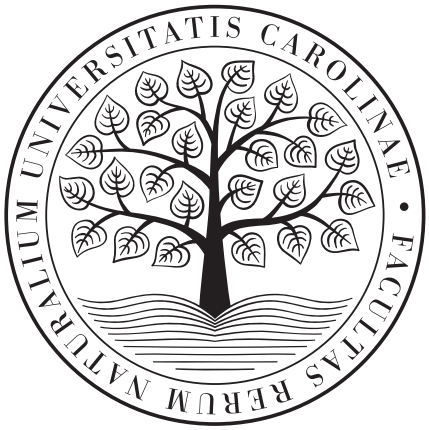
Charles University in Prague
Faculty of Science

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 *breptides* 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-

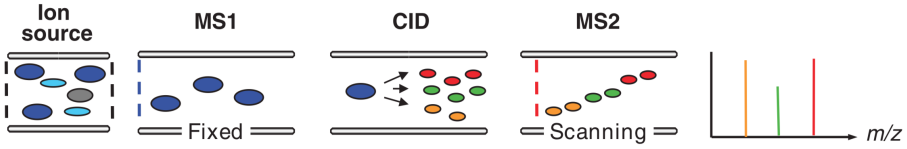


Figure 1.1 An ordinary MS/MS workflow diagram. While the specific intrumentation details differ from spectrometer to spectrometer, the general structure of ionize ⟶ analyse ⟶ fragment ⟶ analyse is common to all of the MS/MS spectrometry experiments. Image taken from Domon and Aebersold [12].

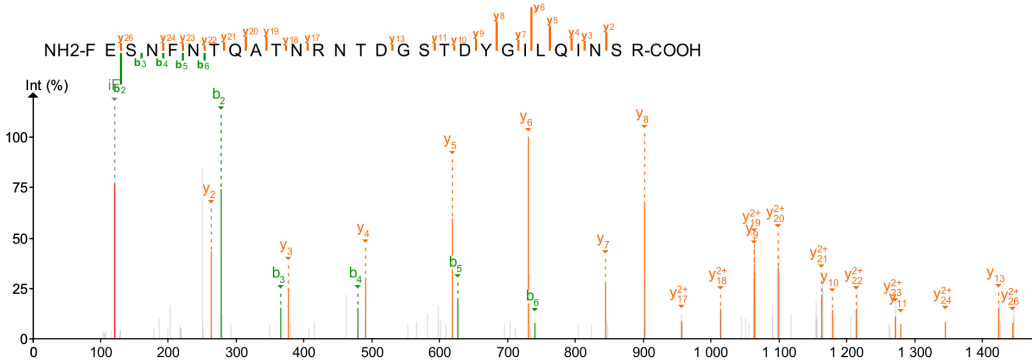


Figure 1.2 An annotated fragmentation spectrum of the precursor *FESNFNTQATNRNTDGSTDYGILQINSR*

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.

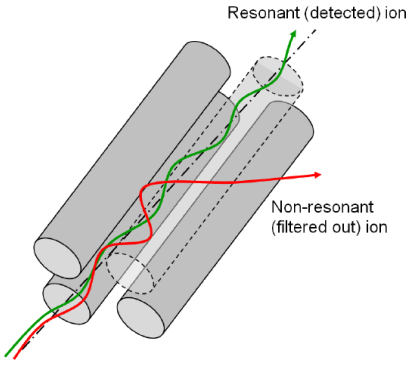


Figure 1.3 A quadrupole with two highlighted classes of ion trajectories. Thanks to its m/z value, the ion with green trajectory passes through the quadrupole and is ultimately detected, while the one with the red trajectory is filtered out. Image taken from *Mass Analyzers (Mass Spectrometry)* [31].

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.

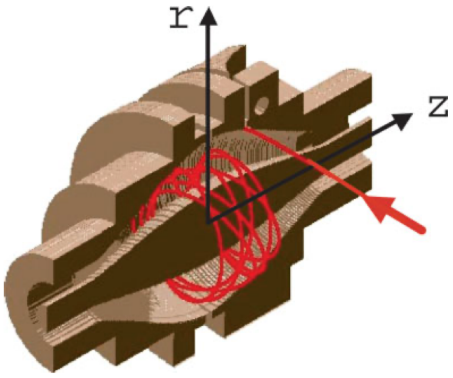


Figure 1.4 An orbitrap mass analyser with a typical ion trajetrory highlighted. The ion circulates around the center while simultanously oscillating along the z-axis. Image taken from Hu et al. [36].

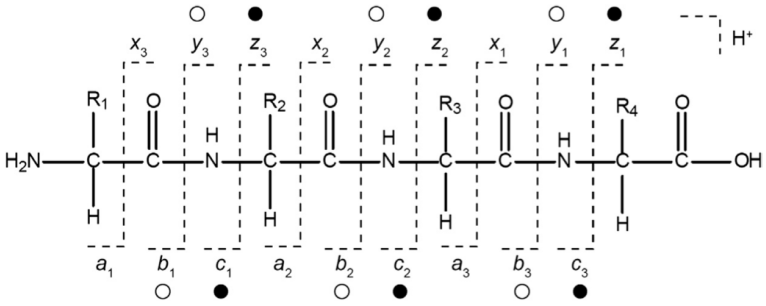


Figure 1.5 A singly positively charged peptide with annotated fragmentation types. Signature CID b/y ion fragments are marked with open circles, while the typical ECD and ETD c/z ion fragments are marked with filled circles. Image taken from Hart-Smith [39].

When aiming to observe post-translational modifications and to preserve 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. These methods require a multiply-charged precursor ion, because the absoption of the elector lowers the overall charge by 1, making singly charged precursor ions undetectable.

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 breptides [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 molecules

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 ac-

[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.

curacy. 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].

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; the accuracy 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 during in silico matching of the spectra due to the combinatoric nature of the problem. Furthermore the detected internal ions can occasionaly be 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

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

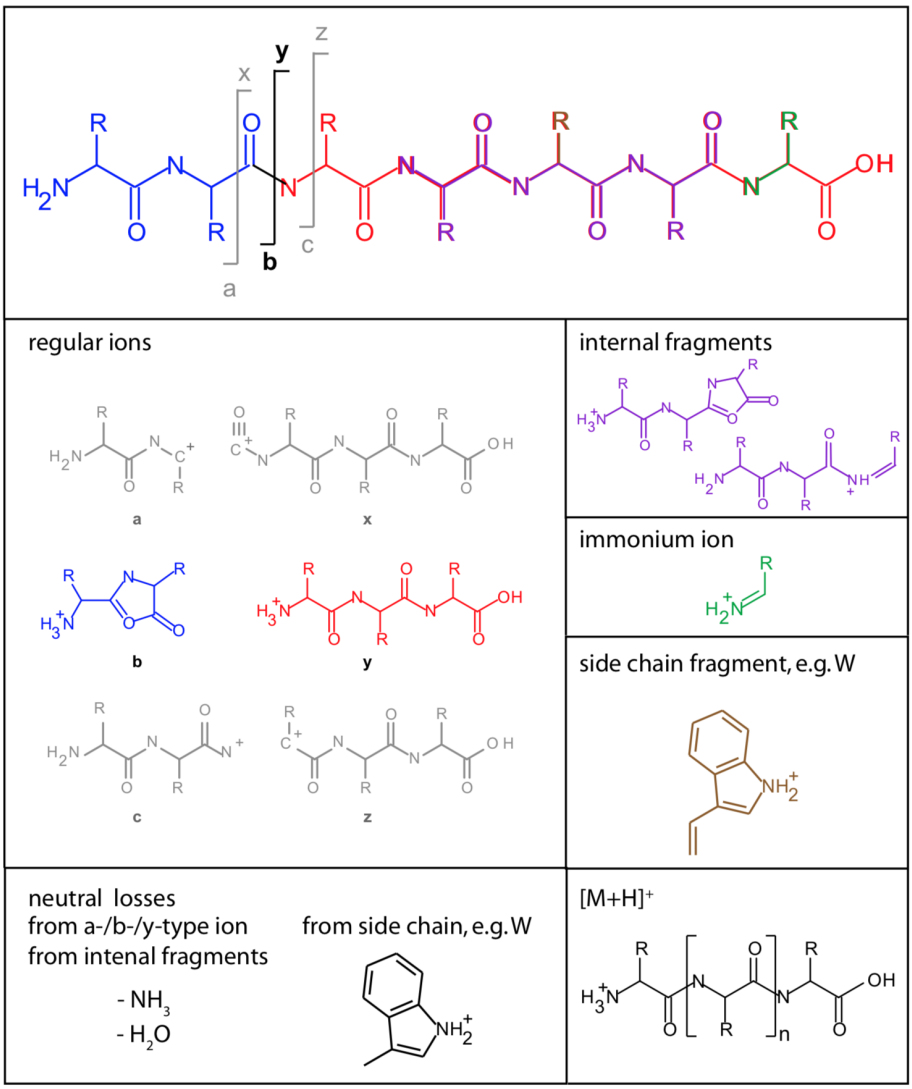


Figure 1.6 During HCD, *b*, *y*, and to a lesser extent *a*, ions are the most common, together with internal ions and immonium ions, and their counterparts with neutral losses. Image taken from Michalski et al. [41].

