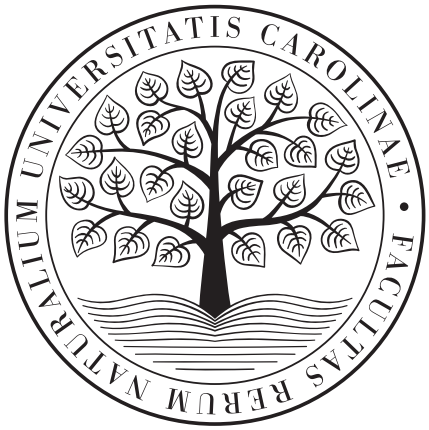
Charles University in Prague
Faculty of Science

**BACHELOR THESIS**




Evžen Wybitul

# Differential discovery of protein features using tandem mass spectrometer

Department of Bruteforcing Hard Problems

Supervisor of the thesis: Miroslav Kratochvíl
Study programme: Bioinformatics
Study branch: Bioinformatics

Prague 2021

I declare that I carried out this bachelor thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date .....  .....

Author's signature

Dedication. It is nice to say thanks to supervisors, friends, family, book authors and food providers.

Title: Differential discovery of protein features using tandem mass spectrometer

Author: Evžen Wybitul

Department: Department of Bruteforcing Hard Problems

Supervisor: Miroslav Kratochvíl, Noxemchâteau Apartment

Abstract: Abstracts are an abstract form of art. Use the most precise, shortest sentences that state what problem the thesis addresses, how it is approached, pinpoint the exact result achieved, and describe the applications and significance of the results. Highlight anything novel that was discovered or improved by the thesis. Maximum length is 200 words, but try to fit into 120. Abstracts are often used for deciding if a reviewer will be suitable for the thesis; a well-written abstract thus increases the probability of getting a reviewer who will like the thesis.

Keywords: key words

# Contents

# Introduction

1. Existují proteiny, jsou klíčové pro funkci organismu, obstarávají většinu procesů v něm. Funkce proteinu je závislá na jejich struktuře, a ta je závislá mimo jiné i na nekovalentních interakcích jednotlivých aminokyselin. Tyto interakce probíhají přes vodíkové nebo disulfidické můstky.

2. Vědět, kde tyto můstky jsou, může pomoci molecular dynamic simulations pro omezení vyhledávacího prostoru, propř. určitě i jiným věcem.

3. Metod určování pozic SS můstků existuje spousta. Jedna z nich využívá tandemovou hmotnostní spektrometrii v kombinaci s kapalinovou chromatografií. To vše na částečně alkylovaném proteinu, který je rozložený trypsinem.

4. V této práci jsme zvolili podobný postup (to jest LC-MSMS na tryptických peptidech), ale přidali jsme k němu in-silico matchování (di)peptidů na naměřená spektra pomocí novel divide and conquer metody. Tato metoda využívá toho, že dipeptidy mají specifický fragmentační pattern a navíc mají i jinou prekurzorovou hmotu.

5. Tuto metodu ověřujeme na několika naměřených proteinech, a máme svkělé výsledky (hopefully).

# Chapter 1

# Tandem mass spectrometry for protein analysis

Proteins are amino acid biopolymers that take part in most natural processes in living organisms. Among other things, they are vital for cell growth, reproduction, metabolism, and movement. [citace] Proteins are also a frequent target of medicine, because they play a key role in most diseases [citace nějakého proteinového léku].

Protein function is highly dependent on its 3D structure [citace], and as was shown by Anfinsen [citace], the information about the structure is in turn encoded in the sequence of the protein.

Protein folding is driven by natural biophysical forces which makes it hard to properly recreate *in silico*, especially when there is no homologous protein with known structure [citace]; this problem is called *de novo* folding.

Techniques for *de novo* folding rely mostly on molecular dynamic simulations. These approaches are often very performance-intensive (?), because they are effectively optimising a complicated scoring function a huge multidimensional problem space. [citace] Any information we have about the structure can be thus very heplful in reducing the problem space when supplied to the algorithm, making the computations faster and more accurate. One type of such information are the positions of disulphide bridges.

Disulphide bridges (DB) in proteins can be formed between the sulfhydryl groups of two cysteines during a thiol–disulfide exchange reaction catalyzed by thioredoxin [citace] In vivo they are oftentimes essential to correct protein folding, because they stabilise the final structure [citace] The knowledge of DB positions can be used, among other things, to constrain the molecular dynamic simulations, as mentioned earlier. In addition, the knowledge of which cysteines do *not* partake in a DB is important, too.

Non-interlinked cysteines have an important pH-regulating function within

proteins [citace] (a něco dalšího ještě?). Consequently, cysteines are scarce compared to other amino acids [citace], but they are usually very well conserved during evolution [citace].

There are many methods aiming to determine the positions of DBs, on of them is tandem mass spectrometry combined with liquid chromatography (LC-MSMS). LC-MSMS is a popular general analysis technique, often used in proteomics for its accuracy and relative straightforwardness of the experiments (?) [citace].

In LC-MSMS, the protein is eventually fragmented to smaller charged peptides whose mass to charge ratio ($m/z$) is measured with atomic precision. The whole experiment can be designed in a way that the DBs are preserved which results in the occurence of *bridgetides* with specific $m/z$ fragmentation signatures. Computational analysis can help us discover these fragmentation spectra and determine the original positions of the DBs in the protein.

## 1.1 Sample preparation

To prepare the protein for the analasis, it needs to be proteolyticaly cleaved; trypsin, and, to a lesser extent, pepsin, are popular choices. Trypsin is a serine protease with very high specificity which makes it very useful for mass spectrometry analyses, because the resulting peptides are predictable.

Trypsin cleaves amino acid chains at the carboxyllic side of lysine and arginine, provided they are not followed by proline [1]. Lysine and arginine are both relatively abundant in most proteins which makes the tryptic digestion peptides — or as we will call them, *tryptides* — reasonably sized for a mass spectrometry analysis [2]. With that being said, the sample protein is not cleaved at every potential cleavage place; so called *missed cleavages* do occur, and their frequency and position depend on neigbouring residues [3], and experimental setup.

After digestion, the resulting peptides undergo separation in liquid chromatography.

## 1.2 Sample separation

In a general proteomic experiment, the signal from more abundant sample proteins may interfere with the other, less frequent proteins. To sidestep this problem, it has become routine to perform separation before the main MS experiment, separating the sample either on the protein level or the peptide level.

One possible method for protein-level separation is two dimensional polyacrylamide gel electrophoresis, during which the proteins are split first by iso-

electric focusing, and then by SDS gel electrophoresis [4]. 2D-PAGE has very high resolution [5]. The proteins are usually digested in-gel after the separation, manually cut out, and then put into the mass spectrometer, causing the method to have relatively low throughput [6], making it unfit for some scenarios.

