Improving Spatial Resolution of and Error Estimation for Radical Probe Mass Spectrometry

by

XiaoFei Zhao

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2015

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The function of a protein depends on the structure of the protein. A commonly used analytical technique for studying protein structure is radical-probe mass spectrometry (RP-MS). RP-MS oxidizes a protein of interest then quantitates the oxidation on the protein. Such quantitations can probe the solvent-accessible surface area (SASA) of the protein. This SASA can be used for studying the structure of the protein. Thus, the spatial resolution of such quantitations of oxidation is the spatial resolution at which protein folding can be studied. This thesis proposes a computational method for increasing, by many times, the spatial resolution of such quantitations of oxidation. Traditional RP-MS can already quantitate the oxidation on a peptide of a protein. MS/MS, which is also known as tandem mass spectrometry, is a technique in analytical chemistry. MS/MS can fragment a peptide into the suffixes of this peptide. Thus, the fraction of such individual suffixes of length i that are oxidized is the relative frequency that one of the last i residues of this peptide is oxidized. Thus, two such suffixes of lengths i and j, where $i > j$, correspond to two such frequencies. Thus, the difference between these two frequencies is the frequency that the oxidation on this peptide is inclusively between the i^{th}-last and $(j+1)^{\text{th}}$-last residues of this peptide. The oxidation between these two residues is used by our computational method to quantitate oxidation at subpeptide level. Such quantitated oxidation extents match the previously published oxidation rates and are computed from an MS/MS dataset. The MS/MS dataset is produced by a specially designed RP-MS experiment. This RP-MS experiment used MS/MS that targeted six tryptic peptides of apomyoglobin (PDB `1WLA`).

However, such quantitations of oxidation are not precise, mostly because random errors exist in such fraction of the suffixes that are oxidized. Such a fraction is a type of peak-area fraction. A peak-area fraction represents, in a sample, the quantity of a type of molecule relative to another type of molecule. To estimate random errors in a peak-area fraction, we made three reasonable assumptions partially justified in the literature. From these assumptions, we mathematically deduced our empirical formula. Our empirical formula estimates random errors in a peak-area fraction that is observed in only one run of mass spectrometry. Such estimated random errors match the empirically observed random errors in a test dataset. The test dataset is generated by three almost repeated runs of MS/MS. To generate the test dataset and the MS/MS dataset, the same instrument analyzed, with similar configurations, two similar samples. Thus, our empirical formula is used for estimating random errors in such a quantitation of oxidation in the MS/MS dataset.

MS^{E} is a technique in analytical chemistry. MS^{E} is similar to MS/MS. However, the throughput of MS/MS is lower than the throughput of MS^{E} by orders of magnitude.

Unfortunately, we showed that, currently, MS^{E} almost certainly cannot improve the spatial resolution of RP-MS presumably because MS^{E} generates too much noise.

Acknowledgements

I would like to first express my sincere gratitude to my supervisor, Professor Bin Ma, for finding such an interesting interdisciplinary research project for me, for guiding me through my project, and especially for improving my scientific-writing skill.

I would like to express my gratitude to Professor Lars Konermann for improving my knowledge in Chemistry.

I would like to express my gratitude to Siavash Vahidi for performing some important wet laboratories, because the foundation of my project is based on these wet laboratories.

I would like to thank my committee members, Professor Ming Li and Professor Forbes Burkowski, for reading and examining this thesis.

I would like to thank all other members of the Bioinformatics research group at the University of Waterloo (Dr Hao Lin, Dr Lin He, Dr XueFeng Cui, Dr Laleh Soltan Ghoraie, ShiWei Li, XiaoBo Li, and Eric Marinier) for having some interesting discussions about my project.

I would like to thank the University of Waterloo for providing a world-class graduate-level education to me.

Dedication

This thesis is dedicated to my parents for the unconditional love that they gave to me.

Table of Contents

List of Tables x

List of Figures xi

1 Introduction 1

2 Fundamentals of protein folding 6

3 Fundamentals of mass spectrometry (MS) 10

3.1 Fast photochemical oxidation of protein (FPOP) 11

3.2 Enzymatic proteolysis 14

3.3 High-performance liquid chromatography (HPLC) 14

3.4 Mass spectrometry (MS) 16

3.4.1 Ion source 17

3.4.2 Mass analyzer 18

3.4.3 Mass detector 20

3.5 MS/MS 20

3.6 Preprocessing of mass spectra 23

3.7 Peptide-spectrum match (PSM) 26

3.8 Mass spectrometry (MS) protocols 27

4 Quantitating mono-oxidation at subpeptide level 32

4.1 Motivation . . . 33

4.2 Related works . . . 33

4.3 The MS/MS dataset . . . 34

4.4 Methods . . . 36

4.5 Results on the MS/MS dataset . . . 40

4.6 Discussion . . . 45

5 Estimating the random error in a peak-area fraction given only one run 46

5.1 Motivation . . . 48

5.2 Related works . . . 48

5.3 Deriving our empirical formula . . . 49

5.3.1 Making and justifying assumptions . . . 50

5.3.2 Estimating the square of the multiplicative-random-error constant . 51

5.3.3 Mathematically deducing our empirical formula from the assumptions 52

5.3.4 Simplifying our empirical formula for use in Chapter 4 . . . 53

5.4 Testing our empirical formula . . . 54

5.4.1 Test dataset . . . 54

5.4.2 Test method . . . 55

5.4.3 Test result . . . 60

5.5 Discussion . . . 61

6 Caveats about using MS^{E} for RP-MS 63

6.1 Background of MS^{E} . . . 63

6.2 The MS^{E} dataset . . . 64

6.3 A lower bound on the interference-to-signal ratios in the MS^{E} dataset . . . 65

6.4 Negative results on the MS^{E} dataset . . . 67

6.5 Discussion about the negative results . . . 67

7 Conclusion **68**

APPENDICES **70**

Definitions specific to this thesis 72

Definitions in Mathematics 73

Definitions in mass spectrometry 73

Acronyms in mass spectrometry 74

References **76**

List of Tables

3.1 The initial rates of the second-order reaction of free amino acids with hydroxyl radical (HO·) at pH=7 and the common mass shifts produced by such reaction. 13

3.2 Some properties of the commonly observed amino acids [1]. 25

4.1 A summary of the MS^1 spectra in the MS/MS dataset. 35

5.1 Important information extracted from the test dataset. 54

List of Figures

1.1 An illustration of the concept of SASA. 2

1.2 A heatmap showing both the unoxidized form of and the mono-oxidized forms of `VEADIAGHGQEVLIR` in MS^1. 3

1.3 A mixture MS^2 spectrum of mono-oxidized `VEADIAGHGQEVLIR`. 4

1.4 A zoomed-in region of Figure 1.3. 4

2.1 An illustration of the landscape of protein structures. 8

3.1 An overview of fast photochemical oxidation of protein (FPOP) [19]. . . . 11

3.2 A mechanism of histidine oxidation [24]. 13

3.3 An example of proteolysis 14

3.4 A schematic of reverse-phase high-performance liquid chromatography (HPLC) (RP-HPLC) 16

3.5 A schematic of a typical mass spectrometer [15]. 17

3.6 Definitions of ΔM, the slight difference between the respective m/z of two adjacent peaks [15]. 19

3.7 A schematic of MS/MS [15]. 21

3.8 Peptide fragmentation in MS^2 [4]. 22

3.9 An interpretation of an MS^2 spectrum by peptide-spectrum match (PSM). 28

3.10 Some workflows of protein mass spectrometry (MS). 30

3.11 A schematic of RP-MS [2]. 31

4.1 The graphical abstract of Chapter 4. 32

4.2 The FASTA sequence of apomyoglobin (PDB `1WLA:A`). 34

4.3 A schematic of MS^1-based quantitation of oxidation at peptide level. 36

4.4 A mixture MS^2 spectrum of `VEADIAGHGQEVLIR` in the MS/MS dataset. . . 38

4.5 The estimated relative frequency that $\mathtt{y}_i$ is mono-oxidized as a function of i. 42

4.6 The relative frequency that a residue is mono-oxidized as a function of the position of this residue. 44

5.1 The graphical abstract of Chapter 5. 46

5.2 A heatmap generated by Algorithm 2. 57

5.3 A heatmap generated by Algorithm 2. 58

5.4 A heatmap generated by Algorithm 2. 59

5.5 The Q-Q plots generated by Algorithm 2. 61

6.1 A pair of consecutive mass spectra in the MS^{E} dataset. 65

Chapter 1

Introduction

Proteins are essential for every life on Earth because the majority of proteins have important biological functions. The structure of a protein is correlated with the function of this protein. The failure of a protein to assume its intended structure can result in diseases, such as Alzheimer's disease and cancer [39, 36]. This thesis proposes a new computational method to derive information for studying protein structure from a specially designed experiment.

Solvent-accessible surface area (SASA) of a protein is the surface area of this protein that is accessible to a solvent such as water. SASA helps for studying protein structure because SASA reduces the number of plausible protein structures to explore. Higher spatial resolution of SASA further reduces such number of plausible protein structures. Figure 1.1 shows the concept of SASA. Fast photochemical oxidation of protein (FPOP), a special experiment for studying protein structure, can probe the SASA of a protein. FPOP oxidizes residues on a protein such that the extent of oxidation on a residue is positively correlated with the solvent accessibility of this residue. Thus, the pattern of oxidation on a protein is correlated with the SASA of this protein. Thus, the spatial resolution of such pattern of oxidation determines the spatial resolution of such SASA. Mono-oxidation is observed as a mass-shift of around +16Da or more precisely around +15.99Da. FPOP can be tuned so that the oxidation caused by such FPOP mainly consists of mono-oxidation [17].

FPOP can oxidize a protein at a precise time, such as a few milliseconds, after the protein starts to fold. Moreover, FPOP is neither labor-intensive nor time-consuming.

Without loss of generality, let us assume that the sequence of our peptide of interest is `VEADIAGHGQEVLIR`. Denote the unoxidized form of `VEADIAGHGQEVLIR` by `(VEADIAGHGQEVLIR)`. Denote all mono-oxidized forms of `VEADIAGHGQEVLIR` by `(VEADIAGHGQEVLIR)(+16)`. The

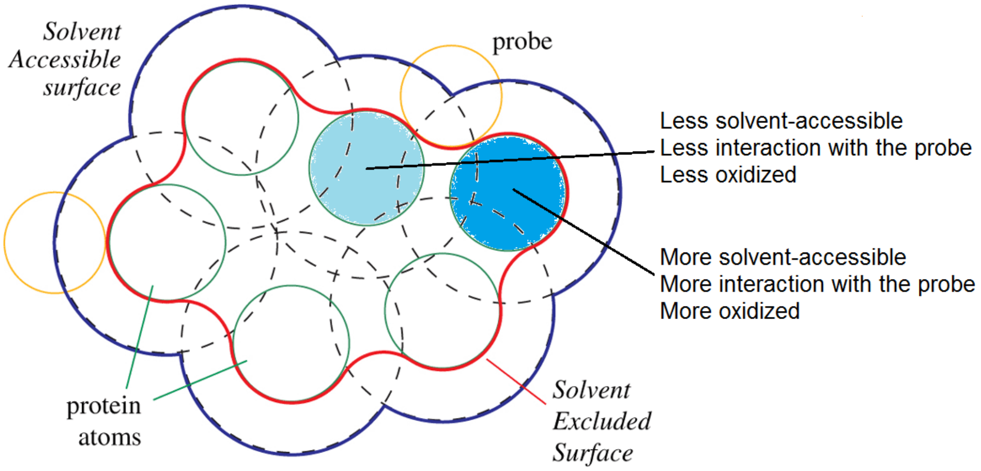


Figure 1.1: An illustration of the concept of solvent-accessible surface area (SASA). SASA changes while the protein folds. Fast photochemical oxidation of protein (FPOP) can detect such change that other methods, such as X-ray crystallography, cannot detect.

oxidation site of `(VEADIAGHGQEVLIR)(+16)` can be located on any of its residues. For example, `(VEADIAGHGQEVLIR)(+16)` can be any of the following: `(V)(+16)EADIAGHGQEVLIR`, `V(E)(+16)ADIAGHGQEVLIR`, `VE(A)(+16)DIAGHGQEVLIR`, etc. LC-MS is an analytical technique that first separates a mixture of analytes by retention time (RT) and then analyzes each analyte by mass spectrometry (MS). Figure 1.2 shows both the unoxidized form and the mono-oxidized forms of `VEADIAGHGQEVLIR` detected in one run of LC-MS. In such a run, the relative frequency that `VEADIAGHGQEVLIR` as a whole is oxidized can be estimated to be the quantity of `(VEADIAGHGQEVLIR)(+16)` divided by the quantity of both `(VEADIAGHGQEVLIR)` and `(VEADIAGHGQEVLIR)(+16)`. Spatially more granular quantitation of mono-oxidation results in higher spatial resolution of the SASA derived from such quantitation of oxidation. Thus, ideally, researchers would like to quantitate the extent of oxidation on each residue of `VEADIAGHGQEVLIR`. Equivalently, researchers would like to quantitate each form of mono-oxidized `VEADIAGHGQEVLIR`. Unfortunately, current technologies and methods can quantitate the following at best: the mixture of all mono-oxidized forms of `VEADIAGHGQEVLIR` as a whole relative to the unoxidized form of `VEADIAGHGQEVLIR`. Such quantitation is qualified to be at peptide level. Any quantitation that is spatially more granular than such peptide-level quantitation is qualified to be at subpeptide level. Quantitation of oxidation at subpeptide level is challenging. My thesis proposes a method for quantitating oxidation at subpeptide level using MS/MS.

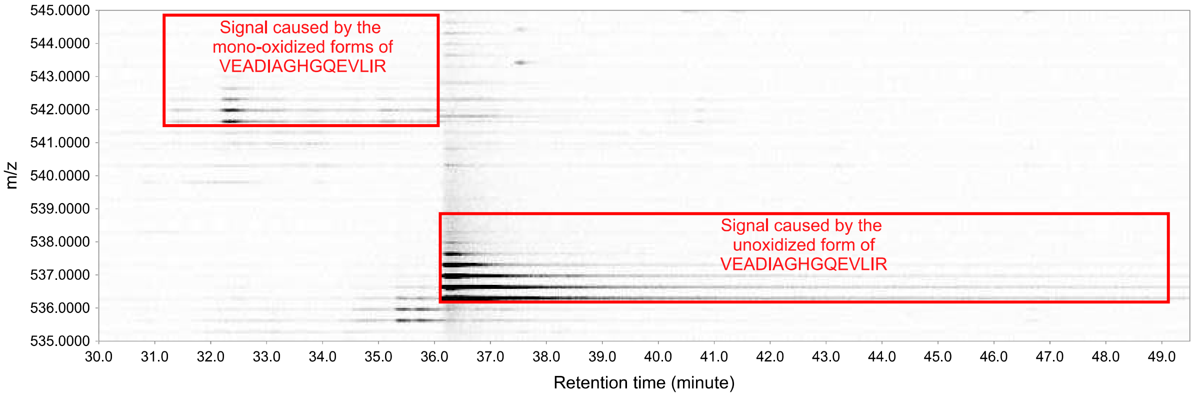


Figure 1.2: A heatmap showing both the unoxidized form of and the mono-oxidized forms of `VEADIAGHGQEVLIR` in MS^1. The unoxidized form and the mono-oxidized forms both have a charge state of 3 and are thus approximately $(15.99 \div 3)$Da apart in m/z.

MS/MS is a commonly used technology in MS. MS/MS can fragment `VEADIAGHGQEVLIR` into the suffixes of `VEADIAGHGQEVLIR` such as `R`, `IR`, `LIR`, etc. `R` is referred to as $\mathtt{y}_1$ ion, `IR` is referred to as $\mathtt{y}_2$ ion, `LIR` is referred to as $\mathtt{y}_3$ ion, etc. Each of these y-ions will form peaks in an MS^2 spectrum at its corresponding mass-to-charge ratio (m/z) (Figure 1.3). When an amino acid is mono-oxidized, the mass of the y-ion containing the mono-oxidized amino acid is shifted by approximately +16Da (Figure 1.3). Thus, the m/z of this y-ion will be shifted by approximately $+\frac{16}{z}$, where z is the charge state of this y-ion. Figure 1.4 illustrates a local region of the MS^2 spectrum of a mono-oxidized peptide. However, the mono-oxidized peptide is a mixture of different forms because different amino acids can be oxidized. Our task is to derive the proportion of each of these forms. Denote unoxidized $\mathtt{y}_i$ by $\mathtt{y}_i$ and mono-oxidized $\mathtt{y}_i$ by $\mathtt{y}'_i$. Let y_i be the quantity of $\mathtt{y}_i$, and let y'_i be the quantity of $\mathtt{y}'_i$. Let $\phi_i = \frac{y'_i}{y_i+y'_i}$. Then, ϕ_i is the proportion of the mono-oxidized forms that have the oxidation on the last i amino acids. Thus, $\phi_i - \phi_{i-1}$ is the proportion of these forms that have the oxidation on the i^{th}-last amino acid. In general, $\phi_i - \phi_j$ is the proportion of these forms that have the oxidation inclusively between the i^{th}-last and $(j+1)^{\text{th}}$-last amino acids. Chapter 4 of this thesis presents a novel algorithm that uses such $\phi_i - \phi_j$ to quantitate oxidation at subpeptide level.

However, ϕ_i and ϕ_j are subject to random errors due to the stochastic nature of a run of MS. Chapter 5 of this thesis presents a novel empirical formula for characterizing such random errors with fewer-than-expected amount of experimental data.

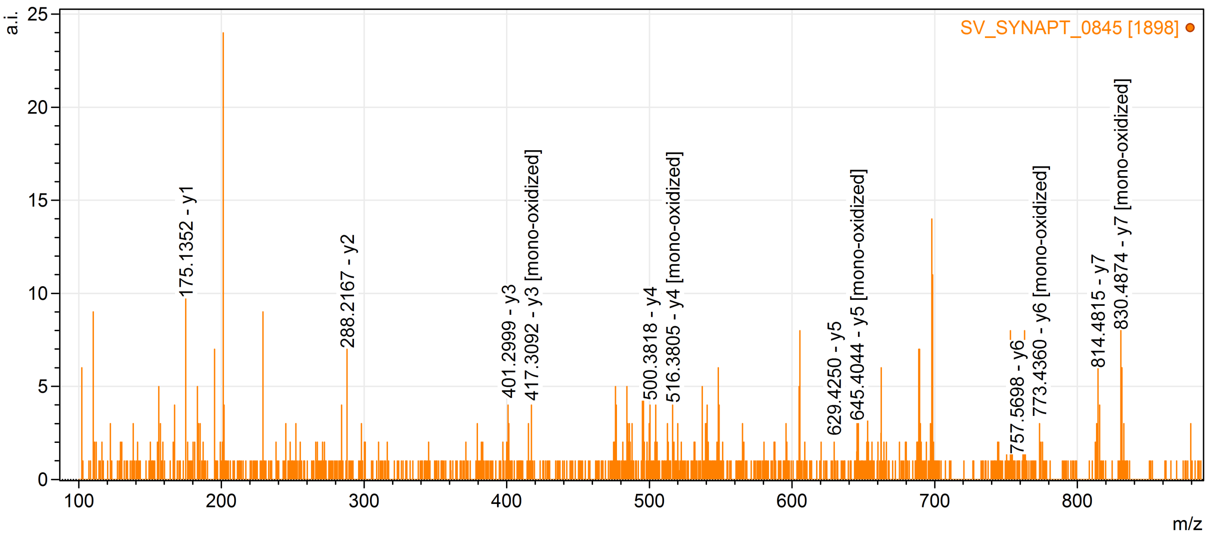


Figure 1.3: A mixture MS^2 spectrum of mono-oxidized `VEADIAGHGQEVLIR`. The vertical axis represents absolute intensity.

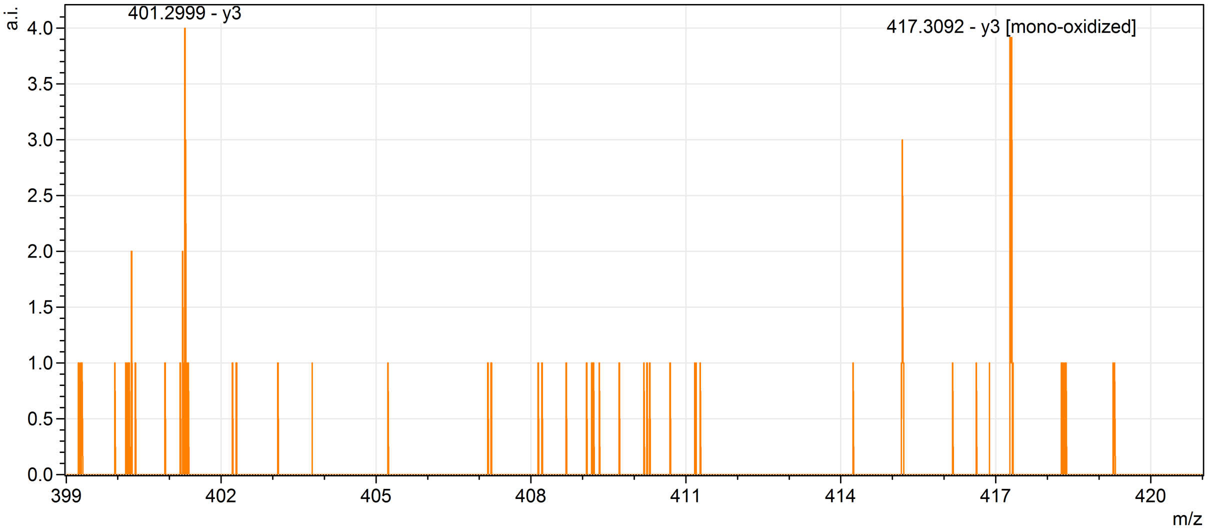


Figure 1.4: A zoomed-in region of Figure 1.3.

The throughput of MS^E is higher than the throughput of MS/MS by orders of magnitude. Unfortunately, Chapter 6 of this thesis shows that currently MS^E almost certainly cannot probe SASA at subpeptide level.

Chapter 2

Fundamentals of protein folding

A peptide is a sequence of n amino-acid residues chained together by $n-1$ peptide bonds, where $n \geq 2$. In this thesis, "residue" refers to only "amino-acid residue" unless explicitly stated otherwise. A polypeptide is a relatively long peptide. A typical polypeptide is composed of the 20 standard amino-acid residues. Different sequences of these 20 residues correspond to different polypeptides. Thus, a lot of different polypeptides can exist, because there are 20^n distinct sequences that have a length of n. Although the number of polypeptides in a typical biological organism is much less than 20^n, this number is still huge. A protein is an assembly of at least one polypeptide. Protein constitutes the building block of life. The metabolism of every biological organism requires numerous proteins.

The structure of a protein is strongly correlated with the function of this protein. Thus, determination of protein structure is a fundamental problem in life science. Protein structure can be observed at different levels of organization. Four levels of organization correspond to the following four levels of protein structures: primary, secondary, tertiary, and quaternary structures. The primary structure of a protein is defined as the sequence of the constituent residues of this protein. The covalent bonds in a protein fully determine the primary structure of this protein. Protein secondary structure is the general three-dimensional form of local segments of proteins. The secondary structure of a protein is defined by the patterns of hydrogen bonds between amine hydrogen and carbonyl oxygen atoms contained in the backbone peptide bonds of this protein. Alpha helices and beta strands are the most common protein secondary structures. Random coil is defined as the lack of any secondary structure. The tertiary structure of a protein is defined as the three-dimensional shape of this protein. Protein quaternary structure is the arrangement of more than one protein molecule in a multi-subunit complex. A single polypeptide chain is a protein if it can function on its own. Many proteins are comprised of several polypeptide chains. These polypeptide chains are referred to as protein subunits.

Determination of the primary structure of a protein is very easy. The state-of-the-art predictors of protein secondary structure achieve an accuracy of approximately 90% [25]. Moreover, a lot of experimental methods such as nuclear magnetic resonance (NMR) can determine protein secondary structure. Thus, determination of the secondary structure of a protein is easy. Experimental methods for determining the tertiary structure of a protein, such as X-ray crystallography, are labor-intensive and time-consuming. The prediction of protein tertiary structure without the use of any experimental data is NP-hard and thus is computationally hard. The prediction of protein tertiary structure with the use of some experimental data is ineffective for certain proteins. Thus, determination of the tertiary structure of a protein is hard. It is very hard to determine the quaternary structure of a protein. Nowadays, determination of the tertiary structure of a protein is still a major challenge. In this thesis, "protein structure" refers to only "tertiary protein structure" unless explicitly stated otherwise.

Protein structure is not static. The structure of a protein depends on the physiological environment that surrounds this protein. For example, in an acidic solution at pH$\approx$2, a protein usually does not assume any shape at all. Thus, every residue of this protein is exposed to this acidic solution, and this protein is referred to as denatured. Denaturation is defined as the process by which a macromolecule loses its quaternary structure, tertiary structure, and then secondary structure. We can view the structure of a protein as a point on a high-dimensional energy landscape (Figure 2.1). The energy landscape is a mapping of all possible conformations of the protein to the potential energies of these conformations. On this landscape, the altitude of a coordinate represents the energy of a conformation. Every protein tends to adopt a low-energy conformation. This tendency is consistent with the second law of thermodynamics because high entropy is associated with low-energy conformation. A change in the coordinate of a point on the energy landscape corresponds to a change in the conformation of a protein. A path from one coordinate to another coordinate on the energy landscape corresponds to a transition from one conformation to another conformation. When a new protein is just synthesized, this new protein usually does not have its intended structure and thus is usually not yet functional. Then, the structure of this new protein changes until this structure stabilizes at a local minimum on the energy landscape. Then, this protein is able to perform its intended function due to its stable structure and is thus functional. The process by which a protein becomes functional by assuming its intended structure is referred to as protein folding. This intended structure is referred to as its native structure.

A few methods have been developed for studying protein structures. Unfortunately,

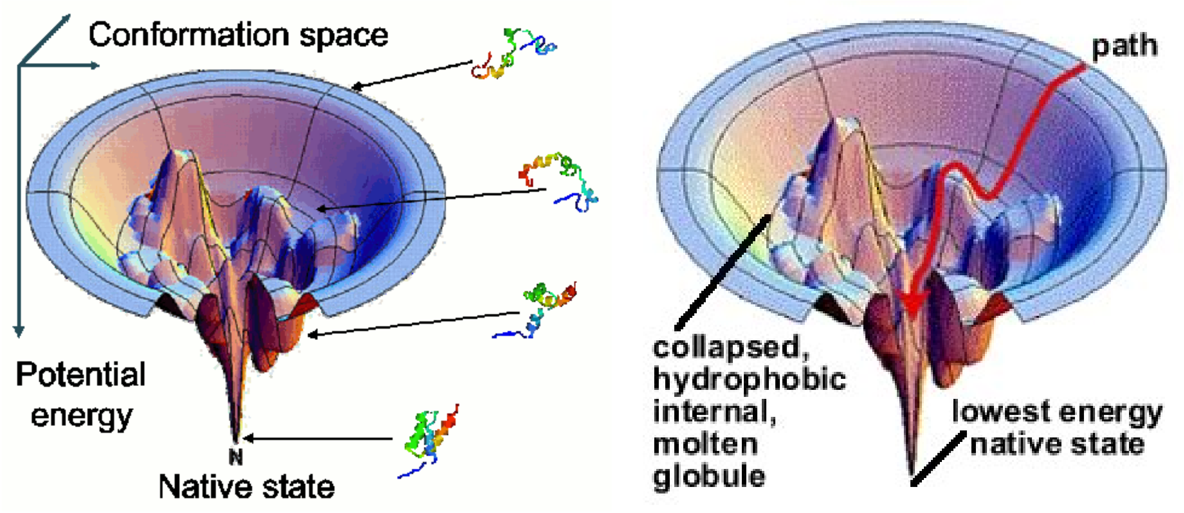

Conformation space
Potential energy
N
Native state
path
collapsed, hydrophobic internal, molten globule
lowest energy native state

Figure 2.1: An illustration of the landscape of protein structures. For a protein, the position of each atom relative to other atoms determines the conformation space. Thus, the dimension of the conformation space is much greater than two.

all these methods have some weaknesses. For example, X-ray crystallography is neither effective for membrane proteins nor effective for studying protein-folding dynamics; NMR is labor-intensive, time-consuming, and not effective for studying protein-folding dynamics.

