described microscan-averaging approach, time-domain transients of MS/MS data from multiple LC-MS/MS runs can be averaged altogether off-line. The benefits of time-domain averaging from multiple LC-MS/MS runs for improved top-down mass spectrometry have already been demonstrated for the above mentioned F(ab) subunit as well as for the analysis of intact, 150 kDa, IgG1 proteins [8], and of their 25 kDa [10] fragments. However, the reported examples were based on proprietary signal processing by Thermo Scientific, not available for general public. As a result, the use of the described approach for TD MS and TD proteomics has been so far limited. An alternative route to increase SNR is to first Fourier transform each of the single microscans (i.e., transients), and then perform spectral averaging of the resulting mass spectra. Traditionally, the disadvantage of this approach, which is believed to be less sensitive, include the possibility to introduce artifacts for analysis of peaks with low SNR values, and the fact that it is normally impossible, when using proprietary software for visualizing mass spectra, to average together MS scans stored in separate files.

Here, we first describe a new workflow for improving top-down MS analysis developed in the attempt of taking advantage of data recorded in separate LC-MS/MS runs and of applying both time-domain and spectral averaging overcoming the above mentioned restrictions. Specifically, we propose two user-accessible approaches: i) a python-based software for spectral averaging capable of using spectra from distinct LC-MS/MS data files; and ii) transient averaging followed by in-house calculated absorption mode FT allowing to reproduce from the averaged transient the resolution provided by eFT signal processing on commercial instruments.

To benchmark this top-down platform, we applied it to the analysis of ~50 kDa F(ab) subunits obtained using the recently commercialized GingisKHAN protease, which is characterized by superior cleavage specificity than papaine, addressing also the other limitation of our original work. Our methodology shows final level of spectral SNR that allow the identification of low abundant product ions, including those that can lead to the confirmation of the cysteines involved in inter-molecular disulfide bridges. This information is fundamental to prove light and heavy chain pairing, and therefore the here presented methodology can be used as a template for future drug-discovery research studies requiring the identification of selected IgGs from complex antibody mixtures derived from natural sources.

Experimental methods

Chemicals. Water, acetonitrile (ACN), formic acid (FA) and trifluoroethanol (TFE) were obtained in LC-MS purity grade. Water and ACN were purchased from Fluka Analytical (Buchs, Switzerland). FA was obtained from Merck (Zug, Switzerland) and TFE from Acros Organics (Geel, Belgium).

Samples. Horse myoglobin was obtained from Sigma Aldrich, therapeutic monoclonal antibody of the IgG1 class, trastuzumab (Herceptin, Genentech) was obtained as the European Medicines Agency approved version and formulation, available commercially to the general public.

GingisKHAN digestion. GingisKHAN (Genovis, Lund, Sweden) digestion of IgG1 was performed in formulation buffer. Two units of GingisKHAN (Tris-HCl) were added to each µg of IgG and left to react for 1 hr at 37 °C in presence of 2 mM Cys solution. The reaction was quenched by acidifying the solution to 1 % TFA. For analysis, samples were diluted with 0.1 % FA in water to a final concentration of 1 µg/µl.

Liquid chromatography – mass spectrometry. The chromatographic separation of IgG proteolytic fragments was performed using an Ultimate 3000 LC ystem (Thermo Scientific, Amsterdam, The Netherlands) under UPLC conditions. A combination of reversed phase C4 guard-column (Acquity UPLC PrST C4 VanGuard pre-column, 2.1 x 5 mm, particle size 1.7 µm, pore size 300 Å, Waters, Baden-Dättwil, Switzerland) and C4 analytical column (Acquity UPLC PrST C4, 1 x 150 mm, particle size 1.7 µm, pore size 300 Å, Waters) was employed to ensure on-line IgG fragment desalting and separation. For each injection, 1 µg of digestion product was loaded on the column, heated at 60 °C. After initial loading at 5 % solution B (organic phase), a gradient of solution B from 10 to 45 % in 15 minutes was applied at a flow rate of 100 µl/min. Solution A consisted of 0.1 % of FA in water, whereas solution B was composed of 99,9 % ACN and 0.1 % FA. The LC column outlet was on-line coupled with the electrospray ionization (ESI) source of the mass spectrometer. MS experiments were performed on an ETD-enabled hybrid linear ion trap high-field Orbitrap FT mass spectrometer (LTQ Orbitrap Elite FTMS, Thermo Scientific, Bremen, Germany). Separate LC-MS experiments were dedicated to record broadband mass spectra and ETD tandem mass spectra. All mass spectra were acquired using ion detection in the Orbitrap FTMS, in the *m/z* range 400-2800 and 200-2000 for broadband and tandem mass spectra, respectively. All mass spectrometry aquisitions were performed in 'protein mode': N_2 gas pressure in the HCD cell was reduced to reach a pressure in the Orbitrap mass analyzer region that is approximately of 0.15E-10 torr higher than the „base pressure“ measured in the same region (obtained with the N_2 flux completely shut down) [10]. Additionally, ions were captured and temporarily stored in the HCD cell before ion transmission to the Orbitrap [11]. Broadband mass spectra were recorded with 15’000 resolution at 400 *m/z*, with a target value for the automatic gain control (AGC) of one million charges in either MS or MS/MS modes. For ETD experiments, AGC target value for fluoranthene radical anions was set to 7E5 charges, with anion maximum injection time of 100 ms. ETD ion-ion interaction time was set to 10 ms. Product ion detection in the Orbitrap mass analyzer was performed with 120’000 resolution at 400 *m/z*. All Orbitrap FTMS scans were recorded averaging 10 microscans. Isolation windows for ETD of IgG fragments included multiple charge states per precursor protein (isolation width: 200 Th).

Data acqusition and signal processing. All experimental data were acquired using standard built-in data acquistion system, DAQ, (Thermo Scientific, Xcalibur). The mass spectra (*.raw) were obtained in either full or reduced profile mode in the separate data sets *via* standard software interface (Thermo Scientific). In parallel to mass spectra acquisition for both data sets, the transients in MIDAS format (*.dat) were acquired using the advanced software interface (Thermo Scientific). Further, the given number of mass spectra were averaged within each single LC-MS/MS run *via* standard data analysis software (Xcalibur, Thermo Scientific). The averaged mass spectra from multiple separate LC-MS/MS runs were produced using MS File Reader (Thermo Scientific) and the pyFTMS data analysis framework (Spectroswiss, Lausanne, Switzerland). The averaging of the corresponding transients in the time-domain followed by the absorption mode Fourier transform (aFT) signal processing were performed using pyFTMS data analysis framework and Autophaser Professional, correspondingly (Spectroswiss) [David P. A. Kilgour, Konstantin O. Nagornov, Anton N. Kozhinov, Konstantin O. Zhurov, Yury O. Tsybin.: Producing absorption mode Fourier transform ion cyclotron resonance mass spectra with non-quadratic phase correction functions. *RCMS*, 1087-1093, 29/11 (2015)]. The averaged mass spectra were converted into mzXML format for further data analysis. Additionally, transients were averaged via proprietary protocols and processed using enhanced Fourier transform (eFT) to yield mass spectra in *.raw format directly at Thermo Scientific [8].

Data analysis. Peak picking and deconvolution was performed using MASH Suite sofftware with following parameters: for myoglobin SNR threshold was set to 5 for transient and spectral averaged runs acquired in full profile mode and 0.2 for spectral averaged runs of reduced profile mode. For IgGthe values were set to 5, 0.6 and 0.2 for transient averaged full profile mode, spectral averaged full profile mode and spectral averaged reduced profile mode, respectively. For both myoglobin and IgG matching against their custom databases (containing primary sequence of protein in question) the minimum score between theoretical and experimental peak cluster was set to 80 %.

In case of the branched ions (z_{HC} + z_{LC}) list of theoretical fragments (neutral mass) was calculated using in-house Python script based on pyteomics library [12]. The internal disulfide bonds as well as the interchain disulfide bond were considered preserved. All probable pairs of heavy and light chains were considered; fragmentation N-terminal to proline or inside the cysteine loops was excluded from the calculations. Maximum of one cleavage site in each chain was allowed.

Results and discussion

Due to the aforementioned reasons, we tailored a pipeline for structural analysis of IgGs presented in **Figure 1.** This procedure entails the use of a novel KGP protease (commercially available as GingisKHAN, Genovis, Sweden) selective towards IgG1 class, and more importantly, highly specific and reproducible (showing primary structure specificity to the motif XXK-TXX above the hinge region). Total ion chromatogram of baseline separated Fc and F(ab) of Trastuzumab, and their subsequent intact mass measurements upon deconvolution with Protein Deconvolution software (Thermo Scientific) returned a molecular weight of 48,xxx Da which is in accordance with ones calculated based on the primary sequence within 10 ppm window. After a rapid digestion step, we performed ten consecutive LC-MS/MS runs followed by data averaging step. The novelty of the proposed pipeline lies in the procedures employed for the averaging. Obtained time-domain data (transients) were processed with enhanced Fourier transform (eFT) within a single LC-MS/MS .RAW file, with resulting single-file spectra further subjected to spectral averaging using an in-house Python script. Alternatively, transients collected from all the LC-MS/MS runs were averaged and converted to mass spectra applying either absorption mode FT (aFT) or the commercial eFT signal processing, as shown in the bottom part of the Figure 1.

LC-MS/MS analysis of model protein and IgG1. The proposed pipeline was tested on intact ~17 kDa myoglobin. Firstly, the LC settings were tested to assure that the elution reproducibility is maintained with respect to the time window set for triggering MS/MS event. This was important to ensure that the variation in number of recorded transients (or related mass spectra in the case of spectral averaging procedure) for each individual LC run is minimized. **Figure 2a** shows broadband mass spectrum of intact myoglobin from a single LC run recorded in Orbitrap analyzer to verify transmission efficiency under particular instrument setup in terms of pressure (so-called 'protein mode') and temporal storage of ions in HCD trap (also known as 'extended trapping'). ETD was optimized by testing different ion-ion interaction time spanning from 3 ms to 15 ms. Based on the intensity of the remaining precursor ion we establish the optimal value for fragmentation efficiency to be 7 ms. For subsequent ETD MS/MS a 'narrow' isolation window of 80 Th was centered around 848 *m/z*, comprising two charge states (20+ and 21+) shown in the inset. At resolution of 120 000 (at *m/z* 400), the charge state envelope of myoglobin is baseline isotopically resolved as seen in a zoom-in of 21+ charge state (Figure 2a, inset). In case of F(ab) subunits of trastuzumab, a wider isolation window of 200 Th was centered at charge state 38+ and it comprised six isotopically unresolved charge states of Fab) subjected to MS/MS as depicted in Figure 2b. Optimal ion-ion interaction time was found to be 10 ms. In case of Fc peak eluting prior to F(ab) an ETD event was not triggered, as Fc portion of antibody does not carry information on chain connectivity, and as such was beyond the scope of this study. Usually Fc subunit can be removed using Protein A so it does not interfere with separation or co-elutes with another F(ab) in mixtures analysis, however here it was kept to monitor LC separation and performance, as well as the one of MS. Its broadband mass spectrum and an expanded view of its portion containing the distribution of the glycoforms in 33-35+ charge states is shown in Figure 2c.

Method validation. Acquired data for myoglobin was further utilized for testing of different averaging procedures on single and multiple LC runs. As previously described, in FTMS the improvement in SNR ratio upon averaging should increase with square root of number of time-domain transients (or microscans) that are averaged. For spectral averaging, similar dependence was observed. However, it is to be noted how in this procedure increment will follow named dependence will be contigent on provided mass accuracy of peaks from the preceding peak picking procedure as well as the accuracy of averaging itself. **Figure 3.** top panel portrays a comparison between spectral intensities for myoglobin tandem mass spectra after absorption mode FT of a transient data from a single and 10 LC-MS/MS runs. It is obvious how the SNR of product ions increases and experimental distributions of peak intensities approach the theoretical ones by increased number of averaged runs. To further investigate viability of novel averaging methods, we addressed another figure of merit, the protein sequence coverage obtained by matching ETD product ions. For that purpose, all acquired and averaged data was subjected to deconvolution and product ion assignment against custom data base (contained only myoglobin sequence, or later trastuzumab one, *vide infra*). Results of the searches were used to construct the fragmentation maps presented in **Figure 4.** This data set represents the initial proving ground for assessment of viability of the herein proposed workflow, with particular emphasis on data averaging procedure for product ion assignment of TD and MD MS data sets, applied to IgG1 subunits analysis (*vide infra*).

ETD product ions (*c*- and *z*- type) vary in distribution throughout all the maps, however their variety does not change drastically the number of bond cleavages they assign. We observe how passing from single to multiple LC run, and from spectral to transient averaging of the same dataset, there is a twofold increase in difference of obtainable sequence coverage (2 % in case of single runs, and 4 % in case of ten LC runs). Increasing sequence coverage trend seems to favor transient averaging.

The ~ 50kDa F(ab) subunit represents a challenge for ETD fragmentation as it is composed by two chains of about 25 kDa each, with a total of five disulfide bridges (4 intra-molecular and 1 inter-molecular). As a consequence, each chain seems to retain high order structure even in the gas phase, as demonstrated by previous works were the fragmentation was localized primarily on the disulfide-free loops of each chain (in the region comprised between the second and third Cys residue). At the same time, the total number of potential fragmentation channels is particularly high, even by counting only the canonical N- and C-terminal-containing product ions. As depicted in **Supplementary Figure S1**, the fragmentation maps obtained from mass spectra resulting from spectral or time-domain transient averaging are very similar when only single LC-MS/MS run is considered. The 9-10 % sequence coverage for each chain seem to indicate that the spectral SNR has to be increased by additional averaging. Furthermore, most of the matched ions in the maps are *c*-type ions. A possible explanation for this phenomenon is given by the fact that the inter-molecular disulfide bridge linking light and heavy chains is located at the C-terminus of each chain (more precisely, it involves the last residue of the light chain and the third lo last residue of the heavy chain fragment, or Fd).

To improve the sequence coverage, we tested both in-house developed methods, spectral averaging and transient averaging with aFT signal processing, on the mass spectra/time-domain transients collected within 10 LC-MS/MS experiments. **Figure 5** shows the results of spectral averaging when starting from reduced profile (Figure 5a) or full profile mode spectra (Figure 5b). In both cases the increase in sequence coverage is substantial, reaching almost 19 and 30 % for light and heavy chain, respectively, in the case of full profile mode. The signal apodization and noise cutoff method commercially implemented obviously reduces the possibility to observe in the average mass spectrum the very low abundant product ions as they are oftentimes cut in the first place in the single spectra used for the averaging procedure; this slightly affects the final sequence coverage which is about 10-15% lower than that obtained using full profile mode spectra. Importantly, the averaged spectra were also used for matching a list of potential branched ions, i.e., product ions which include a portion of the heavy and of the light chain, linked by the above mentioned inter-molecular disulfide bond. Remarkably, we calculated a total theoretical number of z_{HC}+z_{LC} ions (*c*-type ions are excluded for reason already explained) equal to about 7500. Although an unambiguous assignment based solely on MS^2 is not possible for all these ions, as about 40 % of them share the molecular formula and hence the exact mass with at least a second ion, our attempt, shown on the right panels of Figure 5, to match these ions with the light and heavy chain sequences demonstrated the fact that ETD is actually forming *z*-type ions, but they are not of the traditional type but rather have a branched nature; most of matched branched ions are located in those portions of the sequences of each chain that seem poorly covered when looking at the fragmentation maps including only c- and z-type ions, such as the C-terminus of both light and heavy chain. Notably, some very large product ions (>15 kDa) were also matched. It has to be considered, finally, that although this assignment strategy is more sophisticated that the traditional one that simply accounts for non-branched fragments, it still does not include the assignment of internal fragment. A recent paper has demonstrated that, at least in beam-style collisional dissociation, the number of internal fragments generated when fragmenting large protein such as carbonic anhydrase is dramatically high. A similar study for ETD or any similar radical-driven fragmentation technique, however, is currently missing.