In our scenario with one protein per sample, separating on protein-level is not going to be useful; instead, peptide-level separation is preferred. A popular peptide-level separation method is liquid chromatography (LC). In a model MS-based proteomic LC experiment, the proteins are digested without prior separation, and the resulting peptides are separated on reverse-phase liquid chromatography column that is directly connected to a tandem mass spectrometer [7]. Usually the number of different proteins in the sample is high, leading to a large amount of generated spectra and causing a need for automatic processing. This type of identifying sample proteins is sometimes called shotgun processing.

Reverse-phase LC has two main constituents: a mobile liquid phase containing the peptides and a stationary solid phase which is usually a nonpolar column with $C_{18}$ alkyl chains [8]. The mobile phase passes along the stationary phase, the elution time of each individual peptide depending on its hydrophobic interactions with the alkyls. The peptides are eluted with a polar mixture of water and organic solvent, such as acetonitrile [9], the shortest and least hydrophilic eluting the earliest.

## 1.3 Tandem mass spectrometry

Mass spectrometry is an analytical technique with roots deep in the last century that has originally been used for studying small thermostable molecules. However, with the advancements in soft ionization allowing proteins and other biomolecules to be analysed as well [10], mass spectrometry has become an indispesable tool in proteomics research [11].

In the context of proteomics, mass spectrometry experiments can be either single-stage or tandem. During single-stage experiments, the mass distribution of a polypeptide sample is determined. The more frequent of the two, tandem (MS/MS) mass spectrometry is used to learn about certain structural features of a protein, including sequence and post-translational modifications.

Both the single-stage and MS/MS experiments begin similarly: the sample peptides are ionized, the ions travel through an electromagnetic field in an analyser and into a detector, whilst their mass-to-charge ($m/z$) is being calculated [13]. In single-stage mass spectrometry, the experiment ends there, while in MS/MS, some of these *precursors* are selected to undergo fragmentation in the collision cell, as shown in Figure 1.1. The resulting fragments are also analysed and their $m/z$ values noted; the output of the MS/MS experiment are the precur-

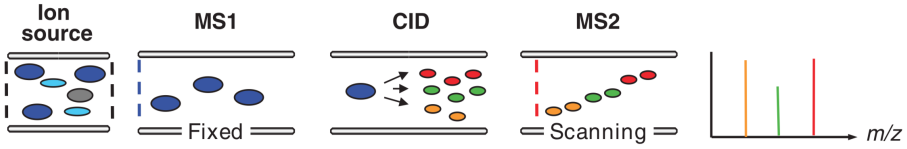



**Figure 1.1** An ordinary MS/MS workflow diagram. While the specific intrumentation details differ from spectrometer to spectrometer, the general structure of ionize ⟶ analyse ⟶ fragment ⟶ analyse is common to all of the MS/MS spectrometry experiments. Image taken from [12].

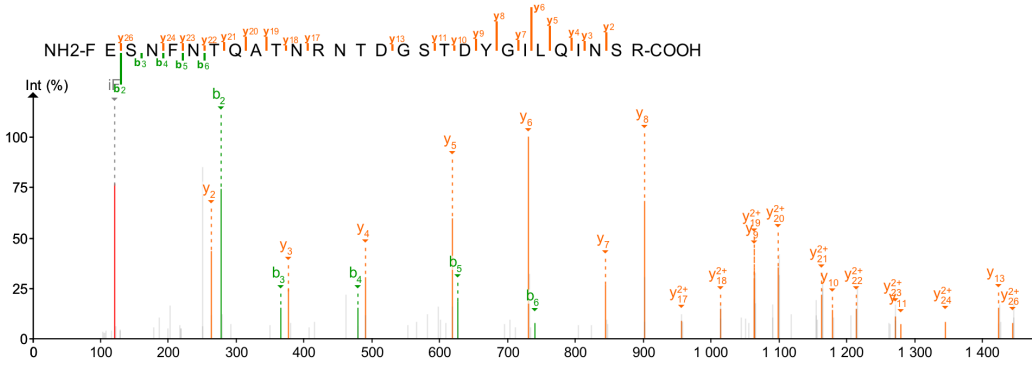



**Figure 1.2** An annotated fragmentation spectrum of the precursor *FESNFNTQATNRNTDGSTDYGILQINSR*

sor masses and their fragmentation spectra, an example of which can be seen one figure Figure 1.2.

We will now discuss some specific approaches to the main phases of MS/MS analysis, putting the focus on those that are relevant for this thesis.

Nás zajímá hlavně vysoký dynamický rozsaha vysoká přesnost. Rozsah je při identifikaci DBs užitečný, jednak jsou to obecně informace navíc, a taky v nízkých hmotnostech bývají krátké interní fragmenty, které by mohly něco o můstcích prozradit. Má také vysokou přesnot, až pod chybovost 1ppm, což se hodí, protože přiřazování DBs je intrinsically problém s kombinatorickým výbuchem a velkou pravděpodobstí přiřazení false positive tím více, čím větší bude tolerance chyby. pomáhá to tedy snížit počet false positives.

### 1.3.1 Sample ionization

Save a few specific exceptions, only charged compounds are detectable by the analyser and detector in mass spectrometer; that means we have to ionize our sample in order be able to analyse it.

There are many sample ionization methods; one of the oldest is electron ionization [14], in which the sample is first transferred to a gas phase and then bombarded with electrons. However, this method is unsuitable for large thermally unstable organic molecules, such as peptides; for proteomics work, the two most popular options are MALDI and ESI.

Matrix-assisted laser desorption/ionization (MALDI) is a ionization technique oft used in proteomics [15, 16]. In MALDI, the sample is placed on a solid light-absorbing crystalline matrix and undergoes several short focused bursts of laser light with specific wavelenghts. The light is absorbed by the sample layer which causes sample evaporation and ionization [17]. Unfortunately for our use case, the whole ionization process has to be done in a vacuum, making it impossible to directly connect the liquid chromatography column to the spectrometer.

### Electrospray ionization

For proteomics experiments that make use of liquid chromatography, electrospray ionization (ESI) is the ionization method of choice. As ESI works under atmospheric pressure, the LC colon can be connected directly to the mass spectrometer, resulting in what is usually called an "online" or "hyphenated" LC-MS system [18].

During ESI, a very fine capillary with a solution containing the sample peptides and charged ions is placed into a strong electrostatic field. Due to the influence of the field, the solution forcibly squirts out of the capillary, creating a mist of miniscule charged droplets. The solution slowly evaporates from the droplets, until eventually the repulsive electric forces inside the droplet overcome its surface tension and the droplet splits into yet smaller droplets [19]. This evaporating and splitting process repeats itself, until we are left with isolated sample ions in the gas phase [20, 21, 10, 22].