The solvent-accessible surface area (SASA), also known as accessible surface area (ASA), of a biomolecule is the part of the surface area of this biomolecule such that this part can be accessed by the solvent in which this biomolecule is dissolved. In protein-related scientific fields, this biomolecule usually refers to a protein of interest and this solvent usually refers to water. Change in the SASA of a protein as a function of the time since the protein started to fold can reveal the following relationship: the protein surface that is exposed to water as a function of time. This relationship can then partially reveal the folding dynamics of the protein.

Any amino-acid residue of any protein can be covalently modified by hydroxyl radical (HO·). Let us suppose that a solution contains HO·. Let us assume that a treatment does not change the composition of the solution for a substantial period of time. If this solution covalently modifies this residue more heavily after this treatment, then this residue is more solvent-accessible after this treatment, and vice versa. Thus, change in the extent of this covalent modification on this residue is positively correlated with change in the solvent-accessibility of this residue. Thus, change in the extent of this covalent modification on every residue of a protein of interest can reveal change in SASA of this protein.

Radical-probe mass spectrometry (RP-MS) can estimate the SASA of a protein. An RP-MS experiment usually proceeds as follows: First, a protein of interest is tuned to be at a given stage of protein folding. Next, a source of energy such as ultraviolet light generates short-lived free radicals such as HO·. Then, these free radicals cause covalent modifications, which mainly consist of mono-oxidation, to the solvent-accessible surface of this protein of interest. Afterwards, a protease such as trypsin cleaves this protein of interest into shorter peptides which could have been modified by these free radicals. Finally, liquid chromatography (LC) elutes and thus separates the mixture of these shorter peptides. While LC is eluting these shorter peptides, mass spectrometry (MS) identifies these peptides and quantitates the extent of oxidation on each of these peptides. A protein of interest goes through different stages in the folding process of this protein. These different stages result in different extents of covalent modification on the residues of this protein of interest, respectively. Thus, changes in the SASA of this protein of interest across these different stages can be inferred, Thus, RP-MS can be used for studying protein-folding dynamics. Moreover, RP-MS is fast, cost-efficient, applicable to any protein, and able to detect a rapid change in protein structure. Unfortunately, a protease usually only cleaves a protein at a few backbones of this protein. Thus, a protease usually cleaves a protein into only a few long peptides. Thus, each of these long peptides is composed of several residues. Thus, the spatial resolution of RP-MS is limited to the peptide level that is determined by this protease.

Chapter 3

Fundamentals of mass spectrometry (MS)

This chapter presents the fundamentals of mass spectrometry (MS) and focuses on MS-related concepts. The order in which we present the sections in this chapter is approximately the order in which a typical radical-probe mass spectrometry (RP-MS) experiment is performed.

Section 3.1 presents fast photochemical oxidation of protein (FPOP), an analytical-chemistry technique for oxidizing a protein to be investigated. Section 3.2 presents proteolysis, a molecular-biology method often applied before performing high-performance liquid chromatography (HPLC). Section 3.2 presents HPLC, an analytical-chemistry technique for separating analytes based on some of their chemical properties. Section 3.4 presents MS, an analytical-chemistry technique for measuring the m/z of analytes. Section 3.5 presents MS/MS, a subclass of MS. In essence, MS/MS fragments analytes and then measures the m/z of each of these fragments. Section 3.6 presents the preprocessing of raw mass spectra. Examples of such preprocessing are centroiding, deconvolution, and deisotoping. Such processing facilitates the subsequent interpretation of these preprocessed mass spectra. Section 3.7 presents peptide-spectrum match (PSM), a key concept in the interpretation of the MS^2 spectra that are produced by protein MS. Section 3.8 presents common protocols of protein MS and focuses on RP-MS.

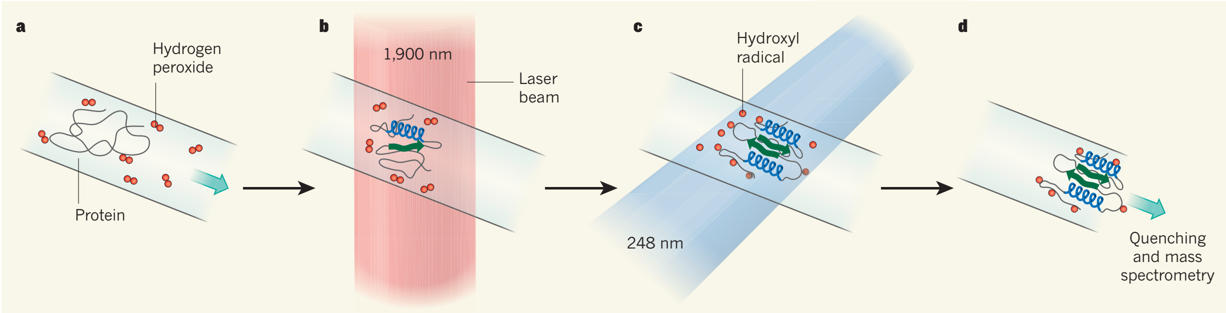


Figure 3.1: An overview of FPOP [19]. A protein experiences four stages while passing through a tube. (a) A denatured protein and hydrogen peroxides (HOOH) flow through a capillary tube. (b) An infrared laser having a wavelength of 1900 nanometers initiates the folding of the denatured protein. (c) After a brief delay, an ultraviolet laser having a wavelength of 248 nanometers splits HOOH into HO·. Some of these HO· almost instantaneously oxidize some residues of the partially folded protein. Longer delay causes the protein to be less denatured before HO· is generated, then this less denatured protein is subject to less modification by HO·. If a residue of the protein is closer to the solvent that dissolved the protein, then this residue is more heavily modified by HO·. The addition of red dots to the protein represents the addition of oxygen atoms to the protein. (d) The modified protein exits the capillary, a radical scavenger removes the remaining HOOH, and MS quantitates the HO·-mediated modification which is positively correlated with solvent-accessible surface area (SASA). Before MS, proteolysis and then liquid chromatography (LC) are often performed on the modified protein.

3.1 Fast photochemical oxidation of protein (FPOP)

FPOP is an analytical-chemistry technique for covalently labeling a protein. FPOP uses ultraviolet light to cause the dissociation of hydrogen peroxide (HOOH) and then the formation of hydroxyl radical (HO·), The action of such ultraviolet light is shown in the following chemical equation.

$$\text{HOOH} + \lightning \;\rightarrow\; \text{HO}\lightning\text{OH} \;\rightarrow\; 2\ \text{HO}\cdot \qquad (\lightning \text{ denotes 248 nm ultraviolet light})$$

The *in vivo* half-life of HO· is approximately 10^{-9}s, and HO· is highly reactive. Thus, HO· virtually damages all types of macromolecules such as carbohydrates, nucleic acids, lipids, and amino acids [37].

HO· can covalently modify a residue through different mechanisms. Thus, different mass shifts of the modified residue can be observed. The time scale of these covalent modifications is usually less than one millisecond. Thus, there is only a sub-millisecond interval between the time that HO· enters in contact with a residue and the time that HO· finishes covalently modifying this residue. While HO· is covalently modifying a protein, HO· may also affect the folding process of this protein. However, virtually all protein folding processes take more than one millisecond to initiate. Thus, the time scale of covalent modification mediated by HO· is short compared with the time scale of protein folding. Thus, before that HO· finishes covalently modifying a protein, the overall folding process of this protein is unlikely to be affected. In fact, HO· indeed does not substantially affect the folding process of a protein before it finishes modifying this protein [17].

Intuitively, if the solvent accessibility of a residue is high, then this residue is more likely to be covalently modified by HO·. Thus, the SASA of a residue of a protein is positively correlated with the extent of HO·-mediated covalent modification on this residue. Moreover, the duration of HO·-mediated covalent modification is short compared with the duration of protein folding. Thus, the HO·-mediated modification to any protein only depends on the structure of this protein at the precise time of this modification, Thus, after a protein refolds for a given amount of time, the extent of the HO·-mediated modification to each residue of this protein reveals the SASA of this protein after this amount of time. Thus, if the time taken for a protein to refold varies, then the extent of the HO·-mediated modification to any residue of this protein as a function of this time is positively correlated with the solvent-accessibility of this residue as a function of this time. Thus, we can characterize the change in the SASA of a protein as a function of time and thus study protein-folding dynamics.

Different residues have different mechanisms of reacting with HO·. Thus, different residues have hugely different reaction rates with HO·. For example, the second order reaction rate of HO· with cysteine, the most reactive one, is approximately 2000 times higher than such rate with glycine, the least reactive one [6]. However, HO· usually oxidizes a residue, and oxidation of a residue by HO· often adds one oxygen atom to this residue. Thus, mono-oxidation is the principal HO·-mediated covalent modification to all residues. For example, all residues can have a mass shift of +15.9949 or +31.9898 Da after reacting with HO· except glycine, which simply does not react with HO·. The mass shifts of +15.9949 and +31.9898 correspond to the addition of one and two oxygen atoms respectively [44]. Table 3.1 shows that reactions between different residues and HO· happen at different speeds. However, these reactions result in similar mass shifts to these different residues. We use +15.9949, +15.99, and +16 interchangeably to denote the mass shift caused by the addition of one oxygen atom.

Free amino acid		$k_{HO\cdot}$ ($M^{-1}s^{-1}$) [6]	Common mass shifts resulting from modification by HO· (Da) [44]				
Cys	(C)	3.5×10^{10}	$-15,9772$	+31.9898	+47.9847		
Trp	(W)	1.3×10^{10}	+3.9949	+15.9949	+31.9898	+47.9847	
Tyr	(Y)	1.3×10^{10}	+15.9949	+31.9898	+47.9847		
Met	(M)	8.5×10^{9}	−32.0085	+15.9949	+31.9898		
Phe	(F)	6.9×10^{9}	+15.9949	+31.9898	+47.9847		
His	(H)	4.8×10^{9}	−23.0160	−22.0320	−10.0320	+4.9789	+15.9949
Arg	(R)	3.5×10^{9}	−43.0534	+13.9793	+15.9949		
Ile	(I)	1.8×10^{9}	+13.9793	+15.9949			
Leu	(L)	1.7×10^{9}	+13.9793	+15.9949			
Val	(V)	8.5×10^{8}	+13.9793	+15.9949			
Pro	(P)	6.5×10^{8}	+13.9793	+15.9949			
Gln	(Q)	5.4×10^{8}	+13.9793	+15.9949			
Thr	(T)	5.1×10^{8}	−2.0157	+15.9949			
Lys	(K)	3.5×10^{8}	+13.9793	+15.9949			
Ser	(S)	3.2×10^{8}	−2.0157	+15.9949			
Glu	(E)	2.3×10^{8}	−30.0106	+13.9793	+15.9949		
Ala	(A)	7.7×10^{7}	+15.9949				
Asp	(D)	7.5×10^{7}	−30.0106	+15.9949			
Asn	(N)	4.9×10^{7}	+15.9949				
Gly	(G)	1.7×10^{7}	n.d.				

Table 3.1: The initial rates of the second-order reaction of free amino acids with HO· at pH=7 and the common mass shifts produced by such reaction. A free molecule is not covalently bound to any other molecule. If an amino acid residue is part of a given protein in a given solution, then the $k_{HO\cdot}$ of this residue depends on the position of this residue with respect to this given protein and on the properties of this given solution.

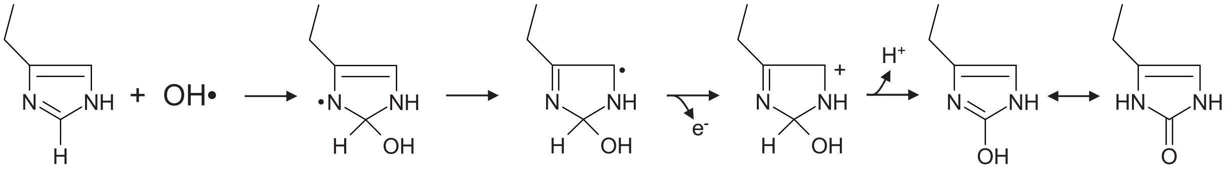


Figure 3.2: A mechanism of histidine oxidation [24].

Before any cleavage: K-A-F-A-R-W-A-R-P-K-P-R-E-Y-M-Q-F-P-W-P-Y-P
After trypsin cleavage: K|A-F-A-R|W-A-R-P-K-P-R|E-Y-M-Q-F-P-W-P-Y-P
After chymotrypsin cleavage: K|A-F|A-R|W|A-R-P-K-P-R|E-Y|M-Q-F-P-W-P-Y-P

Figure 3.3: An example of proteolysis by trypsin and then by high-specificity chymotrypsin. Unlike most proteases, chymotrypsin does not cleave any polypeptide until trypsin activates this chymotrypsin. Trypsin cleaves after any of {K, R} that is not before P. High-specificity chymotrypsin cleaves after any of {F, Y, W} that is not before P.

3.2 Enzymatic proteolysis

A protease (also called peptidase, proteinase, or proteolytic enzyme) is an enzyme that can perform proteolysis. Proteolysis is the cleavage a polypeptide into shorter peptides or amino acids. A specific protease only cleaves at specific peptide bonds in the backbone of a polypeptide. Thus, cleavage of a given polypeptide by a given specific protease produces a predictable set of peptides.

If a carbonyl-carbon is part of lysine (K) or arginine (R) and a nitrogen is not part of proline (P), then the protease trypsin cleaves at the peptide bond between this carbonyl-carbon and this nitrogen, and vice versa. Equivalently, the specificity rule of trypsin cleavage is referred to as follows: after K or R and not before P. Trypsin molecules can cleave each other because trypsin is also a type of protein. Thus, trypsin molecules are stored at below −20°C to prevent them from cleaving each other. Trypsin is the most commonly used protease for MS.

Different specific proteases are subject to different specificity rules. For example, LysN cleaves before K, LysC cleaves after K, GluC cleaves after E, AspN cleaves before D, high-specificity chymotrypsin cleaves after any of {F, Y, W} and not before P, and low-specificity chymotrypsin cleaves after any of {F, Y, W, M, L} and not before P. Different proteases usually cleave the same peptide independently of each other (Figure 3.3).

3.3 High-performance liquid chromatography (HPLC)

Liquid chromatography (LC) is an analytical-chemistry technique. LC separates a mixture into the components constituting this mixture based on the chemical properties of these components. In analytical chemistry, components, analytes, constituents, and substances are all equivalent in meaning. To separate a mixture into the components constituting this mixture, LC uses a column that elutes different components at respectively different speeds.

Mobile phase is defined as the solution that is gradually eluted by the column. Stationary phase is defined as the sorbent of the column which retains components in the solution. The retention time (RT) of a component is defined as the elapsed time during which this component is retained by the LC column, or equivalently the amount of time taken for this component to go through the LC column. The RT of multiple molecules may also refer to the shortest time interval that virtually includes the RT of all these molecules. The RT of a component depends on the interaction between this component and the stationary phase. This interaction depends on the chemistry of the stationary phase, the chemical properties of this component, and on the composition of the mobile phase. Clearly, the more a column retains a component, the higher the RT of this component will be in this column, and vice versa.

High-performance liquid chromatography (HPLC) is a subclass of LC. HPLC is characterized by the high pressure applied to the mobile phase. This high pressure, which is usually between 50 and 350 bars, reduces the RT of all components. Thus, compared with traditional LC, HPLC is characterized by higher resolving power and requires less time per run. LC resolving-power is defined as the ability to distinguish two components with slightly different RTs. Thus, HPLC has gradually replaced traditional LC. Figure 3.4 shows some key characteristics of reverse-phase HPLC (RP-HPLC), a subclass of HPLC. In RP-HPLC, the column preferentially retains hydrophobic components. Thus, in RP-HPLC, the RT of a hydrophobic component should be higher than the RT of a hydrophilic component. Moreover, exposure of the hydrophobic regions of a component depends on the size and shape of this component. Thus, in RP-HPLC, the size and shape of a component affect the RT of this component. Frequently, RP-HPLC and MS are used together.

The following two types of elution exist in LC: isocratic elution and gradient elution. The composition of the mobile phase is relatively constant during isocratic elution and changing during gradient elution. Usually, the variation in the RT of a component in isocratic elution is lower than the variation in RT of this component in gradient elution. Thus, the quantity of an eluted component as a function of RT forms a sharper peak in gradient elution compared with isocratic elution. In gradient elution, the mobile phase consists of mostly water at the beginning. As elution progresses, an organic solvent miscible with water is gradually added to the mobile phase. In the end, the mobile phase consists of mostly this organic solvent. Some commonly used organic solvents for gradient elution are acetonitrile, methanol, and tetrahydrofuran.

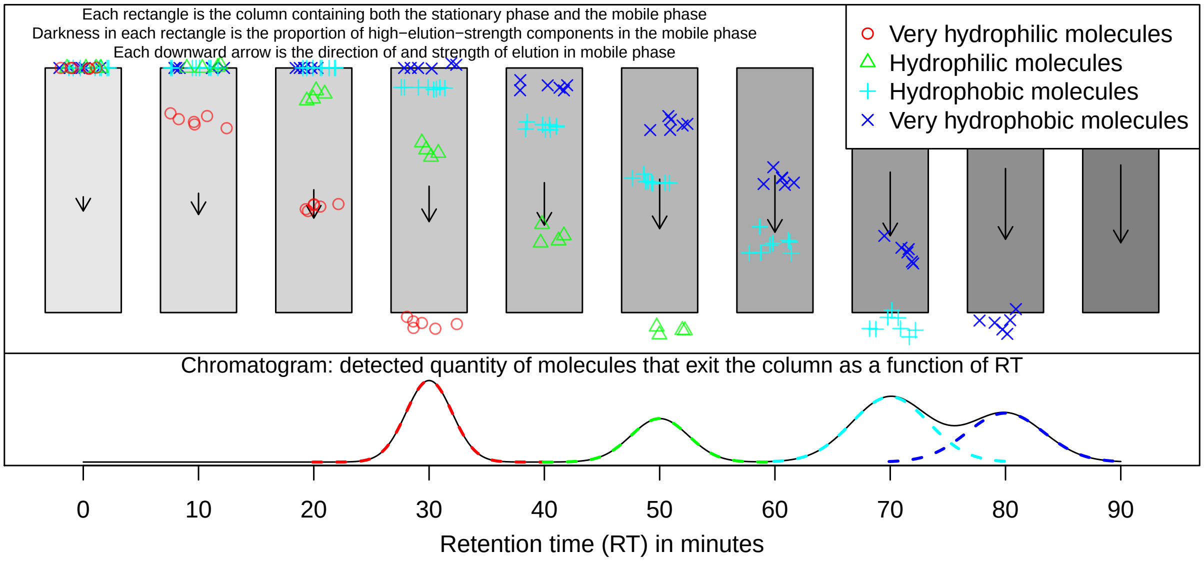


Figure 3.4: A schematic of RP-HPLC featuring a hypothetical RP-HPLC experiment. In this experiment, the RT of hydrophobic molecules and the RT of very hydrophobic molecules overlap. Thus, this experiment cannot separate these two types of molecules.

3.4 Mass spectrometry (MS)

MS is an analytical-chemistry technique based on the use of a mass spectrometer. A mass spectrometer takes as input some analytes and outputs the mass spectra of these analytes. A mass spectrum is a continuum of signal intensity as a function of m/z, where m/z denotes mass-to-charge ratio.

Mass spectra of analytes can reveal some properties of these analytes. Examples of these properties are chemical formula and structural formula. Moreover, mass spectra can distinguish between isotopes of a chemical element because isotopes have different masses.

The history of MS is relatively long. At the beginning of the 20th century, MS has already been used for separating isotopes. However, the development of protein MS, which is the MS for studying proteins, only started at the end of the 20th century. Large biomolecules such as proteins tend to fragment into small molecules after being ionized. Thus, ionization tends to destroy the structural formula of a protein. Thus, protein MS has been a major challenge. In 1984, Yamashita and Fenn [46] developed the electrospray ionization (ESI) method. ESI can ionize large biomolecules such as proteins without breaking these biomolecules. In 1988, Tanaka et al. [40] developed the soft-laser-desorption

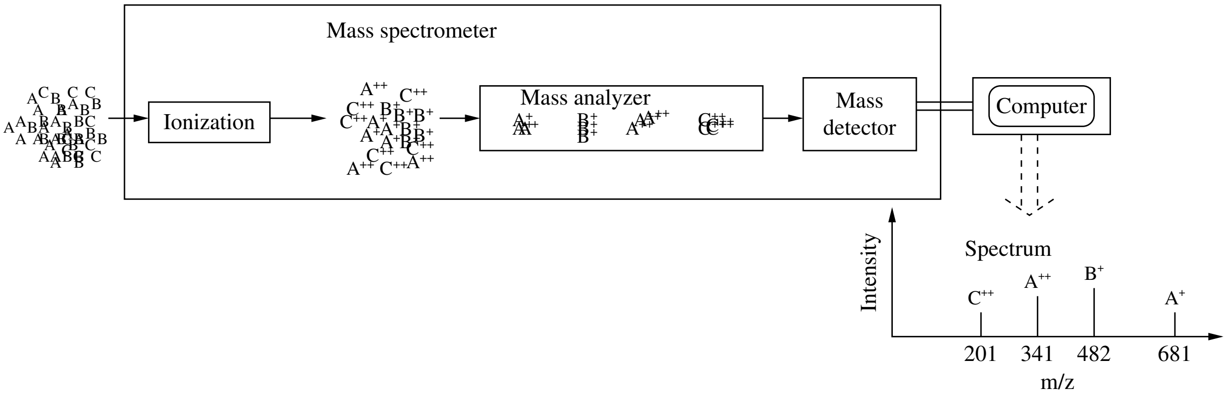


Figure 3.5: A schematic of a typical mass spectrometer [15].

method. Soft laser desorption can also ionize large biomolecules without breaking these biomolecules.

A mass spectrometer mainly consists of the following three components: an ion source, a mass analyzer, and a mass detector. A typical run of MS consists of a sequence of scans. Each of these scans proceeds as follows: First, the ion source ionizes a set of analytes coming from the inlet of the mass spectrometer so that these analytes form ions. Next, an extraction system brings these ions from the ion source to the mass analyzer. Then, the mass analyzer separates these ions according to the m/z of these ions. Afterwards, the mass analyzer sends these separated ions to the mass detector. Finally, the mass detector measures the quantity of ions at each specific m/z to produce a mass spectrum. Figure 3.5 shows a schematic of a typical mass spectrometer.

Section 3.4.1 presents some types of ion sources. Section 3.4.2 presents some types of mass analyzers. Section 3.4.3 presents some types of mass detectors.

3.4.1 Ion source

Ionization can be either hard or soft. Hard ionization usually fragments analytes, and soft ionization usually does not fragment analytes. For example, electron-impact ionization (EI), also known as electron ionization, is a hard ionization technique. Some popular soft ionization techniques are fast atom bombardment (FAB), chemical ionization (CI), matrix-assisted laser desorption/ionization (MALDI), and electrospray ionization (ESI). Among these techniques, only MALDI and ESI can ionize large biomolecules without fragmenting most of these biomolecules. ESI is currently the most popular ionization technique.

The working mechanism of ESI is the following.

1. By mixing water and volatile compounds, a solvent is prepared. By mixing this solvent with sample molecules, a solution is prepared.
2. This solution is dispersed by an electrospray into aerosol.
3. This aerosol is subject to a strong electric field to produce charged droplets.
4. The solvent in these charged droplets evaporates. During this evaporation, each of these droplets can undergo the subsequent Coulomb-fission cycle.
    1. Due to this evaporation, one such droplet continuously decreases in size. However, the charge on this droplet remains constant.
    2. The electrostatic repulsion of the same charge on this droplet becomes too high compared with the surface tension that holds this droplet together.
    3. This droplet explodes and then becomes multiple smaller droplets.
    4. Each of these smaller droplets can recursively undergo another such Coulomb-fission cycle.
5. The solvent is almost completely evaporated. Each of the sample molecules might have one or more charges.

ESI has several advantages. For example, ESI can ionize a protein without denaturing this protein, can analyze a dilute solution, and can ionize analytes in any polar solvent. Most importantly, ESI can generate multiply charged ions. Thus, the m/z of a molecule with high molecular weight can still be within the m/z detection range of a typical mass spectrometer. Thus, ESI is currently the most commonly used ionization technique.

3.4.2 Mass analyzer

Every mass analyzer separates ions according to the m/z of each of these ions. m/z is the mass-to-charge ratio measured in Da. Every mass analyzer is characterized by m/z range, peak shape, mass resolution (resolution), and mass accuracy (accuracy).

If the m/z of an ion is within the m/z range of a mass analyzer, then the mass spectrometer that uses this mass analyzer can detect this ion. Otherwise, this mass spectrometer cannot detect this ion.

A peak is an elevation of intensities within a small interval of m/z in a mass spectrum. As shown in Figure 3.6, a peak is usually bell-shaped.

Resolution is the ability to distinguish between two peaks respectively having two slightly different values of m/z. The IUPAC definition of resolution and the resolving-power definition of resolution coexist. Similarly, the IUPAC definition of resolving power

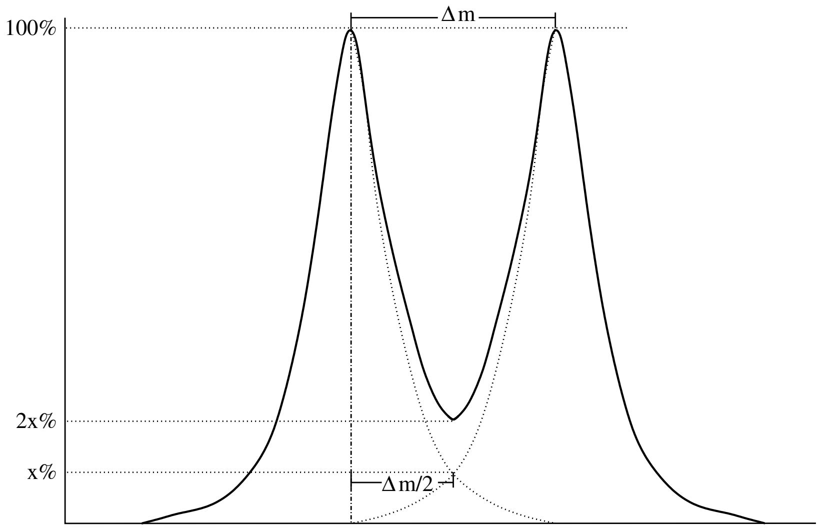


Figure 3.6: Definitions of ΔM, the slight difference between the respective m/z of two adjacent peaks [15]. The two dotted curves are two peaks. The full-line curve is produced by adding the two peaks. In this figure, x%-peak-width ΔM = 2x%-valley ΔM = Δm.

and the resolving-power definition of resolving power coexist. Let M be the m/z range of a mass analyzer. let ΔM be the slight difference between the m/z of a peak and the m/z of another peak. Let us suppose that these peaks are both produced by a mass spectrometer that uses this mass analyzer. According to the IUPAC definition, resolution is defined as $\frac{M}{\Delta M}$, and resolving power is defined as ΔM [31]. According to the resolving-power definition, resolution is defined as ΔM, and resolving power is defined as $\frac{M}{\Delta M}$. The ratio $\frac{M}{\Delta M}$ has no unit, and the unit of ΔM is the unit of m/z. Thus, the unit of resolution or of resolving power can indicate which of these two definitions is used. Similarly, the peak-width definition of ΔM and the valley definition of ΔM coexist. These two definitions of ΔM are described in Figure 3.6. According to the peak-width definition, an $x\%$-peak-width ΔM is defined as the width of a peak measured at $x\%$ of the height of this peak. The overlap between two bell-shaped peaks having the same shape but slightly different m/z values produces a valley between these two bell-shaped peaks. Let us suppose that the minimum height of the valley is $x\%$ of the height of these two bell-shaped peaks. According to the valley definition, an $x\%$-valley ΔM is the difference between the m/z values of these two bell-shaped peaks.

