Protein De novo Sequencing

by

Rong Wang

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, 2016

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

In the proteomic mass spectrometry field, peptide and protein identification can be classified into two categories: database search that relies on existing peptide and protein databases and de novo sequencing with no prior knowledge. There are many unknown protein sequences in nature, especially those proteins that play an vital role in drug development pipelines, such as monoclonal antibodies and venoms. To sequence these unknown proteins, de novo sequencing is a necessity.

There have been standard algorithms for de novo sequencing a short peptide from its tandem mass spectrum (MS/MS). However, the de novo sequencing of a whole protein is still in its infancy.

The most promising method is to digest the protein into overlapping short peptides with different enzymes. After each peptide is de novo sequenced with MS/MS, these overlapping peptides are then assembled together either manually or with a computer algorithm. Such an automated assembly algorithm becomes the main purpose of this thesis.

Compared to the DNA sequence assembly counterpart, the main challenges are the high error rates and the short sequence length of each de novo peptide. To meet these challenges, novel scoring methods and algorithms are proposed and a software program is developed. The program is tested on a standard data set and demonstrates superior performance when compared to the state-of-the-art.

Acknowledgements

I would first like to thank my supervisor, Dr. Bin Ma, for directing the protein de novo sequencing project and providing insightful advices throughout my graduate study. I feel privileged to have such a great advisor.

I would like to thank my committee members, Dr. Lila Kari and Dr. Brendan McConkey, for generously spending time in reviewing my thesis and providing critical comments.

I would like to thank my parents, for their encouragement, endless love and understanding. I would like to thank my boyfriend, Jiasen Xu, for the love and support he gave me.

At last, I would like express my gratefulness to my colleagues in the bioinformatics research group, Chenyu Yao, Jianqiao Shen, Lian Yang, Qi Tang and Tiancong Wang for all the help they gave 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

In the computational proteomics field, most studies are focusing on identifying proteins by digesting sample proteins into peptides with enzymes, generating a tandem mass (MS/MS) spectrum for each peptide precursor, and then identifying the peptide sequence of each MS/ MS spectrum with a database search tool. Protein databases are generated from gene sequences within a given genome. However, due to sequence variation and the existence of unsequenced genomes, many protein sequences remain unknown. The sequences of some proteins, which are crucial in many therapeutic drug development pipelines, are not included in protein databases during the research stage. For example, trastuzumab (Herceptin) and alemtuzumab (MabCampath), which are monoclonal antibodies, have been successfully used on patients with breast cancer and graph-versus-host disease [22]. Captopril, a venom-based drug, has been successfully used on patients with cardiovascular disease[15] [16].

De novo protein sequencing was introduced for sequencing those proteins that are not included in the databases. Sometimes, even when a protein sequence is known, de novo protein sequencing can be utilized to discover novel forms of the protein generated from unexpected mutations, splicing events, and post-translational modifications (PTMs). Over

20 years ago, Johnson and Biemann sequenced a complete protein from rabbit bone marrow by manual interpretation of mass spectrometry data [13]. Edman degradation is another approach for sequencing novel proteins but it has limitations that make it unsuitable for sequencing proteins if the N-terminal amino acid has been chemically modified or if it is concealed within the body of the protein.

Fully automated de novo sequencing of unknown proteins is still a challenging problem. Experimental results are limited by ambiguous de novo interpretations, short peptide length and incomplete peptide fragmentation.

1.2 Research Objectives

In [3], Bandeira et al. proposed the shotgun protein sequencing (SPS) method, which obtains high coverage and accuracy. Sometimes, even if the target protein sequence is not included in the database, its homologous protein sequences are available. Comparative shotgun protein sequencing (cSPS) combines homologous sequences and SPS to improve de novo sequence coverage and accuracy [2]. Another method, meta-SPS [10] was proposed by Guthals et al. by assembling tandem mass (MS/MS) spectra from overlapping peptides. CHAMPS [17] first uses de novo peptide sequencing to interpret bottom-up MS/MS. Then the method uses a homologous sequence to align those peptide sequences to find overlapping peptide sequences and their positions relative to the homologous sequence. Finally, the peptide sequences are assembled to acquire a protein sequence. A de novo protein sequencing combining top-down and bottom-up MS was introduced in [16]. In their method, a top-down tandem mass spectrum is utilized as a scaffold, and bottom-up tandem mass spectra are aligned to the scaffold to increase sequence coverage.

The methods mentioned above either realize protein sequencing by assembling some type of mass spectra or use some kind of references to guide the alignment. However, with the development of mass spectrometry technology, new kinds of spectral data will be introduced. As a result, these mass spectrum based method should be changed accordingly. On top of that, we can not expect that there is always a homologous sequence in the database. Reference based methods are still limited when unknown sequences are encountered. Thus, in the new method, we focus on three objectives:

- Implement an automated assembly method to generate long length de novo protein sequence at high accuracy from mixed protein samples
- Operate directly on de novo peptides rather than mass spectra. By separating the de novo sequencing process from assembly process, we are no longer troubled by any upgrades in the mass spectrometry technology. Any improvements in the de novo sequencing field will be reflected in the results of our method. Besides, compared to the mass spectrum, the de novo peptide is a less complicated object. It is what the peptide originally looks like. Thus, compared to the mass spectrum based methods, de novo peptide based assembly allows the algorithm design effort to be focused on the result accuracy rather than dealing with the complexities of the spectral data.
- Remove the dependency on any reference protein sequences. By discarding reference data, like homologous sequence or top-down tandem mass spectra, this method is very promising in real situations where no related records of the experimental data can be found in the database.

1.3 Overview of the Thesis

The thesis is structured in the following chapters. In Chapter 2, we briefly review fundamentals for MS-based proteomics, such as mass spectrometry technology, database search approach and de novo sequencing approach. Chapter 3 presents some related works, such as CHAMPS [17], Meta-SPS[10] and TBNovo[16]. The details of the design of our novel idea, a de novo peptide based greedy algorithm, are included in Chapter 4. Implementation and experiments results are also provided in Chapter 4. Conclusions are presented in Chapter 5.

Chapter 2

Background

2.1 Fundamentals of mass spectrometry (MS)

Mass spectrometry

Mass spectrometry is an analytical technique that sorts ions based on their mass-to-charge ratio. It works by ionizing chemical compounds to generate charged molecules and measuring their mass-to-charge ratios. The mass is usually measured in Dalton (Da), which is 1/12 of the mass of a carbon atom, and is approximately the mass of a hydrogen atom. A mass spectrum is a plot of the ion signal as a function of the mass-to-charge ratio, which is used for analyzing the elemental composition of a sample or molecule, and for elucidating the chemical structures of molecules, such as peptides and other chemical compounds.

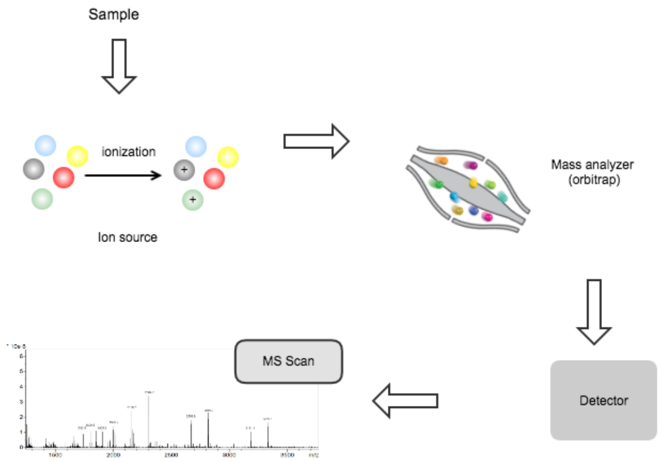


Figure 2.1: The basic components of a mass spectrometer (orbitrap [25])

As shown in Figure 2.1, a mass spectrometer consists of three parts: an ion source, a mass analyzer and a detector. The ionizer converts molecules or atoms into charged particles, which are called ions. In mass analyzer, which is the orbitrap in our example, ions are electrostatically trapped in an orbit around a central, spindle shaped electrode. The electrode confines the ions so that they both orbit around the central electrode and oscillate back and forth along the central electrode's long axis. This oscillation generates an image current in the detector plates which is recorded by the instrument. The frequencies of these image currents depend on the mass to charge ratios of the ions. Mass spectra are obtained by Fourier transformation of the recorded image currents[26].

There are several ionization techniques, depending on the phase (solid, liquid, gas) of the sample and the efficiency of various ionization mechanisms for the unknown species. Two techniques often used with liquid and solid biological samples are electrospray ionization (ESI, invented by John Fenn [5]) and matrix-assisted laser desorption/ionization (MALDI, developed by M. Karas and F. Hillenkamp [14]). ESI produces ions using an electrospray in which a high voltage is applied to a liquid to create an aerosol. It is especially useful in producing ions from macromolecules because it overcomes the propensity of these molecules to fragment when ionized. MALDI is a soft ionization technique used in mass spectrometry, allowing the analysis of biomolecules and large organic molecules, which tend to be fragile and fragment when ionized by more conventional ionization methods. Comparing with ESI, MALDI produces far fewer multiply charged ions.

There are two important parameters of a mass analyzer: mass resolving power, which is the measure of the ability to distinguish two peaks of slightly different mass-to-charge-ratio, and mass accuracy, which is the ratio of the m/z measurement error to the true m/z. It is usually measured in ppm (parts per million, 10^{-6}). Mass analyzers commonly used in proteomics are: quadrapole, time-of-flight (TOF), ion trap and Fourier transform ion cyclotron resonance (FT).

The detector records either the charge induced or the current produced when an ion passes by or hits a surface. In a scanning instrument, the signal produced in the detector during the course of the scan versus where the instrument is in the scan will produce a mass spectrum. In orbitraps, the detector consists of a pair of metal surfaces within the mass analyzer/ion trap region which the ions only pass near as they oscillate. No direct current is produced, only a weak AC image current is produced in a circuit between the electrodes, which is then converted to m/z spectrum.[26].

Tandem Mass Spectrometry

Tandem mass spectrometry, also known as MS/MS, involves two steps of mass spectrometry selection, with some form of fragmentation occurring in between the stages [9]. In a tandem mass spectrometer, in the first stage of mass spectrometry (MS1), ions are formed in the ion source and separated by mass-to-charge ratio. Ions from first stage are also called precursor ions or parent ions. These ions are then separated and detected in a second stage of mass spectrometry (MS2 or MS/MS), as shown in Figure 2.2. The scan which measures the peptides entering the spectrometer during a fixed time interval in the first stage is called survey scan or MS scan. Subsequently, a particular peak in the MS scan is selected. The instrument will fragment the corresponding ion and measure its product ions to form an MS/MS scan. Usually, one MS scan is followed by one to four MS/MS scans, each targeting a different peak in the MS scan.