As a concluding remark, it has to be mentioned that the mass spectrum obtained by transient averaging produced a slightly higher sequence coverage for the F(ab) subunit (**Supplementary Figure S2**). Considering the minimal difference between the two averaging methods for what concerns the canonical *c*- and *z*-type ions, we can conclude that the spectral method works well. However, if we assume that product ions formed by two cleavage on two different chains, the branched $z_{HC}+z_{LC}$ ions, are less likely to be produced by ETD and, hence, are generally characterized by a lower abundance than traditional *c*- and *z* ions, the transient averaging method exibits an advantage in increasing the final SNR of tandem mass spectra, as it allows to increase the sequence coverage from branched ions from the 8-9 % obtained by spectral averaging to about 12 %.

Conclusions

The developed workflow for top-down FT mass spectrometry has demonstrated the envisioned advantages for targeted protein analysis when data from consecutive LC-MS/MS experiments can be analyzed together. Increase in sensitivity, or SNR values, scales comparably for spectral and transient averaging as a function of a number of scans. Performance of the described top-down mass spectrometry was first evaluated on myoglobin analysis and then applied to analysis of monoclonal antibodies. For the latter, a novel enzyme, with a commercial name GingisKHAN, has been employed to produce 50 kDa F(ab) subunits of IgG1 with a single-cleavage site specificity. Top-down analysis of F(ab) subunits allowed rendering pairing of light and heavy chains in IgG. Data analysis confirmed that the most accurate and extensive protein sequence coverage is obtained with transient averaging, followed by spectral averaging of mass spectra acquired in the full profile mode and then by spectral averaging of mass spectra acquired in the reduced profile mode.

Both spectral and transient averaging capabilities are now readily available for the general users. Importantly, absorption mode FT signal processing has been successfully applied here for the first time on transients acquired from Orbitrap FTMS, allowing reaching the performance of commercial mass spectra produced via enhanced FT (eFT) signal processing. Further 2-4 fold increase in sensitivity is expected from the improvements in high-performance data acquisition electronics and on-line signal processing on high-throughput FPGA chips. The accumulated developments should enable transition of top-down mass spectrometry from targeted protein analysis from multiple LC-MS/MS runs to large-scale top-down proteomics from a single LC-MS/MS run.

Acknowledgements

We express our gratitude to Thermo Fisher Scientific Inc. for providing access under license to LTQ Orbitrap Elite FTMS transient signals and related functionality enabled by the developers' kit. We are grateful for financial support through the European Research Council (ERC Starting Grant 280271 to YOT).

References

1. Leavy, O., Therapeutic antibodies: past, present and future. Nat Rev Immunol, 2010. **10**(5): p. 297-297.

2. Kirkpatrick, P., J. Graham, and M. Muhsin, Cetuximab. Nat Rev Drug Discov, 2004. **3**(7): p. 549-550.

3. Smith, M.R., Rituximab (monoclonal anti-CD20 antibody): mechanisms of action and resistance. Oncogene, 0000. **22**(47): p. 7359-7368.

4. Breast cancer: Trastuzumab therapy for small, HER2-positive breast tumours. Nat Rev Clin Oncol, 2015. **12**(3): p. 126-126.

5. Wine, Y., et al., Molecular deconvolution of the monoclonal antibodies that comprise the polyclonal serum response. Proc Natl Acad Sci U S A, 2013. **110**(8): p. 2993-8.

6. Lange, O., et al., Enhanced Fourier transform for Orbitrap mass spectrometry. International Journal of Mass Spectrometry, 2014. **369**: p. 16-22.

7. Marshall, A.G., Theoretical signal-to-noise ratio and mass resolution in Fourier transform ion cyclotron resonance mass spectrometry. Analytical Chemistry, 1979. **51**(11): p. 1710-1714.

8. Fornelli, L., et al., Analysis of intact monoclonal antibody IgG1 by electron transfer dissociation Orbitrap FTMS. Mol Cell Proteomics, 2012. **11**(12): p. 1758-67.

9. Tsybin, Y.O., et al., Structural analysis of intact monoclonal antibodies by electron transfer dissociation mass spectrometry. Anal Chem, 2011. **83**(23): p. 8919-27.

10. Fornelli, L., et al., Middle-down analysis of monoclonal antibodies with electron transfer dissociation orbitrap fourier transform mass spectrometry. Anal Chem, 2014. **86**(6): p. 3005-12.

11. Rosati, S., et al., Exploring an orbitrap analyzer for the characterization of intact antibodies by native mass spectrometry. Angew Chem Int Ed Engl, 2012. **51**(52): p. 12992-6.

12. Goloborodko, A., et al., Pyteomics—a Python Framework for Exploratory Data Analysis and Rapid Software Prototyping in Proteomics. Journal of The American Society for Mass Spectrometry, 2013. **24**(2): p. 301-304.

13. Fellers, R.T., et al., ProSight Lite: graphical software to analyze top-down mass spectrometry data. Proteomics, 2015. **15**(7): p. 1235-8.

14. Cai, W., et al., MASH Suite Pro: A Comprehensive Software Tool for Top-down Proteomics. Molecular & Cellular Proteomics, 2015.

Figure captions.

Figure 1. Schematics of the proposed middle-down workflow for structural analysis of IgGs. Sample preparation entails rapid (one hour) digestion of IgG with a novel IgG1 selective and …K-T… sequence motif specific KP protease (REF Genovis poster) followed by multiple consecutive LC-MS/MS runs. Obtained time-domain data (transients) are: either processed with enhanced Fourier transform (eFT), with resulting mass spectra that are first averaged within a single .RAW file and from there subjected to spectral averaging through the pyFTMS platform; or alternatively collected and averaged prior to in-house absorption mode FT signal processing. The final mass spectrum, obtained either via spectral or transient averaging, is analyzed with ProSight Lite [13] and MASH suite [14] to yield protein sequence coverage maps.

Figure 2. Intact protein and protein subunits mass measurements with - Orbitrap Elite FTMS and precursor ion isolation for subsequent ETD MS/MS. a) Broadband mass spectrum of myoglobin. Inset shows a mass spectrum of its isolated precursor ions (charge states 21+ and 20+, an isolation window of 80 Th centered at *m/z* 848) and baseline-resolved isotopic envelope of charge state 21+. b) Broadband mass spectrum of F(ab) subunit of monoclonal IgG1 trastuzumab. Inset shows a mass spectrum of isolated precursor ions around charge state 38+ obtained using an isolation window of 200 *m/z* centered at *m/z* 1248; and c) Broadband mass spectrum of Fc subunit of monoclonal IgG1 trastuzumab. Inset shows an expanded view of a broadband mass spectrum containing the distribution of the glycoforms in 33-35+ charge states.

Figure 3. Tandem mass spectra (normalized to the base peak) of myoglobin (top panel) and F(ab) subunit of trastuzumab (middle panel) after absorption mode FT of an averaged transient obtained from a single LC-MS/MS run (black lines) or from 10 LC-MS/MS runs (red lines). Insets show representative examples of product ion assignment. The SNR of product ions increases and experimental distributions of peak intensities approach the theoretical ones (shown with green circles) passing from single to multiple LC-MS/MS runs. (Bottom panel) shows an expanded view of a tandem mass spectrum of F(ab) containing assigned *c*- and branched z_{LC}+z_{HC}- product ions after absorption mode FT of an averaged transient from 10 LC-MS/MS runs. The equidistant polymeric peaks visible in mass spectrum originate from sample preparation and do not belong to IgG.

Figure 4. Sequence coverage of myoglobin obtained using LC-MS/MS workflow presented in Figure 1. ETD at 7 ms ion-ion interaction time was employed for MS/MS. Data was acquired with an Orbitrap Elite and recorded in full profile mode. Fragmentation maps shown are obtained from a) a single LC-MS/MS run after conventional spectral averaging (Qual Browser), b) a single LC-MS/MS run after in-house transient averaging, c) ten consecutive LC-MS/MS runs after in-house spectral averaging, and d) ten consecutive LC-MS/MS runs after in-house transient averaging. ETD product ions (*c*- and *z*-type) are color coded. Obtained sequence coverage is indicated below each chain in each panel. For single LC-MS/MS runs xxxx scans were averaged, whereas for ten consecutive LC-MS/MS runs xxx scans were averaged.

Figure 5. Application of workflow described in Figure 1 to IgG analysis using Orbitrap Elite FTMS. Shown are ETD MS/MS fragmentation maps of F(ab) subunit of trastuzumab produced by GingisKHAN digestion,obtained after in-house spectral averaging of a) 10 LC-MS/MS runs recorded in reduced profile mode and b) 10 LC-MS/MS runs recorded in full profile mode. Fragmentation maps are shown for both when the intramolecular disulfide bridge between the light and heavy chains of IgG is preserved (left panel) or cleaved (right panel). ETD product ions (*c*- and *z*- type as well as branched z_{LC}+z_{HC} ions) are represented according to the color coding. Cysteines forming intramolecular disulfide bond are highlighted in orange, and those participating in intermolecular bond in red. CDRs are highlighted in grey. Resulting sequence coverage is indicated below each chain in each panel.

Figure 1.

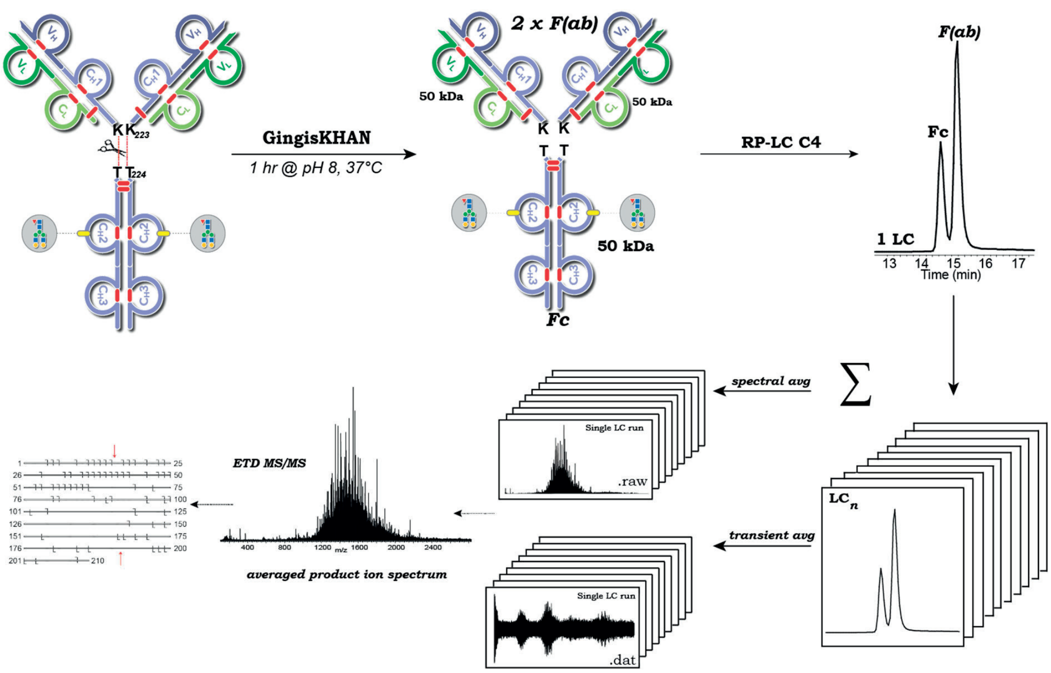

VH
VL
CH1
CL
K K223
T T224
CH2
CH3
GingisKHAN
1 hr @ pH 8, 37°C
2 x F(ab)
50 kDa
50 kDa
K K
T T
50 kDa
Fc
RP-LC C4
F(ab)
Fc
1 LC
13 14 15 16 17
Time (min)
spectral avg
∑
Single LC run
.raw
transient avg
Single LC run
.dat
LCn
ETD MS/MS
400 800 1200 1600 2000 2400
m/z
averaged product ion spectrum
1 25
26 50
51 75
76 100
101 125
126 150
151 175
176 200
201 210

Figure 2.

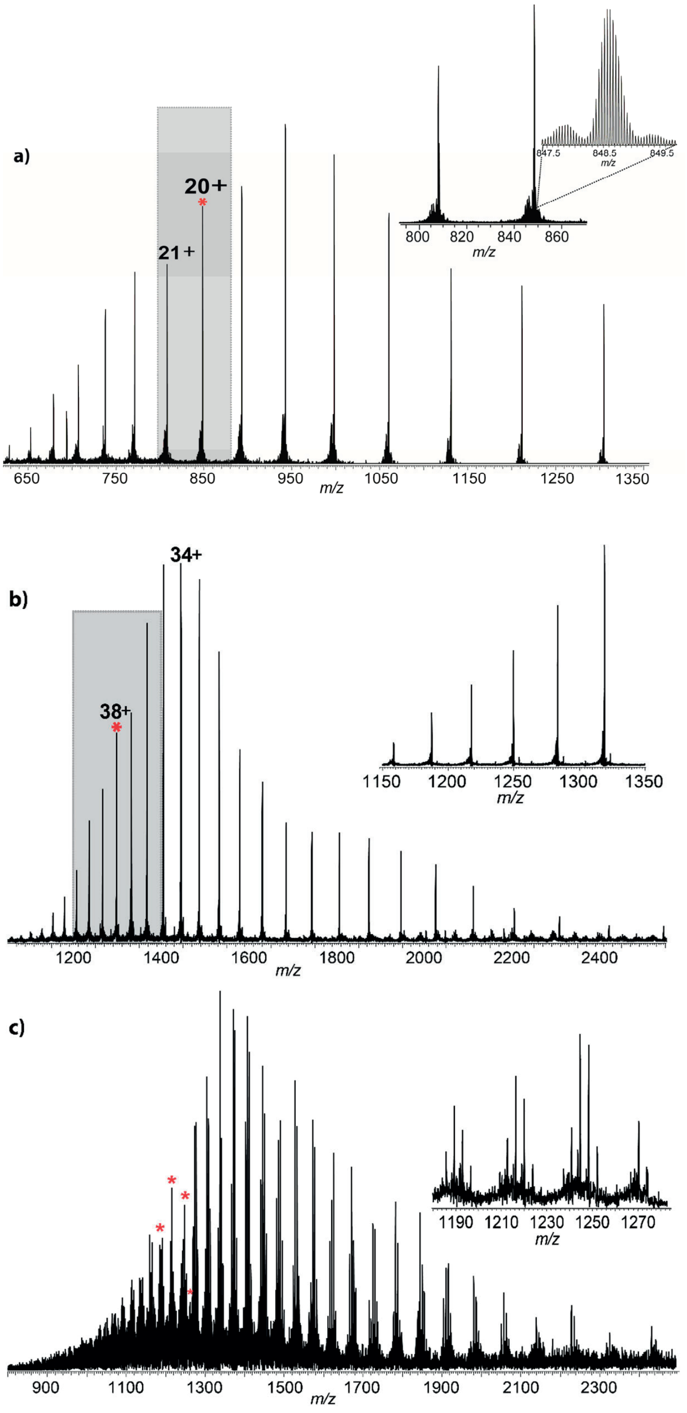

a)
20+
21+
847.5
848.5
849.5
m/z
800
820
840
860
m/z
650
750
850
950
1050
1150
1250
1350
m/z
b)
34+
38+
1150
1200
1250
1300
1350
m/z
1200
1400
1600
1800
2000
2200
2400
m/z
c)
1190
1210
1230
1250
1270
m/z
900
1100
1300
1500
1700
1900
2100
2300
m/z

Figure 3.