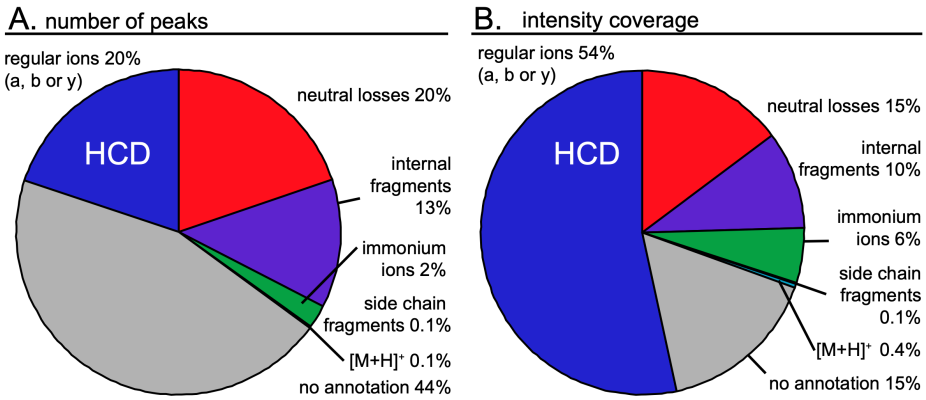


Figure 1.7 (B) The regular *b*, *y*, and *a* ions take up 54% of the measured spectral intensity. Another 25% is explained by fragments with neutral-loss and internal fragments, and another 6% by immonium ions. Together, those four account for 85% of the measured intensity. It is true that almost a half 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 Michalski et al. [41].

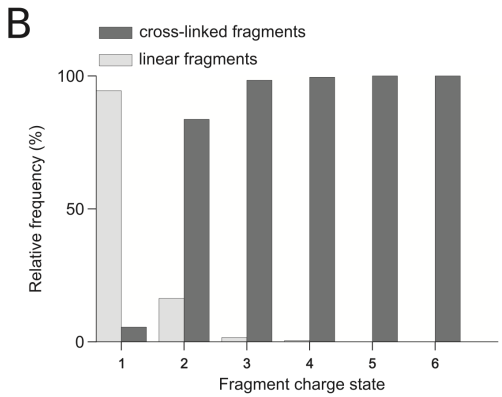


Figure 1.8 Most linear fragments have charge 1, and most crosslinked fragments have a charge of 2 or higher. Image taken from Giese, Fischer, and Rappsilber [48].

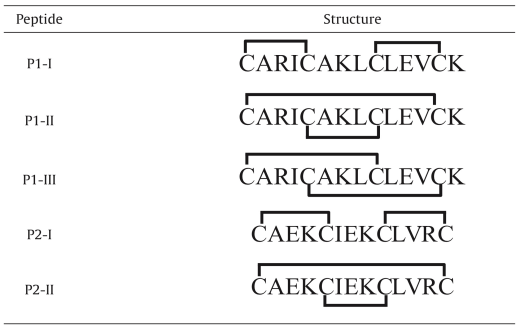


Figure 1.9 An example of different possible configurations of intra-peptide DBs. Image taken from Durand et al. [55].

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.

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 mass 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 a higher frequence of internal ions. 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.

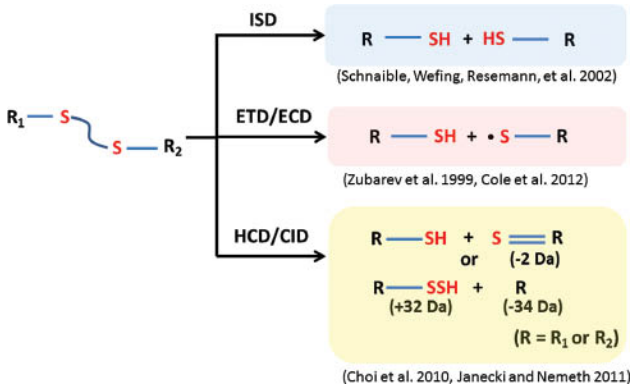


Figure 1.10 Under different dissociation strategies, DBs manifest different cleavage characteristics. Under CID, the cleavage results into two possible assymetrical mass distributions. Image taken from Tsai, Chen, and Huang [52].

1.4 Breptide spectra interpretation

This section concerns itself with the existing methods aiming to determine the quantity and position of DBs in a protein using tandem mass spectrometry. We focus on the computational methods, and finally formulate the precise task that this thesis is trying to solve together with a simple complexity analysis.

As noted by Lakbub, Shipman, and Desaire [56], there are two main types of DB characterization. Profile comparison methods make use of two samples, one reduced with the DBs removed and one nonreduced that has its DBs still intact. Differential analysis is then deployed on a chromatogram profile of these two samples to determine which peptides are in the nonreduced sample, but not in the reduced one. Those peptides are suspect of being breptides and are further analyzed by MS/MS. Intact analysis methods only data from the nonreduced sample. Thus, they simpler from the sample preparation standpoint, but have to work with less information than the profile comparison methods.

In each of the both categories, the protocols further differ in the choice of of sample separation, ionization, and fragmentation methods, the choice of mass analyzer, and whether the bulk of the analysis is performed manually or automatically [56]. Some of the software for automatic DB characterisation is reviewd below; for the sake of completness, an example of a partially-manual method now follows.

Wu et al. [57] propose an intact analysis method based on LC-MS/MS with ETD. The prepared breptides are measured on MS1 and fragmented by ETD. During ETD the DBs are dissociated, resulting in fragmentation spectra with two prominent peaks that represent the original peptides that were connected with the DB. Fragments from these two peaks are put into an MS3 step with

CID fragmentation to gain sequence information about them. This sidesteps the problem of not having data from the non-bound peptides in the intact methods. Software for fragment mass searching and matching is used, but because it is not made with DB research in mind, the method requires manual intervensions when interpreting internal ion peaks, or when the peptide assignment to the birdgetide spectra based on the two prominent ETD peaks is not clear cut. Specialized sofrtware dedicated for DB cahracterisation is reviewed in the next section.

1.4.1 Current computational approaches

Dedicated DB characterisation software usually only needs data from nonreduced samples, however, there are some commercial options that offer the possibility to add data from reduced samples as well, such as PepFinder and BioPharma Finder, as noted by Lakbub, Shipman, and Desaire [56]. Refer to the same publication for a more comprehensive list of past and current manual and compuational DB characterisation methods.

SimXL is tool for general peptide cross-linking analysis [58], including DBs [59], seemingly without having a preferred fragmentation method. SimXL has three main differentiating factors. Frist is the user-friendly UI that allows the reasearchers to view not only the interpreted results, but also the annotated data based on which they were computed. Second is its search space reduction heuristc based on the presence of a reporter ion [60], a peak that is specific to the fragmentation spectra of crosslinked precursor; in the context of our task, alkylated cysteines could be possibly interpreted as reporters [61]. Finally, SimXL employs a further search space-reduction heuristic based on dead-end modifications, but we believe it is probably inapplicable in the context of DB characterisation.