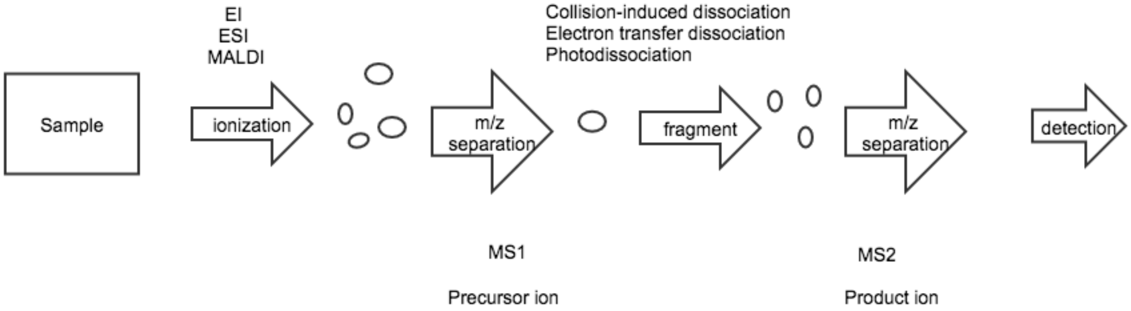


Figure 2.2: Schematic of tandem mass spectrometry

Figure 2.3 illustrates the possible fragmentation sites of a peptide. Fragment ions are labeled consecutively from the N-terminus (amino group) as a, b and c-ions, and also from the C-terminus (carboxyl group) as x, y, and z-ions. The most common and informative ions are generated by fragmentation at the amide bond between amino acids, resulting in b-ions if the charge is retained by the N-terminal part of the peptide and y-ions if the charge is retained by the C-terminal part.

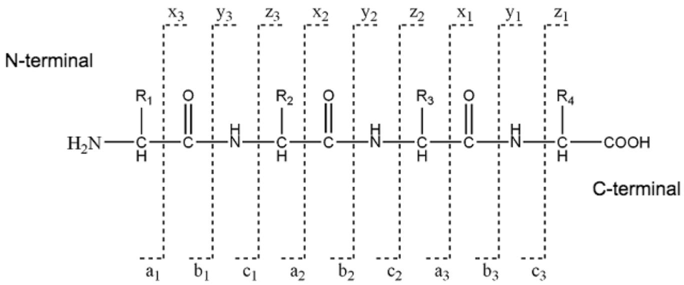

x3 y3 z3 x2 y2 z2 x1 y1 z1
N-terminal
R1 R2 R3 R4
H2N C C N C C N C C N C COOH
C-terminal
a1 b1 c1 a2 b2 c2 a3 b3 c3

Figure 2.3: Fragmentation sites of a peptide.

There are various methods to fragment molecules in MS/MS, including collision induced dissociation (CID), electron transfer dissociation (ETD), higher energy collisional dissociation (HCD) and others.

CID is currently the most commonly used fragment method, while other methods are used to enrich certain types of ions. Under CID condition, the peptide/protein precursor ion undergoes one or more collisions by interactions with neutral gas molecules, contributing to vibrational energy which will redistribute over the peptide/protein ion. The vibrational energy can result in ion dissociation occurring at amide bonds along the peptide backbone, generating b- and y-type fragment ions or leading to losses of small neutral molecules, such as water and/or ammonia or other fragments derived from side chains. In general, CID is more effective for small, low-charged peptides. [12] [21]

Complementary to CID fragmentation, electrontransfer dissociation (ETD) that transfers electron to a multiply protonated peptide/protein, could lead to the cleavage of the N-C α backbone bonds and to generate c- and z-type fragment ions [12]. Different ion types can provide complementary information for the structural characterization of a certain peptide. Another important feature of ETD fragmentation is that it can identify CID-labile post translational modifications (PTMs). Ideally, for peptides with PTMs, ETD can provide both the sequence information and the localization of the modification sites [21].

Another alternative type of fragmentation method is the high-energy collision dissociation (HCD). The fragmentation pattern of HCD is featured with higher activation energy and shorter activation time comparing the traditional ion trap CID. HCD also generates b and y-type fragment ions. While the higher energy for HCD leads to a predominance of y-ions, b-ions can be further fragmented to a-ions or smaller species [8], [21]. Without the low mass cut-off restriction and with high mass accuracy MS2 spectra, HCD has been successfully applied for de novo peptide sequencing, providing more informative ion series. As for PTMs studies, certain diagnostic ions specific for HCD could be recognized for PTMs identification [4].

In this thesis, the fragmentation methods of our data sets is the combination of collision induced dissociation (CID), electron transfer dissociation (ETD) and higher energy collisional dissociation (HCD), which takes advantage of corroborating b/y/c/z ions in

CID/HCD/ETD.

2.2 Interpret MS/MS Data

Identifying peptides from tandem mass spectrometry (MS/MS) data is an important task in proteomics. Protein identification from peptide hits and other following analysis are affected by the accuracy and sensitivity of this task directly. Many software tools have been developed for peptide identification; these tools can be broadly divided into two categories: de novo sequencing and database search, as shown in Figure 2.4.

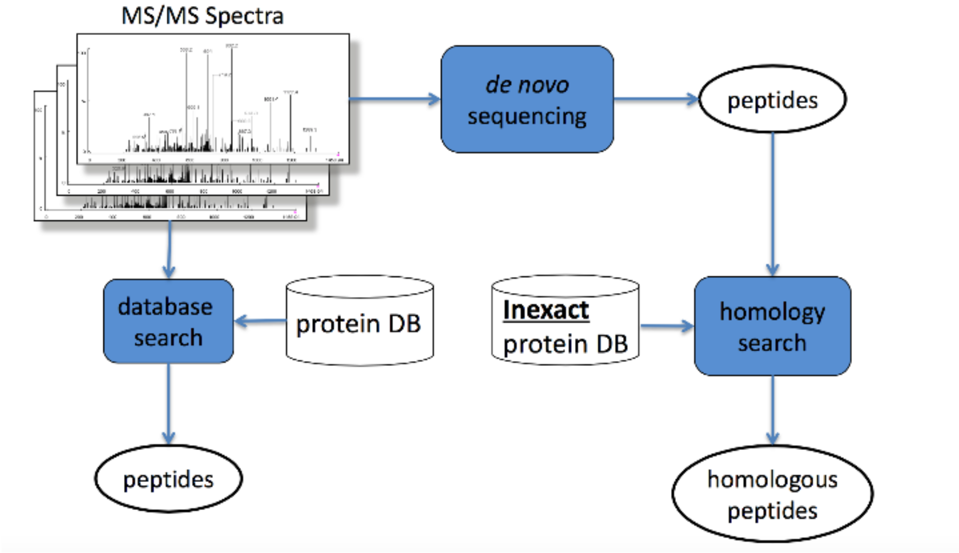


Figure 2.4: Possible ways to interpret MS/MS data [19]

2.2.1 Peptide Identification with DB Search

Peptide Identification with DB Search relies on existing protein databases. In this method, the first step is to digest the protein mixtures into peptides. The resulting peptides are then separated with liquid chromatography(LC) before the mass spectrometry measurement.

Both MS and MS/MS spectra are measured in the experiment. Besides the MS/MS data, a protein sequence database that contains all the target proteins is given. The primary task is selecting the correct proteins from the database [18].

There are two steps in the selecting task: identifying peptide sequences from the database using MS/MS spectrum, and identifying proteins from the grouped peptides. In the first step, the input, which is the acquired experimental MS/MS spectra data, is compared with theoretical spectra generated by peptides digested from the protein sequences that are in the database. A scoring function is then used to evaluate the similarity between the experimental data and the theoretical data. A good scoring function is important for the accuracy of peptide identification. Most commonly used scoring functions compute the theoretical m/z values of the fragment ions of the peptides, and matches the peaks of the spectrum with the m/z values. Higher scores are assigned to well matched spectra. Then the highest scoring peptide is reported as the answer.

After all the peptides are identified, protein identification is still a challenging problem because of several reasons, for example, not all peptides of a protein can be identified and each identified peptide may be shared by a few proteins in the database. Protein identification is the most mature application of mass spectrometry in proteomics. However, the software in use is still not perfect for reasons mentioned above [18].

The database search is generally believed to be a simpler approach because the protein sequence database provides a limited space for the software to search. Therefore, when a protein sequence database is available, a database search is the most common method for peptide identification [29]. There are many software tools using the database search approach, such as Mascot [1], X!Tandem [27], and PEAKS [29]. Until today, database search is still the most widely used method for peptide, protein identification.

2.2.2 Peptide De novo sequencing

De novo sequencing is another approach for peptide identification. It is typically performed without prior knowledge of the amino acid sequence. It is the process of interpreting amino acids from peptide fragment masses of a protein. A de novo sequencing algorithm takes an MS/MS spectrum as input, and outputs a peptide sequence that best matches the spectrum using a scoring function. De novo sequencing computation does not require a protein database, and it has proven successful for confirming and expanding upon results from database searches.

The spectrum graph approach is used in some de novo sequencing softwares, such as SeqMS[6], Lutefisk[23] [24] and PepNovo[7]. This approach converts a spectrum into a graph, where each vertex corresponds to a possible ion related to a peak. Each edge connects two vertices whose corresponding ions have a mass difference approximately equal to the mass of an amino acid. The sequence that will be sought is an optimal path connecting the two termini, as shown in Figure 2.5. The path starts at vertices that correspond to the N and C termini [28].