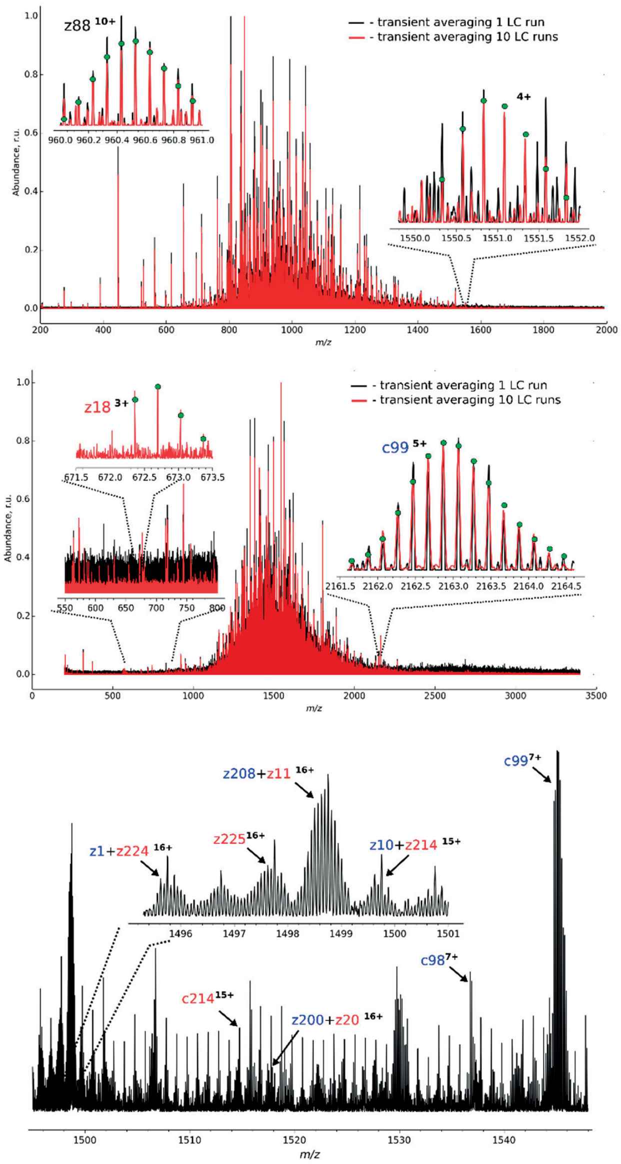

z88 10+
— - transient averaging 1 LC run
— - transient averaging 10 LC runs
960.0 960.2 960.4 960.6 960.8 961.0
4+
1550.0 1550.5 1551.0 1551.5 1552.0
Abundance, r.u.
1.0
0.8
0.6
0.4
0.2
0.0
200 400 600 800 1000 1200 1400 1600 1800 2000
m/z
z18 3+
671.5 672.0 672.5 673.0 673.5
c99 5+
2161.5 2162.0 2162.5 2163.0 2163.5 2164.0 2164.5
550 600 650 700 750 800
0 500 1000 1500 2000 2500 3000 3500
z208+z11 16+
c99 7+
z1+z224 16+
z225 16+
z10+z214 15+
1496 1497 1498 1499 1500 1501
c98 7+
c214 15+
z200+z20 16+
1500 1510 1520 1530 1540

Figure 4.

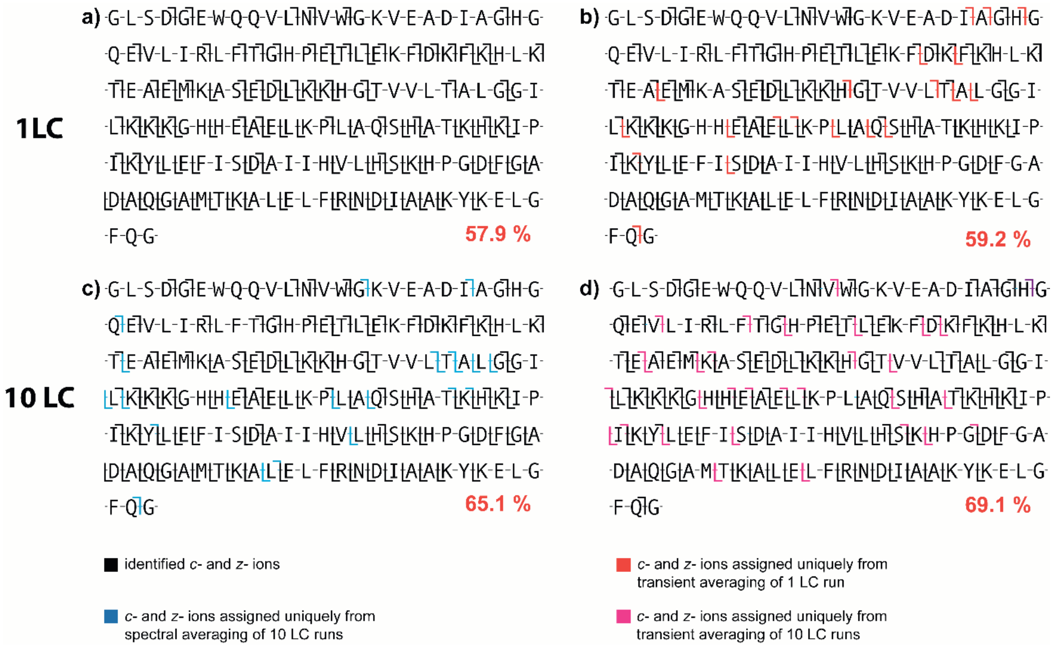
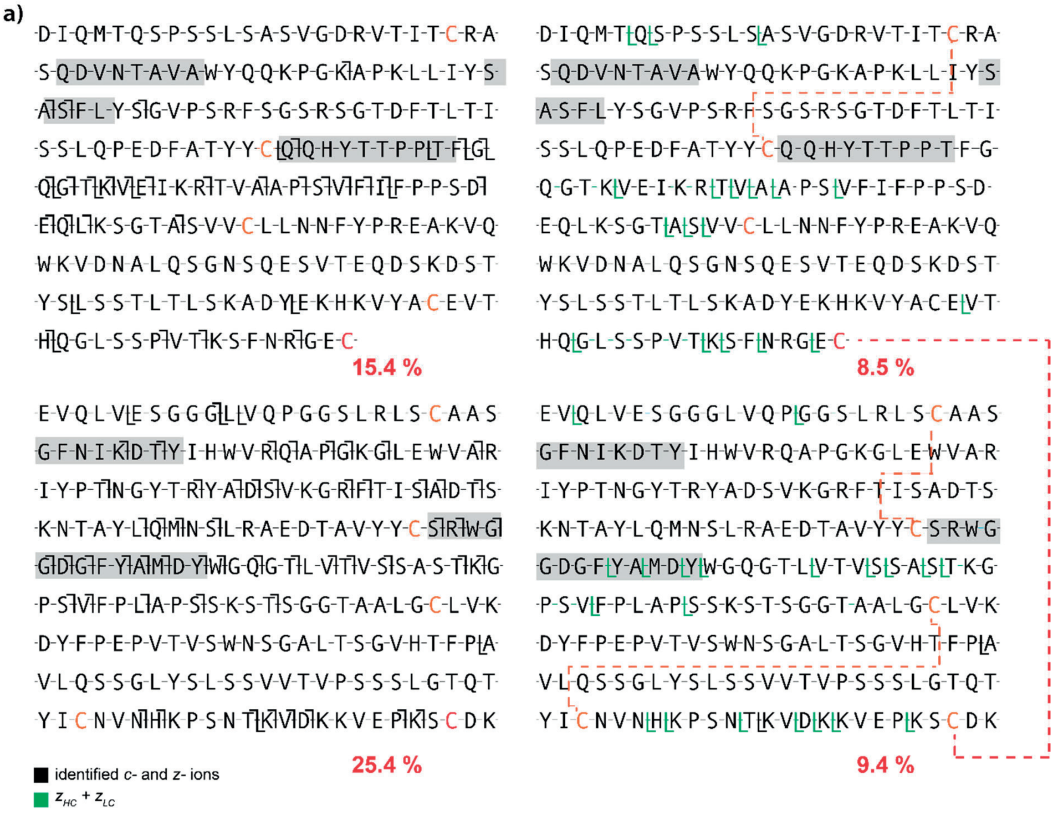
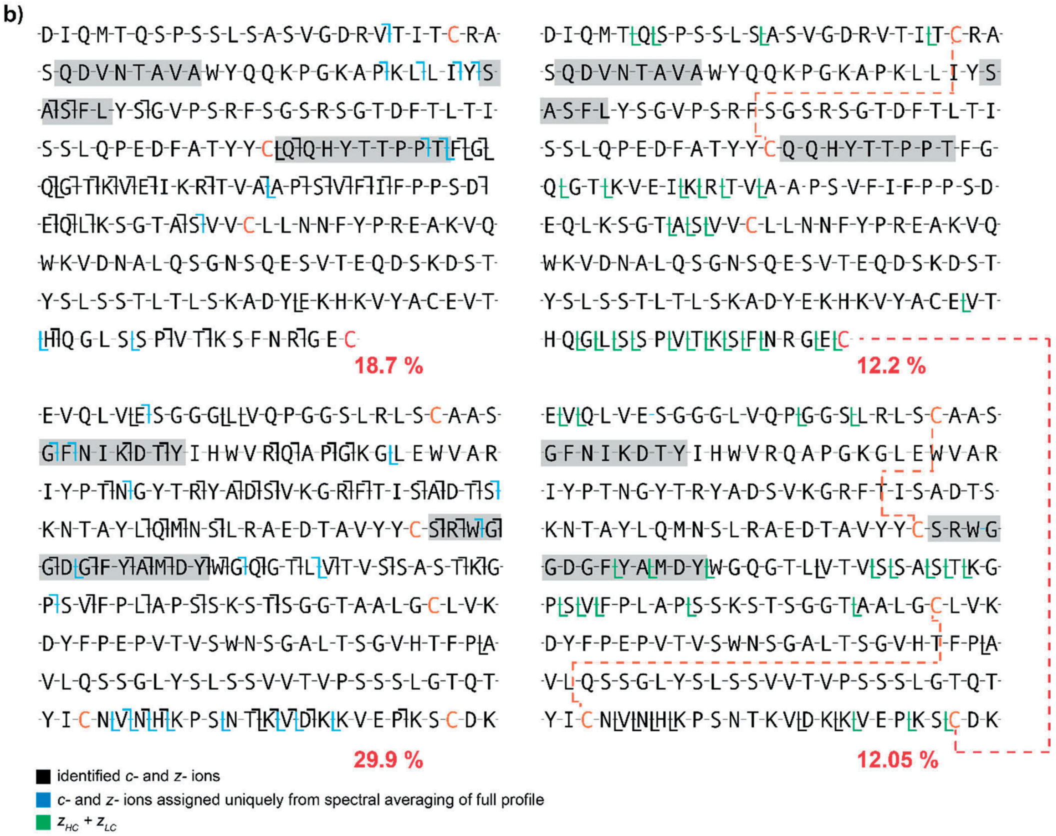

a)
b)
c)
d)
1 LC
10 LC
G-L-S-D-G-E-W-Q-Q-V-L-N-V-W-G-K-V-E-A-D-I-A-G-H-G
Q-E-V-L-I-R-L-F-T-G-H-P-E-T-L-E-K-F-D-K-F-K-H-L-K
T-E-A-E-M-K-A-S-E-D-L-K-K-H-G-T-V-V-L-T-A-L-G-G-I
L-K-K-K-G-H-H-E-A-E-L-K-P-L-A-Q-S-H-A-T-K-H-K-I-P
I-K-Y-L-E-F-I-S-D-A-I-I-H-V-L-H-S-K-H-P-G-D-F-G-A
D-A-Q-G-A-M-T-K-A-L-E-L-F-R-N-D-I-A-A-K-Y-K-E-L-G
F-Q-G
57.9 %
59.2 %
65.1 %
69.1 %
■ identified *c*- and *z*- ions
■ *c*- and *z*- ions assigned uniquely from spectral averaging of 10 LC runs
■ *c*- and *z*- ions assigned uniquely from transient averaging of 1 LC run
■ *c*- and *z*- ions assigned uniquely from transient averaging of 10 LC runs

Figure 5.