Accuracy is the ability to produce a peak whose centroid m/z is near the theoretical m/z of this peak. Let p_O be the average m/z of the observed peak. Let p_E be the theoretical

m/z of the expected peak. Accuracy is defined as either $\frac{|p_O - p_E|}{p_E}$ or $|p_O - p_E|$. Thus, the unit of accuracy can indicate which of these two definitions of accuracy is used.

3.4.3 Mass detector

Every mass detector records the current-or-charge produced by an ion when this ion hits or passes by a surface. The working mechanism of an electron-multiplier detector is as follows: First, an incident ion can cause the ejection of some electrons. Then, each of these ejected electrons can cause a new ejection of more electrons. Afterwards, this electron-ejection amplification continues until a huge quantity of electrons produce a detectable signal. The working mechanism of a Scintillator detector is as follows: First, an incident ion can cause the emission of some electrons. Then, this emission of electrons can cause the emission of light, Afterwards, this emission of light is detected. The working mechanism of a Faraday-Cup detector is as follows: First, an incident ion collides with a metal, and this collision can cause the ejection of secondary electrons. Then, this ejection can generate a flow of electric current. Afterwards, this flow of electric current is detected. Almost every mass detector amplifies the signal generated by an incident ion.

3.5 MS/MS

MS/MS is an analytical-chemistry technique that can fragment a biomolecule. This fragmentation can partially or fully reveal the chemical structure of this fragmented biomolecule. MS/MS has two stages. MS^1 is the first stage of MS/MS, and MS^2 is the second stage of MS/MS. MS^2 is immediately after MS^1. MS^1 proceeds as follows: First, an ionization source ionizes some sample molecules, and these ionized sample molecules are referred to as precursor ions. Then, a mass analyzer separates these precursor ions based on the m/z of these precursor ions. Finally, a mass detector detects the m/z of some of these separated precursor ions. MS^2 proceeds as follows: First, an ion filter selects some MS^1-generated precursor ions that are within a chosen m/z range. Next, some of these selected precursor ions are fragmented. Then, some of these fragments become product ions. Afterwards, a mass analyzer separates these product ions based on the m/z of these product ions. Finally, a mass detector detects the m/z of some of these separated product ions. An MS^1 spectrum is defined as the mass spectrum produced in MS^1. An MS^2 spectrum is defined as the mass spectrum produced in MS^2. MS^1 spectrum, precursor spectrum, survey spectrum, MS^1 scan, precursor scan, and survey scan are all equivalent in meaning. MS^2 spectrum,

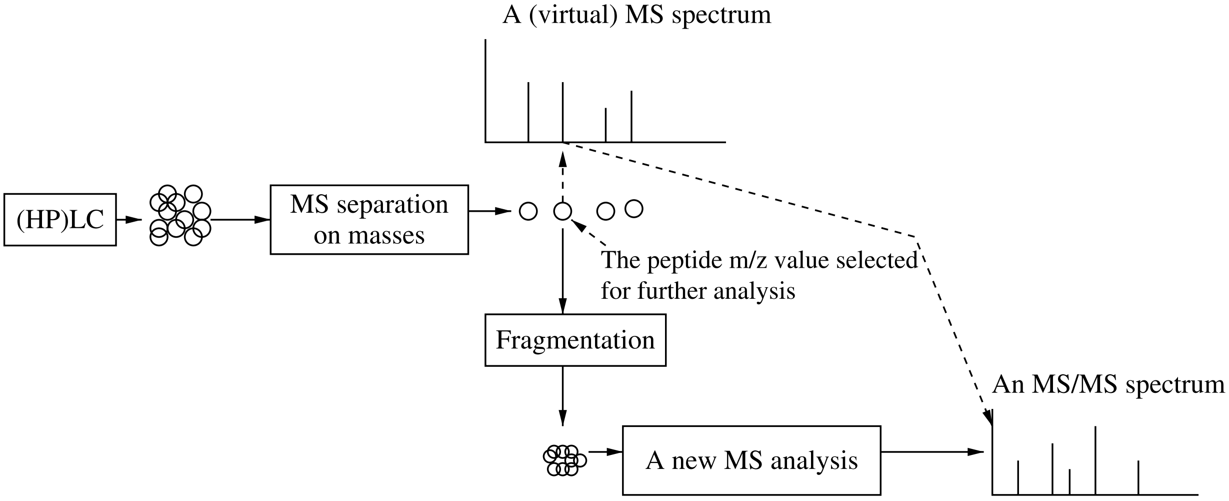


Figure 3.7: A schematic of MS/MS [15]. Figure 3.9 is an example that shows peaks produced by protein MS.

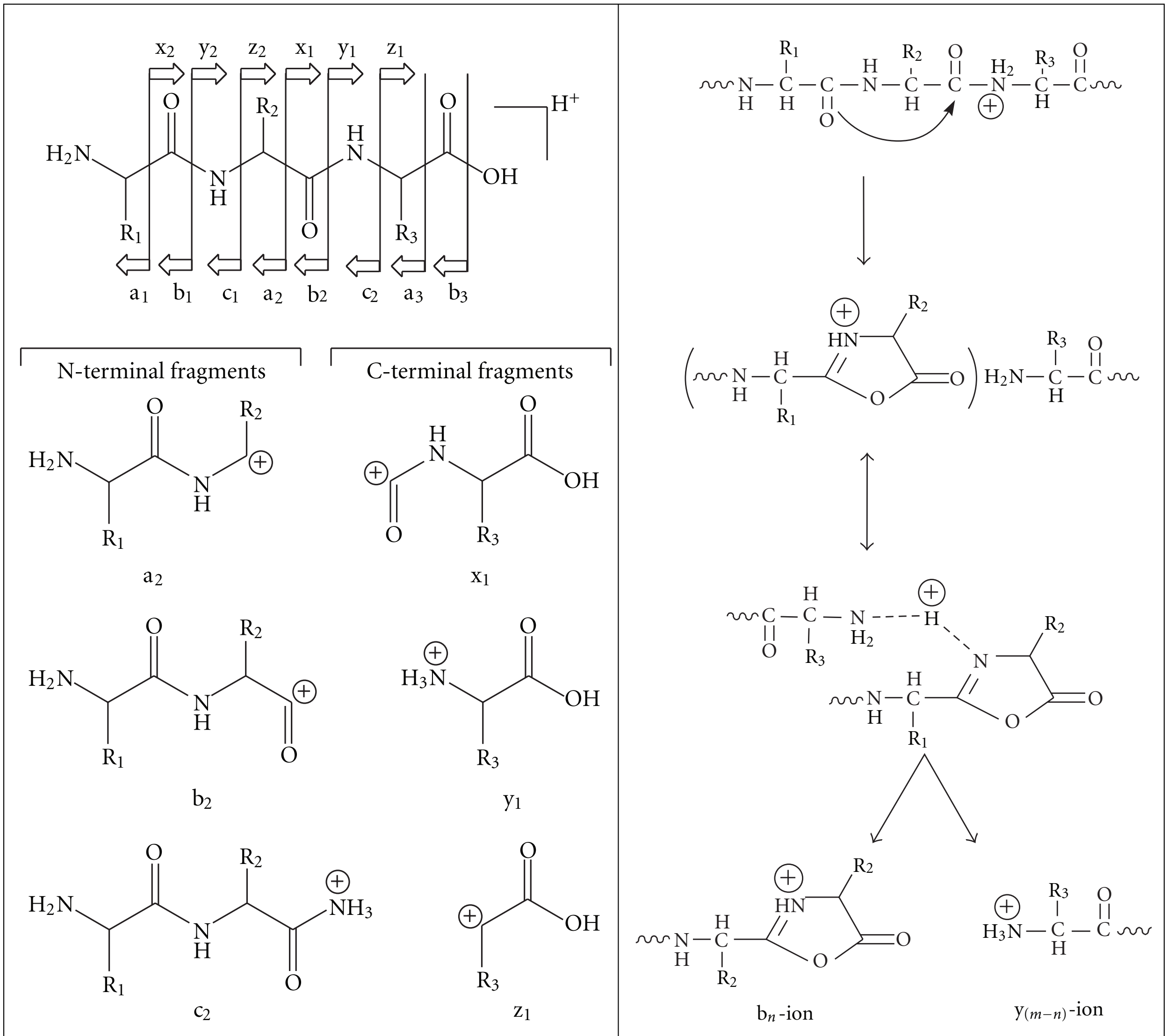
MS/MS spectrum, product spectrum, MS^2 scan, MS/MS scan, and product scan are all equivalent in meaning.

The distribution of the product ions generated by MS/MS depends on the fragmentation method used for this MS/MS. Collision-induced dissociation (CID), also known as collision activated dissociation (CAD), is the most popular fragmentation method. The mechanism of CID is as follows: First, an electromagnetic field accelerates a precursor-ion AB. Next, this precursor ion can collide with at least one neutral gaseous molecule M. Then, this collision can cause this precursor ion to fragment. Afterwards, this precursor ion can fragment into one product ion A and one uncharged molecule B. Finally, A can be detected. The following chemical equation describes the mechanism of CID.

$$\mathrm{AB}^+ + \mathrm{M} \to \mathrm{A}^+ + \mathrm{B} + \mathrm{M}$$

Even if no fragmentation method is used, precursors can still fragment if the energy in these precursors is sufficiently high.

MS^2 can break the backbone of a peptide to produce peptide fragments. Breakage of the bond between alpha-carbon and carbonyl-carbon can generate either an a-ion or an x-ion (Figure 3.8a). Breakage of the bond between carbonyl-carbon and nitrogen can generate either a b-ion or a y-ion (Figure 3.8a). Breakage of the bond between nitrogen and alpha-carbon can generate either a c-ion or a z-ion (Figure 3.8a). Only a positively charged fragment containing the N-terminus of a precursor peptide can become an a-ion, a b-ion,



(a) The notation for the major product ions.

(b) How b-ions and y-ions are formed.

Figure 3.8: Peptide fragmentation in MS^2 [4].

or a c-ion. Only a positively charged fragment containing the C-terminus of a precursor peptide can become an x-ion, a y-ion, or a z-ion. Sufficiently high collision energy in CID can even fragment the side chain of a residue. Then, this fragmentation can generate some product ions that are not shown in Figure 3.8a. However, CID generates mostly b-ions and y-ions.

3.6 Preprocessing of mass spectra

In a mass spectra, the ion intensity at a precise m/z is the strength of the signal that is generated by some ions having this m/z. This strength is often the number of such ions that are detected. A peak is defined as the ion intensities within a small interval of m/z in a mass spectrum. Presumably, some physically and chemically identical ions generate most of the ion intensities in this peak.

A peak is not always bell-shaped. The intensity of a peak is the sum of all intensities in this peak. The centroid of a peak is characterized by a combination of the representative m/z of this peak and the intensity of this peak. A centroid C of a peak is mathematically defined in Equation (3.1) [42].

$$C \overset{\mathsf{def}}{=} \left(\frac{\sum\limits_{a<r<b} y(r)\cdot r}{\sum\limits_{a<r<b} y(r)}, \sum\limits_{a<r<b} y(r) \right) \tag{3.1}$$

where r is m/z, $y(r)$ is the ion intensity at r, $\sum_{a<r<b} y(r)$ is the intensity of this peak, a is the lower m/z border of this peak, and where b is the upper m/z border of this peak. [42] presents several procedures for determining both a and b. Centroiding is defined as the process of replacing a peak by the centroid of this peak. Centroiding transforms a peak into one ion intensity at one m/z. Thus, centroiding reduces, in a mass spectrum, the number of pairs of m/z and ion intensity. Thus, centroiding compresses a mass spectrum although this compression is lossy. Moreover, centroiding partially removes the noise in a mass spectrum. Thus, centroiding facilitates the analysis of a mass spectrum.

Isotopes are defined as atoms having the same number of protons but different number of neutrons. Thus, isotopes have the same chemical properties but different masses. The mass difference between any two isotopes of a chemical element is a multiple of the mass of a neutron, and the mass of a neutron is approximately 1.009Da. Almost every atom has multiple isotopes. Isotopic molecules have the same structural formula but pairwise different masses. This mass difference exists because at least one atom position in this structural formula has isotopes. Isotopes can also refer to isotopic molecules. Thus, all isotopes of every molecule pairwise differ in molecular weight. Each of these pairwise differences is approximately a multiple of 1.009Da.

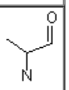
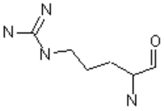
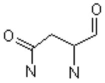
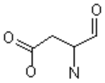
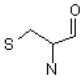
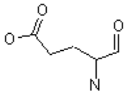
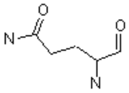
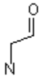
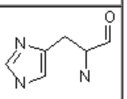
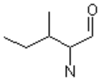
The mass of a molecule is the sum of the respective masses of the individually distinguishable atoms that collectively constitute this molecule. The monoisotopic mass of an atom is defined as the mass of the most abundant isotope of this atom. Nominal mass is defined as the monoisotopic mass rounded to the nearest integer. The average mass of an atom is defined as the average of the respective masses of all isotopes of this atom such that this average is weighted by the natural abundance of each of these isotopes. Table 3.2 shows the respective monoisotopic masses of the commonly observed amino-acid residues and the respective average masses of these residues.

In every mass spectrum, the charge state (z) of a peak is defined as the charge of the ion that generated this peak. Let us suppose that two ions have the same z and differ by a mass difference of Δm. Then, in every mass spectrum, the respective m/z of the peaks respectively generated by these two ions differ by $\frac{\Delta m}{z}$. Thus, n isotopes having the same z and ordered by mass can respectively generate n isotopic peaks ordered by m/z. Every two consecutive peaks in these isotopic peaks differ by approximately $\frac{1.009}{z}$ in m/z. The charge-state determination of some isotopic peaks is defined as the process of inferring the z of these peaks. Deisotoping of some isotopic peaks is defined as the process of converting these peaks into one representative peak. The mass in the m/z of this representative peak is usually the monoisotopic mass of the ions that respectively generated these isotopic peaks. The z of this representative peak is the common z of these isotopic peaks.

Let A_1 and A_2 be two chemically and physically identical molecules. The following can happen: A_1 gains one positive charge to become A_1^+, and A_2 gains two positive charges to become A_2^{++}. Let m_1 the m/z of the peak generated by A_1^+, and let m_2 the m/z of the peak generated by A_2^{++} Then, $m_1 + 1.007 \approx 2 \cdot m_2$. In general, some chemically and physically identical molecules can form different ions during ionization. Then, these different ions generate different peaks having pairwise different m/z. Deconvolution is defined as the process of converting these different peaks into one peak by assuming the following: after ionization, each sample molecule can only gain exactly one proton (electron) instead of being able to gain multiple protons (electrons).

During fragmentation in MS^2, an ion might lose part of this ion such that the z of this ion remains unchanged. This lost part is usually a small molecule, such as H_2O or NH_3. Such loss is referred to as neutral loss.

The mass of a proton is approximately 1.007Da. The mass of a neutron is approxi-

Residue	3-letter code	1-letter code	Mono-isotopic mass	Average mass	Structure
Alanine C_3H_5NO	Ala	A	71.037114	71.0779	
Arginine $C_6H_{12}N_4O$	Arg	R	156.101111	156.1857	
Asparagine $C_4H_6N_2O_2$	Asn	N	114.042927	114.1026	
Aspartic acid $C_4H_5NO_3$	Asp	D	115.026943	115.0874	
Asn or Asp	Asx	B			
Cysteine C_3H_5NOS	Cys	C	103.009185	103.1429	
Glutamic acid $C_5H_7NO_3$	Glu	E	129.042593	129.114	
Glutamine $C_5H_8N_2O_2$	Gln	Q	128.058578	128.1292	
Glu or Gln	Glx	Z			
Glycine C_2H_3NO	Gly	G	57.021464	57.0513	
Histidine $C_6H_7N_3O$	His	H	137.058912	137.1393	
Isoleucine $C_6H_{11}NO$	Ile	I	113.084064	113.1576	

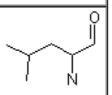
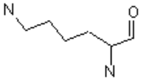
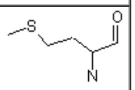
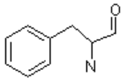
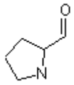
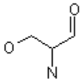
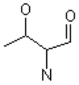
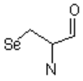
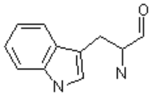
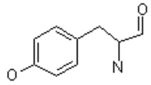
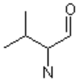
Residue	3-letter code	1-letter code	Mono-isotopic mass	Average mass	Structure
Leucine $C_6H_{11}NO$	Leu	L	113.084064	113.1576	
Lysine $C_6H_{12}N_2O$	Lys	K	128.094963	128.1723	
Methionine C_5H_9NOS	Met	M	131.040485	131.1961	
Phenylalanine C_9H_9NO	Phe	F	147.068414	147.1739	
Proline C_5H_7NO	Pro	P	97.052764	97.1152	
Serine $C_3H_5NO_2$	Ser	S	87.032028	87.0773	
Threonine $C_4H_7NO_2$	Thr	T	101.047679	101.1039	
Selenocysteine C_3H_5NOSe	SeC	U	150.95363	150.0379	
Tryptophan $C_{11}H_{10}N_2O$	Trp	W	186.079313	186.2099	
Tyrosine $C_9H_9NO_2$	Tyr	Y	163.06332	163.1733	
Unknown	Xaa	X			
Valine C_5H_9NO	Val	V	99.068414	99.1311	

Table 3.2: Some properties of the commonly observed amino acids [1].

mately 1.009Da. Let C be any collection of chemically identical molecules. Let m be the monoisotopic mass, in Da, of C. Let us suppose that each isotope of C gains one proton to become a cation. Then, an approximate m/z of the peak generated by this cation is in the following set:

$$\begin{gathered}\{\ \ldots, 1.007 + m - 2\times 1.009, 1.007 + m - 1\times 1.009, \\ 1.007 + m, \\ 1.007 + m + 1\times 1.009, 1.007 + m + 2\times 1.009, \ldots\ \}\end{gathered}$$

where these different m/z correspond to different isotopes. Let us suppose that the most naturally abundant isotope of C forms some pairwise different cations. Then, an approximate m/z of the peak generated by any of these cations is in the following set:

$$\{\frac{m + 1\times 1.007}{1}, \frac{m + 2\times 1.007}{2}, \frac{m + 3\times 1.007}{3}, \ldots\}$$

where these different m/z correspond to different charge states (z). Let M be the set that contains only the respective masses of all isotopes of C. Then, for each $m' \in M$, there exists an integer n such that $m' \approx m + n\times 1.009$. Let Z be the set that contains only the respective z of all cations formed by C. Then, $Z \subset \{1, 2, 3, \ldots\}$, and the numbers in Z are consecutive because ionization efficiency as a function of z is bell-shaped. Thus, for each isotope of C and for each cation formed by this isotope, the m/z of the peak formed by this cation is in the following set.

$$\bigcup_{m' \in M} (\bigcup_{z' \in Z} (\frac{m' + 1.007 \cdot z'}{z'}))$$

Both M and Z can be observed in a mass spectrum.

The combination of deconvolution and deisotoping converts all peaks generated by C into one single peak. This single peak is presumably generated by the addition of one proton to the most naturally abundant isotope of C.

3.7 Peptide-spectrum match (PSM)

In an MS^2 spectrum, the respective masses of some product ions formed by a peptide M can be estimated by the following procedure:

1. Determine the mass offset Δm that is specific to the product ion formed by M. As shown in Figure 3.8b, Both y-ion and b-ion are mostly composed of residues chained together by peptide bonds, y-ion possesses one extra H_2O and one extra

hydrogen atom (H), and b-ion possesses one extra H. Thus, $\Delta m \approx 19.018$ for y-ion and $\Delta m \approx 1.008$ for b-ion.

2. Determine the direction in which the residues of M are processed. This direction is from N-terminus to C-terminus for a-ion, b-ion, and c-ion. This direction is from C-terminus to N-terminus for x-ion, y-ion, and z-ion.
3. Iterate through the residues of M. In each iteration, let Σm be the sum of the respective masses of all residues which are iterated. Add $\Delta m + \Sigma m$ to the output list of the respective masses of these product ion formed by M.

Figure 3.9 shows an example of the application of such a procedure. Such a procedure can be used for calculating a PSM score. The PSM between a peptide and a mass spectrum implies that this peptide is likely to have generated some signals in this mass spectrum. Thus, PSM can be used for identifying the peptide that generated some signals in a given mass spectrum.

De novo sequencing derives information about the sequence of a presumably novel peptide. This sequence presumably has never been discovered before this *de novo* sequencing. Database search derives information about the sequence of a peptide such that this sequence can be extracted from a database. Usually, this peptide can be generated by the proteolysis of at least one protein found in this database. This proteolysis is usually catalyzed by a given protease. *De novo* sequencing explores more peptide sequences than database search. Thus, *de novo* sequencing is more error-prone than database search. Some commonly used database-search software packages are Mascot [9], PEAKS DB [47], Sequest [16], MS-GFDB [22], X!Tandem [10], and OMSSA [18].

Post-translational modification (PTM) is any *in vivo* covalent modification to a protein after this protein has been synthesized. PTM is presented in more detail in [33, Chapter 20]. Almost every PTM can be detected in a mass spectrum as a mass shift. Thus, MS can often identify and sometimes quantitate a PTM. FPOP usually shifts the mass of a biomolecule by +15.99Da (Table 3.1). However, FPOP is not *in vivo*. Thus, FPOP does not generate any PTM.

3.8 Mass spectrometry (MS) protocols

Figure 3.10 shows some workflows of protein MS. Figure 3.11 is an schematic of a typical RP-MS experiment. RP-MS is commonly used for studying protein-folding dynamics. Unfortunately, the spatial resolution of RP-MS is only at peptide level. This peptide level is an intrinsic characteristic of proteolysis. Fortunately, by using a variant of the standard

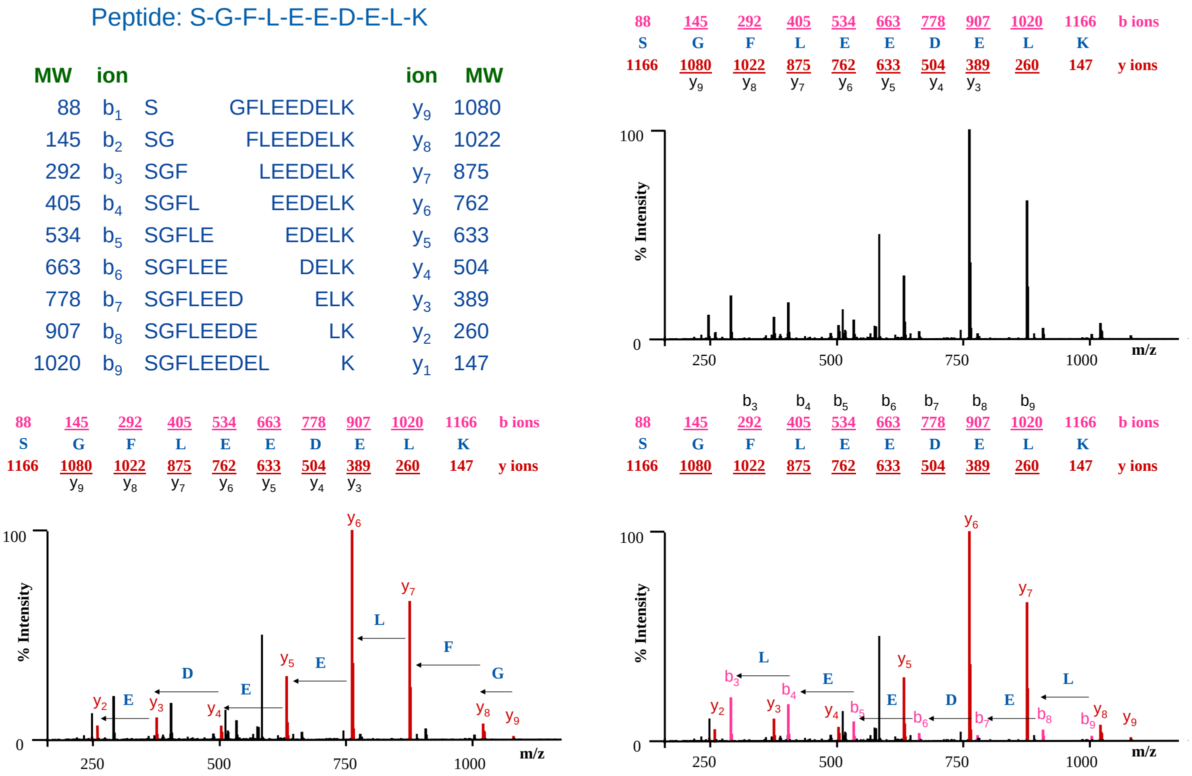


Figure 3.9: An interpretation of an MS^2 spectrum by PSM. The peptide SGFLEEDELK generated this MS^2 spectrum. The peak at the m/z of 583.5 is the doubly charged precursor of SGFLEEDELK. (Upper left) The theoretical molecular weight (MW) in Da, the notation, and the sequence, of each b-ion of SGFLEEDELK and of each y-ion of SGFLEEDELK. (Upper right) An uninterpreted MS^2 spectrum generated by SGFLEEDELK. (Lower left) For some high-intensity peaks, the m/z of this peak matches the theoretical MW of an y-ion of SGFLEEDELK. (Lower right) For some high-intensity peaks, the m/z of this peak matches the theoretical MW of a b-ion of SGFLEEDELK.

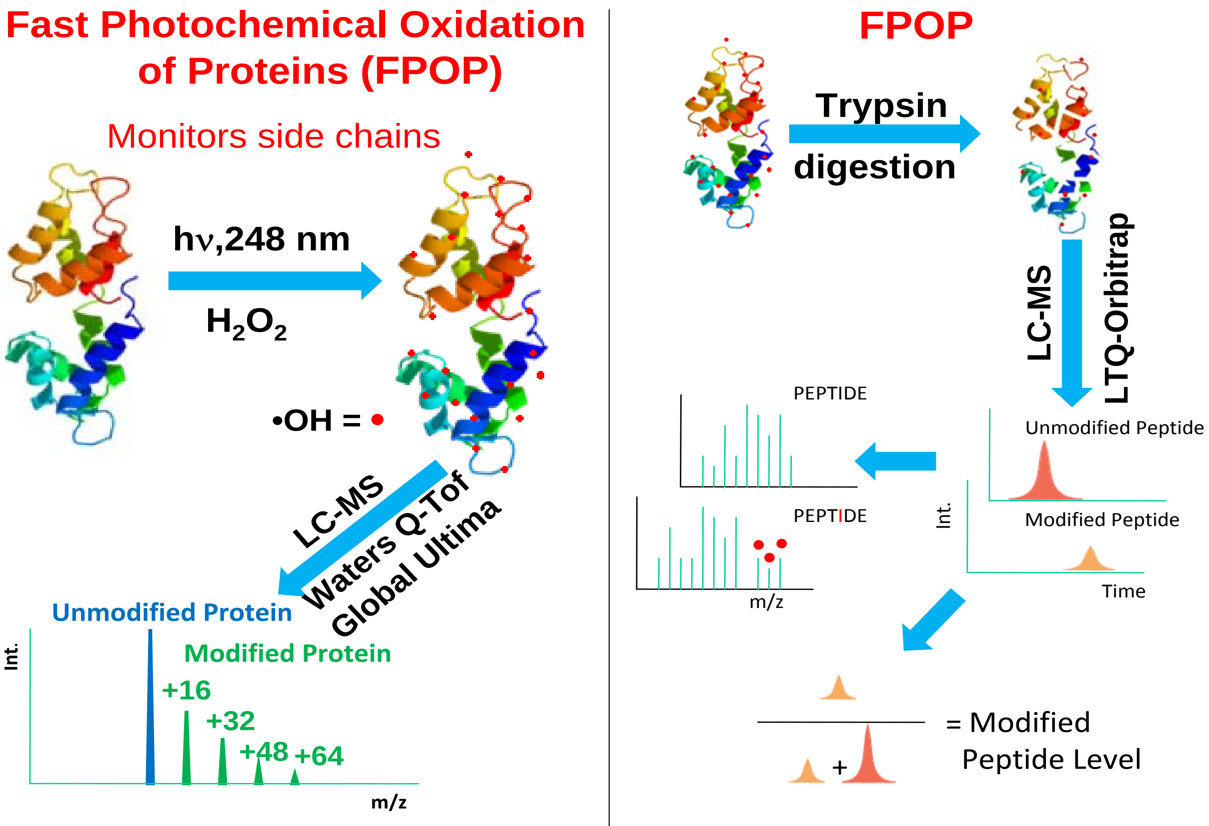
RP-MS, we can improve the spatial resolution of RP-MS to subpeptide level. Chapter 4 presents this improvement.