For our work, two properties of ESI are important. First, ESI is a notably soft ionization technique, owing among other things to the fact it works in atmospheric pressure, which means that the sample undergoes very little to no fragmentation during the ionization [23]. That means that the tryptides traveling to the analyser will be mostly left intact, simplyfying the subsequent analysis. The second property has to do with the typical charge of ions produced by ESI. Ions generated by ESI are often multiply charged [24], bringing their $m/z$ value down and enabling us to analyse peptides with a higher mass in an ordinary mass spectrometer setting.

### 1.3.2 Mass analysers

A mass analyser, together with a detector, measures the $m/z$ ratio of a sample compound. The many existing mass analysers differ in their performance standards, the principle of function, and the sample characterstics they require to function properly.

One of the oldest mass analysers still in use is the time-of-flight (TOF) analyser [25]. It is also one of the simplest to manufacture. In TOF analysers, sample ions are accelerated with an electric field to make them travel along a path with known length. The ions with lower $m/z$ values will arrive sooner than the ones with higher $m/z$ values, as long as all of them are dispersed at a similar-enough point in time. Due to this requirement, TOF analysers are best suited for pulsed ionization techniques such as MALDI. In addition to having a relatively simple construction, TOF analysers have an excellent sensitivity and, at least in theory, their $m/z$ range is unlimited [26].

The linear quadrupole doubles as an analyser and also as a collision cell. As the name suggest, a linear quadrupole consist of four linear rods which are placed parallel to each other and arranged in a square shape, see Figure 1.3. A pair of rods sitting in diagonally opposite corners has the same polarity. However, the pairs periodically switch the polarity. An ion travelling along the rods is periodically repelled and attracted to each of the rods, its precise trajectory depending on its $m/z$ value [27]. In this way, ions with specific $m/z$ values can pass through the quadrupole into a detector [28], while others follow an unstable trajectory and crash into one of the poles or the wall of the quadrupole.

Quadrupoles can also trap specific ions inside for prolonged period of time instead of making them simply pass through. So called linear ion traps are sometimes used as a "staging area" for other analysers, trapping ions and releasing them by clusters based on their $m/z$ values further into the pipeline [29]. Another possibility is to use quadrupoles as collision cells for precursor fragmentation. For a long time, the state of the art in tandem mass spectrometry was the tripple quadrupole spectrometer [30]; it has only lately become dethroned on the basis of accuracy by methods based on Fourier transform.

#### Mass spectrometry based on Fourier transform

The basis of the older of the two Fourier transform based methods, Fourier transform ion cyclotron resonance (FT-ICR), has been cocieved in 1930s by research on ion cyclotron resonance. As Lawrence and Livingston [32] have shown, an ion particle in a magnetic and an electric field can be accelerated by periodically alternating the polarity of the surrounding electric field, and this in turn increases the radius upon which the particle circulates around the center of the chamber.

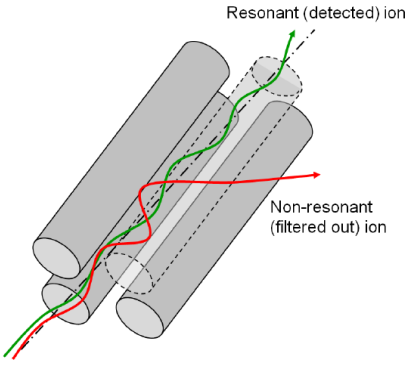



**Figure 1.3** A quadrupole with two highlighted classes of ion trajectories. Thanks to its $m/z$ value, the ion with green trajectory passes through the quadrupole and is ultimately detected, while the one with the red trajectory is filtered out. Image taken from [31].

Once the radius reaches a limit size, the particle can be detected crashing to the wall of the chamber. Later, the $m/z$ values of the ions became measurable even without them crashing into the detector, thanks to Fourier transform that made it possible to decode the signals of passing circulating ions and calculate the $m/z$ values from the frequencies and amplitudes [33]. This also made the measurement faster, as many ions with wildly different $m/z$ values could be measured in parallel. Further improvements increased the mass accuracy and resolution beyond what is attainable by quadrupole analysers [34, 35].

For our work, the most important analyser type is the Orbitrap [36]. It achieves similar accuracy, resolving power and dynamic range to FT-ICR, but does not require an expensive-to-run supraconducting magnet to do so. In orbitrap the ions simultaneously cycle around the centre and oscillate along the z-axis, as is illustrated on Figure 1.4. This oscillation induces a periodically changing electrical current in the detector that is converted to a $m/z$ spectrum of the analyte with the help of FT.

Orbitrap má vysoký dynamický rozsah a chybovost i pod 1ppm, lze jej napojit na ESI (a potažmo na LC-MSMS) je tedy pro naše účely ideální.

### 1.3.3 Precursor fragmentation

In tandem mass spectrometry, once the mass spectrum of the initial sample is analysed (MS1), the *precursors* are selected according to their mass and fragmented, and the fragments are undergo yet another mass analysis (MS2). Again, there are many fragmentation techniques, each useful for a different type of analysis.

When aiming to observe post-translational modifications and to preserve

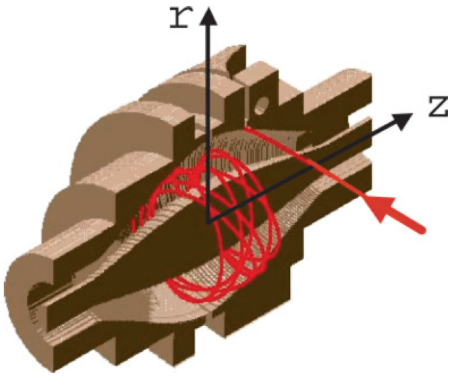



**Figure 1.4** An orbitrap mass analyser with a typical ion trajetrory highlighted. The ion circulates around the center while simultanously oscillating along the z-axis. Image taken from [36].

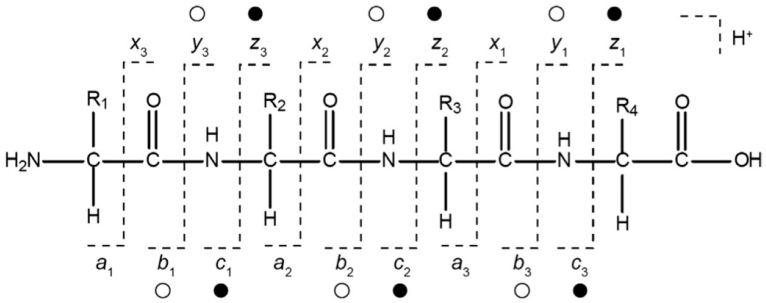



**Figure 1.5** A singly positively charged peptide with annotated fragmentation types. Signature CID b/y ion fragments are marked with open circles, while the typical ECD and ETD c/z ion fragments are marked with filled circles. Image taken from [39].

the volatile bond connecting the PTM to the peptide, electron-capture dissociation (ECD) [37] or electron-transfer dissociation (ETM) [38] are preferred. In ECD a multiply positively charged precursor ion is hit by a beam of low-energy electrons, while in ETD the electron transfer is induced by negatively charged reagent ions, both of these ultimately leading to the creation of a radical cation and amine backbone bond cleavage, resulting in the creation of *c* and *z* ions, as illustrated on Figure 1.5.