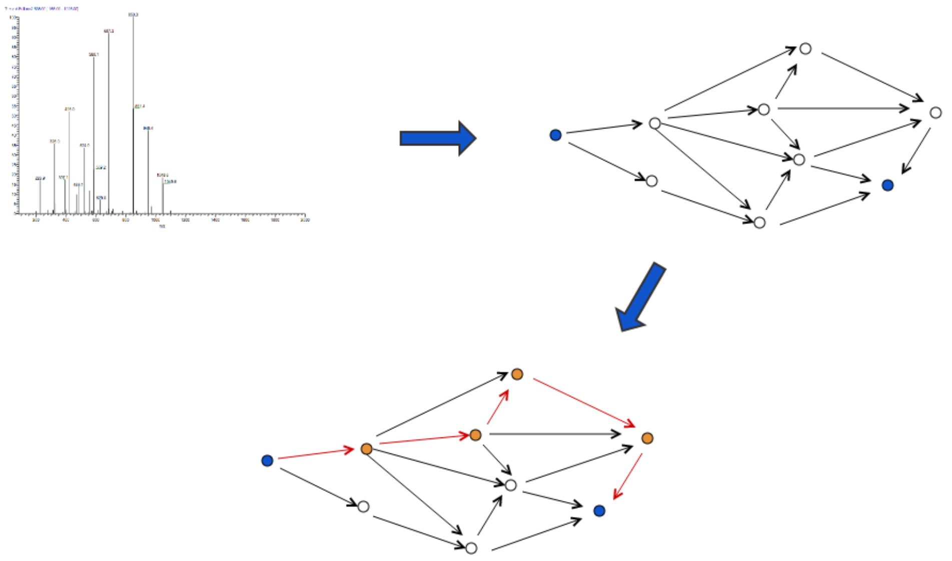


Figure 2.5: Spectrum graph

Another commonly used de novo software package is PEAKS[20]. It uses an algorithm that differs from the graph approach. The algorithm works directly on the spectrum by first computing a y-ion matching score and a b-ion matching score at each mass value according to the peaks around it. A penalty value is assigned if there are no peaks around a mass value. The algorithm then computes many amino acid sequences that maximize the total scores at the mass values of b-ions and y-ions efficiently . A more accurate scoring function is used to further evaluate these candidate sequences. The scoring function also considers other ion types such as immonium ions and internal-cleavage ions. The problem of ion absence is addressed because the PEAKS model assigns a score (or penalty) for each mass value. The software also computes a confidence score for each amino acid in the final result by examining the consensus of the top-scoring peptides [28].

Chapter 3

Related Work

Although full-length de novo sequencing of unknown proteins remains a challenging open problem, progress in this area has already been made by a number of pioneering works. We review some of these works in this section.

3.1 CHAMPS

An automated protein (re)sequencing with MS/MS and a homologous database method is proposed in [17]. The method requires that a homologous sequence of the target protein should be included in a given protein sequence database. Homologous sequence means the sequenced genome that belongs to a close relative of the studied species. In their experiment, the homologous sequence is found using PEAKS software. CHAMPS includes the steps shown in Figure 3.1.

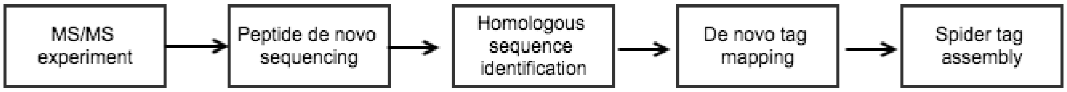

MS/MS experiment
Peptide de novo sequencing
Homologous sequence identification
De novo tag mapping
Spider tag assembly

Figure 3.1: CHAMPS workflow

In the de novo tag mapping step, de novo tags are mapped to the reference sequence using SPIDER algorithm[11]. A similar peptide from the reference sequence should be found for each de novo tag. These similar peptides from the reference sequence are called the homolog tags. During this mapping process, mismatch errors are considered by assign different weights to different error types. By using SPIDER algorithm, for each de novo tag and its corresponding homolog tag found at the reference sequence, a predicted "middle sequence" is generated, which is called spider tag in their method. Many spider tags have already been anchored onto the reference sequence in this step. Pairwise sequence alignment between each spider tag and its corresponding homolog tag is also computed by SPIDER during the mapping. Then in the spider tag assembly step, a score function is used when merging these pairwise alignments together to minimize the error.

Depending upon the level of similarity between reference and target, CHAMPS can correct de novo sequencing errors and anchor sequences to the reference. However, the limitation of this method is that the mapping and assembling step both rely on the condition that a homologous protein of the target protein is included in a known database, which can not be guaranteed every time.

3.2 MetaSPS

Protein sequencing by merging triplet CID/HCD/ETD MS/MS spectra from overlapping peptides is introduced in [10]. The process of this method is shown in Figure 3.2.

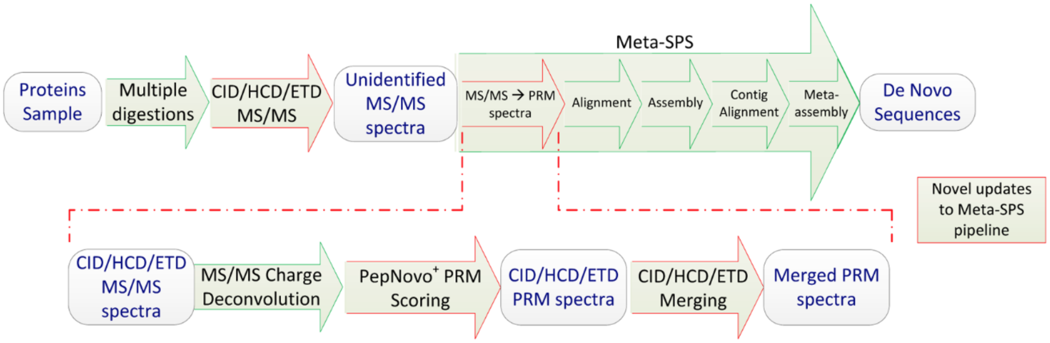


Figure 3.2: MetaSPS workflow

First, Meta-SPS uses PepNovo$^+$ to interpret MS/MS fragmentation patterns and convert MS/MS spectra into PRM (prefix residue mass) spectra rather than processing MS/MS spectra directly. In the spectra, log-likelihood scores are used to replace peak intensities and peak masses are replaced by Prefix-Residue Masses (cumulative amino acid masses of N-term prefixes of the peptide sequence). They trained their scoring models for deconvoluted high-resolution CID, HCD, and ETD MS/MS spectra using multiple data sets. After training the score model, they merged the CID, HCD and/ or ETD PRM spectrum from the same precursor into a single merged PRM spectrum, by extracting corroborating PRMs and SRMs from CID/ ETD and HCD/ETD pairs from PRM spectra and inserting the corresponding combined PRMs into the merged spectrum.

To merge CID/ETD or HCD/ETD pairs, they consider all PRM/PRM matches, SRM/ SRM (Suffix- Residue Masses, cumulative amino acid masses of C-terminal suffixes of the peptide sequence) matches with at least one PRM, PRM/SRM and SRM/PRM pairs and SRM/SRM matches without PRMs in from PRM spectra.

This method assembled tandem mass (MS/MS) spectra from overlapping peptides by using multiple enzymatic digests, combining electron-transfer dissociation (ETD) with collision-induced dissociation (CID) and higher-energy collision-induced dissociation (HCD) fragmentation methods to boost interpretation of long, highly charged peptides and taking advantage of corroborating b/y/c/z ions in CID/HCD/ETD without using reference sequence. However, the process requires complicated mass spectra preprocess and the results are still not satisfactory.

3.3 TBNovo

Top-down tandem mass spectra cover whole proteins. While, top-down tandem mass spectra, even combined, rarely provide full ion fragmentation coverage of a protein. TBNovo, which is proposed in [16], combines top-down and bottom-up MS to assemble mass spectra of overlapping peptides. A top-down tandem mass spectrum is utilized as a scaffold, and bottom-up tandem mass spectra are aligned to the scaffold to increase sequence coverage. The process of TBNovo is shown in Figure 3.3

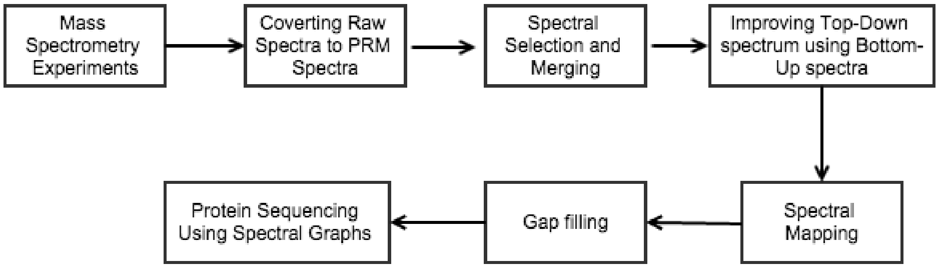

Mass Spectrometry Experiments
Coverting Raw Spectra to PRM Spectra
Spectral Selection and Merging
Improving Top-Down spectrum using Bottom-Up spectra
Spectral Mapping
Gap filling
Protein Sequencing Using Spectral Graphs

Figure 3.3: TBNovo workflow

The method only kept a top-down MS/MS spectrum if its precursor mass is the same (within an error tolerance) to the theoretical precursor mass of the protein. Both top-down and bottom-up PRM spectra are filtered to remove low quality ones in order to provide accurate information for de novo sequencing. In the merge step, all top-down PRM spectra have the same (within an error tolerance) precursor mass are merged into one spectrum to increase protein coverage. If the mass difference between two PRMs in the merged spectrum is smaller than an error tolerance, the two PRMs are merged into one by removing the lower intensity one. Then bottom-up Spectra is used to increase the number of correct PRMs and decrease the number of incorrect PRMs in the top-down spectrum and mapped to top-down spectra. After refining the top-down PRM spectrum, TBNovo find a protein sequence P with a corresponding PRM spectrum that best explains the top-down PRM spectrum by using a spectral graph, where each node represents a PRM.

Combining top-down and bottom-up MS/MS spectra can remove the dependency on the relative positions in the reference sequence reported by some existing approaches using only bottom-up MS/MS spectra. However, top-down MS/MS spectra data is not easy to get, because most laboratories use bottom-up methods to acquire MS/MS spectra.

Chapter 4

Methodology

4.1 Method

As detailed in Chapter 1, fully automated de novo sequencing of unknown proteins is still a challenging problem, because experimental results are mainly limited by ambiguous de novo interpretations, short peptide length and incomplete peptide fragmentation. Existing methods to solve this problem are mainly mass spectra based. However, compared to the mass spectrum, the de novo peptide is a much simpler object because we can simply treat it as a string. As detailed in section 1.2, assembling those de novo peptides directly by using the overlap information between them rather than merging mass spectra is a better and more straightforward method.

To validate our idea, we first design an experiment. In this experiment, database search results and de novo sequencing results for the same protein are obtained by using PEAKS [20]. Then the peptide from database search was replaced by the de novo peptide interpreted from the same MS/MS spectrum. We observed that the information of peptide overlap indeed provides us with some guidance for protein sequencing.