A popular method by Liu, Breukelen, and Heck [62] comes with a whole recommended research protocol. Samples are digested with pepsin to avoid the DB scrambling that is typical low-pH for tryptic digestion, fragmented with ETH followed by HCD, together called *electron transfer higher energy dissociation* (EThcD), and finally analyzed with SlinkS, a dedicated DB matching algorithm. As mentioned previously, it is possible to extract information about the peptides contituting the precursor ion from the ETH fragmentation spectra. The HCD step is employed in order to gain sequenec information about these peptides, similary to how MS3 has been used in [57].

Ultimately, thanks to EThcD, two peaks corresponing to the precursor peptide pair are identified in each fragmentation spectrum, and the whole precursor match is scored by scoring the other individual fragments now that we know from which precursor they (allegedly) came. The method is elegant, but its main shortcoming is the fact that it only works with dipeptide precursors; it also ig-

nores fragments with neutral loss and internal ion fragments. That makes it impossible to identify some of the more complicated DB configurations.

In fact, all of the aforemetioned methods should be able to match simple interpeptide DBs, but they reportedly struggle with more complex intrapeptide bonds or intertwined interpeptides [56]; some of those can be seen on Figure 1.11. A wide array of approaches to DBs is offered, leading to a yet wider array of recommended sample preparation and fragmentaion protocols. We conclude that computational DB characetrisation is a hard problem to solve, a notoin we further formalize in the next section.

1.4.2 Problem statement and complexity

Undisputably, the problem of determining the characteristics of DB linkages in proteins is hard. From biochemical point of view, already sample preparation poses a challenge; there is a need to minimize DB scrambling, but at the same time have the protease be as specific as possible to simplify the subseqeucnt analysis.

Another unresolved problem is the choice of dissociation method. ETD spectra offer the information about the constituent peptides, which is undisputably useful when assigning precursor peptides. On the other hand, CID-based approaches have access to richer but more complicated fragmentation spectra, including interlinked fragments and internal ions with intact DBs; these are useful for precise pinpointing of the DB location.

To continue this discussion further, we need to be specific about the precise task we are attempting to solve. In this thesis, the task is to identify precursor breptides in provided mass-spectrometric data, score them, and use the calculated scores to weigh the information the precursors provide about a part of the analyzed protein cysteines. The scoring will be done based on how well will the a in silico generated fragments of the assigned breptide match the measured fragmentation spectrum. Thus is the split into two similar subproblems: precursor matching, and then scoring by fragment matching. We will describe theoretical complexity of fragment matcing below, and most of the remarks will carry over to precursor matching as well. While the scoring paradigm itself is a very complicated thing, the main algorithmic complexity comes from the matching of fragments which is a prerequisite for the score calculation step.

A simplified variant of the task

We are provided a precursor R comprising of n residues with integer masses, interlined with DBs and peptide bonds, and a target integer fragment mass f. Our task is to find all fragment ions from R whose mass exactly matches f.

No.	Protein	Enzyme(s)	Disulfide bonded peptides	Comment	Ref.
1	RNase A	Asp-N/C + trypsin	CKPVNTFVHESLA YPNCAYK	Two-chain DSBP with one interchain DSB	72
2	IgG3	Trypsin	SCDTPPPCPR SCDTPPPCPR	Two-chain DSBP with two interchain DSBs (disulfide box)	54
3	RNase A	Trypsin	CKPVNTFVHESLADVQAVCSQK YPNCAYK HIIVACEGNP- - -DASV	Three-chain DSBP with two interchain DSBs	72
4	HIV, gp140	Trypsin	NDMVDQMHEDIISLWDQSLKPCVK DNSDNSEYILINCDASTITQACPK LTPLCVTLHCDATFK DCSFDTTTEIR	Four-chain DSBP with three interchain DSBs - PNGase F used for deglycosylation	75
5	CTT, gelatinase inhibitor	N/A	CTTHWGFTLC	Completely cyclisized DSBP	91
6	Chicken lysozyme	Trypsin	NLCNIPCSALLSSDDITASVNCAK WWCNDDR	Two-chain DSBP with both interchain and intrachain DSBs	80
7	rhASA	Lys-C + trypsin + Asp-N	DQGPCQ DLTCFPPATPCDGGC	Two-chain DSBP with both interchain and intrachain DSBs	71
8	rhASA	Pepsin	PALQICCHPGCTPRPACCHCPDPHA	Single-chain DSBP with three intrachain DSBs (nested disulfides)	71

Figure 1.11 Illustrative examles of different ways peptides can be connected by disulfide bonds, ranging from realtively simple examples (1, 3, 5) to complex multipeptide or multibond configurations (4, 8). The more complicated configurations proved to be hard to characterise computationally. Image by Lakbub, Shipman, and Desaire [56]. (DSB = disulfide bond, gp140 = glycoprotein 140, rhASA = recombinant human arylsulfatase A)

We can model the precursor as a graph $G_R = (V, E)$ with weighted nodes, where $|V| = 1 \dots n$ and $(i, j) \in E$ if and only if there is a peptide or a disulfide bond connecting the i-th and j-th residue in R. The weight w_i of the vertex i is the mass of the corresponding i-th residue in the precursor. We will call this graph the *precursor graph of the precursor R*. Even though the bonds in R are constrained in terms of connectivity — there can be at most one DB connected to a given residue, and at most two peptide bonds — G_R can still in theory be a relatively complex non-planar graph, as illustrated on Figure 1.12.

To solve the task, we need to find all contiguous subgraphs of the precursor graph of R in which the sum of weights is exactly f, a problem that is usually called the exact weight subraph problem [63], or, in our case, an exact weight *connected* subgraph problem. A well-studied closely related problem called the maximum weight connected k-subgraph problem was shown to be NP-hard on general graphs, and even on planar and bipartite graphs with integer weights [64]; the same authors propose a polynomial-time $O(k^2 n)$ dynamic programming algorihm for the a restricted version of the problem searching for subgraphs in trees. The exact weight conencted subgraph problem has not been studied quite as thoroughly, but we think it is safe to assume that its complexity will not be much lower. In other words, in the context of our not-necessarily-planar precursor graphs, the problem is probably NP-hard.

Let us now say that given a precursor R a target fragment mass f, we have identified a set F of connected subgraphs in G_R with the target mass (or a set of *fragments* for short). Not all of these fragments can occur in real-world measured spectra, even though they do have the correct mass. The number of edges that have at least one vertex in a fragment, but are not themselves in the set of edges of the fragment, represent fragmentation bond cleavages in the original precursor. In HCD, the most common b and y fragment ions result from single bond cleavage; so-called internal ions also du occur, but are much more rare. The exact percentage of fragments with 3 or higher number of cleavages is hard to determine in linear peptides, and as far as we know has not been determined in crosslinked peptides, either. However, if they do occur, they will probably be exceedingly rare. The number of cleavages is thus an additional contraint on the fragment matching process.

Apart from these constraints, the simplified versoin ignored some of the inherent complexity of the real mass spectra. First of all, the measured fragment masses are rational numbers, not integers. We also do not need an exact match, but have a tolerance range in which we consider the match to be successful; furthermore, the range is not absolute, but is defined relative to the mass of the generated fragment (in ppm). We also have to take into account the possible modifications of amino acid residues, the possibile occurence of a neutral loss, and the asymetrical nature of disulfide bond cleavage under CID.