The fragmentation method we focus on in this work, however, is collision-induced dissociation (CID). It has a different fragmentation signature compared to the abovementioned methods (see Figure 1.5), and it doesn't preserve PTMs nearly as well as they do. Thankfully, DBs are not as labile as the bonds connecting PTMs to the peptide, and thus CID can be safely used to produce fragments

from bridgetides [citace]. A similar fragmentation signature to CID can also be obtained by infrared multiphoton dissociation (IRMPD) [40]. Because IRMPD, and the related UV-MPD, do not require collision gasses to be present for the fragmentaion, they are well suited for analysers operating under high vacuum, such as FT-ICR.

### Dissociation based on collision with neutral gas

Two common fragmentation methods fall under the umbrella of fragmenting by collision with neutral gas: collision-induced dissociation (CID) and higher-energy C-trap dissociation (HCD). Both of them make the accelerated precursor ions collide with neutral gas molecules, ultimately leading to its fragmentation, but use different instrumentation to reach this goal, and have different performance characteristics in different contexts.

The principle of function is not the only thing these two methods have in common; they also share a big portion of the fragmentation signature [41]. In both CID and HCD the dissociation process usually takes place at the more labile bonds, such as the ones connecting PTMs [42], or peptide bonds in the precrusor backbone, resulting in the generation of $b$ and $y$ fragment ions (see Figure 1.5).[1] As a side-effect of the dissociation, a small neutral molecule sometimes breaks off of the fragment, lowering its total mass value without affecting its charge. This dissociated molecule is termed the *neutral loss*; during CID and HCD, the most common neutral losses are water and amonia from the fragment N– and C–termini, and various other small molecules from specific amino acid side-chains.

The similarities do not end there: HCD and a specific subset of CID, a so called beam-type CID, also share a method of inducing the collisions. Precursor ions travel in a beam through a collision cell, and collide with the gas molecules along the way [43]. Because of this passage through the cell, the precursor ions are sometimes dissociated more the once, resulting in the generation and detection of *internal ions*. Many of the internal ions begin with a proline [41], revealing that the double cleavage event prefers some amino acids to others.

The beam-type CID is often connected to a quadrupole analyser (being a quadrupole itself in a 3-quadrupole mass spectrometer), however, which can sometimes make it hard to interpret these spectra due to its relatively low accuracy. As shown by Michalski et al. [41], ion trap CID, unlike the beam-type variant, doesn't lead to the creation of internal ions. Furthermore, it has limits regarding the containment of molecules below a certain mass threshold, leading to a mass cutoff [44].

[1] The fact that many PTM bonds are preferentially dissociated during CID is usually seen as undersirable. However, in our case it simplifies the analysis, as we can safely ignore PTMs and reduce the combinatorial complexity.

In the year 2007, the HCD dissiociation technique has been introduced [45], combining the richer sequence information [43] and lower mass cutoff of beam-type CID with the superior resultion capacity and accuracy of the orbitrap analyser, that was reported to be in the sub 1 ppm levels by the original paper. Data we use in this thesis are generated with HCD, because the high accuracy of its MS2 spectra makes it easier to filter out false positives that occur naturally due to the combinatoric nature of the problem, and furthermore the detected internal ions can be sometimes useful when determining the connectivity of complex DB configrations. Because the HCD fargmentation pathways are key to our work, we will discuss them in more detail below.

**Fragment types** The fragment types of a very pure sample protein were nicely summarized by Michalski et al. [41]. Ions of $b$ and $y$ type comprise most of the spectral intensity (54%, see Figure 1.7), together with $b$ ions with CO neutral loss that are interchangeable with $a$ ions. The ions themselves have different distributions, $y$ ions being the most abundant. Other neutral loss ions, be it a loss of water, amonia, or an amino-acid-specific small molecule[2], together with internal ions, can be attributed a quarter of the total fragment intensity. Immonium ions account for 6% of the intensity, totaling 85% intensity that can be explained with the current understanding of HCD fragmentation pathways. For a visual overview of the many different HCD fragmentation types, please refer to Figure 1.6.

**Charge** Coming from ESI-ionized precursors, fragments can be and indeed often are multiply charged [47, 41]; the only real limit of the fragment charge is the charge of its precursor. Of course, uncharged fragments are undetectable by the mass spectrometer. According the research on CID of crosslinked peptides by Giese, Fischer, and Rappsilber [48], most of the crosslined fragments were found to have a positive charge of at least 2, while an overwhelming majority of linear fragments had a positive charge of 1. Their results are illustrated on Figure 1.8.

The fragmentation pathways are complicated enough even without the presence of DBs, and the addition of peptide crosslinks complicates the matter further. We have to take into account the alkylation of non-bonded cysteines during the analaysis; the DB can be cleaved during the dissociation, resulting in fragments roughly resembling the fragmentation pathway of each connected peptide in the original precursor. If left intact, the crosslinks between precursors (or within one precursor) widen the possibilities of attainable mass values considerably.

[2] If specific amino acid neutral losses are of interest, the wonderful review by [46] lists many of them.