One critical problem we need to solve is how to use this overlap information. One natural thought is to keep assembling these overlapping peptides until there are no peptides to support the assembly process, as shown in Figure 4.1. However, simply assembling

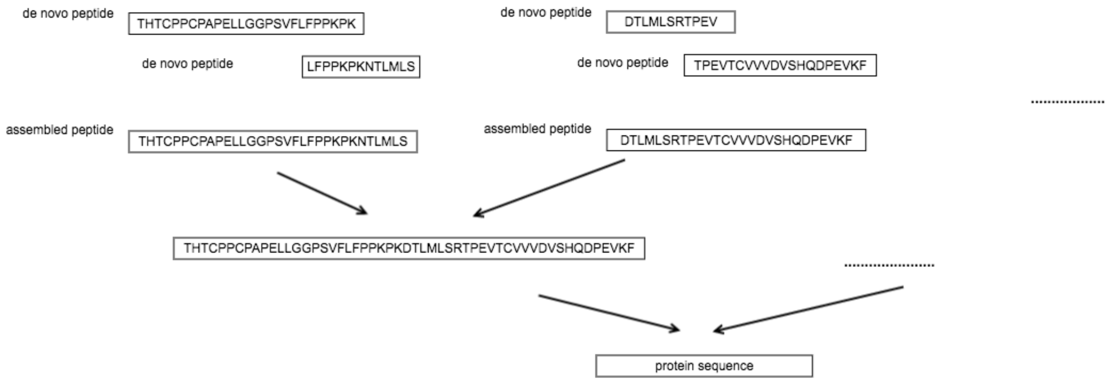


Figure 4.1: Overlap between de novo tags and assembling process

peptides in this way can not solve the following problems :

- The error in de novo peptides may cause the overlapping parts of two proteins are not exactly equal. When this happens, how to detect the overlaps, and how to determine the amino acid sequence of the overlapping part after merging the two peptides?

- Two peptides from different parts of the same protein, or even from two proteins may have significant overlap due to repeats in the protein. This may lead to erroneous merging.

- A single inaccurate assembly can mess up all downstream assembly process. We should have a way to avoid the introduction of errors in the early stage of the assembly.

In this section, we propose a fully automated de novo protein sequencing approach based on a greedy algorithm. In this algorithm, to solve the assembly problems mentioned above, a score function is designed to evaluate the quality of the overlap between two overlapped peptide sequences. Then, the algorithm chooses to assemble the sequence with the highest score and replaces the original peptide sequence with the assembled sequence.

4.1.1 Method Overview

Our method utilizes a two-pass approach. The first pass is a traditional de novo sequencing by using PEAKS software for peptide identification with specified PTMs. Inputs of the first pass are LC-MS/MS, which are from multiple-enzyme digestions of a group of proteins, with three fragmentation modes, CID, HCD and ETD. Then we filter out those identified peptides with a low confidence score. We call these peptides de novo peptide tags. The second pass merges identified peptides in pairs and replaces the original peptide with the merged consensus peptide, while evaluating the merging quality. We call the peptide assembled via overlapping peptide tags that represent a consensus region of protein contig. The second pass consists of three major steps:

- **Peptides pairing and merging:** In this step, overlapped peptides are paired. These peptide pairs are merged using mass match method. After merging, we get peptide merging candidates and we put them in a candidate table.
- **Candidate evaluation using score function:** All candidates are scored by combining the peptides overlap score and the support score.
  - Overlap score: This feature evaluates the candidate based on how similar the overlapped part between two merged peptides is.
  - Support score: This feature evaluates the candidate based on how well the other peptides support the merge of the two peptides.
- **Candidate competing:** The candidate with the highest score is chosen. Then we need to update the candidate table accordingly.
- **Contig merging:** Final sequences with long sequence length and high accuracy are merged.

Our method outputs a confidence score for the final sequence and a local confidence score for each amino acid within it. The confidence score of the candidate is the average score of every amino acid. The score is obtained from PEAKS software for each de novo peptide initially. Then in the merge step, the confidence score for each amino acid in the candidate is inherited from its parents for the non-overlapping parts, and updated carefully for the overlapping amino acids.

4.1.2 Peptide de novo sequencing

For each MS/MS spectrum, the de novo peptide sequence is computed using PEAKS 7.0 software. Each spectrum will associated with a few possible sequences computed by the software. PEAKS makes no distinction between the amino acids L with I, and K with Q. In our analysis, we only use the de novo sequence which is the highest scoring sequence of each spectrum. PEAKS software also outputs a local confidence score for each amino acid in the sequence and the confidence score for the whole sequence. These two scores are also recorded for the following analysis.

To minimize the influence of de novo sequencing errors, we first choose a confidence score threshold to filter out some low quality initial tags obtained from PEAKS. The confidence score reported by PEAKS, which ranges from 0 to 100%, can illustrate the accuracy of the de novo peptide tag to some extent. Although a lower confidence score threshold (*e.g.* 50%) can include more peptide tags, it will also introduce more de novo sequencing mistakes and reduce the efficiency. Higher confidence score threshold (*e.g.* $>$ 80%) can guarantee the correctness, but it will lose a lot of information. We weighed the pros and cons and decided to included de novo peptide tags with the confidence score higher than 70%.

4.1.3 Peptides pairing and merging

In our method, to make assembly process happen, we first need to find overlapped peptides and their overlapped region efficiently. We do not want to waste our time on assembling two unrelated peptides. To limit our search space, only peptides that can be possibly paired together should be considered. It is very unlikely that two peptides without any matching amino acids are actually overlapped in the original protein. So we only considered peptide pairs that have at least n continuously matching amino acids. In our method, we set n

to 3. In our experiment, we also tried 4-mers and 5-mers matching. However, by using a larger n-mer, many possible peptide pairs are missed. If we use a smaller n, false-positive rate would increase, and we need to spend time on evaluating wrong merging choices. By using 3-mers, we reduce our search space and minimize the loss of information as much as possible.

Mass match

Then, to find the border of the overlap region, a mass match method is used. The reason for using mass match method is that peptides are identified from mass spectra and due to ambiguous de novo interpretations, instead of having the same amino acids sequence, the overlapped regions of the peptide pair tends to have the same mass distribution. Rather than extending the overlapping area by comparing the amino acid, we extend the overlapping area by comparing the mass of amino acids, as detailed in Algorithm 1 in section 4.1.7.

After determine the overlap region, we can merge these two peptides, as detailed in section 4.1.7. After merging, we get a new candidate formed by peptide a and b. The new candidate consists of three parts, left end, consensus part and right end, as shown in Figure 4.2.

After merging, we get a candidate table.

Consensus part

Errors will accumulate in the assembled sequence if we only consider the information that comes from one of the overlapped peptide pairs. To guarantee the quality of the assembled sequence, we calculate the consensus part between two overlapped peptides. Consensus part is obtained from the overlap part between peptide tag a and tag b. When doing de novo sequencing, PEAKS assign confidence score to each amino acid. Our method makes use of this property to choose more confident amino acid between tag a and b to generate the consensus part. Details about how to generate consensus part and new candidate are discussed in section 4.1.7.

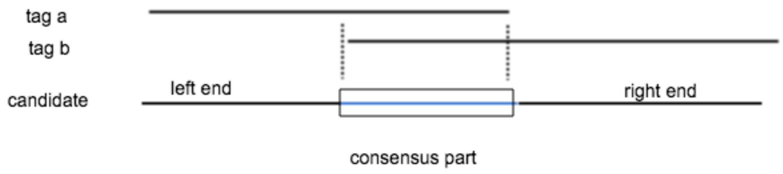


Figure 4.2: Consensus part

The following evaluation is performed on those candidates.

4.1.4 Candidate Evaluation Using Score Function

The algorithm uses a score function to evaluate and choose among peptide merging candidates. The scoring function consists of two parts, the overlap score part and the support score part. Based on our observation, the longer the overlapping part of two peptides is, the more likely that the two peptides should be assembled together. However, the overlap score alone is unable to tell whether an overlap is a real or a false-positive match. Because peptides from different parts of the protein or different proteins may have very similar and sometimes the same amino acid sequence. The support score is designed to make the distinction. The peptide tags used to calculate the support score are called support tags.

Let a, b be the de novo tags, and c_i be the support de novo tag, shown in Figure 4.3. We will define the score function as follows.

$$score(a, b) = score_{overlap}(a, b) + \sum_{i=1}^{3} \frac{1}{i+1} score_{support}(a, b|c_i) \tag{4.1}$$

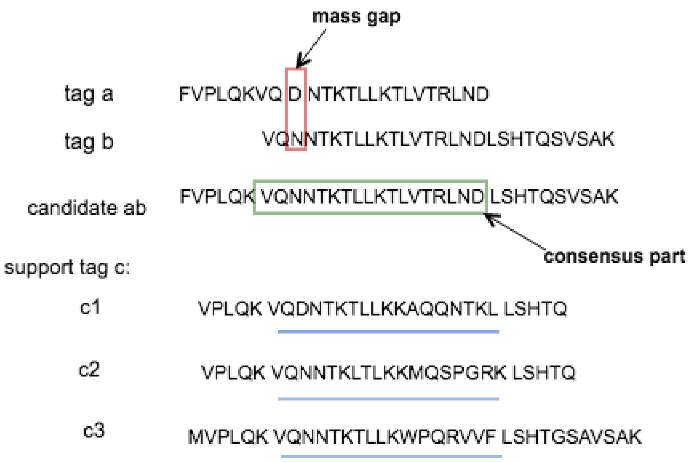

mass gap
tag a
FVPLQKVQ D NTKTLLKTLVTRLND
tag b
VQNNTKTLLKTLVTRLNDLSHTQSVSAK
candidate ab
FVPLQK VQNNTKTLLKTLVTRLND LSHTQSVSAK
consensus part
support tag c:
c1
VPLQK VQDNTKTLLKKAQQNTKL LSHTQ
c2
VPLQK VQNNTKLTLKKMQSPGRK LSHTQ
c3
MVPLQK VQNNTKTLLKWPQRVVF LSHTGSAVSAK

Figure 4.3: score(a,b) explanation

Overlap score

The first term of the definition $score_{overlap}(a, b)$ is the overlap score between a and b. Mass replacement errors are common in peptide de novo sequencing because of the noise in the mass spectrum and PTM. For example, mass(AT) = mass(TA) and mass(RDG) ≈ mass(VTK), as shown in Table 4.1.

Because of the existence of mass gap errors in the de novo peptides, both exact match and mass gap match are considered. The definition of this alignment score is:

$$score_{overlap}(a, b) = \max \begin{cases} 0 \\ (aaMatch + \alpha * massMatch + \\ r\_score - \gamma) * reliability \end{cases} \tag{4.2}$$

- *aaMatch* is the number of exactly matched amino acids between a and b. We prefer the overlap with more exactly matched amino acids. This is reasonable because that the overlap between two peptides is generated by different enzyme cleaving the protein at different sites. Peptides that are from the same segment in the original

Table 4.1: Mass replacement table.