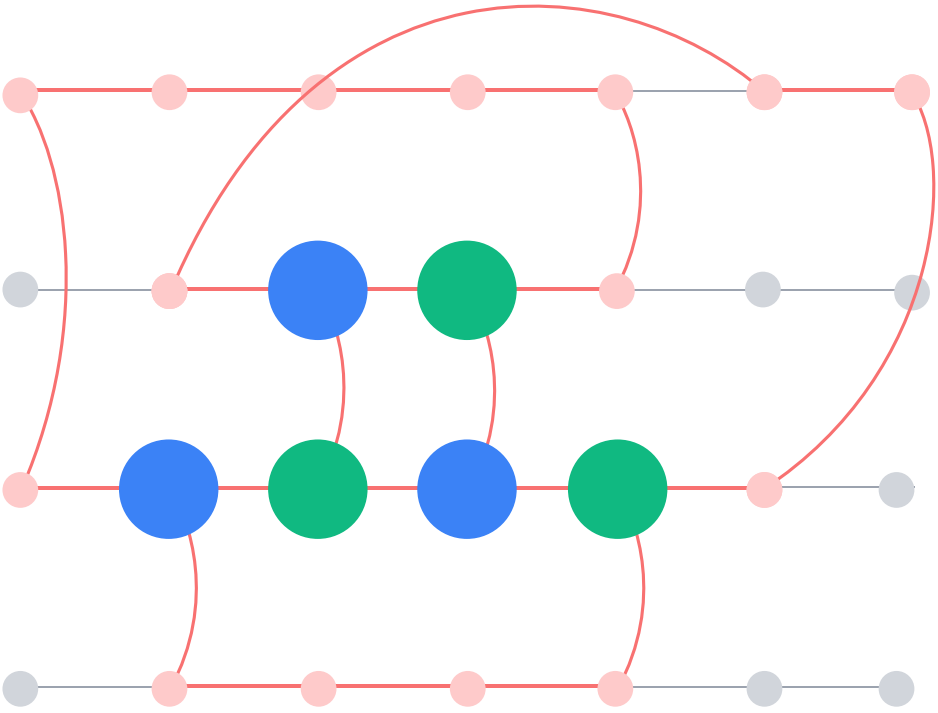


Figure 1.12 An example of a valid precursor graph that is not planar due to the presence of a $K_{3,3}$ subdivision [65] (highlighted). Vertices represent amino acid residues, horizontal lines represent peptide bonds, and vertical curved lines represent disulfide bonds. The precursor could be made out of four different interlinked peptides, but it is alsopossible for it to be a single peptide with many intrapeptide disulfide bonds.

All of these add additional complexity to the problem, and make it hard to devise a general well-performing algorithm. What is more, accounting for all of these possibilities creates a big potential of generating false-positive hits due to the great combinatorial power of the matching algorithm. Nevertheless, in the next chapter we describe a general algorithm that attempts to solve the general version of this problem.

Chapter 2

Differential characterisation of disulfide bonds

Reagents and instrumentation Trypsin was purchased from Roche, lysosyme and lipase were obtained from Sigma-Aldrich. The used liquid chromatography setup was RSLC nano Ultimate 3000 (Thermo Scientific), for tandem mass spectrometry Orbitrap Fusion Lumos (Thermo Scientific) was employed.

Obtaining the data We have anaylzed two proteins, lysosyme (LYS) and lipase (LIP). Two samples of each have been obtained in separate sample preparation pathways; in the AT sample preparation pathway, the samples have been alkylated without prior reduction, while in the RAT pathway, reduction took place before the alkylation. In both cases the alkylation has been done by iodacetamide (IAA), resulting in a covalent addition of a carbamidomethyl group (57.07 Da) onto cysteines without DBs; the reduction was achieved by dithiothreitol (DTT). After that, all four samples have been fully digested by trypsine, resulting in two sets of tryptides for each of the two analysed proteins. The samples have been treated the same for the rest of the experiment. After digestion, the peptides had undergone reverse-LC separation that was directly connected to the mass spectrometer through a nanospray ion source. After the measurement of MS^1 spectra on a quadrupole, the precursors continue on to be fragmented with HCD. The MS^2 spectra were analysed on an orbitrap analyser, as is implied by the use of HCD fragmentation. Finally, raw MS^2 match data was exported to mgf files. Additionaly an artifical testing dataset based on in-silico digestion and fragmentation of he protein ovalbun has been prepared.

The data have been kindly provided by a IOCB mass specrometry research group (J. Cvačka), and have been measured a few years ago without any direct connection to this thesis.

2.1 Dibbi, a program for disulfide bond visualisation

A Python program termed Dibbi has been developed to analyse the positions of DBs in a protein. The input to Dibbi is the sequence of the analysed protein, and the mgf file with the measured MS^2 data. It then provides a visualisation of computed scores of each of the possible DBs, and a file with the in-silico generated matches to the individual fragments and precursors, allowing for further manual analysis of the visualised results. Although Dibbi is perfectly able to operate only with the AT sample data, for best results it is preferable to analyse their RAT counterpart first. This enables the user to manually tune the parameters of the subsequent AT analysis, allowing for lower number of false positive hits.

Dibbi first assigns dynamically generated *precursors* to the measured precursor matches, and then assigns in-silico generated fragments from the assigned precursors to the individual peaks in the spectra. Both of the algorithms make use of the divide and conquer [66] and branch-and-bound [67] techniques. Finally, the precurors are scored according to the quality of their fragment matches, among other things, and their score-weighted contributions are added together to form a general overview of the disulfide bonds in the protein. The precursor matching, fragment matching, and scoring algorthms are described in their own sections below.

2.1.1 Assigning precursor mass

At the start, the protein undergoes a complete in-silico digestion, producting basal peptides that can not be digested any further; the protease of our choosing was trypsin, so we call these basal peptides *tryptides*. As mentioned in the first chapter, a protease can sometimes miss a cleavage point, resulting in a peptie chain that is made from two (or more) contiguous basal tryptides connected with a peptide bond. We call a chain of one or more tryptides a *segment*. Finally, due to disulfide bridges crosslinking, a *precursor* is made out of one or more segments connected with interpeptide disulfide bonds.

There can be one or more DB connecting a segment to another segment, and there can also be intrasegment DBs. Cysteines that do not partake in a DB are alkylated, meaning their mass is effectively higher by some configurable constant. In addition, certain amino acid residues are sometimes modified, too, such as the methionine undergoing oxidation. The algorithm treats the latter as optional, or "variable" modifications.

From the point of view of this part of the algorithm, it matters not where the individual DBs are located, or which exact methionines are modified, because

the individual variations are not distiguishable on the basis of precursor mass. Thus, the precursors in the output of the algorithm only include the information about **which segments are present**, **how many DBs there are** — always at least the number of segments minus 1 — and how many of each kind of variable modifications there are. This information is passed to the next stage of the program.

Given a target mass to which we want to assign a precursor, the algorithm first chooses a beginning of the first segment. After that the FindPrec function iteratively branches out to search the problem space for precursors that have a theoretical mass within an error boundary around the target mass. The precursor is built segment by segment, and the individual segments are build tryptide by tryptide, both from left to right. In each iteration, the following branching points are tried out:

1. Combine the possible modifications and DBs in such a way that the *Selected* segments with the modifications have the correct mass, using the Combine function (see 1.5). Combine implements a divide-and-conquer algorithm for a modified subset sum problem, in which assignments that are within some error boundary of the target are also considered a valid solution.

2. Elongate the current segment by adding the current tryptide to it (see 1.22, and also the whole Algorithm 2).

3. End the current segment (effectively simulating a protease cleavage just before the current tryptide) and begin a new one that begins on the current tryptide or later. See 1.17, and the whole Algorithm 3.

The fact taht the precursors (and their segments) are built strictly from left to right is a simple form of symmetry breaking — that is, not generating the multiple equivalent symmetrical solutions. Some further optimizations were added as well, mainly to prevent traversing a branch once we learn it can not provide any more solutions. The used protease is configurable, as well as the amino acid modifications and the maximum number of segments we want the found precursors to have.

2.1.2 Assigning fragment mass

Our task now is to assign in-silico generated precursor fragments to the peaks from the spectra. The quality of these assignments will later be used to score the precursor matches and ultimately to find out some information about the position of DBs in the protein.

