# A TMT-labeled Spectral Library for Peptide Sequencing

by

Jianqiao Shen

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, 2017

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

Tandem mass spectrometry (MS/MS) is widely used nowadays for bioinformaticians to identify peptides and proteins. Three major peptide identification approaches and tools have been developed over the past two decades: database searching, *de novo* sequencing and spectral library searching. Recently, spectral library searching approach gains increasing popularity because of high-sensitivity and searching speed. Another popular technique is tandem mass tag (TMT). The advantage of this approache is able to simultaneously quantify multiple samples. However, TMT labeling also increases the complexity of spectrum compared to label free spectrum. Currently, there is no TMT-labeled library available and most of the software tools are also optimized for label free data.

In this thesis, we will study the differences between TMT-labeled spectra and label free spectra. We find the intensities of fragmentation ions are changed which proves that the algorithm only considers the mass shift of fragmentation ions and reporter ions is insufficient when searching TMT-labeled query spectra against a label free spectral library. It is necessary to build a spectral library from real TMT-labeled spectra.

We develop a TMT-labeled spectral library. The target library is constructed with JUMP database search tool and generate the decoy library by swapping precursor mass. We then evaluate the library with a small TMT-labeled dataset and compare the performance of spectral library searching with traditional database searching. The test result shows the identification rate is increased using spectral library searching.

# Acknowledgements

First of all, I would like to thank my supervisor Dr. Bin Ma for your guidance during my graduate program in University of Waterloo. Your patience and intelligence helped me work on the project and thesis writing. I also remember that you intrusted me how to accurately express my ideas and discuss with collaborators.

I would also like to thank my committee members, Dr. Lila Kari and Dr. Andrew C. Doxey. Your knowledge and experience in this area helps me improve my thesis.

Besides, I would like to thank my fellow labmates in Bioinformatics Group: Chenyu Yao, Lian Yang, Qi Tang, Rong Wang and Tiancong Wang, for the help and fun in the last two years.

Last but not least, I would like to thank my parents for their continuous support. Thank you for giving birth to me and raising me up to see the world. I know no matter what I do, you are always standing behind me.

# Dedication

This is dedicated to the one I love.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Motivation

Mass spectrometry based peptide sequencing for protein identification has been an important method in proteomics. In a typical bottom-up approach, protein samples will be digested into peptides by enzymes. This peptide mixture will go through liquid chromatography tandem-mass spectrometry LC-MS/MS to generate tandem mass spectra. The amino acid sequences of these peptides will be identified from tandem mass spectra. Finally, the sequences of original proteins can be inferred from the identified peptides.

Three computational approaches are commonly used for peptide sequencing. They are database searching, de *novo* searching, and spectral library searching. Among them, spectral library searching has the advantages of higher accuracy, identification rate, and faster speed compared with the other two approaches. Therefore, it has raised an increasing attention from researchers. Although spectral library searching still has the limitation that it can only identify the peptides present in its library, the growing size of different spectral libraries are improving the use of spectral library searching.

Another important technique is tandem mass tags (TMT) labeling, which helps us measure the quantitation of the proteins from different samples. Combined with LC-MS/MS, researchers can determine the abundance of peptides according to the intensities of reporter ions after fragmentation. Some of the fragmentation ions will be modified by tandem mass tags that results in the mass shift of the fragmentation ions in MS/MS. The appearance of reporter ions and the mass shift of fragmentation ions will make the tandem mass spectra of TMT-labeled peptides different from that of label free peptides. What is more, the intensities of these fragmentation ions can be changed as well. It is difficult for algorithms to convert a label-free spectrum to a TMT-labeled spectrum. Therefore, when identifying the sequences of TMT-labeled peptides, it is better to directly use TMT-labeled spectral library generated from real experiments.

## 1.2 Research Objectives and Contributions

The first objective of my thesis is to show how the intensities of fragmentation ions are changed with TMT-labeled peptides, and to prove the necessity in building a TMT-labeled spectral library from tandem mass spectra generated in real experiments. Another purpose of the thesis is to build a TMT-labeled spectral library and evaluate its equality.

Identified TMT-labeled peptides and label-free peptides are separated by the last amino acids Lysine(K) or Arginine(R). The fragmentation ions' intensities are compared if one spectra is from TMT-labeled peptides and another from label-free peptides but they both have the same amino acid sequence. We find that the intensities of fragmentation ions will be changed, $b_1$ ion of TMT-labeled peptides will appear. The fragment pattern of b-ions and y-ions differs between TMT-labeled peptides and label-free peptides. This proves the necessity in building a TMT-labeled spectral library. The existing conversion algorithms that convert TMT-labeled spectra to label free spectra only target the reporter ions and mass shift of fragmentation ions. Such conversion methods can not perfectly match the TMT-labeled query spectra with candidate label-free spectra in library.

Another contribution of our work is to build a TMT-labeled spectral library. We give a workflow to generate a TMT-labeled spectral library from 105 fractions of tandem mass spectra. The dataset is identified by the database searching software JUMP[25]. All the identified peptide spectrum matches (PSMs) are sorted according to the JUMP's score. All the spectra that filtered by 1% fdr will be selected into the library to ensure the quality. The spectral library is stored in the format of *.splib* which can be easily operated by SpectraST[15]. We design an experiment to prove the decoy algorithm of SpectraST[15] that works for our spectral library and test dataset. The identified results of JUMP database searching and spectral library searching SpectraST[15] are compared. From the comparison result, we can achieve better identification rates using the TMT-labeled spectral library.

## 1.3 Thesis Overview

The thesis is constructed in the following chapters.

In Chapter 2, some basic background knowledge will be introduced including the relationship between proteins and peptides as well as how they are synthesised, mass spectrometry, approaches for peptide sequencing and tandem mass tags. Chapter 3 will show the necessity in building a TMT-labeled spectral library. This will show the difference of fragmentation ions' intensities between TMT-labeled peptides and label free peptides. Chapter 4 will introduce the workflow to generate a TMT-labeled spectral library from real spectra. The evaluation of my TMT-labeled spectral library will be presented in Chapter 5. Finally, we will present a conclusion and future work in Chapter 6.

# Chapter 2

# Background

## 2.1 Proteins and Peptides

Proteins are large biomolecules and consist of one or more long chains of amino acid residues. Proteins can be found in most living organisms and perform an important role in biological activities including forming the structure of the body, catalysing metabolic reaction as different enzymes, and protecting against diseases in the immune system.

The different amino acid sequences of proteins result in different chemical properties that determine their function in biological activities. Identifying the amino acid sequence of a protein is significant in proteomics, as it helps us understand the chemical property of the protein and help us modify or reproduce it. An amino acid is composed of an amine (-$NH_2$), a carboxyl (-COOH) functional group, and a side chain (R group). Proteins consist of 20 standard amino acids, and they are usually represented by a 1-letter code. Two amino acids can react to be linked together. This reaction is called dehydration condensation. Figure 2.1 shows the chemical structure of an amino acid and the process of dehydration condensation. The carboxyl group of one amino acid will lose a hydrogen and oxygen and the amino group of another amino acid will lose a hydrogen to form a peptide bond (-CO-NH-). By this reaction, several amino acids can be linked to form a protein.

The amino acid sequence of a protein assembled by organisms is decided by the genetical information. This information is encoded in Deoxyribonucleic acid (DNA), which exists in the cell nucleus. DNA consists of four different nucleotides. A nucleotide is composed of a nitrogenous base, a five-carbon sugar, and at least one phosphate group. Four different nucleotides differ according to four different nitrogenous bases. They are cytosine (C),

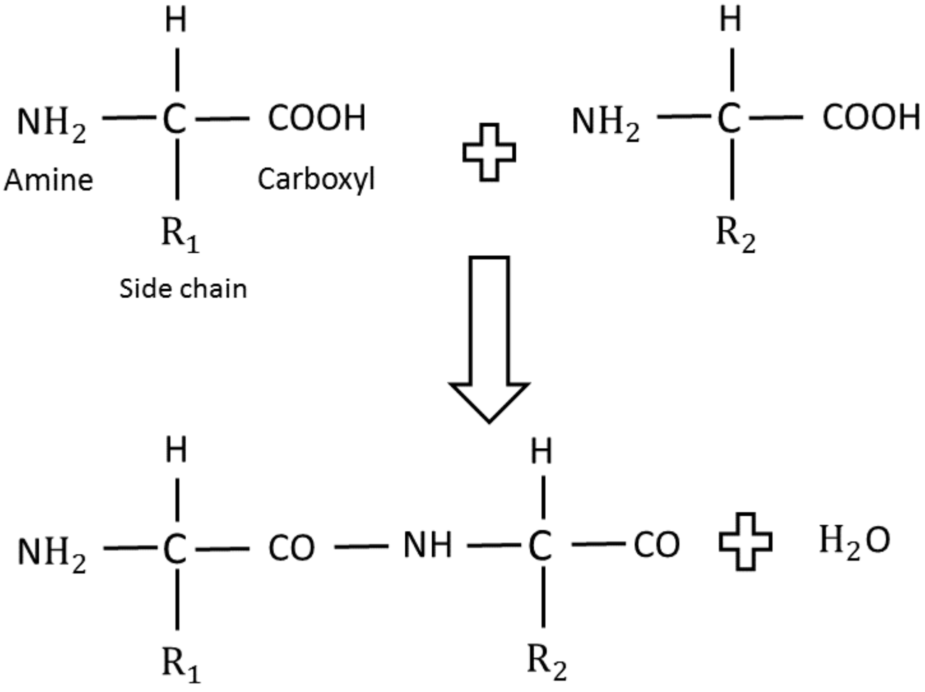

Figure 2.1: Chemical structure of amino acid and dehydration condensation

guanine (G), adenine (A), and thymine (T). The structure of DNA is made up of two helixes. The nitrogenous bases of two DNA strands are bound by pairing rules (A with T and C with G).

The protein synthesis process includes transcription and translation. In the transcription stage, DNA is copied into messenger Ribonucleic acid (mRNA), which is similar to DNA but the nitrogenous base urail (U) replaces thymine (T) in the original DNA and the structure of mRNA is single-stranded. In the copying process, the double-stranded DNA will be unwound and one strand of DNA is used as the template of the mRNA. The nitrogenous base of mRNA and template DNA also obeys the pairing rule in mRNA synthesis. In translation step, proteins are produced from mRNA. There are 20 proteinogenic amino acids encoded in a gene. Each amino acid is encoded by three mRNA nucleotides. This mRNA encoding scheme of amino acids is called the codon table. An amino acid will be transferred to mRNA by transfer RNA (tRNA), which plays the role of a link between mRNA and the amino acid. The nitrogenous bases of mRNA and tRNA are bound according to pairing rules as well. Therefore, the three-nucleotides codon of tRNA is complementary to the codon of mRNA by the pairing rules. The correct amino acids are then linked to form the protein. Figure 2.2 shows how proteins are synthesised.