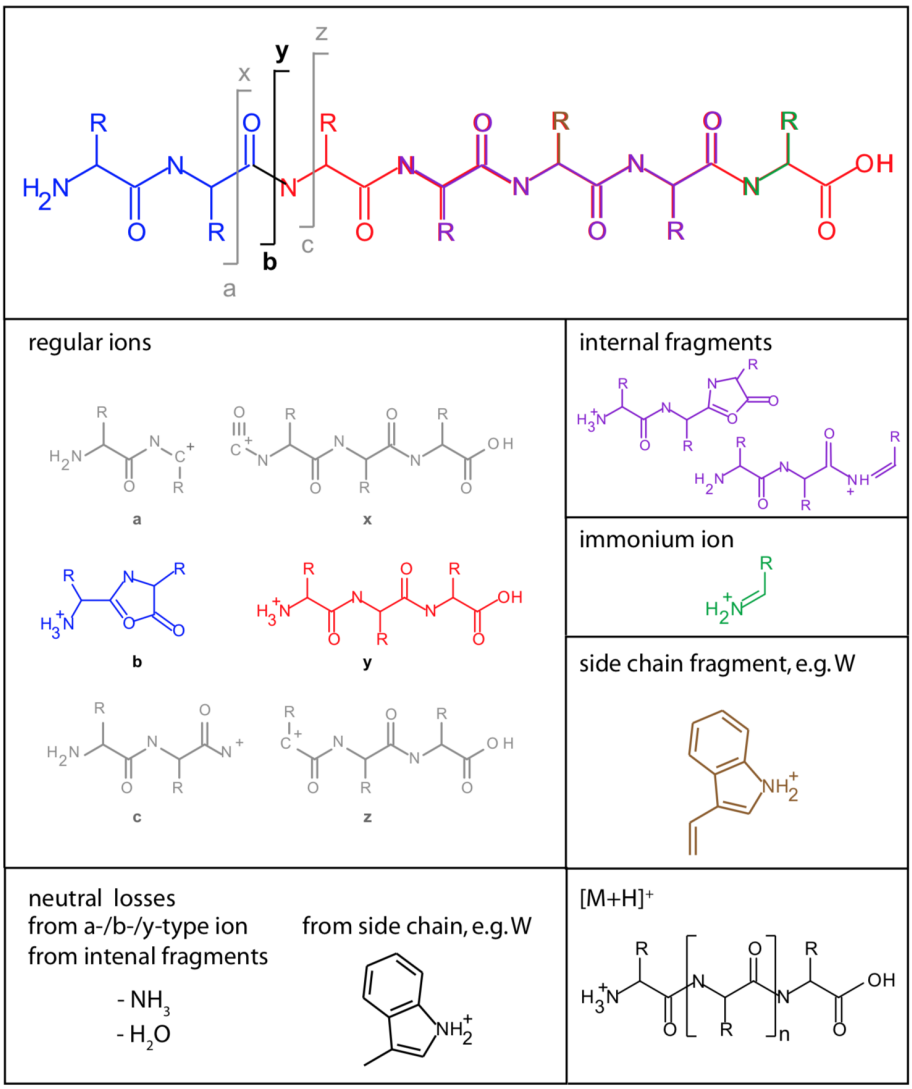



**Figure 1.6** During HCD, *b*, *y*, and to a lesser extent *a*, ions are the most common, together with internal ions and immonium ions, and their counterparts with neutral losses. Image taken from [41].

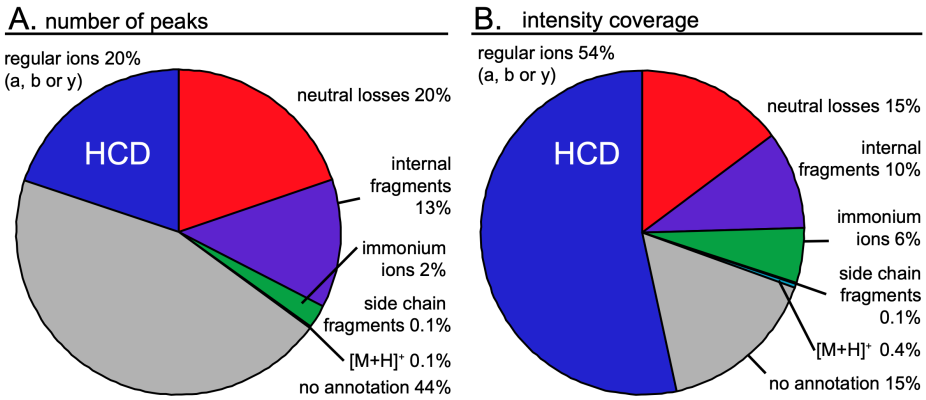



**Figure 1.7** (B) The regular *b*, *y*, and *a* ions take up 54% of the measured spectral intensity. Another 25% is explained by fragments with neutral-loss and internal fragments, and another 6% by immonium ions. Together, those four account for 85% of the measured intensity. It is true that almost a half of the peaks are still left unexplained (A), however, given all of these fragments have to split the remaining 15% of intensity, they are probably rather rare and are only of moderate importance. Image taken from [41].

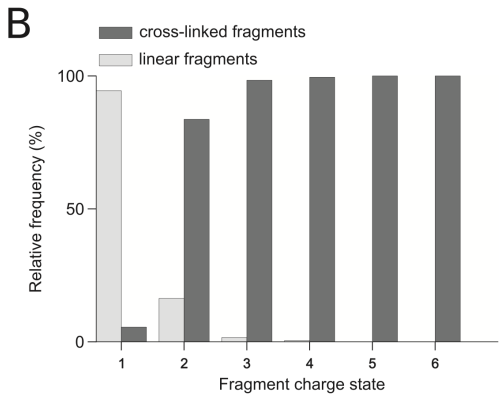



**Figure 1.8** Most linear fragments have charge 1, and most crosslinked fragments have a charge of 2 or higher. Image taken from [48].

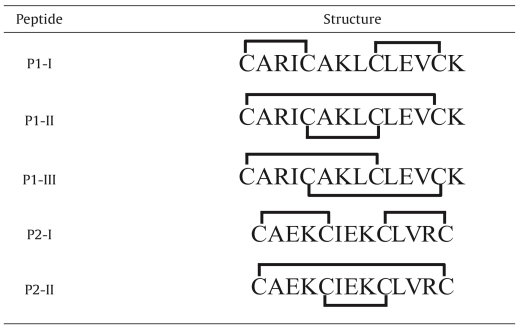



**Figure 1.9** An example of different possible configurations of intra-peptide DBs. Image taken from [55].

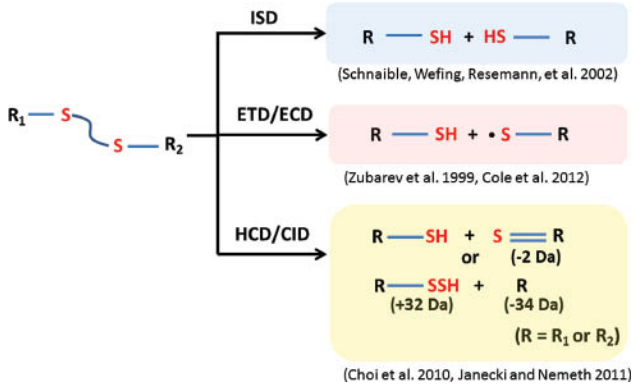



**Figure 1.10** Under different dissociation strategies, DBs manifest different cleavage characteristics. Under CID, the cleavage results into two possible assymetrical mass distributions. Image taken from [52].