The following is the workflow of LC-MS/MS for an investigated protein:
1 A protease, such as trypsin, cleaves the investigated protein into peptides.
2 HPLC elutes these peptides. Each of these peptides exits the HPLC column at the RT of this peptide in this column.
3 While HPLC is eluting, the mass spectrometer repeats the following procedure.
  1 The inlet of the mass spectrometer extracts the eluted peptides from the exit of the HPLC column.
  2 The ion source of the mass spectrometer ionizes these extracted peptides by using a soft ionization technique, such as ESI. Ionized peptides can be referred to as precursors.
  3 The mass analyzer of the mass spectrometer separates these precursors based on the respective m/z of these precursors.
  4 If the conditions for MS^1 are satisfied, then the mass detector of the mass spectrometer measures the m/z-and-intensity of each of these precursors to produce a raw MS^1 spectrum.
  5 Otherwise, if the conditions for MS^2 are satisfied, then the mass spectrometer proceeds as follows:
    1 The mass analyzer selects these separated precursors such that the respective m/z of these selected precursors are within a given range.
    2 The mass analyzer uses a method, such as CID, to fragment these selected precursors. Some of these fragments respectively become product ions.
    3 The mass detector measures the m/z-and-intensity of each of these product ions to produce a raw MS^2 spectrum.
  6 The computer stores this raw MS^1-or-MS^2 spectrum.
4 Optionally, a software application preprocesses the raw mass spectra.
5 A human expert or a software application interprets such raw-or-preprocessed mass spectra.

The following is the workflow of bottom-up proteomics:
1. A protein mixture is prepared from cells or tissues.
2. Proteins of interest are extracted from this protein mixture using a conventional molecular-biology technique, such as 1D gel electrophoresis.
3. LC-MS/MS is performed on these proteins of interest.

The following is the workflow of RP-MS for an investigated protein:
1. FPOP is performed on the investigated protein. FPOP is described in Figure 3.1.
2. LC-MS/MS is performed on the protein modified by FPOP.

Figure 3.10: Some workflows of protein MS.



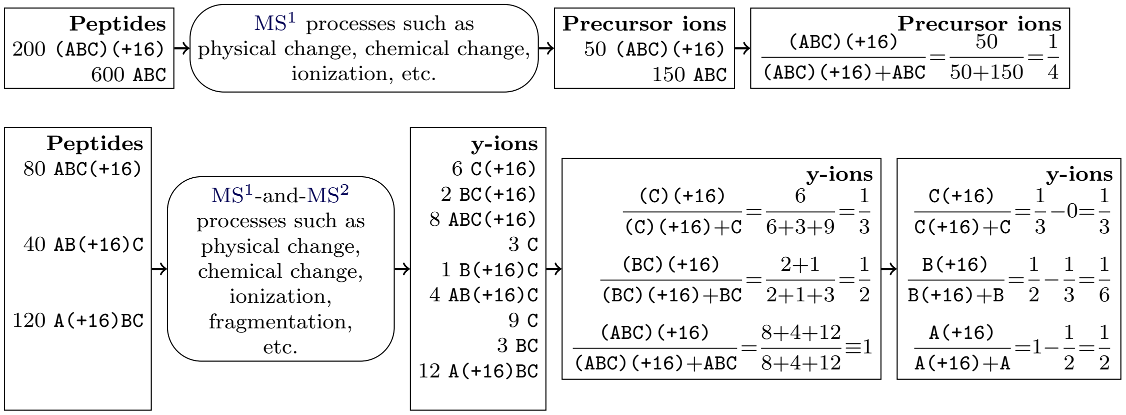
(a) a simple schematic of RP-MS.

(b) a schematic of LC-MS in RP-MS.

Figure 3.11: A schematic of RP-MS [2]. More details are presented in Figure 3.10.

Chapter 4

Quantitating mono-oxidation at subpeptide level


Peptides
200 (ABC)(+16)
600 ABC
MS^1 processes such as physical change, chemical change, ionization, etc.
Precursor ions
50 (ABC)(+16)
150 ABC
Precursor ions
$\frac{\texttt{(ABC)(+16)}}{\texttt{(ABC)(+16)+ABC}}=\frac{50}{50+150}=\frac{1}{4}$
Peptides
80 ABC(+16)
40 AB(+16)C
120 A(+16)BC
MS^1-and-MS^2 processes such as physical change, chemical change, ionization, fragmentation, etc.
y-ions
6 C(+16)
2 BC(+16)
8 ABC(+16)
3 C
1 B(+16)C
4 AB(+16)C
9 C
3 BC
12 A(+16)BC
y-ions
$\frac{\texttt{(C)(+16)}}{\texttt{(C)(+16)+C}}=\frac{6}{6+3+9}=\frac{1}{3}$
$\frac{\texttt{(BC)(+16)}}{\texttt{(BC)(+16)+BC}}=\frac{2+1}{2+1+3}=\frac{1}{2}$
$\frac{\texttt{(ABC)(+16)}}{\texttt{(ABC)(+16)+ABC}}=\frac{8+4+12}{8+4+12}\equiv 1$
y-ions
$\frac{\texttt{C(+16)}}{\texttt{C(+16)+C}}=\frac{1}{3}-0=\frac{1}{3}$
$\frac{\texttt{B(+16)}}{\texttt{B(+16)+B}}=\frac{1}{2}-\frac{1}{3}=\frac{1}{6}$
$\frac{\texttt{A(+16)}}{\texttt{A(+16)+A}}=1-\frac{1}{2}=\frac{1}{2}$

$\Pr[\texttt{C}\text{ is mono-oxidized}]=\frac{1}{4}\cdot\frac{1}{3}$ $\quad\Pr[\texttt{B}\text{ is mono-oxidized}]=\frac{1}{4}\cdot\frac{1}{6}$ $\quad\Pr[\texttt{A}\text{ is mono-oxidized}]=\frac{1}{4}\cdot\frac{1}{2}$

Our algorithm quantitates mono-oxidation at subpeptide level on a dataset produced by targeted LC-MS/MS. This quantitation can improve the spatial resolution at which protein folding is studied with radical-probe mass spectrometry (RP-MS).

Figure 4.1: The graphical abstract of Chapter 4 (hypothetical data used as example).

Section 4.1 presents our motivation, which is to improve the spatial resolution at which protein folding is studied with RP-MS. Section 4.2 presents related works in the literature. Section 4.3 presents an MS/MS dataset produced by a variant of RP-MS. The MS/MS dataset is mainly produced by six runs of targeted MS/MS respectively analyzing six mono-oxidized tryptic peptides. Section 4.4 presents our algorithm. Our algorithm takes as input the data produced by a run of targeted MS/MS, and our algorithm quantitates the oxidation on a subpeptide of the peptide analyzed by this run. Section 4.5 presents the results of running our algorithm on the MS/MS dataset. These results are collectively consistent with some previously published oxidation rates. Section 4.6 presents the discussion about our work.

4.1 Motivation

Quantitating oxidation at peptide level means quantitating the extent of oxidation on a proteolyzed peptide. Quantitating oxidation at subpeptide level means quantitating the extent of oxidation on a short peptide that is part of a proteolyzed peptide. Quantitating oxidation at residue level means quantitating the extent of oxidation on one single residue of a proteolyzed peptide.

We proposed an algorithm for quantitating, at subpeptide level, the mono-oxidation produced by fast photochemical oxidation of protein (FPOP) and detected by targeted LC-MS/MS. Our algorithm can improve the spatial resolution of RP-MS. The spatial resolution of RP-MS is the granularity level at which RP-MS is used for studying protein folding. Moreover, our work is an important step towards quantitation of oxidation at residue level.

4.2 Related works

In 1999, Maleknia et al. [29] used Synchrotron X-rays to generate hydroxyl radical (HO·) within 10 milliseconds. In the same year, they used electrical discharge to oxidize proteins that are introduced into a mass spectrometer. Since then, many analytical methods for generating HO· have been developed. Unfortunately, these methods suffer from the uncertainty that HO· can partially denature an investigated protein. Moreover, these methods cannot efficiently control the extent of HO·-mediated modification to an investigated protein.

```
>1WLA:A|PDBID|CHAIN|SEQUENCE
GLSDGEWQQVLNVWGK VEADIAGHGQEVLIR LFTGHPETLEK FDK
FKHLK TEAEMK ASEDLK K HGTVVLTALGGILK K K GHHEAELKPLAQSHATKHK
IPIKYLEFISDAIIHVLHSK HPGDFGADAQGAMTK ALELFR NDIAAK YK ELGFQG
```

Figure 4.2: The FASTA sequence of apomyoglobin (PDB `1WLA:A`). Each word denotes a tryptic peptide. All tryptic peptides are analyzed in one run of MS^1. However, only the six underlined tryptic peptides are analyzed in six runs of targeted MS^2 respectively.

In 2005, Hambly and Gross [20] developed FPOP. FPOP reduces the chemical effect of HO· to less than 1 microsecond. Moreover, FPOP limits the extent of HO·-mediated modifications by using a radical scavenger, such as glutamine. Since then, FPOP has been extensively used for RP-MS. Unfortunately, RP-MS has only been used to quantitate oxidation at peptide level. Different amino acids respectively have hugely different rates of reaction with HO· (Table 3.1). Thus, if the rate of such reaction of an amino-acid residue is negligible, then this residue is often assumed to be always unoxidized.

Some methods were proposed to quantitate oxidation at subpeptide level. In 2012, Chen et al. [8] used MS^2 spectra to map some peaks in MS^1 spectra to some oxidized residues, then they used this mapping to quantitate the oxidation on each of some selected residues of Barstar. Unfortunately, this mapping requires considerable human effort, and this quantitation requires a high-resolution mass spectrometer. Moreover, this mapping is often compromised by the overlap between the respective retention times (RTs) of differently oxidized isobaric peptides. In 2013, Li et al. [26] used c-ion intensities to quantitate, with some errors, oxidation at subpeptide level. Unfortunately, Li et al. [26] did not discuss the correction of these errors and investigated only two real peptides.

4.3 The MS/MS dataset

The MS/MS dataset is produced by an RP-MS experiment. This experiment was conducted by Siavash Vahidi and Professor Lars Konermann. This experiment was similar to the other experiment described in [43]. This experiment proceeded as follows: First, a solution with pH=2 is prepared, containing denatured apomyoglobin (PDB `1WLA:A`). Next, FPOP oxidized most of this denatured apomyoglobin, although most tryptic peptides of apomyoglobin remain unoxidized. Then, trypsin cleaved oxidized apomyoglobin into tryptic peptides. Each of these tryptic peptides was either oxidized or unoxidized. Afterwards, one run of LC-MS analyzed these tryptic peptides to produce a sequence of MS^1

Sequence of both	For precursor ions of $\boldsymbol{P'}$				For precursor ions of $\boldsymbol{P}$			
$\boldsymbol{P'}$ and $\boldsymbol{P}$	RT in min	z	m/z	peak-area	RT in min	z	m/z	peak-area
`GLSDGEWQQVLNVWGK`	[49.0, 62.0]	3	611.30	12377	[59.0, 64.0]	3	605.97	17390
`VEADIAGHGQEVLIR`	[28.3, 40.3]	3	541.62	29232	[34.0, 83.0]	3	536.29	249193
`LFTGHPETLEK`	[22.5, 34.0]	3	429.89	24164	[28.5, 37.0]	3	424.56	123514
`TEAEMK`	[0.0, 7.0]	2	362.66	4545	[10.0, 14.8]	2	354.67	14512
`HPGDFGADAQGAMTK`	[19.5, 27.0]	3	506.89	20141	[26.7, 33.3]	3	501.56	6005
`ELGFQG`	[21.0, 34.0]	1	666.31	9365	[31.9, 37.9]	1	650.32	81906

Table 4.1: A summary of the MS^1 spectra in the MS/MS dataset. The m/z window that is used for constructing extracted-ion chromatograms (XICs) and thus peak-areas is the m/z of the precursor ion ±0.1Da. The peak-area of all multiply-oxidized (e.g. di-oxidized, tri-oxidized) peptides is at most 10% of the peak-area of the corresponding mono-oxidized peptide (data not shown). $\boldsymbol{P}$ is a chemical species of unoxidized peptides. $\boldsymbol{P'}$ is a chemical superspecies of mono-oxidized peptides that are chemically identical up to mono-oxidation-induced structural isomerism.

spectra. Finally, six runs of targeted LC-MS/MS respectively analyzed six mono-oxidized tryptic peptides among these peptides. Each of these six runs produced a sequence of MS^2 spectra. Figure 4.2 shows the sequence of apomyoglobin and the six mono-oxidized tryptic peptides. The FPOP presumably used a finely tuned quantity of radical scavengers to control oxidation extents. Thus, the sequence of MS^1 spectra shows that a tryptic peptide of apomyoglobin is rarely oxidized more than once. Thus, any tryptic peptide of apomyoglobin is assumed to be either unoxidized or mono-oxidized after the FPOP. We manually verified, by visual inspection, that the mass spectrometer that produced the MS/MS dataset has a mass accuracy of ±0.1Da. Peptides having the same sequence respectively generate precursor ions having the same charge state (Table 4.1).

In this section, $\boldsymbol{P}$ is defined as a chemical species of unoxidized peptides, and $\boldsymbol{P'}$ is defined as a chemical superspecies of mono-oxidized peptides that are chemically identical up to structural isomerism, where this isomerism is only due to the fact that any site on any residue can be mono-oxidized. Peak-area is a function that outputs the area under the curve of an XIC; let M be a class of molecules, let r be a set of mass spectra generated by one run of LC-MS or of LC-MS/MS; then, $\text{peak-area}(M, r) \overset{\text{def}}{=} \sum_{s \in r} \text{XIC}(M, s)$, so $\text{peak-area}(M, r)$ represents the total absolute quantity of M detected in r.

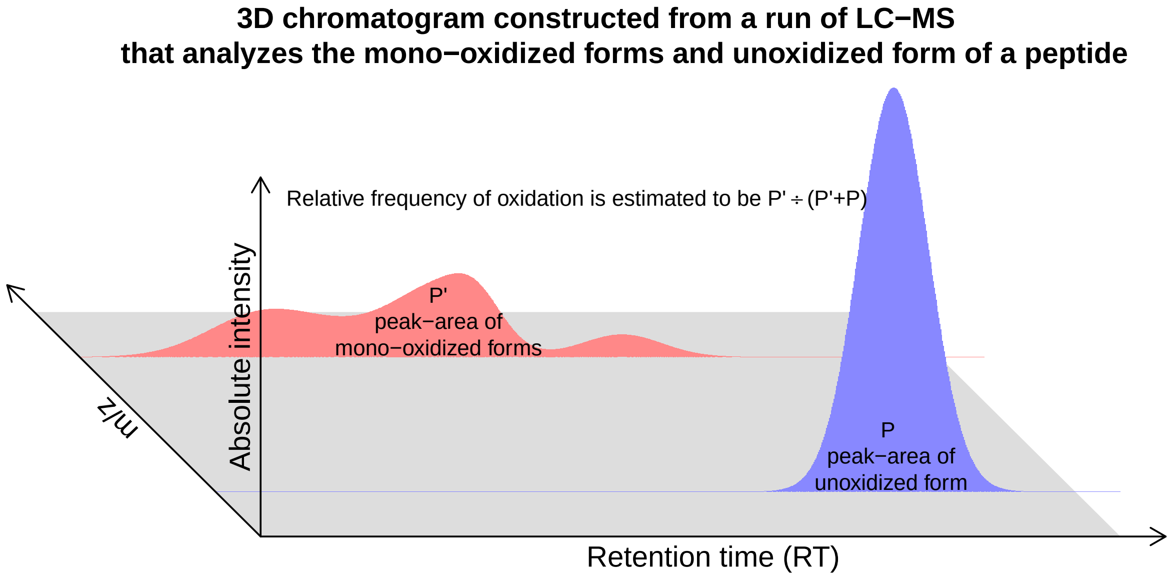

3D chromatogram constructed from a run of LC−MS
that analyzes the mono−oxidized forms and unoxidized form of a peptide
Relative frequency of oxidation is estimated to be P' ÷ (P'+P)
Absolute intensity
m/z
P'
peak−area of
mono−oxidized forms
P
peak−area of
unoxidized form
Retention time (RT)

Figure 4.3: A schematic of MS^1-based quantitation of oxidation at peptide level. A hypothetical run is used as example. The mono-oxidized forms can be chemically different and thus can be eluted at different RT ranges respectively.

4.4 Methods

In brief, our algorithm proceeds as follows: First, oxidation at peptide level is quantitated by using MS^1 spectra. Next, oxidation at subpeptide level is quantitated by using MS^2 spectra. Then, random errors in the quantitation of oxidation at subpeptide level are estimated by our empirical formula presented in Chapter 5, Afterwards, every quantitated oxidation at subpeptide level is processed to be nonnegative, because the unprocessed quantitated oxidation at subpeptide level can be negative due to the errors in such quantitation. Finally, by using both quantitation of oxidation at subpeptide level and quantitation of oxidation at peptide level, our algorithm quantitates oxidation at subpeptide level for multiple peptides in a protein. The input mass spectra were preprocessed by PEAKS 6 [27] before being used by our algorithm.

Every y-ion is indexed from C-terminus to N-terminus, but every residue is indexed from N-terminus to C-terminus (Figure 3.9).

Before quantitating oxidation at subpeptide level, we have to quantitate oxidation at peptide level. In MS^1 spectra, the fraction of the peak-area of mono-oxidized peptide over the peak-area of both mono-oxidized or unoxidized peptide denotes the relative frequency that this peptide is oxidized after FPOP. Thus, this peak-area fraction is used for quantitating oxidation at peptide level. Figure 4.3 shows how MS^1 enables the quantitation of oxidation at peptide level.

In the LC-MS/MS experiment, the instrument is programmed to target a mono-oxidized peptide $\boldsymbol{P'}$. When $\boldsymbol{P'}$ is eluted from liquid chromatography (LC) column, $\boldsymbol{P'}$ is continuously acquired by the mass spectrometer and fragmented. The product ions of $\boldsymbol{P'}$ are detected to produce an MS^2 spectrum. As a result, a sequence of MS^2 spectra are produced. For any index i, we use $\mathtt{y}_i$ to denote the unoxidized $\mathtt{y}_i$ ion and $\mathtt{y}'_i$ to denote the mono-oxidized $\mathtt{y}_i$ ion. The peak-area of $\mathtt{y}_i$ or $\mathtt{y}'_i$ is the total intensities of the corresponding ion in the sequence of MS^2 spectra, and is denoted by y_i or y'_i, respectively. Thus, ϕ_i, the fraction of $\mathtt{y}_i$ ions that are mono-oxidized, can be estimated by the following formula. $\phi_i := \dfrac{y'_i}{y_i + y'_i}$.

Because of the stochastic nature of every run of mass spectrometry and of the algorithmic artifact in the calculation of peak-area, random error exists in the observation of ϕ_i. One run of mass spectrometry cannot empirically assess any random error. Fortunately, Chapter 5 provides an empirical formula that estimates the following: the random error in a peak-area fraction given that this fraction is measured in only one run of mass spectrometry. Thus, we applied our empirical formula to ϕ_i, because ϕ_i is a peak-area fraction. The substitution of ϕ_i into our empirical formula implies that

$$\Phi_i \overset{\text{app}}{\sim} \mathcal{N}\left(\mathrm{E}[\Phi_i], \frac{y_i \cdot y'_i}{(y_i + y'_i)^3}\right), \tag{4.1}$$

where ϕ_i is one realization, or equivalently one observed value, of the hidden random variable Φ_i. Φ_i denotes the hidden stochastic process that generated ϕ_i with random error. ϕ_i is trivially an estimate of $\mathrm{E}[\Phi_i]$. Thus, let $\widehat{\mathrm{E}}[\Phi_i]$ be defined as ϕ_i.

The quantity of every y-ion should be proportional to the quantity of the peptides that can generate this y-ion. Thus, y'_i should be proportional to the quantity of mono-oxidized peptides whose mono-oxidation site is before or at the y-ion index i. Similarly, y_i should be proportional to the quantity of mono-oxidized peptides whose mono-oxidation site is after the y-ion index i. Thus, by definition, $\widehat{\mathrm{E}}[\Phi_i]$ denotes the relative frequency that the oxidation site on $\boldsymbol{P'}$ is before or at the y-ion index i. Thus, $\widehat{\mathrm{E}}[\Phi_i] - \widehat{\mathrm{E}}[\Phi_{i-1}]$ denotes the relative frequency that the oxidation site on $\boldsymbol{P'}$ is at the y-ion index i.

Relative frequency cannot be negative. Thus, for every applicable i, $\widehat{\mathrm{E}}[\Phi_i] - \widehat{\mathrm{E}}[\Phi_{i-1}]$ should be positive. Equivalently, $\widehat{\mathrm{E}}[\Phi_i]$ should monotonically increase as a function of i. For example, this monotonicity almost holds for `(VEADIAGHGQEVLIR)(+16)` (Figure 4.4). This

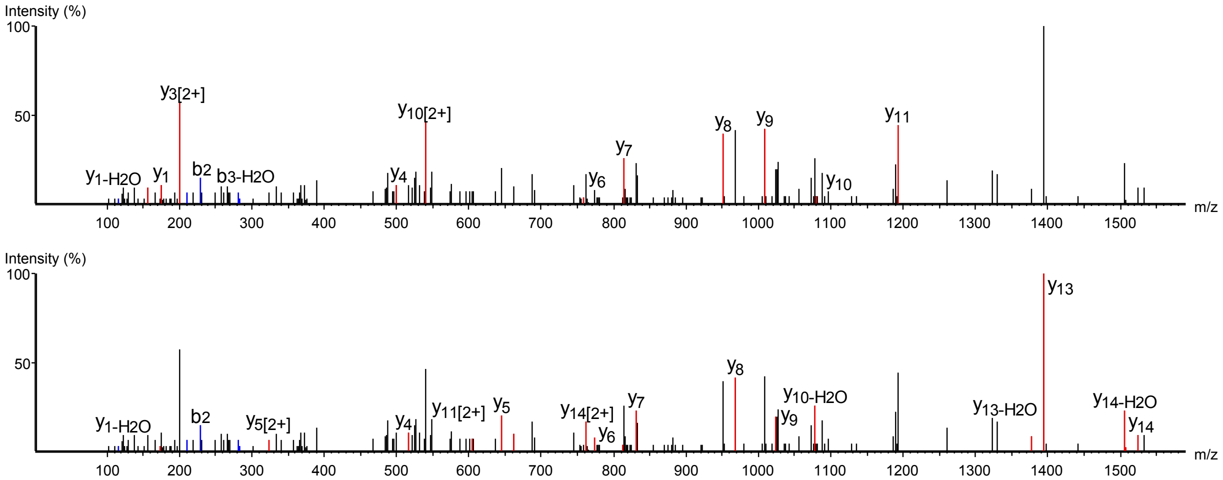


Figure 4.4: A mixture MS^2 spectrum of `VEADIAGHGQEVLIR` in the MS/MS dataset. The top annotation shows $\mathtt{y}_i$ and the bottom annotation shows $\mathtt{y}'_i$. i denotes an y-ion index. $\mathtt{y}$ denotes the unoxidized form of a y-ion. $\mathtt{y}'$ denotes the mono-oxidized forms of a y-ion. Both annotations annotate the same spectrum.

monotonicity is desired but not observed for some pairs of y-ion indexes. This monotonicity can be invalidated by multiple causes. However, the most important cause among these causes seems to be the stochastic nature of Φ_i that generated ϕ_i. This stochastic nature causes random error in the observation of ϕ_i. An estimation of this random error is provided by Equation (4.1). Then, by considering this random error, Line 15 of Algorithm 1 enforces this monotonicity.

The z-score of an observation is defined as follows: the deviation of this observation from the mean of this observation, divided by the standard deviation of this observation. The standard deviation of Φ_i can be estimated by Equation (4.1). Thus, Equation (4.1) can normalize observed deviations respectively to z-scores. Thus, the sum of the respective squares of these z-scores monotonically decreases as a function of the likelihood of observing these z-scores. This monotonic decrease leads to isotonic regression. Thus, the exact formulation of our isotonic regression is as follows given the length n of the sequence of a peptide:

$$\begin{aligned}
&\text{Minimize} && \sum_{i=1}^{n-1}\left(\frac{\widehat{\widehat{\mathrm{E}}}[\Phi_i]-\widehat{\mathrm{E}}(\Phi_i)}{\sqrt{\widehat{\mathrm{var}}(\Phi_i)}}\right)^2 && (4.2)\\
&\text{such that} && \forall i\in\{2,3,\ldots,n-1\}:\left(0\leq\widehat{\widehat{\mathrm{E}}}[\Phi_{i-1}]\leq\widehat{\widehat{\mathrm{E}}}[\Phi_i]\leq 1\right) && (4.3)\\
&\text{where} && \widehat{\mathrm{E}}(\Phi_i)=\phi_i=\frac{y_i'}{y_i+y_i{}'} \quad \text{and} \quad \widehat{\mathrm{var}}(\Phi_i)=\frac{y_i+y_i'}{(y_i\cdot y_i')^3}. && (4.4)
\end{aligned}$$

Our isotonic regression is solved by using the linear-time pool-adjacent-violators algorithm (PAVA) implemented by Turner [41].

The solution to our isotonic regression transforms each $\widehat{\mathrm{E}}[\Phi_i]$ into its corresponding $\widehat{\widehat{\mathrm{E}}}[\Phi_i]$. By definition of isotonic regression, $\widehat{\widehat{\mathrm{E}}}[\Phi_i]-\widehat{\widehat{\mathrm{E}}}[\Phi_{i-1}]\geq 0$ for all valid y-ion indexes i and $i+1$. Thus, every $\widehat{\widehat{\mathrm{E}}}[\Phi_i]-\widehat{\widehat{\mathrm{E}}}[\Phi_{i-1}]$ can denote a valid relative frequency. Thus, $\widehat{\widehat{\mathrm{E}}}[\Phi_i]-\widehat{\widehat{\mathrm{E}}}[\Phi_{i-1}]$ denotes the relative frequency of the following event: the residue located at the y-ion index i of a peptide is mono-oxidized given that this peptide as a whole is mono-oxidized. The random error in each $\widehat{\mathrm{E}}[\Phi_i]$ is estimated to be approximately $\sqrt{\widehat{\mathrm{var}}[\Phi_i]}$, and $\widehat{\mathrm{E}}[\Phi_i]\approx\widehat{\widehat{\mathrm{E}}}[\Phi_i]$. Thus, the random error in $\widehat{\widehat{\mathrm{E}}}[\Phi_i]$ is also estimated to be approximately $\sqrt{\widehat{\mathrm{var}}[\Phi_i]}$. Thus, if Φ_i and Φ_{i-1} are independent, then the random error in $\widehat{\mathrm{E}}[\Phi_i-\Phi_{i-1}]$ is estimated to be approximately $\sqrt{\widehat{\mathrm{var}}[\Phi_i]+\widehat{\mathrm{var}}[\Phi_{i-1}]}$. Thus, the random error in the observed $\widehat{\widehat{\mathrm{E}}}[\Phi_i]-\widehat{\widehat{\mathrm{E}}}[\Phi_{i-1}]$ is estimated to be approximately $\sqrt{\widehat{\mathrm{var}}[\Phi_i]+\widehat{\mathrm{var}}[\Phi_{i-1}]}$.