Algorithm 1 The main part of the precursor matching algorithm, in which all the braching occurs.

```
function FindPrec(I, Selected, Mass, Segments, Cys, Open)
    solutions ← empty list
    ▷ There is no hanging disulfide bond from previous segments
    ▷ So let us try to combine residue modifications to find a solution
    if not Open then
        mods ← calculate possible modifications from Selected
        alks ← calculate alkylation count based on seen non-bonded Cys
        combinations ← Combine(mass, mods, alks)
        return a list of precursors generated from combinations
    end if
    ▷ There are no further solutions, our Mass can not get low enough
    if Mass is too high, or i is at the end then
        return empty list
    end if
    ▷ End this segment, connect next one with a disulfide bond
    ▷ That is, if we have the segments budget and a free cysteine
    if not Open, and Segments > 0, and Cys > 0 then
        S ← NewSegment(I, Selected, Mass, Segments, Cys, Open)
        concatenate solutions with the list S
    end if
    ▷ Elongate the current segment by one tryptide
    S ← Elongate(I, Selected, Mass, Segments, Cys, Open)
    concatenate solutions with the list S
    return the list solutions
end function
```

Algorithm 2 Elongates the currently built precursor segment, adding the current tryptide to it.

function ELONGATE(I, $Selected$, $Mass$, $Segments$, Cys, $Open$)
▷ Prolong the current segment by one tryptide
$tryptide \leftarrow$ the I-th tryptide from the list *TRYPTIDES*
$mass' \leftarrow$ the mass of *tryptide* added to *Mass*
$cys' \leftarrow$ the number of cys in *tryptide* added to *Cys*
$cys' \leftarrow$ lower cys' by one if *Open* is true, using a cys to close the bond
$open' \leftarrow$ False if this tryptide had any cysteines, otherwise *Open*
▷ Call the original function
$S \leftarrow$ FINDPREC($I + 1$, *Selected*, $mass'$, *Segments*, cys', $open'$)
return the list S
end function

Algorithm 3 Ends the currently built precursor segment and begins a new one, beginning with the current tryptide, or any tryptide coming after it.

function NEWSEGMENT(I, *Selected*, *Mass*, *Segments*, *Cys*, *Open*)
$sel' \leftarrow$ the currently ending segment added to *Selected*
$mass' \leftarrow$ subtract mass of H_2 from *Mass*, due to the new DB
▷ Update the budget, because of the newly started segment
$seg' \leftarrow Segments - 1$
▷ Begin the bond with one of our cysteines
$cys' \leftarrow Cys - 1$
▷ The new bond is waiting to get "closed" by a cys in the next run
$open' \leftarrow$ True
▷ Finally, branch out: start a new segment from all possible starting points
for all possible beginings of the next segment from $I + 1$ onward **do**
$i' \leftarrow$ the next beginning
$S \leftarrow$ FINDPREC(i', $selected'$, $mass'$, $segments'$, cys', $open'$)
return the list S
end for
end function

Before the main of this stage of the program, there is an in-between processing step that takes a precursor from the precursor-matching stage, and generates all possible *variants* it could represent. A *variant* is a set of segments with precisely defined DB crosslinks — that differentiates it from a precursor (which is the output of the previous stage), as precursor only holds information about the number of DBs, but not their positions.

Similarly to how precursors are constructed from crosslinked segments, and the segments from basal tryptides, fragments can also be broken down to smaller pieces; we call them *cuts*. A *cut* is a part of a precursor segment; it never spanns more than one segment. All of the fragment cuts form a connected subgraph in the variant graph.

A rough outline of the algorithm can be seen in Algorithm 4. At the start, a beginning is chosen for the first cut of the fragment. In further iterations of the main function, the algorithm branches out. Before we describe the braching, we have to explain the notion of *pivots*. Imagine the algorithm is building a cut, residue by residue, and it encouters a cysteine connected with a DB to another segment in the variant. When the algorihm makes the decision to keep the bond intact, the other cysteine surely will have to be in the final fragment, possibly with some other neighbouring residues from its segment. The cysteine is thus added to the list of *pivots*, and when the current cut is ended, the algorithm jumps to the next pivot from the list and builds a new cut around them.

In this way, the algorithm can jump back and forth between segments, following the directions of DBs in the variant, building cuts around cysteines, until it runs out of pivots, runs out of residues in the variant, or until its mass is too high. The algorithm keeps track of all residues that are part of the current fragment, in order not to add any residue twice by adding cuts with nonempty intersection. Furthermore, the number of peptide bond and disulfide bond breaks is bounded, usually to be at most 1 or 2.

In each iteration, the algorithm attempts to do all of the following:

1. Try to combine the masses of the selected cuts and some of the variable modifications to obtain a fragment with a mass that is within the error boundary around the target mass. In case of a success, add this fragment (that is, this combination of cuts and modifications) to a list of solutions.

2. If the current residue is a cysteine partaking in a DB that we have not seen yet, branch out. Break the bond in one branch, adding assymetric modifications to both of the cysteines, and keep the bond intact in the other, adding a new *pivot* to the list of pivots.

3. Elongate the current cut by adding the next residue to it. The "next residue" is simply the next residue in the segment from which the cut is made. If

there is no "next residue" in this segment, end this cut, and jump to the next pivot. If there is no pivot to jump to, end this branch of the algorithm.

4. End this cut prematurely (in the middle of a segment), that is, simulate a peptide bond dissociation in the precursor. Jump to the next pivot, if there is any, or end this branch of the algorithm.

2.1.3 Precursor scoring and result visualisation

Initially the idnvidiual precursor assignemnts from the first stage are scored, then the fragment assginments from the second stage, and the two scores are put together to score the variants (precursors with concrete bond configurations). After that, we treat each variant as evidence confirming its spefcific DB configuration, and we add the individual pieces of evidence from every variant together, each evidence weighted by the score of its parent variant. We treat evidence abuout alkylation in the same way. Finally, we visualise the collected evidence.

Precursor scoring The score for a precursor P is computed as

$$\text{score}(P) = \frac{1}{1 + (variants_P, mc_P, mass_P, error_P)\boldsymbol{w_P}},$$

where $\boldsymbol{w_P}$ is a vector of weights, *variants* is the number of different variants that can be generated from P, *mc* is the maximum number of missed cleavages among its segments, *mass* is its mass, and *error* is the ppm error of the assignment. All of the attributes of P are normalised before the score is computed. In this thesis we set $\boldsymbol{w_P} = (32, 4, 4, 4)^T$.

Fragment scoring The score for a fragment F is computed as

$$\text{score}(F) = \frac{1}{1 + (charge_F, mods_F, error_F)\boldsymbol{w_F}},$$

where $\boldsymbol{w_F}$ is a vector of weights, *charge* is the charge of the fragment, *mods* is the total number of its mods (including neutral losses), and *error* is the ppm error of the assignment. All of the attributes of F are normalised before the score is computed. In this thesis we set $\boldsymbol{w_F} = (16, 4, 4)^T$. The weights were set according to our expectations based on the biological and informatical background of the problem, as well as according to the patterns in the data.

Algorithm 4 A very high-level overview of the basic functionality of the fragment matching algotihm.

function FragFind(I, *BreaksLeft*, *ValidEndRange*, *Mass*, *Pivots*)
  solutions ⟵ an empty list
  canend ⟵ the end would not be premature, or *BreaksLeft* > 0
  if I is in *ValidEndRange*, and also *canend* **then**
    if can Combine the cuts with some mods to match *TARGET* **then**
      solutions ⟵ the current *solutions* with any of the new ones added
    end if
    if there is some pivot p in *Pivots* **then**
      startrange, *endrange* ⟵ the valid cut range around p
      pivots′ ⟵ *pivots* with the pivot p removed
      breaks′ ⟵ *BreaksLeft*, possibly one less if the end is premature
      for every valid cut start in *startrange* **do**
        S ⟵ FragFind(*start*, *breaks*′, *endrange*, *Mass*, *pivots*)
        solutions ⟵ the current *solutions* concatenated to S
      end for
    end if
  end if
  if the I-th residue is a cys partaking in a DB we have not seen yet **then**
    ▷ Break the bond...
    if we have some breaks to spare **then**
      add the current cysteine to the broken bond counter, then...
      mass′ ⟵ the current *Mass* added to the mass of the I-th residue
      S ⟵ FragFind(I + 1, *BreaksLeft* − 1, *ValidCutRange*, *mass*′, *pivots*)
      solutions ⟵ the current *solutions* concatenated to S
    end if
    ▷ ...or keep it intact
    pivots′ ⟵ add the other end of the bond to *Pivots*
    mass′ ⟵ the mass of the bond, H_2, subtracted from *Mass*
    S ⟵ FragFind(I + 1, *BreaksLeft*, *ValidCutRange*, *mass*′, *pivots*′)
    solutions ⟵ the current *solutions* concatenated to S
  else
    ▷ Continue adding to this cut
    mass′ ⟵ the current *Mass* added to the mass of the I-th residue
    S ⟵ FragFind(I + 1, *BreaksLeft*, *ValidCutRange*, *mass*′, *pivots*)
    solutions ⟵ the current *solutions* concatenated to S
  end if
  return the list *solutions*
end function

Variant scoring The score of a variant V is computed as

$$\text{score}(V) = \text{score}(\mathit{precursor}_P) + w_f \cdot \text{median}(\{\text{score}(F), F \in \mathit{fragments}_V\}).$$

In this thesis we use $w_f = 1/2$.