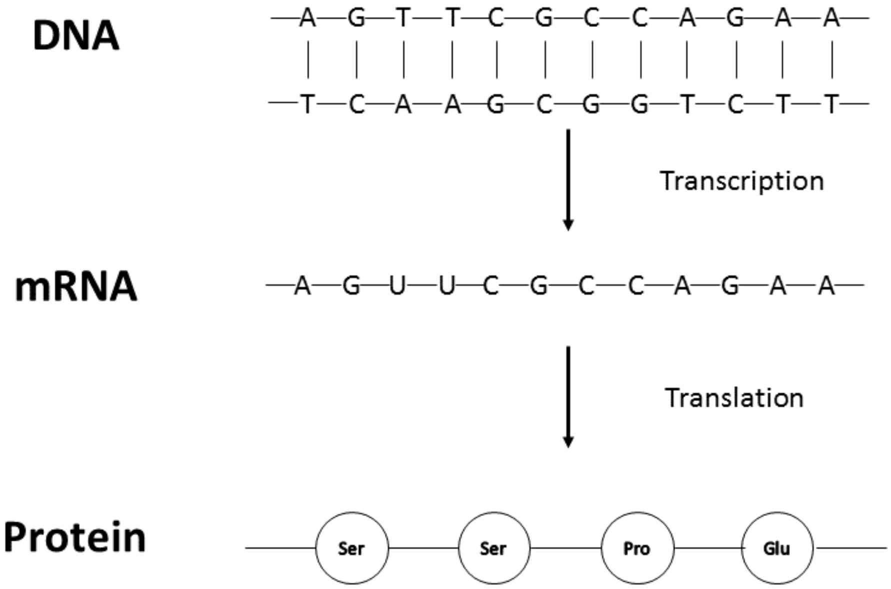
DNA
—A—G—T—T—C—G—C—C—A—G—A—A—
—T—C—A—A—G—C—G—G—T—C—T—T—
Transcription
mRNA
—A—G—U—U—C—G—C—C—A—G—A—A—
Translation
Protein
Ser
Ser
Pro
Glu

Figure 2.2: Process of protein synthesis

Peptides are short chain of amino acids. Their sequence lengths are usually less than

20-30 amino acids residues[16]. Unlike proteins, when focusing on one peptide we are only concerned with the linear amino acid sequence of it but not the three-dimensional structure. The word 'peptide' comes from Greek word '$\pi\varepsilon\pi\tau\acute{o}\varsigma$' which means 'digested' in English. Peptides can be digested from proteins by different enzymes. A peptide will have two termini on either side. One is called N-terminus, referring to the free amine group (-$NH_2$), and another is called C-terminus terminated by a free carboxyl group (-COOH).

Post-translational modification (PTM) can occur after protein synthesis. Most common post-translational modifications are phosphorylation, glycosylation,acetylation, truncation or formation of disulfide bridges[8]. The position can be on the amino acid side chains or at the protein's C- or N- termini. The process will modify an existing functional group or introducing a new one, which will change the mass of the corresponding amino acid and the chemical property of the modified protein. The mass change on amino acid will also have an effect on the mass spectrum of the protein, which we will introduce next.

## 2.2 Mass Spectrometry

Mass Spectrometry is an analytical technique that helps researchers to determine the masses of particles or molecules and to elucidate the chemical structures of molecules, such as peptides. It has been applied in many different fields of chemistry and biology. The basic idea is to ionize the unknown sample and sort the ions according to their mass-to-charge ratio (m/z). The result is a mass spectrum that is the plot of signal intensity of the ions and mass-to-charge ratio.

A mass spectrometer consists of three components: an ion source, a mass analyzer, and a detector. The ion source will ionize the sample and these ions will be accelerated and transported by magnetic or electric fields to the mass analyzer. In the mass analyzer, the motion of ions in space and time only depends on their $m/Q$ where $m$ is the mass and $Q$ is the charge of the ion. Therefore, different ions will be separated by their mass-to-charge ratio. The detector can record the value of an indicator quantity so that we can calculate the abundance or intensity of each ion.

Tandem mass spectrometry (MS/MS) refers to multiple steps of mass spectrometry. In the first stage of mass spectrometry (MS1), samples are ionized and separated according to the mass-to-charge ratio. Ions of a particular mass-to-charge ratio will be selected in the next step and these ions are called precursor ions. Different methods can be used to fragment the precursor ions and create product ions. The resulting product ions will be separated and detected in the second stage of mass spectrometry (MS2). Finally, we can get the MS2 spectra of product ions. Protein sequencing is an important application of tandem mass spectrometry.

In shotgun proteomics, a combination of liquid chromatography combined with tandem mass spectrometry will be used to identify the amino acid sequences of unknown protein samples. The workflow is called Liquid chromatography-tandem mass spectrometry (LC-MS/MS). Since the length and mass of a protein can be large, it is hard for equipment to ionize or to fragment[28]. An intuitive idea is that we can 'cut' the long protein into several short chains. If we are able to identify the amino acid sequences of these shorter chains we can splice the sequences to obtain the original sequence of the protein. This mechanism is called bottom-up proteomics. Figure 2.3 shows the process of LC-MS/MS.

First of all, the mixture of proteins will be digested by particular enzymes such as trypsin to produce peptides. All the peptides will go through liquid chromatography (LC). In this process, all peptides will be dissolved in a certain liquid that is called the mobile phase. Peptides are forced by the liquid through the equipment column. Due to the difference of mass, ph and hydrophobicity between peptides, the time different peptides flow out of the column differs and this time is called the retention time. At each time, the flowed out peptides will be ionized and get the first stage of mass spectra. Precursor ions of peptides will be selected to obtain the second stage of mass spectra. Peptides will be fragmented to create product ions. In fragmentation, energy was introduced to the peptide so that the backbone of the peptide will be cleaved and produce two fragments. Most widely-used fragmentation methods include collision-induced dissociation (CID), electron capture dissociation (ECD) and electron transfer dissociation (ETD). B-ions and y-ions are the most common fragment ions. B-ions are fragmentation ions extended from the amino terminus while y-ions are fragmentation ions extended from the carboxyl terminus. Table 2.1 and 2.2 give the monoisotopic masses of b-ions and y-ions for an example peptide 'LGSLVDEFK'. The mass-to-charge ratio and intensity of b-ions and y-ions are significant features of the tandem mass spectrum of a peptide. Most peptide sequencing algorithms are based on them.

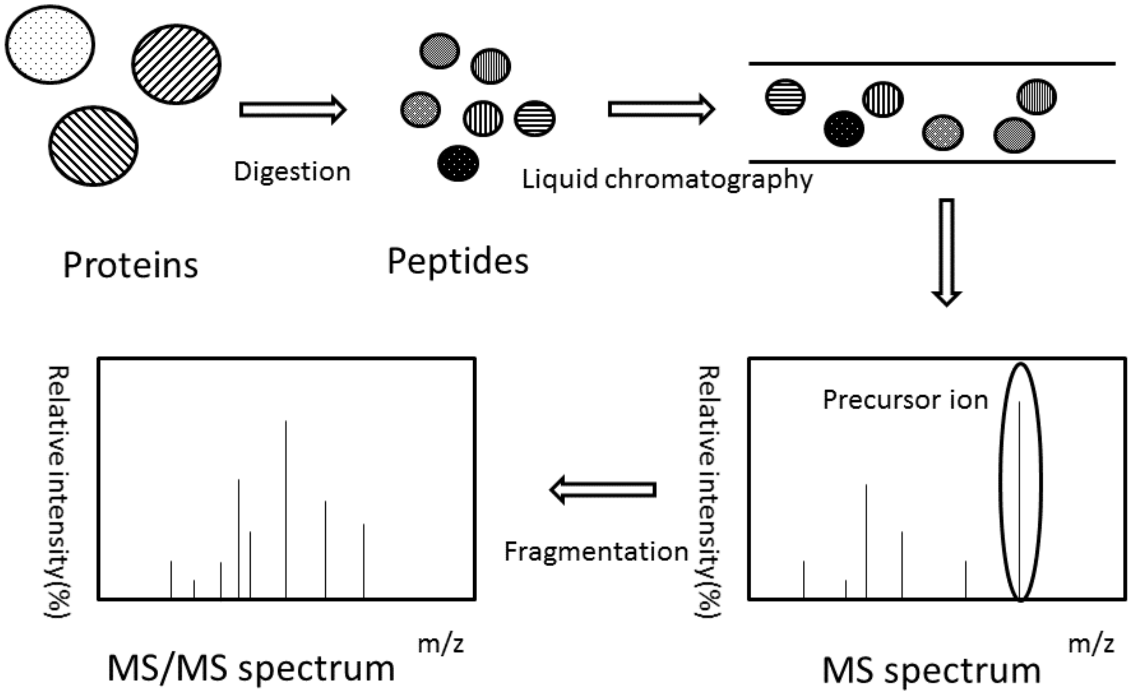

Figure 2.3: Workflow of LC-MS/MS

Table 2.2: Monoisotopic masses of y-ions for peptide LGSLVDEFK

| Sequence | y-ion | Mass |
|---|---|---|
| K | $y_1$ | 147.11285 |
| FK | $y_2$ | 294.18126 |
| EFK | $y_3$ | 423.22386 |
| DEFK | $y_4$ | 538.2508 |
| VDEFK | $y_5$ | 637.31921 |
| LVDEFK | $y_6$ | 750.40328 |
| SLVDEFK | $y_7$ | 837.4353 |
| GSLVDEFK | $y_8$ | 894.45677 |
| LGSLVDEFK | $y_9$ | 1007.54083 |

Table 2.1: Monoisotopic masses of b-ions for peptide LGSLVDEFK

| Sequence | b-ion | Mass |
|---|---|---|
| L | $b_1$ | 114.09139 |
| LG | $b_2$ | 171.11285 |
| LGS | $b_3$ | 258.14488 |
| LGSL | $b_4$ | 371.22894 |
| LGSLV | $b_5$ | 470.29735 |
| LGSLVD | $b_6$ | 585.3243 |
| LGSLVDE | $b_7$ | 714.36689 |
| LGSLVDEF | $b_8$ | 861.4353 |
| LGSLVDEFK | $b_9$ | 989.53027 |

## 2.3 Peptide Sequencing

With the tandem mass spectra of a peptide we want to infer its amino acid sequence. This step has been challenging in the development of proteomics over the past two decades. The difficulty is that the mechanism of fragmentation is not certain and there can be noise peaks in mass spectra, which make the information incomplete. Here are three main computational methods for peptide identification: (1) database searching, (2) de *novo* sequencing, (3) spectral library searching. These three approaches have their own advantages and limitations.

The database searching approach typically identifies peptides through two steps:

- Generating theoretical fragments from peptide sequences in the database according to the peptide fragmentation rules.
- Comparing spectra acquired by LC-MS/MS to theoretical fragments.