We can quantitate both the mono-oxidation on each peptide and the mono-oxidation on each residue of a mono-oxidized peptide. Thus, we can quantitate the mono-oxidation on each residue. Line 15 of Algorithm 1 (page 41) calculates mono-oxidation at residue level by using this quantitation. Afterwards, the relative frequency that each residue is mono-oxidized is estimated. Finally, the random error in this relative frequency is estimated as well.

Some errors exist in our MS^1-based quantitation of oxidation at peptide level. However, a peptide can be divided into multiple subpeptides. Thus, the extent of oxidation on each of these subpeptide is only a small difference between two peak-area fractions. Thus, quantitation of oxidation at subpeptide level has more error than quantitation of oxidation at peptide level. Moreover, the intensity of a product ion is usually much lower than the intensity of the precursor ion that formed this product ion. Thus, MS^2-based quantitation of oxidation has more error than MS^1-based quantitation of oxidation. We used both such MS^2-based quantitation at subpeptide level and such MS^1-based quantitation at peptide level. Therefore, we ignored error in such MS^1-based quantitation because such MS^2-based quantitation has much more error.

As mentioned in Section 3.6, preprocessing of mass spectra is important. Examples of such preprocessing are baseline removal, centroiding, deconvolution, and deisotoping. Thus, before applying any aforementioned procedure, we performed the following: First, we manually determined the charge state of every applicable precursor. Then, we performed a moving average with a window of 20s along RT for the MS^2 spectra in the MS/MS dataset. Finally, we let the software PEAKS 6 [27] preprocess these MS^2 spectra.

4.5 Results on the MS/MS dataset

Algorithm 1 generated Figure 4.5 from the MS/MS dataset which is described in Section 4.3. As mentioned in Section 4.4, the expected pattern is that Φ_i does not substantially decrease as the y-ion index i increases. Figure 4.5 shows the following: The mono-oxidized forms of `GLSDGEWQQVLNVWGK` show the expected pattern at all y-ion indexes without any exception. The mono-oxidized forms of `VEADIAGHGQEVLIR` show the expected pattern at all y-ion indexes except from $i = 9$ to $i = 10$. The mono-oxidized forms of `LFTGHPETLEK` show the expected pattern at all y-ion indexes except from $i = 5$ to $i = 6$ and from $i = 8$ to $i = 9$. The mono-oxidized forms of `TEAEMK` show the expected pattern at all y-ion indexes except from $i = 2$ to $i = 3$ and from $i = 4$ to $i = 5$. The mono-oxidized forms of `HPGDFGADAQGAMTK` show the expected pattern at all y-ion indexes except from $i = 5$ to $i = 6$. The mono-oxidized forms of `ELGFQG` show the expected pattern at all y-ion indexes without any exception.

The native reactivity of a free amino acid with HO· is positively correlated with the percentage of mono-oxidation on this residue. Unfortunately, this correlation is weak mainly because of the following: the reaction of a residue with HO· can cause a mass shift other than +15.99Da to this residue (Table 3.1), so this reaction does not always generate a mono-oxidized peptide. In a protein, the reactivity of a residue may depend on adjacent residues.

The five residues that are top-listed in Table 3.1 are most reactive with HO·. Thus, these five residues are investigated in Figure 4.5.

- Cysteine (`C`) is not in any of the six investigated peptides.
- Tryptophan (`W`) appears twice in `GLSDGEWQQVLNVWGK`. `GLSDGEWQQV` and `W` are the two subtryptic regions that have the majority of the oxidation on `GLSDGEWQQVLNVWGK`. As expected, `GLSDGEWQQV` and `W` both contain `W`. However, `GLSDGEWQQV` is too long.

Algorithm 1 quantitate-oxidation-at-subpeptide-level($\boldsymbol{P}, \boldsymbol{P}', r', r''$)

Input: $\boldsymbol{P}$ is a chemical species of unoxidized peptides. $\boldsymbol{P}'$ is a chemical superspecies of mono-oxidized peptides. Both $\boldsymbol{P}$ and $\boldsymbol{P}'$ have the same sequence and $\boldsymbol{P}'$ is heavier than $\boldsymbol{P}$ by approximately 15.99Da. A sample contains both $\boldsymbol{P}$ and $\boldsymbol{P}'$. A run of LC-MS surveyed this entire sample to produce a sequence r' of MS^1 spectra. A run of LC-MS/MS targeted only $\boldsymbol{P}'$ in this sample to produce a sequence r'' of MS^2 spectra.

Output: The estimated relative frequency that a y-ion is oxidized as a function of the index of this y-ion, such as Figure 4.5, and the estimated relative frequency that a residue is oxidized as a function of this residue, such as Figure 4.6.

```
$\widehat{\Pr}[\boldsymbol{P} \to \boldsymbol{P}'] \overset{\text{def}}{=} \dfrac{\text{peak-area}(\boldsymbol{P}', r')}{\text{peak-area}(\boldsymbol{P}', r') + \text{peak-area}(\boldsymbol{P}, r')}$
    ▷ quantitate mono-oxidation at peptide level by using information in $\mathrm{MS}^1$
Smooth $r''$ by a moving average of 20 $\mathrm{MS}^2$ spectra that are consecutive along RT.
    ▷ Numbers other than 20 yield similar results.
Preprocess smoothed $r''$ using PEAKS 6 [27].
$n \overset{\text{def}}{=}$ the length of the sequence of $\boldsymbol{P}$ or equivalently of $\boldsymbol{P}'$.
for $i \in \{1, 2, \ldots, n-1\}$ do
    $y_i \overset{\text{def}}{=} \text{peak-area}(\mathrm{y}_i(\boldsymbol{P}'), r'') + i$
    $y'_i \overset{\text{def}}{=} \text{peak-area}(\mathrm{y}'_i(\boldsymbol{P}'), r'') + (n - i)$
    $(\widehat{\mathrm{E}}[\Phi_i], \widehat{\mathrm{var}}[\Phi_i]) \overset{\text{def}}{=} \left( \dfrac{y'_i}{y_i + y'_i}, \dfrac{y_i \cdot y'_i}{(y_i + y'_i)^3} \right)$        ▷ Equation (5.15)
end for
Plot $\widehat{\mathrm{E}}[\Phi_i] \pm \sqrt{\widehat{\mathrm{var}}[\Phi_i]}$ as a function of $i$, and this plot is in Figure 4.5.
$(\widehat{\widehat{\mathrm{E}}}[\Phi_1], \widehat{\widehat{\mathrm{E}}}[\Phi_2], \ldots, \widehat{\widehat{\mathrm{E}}}[\Phi_{n-1}]) := \underset{\substack{(\phi_1, \phi_2, \ldots, \phi_{n-1}) \in [0,1]^{n-1} \\ \text{such that } \phi_1 \le \phi_2 \le \cdots \le \phi_{n-1}}}{\arg\min} \left( \sum_{i=1}^{n-1} \left( \dfrac{\phi_i - \widehat{\mathrm{E}}(\Phi_i)}{\sqrt{\widehat{\mathrm{var}}(\Phi_i)}} \right)^2 \right)$
    ▷ Perform isotonic regression of $\widehat{\mathrm{E}}[\Phi_i]$ versus $i$ where each $\widehat{\mathrm{E}}[\Phi_i]$ has weight $(\widehat{\mathrm{var}}[\Phi_i])^{-1}$
    ▷ The PAVA implemented by Turner [41] is used for solving our isotonic regression.
$((\widehat{\mathrm{E}}[\Phi_0], \widehat{\mathrm{var}}[\Phi_0]), (\widehat{\mathrm{E}}[\Phi_n], \widehat{\mathrm{var}}[\Phi_n])) := ((0, 0), (1, 0))$        ▷ Oxidation before 0 and $n$
for $i \in \{1, 2, \ldots, n\}$ do
    $\left( \widehat{\widehat{\mathrm{E}}}[\Phi_i - \Phi_{i-1}], \widehat{\widehat{\mathrm{var}}}[\Phi_i - \Phi_{i-1}] \right) \overset{\text{def}}{=} \left( \widehat{\widehat{\mathrm{E}}}[\Phi_i] - \widehat{\widehat{\mathrm{E}}}[\Phi_{i-1}], \widehat{\mathrm{var}}[\Phi_i] + \widehat{\mathrm{var}}[\Phi_{i-1}] \right)$
    $\widehat{\Pr}[\boldsymbol{P}_{n+1-i} \to \boldsymbol{P}'_{n+1-i}] \overset{\text{app}}{\approx} \widehat{\Pr}[\boldsymbol{P} \to \boldsymbol{P}'] \cdot \mathcal{N} \left( \widehat{\widehat{\mathrm{E}}}[\Phi_i - \Phi_{i-1}], \widehat{\widehat{\mathrm{var}}}[\Phi_i - \Phi_{i-1}] \right)$
end for
Plot $\widehat{\Pr}[\boldsymbol{P}_k \to \boldsymbol{P}'_k]$ as a function of the residue at index $k$ of $\boldsymbol{P}$ or equivalently of $\boldsymbol{P}'$,
    and this plot is in Figure 4.6.
```

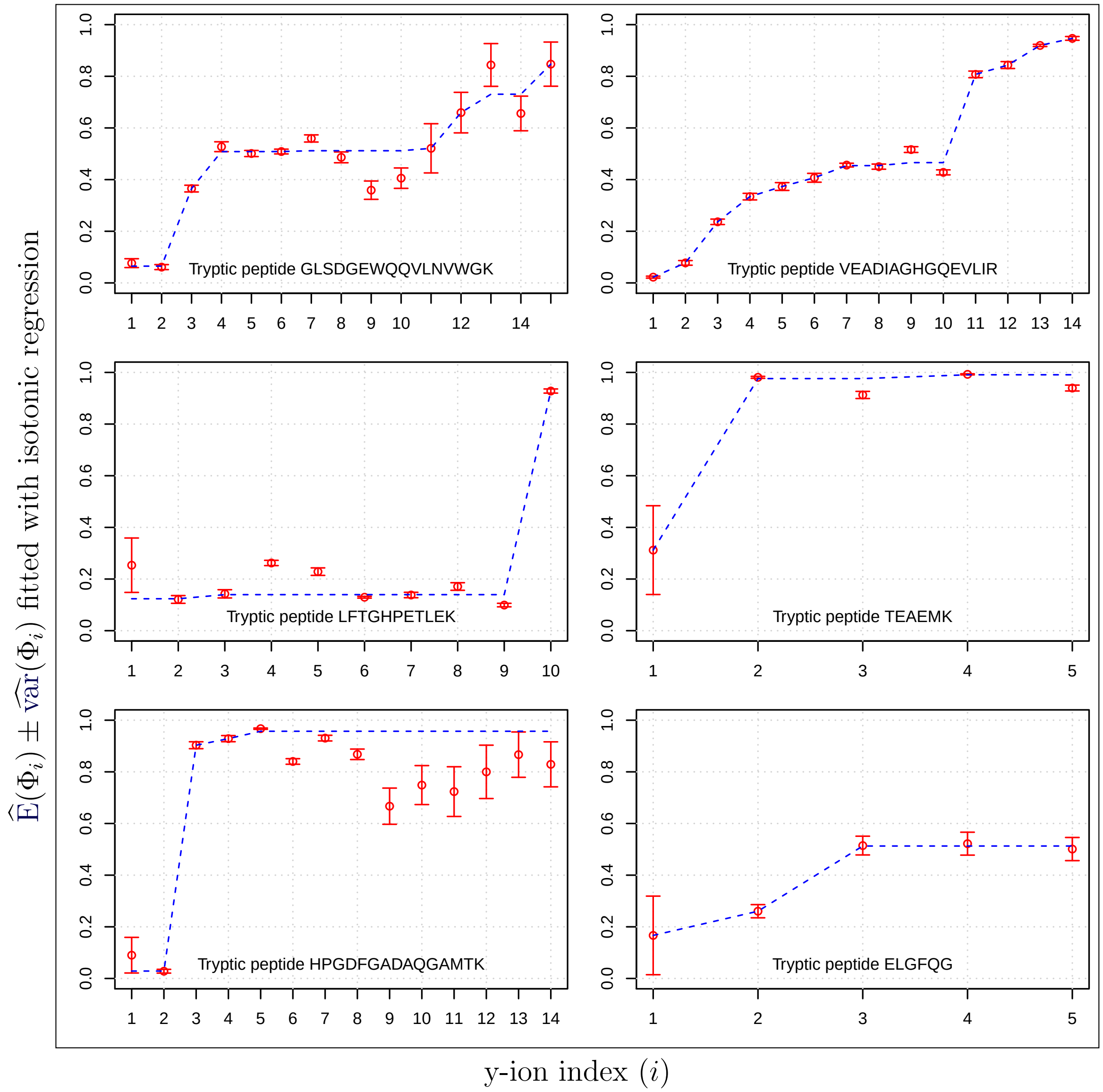


Figure 4.5: The estimated relative frequency that $\mathtt{y}_i$ is mono-oxidized as a function of i. Algorithm 1 generated this plot.

Thus, the precise region of oxidation on GLSDGEWQQV cannot be determined. Thus, the extent of oxidation on W that is part of GLSDGEWQQV cannot be accurately quantitated. For GLSDGEWQQVLNVWGK, the increase in ϕ_i from $i = 2$ to $i = 3$ in Figure 4.5 is likely to be caused by the high extent of oxidation on W. Thus, even if W is in a peptide, other residues in this peptide cannot be excluded for quantitating oxidation.

- Tyrosine (Y) is not in any of the six investigated peptides.
- Methionine (M) appears once in TEAEMK and once in HPGDFGADAQGAMTK. MK is only a small part of TEAEMK, but MK has approximately 98% of the oxidation on TEAEMK. Similarly, AM is only a small part of HPGDFGADAQGAMTK, but AM has approximately 93% of the oxidation on HPGDFGADAQGAMTK. Moreover, $\Phi_i - \Phi_{i-1} \approx 1$ whenever i corresponds to M. Thus, the oxidation on M is sufficiently high compared with other residues. Thus, if one single M is in a peptide and other residues in the peptide all have low reactivity with HO·, we can exclude other residues in this peptide for quantitating oxidation.
- Phenylalanine (F) appears once in LFTGHPETLEK, once in HPGDFGADAQGAMTK, and once in ELGFQG. FTGH has approximately 80% of the oxidation on LFTGHPETLEK and contains F. In HPGDFGADAQGAMTK, the oxidation in the subtryptic region containing F is characterized by huge statistical variation, Thus, we cannot accurately quantitate oxidation near F in HPGDFGADAQGAMTK. F has approximately 25% of the oxidation on ELGFQG; For LFTGHPETLEK, an obvious increase in Φ_i from $i = 9$ to $i = 10$ exists, and 10 is the y-ion index of F. However, for ELGFQG, no significant increase in Φ_i from $i = 2$ to $i = 3$ exists, and 3 is the y-ion index of F in ELGFQG. Thus, even if F is in a peptide, we cannot exclude other residues in this peptide for quantitating oxidation.

Figure 4.6 shows the relative frequency that a residue becomes mono-oxidized as a function of its residue index. The respective second-order reaction rates of the 20 standard amino acids with HO· are listed in Table 3.1.

Let us suppose that the 20 standard amino-acid residues are sorted in descending order based on their relative frequencies. Then, M, W, and F are likely to be ranked first, second, and third, respectively. Let us suppose that the 20 standard amino acids are sorted in descending order based on their reaction rates. Then, M, W, and F are ranked second, forth, and fifth, respectively. Thus, the observed high reactivity of these three residues with HO· is consistent with their intrinsic high reactivity with HO·.

Let us suppose that the 20 standard amino acids are sorted in ascending order based on their reaction rates. Then, G, N, D, A, and E are ranked first, second, third, forth, and fifth respectively. Let us suppose that the 20 standard amino-acid residues are sorted in ascending order based on their relative frequencies. Then, G, D, A, and E are all unlikely to be mono-oxidized, and N is discarded because no observation is made for N. Thus, the

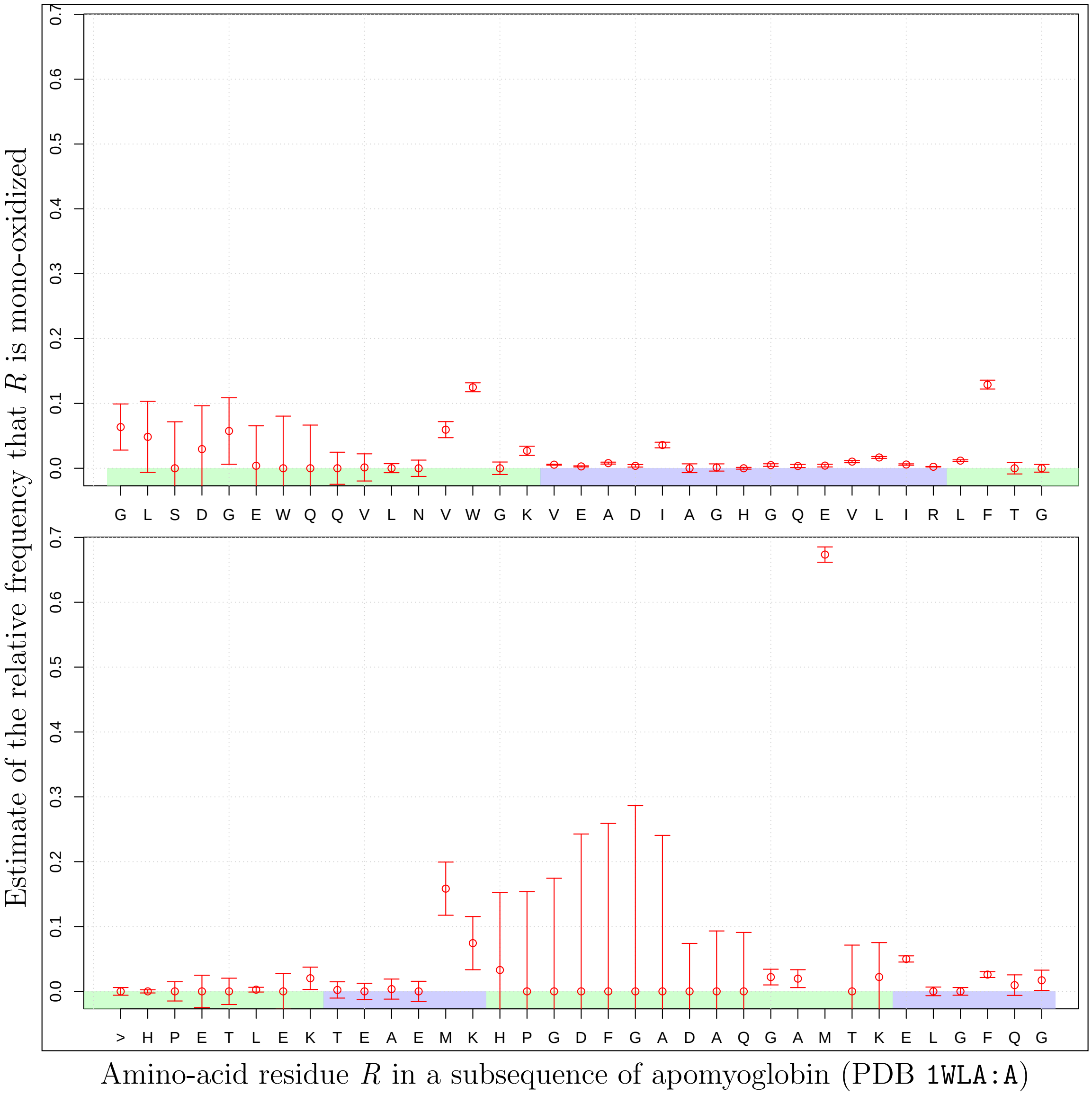


Figure 4.6: The relative frequency that a residue is mono-oxidized as a function of the position of this residue. Algorithm 1 generated this plot.

observed low reactivity of these residues with HO· is consistent with their intrinsic low reactivity with HO·.

We also attempted to use b-ions in addition to using y-ions. Unfortunately, the intensity of a typical b-ion is usually not sufficiently high for quantitating oxidation at subpeptide level. Thus, using b-ions yields worse results than using y-ions.

4.6 Discussion

Traditionally, RP-MS uses MS^2 spectra to identify the residues that are oxidized. We used MS^2 spectra produced from targeted LC-MS/MS to attempt to quantitate mono-oxidation on each residue. We were unable to quantitate the oxidation on every residue of a peptide. However, we presented Algorithm 1 that can quantitate oxidation at subpeptide level. Our algorithm is evaluated on the MS/MS dataset produced by a specially designed RP-MS experiment. In this RP-MS experiment, ultraviolet laser irradiated denatured apomyoglobin during FPOP, and then six runs of targeted MS/MS respectively analyzed six tryptic peptides of apomyoglobin. The evaluation shows the following expected pattern: the estimated oxidation extent before a y-ion index as a function of this y-ion index is monotonically increasing in general. Moreover, the estimated relative frequency that a residue is oxidized approximately matches the expected reactivity of this residue with HO· [28, 17]. Thus, the relative frequency, which is estimated by our algorithm, is approximately correct. Thus, the output produced by our algorithm is likely to be correct.

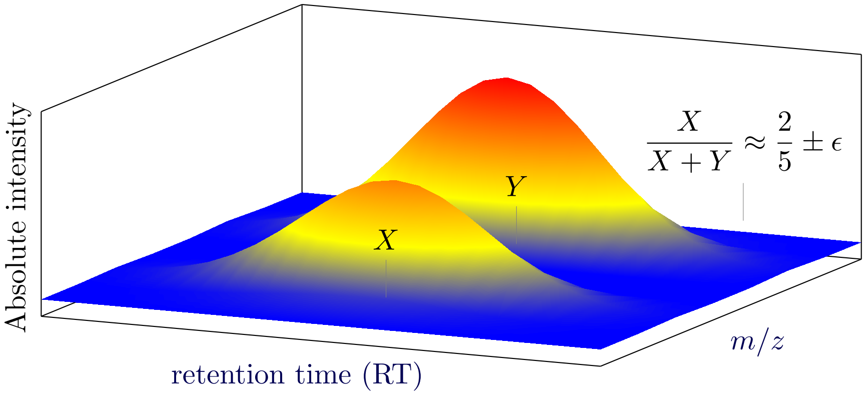
Many aspects of RP-MS are not investigated. First, the evaluation of our algorithm did not consider the other experimental controls of FPOP. For example, these controls include folded protein with irradiation by ultraviolet light and folded protein without irradiation. Moreover, a run of targeted LC-MS/MS only covers one peptide, so multiple runs are required to cover one entire protein. Furthermore, the oxidation site on a peptide should affect the relative frequency that this peptide fragments at a given bond. Thus, almost every peak-area fraction is a biased estimate of a relative frequency. This bias causes systematic errors in quantitation of mono-oxidation at subpeptide level.

In the future, we will first evaluate our algorithm with different experimental controls, then make the specially designed RP-MS experiment less time-consuming and/or less labor-intensive, and finally investigate how the oxidation site on a peptide affects the relative frequency that this peptide fragments at a given bond.

Chapter 5

Estimating the random error in a peak-area fraction given only one run

Part of the 3D-chromatogram of one run of LC-MS/MS
in which X is the peak-area of a product-ion A and Y is the peak-area of a product-ion B
such that B is chemically different from A



Our objective is to estimate the random error ϵ in $\frac{X}{X+Y}$ from only one run of LC-MS/MS. $\frac{X}{X+Y}$ represents the quantity of A relative to B.

Figure 5.1: The graphical abstract of Chapter 5 (hypothetical data used as example).

For any run of LC-MS or of LC-MS/MS, the peak-area of a chemical species is defined as the area under the curve of the extracted-ion chromatogram (XIC) of this chemical species. This peak-area represents the total quantity of this chemical species detected in this run of LC-MS or of LC-MS/MS. Peak-area fraction of a first chemical species relative to a second chemical species is defined as follows: the peak-area of this first chemical species, divided by the sum of the peak-area of this first chemical species and the peak-area of this second chemical species.

Multiple repeated runs can empirically estimate the random error in a measurement of peak-area fraction. Given only one run, this random error seems to be impossible to estimate, because the sample variance of every sample of size one is undefined. However, from some assumptions that are partially supported by evidence in the literature, we mathematically deduced an empirical formula that estimates this random error. We extracted more than 10000 peak-area fractions from a test dataset produced by three runs of LC-MS/MS. Then, for each peak-area fraction in these peak-area fractions and for each applicable run among these three runs, our empirical formula predicted the variance in the single measurement of this peak-area fraction. Then, we compared such predicted variances with the sample variances respectively observed in some pairs of repeated runs. This comparison confirms that each of these peak-area fractions empirically follows the normal distribution with the corresponding predicted variance. Thus, our empirical formula can estimate the random error in one single measurement of peak-area fraction.

Our empirical formula cannot offer every benefit that multiple repeated runs can. For example, multiple repeated runs can respectively provide multiple estimates of the mean of a peak-area, and the average of these estimates is a more precise estimate of this mean. Moreover, the test dataset is produced by only one quadrupole time-of-flight (QTOF) mass spectrometer analyzing only one non-complex sample. However, the more similar a second experiment is to the experiment that produced the test dataset, the more applicable our empirical formula is to the dataset produced by this second experiment. Fortunately, the same mass spectrometer produced both the MS/MS dataset and the test dataset, and two similar samples respectively generated these two datasets. Thus, our empirical formula is very applicable to the MS/MS dataset used in Chapter 4. Thus, Chapter 4 uses Equation (5.15), a key result of this chapter, for estimating the confidence in our quantitation of oxidation. More specifically, for the y-ions that have the same residue sequence and thus the same y-ion index, the peak-area fraction of oxidized y-ions over both oxidized or unoxidized y-ions is the fraction of oxidation that occurred before this y-ion index. Thus, Chapter 4 uses such peak-area fraction to quantitate the extent of oxidation at subpeptide level.

5.1 Motivation

This chapter provides an empirical formula for the following purpose: estimating the random error in a peak-area fraction that is measured once in only one run of LC-MS/MS. A peak-area fraction represents, in a sample, the quantity of a chemical species relative to another chemical species. Thus, our empirical formula can estimate the random error in this relative quantity even if only one run is used for deriving this relative quantity. Our empirical formula performs well on the test dataset. A QTOF mass spectrometer produced the test dataset by analyzing a non-complex sample.

If an instrument similar to this QTOF mass spectrometer analyzes a non-complex sample to produce another dataset, then our empirical formula is likely to be applicable to this other dataset. To produce the MS/MS dataset used in Chapter 4, this same QTOF mass spectrometer analyzed a sample that is almost identical to the test-dataset sample. Thus, our empirical formula is certainly applicable to the MS/MS dataset used in Chapter 4. Thus, in Chapter 4, our empirical formula is used for estimating the random error in quantitation of oxidation at subpeptide level. In Chapter 4, the extent of oxidation before a y-ion index i is estimated to be the following: peak-area fraction of mono-oxidized $\mathtt{y}_i$ over mono-oxidized or unoxidized $\mathtt{y}_i$.

In summary, our empirical formula is unlikely to be applicable to any mass spectrometer analyzing any sample but is certainly applicable to the MS/MS dataset used in Chapter 4. Thus, in Chapter 4, our empirical formula is applied to the MS/MS dataset.

5.2 Related works

Generally, an analytical instrument exhibits both additive error and multiplicative error. Let the random variable ξ be an observed signal intensity. Let ϵ_0 be the additive error in ξ. Let ϵ_1 be the multiplicative error in ξ. Then, mathematically, $\xi \sim \epsilon_1 \cdot \mathrm{E}[\xi] + \epsilon_0$. LC-MS/MS instruments, although highly complex, are characterized by additive error and multiplicative error [21]. Shot noise, also known as Poisson noise, is observed in LC-MS/MS if the quantity of ions detected is an integer representing ion count [3, 12].

In 2004, Anderle et al. [3] developed a noise model to characterize the random error in a peak intensity. This noise model assumes that this random error consists of the following two additive components: a component proportional to the square of the peak intensity and a component proportional to the peak intensity. This noise model is useful for estimating sample preparation noise. Unfortunately, this noise model does not address MS^2 spectra, does not quantitate a first ion relative to a second ion that is chemically different from this first ion, and does not characterize variation in XIC using only one run of LC-MS/MS.