**Disulphide bridges** Although DBs are not affected by low-energy CID as much as the other PTMs [46, 49], in high energy collision fragmentation, cleavage of the S-S bond can be observed with a higher probability [50]. The cleavage of the bond can result in the formation of an asymetrical distribution of masss on the two cysteines [51], that has been nicely ilustrated by Tsai, Chen, and Huang [52], see Figure 1.10 for details. The sole presence of a DB influences the fragmentation pathway of the whole peptide; Mormann et al. [53] reports a low but detectable signal of peptide backbone cleavages in the bonds instide S-S loop of an internal DB, while Clark et al. [54] reports that internal ions occur more frequently. The latter complicates the analysis noticeably as we have to take all combinations of cleavage positions from all interlinked peptides in the precuror, but also in theory allows us to differentiate between different configurations of intra-peptide bonds, such as those on Figure 1.9.

To recapitulate, HCD fragmentation pathways of non-crosslinked peptides

are relatively well-understood. However, the presence of DBs results in complex fragmentation patterns that are hard to analyze. The existing methods for DB identification thus usually involve a lot manual work, or do the bulk of the analysis in silico, but require the researcher to manually discard the many generated false positive matches afterwards. We briefly review some of these methods in the next section.

## 1.4 Bridgetide spectra interpretation

TODO je to komplikované , jak říká tsai2013mass

Výstupem MSMS je sada spekter, u kterých známe jejich prekurzorovou hmotu a prekurzorovu charge, a spektrum jejich fragmentů. Běžný task v bottom-up proteomice je podle těchto fragmentů zjistit, které peptidy z databáze se ve vzorku nacházely. U bridgetidů je to o něco složitější, protože vzniká širší škála fragmentů. na druhou stranu víme, ze kterého proteinu procházejí, což naopak analýzu zjednodušuje.

Přístupy lze rozdělit podle toho, jakou formu separace, ionizace, analýzy, fragmentace a interpretace používají. Interpretace bývá často automatická, někdy ale i manuální, například v tomto experimentu [citace manuální metody]. Dále se však budeme zabývat už jen automatickými přístupy, protože ty manuální trvají dlouho a vyžadují expertní čas.

### 1.4.1 Current approaches

1. Máme jeden protein, ale je to těžké.
2. Metody z Petřina mailu:
3. Metoda 1
4. Metoda 2
5. Metoda 3
6. Jaké mají slabé stránky, proč jim to funguje / nefunguje, co ještě nezkusili. (determinace přes prekurzory)

### 1.4.2 Problem statement and complexity

Pojďme tedy shrnout stávající poznatky pro naši konkrétní variantu problému. (co nás inpirovalo k tomuto přístupu?)

- trypsin, ESI, orbitrap, HCD, a modifikace typu met-oxidace a alkylace.

Teoreticky se ten problém dá převést na subset sum, takhle: ... To znamená, že je NP úplný se složitostí ..., můžeme si ho ale zjednodušit tím, že ho oconstrainujeme na základě biochemických poznatků, které jsme uvedli výše, hlavně v kapitole HCD.

Vzniká obrovská kombinatorická komplexita, kterou si lze zjedodušit tím, že skombinujeme bottom-up přístup s top-down — konkrétně nejdříve namatchujeme prekurzory, protože víme, z jakého proteinu pocházejí a ten task je sám o sobě zjednodušenou verzí tohoto tasku. Poté už nnematchujeme fragmenty z celého univerza/proteinu, ale pouze z několiak málo prekurzorů, které odpovídají konkrétnímu spektru.

# Chapter 2

# Methods

Následující detaily doplní Martin.

1. Trypsin máme z nějaké firmy, lysozym a BSA taky.
2. Používáme orbitrapový analyzátor s přeností 10–15ppm. Pro filtraci a CID používáme neonový quadrupól.
3. Naše metoda stojí na ručně psaném mutuálně rekurzivním biodegradabilním vegan-friendly algoritmu.

## 2.1 Algorithm

1. Algo má dvě části, obě fungují na principu divide and conquer a řeší problém podobný subset sum, ale s více sekvencemi.
2. Je napsaný v Pythonu, ale viz kapitola Discussion, není to ideální volba.

### 2.1.1 Part one, precursor masses

Tady výhledově bude podrobnější a formálnější popis subset sum, jeho řešení, a jak se můj problém a můj algo liší. Taky tady bude pseudokód.

1. Vygenerujeme možné tryptické peptidy.
2. Procházíme všechny prekurzorové hmoty a snažíme se z tryptických peptidů poskládat peptid s vhodnou hmotou.
3. Toto "skládání" bere v úvahu jak missed cleavages, tak možné propojení peptidů SS můstky. Také uvažuje to, že cysteiny, které nejsou v můstku, jsou modifikovány alkylací (+57). Také uvažuje to, že některé (0–všechny) Met mohou být oxidovány (+16).

4. Funguje to v podstatě jako divide and conquer rekurzivní algoritmus na subset sum.

5. Výstupem je pro každé spektrum seznam tryptických peptidů, které je možné spojit do peptidu s danou prekurzorovou hmotou. U tohoto seznamu je také určeno, kolik je v něm Cys můstků (tj vlastně kolik je v něm alkylovaných modifikovaných Cys) a kolik je v něm modifikovaných Met.

6. Běžně pro jedno spektrum získáme 0–3 takovéto peptidy.

7. (hisotogram počtu namatchovaných prekurzorů)

8. SLožitost algortmu je nějak šíleně exponenciální a ještě s velkými multiplikativními konstantami, jak tam je spousta těch maybe-modifikací, ale tím, že to je NP, tak se s tím holt musím smířit. Běží to zhruba minutu pro 13000 spekter a středně velký protein (LYS), což je pro praxi ok.

9. Modifikace jsou konfigurovatelné., stejně jako hmota alkylace, a počet dovolených interpeptidových můstků.

**Differences from subset sum**

1. V sekvenci hledáme $n$ souvislých podsekvencí (v subset sum je to jakoby až $n$, protože tam žádná souvislost není v podmínce). $n$ určuje, kolik interpeptide SS dovolíme — pokud chceme jen dipeptidy, tak $n = 1$, pokud i tripeptidy $n = 2$ atp.

2. Nedá se moc dělat dynamické programování, protože všechny mass jsou floaty. Sice nehledáme přesný sum a máme nějakou toleranci, její velikost ale závisí na současném stavu algoritmu (protože to je relativní chyba). Asi bychom to ale mohli ignorovat, všechny ty floaty vynásobit stem a jet normálně na přesný součet, viz diskuze.

3. Větvení je vícero, protože můžeme skočit na další sekvenci, nebo přibrat modifikace současného rezidua atp.

4. (obrázek větvení, srovnání subset sum a tohohle šílenství)

### 2.1.2 Part two, matching fragments

1. V zásadě to samé, jako v part 1, ale ještě o level kombinatoricky výbuchovatější.

2. V každém spektru procházíme všechny naměřené hmoty fragmentů a snažíme se k nim in-silico vygenerovat fragment z nějakého z proteinů, které jsme k danému spektru vybrali v první části.