a)
D-I-Q-M-T-Q-S-P-S-S-L-S-A-S-V-G-D-R-V-T-I-T-C-R-A
S-Q-D-V-N-T-A-V-A-W-Y-Q-Q-K-P-G-K-A-P-K-L-L-I-Y-S
A-S-F-L-Y-S-G-V-P-S-R-F-S-G-S-R-S-G-T-D-F-T-L-T-I
S-S-L-Q-P-E-D-F-A-T-Y-Y-C-Q-Q-H-Y-T-T-P-P-T-F-G
Q-G-T-K-V-E-I-K-R-T-V-A-A-P-S-V-F-I-F-P-P-S-D
E-Q-L-K-S-G-T-A-S-V-V-C-L-L-N-N-F-Y-P-R-E-A-K-V-Q
W-K-V-D-N-A-L-Q-S-G-N-S-Q-E-S-V-T-E-Q-D-S-K-D-S-T
Y-S-L-S-S-T-L-T-L-S-K-A-D-Y-E-K-H-K-V-Y-A-C-E-V-T
H-Q-G-L-S-S-P-V-T-K-S-F-N-R-G-E-C
15.4 %
D-I-Q-M-T-Q-S-P-S-S-L-S-A-S-V-G-D-R-V-T-I-T-C-R-A
S-Q-D-V-N-T-A-V-A-W-Y-Q-Q-K-P-G-K-A-P-K-L-L-I-Y-S
A-S-F-L-Y-S-G-V-P-S-R-F-S-G-S-R-S-G-T-D-F-T-L-T-I
S-S-L-Q-P-E-D-F-A-T-Y-Y-C-Q-Q-H-Y-T-T-P-P-T-F-G
Q-G-T-K-V-E-I-K-R-T-V-A-A-P-S-V-F-I-F-P-P-S-D
E-Q-L-K-S-G-T-A-S-V-V-C-L-L-N-N-F-Y-P-R-E-A-K-V-Q
W-K-V-D-N-A-L-Q-S-G-N-S-Q-E-S-V-T-E-Q-D-S-K-D-S-T
Y-S-L-S-S-T-L-T-L-S-K-A-D-Y-E-K-H-K-V-Y-A-C-E-V-T
H-Q-G-L-S-S-P-V-T-K-S-F-N-R-G-E-C
8.5 %
E-V-Q-L-V-E-S-G-G-G-L-V-Q-P-G-G-S-L-R-L-S-C-A-A-S
G-F-N-I-K-D-T-Y-I-H-W-V-R-Q-A-P-G-K-G-L-E-W-V-A-R
I-Y-P-T-N-G-Y-T-R-Y-A-D-S-V-K-G-R-F-T-I-S-A-D-T-S
K-N-T-A-Y-L-Q-M-N-S-L-R-A-E-D-T-A-V-Y-Y-C-S-R-W-G
G-D-G-F-Y-A-M-D-Y-W-G-Q-G-T-L-V-T-V-S-S-A-S-T-K-G
P-S-V-F-P-L-A-P-S-S-K-S-T-S-G-G-T-A-A-L-G-C-L-V-K
D-Y-F-P-E-P-V-T-V-S-W-N-S-G-A-L-T-S-G-V-H-T-F-P-A
V-L-Q-S-S-G-L-Y-S-L-S-S-V-V-T-V-P-S-S-S-L-G-T-Q-T
Y-I-C-N-V-N-H-K-P-S-N-T-K-V-D-K-K-V-E-P-K-S-C-D-K
25.4 %
E-V-Q-L-V-E-S-G-G-G-L-V-Q-P-G-G-S-L-R-L-S-C-A-A-S
G-F-N-I-K-D-T-Y-I-H-W-V-R-Q-A-P-G-K-G-L-E-W-V-A-R
I-Y-P-T-N-G-Y-T-R-Y-A-D-S-V-K-G-R-F-T-I-S-A-D-T-S
K-N-T-A-Y-L-Q-M-N-S-L-R-A-E-D-T-A-V-Y-Y-C-S-R-W-G
G-D-G-F-Y-A-M-D-Y-W-G-Q-G-T-L-V-T-V-S-S-A-S-T-K-G
P-S-V-F-P-L-A-P-S-S-K-S-T-S-G-G-T-A-A-L-G-C-L-V-K
D-Y-F-P-E-P-V-T-V-S-W-N-S-G-A-L-T-S-G-V-H-T-F-P-A
V-L-Q-S-S-G-L-Y-S-L-S-S-V-V-T-V-P-S-S-S-L-G-T-Q-T
Y-I-C-N-V-N-H-K-P-S-N-T-K-V-D-K-K-V-E-P-K-S-C-D-K
9.4 %
■ identified c- and z- ions
🟩 z_HC + z_LC


b)
D-I-Q-M-T-Q-S-P-S-S-L-S-A-S-V-G-D-R-V-T-I-T-C-R-A
S-Q-D-V-N-T-A-V-A-W-Y-Q-Q-K-P-G-K-A-P-K-L-L-I-Y-S
A-S-F-L-Y-S-G-V-P-S-R-F-S-G-S-R-S-G-T-D-F-T-L-T-I
S-S-L-Q-P-E-D-F-A-T-Y-Y-C-Q-Q-H-Y-T-T-P-P-T-F-G
Q-G-T-K-V-E-I-K-R-T-V-A-A-P-S-V-F-I-F-P-P-S-D
E-Q-L-K-S-G-T-A-S-V-V-C-L-L-N-N-F-Y-P-R-E-A-K-V-Q
W-K-V-D-N-A-L-Q-S-G-N-S-Q-E-S-V-T-E-Q-D-S-K-D-S-T
Y-S-L-S-S-T-L-T-L-S-K-A-D-Y-E-K-H-K-V-Y-A-C-E-V-T
H-Q-G-L-S-S-P-V-T-K-S-F-N-R-G-E-C
18.7 %
D-I-Q-M-T-Q-S-P-S-S-L-S-A-S-V-G-D-R-V-T-I-T-C-R-A
S-Q-D-V-N-T-A-V-A-W-Y-Q-Q-K-P-G-K-A-P-K-L-L-I-Y-S
A-S-F-L-Y-S-G-V-P-S-R-F-S-G-S-R-S-G-T-D-F-T-L-T-I
S-S-L-Q-P-E-D-F-A-T-Y-Y-C-Q-Q-H-Y-T-T-P-P-T-F-G
Q-G-T-K-V-E-I-K-R-T-V-A-A-P-S-V-F-I-F-P-P-S-D
E-Q-L-K-S-G-T-A-S-V-V-C-L-L-N-N-F-Y-P-R-E-A-K-V-Q
W-K-V-D-N-A-L-Q-S-G-N-S-Q-E-S-V-T-E-Q-D-S-K-D-S-T
Y-S-L-S-S-T-L-T-L-S-K-A-D-Y-E-K-H-K-V-Y-A-C-E-V-T
H-Q-G-L-S-S-P-V-T-K-S-F-N-R-G-E-C
12.2 %
E-V-Q-L-V-E-S-G-G-G-L-V-Q-P-G-G-S-L-R-L-S-C-A-A-S
G-F-N-I-K-D-T-Y-I-H-W-V-R-Q-A-P-G-K-G-L-E-W-V-A-R
I-Y-P-T-N-G-Y-T-R-Y-A-D-S-V-K-G-R-F-T-I-S-A-D-T-S
K-N-T-A-Y-L-Q-M-N-S-L-R-A-E-D-T-A-V-Y-Y-C-S-R-W-G
G-D-G-F-Y-A-M-D-Y-W-G-Q-G-T-L-V-T-V-S-S-A-S-T-K-G
P-S-V-F-P-L-A-P-S-S-K-S-T-S-G-G-T-A-A-L-G-C-L-V-K
D-Y-F-P-E-P-V-T-V-S-W-N-S-G-A-L-T-S-G-V-H-T-F-P-A
V-L-Q-S-S-G-L-Y-S-L-S-S-V-V-T-V-P-S-S-S-L-G-T-Q-T
Y-I-C-N-V-N-H-K-P-S-N-T-K-V-D-K-K-V-E-P-K-S-C-D-K
29.9 %
E-V-Q-L-V-E-S-G-G-G-L-V-Q-P-G-G-S-L-R-L-S-C-A-A-S
G-F-N-I-K-D-T-Y-I-H-W-V-R-Q-A-P-G-K-G-L-E-W-V-A-R
I-Y-P-T-N-G-Y-T-R-Y-A-D-S-V-K-G-R-F-T-I-S-A-D-T-S
K-N-T-A-Y-L-Q-M-N-S-L-R-A-E-D-T-A-V-Y-Y-C-S-R-W-G
G-D-G-F-Y-A-M-D-Y-W-G-Q-G-T-L-V-T-V-S-S-A-S-T-K-G
P-S-V-F-P-L-A-P-S-S-K-S-T-S-G-G-T-A-A-L-G-C-L-V-K
D-Y-F-P-E-P-V-T-V-S-W-N-S-G-A-L-T-S-G-V-H-T-F-P-A
V-L-Q-S-S-G-L-Y-S-L-S-S-V-V-T-V-P-S-S-S-L-G-T-Q-T
Y-I-C-N-V-N-H-K-P-S-N-T-K-V-D-K-K-V-E-P-K-S-C-D-K
12.05 %
■ identified c- and z- ions
🟦 c- and z- ions assigned uniquely from spectral averaging of full profile
🟩 z_HC + z_LC

Supporting Information.

Revealing chain connectivity in monoclonal IgG1 using GingisKHAN proteolysis and top-down electron transfer dissociation Orbitrap FTMS

Kristina Srzentić, Konstantin O. Nagornov, Anna A. Lobas, Daniel Ayoub, Luca Fornelli, Mikhail V. Gorskhov, Konstantin Ayzikov and Yury O. Tsybin

Figure S1. Fragmentation maps of F(ab) subunit of trastuzumab following GingisKHAN digestion and 15 ms ETD MS/MS from a single LC-MS/MS run obtained using (left panel) in-house spectral averaging and (right panel) in-house transient averaging. ETD product ions (*c*- and *z*- type) are indicated in black. Obtained sequence coverage is indicated below each chain in each panel. All LC-MS/MS runs were recorded in full profile mode.

Figure S2. Fragmentation maps of F(ab) subunit of trastuzumab following GingisKHAN digestion obtained from transient averaging (performed with assistance of Thermo Scientific) of 10 LC-MS/MS runs recorded in full profile mode. Panels show fragmentation maps obtained when intramolecular disulfide bridge is cleaved (left) or preserved (right). ETD product ions (*c*- and *z*- type as well as branched z_{LC}+z_{HC} ions) are represented according to the color coding. CDRs are highlighted in grey. Obtained sequence coverage is indicated below each chain in each panel.

Figure S3. Normalized to the base peak tandem mass spectra of F(ab) subunit of IgG1 trastuzumab (top panel) after transient averaging (black line), and after averaging of mass spectra acquired in the full profile mode (red line) and in the reduced profile mode (blue line) of 10 LC-MS/MS runs. Panels A, B and C show expanded views of the mass spectra with assigned *c*- and *z*- product ions of light chain (left column) and of heavy chain (right column) of trastuzumab. The colored star symbol indicates that the experimental mass and abundance of a peak of the corresponding mass spectrum was matched with theoretical values in a certain accuracy range and assigned. Unassigned isotopic peaks can reduce the confidence in product ion assignment. Data show reduction in peak assignment efficiency of spectral averaging in the reduced profile compared to both full profile mode and transient averaging. The transient averaging provides better identification of product ions isotopic peaks compared to spectral averaging in both the full and reduced profile modes.

Figure S4. Mass accuracy distributions (in parts-per-million, ppm) after spectral averaging (left) and transient averaging (right) of myoglobin (top) and F(ab) (bottom). Tandem mass spectra acquired with LTQ Orbitrap FTMS using ETD. Data averaging is performed over 10 LC-MS/MS runs.

Figure S1.

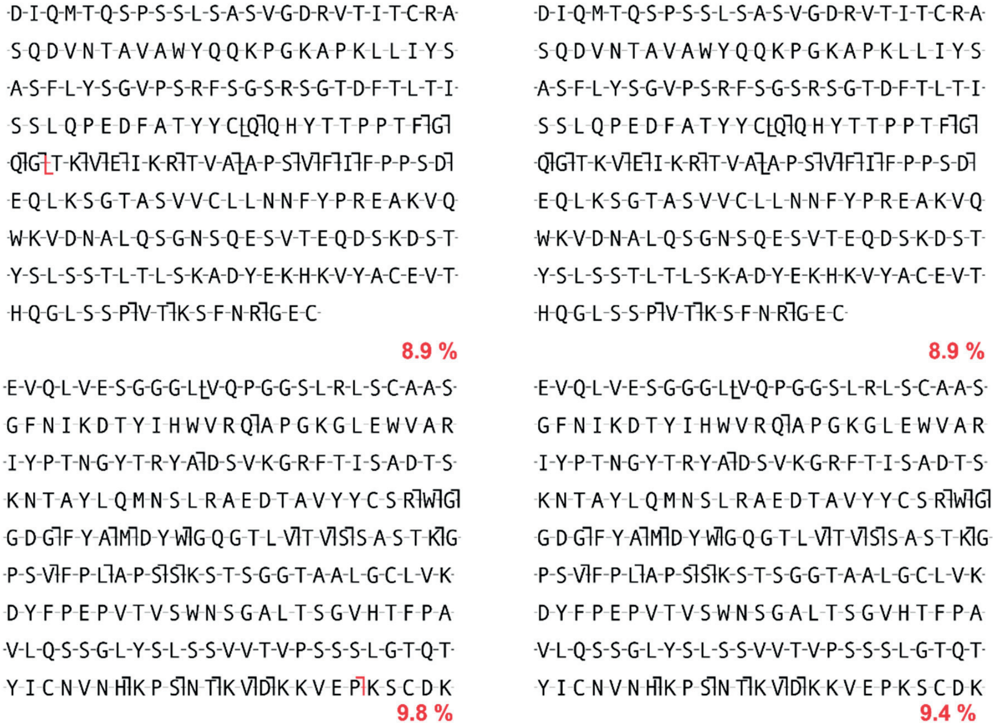

DIQMTQSPSSLSASVGDRVTITCRA
SQDVNTAVAWYQQKPGKAPKLLIYS
ASFLYSGVPSRFSGSRSGTDFTLTI
SSLQPEDFATYYCQQHYTTPPTFG
QGTKVEIKRTVAAPSVFIFPPSD
EQLKSGTASVVCLLNNFYPREAKVQ
WKVDNALQSGNSQESVTEQDSKDST
YSLSSTLTLSKADYEKHKVYACEVT
HQGLSSPVTKSFNRGEC
8.9 %
DIQMTQSPSSLSASVGDRVTITCRA
SQDVNTAVAWYQQKPGKAPKLLIYS
ASFLYSGVPSRFSGSRSGTDFTLTI
SSLQPEDFATYYCQQHYTTPPTFG
QGTKVEIKRTVAAPSVFIFPPSD
EQLKSGTASVVCLLNNFYPREAKVQ
WKVDNALQSGNSQESVTEQDSKDST
YSLSSTLTLSKADYEKHKVYACEVT
HQGLSSPVTKSFNRGEC
8.9 %
EVQLVESGGGLVQPGGSLRLSCAAS
GFNIKDTYIHWVRQAPGKGLEWVAR
IYPTNGYTRYADSVKGRFTISADTS
KNTAYLQMNSLRAEDTAVYYCSRWG
GDGFYAMDYWGQGTLVTVSSASTKG
PSVFPLAPSSKSTSGGTAALGCLVK
DYFPEPVTVSWNSGALTSGVHTFPA
VLQSSGLYSLSSVVTVPSSSLGTQT
YICNVNHKPSNTKVDKKVEPKSCDK
9.8 %
EVQLVESGGGLVQPGGSLRLSCAAS
GFNIKDTYIHWVRQAPGKGLEWVAR
IYPTNGYTRYADSVKGRFTISADTS
KNTAYLQMNSLRAEDTAVYYCSRWG
GDGFYAMDYWGQGTLVTVSSASTKG
PSVFPLAPSSKSTSGGTAALGCLVK
DYFPEPVTVSWNSGALTSGVHTFPA
VLQSSGLYSLSSVVTVPSSSLGTQT
YICNVNHKPSNTKVDKKVEPKSCDK
9.4 %

Figure S2.