In 2008, Du et al. [12] developed a noise model to characterize the noise in a dataset produced by either QTOF or ion-trap mass spectrometers. This noise model assumes that this noise consists of multinomial noise, Poisson noise, and detector noise. According to this noise model, peaks respectively generated by isotopes follow a multinomial distribution, every such isotopic peak follows a Poisson distribution, and the ability of a detector to detect ions is subject to dead-time effect. This noise model is useful for deisotoping. Unfortunately, this noise model does not consider MS^2 spectra, does not address the potentially different variabilities of noise for repeated runs of LC-MS/MS, and does not quantitate a first ion relative to a second ion that is chemically different from this first ion.

In 2010, Karp et al. [21] proposed a methodology for addressing the accuracy-and-precision in isobaric tags for relative and absolute quantitation (iTRAQ). In this methodology, variance heterogeneity refers to the phenomena that low signals have higher relative variability, and ratio compression refers to the phenomena that ratio of quantities in iTRAQ quantitation is compressed towards 1. They mentioned that variance heterogeneity compromises the precision in iTRAQ quantitation, and that ratio compression compromises the accuracy in iTRAQ quantitation. They proposed the following: a correction factor computed from spiked proteins of known ratios to address ratio compression, and an additive-multiplicative error model with variance-stabilizing normalization to address variance heterogeneity. This methodology is useful for quantitating by iTRAQ a protein when the signal intensity of this protein is low. Unfortunately, this methodology does not address any label-free quantitation, is not generally applicable to the quantitation of any product ion, and cannot characterize the random error in any relative quantity using only one single run of LC-MS/MS.

5.3 Deriving our empirical formula

This section presents our empirical formula for estimating the random error in a peak-area fraction. First, we made some reasonable assumptions partially supported by evidence in the literature. Next, we provided a method for estimating an unknown variable in our empirical formula. This estimation does not require any additional experimental data. Then, we mathematically deduced our empirical formula from these assumptions. Afterwards, we showed that, if some conditions are satisfied, then our empirical formula can be simplified. This simplified version of our empirical formula is used in Chapter 4.

5.3.1 Making and justifying assumptions

We made the following three assumptions.

1 If the first and second scans that are sufficiently far apart in retention time (RT) generated a first and a second mass spectra respectively, then the generation of this first mass spectrum does not significantly affect the generation of this second mass spectrum, and vice versa.
2 The correlation between the peak-area of a peptide species and the peak-area of another peptide species is approximately zero.
3 Shot noise and multiplicative random error constitute the majority of random error in almost every observed mass spectrum. In this chapter, the constant δ is defined as the expected value of this multiplicative random error.

Assumptions 1 to 3 have all been made in the literature. Assumption 1 is implicitly made in [38], because the central limit theorem assumes at least one variant of statistical independence. A stronger version of Assumption 2 is made in [45]. This version of Assumption 2 assumes that, in one scan, the XIC of a peptide species and the XIC of another peptide species are independently generated with respect to each other. Assumption 3 is justified in both [3] and [12].

Assumptions 1 to 3 are all reasonable. Autocorrelation of the generation of mass spectrum should become negligible as RT lag becomes sufficiently large. Thus, Assumption 1 is reasonable. A physical or chemical process that affects multiple molecular entities should affect them independently of each other. Thus, Assumption 2 is reasonable. A random signal that is discrete in nature is almost always characterized by shot noise, and the additive error in the property of a process causes some multiplicative error in the quantity of products generated by this process given that the quantity of reactants consumed by this process varies. Thus, Assumption 3 is reasonable, For example, the quantity of ions that hit a mass detector should be characterized by shot noise. And if the chemical reaction rate is subject to additive random error when the quantity of chemical reactants varies, then the quantity of chemical products should be characterized by multiplicative random error.

Thus, we made these three assumptions to mathematically deduce our empirical formula from these three assumptions.

5.3.2 Estimating the square δ^2 of the multiplicative-random-error constant δ defined in Assumption 3

In this chapter, δ is the multiplicative-random-error constant defined in Assumption 3. In a run of LC-MS or of LC-MS/MS, the calibration function continuously applies the same pressure to the same calibrant. Thus, in the calibration function, the same process should generate different mass spectra respectively at different RTs. Thus, the fluctuation of peak intensity in a mass spectrum as a function of RT should be able to empirically estimate δ^2. Thus, let us take the squared coefficient-of-variation of XIC within an RT window as the window moves. Then, the moving squared coefficient-of-variation of XIC in calibration function can empirically estimate δ^2. Let the random variable R be any sequence of consecutive scans. Let r be any sequence of mass spectra that is approximately generated by R. We estimated the coefficient-of-variation between $\text{TIC}(R_i)$ and $\text{TIC}(R_{i+1})$ as $\left|\dfrac{\text{TIC}(r_i) - \text{TIC}(r_{i+1})}{\text{TIC}(r_i)}\right|$. Then, any coefficient-of-variation whose corresponding R_{i+1} is outside of a given RT range of interest is filtered out. Then, δ^2 is empirically estimated to be the half of the average of the remaining squared coefficients-of-variation. This average is halved because both $\text{TIC}(R_i)$ and $\text{TIC}(R_{i+1})$ are random for any valid i. More precisely, we applied the definition of δ in Assumption 3 on the empirical data produced by the calibration function to obtain Equation (5.1). Thus, Equation (5.1) empirically estimates δ^2.

$$\widehat{\delta^2} \approx \frac{1}{|r'|} \cdot \sum_{i=1}^{|r'|} \left(\frac{1}{2} \cdot \left(\frac{\text{TIC}(r'_i) - \text{TIC}(r'_{i+1})}{\text{TIC}(r'_i)} \right)^2 \right) \tag{5.1}$$

In Equation (5.1), r is a sequence of mass spectra produced by the calibration function and ordered by RT so that r_i is the i^{th}-generated mass spectrum in r, and r' is the shortest substring of r such that the mass spectra of interest are all within the RT range spanned by r'.

Some alternative statistical methods estimated δ^2 by using the same calibration functions. The estimate of δ^2 is relatively constant regardless of which statistical method generated this estimate.

5.3.3 Mathematically deducing our empirical formula from the assumptions

Let A be a chemical species. Let the random variable S be a scan that generates a not-yet-observed mass spectrum s. Assumption 3 implies the following.

$$\mathrm{XIC}(A,S) \overset{\mathsf{app}}{\sim} \mathcal{D}_{AS}\left(\mathrm{E}[\mathrm{XIC}(A,S)], \mathrm{E}[\mathrm{XIC}(A,S)] + (\delta \cdot \mathrm{E}[\mathrm{XIC}(A,S)])^2\right). \tag{5.2}$$

In Equation (5.2), $\mathcal{D}_{AS}(\mu,\sigma^2)$ can be any statistical distribution that has a finite mean of μ and a finite variance of σ^2, and XIC is a function that outputs the absolute intensity of some investigated molecules in a mass spectrum; $\mathrm{XIC}(M,s)$ is the sum of the respective intensities of the peaks generated by M in s, given that s is a mass spectrum, and that M is some investigated molecules. Let the random variable R be a sequence of consecutive scans in a run of LC-MS or of LC-MS/MS. The definition of peak-area implies the following.

$$\text{peak-area}(A,R) = \sum_{S\in R} \mathrm{XIC}(A,S). \tag{5.3}$$

The substitution of Equation (5.2) into Equation (5.3) implies the following.

$$\text{peak-area}(A,R) \overset{\mathsf{app}}{\sim} \sum_{S\in R} \mathcal{D}_{AS}\left(\mathrm{E}[\mathrm{XIC}(A,S)], \mathrm{E}[\mathrm{XIC}(A,S)] + (\delta \cdot \mathrm{E}[\mathrm{XIC}(A,S)])^2\right). \tag{5.4}$$

Assumption 1 implies that the generalized central limit theorem presented in [13, Theorem 7.8] is applicable to Equation (5.4). The result of such application is the following.

$$\text{peak-area}(A,R) \overset{\mathsf{app}}{\sim} \mathcal{N}\left(\mathrm{E}[\text{peak-area}(A,R)], \mathrm{E}[\text{peak-area}(A,R)] + \mathrm{E}[\sum_{S\in R}\left((\delta \cdot \mathrm{XIC}(A,S))^2\right)]\right). \tag{5.5}$$

Let B be a chemical species that is different from A. Let X-and-Y be respectively the peak-areas of A-and-B in a sequence R of scans produced by one run of LC-MS or of LC-MS/MS. Equivalently, let $X \overset{\mathsf{def}}{=} \text{peak-area}(A,R)$ and let $Y \overset{\mathsf{def}}{=} \text{peak-area}(B,R)$. Assumption 2 implies that the covariance between X and Y is small compared with their respective variances. Thus, the application of the multivariate delta method presented in [34] to $X \div (X+Y)$, the application of the Taylor expansion for moments of function of random variables presented in [23, Chapter 4] to the equation resulting from this multivariate delta method, and then the substitution of Equation (5.5) into the equation resulting from this

Taylor expansion results in the following.

$$\frac{X}{X+Y} \overset{\mathsf{app}}{\sim} \mathcal{N}\left(\frac{\mu_X}{\mu_X+\mu_Y}, \left(\frac{\mu_X}{\mu_X+\mu_Y}\right)^2 \cdot \left(\frac{\sigma_X^2}{\mu_X^2} + \frac{\sigma_X^2+\sigma_Y^2}{(\mu_X+\mu_Y)^2} - \frac{2\cdot\sigma_X^2}{\mu_X\cdot(\mu_X+\mu_Y)}\right)\right) \tag{5.6}$$

$$\text{where} \quad \mu_X \overset{\mathsf{def}}{=} \mathrm{E}[\text{peak-area}(A,R)] \tag{5.7}$$

$$\mu_Y \overset{\mathsf{def}}{=} \mathrm{E}[\text{peak-area}(B,R)] \tag{5.8}$$

$$\sigma_X^2 \overset{\mathsf{def}}{=} \mathrm{E}[\text{peak-area}(A,R)] + \mathrm{E}[\sum_{S\in R}\left((\delta\cdot \mathrm{XIC}(A,S))^2\right)] \tag{5.9}$$

$$\sigma_Y^2 \overset{\mathsf{def}}{=} \mathrm{E}[\text{peak-area}(B,R)] + \mathrm{E}[\sum_{S\in R}\left((\delta\cdot \mathrm{XIC}(B,S))^2\right)]. \tag{5.10}$$

5.3.4 Simplifying our empirical formula for use in Chapter 4

In Equation (5.6), if δ^{-2} is large compared with the intensity of XIC in most mass spectra of interest, then

$$\mathrm{E}[\text{peak-area}(A,R)] \gg \mathrm{E}[\sum_{S\in R}\left((\delta\cdot \mathrm{XIC}(A,S))^2\right)], \tag{5.11}$$

$$\mathrm{E}[\text{peak-area}(B,R)] \gg \mathrm{E}[\sum_{S\in R}\left((\delta\cdot \mathrm{XIC}(B,S))^2\right)]. \tag{5.12}$$

Then, the substitution of Equations (5.11) and (5.12) into Equation (5.6) implies the following.

$$\sigma_X^2 \approx \mathrm{E}[\text{peak-area}(A,R)] = \mu_X \quad \text{and} \quad \sigma_Y^2 \approx \mathrm{E}[\text{peak-area}(B,R)] = \mu_Y. \tag{5.13}$$

Then, the substitution of Equation (5.13) into Equation (5.6) implies the following simplification.

$$\frac{X}{X+Y} \overset{\mathsf{app}}{\sim} \mathcal{N}\left(\frac{\mu_X}{\mu_X+\mu_Y}, \left(\frac{\mu_X}{\mu_X+\mu_Y}\right)^2 \cdot \left(\frac{\mu_X}{\mu_X^2} + \frac{\mu_X+\mu_Y}{(\mu_X+\mu_Y)^2} - \frac{2\cdot\mu_X}{\mu_X\cdot(\mu_X+\mu_Y)}\right)\right) \tag{5.14}$$

$$\overset{\mathsf{app}}{\sim} \mathcal{N}\left(\frac{\mu_X}{\mu_X+\mu_Y}, \frac{\mu_X\cdot\mu_Y}{(\mu_X+\mu_Y)^3}\right). \tag{5.15}$$

δ is indeed sufficiently small in the MS/MS dataset that is used in Chapter 4. Thus, Chapter 4 utilizes Equation (5.15) instead of Equation (5.6).

Peptide sequence of $\boldsymbol{P}$	RT in min	m/z	Scans in which MS^2 is performed for $\boldsymbol{P}$
`ALELFRNDIAAK`	[30.7, 31.1]	454.26	10 scans in 1$^{\text{st}}$ run and 10 scans in 2$^{\text{nd}}$ run
`HGTVVLTALGGILK`	[37.6, 38.0]	460.29	10 scans in 2$^{\text{nd}}$ run and 10 scans in 3$^{\text{rd}}$ run
`HGTVVLTALGGILKK`	[34.2, 35.9]	502.98	10 scans in 1$^{\text{st}}$ run and 20 scans in 3$^{\text{rd}}$ run

Table 5.1: Important information extracted from the test dataset. Each of these three peptides is selected for MS^2 in two repeated runs of LC-MS/MS, has an RT range that is defined as the smallest range covering all scans in these two runs, and is used with these two runs as the input to Algorithm 2.

5.4 Testing our empirical formula

5.4.1 Test dataset

A complete iterative-exclusion mass spectrometry (IE-MS) dataset that includes the test dataset was produced by the radical-probe mass spectrometry (RP-MS) experiment described in [43]. In this RP-MS experiment, all runs of LC-MS/MS analyzed the same sample with almost identical configurations. Thus, these runs of LC-MS/MS are repeated.

Let X be the peak-area of a product ion. Let Y be the peak-area of another chemically different product ion. Our empirical formula estimates the random error in the peak-area fraction $X \div (X + Y)$. However, only the runs of LC-MS/MS that select chemically identical precursors for MS^2 can reveal such random error. Thus, the test dataset used for testing our empirical formula is only produced by three runs in this RP-MS experiment. These three runs of LC-MS/MS analyzed the same sample and were repeated with almost identical configurations. In each of these runs, a Synapt QTOF mass spectrometer (Waters, Milford, MA) performed MS/MS by using collision-induced dissociation (CID).

IE-MS avoids selecting the same precursor-ion species for MS^2 in multiple runs. However, the test dataset shows that IE-MS sometimes still selects the same precursor-ion species in two runs. In this sub-experiment, all these precursor-ion species are respectively formed by the three peptide species listed in Table 5.1. For each of these peptide species, Table 5.1 shows the two repeated runs that produced the MS^2 spectra of this peptide species.

5.4.2 Test method

To calculate peak-area, peaks must be detected first. Automated peak detection is both less labor-intensive and less error-prone than manual peak detection. Unfortunately, a typical peak-detection algorithm detects only high-intensity peaks. Thus, we designed a peak-detection algorithm (Algorithm 3) that selects both high-intensity peaks and low-intensity peaks. We manually verified, by careful visual inspection, that our peak-detection algorithm is correct. Basically, our peak-detection algorithm takes as input a sequence of MS^2 spectra and outputs values of m/z. The intensity at any of these values of m/z in any of these MS^2 spectra is mostly generated by one product-ion species, and some peak-areas that respectively have some of these values of m/z pairwise differ by several orders of magnitude.

The interval [30.7 min, 38.0 min] is the smallest RT range that covers all RT ranges listed in Table 5.1. Thus, this RT range is used for estimating δ^2. The three runs described in Table 5.1 respectively have three calibration functions. Section 5.3.2 describes how to estimate δ^2, which is estimated to be 0.00334 from the first run, 0.00525 from the second run, and 0.00555 from the last run. Finally, we estimated δ^2 to be the average of these three individual estimates. Thus, $\widehat{\delta^2} \approx 0.0047$.

Let Φ_1 and Φ_2 be two values of the same peak-area fraction that are respectively observed in two repeated runs. Then, $\text{E}[\Phi_1] = \text{E}[\Phi_2]$. Let p_1 be an estimate of $\text{var}[\Phi_1]$. Let p_2 be an estimate of $\text{var}[\Phi_2]$. Let p_{12} be an estimate of $\text{var}[\Phi_1 - \Phi_2]$. If both p_1 and p_2 are correct, then p_{12} is also correct. Otherwise, p_{12} is likely to be incorrect. Thus, if p_{12} is correct, then both p_1 and p_2 are likely to be correct. Thus, Algorithm 2 uses p_{12} to verify that both p_1 and p_2 are correct.

Let A and B be two different chemical species. Let X and Y represent the quantities of A and B. Let the random-variable Φ_1 be the $X/(X+Y)$ observed in a first run. Let the random-variable Φ_2 be the $X/(X+Y)$ observed in a second run. Let us suppose that these two runs are repeated. If $\Phi_1 \overset{\mathsf{app}}{\sim} \mathcal{N}(\mu, \sigma_1^2)$ and $\Phi_2 \overset{\mathsf{app}}{\sim} \mathcal{N}(\mu, \sigma_2^2)$, then $\Phi_1 - \Phi_2 \overset{\mathsf{app}}{\sim} \mathcal{N}(0, \sigma_1^2 + \sigma_2^2)$. Otherwise, it is unlikely that $\Phi_1 - \Phi_2 \overset{\mathsf{app}}{\sim} \mathcal{N}(0, \sigma_1^2 + \sigma_2^2)$. Thus, if $\Phi_1 - \Phi_2 \overset{\mathsf{app}}{\sim} \mathcal{N}(0, \sigma_1^2 + \sigma_2^2)$, then it is likely that $\Phi_1 \overset{\mathsf{app}}{\sim} \mathcal{N}(\mu, \sigma_1^2)$ and that $\Phi_2 \overset{\mathsf{app}}{\sim} \mathcal{N}(\mu, \sigma_2^2)$ for some μ. Thus, Algorithm 2 assesses our empirical formula by using the following procedure: First, our empirical formula is applied to this first run to estimate σ_1 and then to this second run to estimate σ_2. Then, σ_1 and σ_1 are respectively used to estimate Φ_1 and Φ_2. Finally, the distribution of $\Phi_1 - \Phi_2$, visualized with density plots and Q-Q plots, assesses our empirical formula.

Algorithm 2 test-empirical-formula($r_1, r_2, \mathbb{M}$)

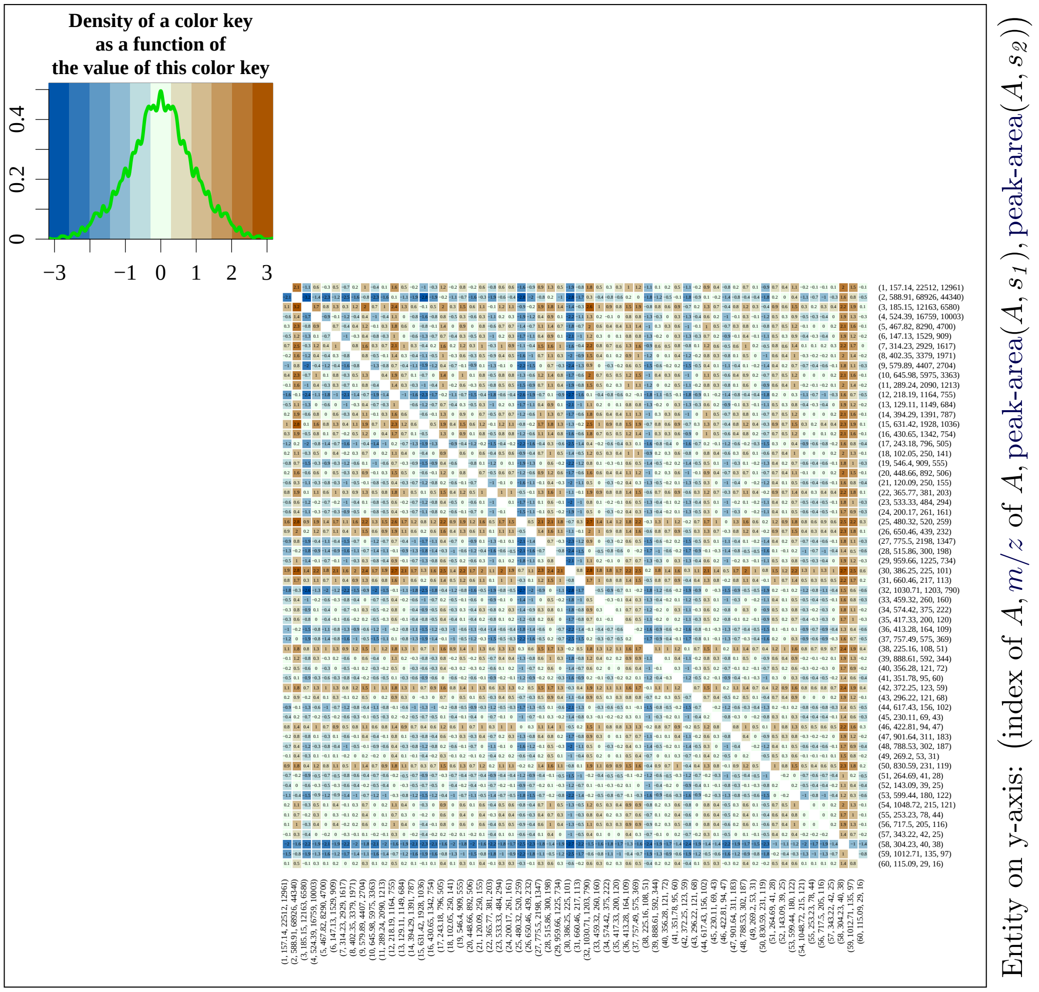
Input: r_1 is a sequence of MS^2 spectra produced by a first run R_1 of LC-MS/MS. r_2 is a sequence of MS^2 spectra produced by a second run R_2 of LC-MS/MS. $\left(R_1, R_2\right)$ is approximately iid. Equivalently, R_1 and R_2 are repeated. $\mathbb{M}$ is a set of many product-ion species generated by one chemical species of precursor ions and detected in both r_1 and r_2. The respective m/z of these product-ion species are respectively the m/z in detect-peaks(r_1, r_2) that is presented in Algorithm 3. Table 5.1 summarizes the three inputs that are given one-by-one to this algorithm.

Output: Each input generates a heatmap of the observed deviation divided by the predicted standard deviation. Examples of such heatmaps are Figures 5.2 to 5.4. Each input generates a Q-Q plot of the intensities in the heatmap as well. Examples of such Q-Q plots are in Figure 5.5.

```
Let $R$ be the distribution such that $(R_1, R_2) \overset{\text{app}}{\sim} R$.
for $i \in \{1, 2\}$ do
    for $A \in \mathbb{M}$, $B \in \mathbb{M} \setminus \{A\}$ do                ▷ apply Equation (5.6)
        $\widehat{\mu_X} := \text{peak-area}(A, r_i)$                            ▷ because $X \overset{\text{app}}{\sim} \text{peak-area}(A, R)$
        $\widehat{\sigma^2_X} := \sum_{s \in r_i} \left( \widehat{\delta^2} \cdot (\text{XIC}(A, s))^2 + \text{XIC}(A, s) \right)$
        $\widehat{\mu_Y} := \text{peak-area}(B, r_i)$                            ▷ because $Y \overset{\text{app}}{\sim} \text{peak-area}(B, R)$
        $\widehat{\sigma^2_Y} := \sum_{s \in r_i} \left( \widehat{\delta^2} \cdot (\text{XIC}(B, s))^2 + \text{XIC}(B, s) \right)$
        $\widehat{\mathrm{E}}_i[\Phi_{AB}] := \frac{\widehat{\mu_X}}{\widehat{\mu_X} + \widehat{\mu_Y}}$              ▷ because $\Phi_{AB} \overset{\text{app}}{\sim} \frac{X}{X+Y}$
        $\widehat{\text{var}}[\widehat{\mathrm{E}}_i[\Phi_{AB}]] := \left( \frac{\widehat{\mu_X}}{\widehat{\mu_X} + \widehat{\mu_Y}} \right)^2 \cdot \left( \frac{\widehat{\sigma^2_X}}{\widehat{\mu^2_X}} + \frac{\widehat{\sigma^2_X} + \widehat{\sigma^2_Y}}{(\widehat{\mu_X} + \widehat{\mu_Y})^2} - \frac{\widehat{\sigma^2_X}}{\widehat{\mu_X} \cdot (\widehat{\mu_X} + \widehat{\mu_Y})} \right)$
    end for
end for
for $A \in \mathbb{M}$, $B \in \mathbb{M} \setminus \{A\}$ do
    $\widehat{z}_{AB} := \frac{\widehat{\mathrm{E}}_1[\Phi_{AB}] - \widehat{\mathrm{E}}_2[\Phi_{AB}]}{\sqrt{\widehat{\text{var}}[\widehat{\mathrm{E}}_1[\Phi_{AB}]] + \widehat{\text{var}}[\widehat{\mathrm{E}}_2[\Phi_{AB}]]}}$     ▷ $\widehat{z}_{AB}$ denotes $\frac{\text{observed deviation}}{\text{predicted standard deviation}}$.
end for
Plot $\widehat{z}_{AB}$ as a function of $A$ and $B$. This plot is in Figures 5.2 to 5.4.
Create normal Q-Q plot for $\{\widehat{z}_{AB} : \text{index of } A < \text{index of } B\}$. This plot is in Figure 5.5.
```

Heatmap of $\widehat{z}$ as a function of a pair (A, B) of chemical species of product ions where $\widehat{z}_{AB} \overset{\text{def}}{=} \dfrac{\widehat{\mathrm{E}}_1[\Phi_{AB}] - \widehat{\mathrm{E}}_2[\Phi_{AB}]}{\sqrt{\widehat{\mathrm{var}}[\widehat{\mathrm{E}}_1[\Phi_{AB}]] + \widehat{\mathrm{var}}[\widehat{\mathrm{E}}_2[\Phi_{AB}]]}}$ and $\Phi_{AB} \overset{\text{def}}{=} \dfrac{\text{peak-area}(A, S)}{\text{peak-area}(A, S) + \text{peak-area}(B, S)}$.

(S_1, S_2) is a pair of repeated runs of LC-MS/MS observed as (s_1, s_2) where $(S_1, S_2) \overset{\text{iid}}{\sim} S$.
The peptide-species $\boldsymbol{P}$ whose sequence is `ALELFRNDIAAK` generated both A and B.



Entity on x-axis: $\big(\text{index of } B, m/z \text{ of } B, \text{peak-area}(B, s_1), \text{peak-area}(B, s_2)\big)$

Figure 5.2: A heatmap generated by Algorithm 2. Φ_{AB} denotes the fraction of A in a mix of both A and B; $\widehat{z}_{AB}$ denotes $\dfrac{\text{first estimate of } \mathrm{E}[\Phi_{AB}] \ - \ \text{second estimate of } \mathrm{E}[\Phi_{AB}]}{\sqrt{\text{estimated variance of first estimate} \ + \ \text{estimated variance of second estimate}}}$ or equivalently $\dfrac{\text{predicted } \mathrm{E}[\Phi_{AB}] \ - \ \text{true } \mathrm{E}[\Phi_{AB}]}{\text{predicted } \sqrt{\mathrm{var}[\Phi_{AB}]}}$.