In the first step, the database usually stores the amino acid sequences of proteins instead of peptides. Therefore, the type of enzyme used in the experiment is a parameter for the searching tool to digest the database proteins in silico. If the masses of the candidate peptides are in the set mass tolerance of the query peptides, the mass-to-charge ratio of the theoretical fragment ions are calculated. Some algorithms will also predict the relative intensities of the fragment ions[4]. According to the fragmentation used in the experiment, the types of theoretical fragment ions are parameters for the searching tools as well. Moreover, if the sample protein has PTMs, the type of PTMs should be passed to the searching tool since the mass of fragmentation ions will be changed. In the second step, different searching tools will use different score schemes to measure the similarity between the query spectrum and the theoretical fragmentation patterns of the candidate peptides[20]. The candidate peptides will be sorted by the score and the peptide digested from the database with the highest score has the higher possibility to be the query peptide. Some widely used software are SEQUEST[9], Mascot[5], X!Tandem[6], and OMSSA[11].

Although database searching is a common method for peptide identification, it has several limitations. The searching space is limited by the protein database, which means if we have an example and its sequence is not in our database, it is impossible for us to identify it. Another problem is that since we do not fully understand the factors that determine peptide fragmentation, the database searching tool can only predict theoretical fragment ions and their corresponding rudimentary intensity values[12]. These predicted fragments can be quite different from the 'true' spectrum. When scoring the query spectrum with the theoretical fragments, the information in the query spectrum is not completely taken into account. Finally, it is difficult to judge whether the relative intensities of the fragment ions predicted by the database searching tool are reasonable and accords with that in the real spectrum.

Another approach is de *novo* sequencing. This approach tries to identify the amino acid sequence of the peptide directly from its tandem mass spectra without a protein database. The basic idea is that we have fragmentation ions in a tandem mass spectrum of a peptide, and from the mass difference between neighbouring b-ions or y-ions we can calculate the mass of the amino acid on that position. The relative intensity of these ions are also considered in the score scheme of de *novo* sequencing. With that mass, we can determine what amino acid it is. The de *novo* sequencing can take all possible amino acid combinations into consideration and return the most probable one. Two widely-used software are PEAKS[18] and Novor[17].

The advantage of de *novo* sequencing is that one can identify a peptide unknown to the protein database. However, the accuracy and correctness of de *novo* sequencing mainly depends on the quality of the tandem mass spectra. Because the mass of one amino acid in the peptide is calculated by the mass difference of two neighbouring fragmentation ions, we require a high-resolution of these ions including their mass-to-charge ratio and relative intensity. What is more, there is no annotation indicating which peaks in the tandem mass spectrum are fragmentation ions. Different algorithms have been designed to assign the peaks to possible fragmentation ions. However, internal fragmentation ions may appear in tandem mass spectra that have neither N-terminus nor C-terminus. This is caused by double backbone cleavage of the peptide. The mass of internal fragmentation ions will confuse the de *novo* sequencing software. Therefore, the appearance of internal fragmentation ions should be considered by the algorithm making the de *novo* sequencing more challenging. Compared to the results of database searching with an accurate and complete database, the results of de *novo* sequencing are usually less satisfactory. The combination of these two peptide sequencing methods have received great attention and achieve better results. Peaks DB[27] is such an example software.

The third approach is spectral library searching, which will be focused on in this thesis. Similar to database searching introduced above, spectral library searching use a database to identify the peptides. However, the entry in the database is identified by the tandem mass spectra of peptides instead of protein sequences. Therefore, the spectral library will have the information of peptide sequences and all peaks' mass-to-charge ratios and actual intensity in tandem mass spectra. Several examples of spectral library searching software are SpectraST[15], X!Hunter[7], and Bibliospec[10].

Compared to the theoretical fragmentation ions generated by traditional database searching, spectral library searching can make the best use of real information in tandem mass spectra in the library and achieve a better measurement of similarity between query spectrum and candidate spectra. Another advantage of spectral library searching is its faster speed. The search space of spectral library searching is smaller than database searching, since many peptides in the sequence database are supposedly not detectable in experiments[1], which means these peptides will not appear in the spectral library. What is more, it is more convenient to identify peptides with unusual post-translational modifications (PTMs), which can be challenges to sequence database searching[26]. However, spectral library searching still has its own limitations. Firstly, like traditional database searching it is impossible for spectral library searching to identify an unknown peptide if it is not in the library. Secondly, the fragmentation pattern of a peptide depends on the fragmentation method used in the experiment. For example, if we have a spectral library and its tandem mass spectra are generated by collision-induced dissociation (CID) while the query peptides are fragmented by electron capture dissociation (ECD), it is not suitable to use spectral library searching. Last but no least, in order to store the information of tandem mass spectra, the database of the spectral library is much larger than the traditional sequence database.

An important and powerful method to evaluate the results of peptide sequencing is "target-decoy" searching strategy. We can generate a decoy peptide sequences library and attach it to our target library. Each match of query spectrum and the sequence in decoy library is considered as an incorrect hit. The decoy library is generated to ensure that the probability of an incorrect hit from decoy library is the same as an incorrect hit from target library. We can not know a match from target library is correct or not, but with the help of decoy library, we can estimate the false discovery rate (fdr). The number of identified spectra under certain fdr is a evaluation of different searching results.

## 2.4 Tandem Mass Tags Labeling

The shotgun proteomics has a main disadvantage: the quantitation of the proteins is less straightforward[23], since all the peptides in a sample are digested into peptides, we can not directly deduce the the quantitation of the proteins from the digested peptides. Different methods have been developed to solve this problem, including stable isobaric isotope labeling. Isobaric tags have the same mass but differ in the number and combination of heavy isotopes in their structure[22]. Tandem mass tags (TMT) and isobaric tags for relative and absolute quantitation (iTRAQ) are two fundamental isobaric tags. In this thesis, we will focus on tandem mass tags (TMT) labeling.

TMT products inculde TMT-duplex, TMT-sixplex, and TMT-10plex. We will use TMT-sixplex as an example to introduce tandem mass tags (TMT). A TMT reagent has three functional groups[22]: (1) An amine-reactive group, which will label the N-terminus and lysine(K) of peptides. (2) A reporter group, which will produce the reporter ion after fragmentation provides the abundance of the peptide. The reporter ion will appear at m/z 126, 127, 128, 129, 130, and 131. (3) A mass normalization (balance) group, which balances the mass differences of the reporter group to ensure the same peptides labeled by different tags have the same mass. TMT-10plex has similar structure with TMT-sixplex and it can produce 10 reporter ions by creating a mass difference of 6.3 mDa.

A general workflow of TMT labeling to measure the quantitation of the proteins includes following steps:

- Protein samples are digested into peptides and other sample preparations are operated.
- Label the samples with tandem mass tags. For TMT-sixplex, it can label up to six different samples.
- Analyze the labeled peptides by LC-MS/MS. Since the overall masses of tags are the same, the precursor ion of peptides with the same sequences will appear at the same mass-to-ratio. However, in MS2, the bond between the reporter group and the normalization group will be broken producing reporter ions at different mass-to-ratio.
- The quantitation of the peptides is indicated by the intensity of reporter ions. Knowing the quantitation of the peptides can increase the confidence in the quantification of the protein.

It should be emphasised that with N-terminus and lysine(K) labeled by the tandem mass tag, the mass of the precursor ion will be increased in the first stage of tandem mass spectra. Moreover, the mass of b-ions from all peptides and y-ions from peptides ending with lysine(K) will be increased in the second stage of tandem mass spectra as well. These mass changes will lead to the mass shift of fragmentation ions in MS/MS. The change of fragment pattern and extra reporter ions will make it impossible for a spectral library searching algorithm to match the query TMT-lableled spectrum with the correct label-free spectrum in the library. Necessary conversion should be done to ensure the consistency with query spectra and candidate spectra in the library. A more intuitive idea is to build a spectral library with tandem mass spectra from TMT-labeled peptides in real experiments.

# Chapter 3

# Tandem Mass Spectra of TMT-labeled and Label-Free Peptides

We have discussed the differences between tandem mass spectra of TMT-labeled peptides and that of label-free peptides. We know that a label-free spectral library can not directly be used to identify a TMT-labeled peptide. Although some algorithms can be applied to generate an artificial label free spectrum from a TMT-labeled spectrum, which means we can convert TMT-labeled spectra in library into label-free ones[29]. The conversion algorithm simply shifts the mass-to-charge ration of the fragmentation ion peaks according to the mass of TMT tag and discards three ions:(1) TMT report ions, (2) TMT fragments, (3) TMT ion losses from the precursor. However, the problem still exists of whether the relative intensities of these fragmentation ions will be changed. If the relative intensities are changed it is hard for algorithms to decide how to adjust these intensities.

Take a peptide 'QGAIVAVTGDGVNDSPALK' as an example. The top half of Figure 3.1 shows the MS/MS of a label-free example peptide while the bottom half shows the MS/MS of an TMT-labeled example peptide. The b-ions and y-ions are marked to be compared. From the fragment pattern, we can see despite the mass shift of fragmentation ions caused by tandem mass tags, $b_1$ ion appears in the MS/MS of the TMT-labeled example peptide while they do nor appear in the label-free example peptide. Moreover, the relative intensities of b-ions from the TMT-labeled example peptide are higher than that of the label-free example peptide. Oppositely, the relative intensities of y-ions from TMT-labeled example peptide are lower except $y_1$ and $y_2$. We can see some differences of fragmentation ions' relative intensities from the label-free and TMT-labeled peptides.