Score aggregation For every theoretically possible disulfide bond $B = (u, v)$ the following evidence score is computed from the assigned fragments,

$$\text{score}(B) = \sum_{\text{assigned fragment } F} \text{score}(\mathit{variant}_F) \cdot [B \in \mathit{bonds}_F],$$

where *bonds* is a set of DBs that are present in the fragment. Similary for every cysteine C an alkylation score is computed,

$$\text{score}(C) = \sum_{\text{assigned fragment } F} \text{score}(\mathit{variant}_F) \cdot [C \in \mathit{alkcys}_F],$$

where *alkcys* is a set of alkylated cysteines that are present in the fragment.

Visualisation The aggregated evidence is plotted independenlty for each disulfide bond and — in the case of alkylation — for each cysteine. The evidence for a bond $B = (u, v)$ is visualised separately in each direction, (u, v) and (v, u). The evidence for bond (u, v) in that specific direction is normalised in the context of every other bond u, X, and the evidence for alkylation of the u cysteine. No further automatic interpretation is done, to keep the plots as close to the raw assignment data as possible. Along with a plot of the aAT sample for the protein, there is also a plot for the prediction of bonds in the RAT sample — in this way, the researcher can deduce which parts of the given protein are prone to generating false positives, and can adjust the weights for scoring the AT sample accordingly.In addition to those two plots, plots of "gold" data are shown, representing the theoretical true state of the analysed protein.

In both the fragment and the precursor scoring, the numerators can be dynamically adjusted based on what bond is the algorithm scoring. The users are encouraged to perform a differential analysis by first runnning the program as-is, and then adjusting for false-positive-prone cysteine bonds by lowering the weights of the bonds that the algorithm (falsely) identified in the RAT sample.

Chapter 3

Results and discussion

Scoring evaluation Data on which fragments generated which peaks in the measured data are unavailable, so its is impossible to directly test how well the metric separates true positives from false positives. However, we do know where the DBs are located in the protein, and a fragment (and by extension, the whole variant) that contradicts our knowledge about DB position should in general have a lower score than a fragment that does not. We have evaluated our variant scoring metric on four different datasets, showing its values separately for the two categories of variants; see Figure 3.1 for the results.

We have found that the metric in general satisfactorily spearates the "good" variants from the bad ones, nonetheless there are still many "bad" variants with a high score. In the in-silico generated GENOVA dataset the upper tail end of the bad variants ends much lower than it does in the other samples. We interpret this as follows: the upper part of the tail can be attributed to DB scrambling, a process that occured duting the preapartion of the real-world datasets, but didn't occur during the generation of GENOVA. In other words, the high-scoring bad variants should not exist in theory, but they were observed in the data with great condifence, and scored appropriately. On the other hand, the lower end of the tail are probably false positives probably caused by imprefections in our matching and scoring system.

Disulfide bond characterisation The complete analysis was performed on three proteins, namely LYS, LIP, and GENOVA. In all three cases, the samples were analysed in a two step process. The analysis was first run with vanilla settings. If there was a strong evidence for a bond in the RAT data, the bond was deemed to be prone to generating false positive signals. Its weight has been adjusted to 0.1, or 0.8, depending on the strength of its RAT evidence, and the analysis was run again.

The results provided in this thesis are already post-adjustment, and can bee

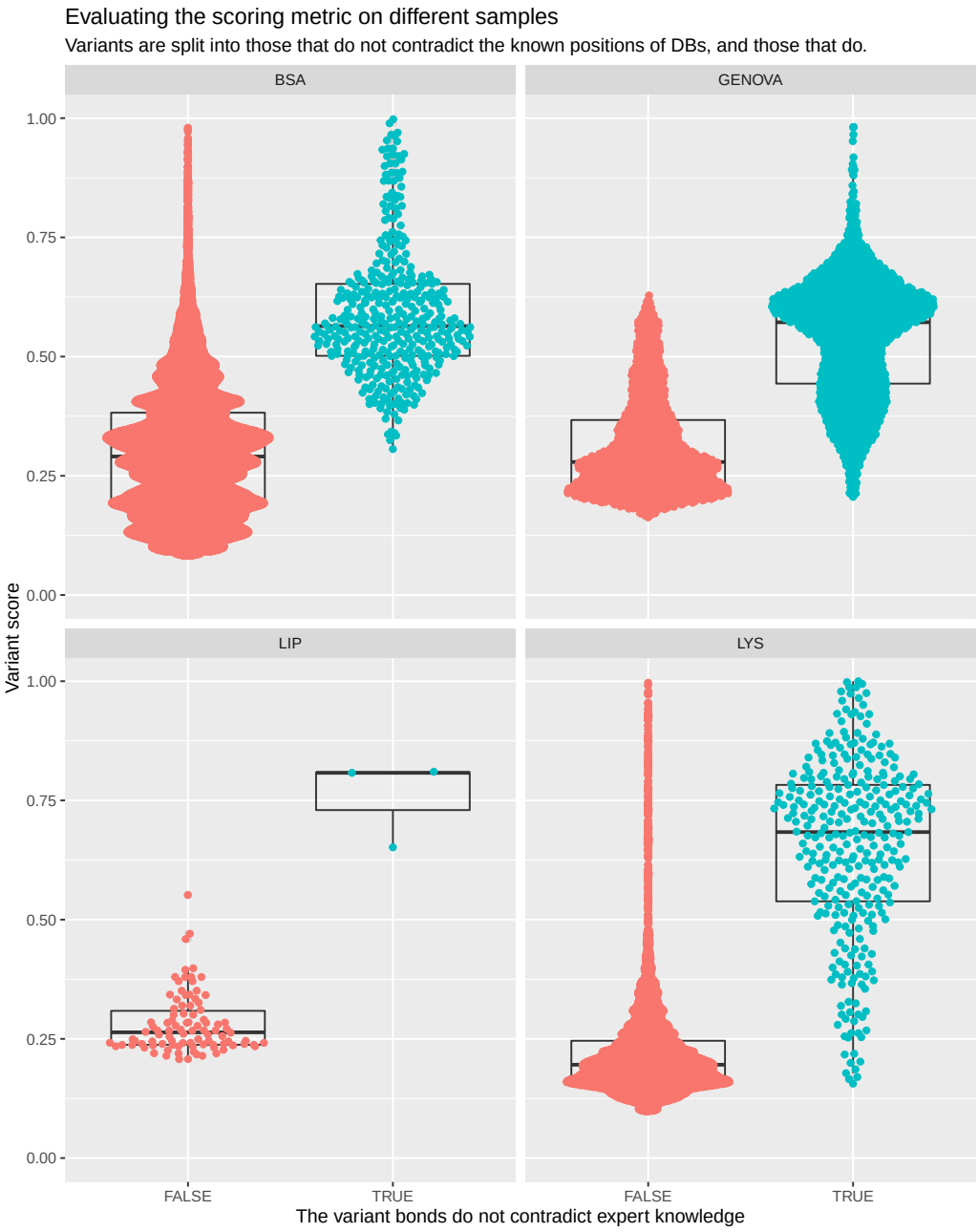


Figure 3.1 Scoring metric evaluation on datasets from four different samples. Individual datapoints represent different assigned variants, that are grouped based on whether they contradict expert knowledge about DB positions. The boxplots illustrate that the separation is usually relatively good, but as we can see from the datapoints, there are still a lot of (in theory) nonexsting variants with a high score. This can partly be attributed to DB scrambling. The lack of data in the LIP sample is explained in the text. (BSA = bovine serum albumin, LYS = lysosyme, LIP = lipase, GENOVA = an in-silico generated ovalbumin dataset)

seen on the following figures: Figure 3.2 (lysosyme, LYS), Figure 3.3 (lipase, LIP), and Figure 3.4 (in-silico generated ovalbumin, GENOVA). The top row shows data from RAT samples, the bottom shows data from AT samples, the left column shows the theoretical positions of the cysteine alkylations and disulfide bonds, and the right column shows the evidence we have gathered from the assigned fragments and variants. Alkylation evidence is illustrated by a border around the cysteine. Only precursor and fragment assignments that had an error of 10ppm or below were considered; we also allowed at most 2 dissociation events during fragmentation, and at most three segments in the precursor.