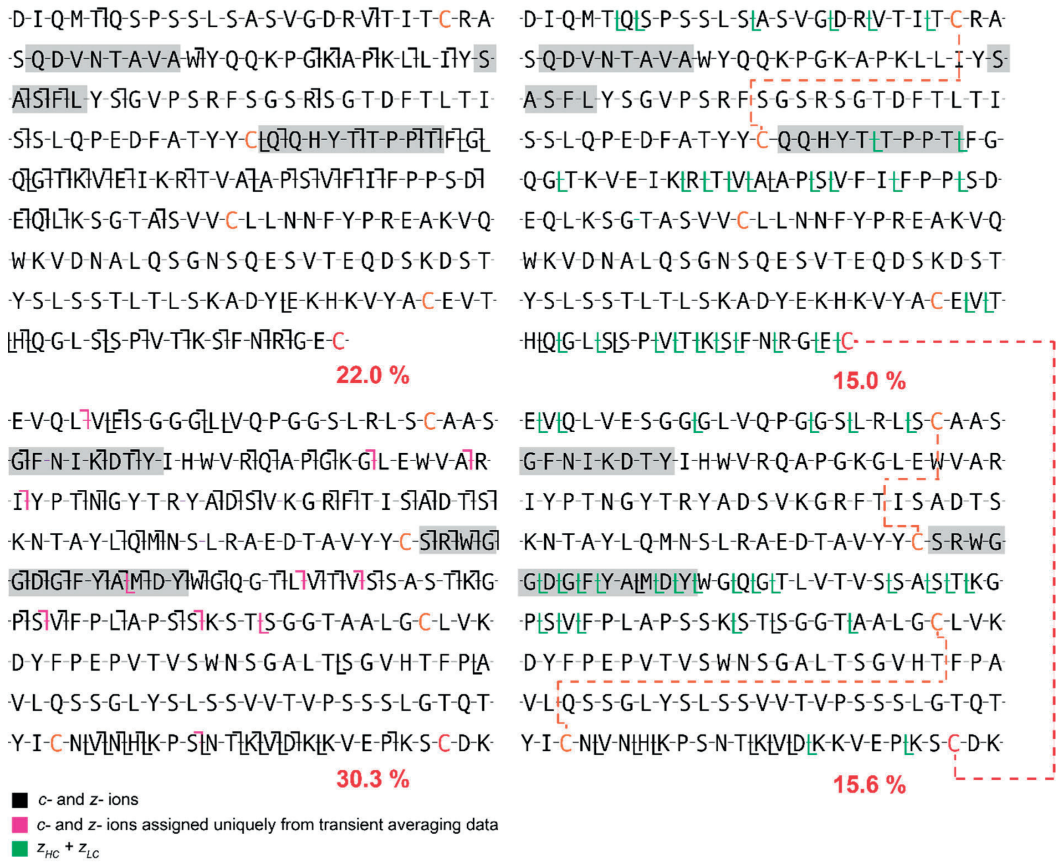

22.0 %
15.0 %
30.3 %
15.6 %
■ c- and z- ions
🟥 c- and z- ions assigned uniquely from transient averaging data
🟩 z_{HC} + z_{LC}

Figure S3.

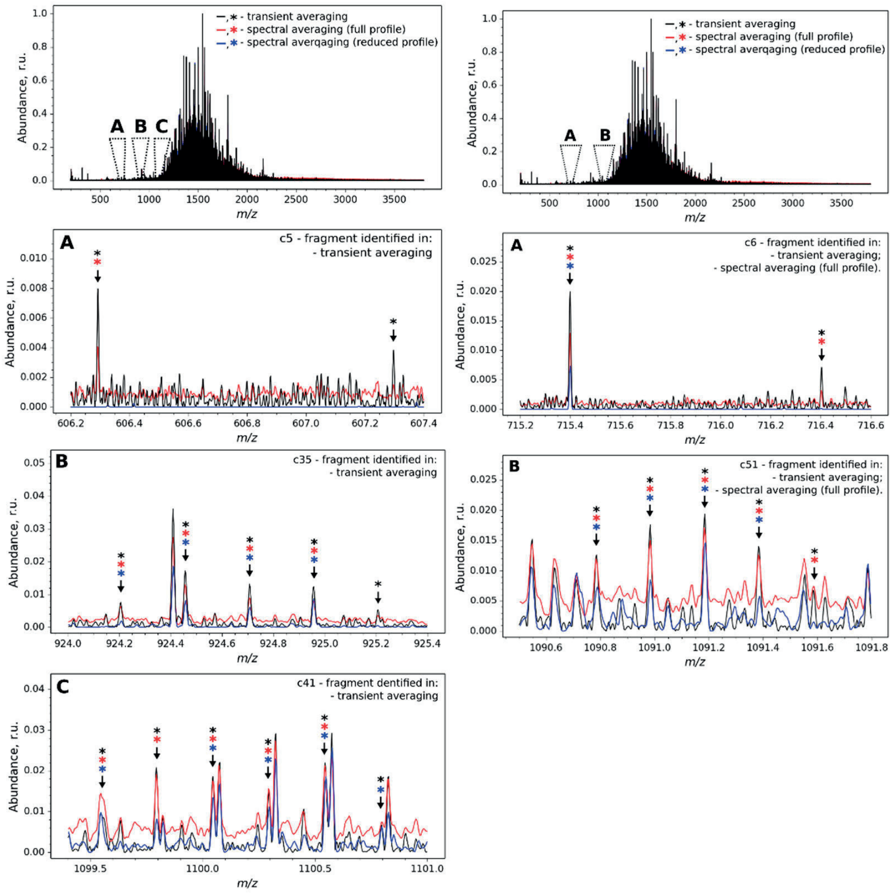
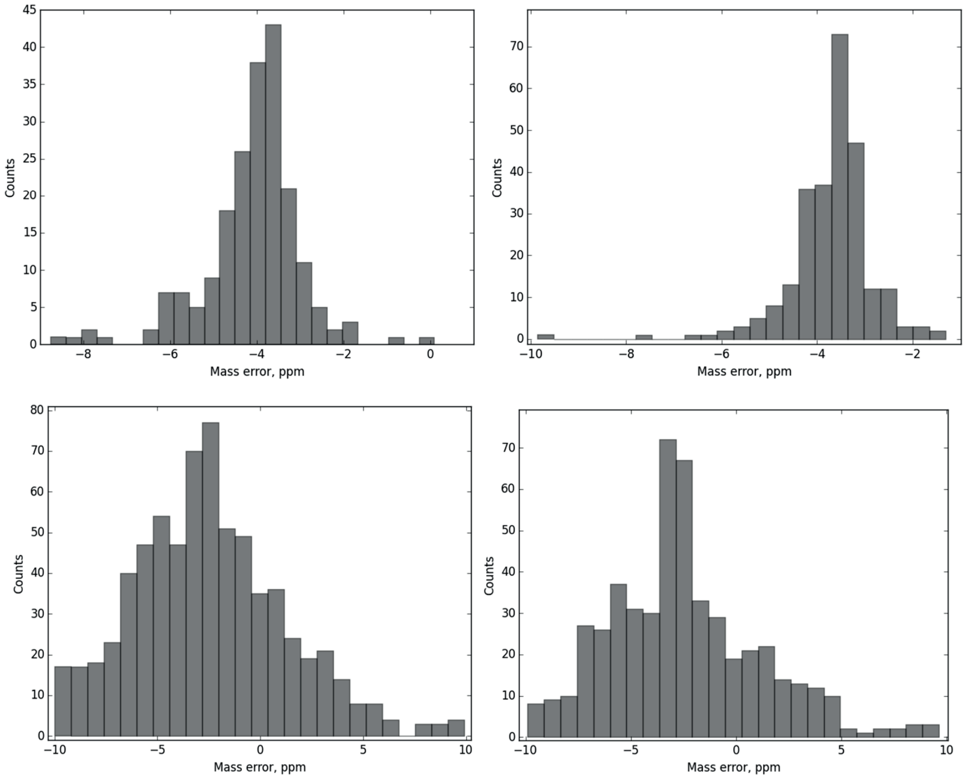

–,* - transient averaging
–,* - spectral averaging (full profile)
–,* - spectral averqaging (reduced profile)
Abundance, r.u.
m/z
A
B
C
500
1000
1500
2000
2500
3000
3500
0.0
0.2
0.4
0.6
0.8
1.0
c5 - fragment identified in:
- transient averaging
606.2
606.4
606.6
606.8
607.0
607.2
607.4
c6 - fragment identified in:
- transient averaging;
- spectral averaging (full profile).
715.2
715.4
715.6
715.8
716.0
716.2
716.4
716.6
c35 - fragment identified in:
- transient averaging
924.0
924.2
924.4
924.6
924.8
925.0
925.2
925.4
c51 - fragment identified in:
- transient averaging;
- spectral averaging (full profile).
1090.6
1090.8
1091.0
1091.2
1091.4
1091.6
1091.8
c41 - fragment dentified in:
- transient averaging
1099.5
1100.0
1100.5
1101.0

Figure S4.


Counts
Mass error, ppm
0
5
10
15
20
25
30
35
40
45
−8
−6
−4
−2
0
Counts
Mass error, ppm
0
10
20
30
40
50
60
70
−10
−8
−6
−4
−2
Counts
Mass error, ppm
0
10
20
30
40
50
60
70
80
−10
−5
0
5
10
Counts
Mass error, ppm
0
10
20
30
40
50
60
70
−10
−5
0
5
10

]

Chapter 8. Conclusions

8.1. Summary of the results.

The experimental studies presented in this Thesis are a comprehensive account of contemporary considerations including: i) validated or disproved 'educated guesses', ii) development of both practical and instrumental methods that aim at inception and further advancement of the *middle-down* (MD) mass spectrometry (MS)-based approach to proteomics for *in-depth* proteome analysis. In tailoring this novel domain, steered by mid-size range peptide analysis, throughout all studies presented herein, special attention to details of each step was given, from theoretical bioinformatics calculations, sample preparation and wet lab protocol optimization to data acquisition, processing and analysis.

With advancement of the field, analytical characteristics, such as increased mass resolution and mass accuracy, became accessible in many laboratories across the world. Hence, initial proof-of-principle experiments that attest the viability and importance of the MDP commenced, and the field of proteomics has already defined a set of proteases capable of cleaving proteins with different *specificities*. However, there are very few proteases that were suggested for MD [117, 153].This may be because, from an evolutionary point of view, a biologically efficient protease can accomplish protein degradation readily. Only in a few cases is the degradation process aimed at targeting rare amino acids or amino acid sequences; oftentimes the final biological function of a protease is simply facilitating protein turnover within a cell or, in the case of a secreted protease, degrading proteins present in the environment into smaller peptides. For both of these tasks, the specificity of the protease should be either very limited or directed towards non-rare amino acids. Choice of the agent for MD that would produce peptides in a desired way with respect to the selected mass bin requires a thorough theoretical cross-kingdoms study to reveal a good target residue, or a combination of good target residues.

In particular, we considered that it is of fundamental importance to tackle selecting the best *theoretical targets* for protein cleavage in MD from two different points of view. The first is the *frequency* of occurrence of each of the

20 amino acids in all proteomes. *Paper I* of this Thesis focuses on bioinformatics studies in order to define the set of residues to target for performing middle-down experiments. This bioinformatics survey was crucial for understanding: which residue(s) would be a good target for obtaining long peptides, if that residue comes as a first candidate across all kingdoms, and finally, how much of a proteome (expressed as a percentage of the total number of proteins within) we can map with these peptides. The latter one is particularly interesting, as it is not obvious how if one yields a pool of peptides with an average mass which fits the selected mass bin, this will not necessarily mean that all of the peptide masses are equally partitioned or that the distribution of masses is uniform. This means that not all peptides will be successfully analyzed and lead to an identification in MD experimental setup. Interestingly, the survey revealed how there is no single residue, or any combination of residues that we could target to obtain 100 % proteome coverage in any proteolytic-based working regime (BUP, the newly-defined eBUP and MDP). However, results indicated that targeting dibasic and rare amino acids such as methionine, tryptophan or cysteine would identify up to 90 % of the proteome with a drastic reduction in the number of yielded peptides (five-fold compared to BUP). Reduction in eluting peptides per LC run is a desirable characteristic that our approach could provide, and, from there, we wanted to evaluate how these peptides ionize and how their terminal residues can aid in obtaining a successful fragmentation ladder upon different ion activation and dissociation techniques, particularly the charge-dependent ones such as ETD and HCD.

Hence, the second consideration we made for selection of MD cleavage site was the *position of specific amino acid residues*, such as the basic ones, within the sequence of the obtained peptide. The mentioned aspect, in combination with the average size of obtained polypeptides, is important as it might potentially affect the peptide fragmentation, defining which ion activation technique is more suitable for this category of proteolytic peptides. *Paper II* discusses the effect of specific positioning of basic residues (namely lysines) on the fragmentation of large (>20 amino acids) synthetic peptides.

Specifically, we investigated the result of different cleavages at dibasic sites (before the site, in-between the basic residues, and after the dibasic pair), to observe the effect produced by charge location on the radical-mediated (ETD) and energy threshold-based (HCD) fragmentation. The study also aimed at determining optimal HCD and ETD parameters for fragmentation of middle-down range peptides. Interestingly, for both HCD and ETD, we found a direct correlation between precursor charge state and obtained sequence coverage, regardless of the position of the basic residues belonging to the dibasic site. The analysis of the observed product ions revealed that in HCD the distal positioning of the two fixed charges is detrimental for the cleavage of peptide bonds in the middle of the peptide sequence, even at the collision energies where virtually no precursor ion remains. Conversely, in the same condition in ETD, we observed a charge reduction phenomenon at the N-terminal basic amino acid, as demonstrated by the absence of light *c*-ions (up to *c6*). Final results showed how optimal ETD ion-ion interaction time changes from 10-15 ms for the longest peptides to around 50 ms for 2 kDa peptides. The optimal NCE value we identified for HCD for MD was 27 %, which is the value used in all subsequent experimental studies conducted.

Next, we found agents, developed and proposed workflows for each of the identified targets (dibasic, M, W, and C). The first study characterized a novel protease that targets basic/dibasic residue at scissile bond. Results of *Paper III* present a detailed assessment of such a candidate: a secreted aspartic protease Sap9. Biochemical properties and activity of Sap9 are better than of other proteases in the acidic pH, but Sap9 was found in our studies to be not highly specific without engineering. The generated peptides are almost equally distributed in 3-7 kDa; hence, even though it was not the original aim of the research, the study of *Paper IV* describes the result of a specific application of Sap9 for the structural analysis of immunoglobulins G. The study describes the experimental results of a novel Sap9-based eBUP approach aiming to further improve in-depth structural analysis of monoclonal antibodies and their mixtures. Key features of Sap9-based eBUP IgG analysis include extensive antibody sequence coverage with up to 100% for light chain and up to 99-

100% for heavy chain in a single LC-MS/MS run, sequence information on connectivity of complementarity determining regions (CDRs) and, importantly, reduced artifact introduction (e.g., deamidation) during proteolysis with Sap9 compared to conventional bottom-up proteomics workflows, e.g., with trypsin. This allowed us to investigate endogenous deamidation in CDR regions in therapeutic IgG1, trastuzumab. Importantly, this was the first attempt of a proteolytic MS-based quantitation, and we obtained comparable results to other orthogonal methods. Since the publication of this paper, a particularly strong interest was given to Sap9-based IgG analysis from both academic and industrial (pharmaceutical) groups, with a number of them now using Sap9 for IgG analysis in their drug development workflows.

In the following, we developed new protocols and workflows for chemical-mediated middle-down mass spectrometry and proteomics, which could offer similarly successful results to those related to Sap9, but potentially more specific and able to be used for quantitative and qualitative proteomics. *Paper V* describes a thorough evaluation of different protocols targeting rare amino acids and elucidates cysteine cleavage as the novel selective MD avenue. During this study, we successfully defined novel cleavage rules and strategy for tailoring a dedicated MD search engine, which is to date still not available. This methodology was of particular interest when applied to IgG, as it allowed confident paratope ID with significant reduction in the number of peptides whilst their average size was true MD range (5-7 kDa). Additionally, one of the important variable domains (CDR3) would with this approach always be located on the N-terminus of the protein (starting with chemically-tagged cysteine) and would therefore be more susceptible to complete sequencing. These considerations encompass two novel methods for MD we proposed: primary-sequence dependent enzymatic proteolysis with Sap 9 and chemical-mediated hydrolysis N-terminal to cysteine.