Mass	Sequence
113.0841	I, L
114.0429	N, GG, D
128.0586	Q, AG
160.0307	C(+57), CG
129.0426	E, Q(+0.98)
170.1055	AV, GL
...	

protein will share some common segments. Therefore, the more exactly matched amino acids these two peptides share, the more likely they are from the same segment.

- *massMatch* is the number of matched mass gap, as shown in Table 4.1 and Figure 4.3. We include matched mass gap because we are assembling de novo sequencing tags. De novo sequencing sometimes gives only partially correct tags. The most common error is that a segment of amino acids is replaced by another segment with approximately the same masses. As a result, two peptides that are from the same region may only share limited number of exactly matched amino acids. We will miss many potential peptide merging candidates if we consider the exact amino acids match only in this case. We do not count the mass gap that expands more than three amino acids, because the longer the mass match region is, the more unreliable it is. According to our observation, matching mass gap that expands more than ten amino acids can occur between two unrelated sequences, which is obviously not reliable. We value exact amino acid match more since it is more accurate. We assign a weight α to *massMatch*, which is 0.9 in our method.

- *r_score* is served as a reward for the overlap score. Peptide pairs with unmatched tails or heads are not very reliable. Since de novo sequencing sometimes gives only partially correct tags and sequencing errors tend to distribute towards to both ends of the sequence, we cannot determine whether the reason for unmatched left ends or right ends is the bad pairing choice or the sequencing error. We prefer peptide pairs with matched heads or tails and we will give rewards to the peptide merging candidates that satisfy the following conditions:

  - Overlap parts that ends at the end of tag a
  - Overlap starts from the beginning of tag b, meanwhile, the amino acid at the beginning of tag b and the amino acid before the overlap beginning position of tag a form a cleavage site

- γ is the threshold that can filter some poor-quality assembling. We prefer peptide pairs with better matching, which means more exactly matched amino acids and high quality mass gaps. By tuning γ, we can control the merging quality.

- *reliability* equals to the minimum value between *conf_b* and *conf_a*, which are the confidence score of tag a and b respectively. In PEAKS[29] software, the confidence score reflects the accuracy for each de novo tag to a certain extent. In our method, we hope to obtain assembled sequence with high accuracy, so we tend to choose peptide pairs having high confidence score as its parents.

Support score

When the overlap information is strong enough, we can get a very promising result. However, in a real situation, similar and sometimes the same amino acid sequences can appear at the different segments within the same protein. Besides, de novo sequencing errors are quite common. We may merge the tag which is in the N-terminal part of the original protein with the tag that belongs to the C-terminal part of the original protein by mistake if considering the overlap information only. This problem can be more challenging when dealing with several proteins at the same time.

However, based on our observation, similar or the same amino acid sequences of two unrelated peptide tags are unlikely to be very long, and it is very unlikely to find the third sequence that cover both the matching area and the neighbourhood of the matching area. On the contrary, if the two peptide tags can form a true merging pairs, either the overlap between them can become quite long or there are several tags covering both the overlap area and the neighbourhood of the overlap area.

Based on the observation, we introduced the support score, $score_{support}(a, b|c_i)$ in formula 4.1. Peptide merging candidates are chosen depending on both the overlap score and the support from other peptide tags that cover both the overlap region and the neighbourhood of the overlap area. Support score is served as a complement score when it is not enough to distinguish the real match from the false-positive match by overlap score. Peptide tags used to calculate support scores are called support tags. The support score is defined as follows:

Assuming the overlap part between a and b is *ab_overlap*, $score_{overlap}(c_i, ab_{overlap})$ is the overlap score between support tag *c_i* and the *ab_overlap* using the formula 4.3. $score(c_i, ab_{left})$ and $score(c_i, ab_{right})$ are illustrated in Figure 4.4

$$score_{overlap}(c_i, ab_{overlap}) = \max \begin{cases} 0 \\ (aaMatch + \alpha * massMatch - \gamma) * reliability \end{cases} \tag{4.3}$$

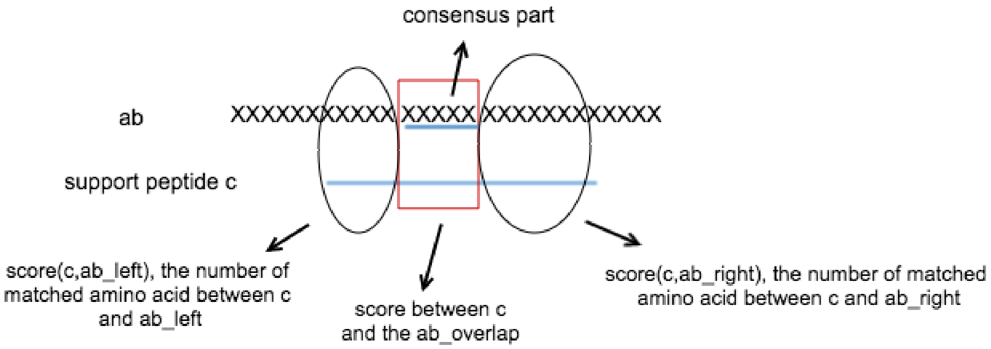

consensus part
ab
XXXXXXXXXXX XXXXX XXXXXXXXXXXX
support peptide c
score(c,ab_left), the number of matched amino acid between c and ab_left
score between c and the ab_overlap
score(c,ab_right), the number of matched amino acid between c and ab_right

Figure 4.4: support score

$$score_{Support} = \min(score_{overlap}(c_i, ab_{overlap}), score(c_i, ab_{left}), score(c_i, ab_{right})) \tag{4.4}$$

The first method we used to calculate the support score is different than the final version, which is illustrated as follows. Assuming there is peptide tag c that overlap with both peptide a and peptide b, we calculate the overlap score between a and c excluding the overlapping part between a and b using the formula 4.3. We use the same method to calculate the score between b and c. Then we choose the smaller one from *score_ac* and *score_bc* as the support score contributed by tag c.

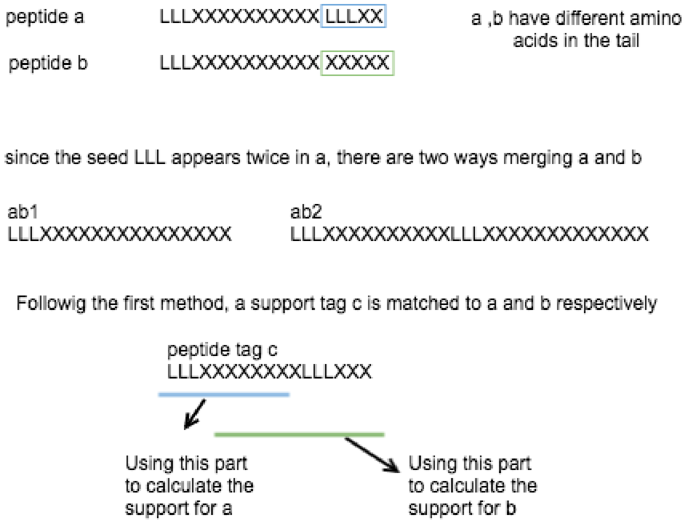

peptide a
LLLXXXXXXXXXX LLLXX
peptide b
LLLXXXXXXXXXX XXXXX
a ,b have different amino acids in the tail
since the seed LLL appears twice in a, there are two ways merging a and b
ab1
LLLXXXXXXXXXXXXXXX
ab2
LLLXXXXXXXXXXLLLXXXXXXXXXXXXX
Followig the first method, a support tag c is matched to a and b respectively
peptide tag c
LLLXXXXXXXXLLLXXX
Using this part to calculate the support for a
Using this part to calculate the support for b

Figure 4.5: support score using first method

However, from our experiments, we found out that the first method we used here cannot deal with the situation shown in Figure 4.5, peptide a and peptide b have almost the same amino acid sequences except for the ending part.

In this case, although the matching score between a,b is low, the support score for ab2 is high. Therefore, using the formula 4.4 to evaluate the support from c to ab2 is more appropriate. For all the support tags to a candidate, we use formula 4.4 to calculate the support contributed by one tag and then sort these support tags based on the support they give. Then we only consider the support from the first three support tags, because we do not want to give those candidates that have poor overlap quality but are supported by many support tags priority, like the situation in Figure 4.6. Candidate ab2 is more reliable than candidate ab1, however, support tags give more support to candidate ab1. By limiting the number of support tags to 3, we minimize this side effect by introducing support tags.

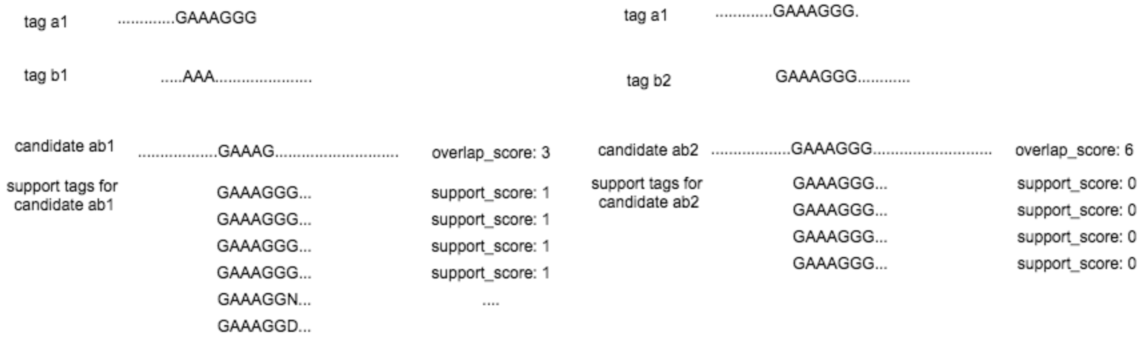


Figure 4.6: A candidate with poorer overlap quality but higher support score

4.1.5 Candidate competing

Guaranteeing the correctness of each step is quite important. If we merge a false-positive match first, we will end up with a mess. As we mentioned in previous section, our score function can distinguish the real match from a false-positive match by assigning a higher support score when overlap scores are very close. Thus, by choosing candidate with higher score first, we exclude the false-positive match. Meanwhile, we prefer the candidate with higher quality. According to our score function, candidates generated from merging peptide pairs that share longer and more reliable overlap regions and those that are supported by support de novo tags will get higher scores. Thus, choosing candidates with higher scores first, we give high quality matches priority.

In our method, candidates with different scores compete with each other and we choose the assembled sequence candidate with the highest score first. After choosing the candidate with the highest score, we do not consider those candidates involving either one of the peptide tags which the chosen candidate is generated from anymore.