3. Konkrétní řešení je zase rekurzivní a opět podobné subset sum — procházíme postupně peptid a rozhodujeme se, jestli uděláme break, nebo ne.

Problémy jsou následující.

1. V první části nevíme, kde přesně jsou můstky a modifikace, jen kolik jich je dohromady

2. Nevíme, jakou charge měl fragment, který danou stopu vygeneroval, musíme tedy vyzkoušet generovat fragmenty pro všechny charge (1 až charge prekurzoru, tu známe).

3. Vznikají nám neutral lossy, ale nemusí.

4. Mohou se breakovat SS, ale nemusí. Pokud se breaknou, existují čtyři možnosti, co tam po SS můstku zbylo.

5. Fragmenty mohou vznikat dvěma breaky — můžeme tedy mít kombinované b+y fragmenty.

6. Opět (prakticky) nemůžeme použít dynamické programování.

7. (obrázek nějakého složitého fragmentu?)

Níže bude opět pseudokód a popis algoritmu podobbný tomu v části 1.

## 2.2 Evaluation

1. (jak zpracováváme výstup algo?)

2. (srovnáváme ho s něčím? jak?)

3. (jaké je kritérium toho, že řekneme "tady to je podezřelé, mrkni, jeslti tam není můstek")

# Chapter 3

# Results and discussion

# Conclusion

In the conclusion, you should summarize what was achieved by the thesis. In a few paragraphs, try to answer the following:

- Was the problem stated in the introduction solved? (Ideally include a list of successfully achieved goals.)
- What is the quality of the result? Is the problem solved for good and the mankind does not need to ever think about it again, or just partially improved upon? (Is the incompleteness caused by overwhelming problem complexity that would be out of thesis scope, or any theoretical reasons, such as computational hardness?)
- Does the result have any practical applications that improve upon something realistic?
- Is there any good future development or research direction that could further improve the results of this thesis? (This is often summarized in a separate subsection called 'Future work'.)

This is quite common.

# Bibliography


[1] Jesper V Olsen, Shao-En Ong, and Matthias Mann. "Trypsin cleaves exclusively C-terminal to arginine and lysine residues". In: *Molecular & cellular proteomics* 3.6 (2004), pp. 608–614.

[2] Rune Matthiesen. "Introduction to Mass Spectrometry-Based Proteomics". In: *Mass spectrometry data analysis in proteomics*. Humana Press, 2020, pp. 12–12.

[3] Paul D Gershon. "Cleaved and missed sites for trypsin, Lys-C, and Lys-N can be predicted with high confidence on the basis of sequence context". In: *Journal of proteome research* 13.2 (2014), pp. 702–709.

[4] Patrick H O'Farrell. "High resolution two-dimensional electrophoresis of proteins." In: *Journal of biological chemistry* 250.10 (1975), pp. 4007–4021.

[5] Joachim Klose and Ursula Kobalz. "Two-dimensional electrophoresis of proteins: an updated protocol and implications for a functional analysis of the genome". In: *Electrophoresis* 16.1 (1995), pp. 1034–1059.

[6] Wayne F Patton, Birte Schulenberg, and Thomas H Steinberg. "Two-dimensional gel electrophoresis; better than a poke in the ICAT?" In: *Current opinion in biotechnology* 13.4 (2002), pp. 321–328.

[7] Michael P Washburn, Dirk Wolters, and John R Yates. "Large-scale analysis of the yeast proteome by multidimensional protein identification technology". In: *Nature biotechnology* 19.3 (2001), pp. 242–247.

[8] Shung Ho Chang, Karen M Gooding, and Fred E Regnier. "High-performance liquid chromatography of proteins". In: *Journal of Chromatography A* 125.1 (1976), pp. 103–114.

[9] T Fröhlich and GJ Arnold. "Proteome research based on modern liquid chromatography–tandem mass spectrometry: separation, identification and quantification". In: *Journal of Neural Transmission* 113.8 (2006), pp. 973–994.

[10] John B Fenn et al. "Electrospray ionization for mass spectrometry of large biomolecules". In: *Science* 246.4926 (1989), pp. 64–71.

[11] Francis S Collins, Michael Morgan, and Aristides Patrinos. “The Human Genome Project: lessons from large-scale biology”. In: *Science* 300.5617 (2003), pp. 286–290.

[12] Bruno Domon and Ruedi Aebersold. “Mass spectrometry and protein analysis”. In: *science* 312.5771 (2006), pp. 212–217.

[13] Jürgen H Gross. *Mass spectrometry: a textbook*. Springer Science & Business Media, 2006, pp. 4–5.

[14] Frank Henry Field and Joseph Louis Franklin. *Electron impact phenomena: and the properties of gaseous ions*. Vol. 1. Academic Press, 2013.

[15] Richard M Caprioli, Terry B Farmer, and Jocelyn Gile. “Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS”. In: *Analytical chemistry* 69.23 (1997), pp. 4751–4760.

[16] Philip L Ross, Katherine Lee, and Phillip Belgrader. “Discrimination of single-nucleotide polymorphisms in human DNA using peptide nucleic acid probes detected by MALDI-TOF mass spectrometry”. In: *Analytical Chemistry* 69.20 (1997), pp. 4197–4202.

[17] Michael Karas, Doris Bachmann, and Franz Hillenkamp. “Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules”. In: *Analytical chemistry* 57.14 (1985), pp. 2935–2939.

[18] Gregory J Opiteck et al. “Comprehensive on-line LC/LC/MS of proteins”. In: *Analytical chemistry* 69.8 (1997), pp. 1518–1524.

[19] Lord Rayleigh. “XX. On the equilibrium of liquid conducting masses charged with electricity”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 14.87 (1882), pp. 184–186.

[20] Malcolm Dole et al. “Molecular beams of macroions”. In: *The Journal of chemical physics* 49.5 (1968), pp. 2240–2249.

[21] Malcolm Dole et al. “Gas phase macroions”. In: *Macromolecules* 1.1 (1968), pp. 96–97.

[22] John B Fenn et al. “Electrospray ionization–principles and practice”. In: *Mass Spectrometry Reviews* 9.1 (1990), pp. 37–70.

[23] William J Griffiths et al. “Electrospray and tandem mass spectrometry in biochemistry”. In: *Biochemical Journal* 355.3 (2001), pp. 545–561.

[24] Natalia Felitsyn, Michael Peschke, and Paul Kebarle. “Origin and number of charges observed on multiply-protonated native proteins produced by ESI”. In: *International Journal of Mass Spectrometry* 219.1 (2002), pp. 39–62.