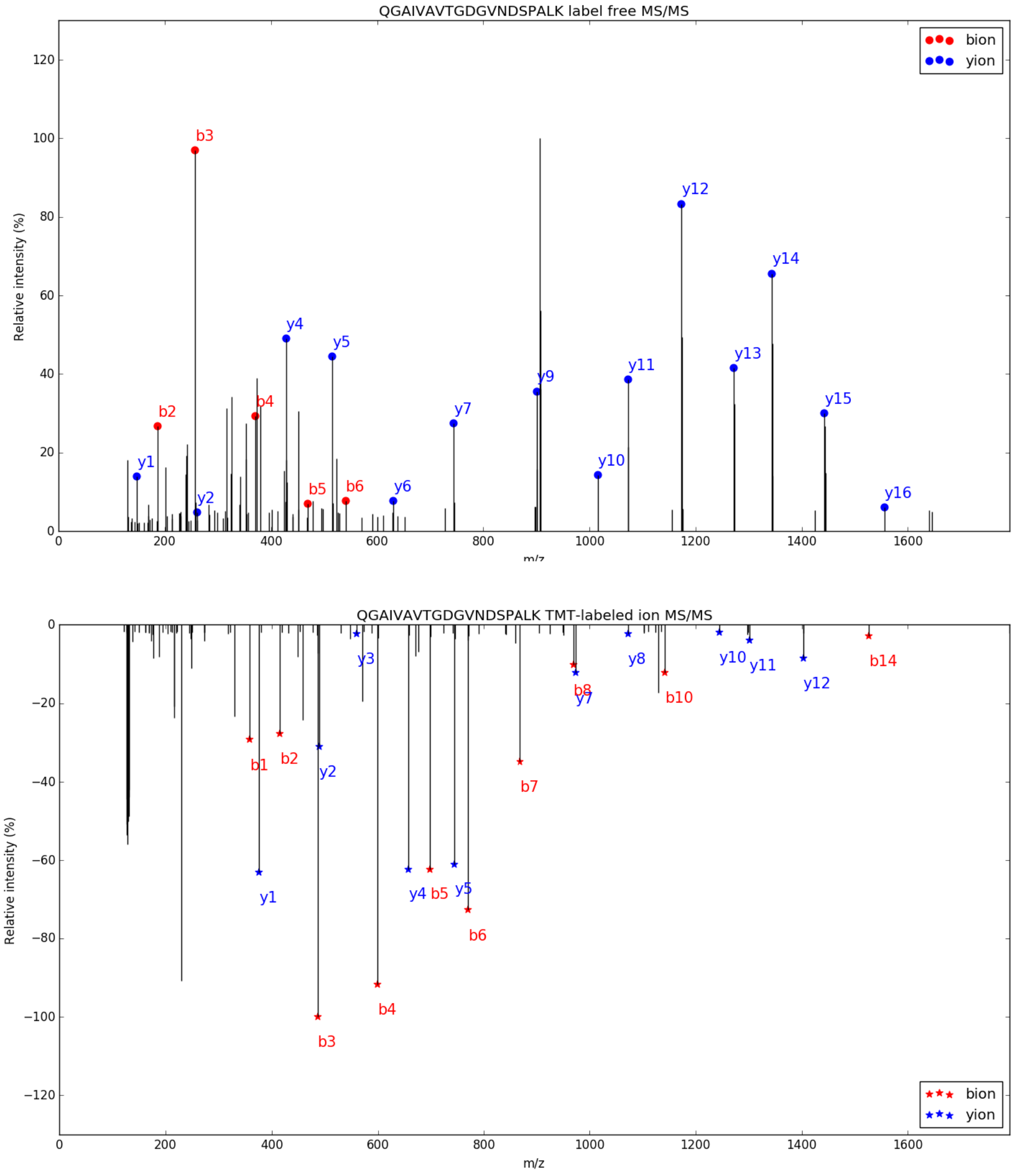

Figure 3.1: Label free and TMT-labeled MS/MS of the example peptide QGAIVAVTGDG-VNDSPALK

We designed an experiment to find the statistical differences. Two datasets are searched by the same database searching tool, JUMP[25] with the same parameters except the existence of TMT modifications. We focus on two spectra if one is from a TMT-labeled peptide, another from a label-free peptide, and their identified peptide sequence is the same. We also separate the spectra according to the last amino acid of their peptide sequences. Since they are digested by trypsin, the peptides can end with Lysine(K) or Arginine(R). Therefore, we can get four groups of peptides:

- Label free peptides ending with Lysine(K).
- TMT-labeled peptides ending with Lysine(K).
- Label free peptides ending with Arginine(R).
- TMT-labeled peptides ending with Arginine(R).

We only focus on the relative intensity of matched $b_1$-$b_{10}$ ion peaks and $y_1$-$y_{10}$ ion peaks. The reference peak is the peak among $b_1$-$b_{10}$ and $y_1$-$y_{10}$ with the highest intensity. The intensities of other matched b-ions and y-ions are divided by the intensity of the reference peak and normalized to get the relative intensities. All the relative intensities of $b_x$ and $y_x$ ion peaks are summed to calculate the corresponding average intensities of the four groups. This is shown in Figures 3.2 to 3.5, below.

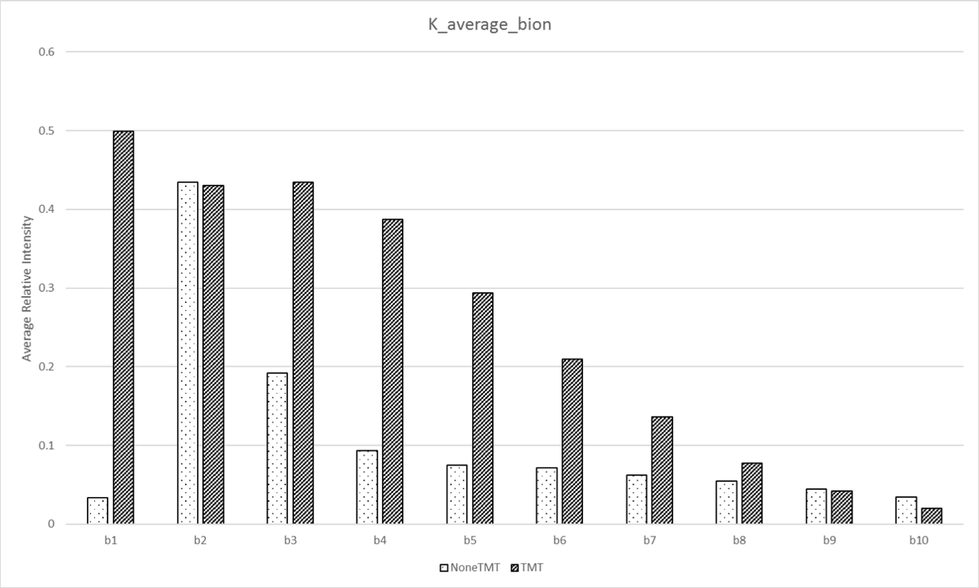

Figure 3.2: Relative intensity of b-ions from peptides ended with Lysine(K)

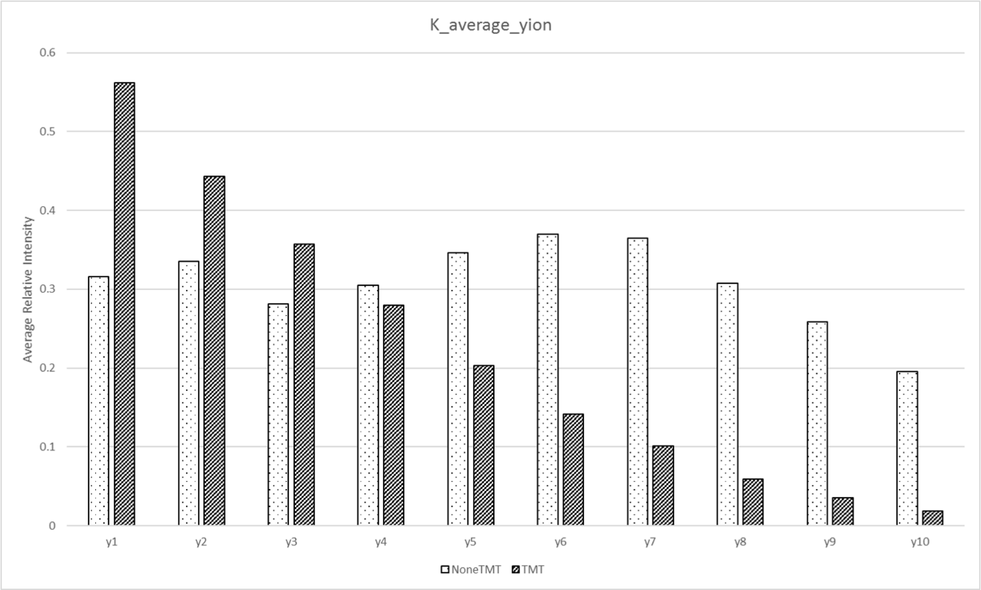

Figure 3.3: Relative intensity of y-ions from peptides ended with Lysine(K)

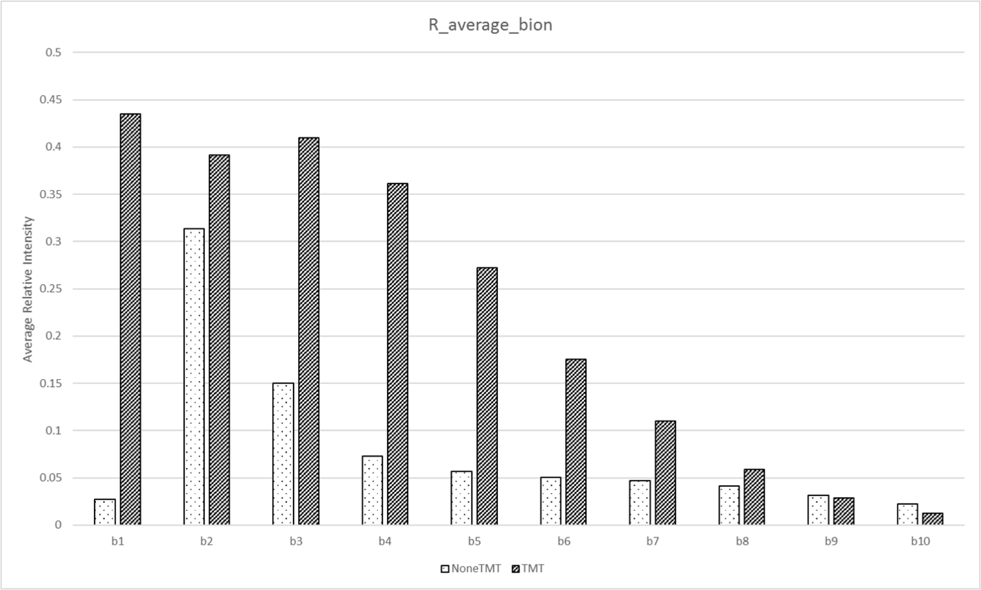

Figure 3.4: Relative intensity of b-ions from peptides ended with Arginine(R)

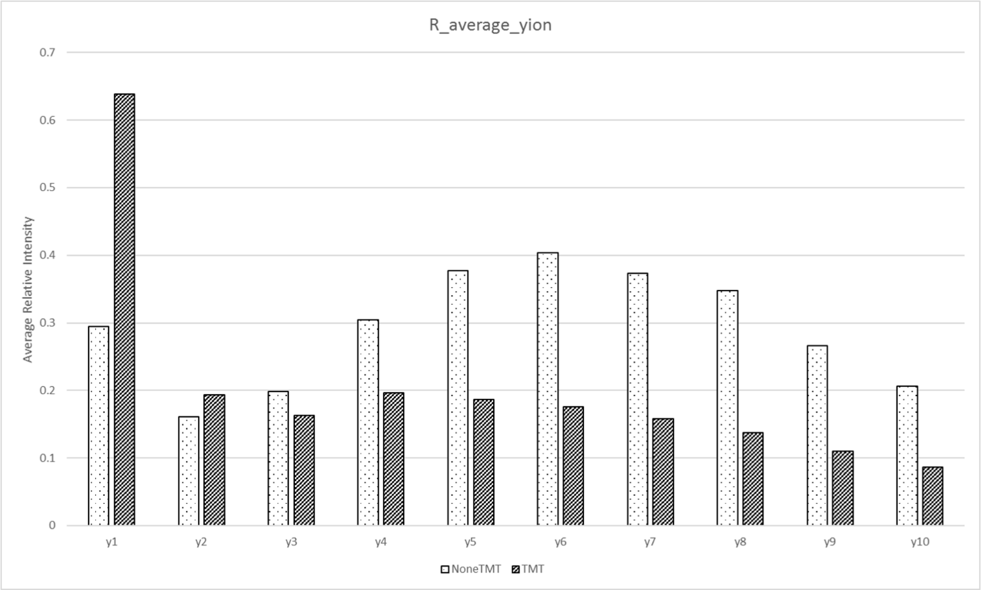

Figure 3.5: Relative intensity of y-ions from peptides ended with Arginine(R)

From the figures shown above, we can conclude that for label-free and TMT-labeled peptides digested by trypsin, the most four important differences between their MS/MS are: (1) The $b_1$ ion will appear in the tandem mass spectra of TMT-labeled peptides while not in label-free peptides. (2) The relative b-ion intensities of TMT-labelled peptides are higher than those of label-free peptides. (3) For peptides ending with Arginine(R), the relative intensities of the $y_1$ ion of TMT-labeled peptides are higher than that of label-free peptides while the relative intensities of other y-ions are lower. (4) For peptides ending with Lysine(K), the relative intensities of front y-ions of TMT-labelled peptides are higher than those of label-free peptides while the relative intensities of rear y-ions are lower oppositely. These differences will make it harder for a spectral library searching tool to search a TMT-labeled query spectrum against a label-free spectral library. Therefore, it is necessary for us to build a specific TMT-labeled spectral library.