Finally, *papers VI* and *VII* bring a structure-dependent MD approach integrated with TD-like MS settings we developed for characterization of antibodies. The main objective of this work was improving IgG drug discovery pipeline via matching the pairs of light and heavy chains. The applied MD proteolysis is based on a single structure-specific cleavage either with papain (*Paper VI*) or a novel protease GinigisKHAN (*Paper VII*) and analyzed with ETD MS/MS in TD fashion. Both of these approaches were developed and applied to overcome the intact IgG sequence coverage limit of TD (33 %) while obtaining important chain pairing information. These methods are first attempts for a novel workflow that assesses pairing information on 50 kDa subunits. Additionally, these studies were focused on data acquisition and subsequent analysis, and they represent the first time evaluation of spectral full and reduced profile mode and transient averaging for protein analysis using Orbitrap FTMS. This pipeline was successfully tested on IgG (up to 34 % and 41 % sequence coverage for LC and HC respectively), and we believe that these tests could be considered *proof-of-principle* for future application in TD analysis of large proteins. Overall, these developments complete the triad of methods for structural analysis of antibodies introduced by our group: top-down, middle-down, and extended bottom-up, allowing for their comprehensive structural analysis.

8.2. Concluding remarks and future perspectives.

Advancing a new field is always a challenging task which requires a lot of considerations that ultimately can be proven wrong. Oftentimes it happens that a series of time-consuming and expensive experiments does not offer expected results, and, less often, those lead to *positive experimental results* that are conclusive and definite. Middle-down proteomics has the potential for proposing new avenues in digestion-based shotgun studies. The peculiar feature of MD is the size of the generated peptides. The research project presented in this Thesis as *Paper IV* clearly illustrates that, at least for selected applications, the unambiguous identification of a certain protein is not possible using single (or even multiple, in specific cases) short peptides like tryptic ones. Conversely, MD-sized peptides can allow confident identification which, passing from a targeted to a large-scale type of experiment, could translate into the possibility of including in the identification additional information, (i.e., about splicing variants or isoforms, which is generally lost in BUP). Effectively, the Human Proteome Project Organization (HUPO) is proposing progressively more stringent criteria for the validation of BUP data. Particularly, the guidelines for the 2015 version of the Project propose the need not only to apply a 1% false discovery rate threshold at the protein level to filter the data, but also to include in the list of identified proteins only those described by *two uniquely mapping identified peptides* of at least nine amino acid long [154]. In this context, MD could be seen as a means to create a level of high-throughput analysis of the human (and others) proteomes similar to the level currently reached by shotgun BUP [155, 156], but with the possibility of confidently identifying a protein using *one long peptide* only. Although the effectiveness of MD in research projects focused on extremely complex proteomes such as the eukaryotic ones is not fully demonstrated, some pioneering studies suggest that MDP is applicable to large-scale studies [117]. However, the utility of developing a new proteomic pipeline with a novel set of specific rules exceeds the simple practical result of having, at the end of the process, a new tool for the investigation of protein mixtures. Establishing new paradigms, in fact, might help also to put into discussion old ones, so that eventually already established technology can be further optimized and refined. For instance, the widely accepted opinion is that for proteomics we should need one protease for all organisms and one method for all biological problems to investigate. Although the attention of scientists in the world is currently focused on increasing the number of protein identifications and accuracy of protein quantification, considering as an example the human proteome and its sophisticated organization, it is probable how we need to address each biological question as an individual one, and not try to apply a default patch to all, with the result that when that does not work for all problems, it is discarded as insufficient. Not a single MDP approach for all, but rather a myriad of approaches for particular targeted protein analysis. The work presented in this Thesis not only puts forward the equation "*larger peptides=more information*" and advocates for it, but also introduces different methods, targeting different amino acid residues (including some, like cysteines, that have rarely or never been used in proteomics) so that it could be potentially possible to obtain MD-sized peptides for specific sub-proteomes. Effectively this approach is already applied to histones, which include a large number of basic amino acids and can be thus better treated with proteases specific for acid residues (such as Glu-C) for generating peptides bigger than 5 kDa [157]. Major bottleneck in validation of MDP data analysis is *the limitation* in the currently available workflow for top-down mass spectrometry.

This Thesis showed clear potential of MD and contributed to its further advancement towards a well-defined stand-alone approach. In particular, we clarified in detail how to generate longer-size peptides, proposed analyte-dependent proteolysis, offered novel pipelines and proposed a set of general guidelines: residues to target, methods and agents to use, ion activation techniques and their optimal parameters, all the way to data analysis and software requirements. We showed experimentally the potential, robustness, efficiency and real-life applicability of proposed methodologies in a targeted proteomics survey on a presently very important class of proteins (IgGs). Besides some practical aspects to improve, in parallel with instrumentation, method development and data analysis tools, it is easy to envision MD as a mature approach in the near future. The position of MD alongside BUP and TD with respect to the scientific awareness and feasibility will fluctuate with time. Today MDP could be benchmarked for targeted analysis of limited sample pool such as sequence verification of IgG candidates for drug delivery, mapping of mutually dependent adjacent modifications and their subsequent quantitation (e.g., 'cross-talk' analysis of histones), detection of single point mutations from low-abundant proteins, such as those present in biological, clinically relevant samples due to the potential for a complete or close to complete sequence coverage from a single proteomic experiment. Considering the variety of mutations that may be present on a single protein, identification of the position of the mutation may be critical for the correct diagnosis, for determining the outcome of the disease and for finding the most suitable treatment.

In my opinion, optimizing current shortcomings (i.e., increased spectral complexity despite the sample pool reduction), optimizing front-end separations and tweaking the proteolysis to the analyte characteristics, is a challenging initiative and a ‘must do’ task that will prove to be very important for the future MD applications. While TD remains the only approach that can provide identification of proteoforms, contribution of proteolysis-based approaches is still very important, and I envision in the future integration of BUP, MD and TD approaches, where MD will have a significant complementary contribution in large–scale characterization of protein families and extended PTM studies.

References

1. Heisenberg, W., *Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik.* Zeitschrift für Physik, 1927. **43**(3-4): p. 172-198.

2. Strippoli, P., et al., *Uncertainty principle of genetic information in a living cell.* Theor Biol Med Model, 2005. **2**: p. 40.

3. Clamp, M., et al., *Distinguishing protein-coding and noncoding genes in the human genome.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(49): p. 19428-19433.

4. Collins, F.S., M. Morgan, and A. Patrinos, *The human genome project: Lessons from large-scale biology.* Science, 2003. **300**(5617): p. 286-290.

5. Nørregaard Jensen, O., *Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry.* Current Opinion in Chemical Biology, 2004. **8**(1): p. 33-41.

6. Yates, J.R., C.I. Ruse, and A. Nakorchevsky, *Proteomics by Mass Spectrometry: Approaches, Advances, and Applications.* Annual Review of Biomedical Engineering, 2009. **11**: p. 49-79.

7. Corthals, G.L., et al., *The dynamic range of protein expression: A challenge for proteomic research.* Electrophoresis, 2000. **21**(6): p. 1104-1115.

8. Cooks, R.G.a.R., A. L., *The Thompson'. A suggested unit for mass spectroscopists.* Rapid Commun. Mass spectrom, 1991. **5**(93).

9. Thompson, J.J., *Ray of Positive Electricity and Their Application to Chemical analysis.* 1913: Longmans Green, London.

10. Tsybin, Y.O., et al., *High-Resolution and Tandem Mass Spectrometry &#8211; the Indispensable Tools of the XXI Century.* CHIMIA International Journal for Chemistry, 2011. **65**(9): p. 641-645.

11. Aebersold, R. and M. Mann, *Mass spectrometry-based proteomics.* Nature, 2003. **422**(6928): p. 198-207.

12. Chait, B.T., *Mass Spectrometry: Bottom-Up or Top-Down?* Science, 2006. **314**(5796): p. 65-66.

13. Chait, B.T., *Mass spectrometry in the postgenomic era.* Annu Rev Biochem, 2011. **80**: p. 239-46.

14. Cox, J. and M. Mann, *Quantitative, high-resolution proteomics for data-driven systems biology.* Annu Rev Biochem, 2011. **80**: p. 273-99.

15. Huttlin, E.L., et al., *A tissue-specific atlas of mouse protein phosphorylation and expression.* Cell, 2010. **143**(7): p. 1174-89.

16. Lemeer, S. and A.J. Heck, *The phosphoproteomics data explosion.* Curr Opin Chem Biol, 2009. **13**(4): p. 414-20.

17. Ewing, R.M., et al., *Large-scale mapping of human protein-protein interactions by mass spectrometry.* Mol Syst Biol, 2007. **3**: p. 89.

18. Smith, L.M., N.L. Kelleher, and P. Consortium for Top Down, *Proteoform: a single term describing protein complexity.* Nat Methods, 2013. **10**(3): p. 186-7.

19. Cherry, J.M., et al., *Saccharomyces Genome Database: the genomics resource of budding yeast.* Nucleic Acids Research, 2012. **40**(D1): p. D700-D705.

20. Collins, F.S., et al., *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-945.

21. Creasy, D.M. and J.S. Cottrell, *Unimod: Protein modifications for mass spectrometry.* Proteomics, 2004. **4**(6): p. 1534-1536.

22. Kampa, D., et al., *Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22.* Genome Res, 2004. **14**(3): p. 331-42.

23. Nielsen, M.L., M.M. Savitski, and R.A. Zubarev, *Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics.* Molecular & Cellular Proteomics, 2006. **5**(12): p. 2384-2391.

24. Ghaemmaghami, S., et al., *Global analysis of protein expression in yeast.* Nature, 2003. **425**(6959): p. 737-741.

25. Beck, M., et al., *The quantitative proteome of a human cell line.* Molecular Systems Biology, 2011. **7**.

26. Anderson, N.L. and N.G. Anderson, *The human plasma proteome - History, character, and diagnostic prospects.* Molecular & Cellular Proteomics, 2002. **1**(11): p. 845-867.

27. Hortin, G.L. and D. Sviridov, *The dynamic range problem in the analysis of the plasma proteome.* Journal of Proteomics, 2010. **73**(3): p. 629-636.

28. Zubarev, R.A., *The challenge of the proteome dynamic range and its implications for in-depth proteomics.* Proteomics, 2013. **13**(5): p. 723-726.

29. Schwanhausser, B., et al., *Global quantification of mammalian gene expression control.* Nature, 2011. **473**(7347): p. 337-342.

30. Michalski, A., J. Cox, and M. Mann, *More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS.* J Proteome Res, 2011. **10**(4): p. 1785-93.

31. Nesvizhskii, A.I. and R. Aebersold, *Interpretation of shotgun proteomic data: the protein inference problem.* Mol Cell Proteomics, 2005. **4**(10): p. 1419-40.

32. Chamot-Rooke, J., et al., *Posttranslational modification of pili upon cell contact triggers N. meningitidis dissemination.* Science, 2011. **331**(6018): p. 778-82.

33. Catherman, A.D., O.S. Skinner, and N.L. Kelleher, *Top Down proteomics: facts and perspectives.* Biochem Biophys Res Commun, 2014. **445**(4): p. 683-93.

34. Mao, Y., et al., *Top-down structural analysis of an intact monoclonal antibody by electron capture dissociation-fourier transform ion cyclotron resonance-mass spectrometry.* Anal Chem, 2013. **85**(9): p. 4239-46.

35. Fornelli, L., et al., *Analysis of intact monoclonal antibody IgG1 by electron transfer dissociation Orbitrap FTMS.* Mol Cell Proteomics, 2012. **11**(12): p. 1758-67.

36. Tsybin, Y.O., et al., *Structural analysis of intact monoclonal antibodies by electron transfer dissociation mass spectrometry.* Anal Chem, 2011. **83**(23): p. 8919-27.

37. Cannon, J., et al., *High-throughput middle-down analysis using an orbitrap.* J Proteome Res, 2010. **9**(8): p. 3886-90.

38. Taverna, S.D., et al., *Long-distance combinatorial linkage between methylation and acetylation on histone H3N termini.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(7): p. 2086-2091.

39. Wu, S.L., et al., *On-line LC-MS approach combining collision-induced dissociation (CID), electron-transfer dissociation (ETD), and CID of an isolated charge-reduced species for the trace-level characterization of proteins with post-translational modifications.* Journal of Proteome Research, 2007. **6**(11): p. 4230-4244.

40. Hohmann, L., et al., *Proteomic Analyses Using Grifola frondosa Metalloendoprotease Lys-N.* Journal of Proteome Research, 2009. **8**(3): p. 1415-1422.

41. Hauser, N.J., et al., *Electron transfer dissociation of peptides generated by microwave D-cleavage digestion of proteins.* Journal of Proteome Research, 2008. **7**(5): p. 1867-1872.

42. Swatkoski, S., et al., *Evaluation of microwave-accelerated residue-specific acid cleavage for proteomic applications.* Journal of Proteome Research, 2008. **7**(2): p. 579-586.

43. Johnson, G. and T.T. Wu, *Kabat Database and its applications: 30 years after the first variability plot.* Nucleic Acids Research, 2000. **28**(1): p. 214-218.

44. Stephenson, R.C. and S. Clarke, *Succinimide formation from aspartyl and asparaginyl peptides as a model for the spontaneous degradation of proteins.* Journal of Biological Chemistry, 1989. **264**(11): p. 6164-6170.

45. Hurtado, P.P. and P.B. O'Connor, *Differentiation of isomeric amino acid residues in proteins and peptides using mass spectrometry.* Mass Spectrom Rev, 2012. **31**(6): p. 609-25.

46. Cournoyer, J.J., et al., *Deamidation: Differentiation of aspartyl from isoaspartyl products in peptides by electron capture dissociation.* Protein Sci, 2005. **14**(2): p. 452-63.

47. Li, X., C. Lin, and P.B. O'Connor, *Glutamine deamidation: differentiation of glutamic acid and gamma-glutamic acid in peptides by electron capture dissociation.* Anal Chem, 2010. **82**(9): p. 3606-15.

48. O'Connor, P.B., et al., *Differentiation of aspartic and isoaspartic acids using electron transfer dissociation.* Journal of the American Society for Mass Spectrometry, 2006. **17**(1): p. 15-19.

49. Chelius, D., et al., *Automated tryptic digestion procedure for HPLC/MS/MS peptide mapping of immunoglobulin gamma antibodies in pharmaceutics.* Journal of Pharmaceutical and Biomedical Analysis, 2008. **47**(2): p. 285-294.

50. Ayoub, D., et al., *Correct primary structure assessment and extensive glyco-profiling of cetuximab by a combination of intact, middle-up, middle-down and bottom-up ESI and MALDI mass spectrometry techniques.* mAbs, 2013. **5**(5): p. 699-710.

51. Debaene, F., et al., *Time resolved native ion-mobility mass spectrometry to monitor dynamics of IgG4 Fab arm exchange and "bispecific" monoclonal antibody formation.* Anal Chem, 2013. **85**(20): p. 9785-92.

52. Rosati, S., et al., *Exploring an orbitrap analyzer for the characterization of intact antibodies by native mass spectrometry.* Angew Chem Int Ed Engl, 2012. **51**(52): p. 12992-6.

53. Marshall, A.G. and C.L. Hendrickson, *High-Resolution Mass Spectrometers.* Annual Review of Analytical Chemistry, 2008. **1**(1): p. 579-599.