[25] WE Stephens. “A Pulsed Mass Spectrometer with Time Disaersion”. In: *Phys. Rev.* 69 (1946), p. 691.

[26] Stephen D Fuerstenau and W Henry Benner. “Molecular weight determination of megadalton DNA electrospray ions using charge detection time-of-flight mass spectrometry”. In: *Rapid Communications in Mass Spectrometry* 9.15 (1995), pp. 1528–1538.

[27] Wolfgang Paul. “Electromagnetic traps for charged and neutral particles (Nobel lecture)”. In: *Angewandte Chemie International Edition in English* 29.7 (1990), pp. 739–748.

[28] Wolfgang Paul and Helmut Steinwedel. “Ein neues massenspektrometer ohne magnetfeld”. In: *Zeitschrift für Naturforschung A* 8.7 (1953), pp. 448–450.

[29] Dunmin Mao and DJ Douglas. “H/D exchange of gas phase bradykinin ions in a linear quadrupole ion trap”. In: *Journal of the American Society for Mass Spectrometry* 14.2 (2003), pp. 85–94.

[30] RA Yost and CG Enke. “Selected ion fragmentation with a tandem quadrupole mass spectrometer”. In: *Journal of the American Chemical Society* 100.7 (1978), pp. 2274–2275.

[31] *Mass Analyzers (Mass Spectrometry)*. [Online; accessed 2021-08-03]. May 18, 2021. URL: https://chem.libretexts.org/@go/page/324.

[32] Ernest O Lawrence and M Stanley Livingston. “The production of high speed light ions without the use of high voltages”. In: *Physical Review* 40.1 (1932), p. 19.

[33] Melvin B Comisarow and Alan G Marshall. “Fourier transform ion cyclotron resonance spectroscopy”. In: *Chemical physics letters* 25.2 (1974), pp. 282–283.

[34] I Jonathan Amster. “Fourier transform mass spectrometry”. In: *Journal of mass spectrometry* 31.12 (1996), pp. 1325–1337.

[35] Michael L Easterling, Todd H Mize, and I Jonathan Amster. “Routine part-per-million mass accuracy for high-mass ions: Space-charge effects in MALDI FT-ICR”. In: *Analytical chemistry* 71.3 (1999), pp. 624–632.

[36] Qizhi Hu et al. “The Orbitrap: a new mass spectrometer”. In: *Journal of mass spectrometry* 40.4 (2005), pp. 430–443.

[37] Roman A Zubarev et al. “Electron capture dissociation for structural characterization of multiply charged protein cations”. In: *Analytical chemistry* 72.3 (2000), pp. 563–573.

[38] John EP Syka et al. “Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry”. In: *Proceedings of the National Academy of Sciences* 101.26 (2004), pp. 9528–9533.

[39] Gene Hart-Smith. “A review of electron-capture and electron-transfer dissociation tandem mass spectrometry in polymer chemistry”. In: *Analytica chimica acta* 808 (2014), pp. 44–55.

[40] Jos Oomens et al. “Gas-phase infrared multiple photon dissociation spectroscopy of mass-selected molecular ions”. In: *International Journal of Mass Spectrometry* 254.1-2 (2006), pp. 1–19.

[41] Annette Michalski et al. “A systematic investigation into the nature of tryptic HCD spectra”. In: *Journal of proteome research* 11.11 (2012), pp. 5479–5491.

[42] Lingdong Quan and Miao Liu. “CID, ETD and HCD fragmentation to study protein post-translational modifications”. In: *Mod Chem Appl* 1.1 (2013), pp. 1–5.

[43] Yu Xia, Xiaorong Liang, and Scott A McLuckey. “Ion trap versus low-energy beam-type collision-induced dissociation of protonated ubiquitin ions”. In: *Analytical chemistry* 78.4 (2006), pp. 1218–1227.

[44] John N Louris et al. “Instrumentation, applications, and energy deposition in quadrupole ion-trap tandem mass spectrometry”. In: *Analytical Chemistry* 59.13 (1987), pp. 1677–1685.

[45] Jesper V Olsen et al. “Higher-energy C-trap dissociation for peptide modification analysis”. In: *Nature methods* 4.9 (2007), pp. 709–712.

[46] Béla Paizs and Sándor Suhai. “Fragmentation pathways of protonated peptides”. In: *Mass spectrometry reviews* 24.4 (2005), pp. 508–548.

[47] Viswanatham Katta, Swapan K Chowdhury, and Brian T Chait. “Use of a single-quadrupole mass spectrometer for collision-induced dissociation studies of multiply charged peptide ions produced by electrospray ionization”. In: *Analytical chemistry* 63.2 (1991), pp. 174–178.

[48] Sven H Giese, Lutz Fischer, and Juri Rappsilber. “A study into the collision-induced dissociation (CID) behavior of cross-linked peptides”. In: *Molecular & Cellular Proteomics* 15.3 (2016), pp. 1094–1104.

[49] Hadi Lioe and Richard AJ OrsHair. “A novel salt bridge mechanism highlights the need for nonmobile proton conditions to promote disulfide bond cleavage in protonated peptides under low-energy collisional activation”. In: *Journal of the American Society for Mass Spectrometry* 18.6 (2007), pp. 1109–1123.

[50] Mark F Bean and Steven A Carr. “Characterization of disulfide bond position in proteins and sequence analysis of cystine-bridged peptides by tandem mass spectrometry”. In: *Analytical biochemistry* 201.2 (1992), pp. 216–226.

[51] Mingxuan Zhang and Igor A Kaltashov. “Mapping of protein disulfide bonds using negative ion fragmentation with a broadband precursor selection”. In: *Analytical chemistry* 78.14 (2006), pp. 4820–4829.

[52] Pei Lun Tsai, Sung-Fang Chen, and Sheng Yu Huang. “Mass spectrometry-based strategies for protein disulfide bond identification”. In: *Reviews in Analytical Chemistry* 32.4 (2013), pp. 257–268.

[53] Michael Mormann et al. “Fragmentation of intra-peptide and inter-peptide disulfide bonds of proteolytic peptides by nanoESI collision-induced dissociation”. In: *Analytical and bioanalytical chemistry* 392.5 (2008), pp. 831–838.

[54] Daniel F Clark et al. “Collision induced dissociation products of disulfide-bonded peptides: ions result from the cleavage of more than one bond”. In: *Journal of the American Society for Mass Spectrometry* 22.3 (2011), pp. 492–498.

[55] Kirt L Durand et al. “Tandem mass spectrometry (MSn) of peptide disulfide regio-isomers via collision-induced dissociation: utility and limits in disulfide bond characterization”. In: *International Journal of Mass Spectrometry* 343 (2013), pp. 50–57.

# Appendix A

# Using CoolThesisSoftware