## 3.1 $b_1$ ion

When using Higher-energy collisional dissociation (HCD) to generate the tandem mass spectra, the $b_1$ ion is usually not detected. One of the possible reasons for this is that the mass of the $b_1$ ion is too small and it is out of the detectable mass range of the equipment.

Another important reason is that the carbonyl oxygen of the residue N-terminal to the cleavage site is involved in the formation of b-type ions[19]. This explains why there is no $b_1$ ion in label-free spectra. However, with N-terminal of the original peptide modified by the TMT tag, the $b_1$ ion can be produced by the cleavage reaction and the mass of the $b_1$ ion increases by the mass of TMT tag so that it can fall within the detectable mass range.

## 3.2 Other b-ions

The relative intensities of b-ions of TMT-labeled peptides will rise. One of the important factors is that b-ions are less stable due to higher collision energy[24] when using higher-energy collisional dissociation (HCD), but with the N-terminus modified by TMT tags b-ions will be more stable to be detected by the equipment. Another important factor is that, the mass of b-ions increases and the front b-ions can be located in the middle detectable mass range while the rear ones may be out of the detectable mass range. Combined with the first factor, the front b-ions will have higher relative intensities while the relative intensities of the rear b-ions will not change significantly.

## 3.3 y-ions

For TMT-labeled peptides ending with Arginine(R), the relative intensities of y-ions are lower than label-free ones. Generally, within the fragmentation ions generated by higher-energy collisional dissociation (HCD), the reference peak is usually a certain y-ion. However, with the increase of the relative intensities of b-ions from TMT-labeled peptides, the reference peak can be a b-ion and its corresponding intensity is higher than the intensity of the reference peak from label-free peptides. This will make the relative intensities of y-ions lower.

For TMT-labeled peptides ending with Lysine(K), two factors affect the relative intensities of y-ions from TMT-labeled peptides. The first one is the same as the factor mentioned above, which make the relative intensities of y-ions lower. Another factor is that with the TMT modification at the last amino acid Lysine(K), y-ions of TMT-labeled peptides will have the same mass shift as b-ions. The mass change of y-ions locates the front y-ions in middle range of the detectable mass range while the rear y-ions are located far from the middle of the detectable mass range. These two factors explains the fragment patterns of TMT-labeled peptides ending with Lysine(K) together.

# Chapter 4

# Methodology

## 4.1 Method Overview

The overall procedure to generate the spectral library from TMT-labeled tandem mass spectra includes the following steps: peptide identification with JUMPg[25], generating the initial spectra library in .msp format, convert the .msp spectra library into .splib format with SpectraST[15], generate the decoy library. Figure 4.1 shows the flow chart of the overall procedure. The detailed information will be provided in the next several sections.

## 4.2 Materials and Experiment

The analysis was performed with a previously optimized protocol[2][21]. Quantified protein (∼1 mg in the lysis buffer with 8 M urea) for each TMT channel were proteolyzed with Lys-C (Wako, 1:100 w/w) at 21°C for 2 h, diluted 4-fold to reduce urea to 2 M, and further trypsinized at 21°C overnight (Promega, 1:50 w/w). The insoluble debris was kept in the lysates, which allowed the recovery of insoluble proteins during digestion. The digestion was stopped with the addition of 1% trifluoroacetic acid followed by centrifugation. The supernatant was desalted with Sep-Pak C18 cartridge (Waters), and dried by speedvac. Each sample was resuspended in 50 mM HEPES, pH 8.5, labeled with TMT reagents, equally mixed, and desalted again for subsequent fractionation.

The TMT labeled samples were fractionated by offline basic pH reverse phase LC followed by acidic pH reverse phase LC-MS/MS analysis. Considering that unmodified

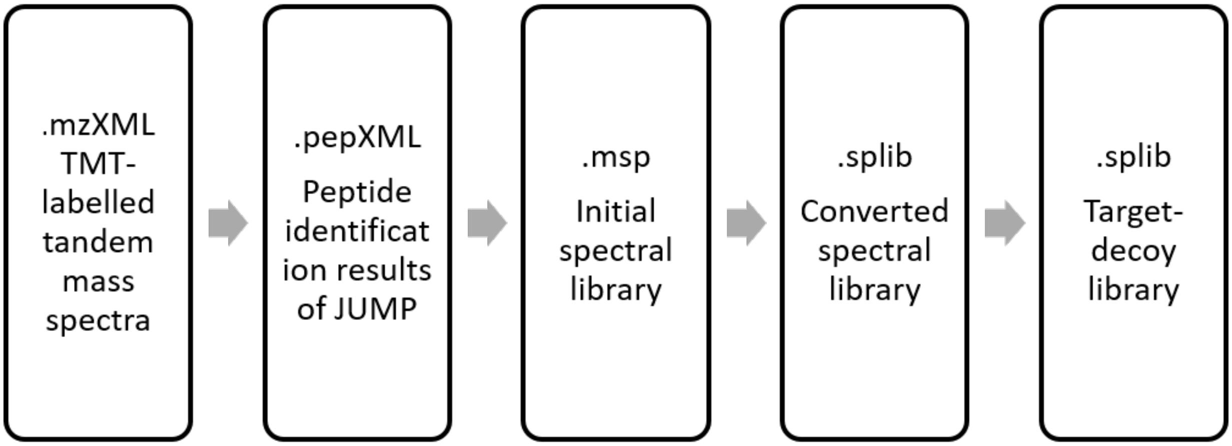
.mzXML
TMT-labelled tandem mass spectra
.pepXML
Peptide identification results of JUMP
.msp
Initial spectral library
.splib
Converted spectral library
.splib
Target-decoy library

Figure 4.1: Flow chart of the overall procedure to generate the spectral library

peptides have different elution profiles from phosphopeptides, we performed two offline LC runs: 5% of the sample for whole proteome (∼2 h gradient, ∼120 fractions) and 95% for phosphoproteome (∼1 h gradient, ∼60 fractions) on a XBridge C18 column (3.5 m particle size, 4.6 mm x 25 cm, Waters; buffer A: 10 mM ammonium formate, pH 8.0; buffer B: 95% acetonitrile, 10 mM ammonium formate, pH 8.0).

In the LC-MS/MS analysis, each fraction was run sequentially on a LC column (75 m x ∼40 cm for whole proteome, 50 m x ∼30 cm for phosphoproteome, 1.9 m C18 resin from Dr. Maisch GmbH, at 65°C to reduce backpressure) interfaced with an Q Exactive HF Orbitrap MS (Thermo Fisher). Peptides were eluted by in ∼3 h gradient (buffer A: 0.2% formic acid, 5% DMSO; buffer B: buffer A plus 65% acetonitrile). MS settings included MS1 scans (60,000 resolution, 1 x 106 AGC and 100 ms maximal ion time) and 20 data-dependent MS2 scans (410-1600 m/z, 60,000 resolution, 1 x 105 AGC, ∼125 ms maximal ion time, HCD, 38% normalized collision energy, 1.0 m/z isolation window with 0.3 m/z offset, and ∼15 s dynamic exclusion).

## 4.3 Peptide Identification

A total of 105 fraction of TMT-labeled tandem mass spectra are searched against the human protein reference database using JUMP. JUMP is a tag-based database search tool for peptide identification. Compared with other database search tools such as SEQUEST, Mascot, InsPecT, and PEAKS DB, it can achieve better sensitivity and specificity[25]. Besides the TMT modification, Cysteine Alkylation is set as a static modification and Oxidation at Methionine is set as a dynamic modification. The identification results are stored in .pepXML format, which can be easily viewed or operated by other software.

## 4.4 Spectra Library in .msp

Although SpectraST can create a spectral library from a .pepXML file, which contains peptide identifications from a previous shotgun proteomics experiment, it is preferable that the .pepXML has been processed with PeptideProphet and/or iProphet. Since we are using JUMP to do the database search and for the purpose of operating the spectra directly, we wrote codes to scan through the .pepXML file for identifications and extract corresponding experimental spectra from the .mzXML files.

From the peptide identification results of JUMP, we generate the initial spectra library in .msp format, which records the information of the identified peptide sequence and its corresponding original spectrum as well as annotations. To be specific, we record the precursor m/z of the identified spectrum, the corresponding peptide sequence and post-translational modification, the relative intensity and m/z of each peak of the spectrum. TMT modification: +229.162932 at N-term and Lysine and Alkylation: +57.02146 Cysteine are included. The dynamic modification: +15.99492 will also be added if it exists. The positions of modifications appearing in the peptide sequences will be labelled. All the intensity of peaks will be normalized to a maximum 10000 to follow the .msp format. More detailed information can also be annotated, such as the type of enzyme, and number of amino acids in the peptide and the original organism. Since this information will not influence the results of the searching algorithm of SpectraST, for simplicity we will not include them in our test.

To ensure the accuracy of the library, only identified spectra within 1% fdr will be selected. There are mixed spectra, which means the original spectrum can contain the fragmentation ions from two or more different peptides if these peptides have similar precursor m/z. In order to ensure the unique explanation of each spectrum in our library, we will discard these mixed spectra.

## 4.5 Convert Spectra Library into .splib

Now we have the initial spectra library in .msp format and it contains the necessary information for further processing. In this step, we convert our .msp spectra library into .splib format, which is a binary format for SpectraST and which is convenient for SpectraST when performing some other operations. The ion peaks of the spectrum in the .splib library will be annotated if they are identified as fragmentation ions. The annotations will indicate the ion type of the peak and cleaved positions. Besides post-translational modifications, neutral molecule losses are also taken into account.