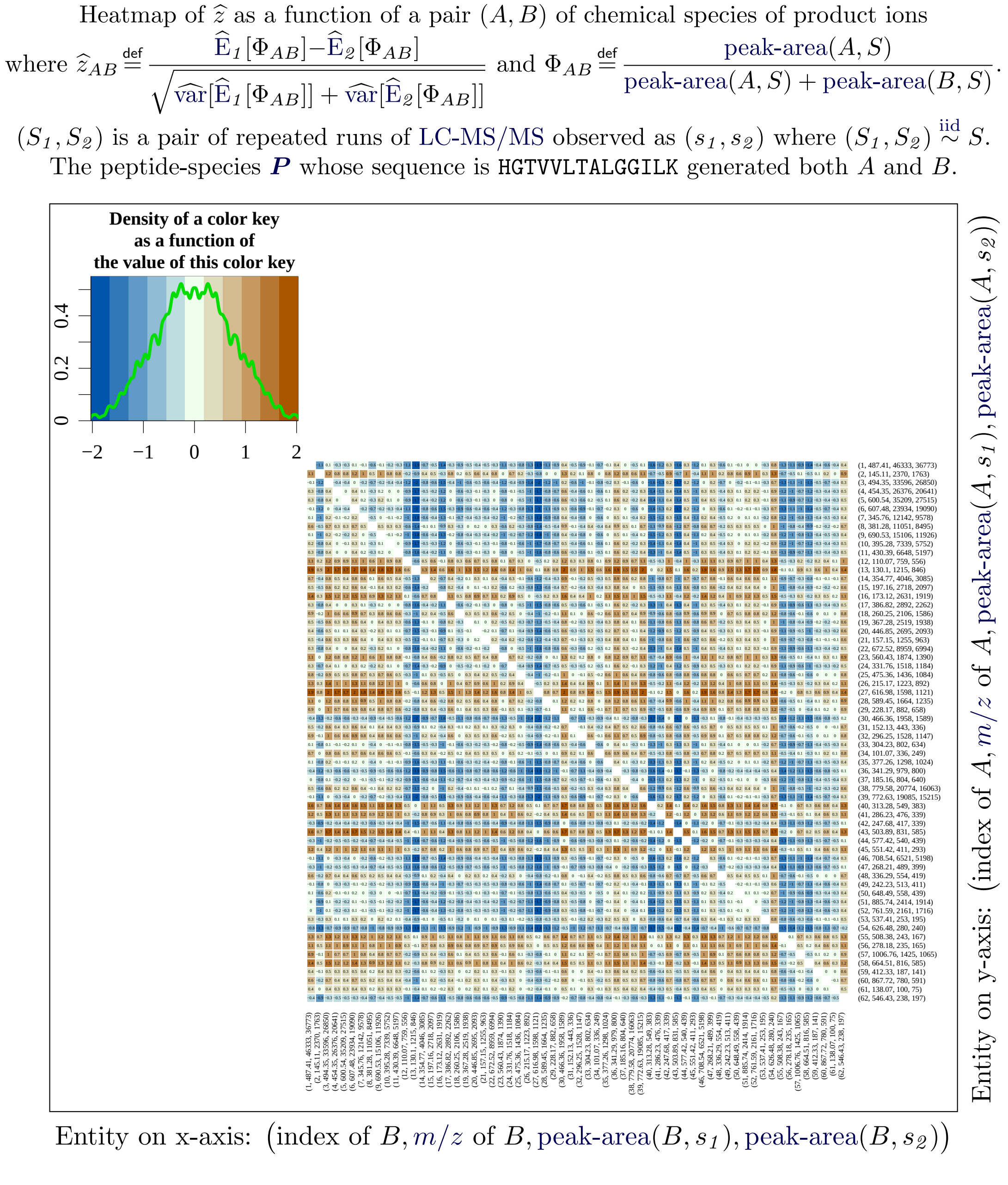


Figure 5.3: A heatmap generated by Algorithm 2. Φ_{AB} denotes the fraction of A in a mix of both A and B; $\widehat{z}_{AB}$ denotes $\frac{\text{first estimate of } \mathrm{E}[\Phi_{AB}] \;-\; \text{second estimate of } \mathrm{E}[\Phi_{AB}]}{\sqrt{\text{estimated variance of first estimate } + \text{ estimated variance of second estimate}}}$ or equivalently $\frac{\text{predicted } \mathrm{E}[\Phi_{AB}] \;-\; \text{true } \mathrm{E}[\Phi_{AB}]}{\text{predicted } \sqrt{\mathrm{var}[\Phi_{AB}]}}$.

Heatmap of $\widehat{z}$ as a function of a pair (A, B) of chemical species of product ions where $\widehat{z}_{AB} \overset{\text{def}}{=} \dfrac{\widehat{\mathrm{E}}_1[\Phi_{AB}] - \widehat{\mathrm{E}}_2[\Phi_{AB}]}{\sqrt{\widehat{\mathrm{var}}[\widehat{\mathrm{E}}_1[\Phi_{AB}]] + \widehat{\mathrm{var}}[\widehat{\mathrm{E}}_2[\Phi_{AB}]]}}$ and $\Phi_{AB} \overset{\text{def}}{=} \dfrac{\text{peak-area}(A, S)}{\text{peak-area}(A, S) + \text{peak-area}(B, S)}$.

(S_1, S_2) is a pair of repeated runs of LC-MS/MS observed as (s_1, s_2) where $(S_1, S_2) \overset{\text{iid}}{\sim} S$.

The peptide-species $\boldsymbol{P}$ whose sequence is HGTVVLTALGGILKK generated both A and B.

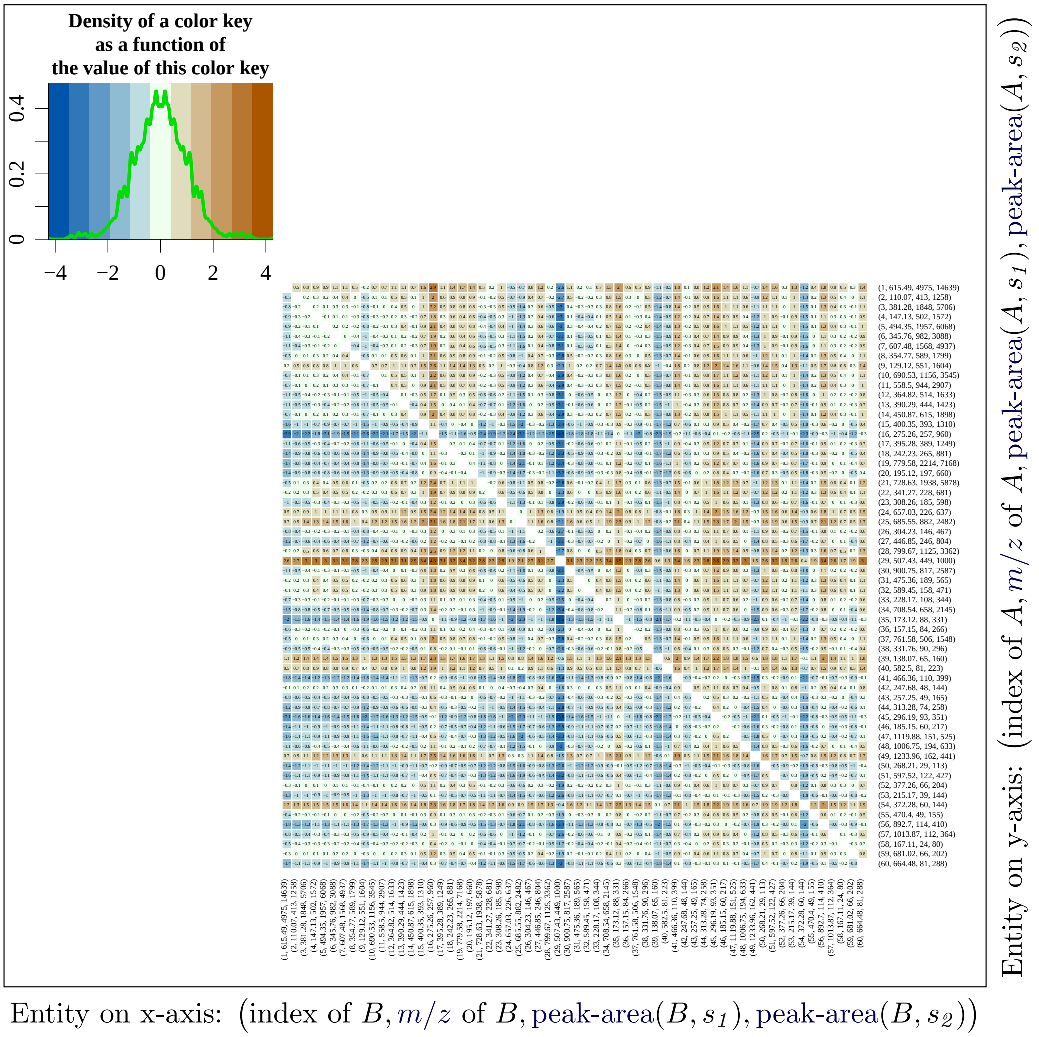


Figure 5.4: A heatmap generated by Algorithm 2. Φ_{AB} denotes the fraction of A in a mix of both A and B; $\widehat{z}_{AB}$ denotes $\dfrac{\text{first estimate of } \mathrm{E}[\Phi_{AB}] \;-\; \text{second estimate of } \mathrm{E}[\Phi_{AB}]}{\sqrt{\text{estimated variance of first estimate} \;+\; \text{estimated variance of second estimate}}}$ or equivalently $\dfrac{\text{predicted } \mathrm{E}[\Phi_{AB}] \;-\; \text{true } \mathrm{E}[\Phi_{AB}]}{\text{predicted } \sqrt{\mathrm{var}[\Phi_{AB}]}}$.

5.4.3 Test result

In the test dataset, we observed three pairs of runs of LC-MS/MS. For each of these three pairs, at least one precursor-ion species is selected for MS^2 by both runs. Algorithm 2 takes as input these two runs and the product-ion species observed in these two runs. Algorithm 2 outputs a heatmap (see Figures 5.2 to 5.4). The values of the color key in Figure 5.2 very closely follow the standard normal. The values of the color key in Figure 5.3 closely follow the standard normal, because the distribution of these values is slightly less heavy-tailed than the standard normal. The values of the color key in Figure 5.4 closely follow the standard normal, because the distribution of these values is slightly more heavy-tailed than the standard normal.

In each heatmap (Figures 5.2 to 5.4), the intensities are evenly distributed in a typical random subregion of this heatmap. Thus, the skewness in each heatmap is approximately zero. The heatmap in Figure 5.2 has no outlier. The heatmap in Figure 5.3 has no outlier. The heatmap in Figure 5.4 has only one weak outlier. This weak outlier is at the 29th row, or equivalently the 29th column, of this heatmap.

For any A_1, any A_2, any B_1, and any B_2,

$$\frac{A_1}{A_1 + A_2} - \frac{B_1}{B_1 + B_2} \equiv -\left(\frac{A_2}{A_1 + A_2} - \frac{B_2}{B_1 + B_2}\right).$$

Thus, in each heatmap in Figures 5.2 to 5.4, the upper right triangle is symmetric to the additive inverse of the lower left triangle, and vice versa. Thus, the plot of the density of a color key as a function of the value of this color key is always symmetric with respect to the zero of this value. Thus, the skewness in the distribution of the values of the color key cannot be assessed in any heatmap in Figures 5.2 to 5.4. However, this skewness can be assessed in one of these two triangles. Thus, for each heatmap in Figures 5.2 to 5.4, Algorithm 2 selected only the intensities in the upper right triangle of this heatmap to generate Figure 5.5. Figure 5.5 shows that the intensities in every heatmap approximately follow the standard normal.

Each heatmap in Figures 5.2 to 5.4 has more than 100 degrees of freedom. Thus, Figures 5.2 to 5.4 have in total more than 300 degrees of freedom. And our model from which we derived our empirical formula does not have any free parameter. Thus, our empirical formula is not subject to overfitting. Thus, in Figure 5.5, the approximate match between the observed distribution and the expected standard normal is significant.

To further prove the significance of this approximate match, we repeated the following procedure four times: First, we randomly selected from our test dataset some MS^2

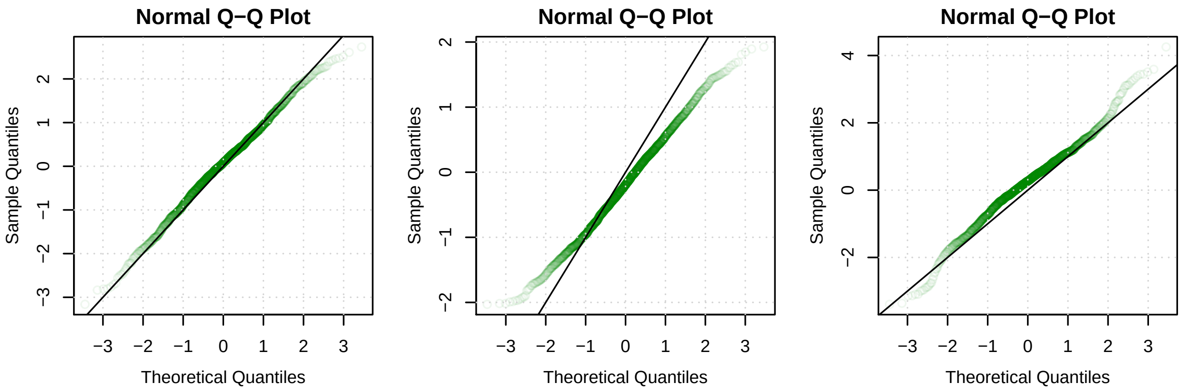


Figure 5.5: The Q-Q plots generated by Algorithm 2. These Q-Q plots correspond to Figures 5.2 to 5.4 respectively from left to right. Each of these Q-Q plots is generated with only the intensities above the diagonal of the corresponding heatmap.

spectra generated by a mixture of pairwise different precursors. Then, we evaluated our empirical formula on these MS^2 spectra. Next, to visualize the result of such evaluation, we constructed a heatmap that is similar to the heatmap shown in Figure 5.2. In each heatmap constructed with this procedure, the intensities are not bell-shaped. Moreover, more than 20% of these intensities are not in the z-score range between -5 and 5. Thus, this approximate match is unlikely to occur by chance.

The calculation of both peak-area and XIC runs in time that is linear with respect to input size. Thus, the running time for evaluating our empirical formula is linear with respect to input size.

5.5 Discussion

Let A and B be two different chemical species. In one given run of LC-MS/MS, let X be the peak-area of A and let Y be the peak-area of B. X and Y denote quantity of A and quantity of B that are both detected in this run of LC-MS/MS respectively. Let us suppose that the $\frac{X}{X+Y}$ of the same sample are observed in multiple repeated runs of LC-MS/MS. Then, the multiple $\frac{X}{X+Y}$, observed from these multiple repeated runs respectively, all estimate the expected value of $\frac{X}{X+Y}$. The expected value of $\frac{X}{X+Y}$ represents the quantity of A relative to B in this same sample. However, every observed $\frac{X}{X+Y}$ is characterized by some random error because every run of LC-MS/MS is inherently stochastic.

Sample variance is undefined for one observation. Similarly, empirical estimation of random error is undefined for one run. Thus, if only one run of LC-MS/MS is used for estimating $\frac{X}{X+Y}$, then the estimation of the random error in this estimate is challenging. However, from some reasonable assumptions that are partially supported by evidence in the literature, we mathematically deduced an empirical formula that estimates, by using only one run of LC-MS/MS, the random error in such $\frac{X}{X+Y}$.

We tested our empirical formula with some pairs of repeated runs of LC-MS/MS. Our empirical formula estimated the random error in $\frac{X}{X+Y}$ for more than 10000 (X, Y) pairs. Both X and Y in these pairs assumed values from below 100 to above 40000. Then, the estimated random errors are compared with the actual random errors observed in two repeated runs. This comparison confirms that our empirical formula can approximately estimate the random error in $\frac{X}{X+Y}$. Our empirical formula is not extensively tested on multiple datasets that are respectively produced by multiple LC-MS/MS instruments, However, our empirical formula is likely to be applicable to a dataset that is produced by a similar instrument analyzing a non-complex sample.

Our work has several limitations. First, compared with a QTOF mass spectrometer, other mass spectrometers, such as Fourier transform ion cyclotron resonance (FTICR) mass spectrometer, have different working mechanisms. Thus, our empirical formula may not be applicable to an arbitrary dataset. Moreover, our empirical formula estimates only the random error in measuring the quantity of a chemical species relative to another chemical species. Thus, our empirical formula does not estimate the random error in measuring the absolute quantity of any chemical species, does not address any systematic error, and cannot reduce the random error in the estimate of the mean value of $\frac{X}{X+Y}$ in the same way as repeated runs. Despite all these limitations, our empirical formula is still useful. Let us suppose that, by using only one run of LC-MS/MS, a QTOF mass spectrometer analyzed a non-complex sample. Then, our empirical formula can estimate the random error in the measured quantity of a chemical species in this sample relative to another chemical species.

Chapter 6

Caveats about using MS^E for RP-MS

One run of targeted LC-MS/MS usually can cover only one peptide. However, one run of LC-MS^E covers all peptides in the sample analyzed by this run. Thus, we hypothesized that MS^E can improve the spatial resolution of radical-probe mass spectrometry (RP-MS), because MS^E could make RP-MS at subpeptide resolution less labor-intensive and less time-consuming if our hypothesis is true. Unfortunately, our hypothesis is wrong. However, we learned some important lessons that can be shared. This chapter presents the background on MS^E, an MS^E dataset, a lower bound on the interference to desired signal in MS^E spectra, how the MS^E dataset failed to confirm our hypothesis, and why our hypothesis is wrong. Past works related to RP-MS are presented in Section 4.2 and are thus omitted in this chapter.

6.1 Background of MS^E

MS^E, a technology in mass spectrometry, was pioneered by the Waters Corporation [35]. The superscripted letter E in MS^E stands for varying levels of energy. In MS^E, the collision-induced dissociation (CID) alternates between the low-energy mode and the high-energy mode. In low-energy mode, collision energy is low. Thus, the percentage of precursor ions that fragment and subsequently become product ions is low. Thus, low-energy CID produces MS^1-like spectra. In high-energy mode, collision energy is high. Thus, the percentage of precursor ions that fragment and subsequently become product ions is high. Thus, high-energy CID produces MS^2-like spectra. In MS^E, all molecules coming from the inlet of a mass spectrometer are selected for fragmentation regardless of CID mode. Thus, the precursor selectivity in MS^E is low.

MS^E has several advantages compared with MS/MS. First, MS/MS selects at each retention time (RT) only the precursors that satisfy certain predefined conditions for MS^2. This satisfaction is highly non-reproducible. However, MS^E selects at each RT all precursors for high-energy CID. Thus, precursor selectivity is relatively constant across MS^E experiments but varies across MS/MS experiments. Thus, the result generated by MS^E is generally more reproducible than the result generated by MS/MS. Moreover, MS/MS selects at each RT only the precursors that are within a narrow window of m/z for MS^2. However, MS^E selects at each RT all precursors for high-energy CID. Thus, the analysis by MS^E is more comprehensive than the analysis by MS/MS.

Unfortunately, given the same precursors to be selected for either high-energy CID or MS^2, MS^E selects a larger quantity of more-chemically-heterogeneous precursors than MS/MS would select. Thus, MS^2-like spectra produced by MS^E are both more complex and noisier than MS^2 spectra produced by MS/MS.

Mass spectra produced by one run of MS^E can identify endogenous metabolites in rat urines [35]. Moreover, appropriate processing of MS^E spectra can enhance the discovery of metabolites [5]. Unfortunately, MS^E has been used for only characterizing small metabolites. Thus, performance of MS^E for protein mass spectrometry (MS) is unknown, and performance of MS^E for RP-MS is completely unknown. However, one run of MS^E can potentially cover all peptides that would require multiple runs of conventional MS/MS to cover. Thus, evaluating the performance of MS^E for studying proteins is important. For example, to cover a peptide by LC-MS/MS, one run has to target this peptide during the entire RT range of this peptide. Thus, the coverage of multiple peptides of interest requires multiple runs of LC-MS/MS. However, if the MS^2-like spectra produced by MS^E are not much noisier and not much more complex than the MS^2 spectra produced by MS/MS, then the coverage of multiple peptides of interest would require only one run of MS^E. Moreover, MS^E selects all precursors for high-energy CID. Thus, MS^E can potentially cover all oxidized products of this peptide of interest in only one run.

6.2 The MS^E dataset

The MS^E dataset was generated by an RP-MS experiment conducted by Siavash Vahidi and Professor Lars Konermann. The mass spectrometer performed MS^E instead of MS/MS in this RP-MS experiment that is otherwise standard. This RP-MS experiment proceeded as follows: First, fast photochemical oxidation of protein (FPOP) was performed on apomyoglobin (PDB `1WLA`). After this FPOP, some apomyoglobins were covalently modified. Next, trypsin cleaved all apomyoglobins into peptides. Then, these peptides were eluted and thus

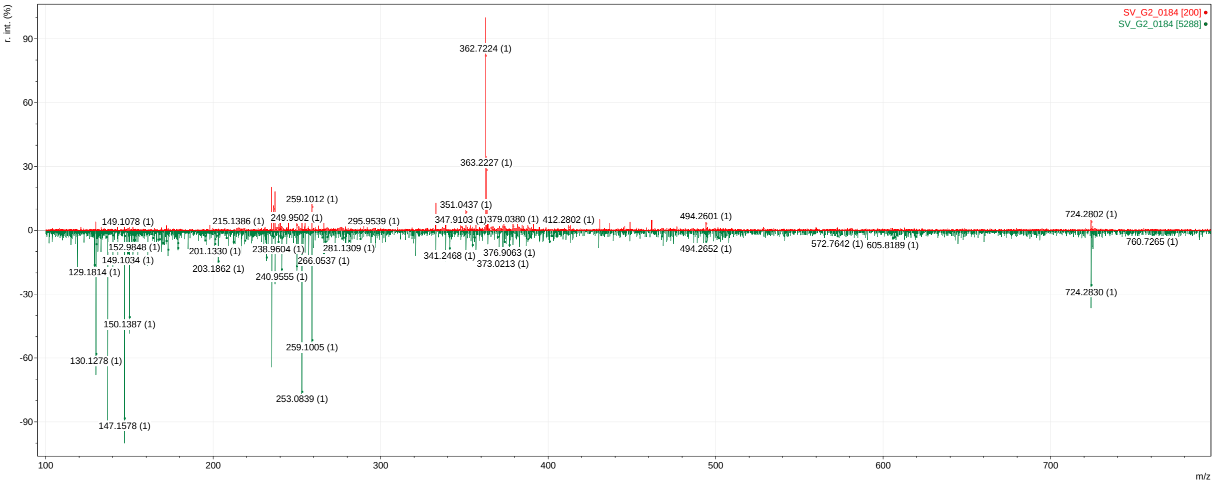


Figure 6.1: A pair of consecutive mass spectra in the $\mathrm{MS^E}$ dataset. First, the low-energy CID of `TEAE(M)(+15.99)K` generated the upper MS^1-like spectrum. Immediately afterwards, the high-energy CID of `TEAE(M)(+15.99)K` generated the lower MS^2-like spectrum. The y-axis represents relative intensity.

separated by high-performance liquid chromatography (HPLC). While HPLC is eluting these peptides, the peptides that HPLC finished eluting were ionized by, analyzed by, and then detected by a Synapt G2 mass spectrometer (Waters, Milford, MA). This mass spectrometer was always in $\mathrm{MS^E}$ mode. The CID energy inside this mass spectrometer was alternating between 20.0eV and 30.0eV. Finally, a sequence of raw $\mathrm{MS^E}$ spectra was generated by this mass spectrometer. We converted this sequence of raw $\mathrm{MS^E}$ spectra into the mzML format by using MSConvert [7].

6.3 A lower bound on the interference-to-signal ratios in the $\mathrm{MS^E}$ dataset

Let us define the following:

1. Let Δm be the resolution of the mass spectrometer.
2. Let M be the length of the continuous m/z range of the mass spectrometer such that almost all peak intensities are within this range. The unit of m/z is dalton.
3. Let $n{+}1$ be the number of residues in a peptide of interest.

4. Let r be the following proportion in an MS^1-like spectrum: the sum of the intensities in the desired m/z intervals to the sum of all intensities. Let us suppose that $M=$ 1000Da. Let us suppose that the desired signal intensity can be in the following m/z intervals: [200,200.1] and [201,201.1]. Then, $r=((200.1-200)+(201.1-201))/1000$.
5. Signaling peaks are the peaks that we are interested in. Signal is the sum of the respective intensities of all signaling peaks. Noisy peaks are the peaks that we are not interested in. Noise is the sum of the respective intensities of all noisy peaks. An interfering peak is a noisy peak whose m/z overlaps with the m/z of any signaling peak. Interference is the sum of the respective intensities of all interfering peaks.

Let us make the following optimistic assumptions.

1. The respective m/z of noisy peaks are evenly distributed in a range of length M.
2. Every precursor ion forms at most one singly charged y-ion. This y-ion always has two isotopes.
3. The accuracy of the mass spectrometer is perfect.

For both MS^1-like spectra and MS^2-like spectra, r denotes the ratio of signal to noise. In RP-MS, signaling peaks are generated by only oxidized or unoxidized y-ions. These y-ions are formed by only mono-oxidized precursors in a typical RT range of interest. For each $i\in[1...n]$, $\mathtt{y}_i$ can be either unoxidized or mono-oxidized and generates two isotopic peaks. Thus, $\mathtt{y}_i$ can generate $2\times2\times n$ signaling peaks. Thus, the signal is within a noncontiguous m/z range of $2\times2\times n\times\Delta m$, because each signaling peak has a width of Δm. Thus, $\frac{1-r}{M}\div\frac{r}{2\times2\times n\times\Delta m}$ denote the ratio of interference to signal, because all noisy peaks are distributed over an m/z range of length M.

The following is observed in the MS$^{\mathrm{E}}$ dataset. $M\approx1000$Da because almost all peaks are in the m/z range from 100Da to 1100Da. $\Delta m\approx0.1$Da. $n=10$ for a typical tryptic peptide of apomyoglobin (PDB `1WLA`). $r\approx0.01$ for a typical mass spectrum, although the respective r of two mass spectra can differ by orders of magnitude. Thus, $\frac{1-r}{M}\div\frac{r}{2\times2\times n\times\Delta m}\approx\frac{2}{5}$. Thus, on average, the interference is 40% of the signal.

Worse still, our assumptions are overly optimistic. For example, high-energy CID can generate an ion that is not a standard y-ion, sources other than irrelevant precursor ions can generate noisy peaks, and the accuracy of the mass spectrometer is not perfect. Thus, $y'_i\div(y_i+y'_i)$ as a function of i is unlikely to be generally increasing, where y_i and y'_i are defined in Section 4.4.

6.4 Negative results on the MS$^{\mathrm{E}}$ dataset

We attempted to reduce interference and to amplify the signal. Unfortunately, even our optimistic assumptions imply that the ratio of interference to signal is at least 40%. In reality, we observed that, for almost all signaling peaks, the ratio of interference to signal is much higher than this lower bound of 40%. Moreover, different y-ions respectively generated by different precursors sometimes overlap with each other in both m/z and RT. We observed that, as i increases, $(y'_i) \div (y_i + y'_i)$ randomly fluctuates instead of generally increasing, presumably because y_i and/or y'_i are subject to too much interference. The MS^2-like spectrum in Figure 6.1 is one of the best-quality MS^2-like spectra in the MS$^{\mathrm{E}}$ dataset. Still, in the MS^2-like spectrum in Figure 6.1, only $\mathtt{y}_1$, $\mathtt{y}'_2$, and $\mathtt{y}'_6$ can be detected by meticulous and labor-intensive visual inspection after zooming into the respective m/z of these y-ions.

6.5 Discussion about the negative results

RP-MS that uses MS$^{\mathrm{E}}$ is both less time-consuming and less labor-intensive than RP-MS that uses targeted MS/MS. Thus, we attempted to use MS$^{\mathrm{E}}$ for improving the spatial resolution of RP-MS. Unfortunately, our attempt failed, presumably because the MS^2-like spectra produced by MS$^{\mathrm{E}}$ have too much noise-induced interference. By making several optimistic assumptions, we established a lower bound on this interference. Methods for reducing this interference may exist. Still, we suspect that current MS$^{\mathrm{E}}$ technology cannot reliably quantitate most product ions generated by high-energy CID.