54. Fenyo, D., J. Qin, and B.T. Chait, *Protein identification using mass spectrometric information.* Electrophoresis, 1998. **19**(6): p. 998-1005.

55. Zubarev, R.A., P. Håkansson, and B. Sundqvist, *Accuracy Requirements for Peptide Characterization by Monoisotopic Molecular Mass Measurements.* Analytical Chemistry, 1996. **68**(22): p. 4060-4063.

56. McNaught, A.D., Wilkinson, A., *International Union of Pure and Applied Chemistry, Compendium of chemical terminology: IUPAC recommendations 2nd ed.* Vol. vii. 1997, Oxford Oxfordshire, Malden, MA: Blackwell Science.

57. Zhurov, K.O., et al., *Distinguishing Analyte from Noise Components in Mass Spectra of Complex Samples: Where to Cut the Noise?* Analytical Chemistry, 2014. **86**(7): p. 3308-3316.

58. Horn, D.M., R.A. Zubarev, and F.W. McLafferty, *Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules.* Journal of the American Society for Mass Spectrometry, 2000. **11**(4): p. 320-332.

59. Marshall, A.G., *Theoretical signal-to-noise ratio and mass resolution in Fourier transform ion cyclotron resonance mass spectrometry.* Analytical Chemistry, 1979. **51**(11): p. 1710-1714.

60. Makarov, A., et al., *Dynamic range of mass accuracy in LTQ Orbitrap hybrid mass spectrometer.* J Am Soc Mass Spectrom, 2006. **17**(7): p. 977-82.

61. K.R., J., *Collision-induced decompositions of aromatic molecular ions.* Int. J. Mass. Spectrom.Ion Phys., 1968. **1**: p. 227-235.

62. McLafferty, F.W., *Tandem Mass Spectrometry.* Wiley-Interscience, New York, 1983.

63. Hoffmann, E.d., Stroobant, V., Hoffmann, E. d., Stroobant, V., *Mass spectrometry : principles and applications. 3rd ed.* . Chichester, England ; Hoboken, NJ: J. Wiley, 2007. **xii**: p. 489 p.

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

65. Zhurov, K.O., et al., *Principles of electron capture and transfer dissociation mass spectrometry applied to peptide and protein structure analysis.* Chem Soc Rev, 2013. **42**(12): p. 5014-30.

66. Little, D.P., et al., *Infrared multiphoton dissociation of large multiply charged ions for biomolecule sequencing.* Anal Chem, 1994. **66**(18): p. 2809-15.

67. Hofstadler, S.A., K.A. Sannes-Lowery, and R.H. Griffey, *Infrared multiphoton dissociation in an external ion reservoir.* Anal Chem, 1999. **71**(11): p. 2067-70.

68. Bowers, W.D., S.S. Delbert, and R.T. McIver, Jr., *Consecutive laser-induced photodissociation as a probe of ion structure.* Anal Chem, 1986. **58**(4): p. 969-72.

69. Brodbelt, J., *Shedding Light on the Frontier of Photodissociation.* Journal of The American Society for Mass Spectrometry, 2011. **22**(2): p. 197-206.

70. Hunt, D.F., et al., *Sequence analysis of polypeptides by collision activated dissociation on a triple quadrupole mass spectrometer.* Biological Mass Spectrometry, 1981. **8**(9): p. 397-408.

71. Johnson, R.S., S.A. Martin, and K. Biemann, *Collision-induced fragmentation of (M + H)+ ions of peptides. Side chain specific sequence ions.* International Journal of Mass Spectrometry and Ion Processes, 1988. **86**: p. 137-154.

72. McLuckey, S.A., *Principles of collisional activation in analytical mass spectrometry.* J Am Soc Mass Spectrom, 1992. **3**(6): p. 599-614.

73. Wells, J.M. and S.A. McLuckey, *Collision-induced dissociation (CID) of peptides and proteins.* Methods Enzymol, 2005. **402**: p. 148-85.

74. de Godoy, L.M.F., et al., *Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast.* Nature, 2008. **455**(7217): p. 1251-1254.

75. Sleno, L. and D.A. Volmer, *Ion activation methods for tandem mass spectrometry.* J Mass Spectrom, 2004. **39**(10): p. 1091-112.

76. Olsen, J.V., et al., *Higher-energy C-trap dissociation for peptide modification analysis.* Nat Methods, 2007. **4**(9): p. 709-12.

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

78. Zubarev, R.A., et al., *Electron capture dissociation for structural characterization of multiply charged protein cations.* Anal Chem, 2000. **72**(3): p. 563-73.

79. Zubarev, R.A., et al., *Electron Capture Dissociation of Gaseous Multiply-Charged Proteins Is Favored at Disulfide Bonds and Other Sites of High Hydrogen Atom Affinity.* Journal of the American Chemical Society, 1999. **121**(12): p. 2857-2862.

80. Sawicka, A., et al., *Model Calculations Relevant to Disulfide Bond Cleavage via Electron Capture Influenced by Positively Charged Groups.* The Journal of Physical Chemistry B, 2003. **107**(48): p. 13505-13511.

81. Leymarie, N., C.E. Costello, and P.B. O'Connor, *Electron capture dissociation initiates a free radical reaction cascade.* J Am Chem Soc, 2003. **125**(29): p. 8949-58.

82. Syrstad, E.A. and F. Turecek, *Toward a general mechanism of electron capture dissociation.* J Am Soc Mass Spectrom, 2005. **16**(2): p. 208-24.

83. Breuker, K., et al., *Nonergodic and conformational control of the electron capture dissociation of protein cations.* Proc Natl Acad Sci U S A, 2004. **101**(39): p. 14011-6.

84. Turecek, F. and R.R. Julian, *Peptide radicals and cation radicals in the gas phase.* Chem Rev, 2013. **113**(8): p. 6691-733.

85. Wodrich, M.D., et al., *Heterolytic N-Calpha bond cleavage in electron capture and transfer dissociation of peptide cations.* J Phys Chem B, 2012. **116**(35): p. 10807-15.

86. Zhurov, K.O., et al., *Ping-pong protons: how hydrogen-bonding networks facilitate heterolytic bond cleavage in peptide radical cations.* J Phys Chem B, 2014. **118**(10): p. 2628-37.

87. Wodrich, M.D., et al., *On the viability of heterolytic peptide N-C(alpha) bond cleavage in electron capture and transfer dissociation mass spectrometry.* J Phys Chem B, 2014. **118**(11): p. 2985-92.

88. Syka, J.E., et al., *Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry.* Proc Natl Acad Sci U S A, 2004. **101**(26): p. 9528-33.

89. Pitteri, S.J., et al., *Electron Transfer Ion/Ion Reactions in a Three-Dimensional Quadrupole Ion Trap: Reactions of Doubly and Triply Protonated Peptides with SO2•*. Analytical Chemistry, 2005. **77**(6): p. 1831-1839.

90. Tsybin, Y.O., et al., *Combined infrared multiphoton dissociation and electron capture dissociation with a hollow electron beam in Fourier transform ion cyclotron resonance mass spectrometry.* Rapid Communications in Mass Spectrometry, 2003. **17**(15): p. 1759-1768.

91. Swaney, D.L., et al., *Supplemental Activation Method for High-Efficiency Electron-Transfer Dissociation of Doubly Protonated Peptide Precursors.* Analytical Chemistry, 2007. **79**(2): p. 477-485.

92. Good, D.M., et al., *Performance characteristics of electron transfer dissociation mass spectrometry.* Mol Cell Proteomics, 2007. **6**(11): p. 1942-51.

93. Xia, Y., et al., *Effects of cation charge-site identity and position on electron-transfer dissociation of polypeptide cations.* J Am Chem Soc, 2007. **129**(40): p. 12232-43.

94. Mirgorodskaya, E., P. Roepstorff, and R.A. Zubarev, *Localization of O-glycosylation sites in peptides by electron capture dissociation in a Fourier transform mass spectrometer.* Anal Chem, 1999. **71**(20): p. 4431-6.

95. Creese, A.J. and H.J. Cooper, *The effect of phosphorylation on the electron capture dissociation of peptide ions.* J Am Soc Mass Spectrom, 2008. **19**(9): p. 1263-74.

96. Kelleher, N.L., et al., *Localization of labile posttranslational modifications by electron capture dissociation: the case of gamma-carboxyglutamic acid.* Anal Chem, 1999. **71**(19): p. 4250-3.

97. Stensballe, A., et al., *Electron capture dissociation of singly and multiply phosphorylated peptides.* Rapid Commun Mass Spectrom, 2000. **14**(19): p. 1793-800.

98. Kleinnijenhuis, A.J., et al., *Analysis of histidine phosphorylation using tandem MS and ion-electron reactions.* Anal Chem, 2007. **79**(19): p. 7450-6.

99. Kweon, H.K. and K. Hakansson, *Metal oxide-based enrichment combined with gas-phase ion-electron reactions for improved mass spectrometric characterization of protein phosphorylation.* J Proteome Res, 2008. **7**(2): p. 749-55.

100. Ong, S.E., G. Mittler, and M. Mann, *Identifying and quantifying in vivo methylation sites by heavy methyl SILAC.* Nat Methods, 2004. **1**(2): p. 119-26.

101. Zhang, K., et al., *Histone acetylation and deacetylation: identification of acetylation and methylation sites of HeLa histone H4 by mass spectrometry.* Mol Cell Proteomics, 2002. **1**(7): p. 500-8.

102. Choudhary, C. and M. Mann, *Decoding signalling networks by mass spectrometry-based proteomics.* Nat Rev Mol Cell Biol, 2010. **11**(6): p. 427-39.

103. Yang, H., et al., *Toward proteome-scale identification and quantification of isoaspartyl residues in biological samples.* J Proteome Res, 2009. **8**(10): p. 4615-21.

104. Yang, H. and R.A. Zubarev, *Mass spectrometric analysis of asparagine deamidation and aspartate isomerization in polypeptides.* Electrophoresis, 2010. **31**(11): p. 1764-72.

105. Kyte, J. and R.F. Doolittle, *A simple method for displaying the hydropathic character of a protein.* Journal of Molecular Biology, 1982. **157**(1): p. 105-132.

106. Taouatas, N., et al., *Straightforward ladder sequencing of peptides using a Lys-N metalloendopeptidase.* Nat Methods, 2008. **5**(5): p. 405-7.

107. Scholten, A., et al., *In-depth quantitative cardiac proteomics combining electron transfer dissociation and the metalloendopeptidase Lys-N with the SILAC mouse.* Mol Cell Proteomics, 2011. **10**(10): p. O111 008474.

108. Gauci, S., et al., *Lys-N and trypsin cover complementary parts of the phosphoproteome in a refined SCX-based approach.* Anal Chem, 2009. **81**(11): p. 4493-501.

109. Nesvizhskii, A.I., O. Vitek, and R. Aebersold, *Analysis and validation of proteomic data generated by tandem mass spectrometry.* Nat Meth, 2007. **4**(10): p. 787-797.

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

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

112. Craig, R. and R.C. Beavis, *A method for reducing the time required to match protein sequences with tandem mass spectra.* Rapid Commun Mass Spectrom, 2003. **17**(20): p. 2310-6.

113. Geer, L.Y., et al., *Open mass spectrometry search algorithm.* J Proteome Res, 2004. **3**(5): p. 958-64.

114. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nat Biotechnol, 2008. **26**(12): p. 1367-72.

115. Laskay, U.A., et al., *Extended bottom-up proteomics with secreted aspartic protease Sap9.* J Proteomics, 2014. **110**: p. 20-31.

116. Fenselau, C., O. Laine, and S. Swatkoski, *Microwave assisted acid cleavage for denaturation and proteolysis of intact human adenovirus.* Int J Mass Spectrom, 2011. **301**(1-3): p. 7-11.

117. Wu, C., et al., *A protease for 'middle-down' proteomics.* Nat Meth, 2012. **9**(8): p. 822-824.

118. Senko, M.W., S.C. Beu, and F.W. McLaffertycor, *Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions.* J Am Soc Mass Spectrom, 1995. **6**(4): p. 229-33.

119. Valkenborg, D., I. Jansen, and T. Burzykowski, *A model-based method for the prediction of the isotopic distribution of peptides.* J Am Soc Mass Spectrom, 2008. **19**(5): p. 703-12.

120. Liu, X.W., et al., *Protein Identification Using Top-Down.* Molecular & Cellular Proteomics, 2012. **11**(6).

121. Fellers, R.T., et al., *ProSight Lite: graphical software to analyze top-down mass spectrometry data.* Proteomics, 2015. **15**(7): p. 1235-8.

122. Cai, W., et al., *MASH Suite Pro: A Comprehensive Software Tool for Top-down Proteomics.* Molecular & Cellular Proteomics, 2015.

123. Mann, M., Meng, C.K., Fenn, J.B., *Interpreting mass spectra of multiply charged ions.* Anal. Chem, 1989. **61**: p. 1702-8.

124. Wilm, M.S. and M. Mann, *Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last?* International Journal of Mass Spectrometry and Ion Processes, 1994. **136**(2): p. 167-180.

125. Körner, R., et al., *Nano electrospray combined with a quadrupole ion trap for the analysis of peptides and protein digests.* Journal of the American Society for Mass Spectrometry, 1996. **7**(2): p. 150-156.

126. Rayleigh, *XX. On the equilibrium of liquid conducting masses charged with electricity.* Philosophical Magazine Series 5, 1882. **14**(87): p. 184-186.

127. Taflin D.C. Ward, T.L.a.D., E.J., *Electrified Droplet Fission and the Rayleigh Limit.* Langmuir, 1989. **5**: p. 376-384.

128. Kelly, M.A., et al., *Electrospray analysis of proteins: A comparison of positive-ion and negative-ion mass spectra at high and low pH.* Organic Mass Spectrometry, 1992. **27**(10): p. 1143-1147.

129. Wilm, M., et al., *Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry.* Nature, 1996. **379**(6564): p. 466-9.

130. Onisko, B., et al., *Mass Spectrometric Detection of Attomole Amounts of the Prion Protein by nanoLC/MS/MS.* Journal of the American Society for Mass Spectrometry, 2007. **18**(6): p. 1070-1079.

131. Glish, G.L., et al., *A New Hybrid Sector Quadrupole Mass-Spectrometer for Mass-Spectrometry Mass-Spectrometry.* International Journal of Mass Spectrometry and Ion Processes, 1982. **41**(3): p. 157-177.

132. Schwartz, J.C., M.W. Senko, and J.E. Syka, *A two-dimensional quadrupole ion trap mass spectrometer.* J Am Soc Mass Spectrom, 2002. **13**(6): p. 659-69.

133. Syka, J.E.P. and W.J. Fies, *Fourtier transform quadrupole mass spectrometer and method.* 1988, Google Patents.

134. Schwartz, J.C. and M.W. Senko, *Two-dimensional quadrupole ion trap operated as a mass spectrometer.* 2004, Google Patents.

135. Guilhaus, M., D. Selby, and V. Mlynski, *Orthogonal acceleration time-of-flight mass spectrometry.* Mass Spectrometry Reviews, 2000. **19**(2): p. 65-107.

136. Chernushevich, I.V., A.V. Loboda, and B.A. Thomson, *An introduction to quadrupole-time-of-flight mass spectrometry.* Journal of Mass Spectrometry, 2001. **36**(8): p. 849-865.