Each peptide sequence in the library may have several corresponding spectra. In other words, different experimental spectra can be identified as one peptide sequence. These spectra can be similar but have slight differences. It is difficult to judge which spectrum is the best match with the query spectrum. To solve this problem, two methods can be applied:

- Only keep the spectrum which gets the highest score from JUMP and discard other spectra. We note this library as the highest-score library. It has 244,625 unique peptides in total.
- Consensus those spectra into one spectrum. This operation is implemented by SpectraST. By this process, we want to generate a consensus spectrum which can represent the replicate and keep the important features that these replicate spectra share. We note this library as the consensus library. It has 1,118,494 spectra before consensus. The detailed algorithm is described below.

(1) For all the replicate spectra of one peptide sequence, they will be ranked by their signal-to-noise ratio, which is defined as the average intensity of the 2nd to 6th highest peaks divided by the median intensity[13].

(2) All peak intensities are placed into the same wide bins and these bins will be normalized. The dot products between the top-ranked replicate spectrum and all other

replicates are calculated[13]. For one replicate spectrum, if its dot product is greater than 0.6, then it will be clustered with the top-ranked replicate. This operation is maint ained until no more replicates can be clustered and it remains the largest cluster.

(3) Align peaks for all the remaining replicate. Starting from the base peak, for each peak if there is another peak in other replicates and within an adaptive m/z tolerance, they are matched and aligned.

(4) Remove noise peaks. Remove the aligned peaks if they are not present in 60% of the replicate spectra.

(5) Calculate the weighted average m/z and intensity of the consensus spectrum. The weight is the signal-to-noise ratio.

## 4.6 Generating Decoy Library

One of the effective methods to evaluate the peptide identification quality is performing a decoy search. Decoy searching is widely used in database searching and can achieve good results. However, it is not in spectral library searching. The difficulty is how to generate a good decoy library. There are four main problems[14]. First, the decoy spectrum should not be too similar to a real spectrum from a peptide that the user is interested in, otherwise, this correct peptide will be difficult to identify and results in false negatives. Secondly, the decoy library should have realistic features of the target library so that it can have the same chance to be matched for an incorrect identification. Third, the distribution of precursor m/z, enzymatic termini and post-translational modifications should remain the same as the target library. Lastly, we should have the same number of target and decoy spectra to be search candidates for any subdivision of the search space. We first try to use the spectral library of yeast as a decoy library. The procedure used to generate the yeast spectral is the same with the procedure to generate the target TMT-labelled spectral library except we discard those spectra if their corresponding sequences are in our target library so that the yeast spectral can be considered as the decoy library. Since the size of these two libraries are different the rate will be multiplied when calculating the false discovery rate. However, we do not achieve a satisfying result using this method. Another method we used to generate the decoy library is to randomly shuffle the peptide sequences of the target library and then shift the corresponding bion and yion peaks of the fragments to generate a decoy spectrum[3]. Other unassigned peaks remain on the original position. This process is included in the software SpectraST. However, this method still does not give us a satisfactory result. The method used in our paper is precursor swap. The basic idea is to swap the precursor m/z of two target spectra as their decoy spectra respectively. This method gives the most satisfying results compared to the other methods.

To be more detailed, for one spectra with precursor $M$ m/z in the library, the algorithm will find candidate spectra with the same charge in the library whose precursors are in the range between $M + d$ m/z and $M + D$ m/z where $d < D$ and $d > 2\Delta$, $\Delta$ is the precursor m/z tolerance of the SpectraST search mode. We restrict $d$ to ensure that the swapped decoy spectra will not be matched to query spectrum with precursor $M$. Then we pick the spectrum with the highest m/z and swap this spectrum with the target spectrum as their corresponding decoy spectra. Since we only swap the precursor of the target library, this ensures the decoy spectrum retains the realistic features. Moreover, this method preserves the distribution of precursor m/z of the original spectra data set.

## 4.7 Searching

We use SpectraST to conduct the searching. The basic idea of the searching algorithm of SpectraST is that: given a query spectrum with precursor $M'$ m/z, for all the peaks of the candidate target spectra in the library within the mass tolerance will be rescaled by taking the square root of raw intensities[15] to deemphasize the dominant peaks. Those peaks not matched with the fragment ions will multiply their intensities by a certain factor as a punishment. Then all the peaks will be placed into the same wide bins, the default width is 1 Th. The intensity of each bin is simply the sum of the intensity of the peaks in that bin. We can consider the spectrum as a vector and the element is the intensity of each bins. The bins will be normalized so that the vector norm equals 1. The query spectrum will be processed by such operations as well. SpectraST can return three different scores to measure the similarity between the query spectrum and the candidate spectrum in the library. First one is the dot score, it is the dot product of two spectral vectors. The second is the delta score, which is the normalized difference between the dot score of the top hit and the second hit of the query spectrum. The last is called *fval*, it is the liner combinations of the dot score and the delta score. The use of *fval* is recommended by the developer of SpectraST, it may not be optimal for all possible situations, but it should be generally adequate for a wide variety of applications[15]. In my thesis and experiments introduced below, *fval* will be used to sort the results of the identified peptides searched by SpectraST.

# Chapter 5

# Results and Discussion

To evaluate the equality of spectral libraries, we design several experiments to compare the peptide identification results of our libraries with that of JUMP. The spectral library search is operated by SpectraST.

For the consensus spectral library, we use another 56 shallow spectra fraction samples for testing to ensure that no testing spectrum is included in our library. We pick 5 fractions as test data and it contains 249,725 MS/MS. They are w005.mzXML, w015.mzXML, w025.mzXML, w035.mzXML, and w045.mzXML.

## 5.1 Decoy Library

In order to prove the precursor swap method is also useful for our data, we design an experiment to ensure that our target library does not have bias against the artificial decoy library for a query spectrum.

For each spectrum in test data we add $d'$ m/z to its precursor where $\Delta < d' < d - \Delta$ and generate a new precursor shifted dataset. Assuming that in the SpectraST parameter setting, we set the precursor m/z tolerance to $\Delta$ m/z. For a query spectrum with the precursor $M'$ m/z in the original test dataset, if there is a matched spectrum in the target library, its precursor should be in the range of $M' - \Delta$ and $M' + \Delta$ and the precursor of the corresponding decoy spectrum should be in the range of $M' + d - \Delta$ and $M' + D + \Delta$. The precursor of the query spectrum will be $M' + d'$ m/z in the precursor shifted dataset. The shifted query spectrum will not be matched to either the target library spectrum or the decoy library spectrum because $\Delta < d' < d - \Delta$ and we can deduce that $M' + d' > M' + \Delta$

and $M'+d' < M'+d-\Delta$ which means both are not in the precursor tolerance of the shifted query spectrum when searching with SpectraST. To be concluded, we want to show that the number of false positive with the target library can be estimated by the decoy library generated by the precursor swap method and that our target library does not have any bias against the test data set. Figure 5.1 shows that the score distributions of the target and decoy libraries are highly consistent, which means for an incorrect identification, the decoy spectrum has the same chance to be matched with the query spectrum.

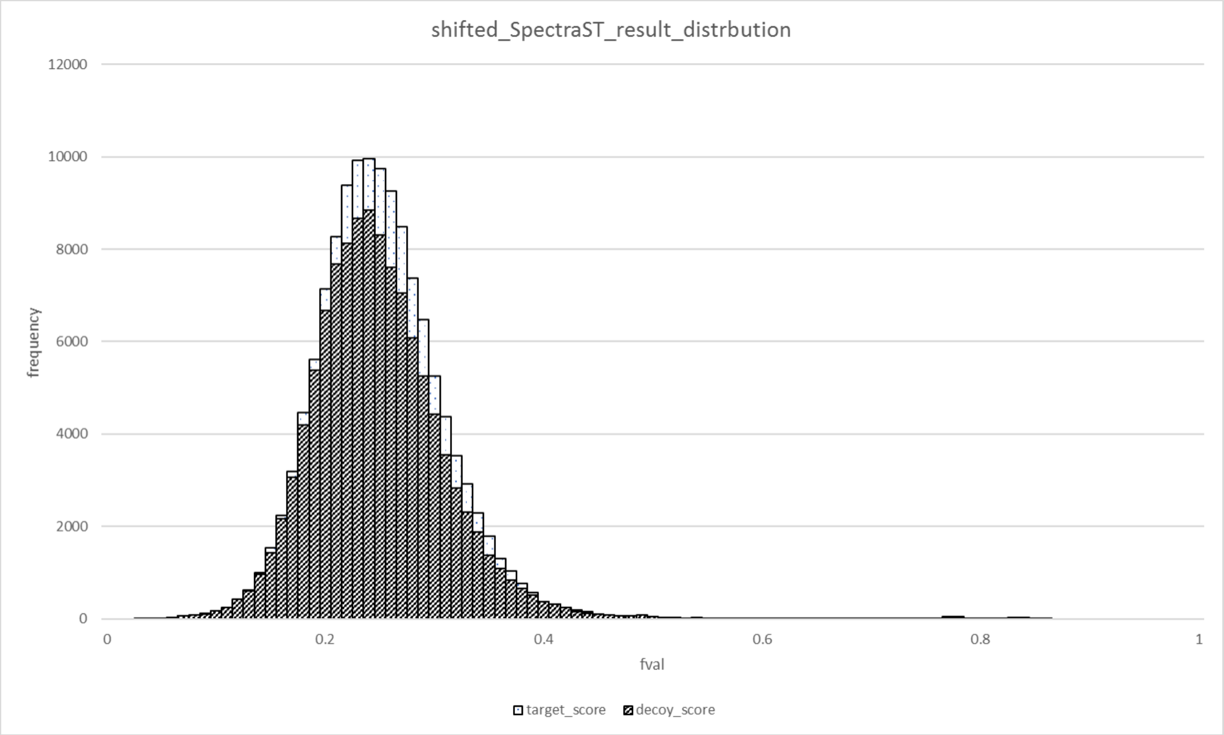

Figure 5.1: Score distribution of peptide identification results of precursor shifted dataset against target library and decoy library.

## 5.2 Peptides not in Library

There are spectra from the test data whose identified peptide sequences do not appear in the library. Therefore it is impossible for SpectraST to identify those spectra. This is a limitation of library searching, it cannot identify the query spectra if its corresponding peptide sequence and spectrum are not in the library. When searching, we will discard such spectra to ensure the fair comparison of the sequence database search and spectral library search. Another problem is that SpectraST cannot match the candidate spectral library with the query spectrum if they do not have same charge state. For example, if our query spectrum is identified as a certain peptide sequence with a charge state of 3 by JUMP, however our library only has the spectrum of that peptide sequence with a charge state of 2, then SpectraST will not try to match the candidate spectral library with our query spectrum and causes mismatch. We will also discard such a spectrum if the charge state of JUMP identified result is different from that of the library spectrum with the same peptide sequence.