The additional dataset described in [32] is also generated by a Synapt G2 mass spectrometer that runs in MS$^{\mathrm{E}}$ mode. Thus, we looked at this additional dataset. The experiment that generated this additional dataset has the following characteristics compared with the MS$^{\mathrm{E}}$ dataset: First, the duration of each scan was increased to produce mass spectra with higher mass resolution. Second, peptides are eluted more slowly to better separate these peptides. Third, CID seems to be more optimized. Presumably because of these characteristics, we can manually identify some product ions in this additional dataset. Unfortunately, the MS^2-like spectra in this additional dataset are generally still too noisy and too complex. Thus, we can manually quantitate only very few product ions of interest in this dataset. Thus, MS$^{\mathrm{E}}$ currently seems to be unable to improve the spatial resolution of RP-MS.

Chapter 7

Conclusion

The spatial resolution of the quantitation of oxidation by radical-probe mass spectrometry (RP-MS) is low. The low spatial resolution of such quantitations result in the low spatial resolution of the solvent-accessible surface area (SASA) derived from such quantitations. The low spatial resolution of this SASA results in the low spatial resolution at which protein folding is studied. We showed that targeted LC-MS/MS can improve the spatial resolution of such quantitations of oxidation. Moreover, we designed an algorithm that automates such quantitations of oxidation at this improved spatial resolution. MS/MS can fragment a mono-oxidized peptide into the suffixes of this peptide. Thus, one such suffix is oxidized if and only if the oxidation site on this peptide is on this suffix. Thus, one such suffix of length i is oxidized if and only if this oxidation site is in the last i amino-acid residues of this peptide. Let ϕ_i be the relative frequency that one such suffix of length i is oxidized. Without loss of generality, let $i > j$. Then, $\phi_i - \phi_j$ denotes the relative frequency that the oxidation site in a given mono-oxidized peptide is between the i^{th}-last and $(j+1)^{\text{th}}$-last amino-acid residues of this peptide. Thus, $\phi_i - \phi_j$ can be used for quantitating oxidation at subpeptide level, and $\phi_i - \phi_{i-1}$ can be used for quantitating oxidation at residue level. We evaluated our algorithm on an MS/MS dataset, most of which is produced by six runs of targeted MS/MS. Our algorithm quantitated oxidation near residue level. The extents of oxidation computed by our algorithm agree with the corresponding theoretical extents of oxidation. Thus, our algorithm is sufficiently correct. The throughput of targeted LC-MS/MS is low. Also, the fragmentation chemistry in MS^2 can result in a bias in the oxidation quantitated by our algorithm. However, our algorithm is still sufficiently useful.

However, random errors exist in such quantitation of oxidation. Worse yet, only multiple repeated runs can empirically estimate such random errors, but we have only one run of targeted LC-MS/MS per peptide. To estimate such random errors using insufficient experimental data, we made some assumptions partially supported by evidence in the literature. Then, from these assumptions, we mathematically deduced an empirical formula. Our empirical formula estimates the random error in the peak-area fraction that is calculated from only one run of LC-MS/MS. A peak-area fraction represents, in a sample of interest, the quantity of a type of molecule relative to another type of molecule. Peak-area fraction is a generalized version of ϕ_i. Three repeated runs of LC-MS/MS confirmed that our empirical formula is sufficiently correct. These three runs were all performed by a quadrupole time-of-flight (QTOF) mass spectrometer that analyzed a non-complex sample. Multiple runs provide more information than one run. Thus, one run generally cannot replace repeated runs. For example, multiple repeated runs result in multiple estimates of the same expected value of a peak-area fraction. Then, the average of these multiple estimates has less random error than any one of these multiple estimates. However, our empirical formula is still sufficiently useful.

The throughput of targeted MS/MS is lower than the throughput of $\mathrm{MS^E}$ by orders of magnitude. Thus, we hypothesized that $\mathrm{MS^E}$ can also improve the spatial resolution of RP-MS. Unfortunately, an $\mathrm{MS^E}$ dataset shows that our hypothesis is likely to be wrong. Moreover, an additional $\mathrm{MS^E}$ dataset shows that our hypothesis is very likely to be wrong.

$\mathrm{MS^E}$ does not seem to be able to achieve the purpose of improving the spatial resolution of RP-MS. Thus, in the future, we will try some alternative approaches for this purpose. Ideally, these alternative approaches should be neither labor-intensive nor time-consuming. For example, the following experimental methods all outperform $\mathrm{MS^E}$ in protein identification: ion mobility spectrometry (IMS) assisted $\mathrm{MS^E}$ (HD-$\mathrm{MS^E}$) [11], ultra-definition $\mathrm{MS^E}$ (UD-$\mathrm{MS^E}$) [11], and multiplexed MS/MS [14]. Thus, these methods can be the basis of these alternative approaches.

In the future, we will also test our empirical formula on additional datasets. For example, one such additional dataset can be produced by a mass spectrometer of another type, and this mass spectrometer can analyze a complex sample to produce this dataset. In isobaric tags for relative and absolute quantitation (iTRAQ) experiments, the ratio of iTRAQ reporter ions is also a peak-area fraction. Thus, our empirical formula has the potential to estimate the random error in iTRAQ when fewer-than-expected runs of LC-MS/MS are performed.

APPENDICES

Our custom peak-detection algorithm

The intensity of a candidate peak is assumed to be proportional to the probability that this candidate peak is a true peak. And the respective generations of different peaks are assumed to be pairwise independent. Thus, the product of the respective intensities of different candidate peaks is assumed to be proportional to the probability that these candidate peaks are all true peaks. These candidate peaks can be in different mass spectra but all have the same m/z. The continuous range of applicable m/z is partitioned into connected and pairwise non-overlapping subranges of m/z. Each of these subranges spans 0.01 m/z. Let p be probability that a first subrange contains at least one true peak. Let p' be the probability that another subrange near this first subrange contains at least one true peak. Then, p relative to p' is the likelihood that this first subrange contains at least one significant true peak. If this significant true peak does not overlap with any other peak that has already been picked, then this significant true peak is picked.

Additional justification for using our empirical formula

Table 5.1 summarizes the dataset used for testing our empirical formula. Table 4.1 summarizes the dataset to which our empirical formula is applied. The comparison between these two datasets reveals the following discrepancy: the RT ranges in Table 4.1 are typically much larger than the RT ranges in Table 5.1. Thus, our empirical formula is perhaps not applicable to the dataset summarized in Table 4.1, because the errors in a smaller RT range cannot be extrapolated to the errors in a larger RT range.

Fortunately, our empirical formula is not affected by this potential problem of extrapolation. The following two paragraphs explain why this potential problem is not an issue.

Algorithm 3 detect-peaks(r_1, r_2)

```
Input: $r_1$ is a sequence of MS$^2$ spectra produced by a first run of LC-MS/MS, $r_2$ is a
    sequence of MS$^2$ spectra produced by a second run of LC-MS/MS; $\big(r_1, r_2\big)$ should be
    iid, or equivalently the run that produced $r_1$ and the run that produced $r_2$ should be
    repeated.
Output: a set such that each element in this set is the $m/z$ of a chemical species of
    product ions that is detected in both $r_1$ and $r_2$.
    **We manually validated by visual inspection the output of this algorithm.**
find a subsequence $r'_1$ of $r_1$ and a subsequence $r'_2$ of $r_2$ such that
    1. the RT of any spectrum $s_1$ in $r'_1 \approx$ the RT of any spectrum $s_2$ in $r'_2$ and
    2. the precursor $m/z$ of any $s_1$ in $r'_1 \approx$ the precursor $m/z$ of any $s_2$ in $r'_2$.
$\Delta \overset{\text{def}}{=} \{\frac{-50}{100}, \frac{-49}{100}, ..., \frac{49}{100}, \frac{50}{100}\}$    ▷ Discretized values of $m/z$
$\text{sumLnInts}(\frac{m}{z}, r') \overset{\text{def}}{=} \sum_{s \in r'} \left( \ln \left( 1 + \sum_{p \in s} \left( \text{Intensity}(p) \cdot \mathbb{1}[\frac{m}{z} < m/z \text{ of } p \leq \frac{m}{z} + \frac{1}{100}] \right) \right) \right)$
    ▷ sum of logarithm of peak intensity in every spectrum with add-one Laplace smoothing, where $p$ means peak, $s$ means spectrum, and where $r'$ means sequence of spectra
$\text{lnLike}(\frac{m}{z}, r') \overset{\text{def}}{=} \text{sumLnInts}(\frac{m}{z}, r') - \frac{1}{101} \cdot \sum_{\delta \in \Delta} \big(\text{sumLnInts}(\frac{m}{z} + \delta, r')\big)$
    ▷ log-likelihood at an $m/z$ relative to the background log-likelihood near this $m/z$
$\frac{M}{Z} \overset{\text{def}}{=} \{\frac{1}{100}, \frac{2}{100}, ..., \frac{199999}{100}, \frac{200000}{100}\}$, $\frac{M}{Z}' \overset{\text{def}}{=} \emptyset$    ▷ Discretized values of $m/z$
while $\max_{\frac{m}{z} \in \frac{M}{Z}} \text{lnLike}(\frac{m}{z}, r'_1) > \frac{\ln(200000)}{10} \land |\frac{M}{Z}'| < 1000$ do
    $\frac{m}{z} \overset{\text{def}}{=} \arg\max_{\frac{m}{z} \in \frac{M}{Z}} \text{lnLike}(\frac{m}{z}, r'_1)$
    $\frac{M}{Z} \overset{\text{def}}{=} \frac{M}{Z} \setminus \left( \bigcup_{\delta \in \Delta} \big(\{\frac{m}{z} + \delta\}\big) \right)$    ▷ eliminate peaks that are adjacent in $m/z$
    if $\text{lnLike}(\frac{m}{z}, r'_2) > \frac{\ln(200000)}{10} \land (\forall \frac{m}{z}' \in \frac{M}{Z}' : -3 < \frac{m}{z}' - \frac{m}{z} < 3)$ then
        ▷ select intensity with high relative log-likelihood and naively avoid isotope
        $\frac{M}{Z}' \overset{\text{def}}{=} \frac{M}{Z}' \cup \{\frac{m}{z}\}$
    end if
end while
return $\frac{M}{Z}'$
```

First, each of the peptides in Table 4.1 has most of its peak intensities concentrated at a few short RT intervals. The length of each of these RT intervals is similar to the length of any RT range in Table 5.1 (data not shown). Thus, the union of these RT intervals is still larger, but not much larger, than a typical RT range in Table 5.1. Thus, this problem is alleviated.

Second, we merged the MS^2 spectra of each peptide in Table 4.1. The number of these merged MS^2 spectra is approximately the number of the MS^2 spectra of each peptide in Table 5.1. This merging step does not increase any error. After this merging step, the multiplicative random error captured by the constant δ is still negligible compared with the shot noise captured by Equation (5.15). Also, this multiplicative random error and this shot noise should constitute most of the error in these merged mass spectra. Thus, the shot noise captured by Equation (5.15) is sufficient for characterizing the random errors in the dataset summarized in Table 4.1.

Definitions specific to this thesis

$\boldsymbol{P'}$ is defined as a chemical superspecies of mono-oxidized peptides that are chemically identical up to structural isomerism, where this isomerism is only due to the fact that any site on any residue can be mono-oxidized; for example, $\boldsymbol{P'}$ can be any of the following: {`F[+16]DK`}, {`Y[+16]K`, `Y[+16]K`}, {`Y[+16]K`, `YK[+16]`}, and {`Y[+15.99]K`, `Y[+16]K`}; however, $\boldsymbol{P'}$ cannot be any of the following: {`Y[+16]K`, `Y[+16]K[+16]`}, {`Y[+16]K`, `YK`}, {`Y[+32]K`}, {`Y[+16]K`, `Y[+16]K[+14]`}, and {`Y[+16]K`, `Y[-16]K`}. 35, 37, 41, 72

$\boldsymbol{P}$ is defined as a chemical species of unoxidized peptides; for example, $\boldsymbol{P}$ can be {`FDK`, `FDK`}, can be {`ALELFR`}, cannot be {`FDK`, `FKD`}, and cannot be {`FDK`, `F[+16]DK`}. 35, 41, 54, 57–59, 72

TIC is a function such that $\mathrm{TIC}(s)$ is the sum of the respective intensities at all applicable m/z values in the mass spectrum s; $\mathrm{TIC}(s)$ represents the intensity of all ions detected in s. 51, 72

XIC is a function that outputs the absolute intensity of some investigated molecules in a mass spectrum; $\mathrm{XIC}(M, s)$ is the sum of the respective intensities of the peaks generated by M in s, given that s is a mass spectrum, and that M is some investigated molecules. 35, 50, 52, 53, 56, 61, 72, 73

peak-area is a function that outputs the area under the curve of an extracted-ion chromatogram (XIC); let M be a class of molecules, let r be a set of mass spectra generated by one run of LC-MS or of LC-MS/MS; then, $\text{peak-area}(M, r) \overset{\text{def}}{=} \sum_{s \in r} \text{XIC}(M, s)$, so $\text{peak-area}(M, r)$ represents the total absolute quantity of M detected in r. iii, viii, 35–37, 39, 41, 45–50, 52–59, 61, 69, 73

Definitions in Mathematics

$\mathcal{N}$ is the random-variable generator for the normal distribution. $\mathcal{N}(\mu, \sigma^2)$ has a mean of μ and a variance of σ^2. 37, 41, 52, 53, 55, 73

iid denotes "independent and identically distributed". 56–59, 71

E is the expectation operator; $\mathrm{E}[X]$ is the expected value of the random variable X or equivalently the mean of X. 37, 39, 41, 42, 48, 52, 53, 55–59, 73

var is the variance operator; $\mathrm{var}[X]$ is the statistical variance of the random variable X. 39, 41, 42, 55–59, 73

$\overset{\text{app}}{\sim}$ denotes "is approximately distributed as". 37, 41, 52, 53, 55, 56, 73

Definitions in mass spectrometry

m/z is the mass-to-charge ratio measured in Da. xi, 3, 10, 16–20, 23, 24, 26, 28, 30, 35, 46, 54–59, 64–67, 70–72

HO· is the symbol for hydroxyl radical. x, 8, 9, 11–13, 33, 34, 40, 43, 45

LC-MS/MS is an analytical-chemistry technique using liquid chromatography (LC) and MS/MS such that the outlet of the LC column is connected to the inlet of the MS/MS instrument. 30, 32, 33, 35, 37, 41, 45–49, 51, 52, 54, 56–64, 68, 69, 71, 73

LC-MS is an analytical-chemistry technique using LC and mass spectrometry (MS) such that the outlet of the LC column is connected to the inlet of the MS instrument. 2, 31, 34, 35, 41, 47, 51, 52, 73

MS/MS denotes “tandem mass spectrometry”. iii, vii, viii, x–xii, 2, 3, 5, 10, 20, 21, 33–35, 38, 40, 45, 47, 48, 53, 54, 64, 67–69, 73, 74

MS^1 is the first stage in MS/MS or the only stage in MS; MS^1 generates precursor ions and survey scans. x–xii, 3, 20, 30, 32, 34–37, 39–41, 63, 65, 66, 74

MS^2 is the second stage in MS/MS; MS^2 generates product ions and product scans. xi, xii, 3, 4, 10, 20–22, 24, 26, 28, 30, 32, 34–41, 45, 48, 49, 54–56, 60, 61, 63–68, 71, 72, 74

MS$^{\mathbf{E}}$ is a mass-spectrometry technology pioneered by the Waters Corporation [35]; in MS^E, collision-induced dissociation (CID) alternates between low energy mode which produces MS^1-like spectra and high energy mode which produces MS^2-like spectra. iii, iv, viii, xii, 5, 63–67, 69, 74

mono-oxidation is defined as a modification characterized by a mass shift of approximately 15.99Da to a biomolecule; 15.99Da is approximately equal to the mass of one oxygen atom. viii, 1, 2, 9, 12, 32, 33, 35, 37, 39–41, 45, 74

mono-oxidized denotes “affected by mono-oxidation”. xi, xii, 1–4, 32, 33, 35–44, 48, 66, 68, 72

Acronyms in mass spectrometry

CID Collision-Induced Dissociation 21, 23, 30, 54, 63–67, 74

ESI ElectroSpray Ionization 16–18, 30

FPOP Fast Photochemical Oxidation of Protein vii, xi, 1, 2, 10, 11, 27, 30, 33–35, 37, 45, 64

HPLC High-Performance Liquid Chromatography vii, xi, 10, 14, 15, 30, 65, 75

IE-MS Iterative-Exclusion Mass Spectrometry 54

LC Liquid Chromatography 9, 11, 14, 15, 37, 73

MS Mass Spectrometry vii, xi, 2, 3, 9–11, 14–17, 21, 27, 30, 64, 73, 74

PDB Protein Data Bank iii, xii, 34, 66

PSM Peptide-Spectrum Match vii, xi, 10, 26–28

PTM Post-Translational Modification 27

QTOF Quadrupole Time-Of-Flight 47–49, 54, 62, 69

RP-HPLC Reverse-Phase High-Performance Liquid Chromatography xi, 15, 16

RT Retention Time 2, 15, 16, 30, 34–36, 40, 41, 46, 50, 51, 54, 55, 64, 66, 67, 70–72

SASA Solvent-Accessible Surface Area iii, xi, 1, 2, 5, 8, 9, 11, 12, 68

XIC eXtracted-Ion Chromatogram 35, 47, 49, 51, 53, 73

RP-MS Radical-Probe Mass Spectrometry iii, iv, viii, xi, 9, 10, 27, 29–34, 45, 54, 63, 64, 66–69

References

[1] Mascot database search amino acid reference data. http://www.matrixscience.com /help/aa_help.html. Accessed: 2015-07-05. x, 25

[2] Monday morning oral sessions. Journal of the American Society for Mass Spectrometry, 20(1):S7–S9, 2009. ISSN 1044-0305. doi: 10.1007/BF03215749. URL http://dx.doi.org/10.1007/BF03215749. xi, 31

[3] Markus Anderle, Sushmita Roy, Hua Lin, Christopher Becker, and Keith Joho. Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum. Bioinformatics, 20(18):3575–3582, 2004. 48, 50

[4] Shibdas Banerjee and Shyamalava Mazumdar. Electrospray ionization mass spectrometry: a technique to access the information beyond the molecular weight of the analyte. International journal of analytical chemistry, 2012, 2012. xi, 22

[5] Kevin P Bateman, Jose Castro-Perez, Mark Wrona, John P Shockcor, Kate Yu, Renata Oballa, and Deborah A Nicoll-Griffith. Mse with mass defect filtering for in vitro and in vivo metabolite identification. Rapid communications in mass spectrometry, 21(9):1485–1496, 2007. 64

[6] George V Buxton, Clive L Greenstock, W Phillips Helman, and Alberta B Ross. Critical review of rate constants for reactions of hydrated electrons, hydrogen atoms and hydroxyl radicals (oh/ o- in aqueous solution. Journal of physical and chemical reference data, 17(2):513–886, 1988. 12, 13

[7] Matthew C Chambers, Brendan Maclean, Robert Burke, Dario Amodei, Daniel L Ruderman, Steffen Neumann, Laurent Gatto, Bernd Fischer, Brian Pratt, Jarrett Egertson, et al. A cross-platform toolkit for mass spectrometry and proteomics. Nature biotechnology, 30(10):918–920, 2012. 65

[8] Jiawei Chen, Don L Rempel, Brian C Gau, and Michael L Gross. Fast photochemical oxidation of proteins and mass spectrometry follow submillisecond protein folding at the amino-acid level. Journal of the American Chemical Society, 134(45):18724–18731, 2012. 34

[9] John S Cottrell and U London. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 20(18):3551–3567, 1999. 27

[10] Robertson Craig and Ronald C Beavis. Tandem: matching proteins with tandem mass spectra. Bioinformatics, 20(9):1466–1467, 2004. 27

[11] Ute Distler, Jörg Kuharev, Pedro Navarro, Yishai Levin, Hansjörg Schild, and Stefan Tenzer. Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. Nature methods, 2013. 69

[12] Peicheng Du, Gustavo Stolovitzky, Peter Horvatovich, Rainer Bischoff, Jihyeon Lim, and Frank Suits. A noise model for mass spectrometry based proteomics. Bioinformatics, 24(8):1070–1077, 2008. 48, 49, 50

[13] Rick Durrett. Probability: theory and examples. Cambridge university press, 2010. 52

[14] Jarrett D Egertson, Andreas Kuehn, Gennifer E Merrihew, Nicholas W Bateman, Brendan X MacLean, Ying S Ting, Jesse D Canterbury, Donald M Marsh, Markus Kellmann, Vlad Zabrouskov, et al. Multiplexed ms/ms for improved data-independent acquisition. Nature methods, 10(8):744–746, 2013. 69

[15] Ingvar Eidhammer, Kristian Flikka, Lennart Martens, and Svein-Ole Mikalsen. Computational methods for mass spectrometry proteomics. John Wiley & Sons, 2008. xi, 17, 19, 21

[16] Jimmy K Eng, Ashley L McCormack, and John R Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry, 5(11):976–989, 1994. 27

[17] Brian Craig Gau. The advancement of mass spectrometry-based hydroxyl radical protein footprinting: Application of novel analysis methods to model proteins and apolipoprotein E. WASHINGTON UNIVERSITY IN ST. LOUIS, 2011. 1, 12, 45

[18] Lewis Y Geer, Sanford P Markey, Jeffrey A Kowalak, Lukas Wagner, Ming Xu, Dawn M Maynard, Xiaoyu Yang, Wenyao Shi, and Stephen H Bryant. Open mass spectrometry search algorithm. Journal of proteome research, 3(5):958–964, 2004. 27

[19] Martin Gruebele. Analytical biochemistry: weighing up protein folding. Nature, 468 (7324):640–641, 2010. xi, 11

[20] David M Hambly and Michael L Gross. Laser flash photolysis of hydrogen peroxide to oxidize protein solvent-accessible residues on the microsecond timescale. Journal of the American Society for Mass Spectrometry, 16(12):2057–2063, 2005. 34

[21] Natasha A Karp, Wolfgang Huber, Pawel G Sadowski, Philip D Charles, Svenja V Hester, and Kathryn S Lilley. Addressing accuracy and precision issues in itraq quantitation. Molecular & Cellular Proteomics, 9(9):1885–1897, 2010. 48, 49

[22] Sangtae Kim, Nikolai Mischerikow, Nuno Bandeira, J Daniel Navarro, Louis Wich, Shabaz Mohammed, Albert JR Heck, and Pavel A Pevzner. The generating function of cid, etd, and cid/etd pairs of tandem mass spectra: applications to database search. Molecular & Cellular Proteomics, 9(12):2840–2852, 2010. 27

[23] Eun Sul Lee and Ronald N Forthofer. Analyzing complex survey data. Sage, 2006. 52

[24] Jin-Won Lee and John D Helmann. The perr transcription factor senses h2o2 by metal-catalysed histidine oxidation. Nature, 440(7082):363–367, 2006. xi, 13

[25] Leong Lee, J.L. Leopold, and R.L. Frank. Protein secondary structure prediction using blast and exhaustive rt-rico, the search for optimal segment length and threshold. In Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2012 IEEE Symposium on, pages 35–42, May 2012. doi: 10.1109/CIBCB.2012.6217208. 7

[26] Xiaoyan Li, Zixuan Li, Boer Xie, and Joshua S Sharp. Improved identification and relative quantification of sites of peptide and protein oxidation for hydroxyl radical footprinting. Journal of The American Society for Mass Spectrometry, 24(11):1767–1776, 2013. 34

[27] Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, and Gilles Lajoie. Peaks: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid communications in mass spectrometry, 17(20):2337–2342, 2003. 36, 40, 41

[28] Simin D Maleknia and Kevin M Downard. Advances in radical probe mass spectrometry for protein footprinting in chemical biology applications. Chemical Society Reviews, 43(10):3244–3258, 2014. 45

[29] Simin D Maleknia, Michael Brenowitz, and Mark R Chance. Millisecond radiolytic modification of peptides by synchrotron x-rays identified by mass spectrometry. Analytical chemistry, 71(18):3965–3973, 1999. 33

[30] Simin D Maleknia, Mark R Chance, and Kevin M Downard. Electrospray-assisted modification of proteins: a radical probe of protein structure. Rapid communications in mass spectrometry, 13(23):2352–2358, 1999.

[31] Alan D McNaught and Alan D McNaught. Compendium of chemical terminology, volume 1669. Blackwell Science Oxford, 1997. 19

[32] Jan Muntel, Vincent Fromion, Anne Goelzer, Sandra Maaβ, Ulrike Mäder, Knut Büttner, Michael Hecker, and Dörte Becher. Comprehensive absolute quantification of the cytosolic proteome of bacillus subtilis by data independent, parallel fragmentation in liquid chromatography/mass spectrometry (lc/mse). Molecular & Cellular Proteomics, 13(4):1008–1019, 2014. 67

[33] Pankaja Naik. Essentials of Biochemistry. Jaypee Bros. Medical Publishers, 2012. 27

[34] Gary W Oehlert. A note on the delta method. The American Statistician, 46(1): 27–29, 1992. 52

[35] Robert S Plumb, Kelly A Johnson, Paul Rainville, Brian W Smith, Ian D Wilson, Jose M Castro-Perez, and Jeremy K Nicholson. Uplc/mse; a new approach for generating molecular fragment information for biomarker structure elucidation. Rapid communications in mass spectrometry, 20(13):1989–1994, 2006. 63, 64, 74

[36] Eduard Porta-Pardo, Thomas Hrabe, and Adam Godzik. Cancer3d: understanding cancer mutations through protein structures. Nucleic acids research, page gku1140, 2014. 1

[37] Russel J. Reiter, Daniela Melchiorri, Ewa Sewerynek, Burkhard Poeggeler, Lorneli Barlow-Walden, Jihing Chuang, Genaro Gabriel Ortiz, and Dario AcuaCastroviejo. A review of the evidence supporting melatonin's role as an antioxidant. Journal of Pineal Research, 18(1):1–11, 1995. ISSN 1600-079X. doi: 10.1111/j.1600-079X.1995.tb00133.x. URL http://dx.doi.org/10.1111/j.1600-079X.1995.tb00133.x. 11

[38] Rovshan Sadygov, James Wohlschlegel, Sung Kyu Park, Tao Xu, and John R Yates. Central limit theorem as an approximation for intensity-based scoring function. Analytical chemistry, 78(1):89–95, 2006. 50

[39] Dennis J Selkoe. Cell biology of protein misfolding: the examples of alzheimer's and parkinson's diseases. Nature cell biology, 6(11):1054–1061, 2004. 1

[40] Koichi Tanaka, Hiroaki Waki, Yutaka Ido, Satoshi Akita, Yoshikazu Yoshida, Tamio Yoshida, and T Matsuo. Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. Rapid communications in mass spectrometry, 2(8):151–153, 1988. 16

[41] Rolf Turner. Package iso. 2013. 39, 41

[42] Jan Urban, Nils Kristian Afseth, and Dalibor Štys. Fundamental definitions and confusions in mass spectrometry about mass assignment, centroiding and resolution. TrAC Trends in Analytical Chemistry, 53:126–136, 2014. 23

[43] Siavash Vahidi, Bradley B Stocks, Yalda Liaghati-Mobarhan, and Lars Konermann. Mapping ph-induced protein structural changes under equilibrium conditions by pulsed oxidative labeling and mass spectrometry. Analytical chemistry, 84(21):9124–9130, 2012. 34, 54

[44] Guozhong Xu and Mark R Chance. Hydroxyl radical-mediated modification of proteins as probes for structural proteomics. Chemical reviews, 107(8):3514–3543, 2007. 12, 13

[45] Hua Xu and Michael A Freitas. A mass accuracy sensitive probability based scoring algorithm for database searching of tandem mass spectrometry data. BMC bioinformatics, 8(1):133, 2007. 50

[46] Masamichi Yamashita and John B Fenn. Electrospray ion source. another variation on the free-jet theme. The Journal of Physical Chemistry, 88(20):4451–4459, 1984. 16

[47] Jing Zhang, Lei Xin, Baozhen Shan, Weiwu Chen, Mingjie Xie, Denis Yuen, Weiming Zhang, Zefeng Zhang, Gilles A Lajoie, and Bin Ma. Peaks db: de novo sequencing assisted database search for sensitive and accurate peptide identification. Molecular & Cellular Proteomics, 11(4):M111–010587, 2012. 27