Lysosyme From almost 14,000 individual spectra, 1,668 have been assigned at least one precursor. Out of the 2,219 total assigned precursors, 1,985 have generated a variant in which at least one fragment matched a peak in the spectra. Most of generated variants have in some shape or form diverged from the expert knowledge of where the bons and alkylations should be present. However, thanks to the use of our scoring metric, we have successfully identified three of the four lysosyme disulfide bonds. The cysteine pair from the last unidentified bond has been seen only in 7 fragments whose variants did not contradict our knowledge about bonds, compared for example to the 2,534 fragments of the bond $(5, 126)$. We are not sure what caused this disparity, and the causes should be more deeply investigated in the future.

Lipase As is evident from the metric evaluation plot (Figure 3.1), there has been a lot less data for lipase than for the other proteins. From almost 12,000 individual spectra, only 67 have been assigned a precursor, compared to over 2,000 in lysosyme. This major disparity inspired us to construct our own control dataset. We have found that our algorithms were very prone to producting false positives, and had produced close-to-zero false negatives. Based on this fact we have concluded that a part of the lipase data has been lost, or that the data have been damaged in some uknown way. Despite these problems, Dibby successfully identifies one out of three disulfide bonds that are present in the protein.

To conclude, DB scrambling during sample preparation has proven to be a challenge, as did some "hotspot" cysteines that generated a lot of false signals. Despite these obstacles, we have successfully identified the majority of disulfide bonds in lysosyme, all of the bonds in our generated control, and one third of the bonds in lipase. The visualisation requires rather heavy manual interpretation, and the differential analysis step is still not automated, these two things could be made better.

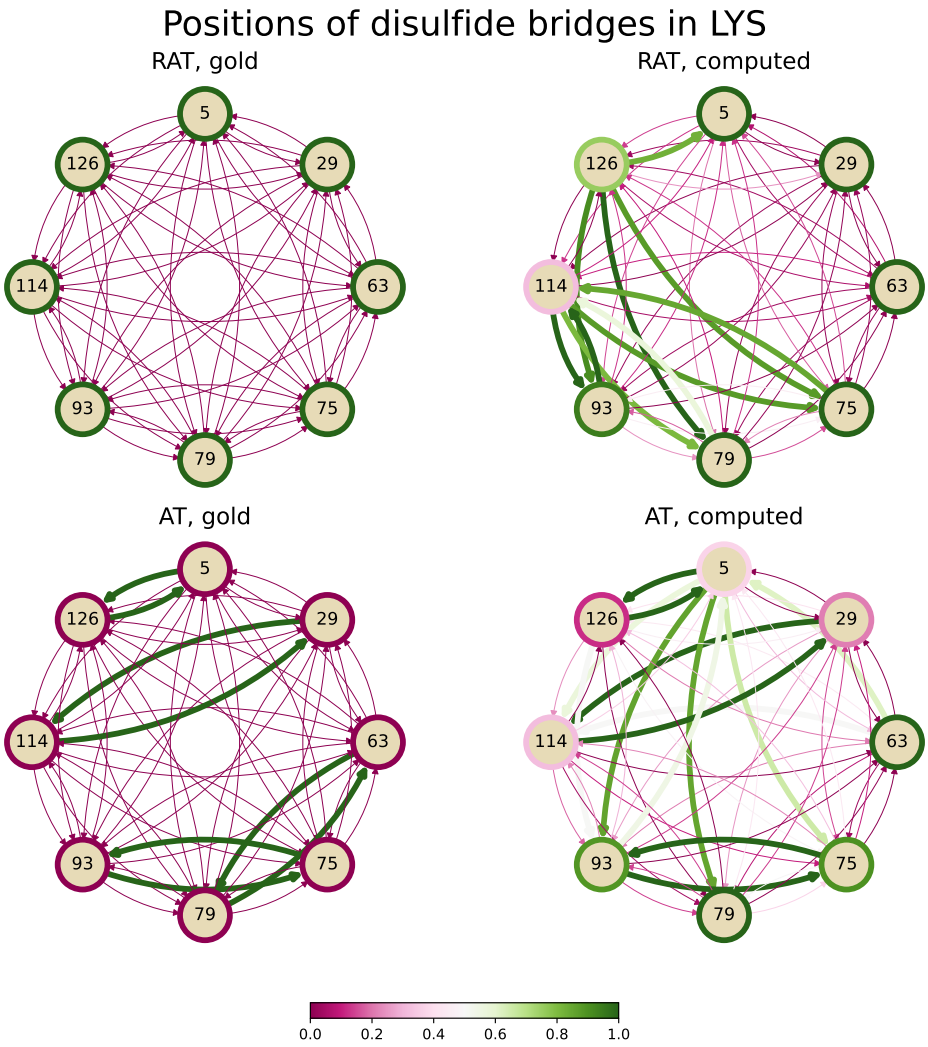


Figure 3.2 The evidence for positions of DBs and alkylations in lysosyme, weighted by variant score, and normalised. There are quite a few false positives, though not as many as to overpower all the true evidence. The bonds $(5, 126)$ and $(29, 114)$ are quite clearly visible. We can safely ignore the other directed edges stemming from 5, because their scores are relatively low compared to $(5, 126)$, and because they are not bidirectional — that means that the supposed bond-partners are seen more commonly in other configurations. The presence of the bond $(75, 93)$ is not so clear-cut — we also have strong signals about the cysteines being alkylated. This varrants further investigation. Finally, the evidence for bond $(63, 79)$ is practically not present, and both cysteines were deemed as alkylated.