4.1.6 Contig merging

De novo sequencing software may generate some low quality de novo peptide sequences and sequencing errors occurred in the initial peptide tags will lead to the inaccuracy in our final contigs. This kind of inaccuracy tends to appear at the begin and the end part of the final contigs. And because of this kind of inaccuracy, contigs with long overlap region may not be merged because of the long non-overlapping head or long non-overlapping tail. To increase the length our final contigs, we merge more reliable contigs by relaxing merging conditions.

4.1.7 Algorithms

The overall workflow of our algorithm is shown in Figure 4.7:

- Step 1: Filter out the low quality de novo tags using a chosen threshold. Build a hash table with remaining peptides to find peptide merging pairs.
- Step 2: Merge peptide pairs and evaluate each resulting candidate.
- Step 3: Choose candidate with highest score and update the contig table and the candidate table
- Step 4: Repeat step 2 and 3 until the highest score of the candidate is smaller than a preset threshold

- Step 5: Merge final contigs by relaxing the merging condition

We will next cover the various parts of the algorithm details that are necessary to implement the theory introduced in section 4.1.

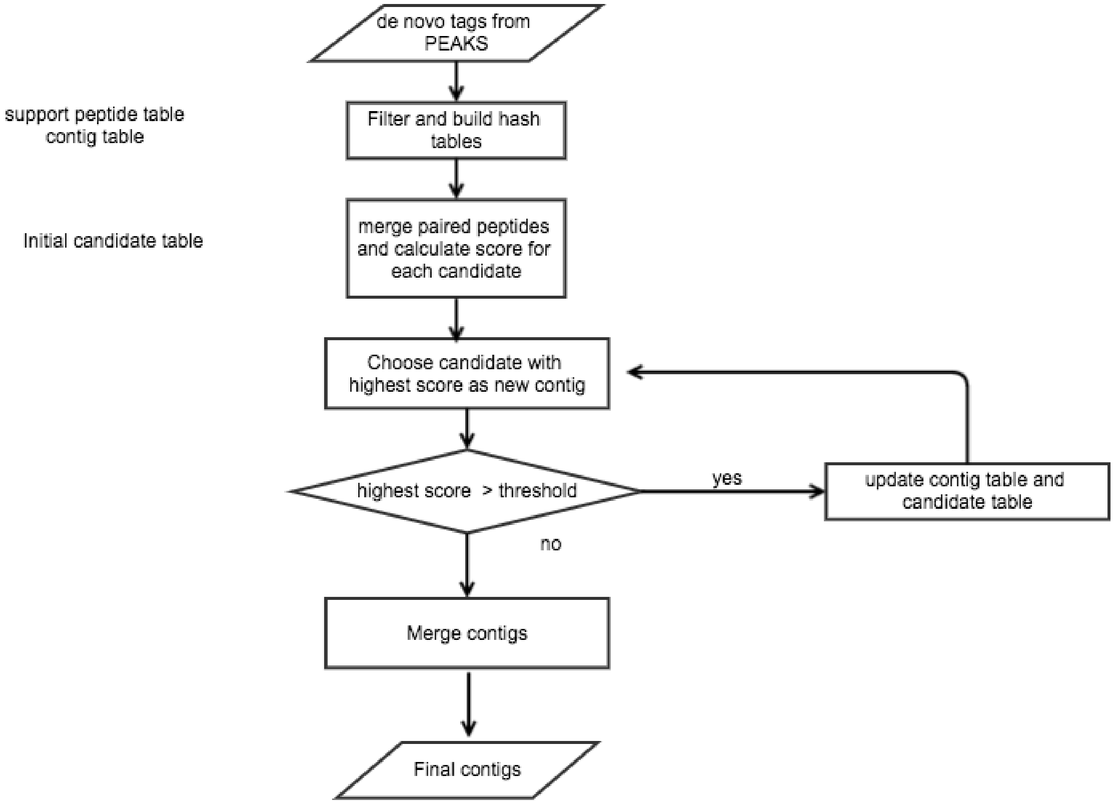

de novo tags from PEAKS
support peptide table
contig table
Filter and build hash tables
Initial candidate table
merge paired peptides and calculate score for each candidate
Choose candidate with highest score as new contig
highest score > threshold
yes
update contig table and candidate table
no
Merge contigs
Final contigs

Figure 4.7: workflow

How to find peptide pairs

We first filter those initial peptides obtained from PEAKS software. We then build a contig hash table and a support peptide hash table using 3-mer. The contig table is used to store the chosen candidates and it is initialized with those initial peptides. The support peptide table stores these initial de novo peptide tags for calculating the support score for each candidate.

Then for a peptide tag a in the contig table, by searching the contig table using each 3-mers in tag a, we can find the possible pairing peptide tag b efficiently. After merging paired peptides a,b, we can use the support peptide table to find support peptides for candidate. The process is illustrated in Figure 4.8.

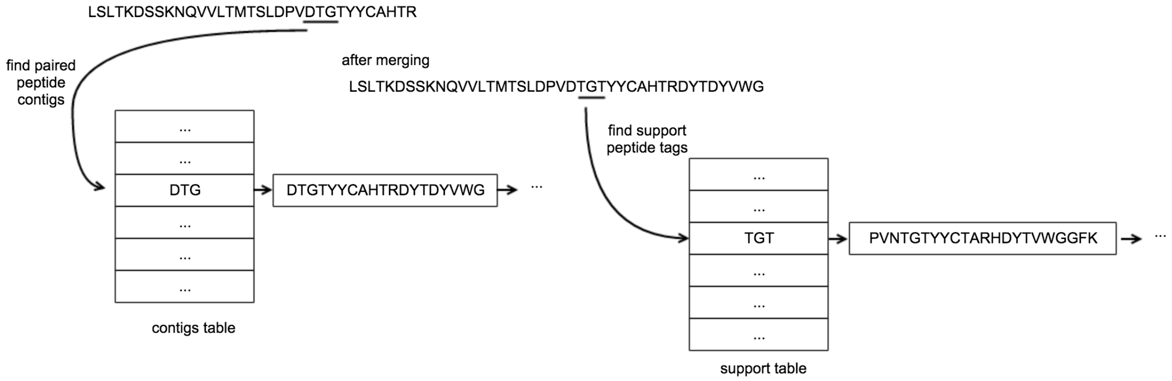

LSLTKDSSKNQVVLTMTSLDPVDTGTYYCAHTR
find paired peptide contigs
...
...
DTG
...
...
...
DTGTYYCAHTRDYTDYVWG
...
contigs table
after merging
LSLTKDSSKNQVVLTMTSLDPVDTGTYYCAHTRDYTDYVWG
find support peptide tags
...
...
TGT
...
...
...
PVNTGTYYCTARHDYTVWGGFK
...
support table

Figure 4.8: Find paired peptides and support peptides

How to merge peptide pairs and calculate the score for each candidate

For peptide tag a, after finding paired peptide tag b, we use the 3-mer as seed, extending to both sides of a and b by comparing the mass of a to the mass of b until we reach the end or the mass difference exceeds the threshold (*e.g.* 1 Dalton in our algorithm). We record the match begin position and the end position for both a and b, as shown in Algorithm 1. This mass match algorithm is required in almost every steps in our method.

Algorithm 1 MassMatch

```
procedure MassMatch
    $pos \leftarrow$ positions of $seed$ in tag b
    for each position $p \in pos$ do
        $mass\_a \leftarrow mass\_of\_seed$
        $mass\_b \leftarrow mass\_of\_seed$
        $pos\_a\_left \leftarrow$ beginning position of $seed$ in tag a
        $pos\_a\_right \leftarrow$ end position of $seed$ in tag a
        $pos\_b\_left \leftarrow p\_beginning$
        $pos\_b\_right \leftarrow p\_end$
        $i\_l \leftarrow pos\_a\_left$
        $i\_r \leftarrow pos\_a\_right$
        $j\_l \leftarrow pos\_b\_left$
        $j\_r \leftarrow pos\_b\_right$
        while $j\_l \geq 0$ and $i\_l \geq 0$ do                              ▷ Left side
            if $abs(mass\_a - mass\_b) < threshold$ then
                $pos\_a\_left \leftarrow i\_l$
                $pos\_b\_left \leftarrow j\_l$
                $i\_l \leftarrow i\_l$ - $1$
                $j\_l \leftarrow j\_l$ - $1$
                $mass\_a = mass\_b$
                if $j\_l \geq 0$ and $i\_l \geq 0$ then
                    $mass\_a \leftarrow mass\_a$ + $mass$ $of$ $a[i\_l]$
                    $mass\_b \leftarrow mass\_b$ + $mass$ $of$ $b[j\_l]$
            else if $mass\_a > mass\_b$ then
                $j\_l \leftarrow j\_l$ - $1$
                if $j\_l \geq 0$ then
                    $mass\_b \leftarrow mass\_b$ + $mass$ $of$ $b[j\_l]$
            else if $mass\_a < mass\_b$ then
                $i\_l \leftarrow i\_l$ - $1$
                if $i\_l \geq 0$ then
                    $mass\_a \leftarrow mass\_a$ + $mass$ $of$ $a[i\_l]$
        $mass\_a \leftarrow$ mass of $seed$    $mass\_b \leftarrow$ mass of $seed$
        while $j\_r \leq a\_length$ and $i\_r \leq b\_length$ do             ▷ Expand to right side
            ...
```

Once we get the overlap region, we can calculate the overlap score between a and b according to formula 4.2. There are different cases when we merge peptide tag a and tag b:

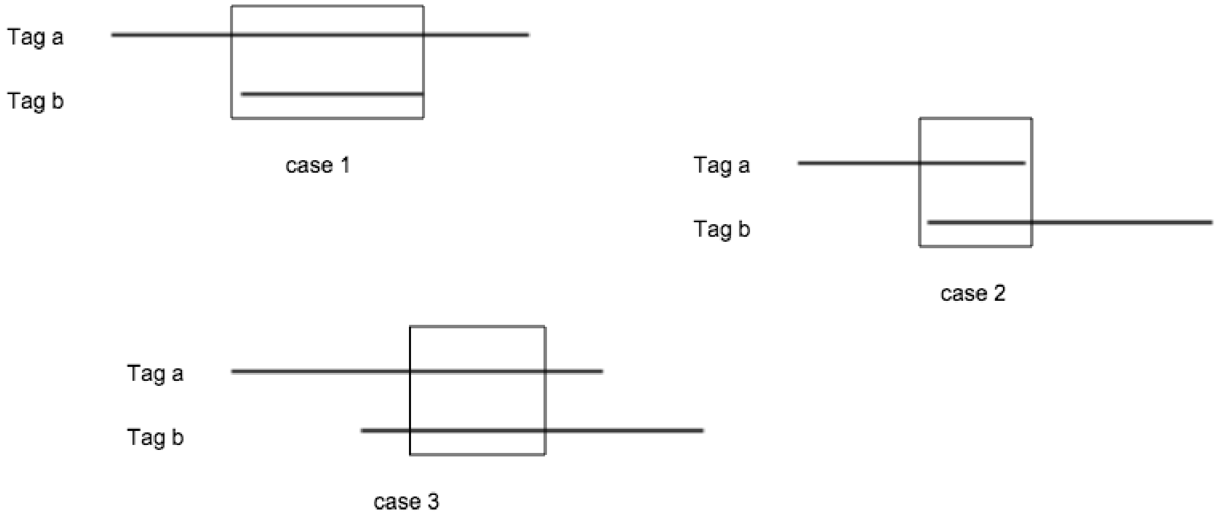


Figure 4.9: Overlap cases

Case 1: Tag a contains tag b

Case 2: The right end of tag a overlaps with the left end of tag b

Case 3: The internal tag a overlaps with the internal tag b

For Case 1, we just replaced the overlap part in tag a with the consensus part between tag a and tag b. Consensus part between a and b is generated as follows:

```
Algorithm 2: Calculate consensus sequence

Input: overlap part from tag a: $overlap\_a$, overlap part from tag b: $overlap\_b$.
Output: Consensus sequence between tag a and b
consensus_seq ← " "

    aa stands for amino acid and mm stands for mass match part
    $aa\_a$ stands for amino acid from tag a and $aa\_b$ stands for amino acid from tag b
    $mm\_a$ stands for mass match part from tag a and $mm\_b$ stands for mass match part from tag b
    $a\_conf$ stands for confidence score of tag a, $b\_conf$ stands for confidence score of tag b, $mm\_conf$ stands for confidence score of mass match part and $aa\_conf$ stands for confidence score of amino acid

Loop through the overlap part between tag a and tag b, still using the mass match method

  • for exact matching aa:
        $consensus\_seq \leftarrow consensus\_seq + \text{aa}$
        $aa\_score \leftarrow max(aa\_a\_conf, aa\_b\_conf)$
  • for mass match part between a and b:
        $\text{mm} \leftarrow a\_conf > b\_conf\ ?\ mm\_a : mm\_b$
        $consensus\_seq \leftarrow consensus\_seq + \text{mm}$
        $mm\_conf \leftarrow a\_conf > b\_conf\ ?\ mm\_a\_conf : mm\_b\_conf$
```

In this way, we get the consensus sequence as well as the confidence score of each amino acid in it. In the actual method, we combine this step with overlap score calculation, that is after calculating overlap score between tag a and tag b, we get the consensus sequence with confidence score as well.

Case 2 is separated into two parts. The consensus sequence calculation step follows the same steps above. Then we need to connect the head of tag a and the tail of tag b with this consensus sequence. Head and tail depend on the longer peptide tag. If tag a has a longer head, then the head of the candidate comes from tag a. Otherwise, the head of the candidate comes from tag b. If the heads of the two tags have the same length, then we choose the head from the more confident tag. Same for the tail.

Case 3 is similar to Case 2, however since de novo sequencing error distributes towards to both end of the sequence, peptide pairs with unmatched head or tail are not reliable. We filtered out the merging candidate if the end position of overlap region is too far away from the end of tag a or the begin position of overlap region is too far away from the start of tag b. To distinguish between Case 2 and Case 3, we give some rewards to merge scenarios in Case 2, as mentioned in section 4.1.4.

For each candidate, we use support peptides table to find possible support peptides and then calculate the support score for each support peptide tag using the mass match algorithm and formula 4.4.

How to update contig and candidate table

By using formula 4.1, each candidate is associated with a score. We then greedily choose the candidate with highest score and update our contig table and candidate table until the highest score is below our threshold. Assuming that the candidate is merged from peptide tag a and peptide tag b, then we remove tag a and tag b from the contig table, and add the new candidate into the contig table. Accordingly, we remove all the candidates that merged from either tag a or tag b from candidate table. We update the candidate table by using the new candidate as peptide tag, merging it with its paired peptides and adding the new results into the candidate table. In this way, we can make sure that peptides that should be merged together with high possibility are chosen first.

How to merge contigs

In our method, we evaluated final sequences by multiplying their length and confidence score, because we expect longer sequence with higher accuracy. Then we choose those sequences with this attribute ranked within top 30 and with confidence score higher than 70. Then we run contig merging on them. The contig merging process is similar to the peptide pairing and merging. The difference is we relax the merge condition in case3 in 4.9. Because the de novo sequencing errors tend to appear at the head and tail part of the contigs, instead of filtering the merging candidate, we consider paired contigs if the match score between them is higher than 10.

Summary

By using the overlap information among de novo peptide tags to merge those pairwise peptides and choose the most promising one to update our dataset, we designed an automated protein sequencing approach. Experimental results are reported in the next section.

4.2 Experiments and Results

4.2.1 Experiment Overview

In this section, we present experiments on a protein dataset. The dataset includes six target proteins. Proteins were digested using several enzymes including Arg-C, Asp-N, CNBr, Glu-C, Lys-C, trypsin and chymotrypsin. Each digest is measured with LC-MS/MS, with three fragmentation modes, CID, HCD and ETD, respectively. Because different enzyme cleave the protein at different sites, the peptides generated from a enzyme may overlap with those from another enzyme. Multiple fragmentation are utilized to produce three MS/MS spectra for each peptide. This helps increase the de novo sequencing accuracy of each peptide. The same dataset has been previously used for developing and demonstrating the performance of the Meta-SPS tool by Guthals et al.

The performance of our method is assessed in terms of de novo sequencing length, coverage, and accuracy. Coverage, length and accuracy are determined by comparing the algorithms's results with the original proteins. Coverage is calculated by counting the percentage of amino acids covered in the reference sequence by de novo sequence contig via error-tolerant alignment. Error-tolerant here means the acceptable mass replacement mentioned in section 4.1.4.

4.2.2 Results on six target proteins

Overview of sequencing coverage of target proteins

The longest de novo sequencing result, average sequence length and de novo coverage are reported in the Table 4.2. The longest de novo sequence is the maximum number of amino acids covered by a single de novo contig. Average sequence length is the average number of amino acids covered by each aligned de novo contig and contig coverage is the percent of amino acids in the protein covered by at least one aligned de novo contig.

Table 4.2: De novo sequencing length and coverage

Protein	leptin	kallikrein	groEL	myoglobin	aprotinin	peroxidase
Protein Length(AA)	167	261	548	154	100	353
Longest de novo Sequence(AA)	97	136	170	96	52	65
DB Search Coverage(%)	87.4	89.7	99.8	99.3	67	64
Contig Coverage(%)	87.4	83.5	94.2	99.3	67	56.9
Average Seq. Length(AA)(%)	74.5	82.3	72.4	52.7	31.5	25

In our evaluation, we pay attention to two aspects:

- Single contig length and accuracy. Here we considered the the longest de novo sequence and its accuracy for each protein.

- Protein coverage and accuracy. For protein coverage and accuracy, we considered all de novo contigs that belong to that protein.

Sequencing coverage comparison

Here we compare our results with Meta-SPS[10], since we are using the data from this paper. Table 4.3 compares our results with theirs in terms of protein coverage. Our method results in better sequence coverage for each of the six proteins. Table 4.4 compares our longest contig length with their longest contig length.

Table 4.3: Multiple de novo contigs coverage comparison

Protein	leptin	kallikrein	groEL	myoglobin	aprotinin	peroxidase
our method(%)	87.4	83.5	94.2	99.3	67	56.9
Meta-SPS (%)	86.2	79.3	80.5	84.4	59	39.9

Table 4.4: Longest de novo contig coverage comparison

Protein	leptin	kallikrein	groEL	myoglobin	aprotinin	peroxidase
our method(AA)	97	136	170	96	52	65
Meta-SPS (AA)	93	134	194	80	59	58

Sequencing accuracy comparison

In de novo peptide sequencing, single amino acid mass replacements are likely to happen, so we label the exact matched amino acids and single amino acid mass replacements in our final de novo sequences as correct. Sequencing accuracy is the percentage of all amino acids that were labeled correct. Meta-SPS[10] only provided their longest de novo sequence, so we are unable to compare our results with theirs in terms of multiple de novo contig accuracy. Here, we report the sequence accuracy of our method for each of the six proteins in table 4.5.

Table 4.5: De novo sequencing accuracy

Protein	leptin	kallikrein	groEL	myoglobin	aprotinin	peroxidase
Sequencing Accuracy(%)	97.2	87.5	89.4	98.5	97.3	80.1

Table 4.6 compares our longest de novo contig accuracy with the longest sequence accuracy from Meta-SPS[10]. In Meta-SPS[10], the reversed amino acids were labeled correct. However, in our calculation, we did not count the reversed amino acids. The reversed amino acids were shown in Figure 4.10.

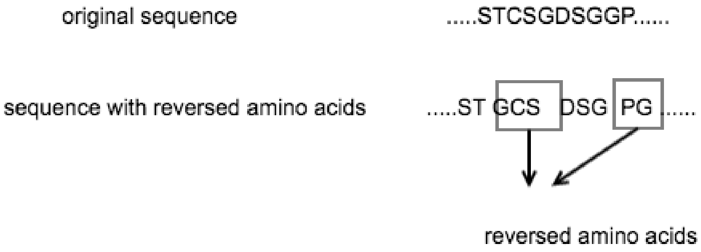


Figure 4.10: Reversed amino acids

Table 4.6: Longest de novo contig accuracy

Protein	leptin	kallikrein	groEL	myoglobin	aprotinin	peroxidase
Longest contig accuracy(%)	100	82.7	95.7	100	96.1	61.5
Meta-SPS (%)	82.1	82.2	93.9	100	75.8	90

Confidence score

As mentioned in the previous section, our method outputs a confidence score for the final sequence and a local confidence score for each amino acid within it. The confidence score of the sequence is the average score of every amino acid. From our observation, de novo sequence with higher confidence score tends to be more accurate. In our results, we set sequence confidence score as x-axis and sequence accuracy as y-axis, as shown in Figure 4.11. Generally, the sequence accuracy becomes higher as the sequence confidence score becomes higher.