## 5.3 Searching Parameters and Results

We search the test data set against our target-decoy consensus library, the parameters are set as below:

- TMT reporter peaks between 126 Th and 132 Th are removed. We do not expect these peaks will contribute to our spectra matching.
- The Scaling factor for unannotated peaks in the library spectra is 0.2, which means all peaks of the library spectra will be taken into account but the unannotated scale 0.2 as a punishment when calculating the matching score. We use 0.2 as our punishment factor because we tried 0, 0.2, 0.5 and 1 for testing and 0.2 gives us the best results. We believe this is because although most unannotated peaks are noise peaks, it is possible that peaks of internal fragment ions will appear and they will not be annotated. But these peaks can actually help us match the query spectrum with the target. Therefore, 0.2 is a reasonable punishment factor to be chosen.
- Precursor m/z tolerance is 0.1 m/z, which means only library spectra with precursors no less or no more than 0.1 m/z compared to the query spectrum will be searched. Both results of JUMP and SpectraST are filtered within 1% fdr to ensure correctness. The score we use here is the SpectraST *fval* score, which is the combination of the dot product score of candidate library spectra and query spectrum and the score rank of the candidate library spectra.

Figure 5.2 shows the *fval* score distribution of the test dataset searching against target-decoy consensus library. For target score, we can clearly see two peaks which we believe one

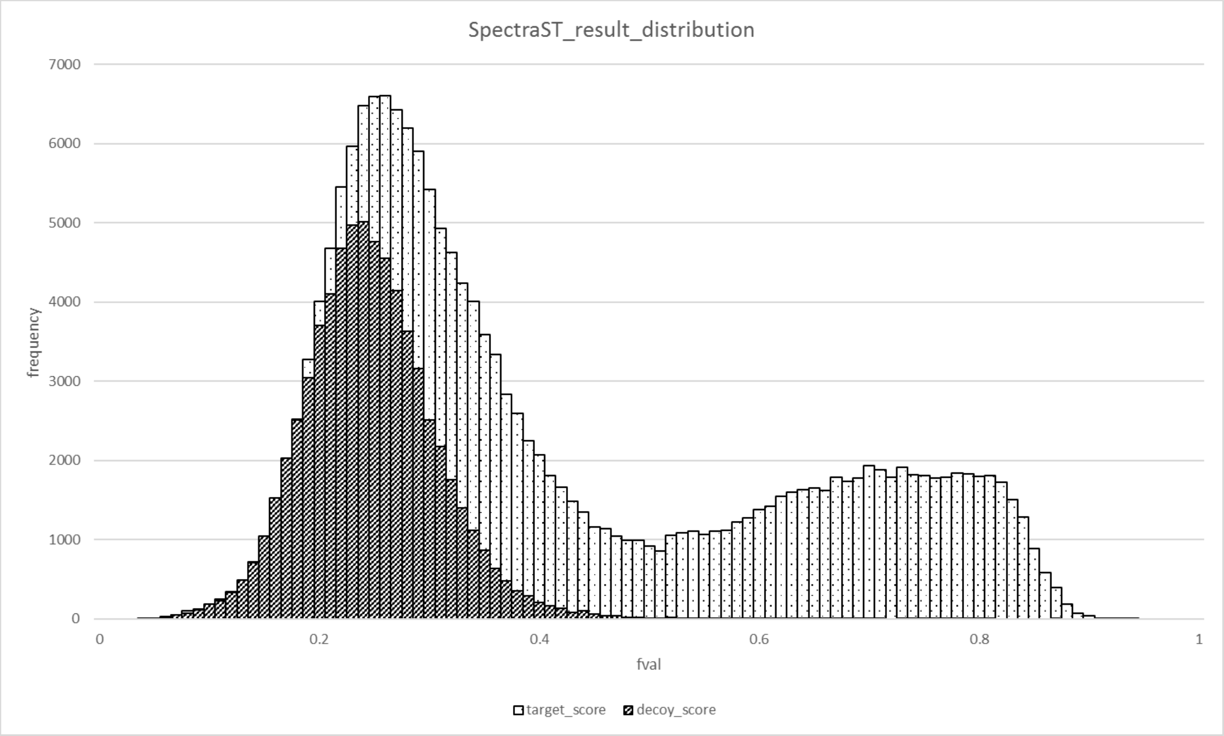

Figure 5.2: Score distribution of peptide identification results of test data against target-decoy consensus library generated by the precursor swap method.

is from the correct identified spectra and another is incorrect. These two peaks can be easily separated. The peak of the incorrect target score is higher than that of the decoy score, and we believe that because there are some correct target identifications with relatively low scores between 0.2 and 0.4 which makes the corresponding peak higher. Another possible cause is that as the query spectra will have mixed spectra, the score distribution of these spectra is different to that of normal spectra. We use the results of JUMP to distinguish the mixed spectra from our test data. There are a total of 88,728 mixed spectra in the five testing fractions. From Figure 5.3 and 5.4 we can see that the target score peak of the mixed spectra is lower than that of the normal spectra.

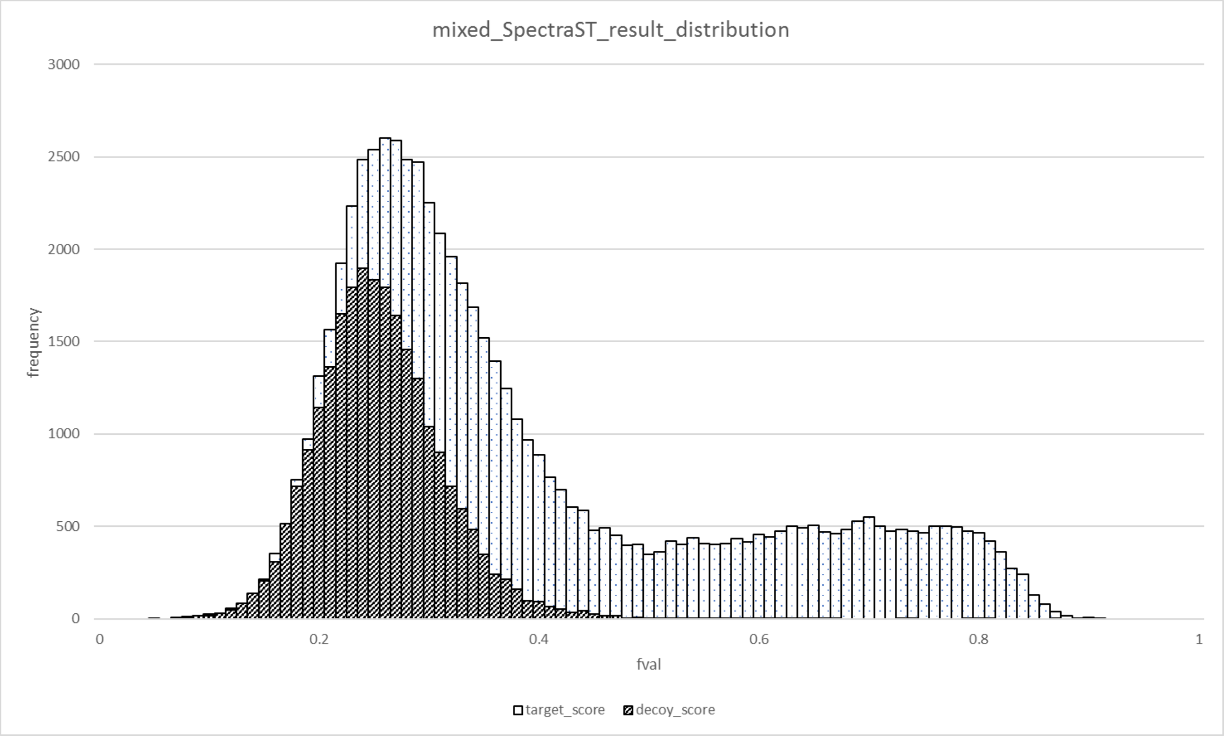

Figure 5.3: Score distribution of mixed spectra.

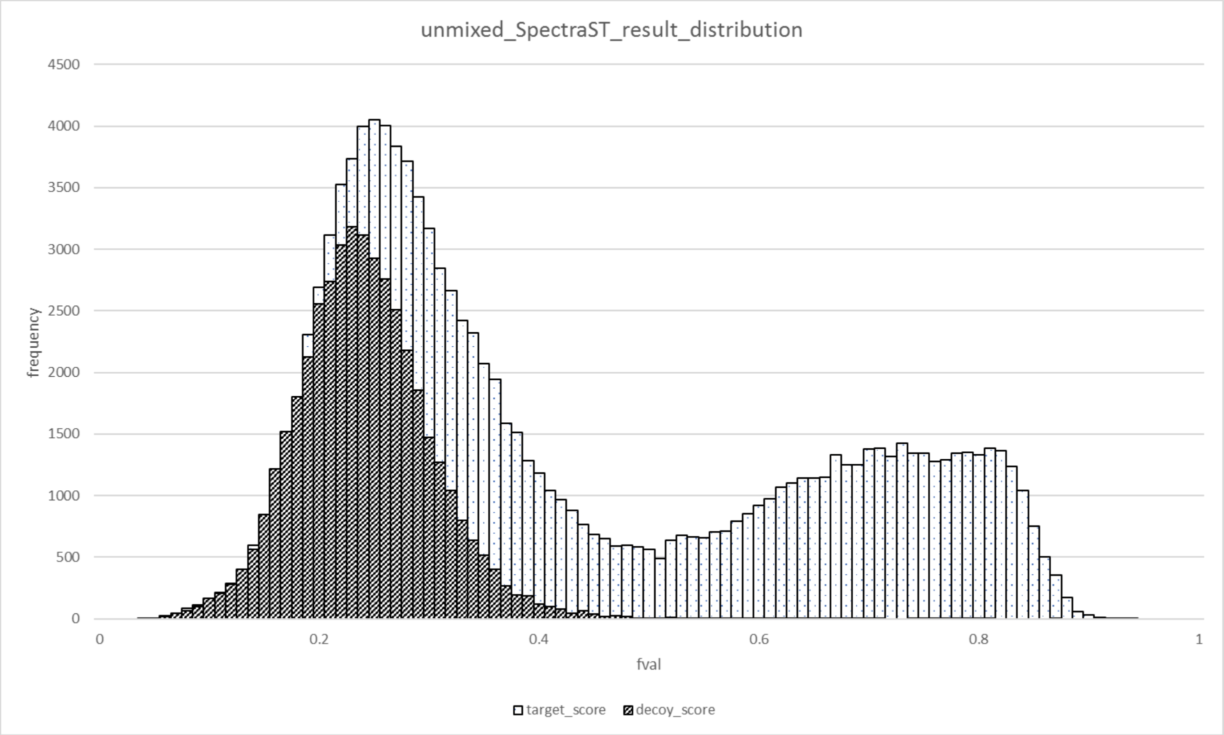

Figure 5.4: Score distribution of unmixed spectra.

Within 1% fdr, JUMP can identify 36,415 peptide-spectrum matches while SpectraST can identify 65,437 PSMs. These two results share 32,636 of the same identified results. About 89.6% of the JUMP results are included in the results of our spectral library search.