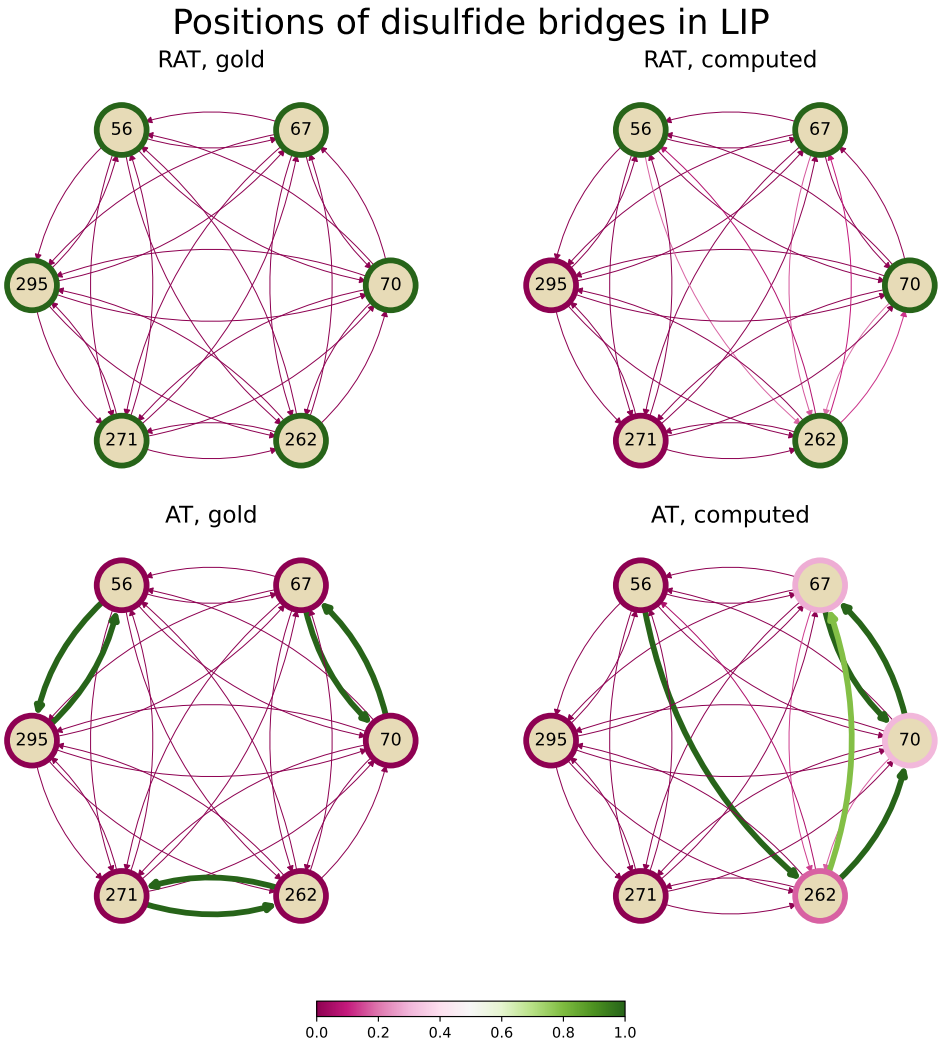


Figure 3.3 The evidence for positions of DBs and alkylations in in-silico generated ovalbumin, weighted by variant score, and normalised. No fragments containing cysteines 271 and 295 have been assigned (neither in alkylated form, nor as a part of any disulfide bond); this leads us to believe that part of the data has been lost, or otherwise damaged.

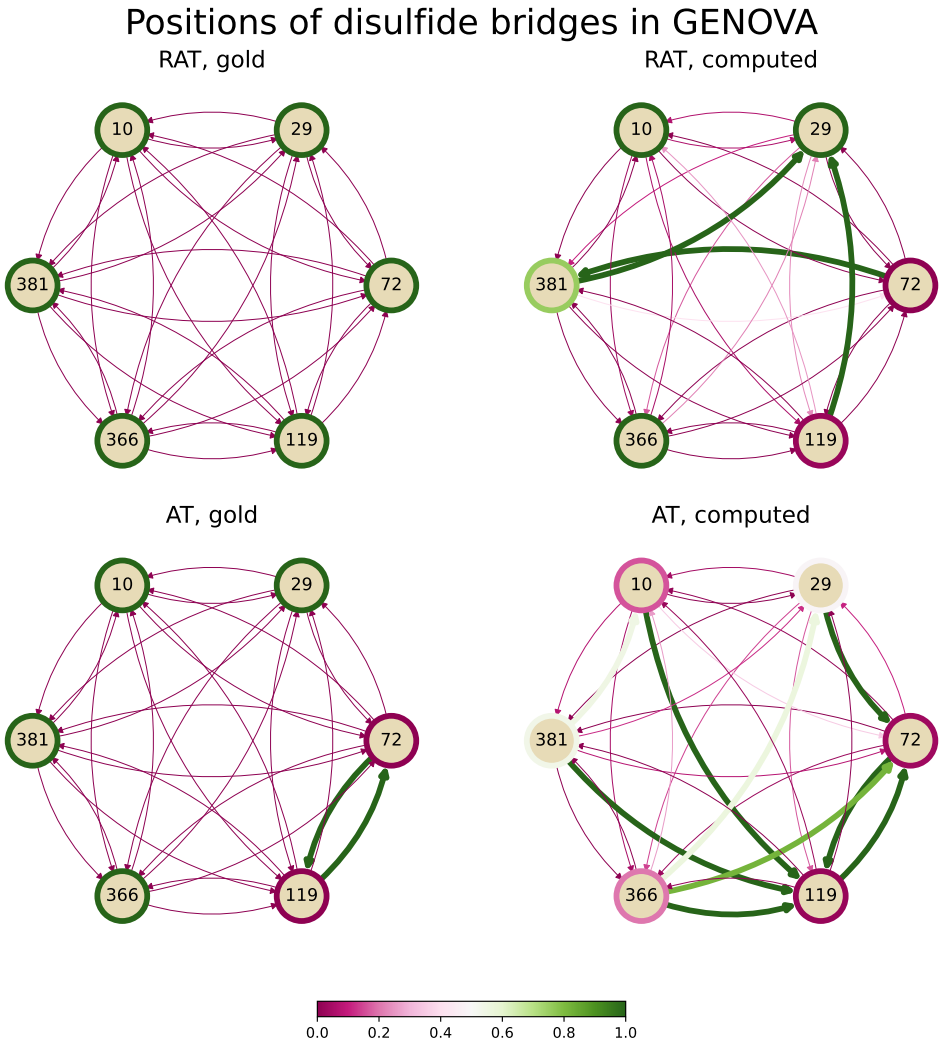


Figure 3.4 The evidence for positions of DBs and alkylations in lipase, weighted by variant score, and normalised. Although there are some false positives, leading mainly to cysteine 119, the only confirmed bidirectional bond is (72, 119).

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?)

  This is quite common.
- 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'.)

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.

[56] Jude C Lakbub, Joshua T Shipman, and Heather Desaire. "Recent mass spectrometry-based techniques and considerations for disulfide bond characterization in proteins". In: *Analytical and bioanalytical chemistry* 410.10 (2018), pp. 2467–2484.

[57] Shiaw-Lin Wu et al. "Mass spectrometric determination of disulfide linkages in recombinant therapeutic proteins using online LC- MS with electron-transfer dissociation". In: *Analytical chemistry* 81.1 (2009), pp. 112–122.

[58] Diogo B Lima et al. "SIM-XL: A powerful and user-friendly tool for peptide cross-linking analysis". In: *Journal of proteomics* 129 (2015), pp. 51–55.

[59] Chuanlong Cui et al. "Comprehensive identification of protein disulfide bonds with pepsin/trypsin digestion, Orbitrap HCD and Spectrum Identification Machine". In: *Journal of proteomics* 198 (2019), pp. 78–86.

[60] Amadeu H Iglesias, Luiz Fernando A Santos, and Fabio C Gozzo. "Identification of cross-linked peptides by high-resolution precursor ion scan". In: *Analytical chemistry* 82.3 (2010), pp. 909–916.

[61] Bo Wei et al. “Identification of sulfenylated cysteines in Arabidopsis thaliana proteins using a disulfide-linked peptide reporter”. In: *Frontiers in plant science* 11 (2020), p. 777.

[62] Fan Liu, Bas van Breukelen, and Albert JR Heck. “Facilitating protein disulfide mapping by a combination of pepsin digestion, electron transfer higher energy dissociation (EThcD), and a dedicated search algorithm SlinkS”. In: *Molecular & Cellular Proteomics* 13.10 (2014), pp. 2776–2786.

[63] Amir Abboud and Kevin Lewi. “Exact weight subgraphs and the k-sum conjecture”. In: *International Colloquium on Automata, Languages, and Programming*. Springer. 2013, pp. 1–12.

[64] Dorit S Hochbaum and Anu Pathria. “Node-optimal connected k-subgraphs”. In: *manuscript, UC Berkeley* (1994).

[65] Kazimierz Kuratowski. “Sur le probleme des courbes gauches en topologie”. In: *Fundamenta mathematicae* 15.1 (1930), pp. 271–283.

[66] Douglas R Smith. “The design of divide and conquer algorithms”. In: *Science of Computer Programming* 5 (1985), pp. 37–58.

[67] Stephen Boyd and Jacob Mattingley. “Branch and bound methods”. In: *Notes for EE364b, Stanford University* (2007), pp. 2006–07.