137. Marshall, A.G. and C.L. Hendrickson, *High-Resolution Mass Spectrometers.* Annual Review of Analytical Chemistry, 2008. **1**: p. 579-599.

138. Makarov, A., *Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis.* Anal Chem, 2000. **72**(6): p. 1156-62.

139. Michalski, A., et al., *Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes.* Mol Cell Proteomics, 2012. **11**(3): p. O111 013698.

140. Y. Naito, M.I., *Peak confluence phenomenon in Fourier transform ion cyclotron resonance phenomenon.* J. Mass Spectrom. Soc. Jpn, 1994. **42**(1).

141. Gorshkov, M.V., L. Fornelli, and Y.O. Tsybin, *Observation of ion coalescence in Orbitrap Fourier transform mass spectrometry.* Rapid Commun Mass Spectrom, 2012. **26**(15): p. 1711-7.

142. Tarasova I. A, Surin A, K., Fornelli L., Pridatchenko M. L., Suvorina M. Y., Gorshkov M. V., *Ion coalescence in Fourier transform mass spectrometry: should we worry about this in shotgun proteomics?* Eur J Mass Spectrom (Chichester, Eng). 2015. **21**(3): p. 459-70.

143. Werner, T., et al., *Ion Coalescence of Neutron Encoded TMT 10-Plex Reporter Ions.* Analytical Chemistry, 2014. **86**(7): p. 3594-3601.

144. Laskay, U.A., et al., *Practical considerations for improving the productivity of mass spectrometry-based proteomics.* Chimia (Aarau), 2013. **67**(4): p. 244-9.

145. Desmet, G., D. Cabooter, and K. Broeckhoven, *Graphical Data Representation Methods To Assess the Quality of LC Columns.* Analytical Chemistry, 2015. **87**(17): p. 8593-8602.

146. Vaast, A., et al., *Gradient-elution parameters in capillary liquid chromatography for high-speed separations of peptides and intact proteins (vol 1355, pg 149, 2014).* Journal of Chromatography A, 2014. **1366**: p. 137-137.

147. Vandeemter, J.J., F.J. Zuiderweg, and A. Klinkenberg, *Longitudinal Diffusion and Resistance to Mass Transfer as Causes of Nonideality in Chromatography.* Chemical Engineering Science, 1956. **5**(6): p. 271-289.

148. Gusev, I., X. Huang, and C. Horvath, *Capillary columns with in situ formed porous monolithic packing for micro high-performance liquid chromatography and capillary electrochromatography.* Journal of Chromatography A, 1999. **855**(1): p. 273-290.

149. Sweredoski, M.J., et al., *High Resolution Parallel Reaction Monitoring with Electron Transfer Dissociation for Middle-Down Proteomics.* Analytical Chemistry, 2015. **87**(16): p. 8360-8366.

150. Frese, C.K., et al., *Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry.* Anal Chem, 2012. **84**(22): p. 9668-73.

151. Brunner, A.M., et al., *Benchmarking multiple fragmentation methods on an orbitrap fusion for top-down phospho-proteoform characterization.* Anal Chem, 2015. **87**(8): p. 4152-8.

152. Riley, N.M., et al., *Enhanced Dissociation of Intact Proteins with High Capacity Electron Transfer Dissociation.* J Am Soc Mass Spectrom, 2015.

153. Tsiatsiani, L. and A.J. Heck, *Proteomics beyond trypsin.* FEBS J, 2015. **282**(14): p. 2612-26.

154. Omenn, G.S., et al., *Metrics for the Human Proteome Project 2015: Progress on the Human Proteome and Guidelines for High-Confidence Protein Identification.* J Proteome Res, 2015. **14**(9): p. 3452-60.

155. Hebert, A.S., et al., *The one hour yeast proteome.* Mol Cell Proteomics, 2014. **13**(1): p. 339-47.

156. Richards, A.L., et al., *One-hour proteome analysis in yeast.* Nat Protoc, 2015. **10**(5): p. 701-14.

157. Sidoli, S., et al., *Middle-down hybrid chromatography/tandem mass spectrometry workflow for characterization of combinatorial post-translational modifications in histones.* Proteomics, 2014. **14**(19): p. 2200-11.

Acknowledgments

By now I have read and frantically scrolled through many, many different thesis of my colleagues, friends, even strangers, and I couldn't help not reading the 'thank you' part ☺. Most of these started this section with similar sentences, and it made me wonder; is there a right way to do this, a right order to follow? Really, how many ways there are to open a chapter of acknowledgments? –Probably as many as you can think of and I have decided to start mine with a simple, to the point and sincere; '*THANK YOU !*'

During the past 4 years, I had a privilege and a pleasure to meet, work and spend time with many wonderful people who made my life not only better, but also fun, and it's my turn to thank them for it.

So first and foremost, not only that I had a great supervisor, I had two of them! I would like to thank Dr. Yury O. Tsybin, for all of his guidance and support, always pushing us forward. I am thankful for having a supervisor who was always there when questions needed answers, when the day started just wrong and all results were bad, always ready to lend a hand, but mostly I am thankful for his patience when my concerns needed to be resolved and silenced. Your friendship, support, and fruitful discussion have been indispensable, *БОЛЬШОЕ СПАСИБО.*

Dr. Martin Kussmann, for all of his support in crucial and time-sensitive situations, for opening his door and sharing a different, exciting R&D from industrial prospective. I greatly appreciate your advices and insight We have always practiced to start emails in Italian, but this time I wish to tell you *vielen Dank* for being there.

A special *merci* I owe to Ms. Christine Kupper, the person who always has a solution in her pocket for any bureaucratic and administrative task that you have no clue about. Thank you for every time I stepped into your office with a silly question that you resolved, for every constructive advice you shared.

Every lab should have a Christine, but we were lucky enough to have THE Christine.

For all their support, help and fun times, I would like to say big thank you to my ‘LSMB family’:

To Anton, for being a man of few words but of great deeds. I wish you all the best in your career, I have no doubt you will do amazing things.

To Matt, for interesting and fun discussions, for the best English in the lab and all the good food you suggested during ASMS; Juicy KLucy and Cinnabons of course ☺

To Konstantin, always cheerful and optimistic, hard-working and humble. You are our *‘Jack of all trades’* (the closest translation of Croatian: *‘Katica za sve’*) in the best possible way. *СПАСИБО.*

To Daniel, who taught me all I know about the IgGs. I will never forget ‘in the pines’ and ‘George, pass me the …. chopper’ which still makes me laugh to tears ☺.

To Üni, only girl in ‘all boys club’, who taught me how to deal with the LC and Orbi and NOT pulling your hair out. Thank you for being my friend, my sister in arms and my complaining buddy. I will never stop laughing to ‘Ding-dong’ and I will never again do ‘Insanity’ ☺

To Kostya, my comrade, goes a big thank you for countless coffees in l’Arcadie and the ‘nerd-alert’ portion of the day! It helped me survive my last year in Lausanne.

Finally, to Luca, I owe a thank you that no words can express. You’ve been my colleague, my third co-advisor, my most severe critic, my family and above all my true friend I could always come to and count on, who always had my back. Over the ocean or right next door, *‘Siamo due legati dentro con una amicizia che ci dà una profonda convinzione che nessuno ci dividerà’.*

My next *merci beaucoup* goes to Dr. Laure Menin. I greatly appreciate all your help and advices with instrument maintenance, and I hope we will have again a chance to clean S-lenses and tune ETD together!

Much of the work in this dissertation would not have been possible without the collaborative efforts of many people. Many thanks to Dr. Manfredo Quadroni and Dr. Patrice Waridel for their continuous support, lending us consumables and their instruments. It has been a pleasure working with you. For promptly fixing our instruments many thanks are owed to Jerome Belec, Damien Viallon, Alain Siegrist and Myriam Demant from Thermo Scientific. Thanks to Dr. Michel Monod for his magic fridge and Sap9 protease which started off my PhD project. I extend my gratitude to Ms. Gladys Pache and all super nice staff of BCH magasin for being always so kind. To Ms. Anne Lene Odegaard, the best ever administrative assistant of doctoral school of chemistry.

A special thanks to my favorite Russian girls Alexandra and Natalia (AKA Topolona and my Russian Grace Kelly), my first friends in Lausanne, with whom I enjoyed many escapes from a PhD routine all over Europe. To Milena- my 'Swiss mum' who has put a cast over my broken leg, and has been my dear friend ever since. To Angelica, mia stella gemella, grazie di cuore. To my Balkan gang- especially Maja & Petar and Sanja & Mićo, who made my last year here fun and close to home.

My 'sisters', Sandra, Adrijana and Ana have been my friends, fans and my pillars for as long as I can remember. I am blessed having you in my life, and I am forever grateful to have met you.

To Davor, thank you for being you, and for always being there for me. I would not be who I am today without you.

Finally, the biggest thank you of all is reserved for my family. I will forever remember, through all adventures of my life, through every moment of every day, that you always walked beside me, hoping for me, loving me, cheering me on.

Curriculum Vitae

Kristina Srzentić

Address	Chemin des cedres 1, 1004 Lausanne, Switzerland
Phone number	+41 78 65 44 661
Date of birth	24th November 1986
Nationality	Croatian
Email address	ksrzentic@gmail.com
Skype	Kristina Srzentić *(Switzerland)*

Education and training

Philosophy Doctorate, Chemistry (May 2012 – Jan 2016)

- Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
- Doctoral Program in Chemistry and Chemical Engineering (EDCH)
    - Thesis title: “Middle-down approaches for mass spectrometry-based protein identification and characterization”
- Supervisors: Dr. Yury O. Tsybin
  Prof. Martin Kussmann

Visiting Scholar (Jun 2010-Mar 2012)

- Rudjer Boskovic Institute, Zagreb, Croatia
- Host: Prof. Mario Cindrić
- Project: ‘Prostate cancer biomarkers’ – National funds grant, Ministry of Science, Education and Sports of Republic of Croatia 098-0000000

Project Assistant (June 2014)

- Northwestern University, Evanston, IL, USA
- Host: Prof. Neil L. Kelleher

Master of science, Chemical Engineering and Technology (Oct 2009 – Sep 2012)

- University of Zagreb, Croatia
- Rudjer Boskovic Institute, Division of Molecular Medicine, Centre for proteomics and Mass Spectrometry, Zagreb, Croatia
- Analitical chemistry, organic chemistry, biochemistry
- Major in “Applied Chemistry”
- Thesis title: “Application of Derivatization Methods in Prostate Cancer Biomarker Discovery”

- Supervisors: Dr. Mario Cindrić
  Prof. Tomislav Bolanča
- Final mark: 4,45/5

<u>Bachelor of Science, Biotechnology</u> (Sep 2005 – Sep 2009)

- University of Zagreb, Croatia
- Mathematics, physics, engineering, biochemistry, physical chemistry, organic chemistry, analytical chemistry
- Thesis title: "Chemometrics in the Development of Analytical Methods"
- Supervisor: Prof. Tomislav Bolanča
- Final mark: 4,4/5

Publications and awards

Publications

- **Srzentić K.,** Fornelli L., Laskay Ü. A., Monod M., Beck A., Ayoub D., Tsybin Y.O. Advantages of extended bottom-up proteomics using Sap9 for analysis of monoclonal antibodies. *Anal. Chem* (2014), 86 (19), pp 9945–9953
- Laskay Ü. A., **Srzentić K.**, Tsybin Y.O. Extended Bottom-Up Proteomics with Secreted Aspartic Protease Sap9. *J. Proteomics* (2014), 110, pp 20-31
- Laskay Ü. A., Lobas A. A., **Srzentić K.**, Gorshkov M. V., Tsybin Y.O. Proteome digestion specificity analysis for rational design of extended bottom-up and middle-down proteomics experiments. *J. Proteome Res.* (2013), 12 (12), p 5558–5569
- Laskay Ü. A., **Srzentić K.**, Fornelli L., Upir O., Kozhinov A. N., Monod M., Tsybin Y. O. Practical Considerations for Improving the Productivity of Mass Spectrometry-based Proteomics. *Chimia* (2013), 67 (4) p 1–6
- **Srzentić K.**, Zhurov K. O., Tsybin Y. O. Optimization of chemical hydrolysis for protein identification by mass-spectrometry based proteomics. *Manuscript in preparation*
- **Srzentić K.**, Gasilova N., Kussmann M. and Tsybin Y.O. Chemoselective digestion for middle-down proteomics and structural analysis of IgGs. *Manuscript in preparation*
  - **Srzentić K.**, Nagornov K, Tsybin Y. O. Revealing chain connectivity in monoclonal IgG1 using GingisKHAN proteolysis and top-down electron transfer dissociation Orbitrap FTMS

- **Srzentić K.***, Schumann K.*, Tsybin Y.O. and Shevchenko A. Quantification of the membrane lipidome turnover by metabolic 15N labeling and ultra-high resolution Orbitrap FTMS. *Manuscript in preparation*

2. Gasilova N., **Srzentić K.**, Qiao L., Tsybin Y. O., Girault H. H. On-chip mesoporous functionalized magnetic microspheres for extended bottom-up proteomics, *in print*

Awards

- Dean's award for outstanding student scientific work 2010, University of Zagreb
- SCS/SCNAT Travel award 2014
- ASMS Travel award 2014
- SGMS Travel award 2015

Oral presentations

- Fall meeting of the Swiss Chemical Society, 2013, Lausanne, Switzerland
- IX Mass Spectrometry in Biotechnology and Medicine (MSBM), 2015, Dubrovnik, Croatia
- 33rd meeting of the Swiss Group for Mass Spectrometry, Beatenberg, Switzerland, October 29-30 2015

Poster presentations

- Mass Spectrometry in Biotechnology and Medicine (MSBM) Summer School, 2012, Dubrovnik, Croatia
- American Society for Mass Spectrometry (ASMS) Conference, 2013, Minneapolis, MN, USA
- American Society for Mass Spectrometry (ASMS) Conference, 2014, Baltimore, MD, USA
- International Mass Spectrometry Conference, (IMSC), 2014, Geneva, Switzerland
- American Society for Mass Spectrometry (ASMS) Conference, 2015,St. Louis, MO, USA

Technical skills

- Work on: FT-based mass spectrometer, Orbitrap Elite, Orbitrap Fusion, basic use of Q Exactive (Thermo Scientific); time-of-flight, MALDI TOF/TOF (ABSciex); basic use of QTOF (Agilent); bottom-up and middle down LC-MS/MS analysis
- Protein purification and/or separation techniques: ion chromatography (IC), liquid chromatography (HPLC); SDS-PAGE (1D and 2D), Western blot; dialysis, solid-phase extraction (SPE), GELFrEE, Off-Gel (Agilent), cell culture, familiar with PCR and rt-PCR

- Sample preparation for bottom-up and middle-down proteomics: cell lysates from human cancer tissue, protein enzymatic digestion (trypsin, Lys-C, *etc.*), protein chemical digestion with toxic chemicals (NTCB, CNBr, BNPS-Skatole), Nanodrop concentration assays (BCA, Bradford, Lowry), fluorescence-based activity assay
- Immunoglobulin G purification and digestion with different enzymes
- Maintenance of FT-ICR (cryogens refill, cleaning optics), Orbitrap Elite (cleaning optics, vacuum system, ETD ion volume, *etc.*), nano-LC system

Personal skills

- Fluent in English and Italian, basic knowledge of French
- Ability to work in an international research team and in a multidisciplinary environment