Figure 5.5 shows the Venn diagram of the two results from JUMP and SpectraST. All decoy results are discarded. The above set shows the number of spectra we discarded from test data because the sequence identified from JUMP is not in our library or the sequence is in the library but has a different charge state. Compared with the number of identified spectra by JUMP and SpectraST, the number of spectra we discard will not have a fundamental effect on the conclusion. Moreover, this problem can be solved or improved with the extension of our spectral library. For the 3,779 spectra that only appear in the result of JUMP, 411 of them also appear in the result of SpectraST and have the same identified peptide sequences but they are not in the 1% fdr cutoff of SpectraST. We believe that for the results of SpectraST below the 1% fdr cutoff, there are some correctly identified peptide sequences. Further effort can be done to improve the SpectraST *fval* score of such identifications.

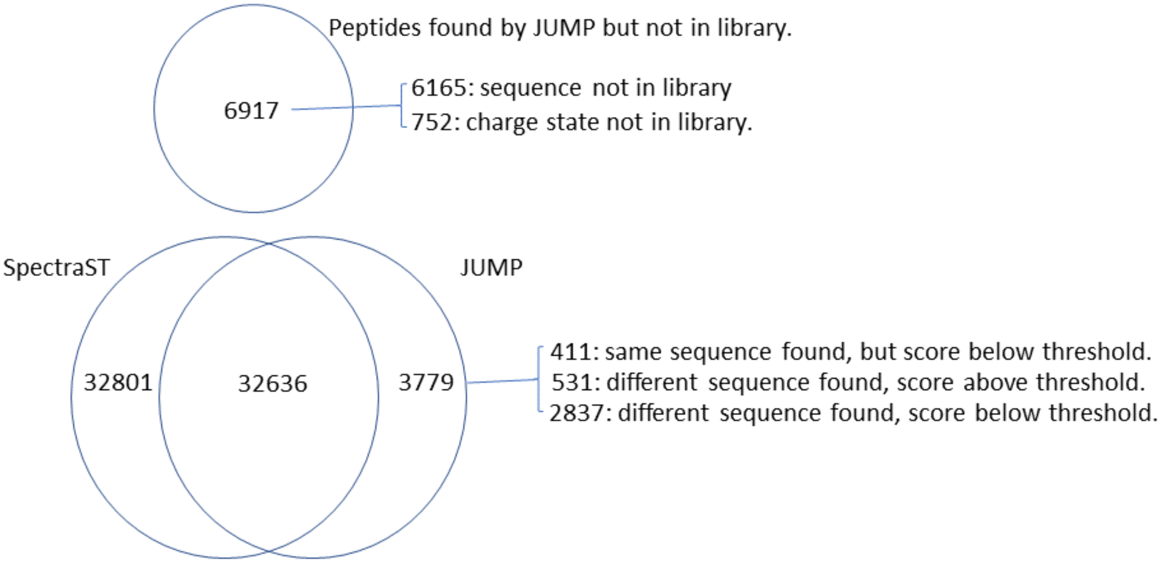
Peptides found by JUMP but not in library.
6917
6165: sequence not in library
752: charge state not in library.
SpectraST
JUMP
32801
32636
3779
411: same sequence found, but score below threshold.
531: different sequence found, score above threshold.
2837: different sequence found, score below threshold.

Figure 5.5: Venn diagram of peptide identification results of sequence database search and spectral library search.

# Chapter 6

# Conclusion and Future Work

## 6.1 Conclusion

In this thesis, we find the fragmentation ions' intensities differ between TMT-labeled peptides and label-free peptides. Therefore, we show the necessity in building a TMT-labeled spectral library from tandem mass spectra instead of simply using a conversion algorithm to convert TMT-labeled spectra and search them against a label-free spectral library. We also build a TMT-labeled spectral library from real tandem mass spectra using SpectraST[15]. We evaluate its quality by comparing it with the database searching tool JUMP[25]. The designed experiment shows that the decoy algorithm works for our test dataset and TMT-labeled spectral library. Therefore, we believe that using a false discovery rate to evaluate the performance of a spectral library is reasonable. The result shows that using a TMT-labeled spectral library helps improve the identification rate.

## 6.2 Future Work

Future work includes solving the problem of mixed spectra. In order to ensure the uniqueness of each spectrum in our library we discard all the mixed spectra. However, this operation will decrease the number of peptides we have in library because some peptide sequences can only be identified from those mixed spectra. We propose two possible solutions. The first is to build another spectral library that only contains those identified mixed spectra. If the query spectra can not be identified by the first normal spectral library, we can try to search it against the mixed spectra library. However, the problem still exists that the fragmentation ions from another precursor peptide will affect the matches with the query spectrum. Another approach is to extract those fragmentation ion peaks and generate several new spectra from the mixed spectrum. Each new spectrum can only correspond to one peptide sequence. The problem of this approach is that too many peaks are discarded from the mixed spectrum, which means we can only completely use the information of the original spectrum. More work can be focused on how to include mixed spectra into our library.

Another future work is to solve the problem of mixed query spectra. If the query spectrum is mixed, which means it contains the fragmentation ions from at least two peptides. Current spectral library searching can only return the identified sequences sorted by the score. However, the appearance of fragmentation ions from different peptides in a query spectrum will affect this score. One approach is to do a second search. To be more detailed, after we do the first round search for the query spectrum and find the best match in the spectral library, we will discard these matched fragmentation ion peaks in the query spectrum and search it again with the spectral library.

# References

[1] Erik Ahrné, Alexandre Masselot, Pierre-Alain Binz, Markus Müller, and Frederique Lisacek. A simple workflow to increase ms2 identification rate by subsequent spectral library search. *Proteomics*, 9(6):1731–1736, 2009.

[2] Vishwajeeth R Pagala Anthony A. High Viraj P. Ichhaporia Linda Hendershot Bing Bai, Haiyan Tan and Junmin Peng. Deep profiling of proteome and phosphoproteome by isobaric labeling, extensive liquid chromatography, and mass spectrometry. *Methods in Enzymol*, pages 377–395, 2017.

[3] Chia-Ying Cheng, Chia-Feng Tsai, Yu-Ju Chen, Ting-Yi Sung, and Wen-Lian Hsu. Spectrum-based method to generate good decoy libraries for spectral library searching in peptide identifications. *Journal of proteome research*, 12(5):2305–2310, 2013.

[4] John S Cottrell. Protein identification using ms/ms data. *Journal of proteomics*, 74(10):1842–1851, 2011.

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

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

[7] Robertson Craig, JC Cortens, David Fenyo, and Ronald C Beavis. Using annotated peptide mass spectrum libraries for protein identification. *Journal of proteome research*, 5(8):1843–1849, 2006.

[8] Costel C Darie. Post-translational modification (ptm) proteomics: challenges and perspectives. *Modern Chemistry & Applications*, 2013.

[9] 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.

[10] Barbara E Frewen, Gennifer E Merrihew, Christine C Wu, William Stafford Noble, and Michael J MacCoss. Analysis of peptide ms/ms spectra from large-scale proteomics experiments using spectrum libraries. *Analytical chemistry*, 78(16):5678–5684, 2006.

[11] 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. *arXiv preprint q-bio/0406002*, 2004.

[12] Johannes Griss. Spectral library searching in proteomics. *Proteomics*, 16(5):729–740, 2016.

[13] Henry Lam. Building and searching tandem mass spectral libraries for peptide identification. *Molecular & Cellular Proteomics*, 10(12):R111–008565, 2011.

[14] Henry Lam, Eric W Deutsch, and Ruedi Aebersold. Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. *Journal of proteome research*, 9(1):605–610, 2009.

[15] Henry Lam, Eric W Deutsch, James S Eddes, Jimmy K Eng, Nichole King, Stephen E Stein, and Ruedi Aebersold. Development and validation of a spectral library searching method for peptide identification from ms/ms. *Proteomics*, 7(5):655–667, 2007.

[16] Harvey Lodish, Arnold Berk, S Lawrence Zipursky, Paul Matsudaira, David Baltimore, James Darnell, et al. *Molecular cell biology*, volume 3. Scientific American Books New York, 1995.

[17] Bin Ma. Novor: real-time peptide de novo sequencing software. *Journal of the American Society for Mass Spectrometry*, 26(11):1885–1894, 2015.

[18] 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.

[19] Simin D Maleknia and Richard Johnson. Mass spectrometry of amino acids and proteins. *Amino Acids, Peptides and Proteins in Organic Chemistry: Analysis and Function of Amino Acids and Peptides, Volume 5*, pages 1–50, 2011.

[20] Alexey I Nesvizhskii. Protein identification by tandem mass spectrometry and sequence database searching. *Mass Spectrometry Data Analysis in Proteomics*, pages 87–119, 2007.

[21] Vishwajeeth R Pagala, Anthony A High, Xusheng Wang, Haiyan Tan, Kiran Kodali, Ashutosh Mishra, Kanisha Kavdia, Yanji Xu, Zhiping Wu, and Junmin Peng. Quantitative protein analysis by mass spectrometry. *Protein-Protein Interactions: Methods and Applications*, pages 281–305, 2015.

[22] Navin Rauniyar and John R Yates III. Isobaric labeling-based relative quantification in shotgun proteomics. *Journal of proteome research*, 13(12):5293, 2014.

[23] Roberto Romero, Juan Pedro Kusanovic, Francesca Gotsch, Offer Erez, Edi Vaisbuch, Shali Mazaki-Tovi, Allan Moser, Sunny Tam, John Leszyk, Stephen R Master, et al. Isobaric labeling and tandem mass spectrometry: a novel approach for profiling and quantifying proteins differentially expressed in amniotic fluid in preterm labor with and without intra-amniotic infection/inflammation. *The Journal of Maternal-Fetal & Neonatal Medicine*, 23(4):261–280, 2010.

[24] Chen Shao, Yang Zhang, and Wei Sun. Statistical characterization of hcd fragmentation patterns of tryptic peptides on an ltq orbitrap velos mass spectrometer. *Journal of proteomics*, 109:26–37, 2014.

[25] Xusheng Wang, Yuxin Li, Zhiping Wu, Hong Wang, Haiyan Tan, and Junmin Peng. Jump: a tag-based database search tool for peptide identification with high sensitivity and accuracy. *Molecular & Cellular Proteomics*, 13(12):3663–3673, 2014.

[26] Ding Ye, Yan Fu, Rui-Xiang Sun, Hai-Peng Wang, Zuo-Fei Yuan, Hao Chi, and Si-Min He. Open ms/ms spectral library search to identify unanticipated post-translational modifications and increase spectral identification rate. *Bioinformatics*, 26(12):i399–i406, 2010.

[27] 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.

[28] Yaoyang Zhang, Bryan R Fonslow, Bing Shan, Moon-Chang Baek, and John R Yates III. Protein analysis by shotgun/bottom-up proteomics. *Chemical reviews*, 113(4):2343, 2013.

[29] Zheng Zhang, Xiaoyu Yang, Yuri A Mirokhin, Dmitrii V Tchekhovskoi, Weihua Ji, Sanford P Markey, Jeri Roth, Pedatsur Neta, Deniz Baycin Hizal, Michael A Bowen, et al. Interconversion of peptide mass spectral libraries derivatized with itraq or tmt labels. *Journal of proteome research*, 15(9):3180–3187, 2016.