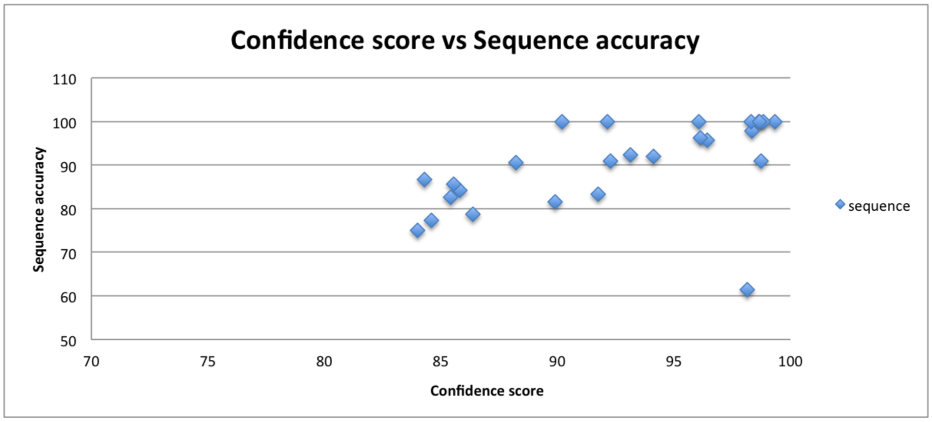
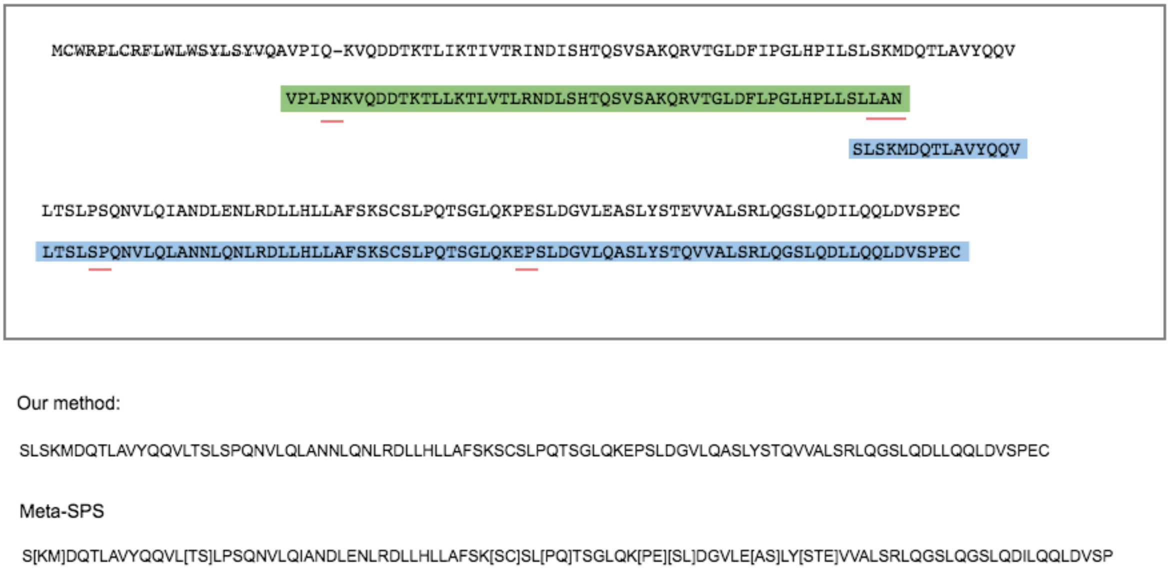


Figure 4.11: Relationship between de novo sequence confidence and de novo sequence accuracy

Final de novo sequencing results of six target proteins

In the following, we present the experimental results. Each colored row corresponds to a de novo sequence, mapped and aligned with the reference protein sequence (without colored background). In the reference sequences, regions covered with dotted lines indicate the lack of coverage by database search. Mass gaps are indicated by dashes in sequences. Missmatches or mass gaps of de novo sequences that expand more than one amino acid are indicated by red underlines. Below each protein map is the longest de novo sequence covering that protein from our results and the results from Meta-SPS[10]. Amino acids within the brackets correspond to reversed animo acids or incorrect mass interpretation.



MCWRPLCRFLWLWSYLSYVQAVPIQ-KVQDDTKTLIKTIVTRINDISHTQSVSAKQRVTGLDFIPGLHPILSLSKMDQTLAVYQQV

VPLPNKVQDDTKTLLKTLVTLRNDLSHTQSVSAKQRVTGLDFLPGLHPLLSLLAN

SLSKMDQTLAVYQQV

LTSLPSQNVLQIANDLENLRDLLHLLAFSKSCSLPQTSGLQKPESLDGVLEASLYSTEVVALSRLQGSLQDILQQLDVSPEC

LTSLSPQNVLQLANNLQNLRDLLHLLAFSKSCSLPQTSGLQKEPSLDGVLQASLYSTQVVALSRLQGSLQDLLQQLDVSPEC

Our method:

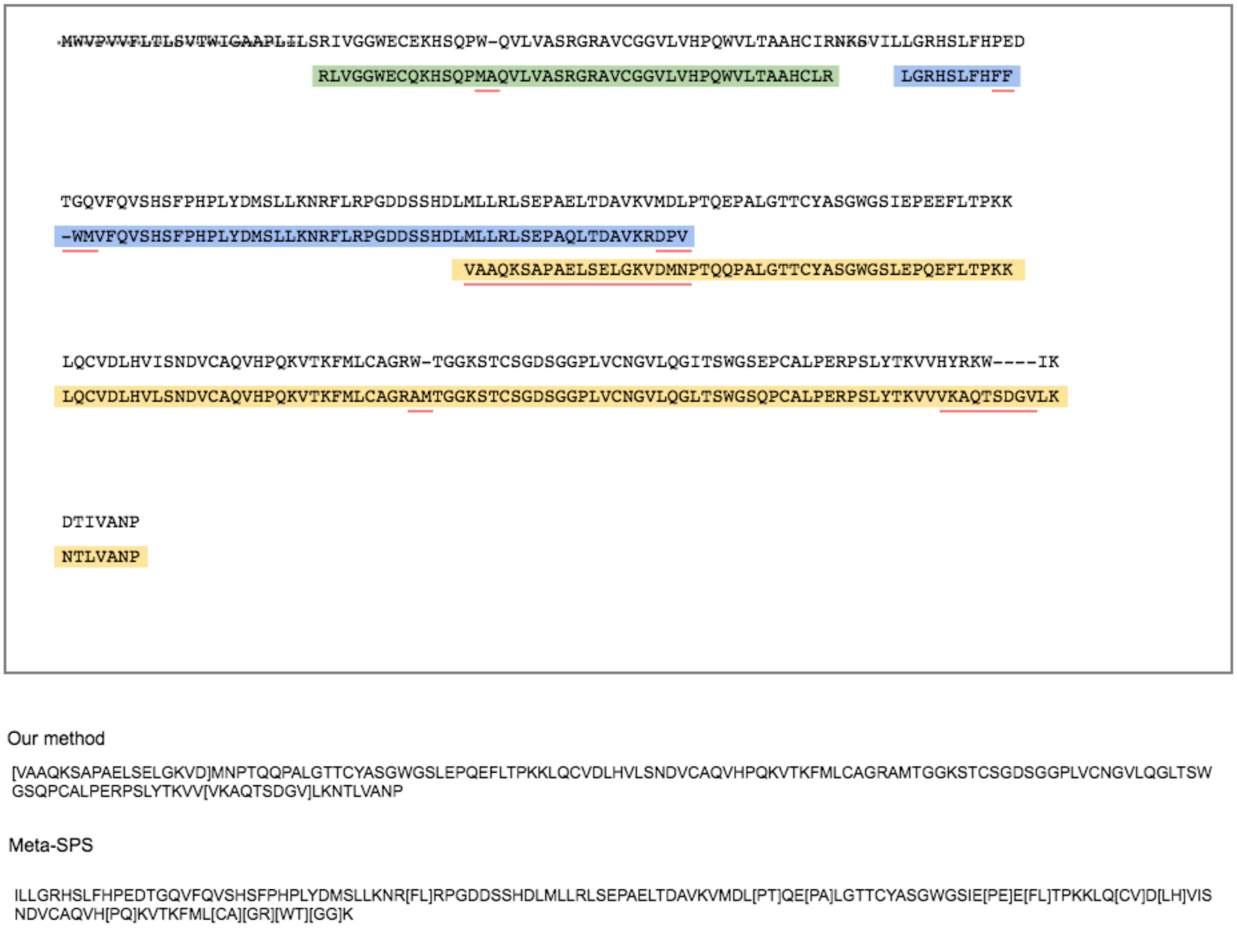
SLSKMDQTLAVYQQVLTSLSPQNVLQLANNLQNLRDLLHLLAFSKSCSLPQTSGLQKEPSLDGVLQASLYSTQVVALSRLQGSLQDLLQQLDVSPEC

Meta-SPS

S[KM]DQTLAVYQQVL[TS]LPSQNVLQIANDLENLRDLLHLLAFSK[SC]SL[PQ]TSGLQK[PE][SL]DGVLE[AS]LY[STE]VVALSRLQGSLQGSLQDILQQLDVSP

Figure 4.12: Sequence coverage of leptin

As shown in Figure 4.12, compared with Meta-SPS, we achieve better sequence coverage and accuracy. The longest sequence generated for leptin is 97 AA at nearly 100% accuracy. Errors in the sequence are mainly caused by the mass replacements happened in the de novo peptide sequencing process.



```
~~MWVPVVFLTLSVTWIGAAPLIL~~SRIVGGWECEKHSQPW-QVLVASRGRAVCGGVLVHPQWVLTAAHCIRNKSVILLGRHSLFHPED
                        RLVGGWECQKHSQPMAQVLVASRGRAVCGGVLVHPQWVLTAAHCLR     LGRHSLFHFF

TGQVFQVSHSFPHPLYDMSLLKNRFLRPGDDSSHDLMLLRLSEPAELTDAVKVMDLPTQEPALGTTCYASGWGSIEPEEFLTPKK
-WMVFQVSHSFPHPLYDMSLLKNRFLRPGDDSSHDLMLLRLSEPAQLTDAVKRDPV
                                     VAAQKSAPAELSELGKVDMNPTQQPALGTTCYASGWGSLEPQEFLTPKK

LQCVDLHVISNDVCAQVHPQKVTKFMLCAGRW-TGGKSTCSGDSGGPLVCNGVLQGITSWGSEPCALPERPSLYTKVVHYRKW----IK
LQCVDLHVLSNDVCAQVHPQKVTKFMLCAGRAMTGGKSTCSGDSGGPLVCNGVLQGLTSWGSQPCALPERPSLYTKVVVKAQTSDGVLK

DTIVANP
NTLVANP
```

Our method

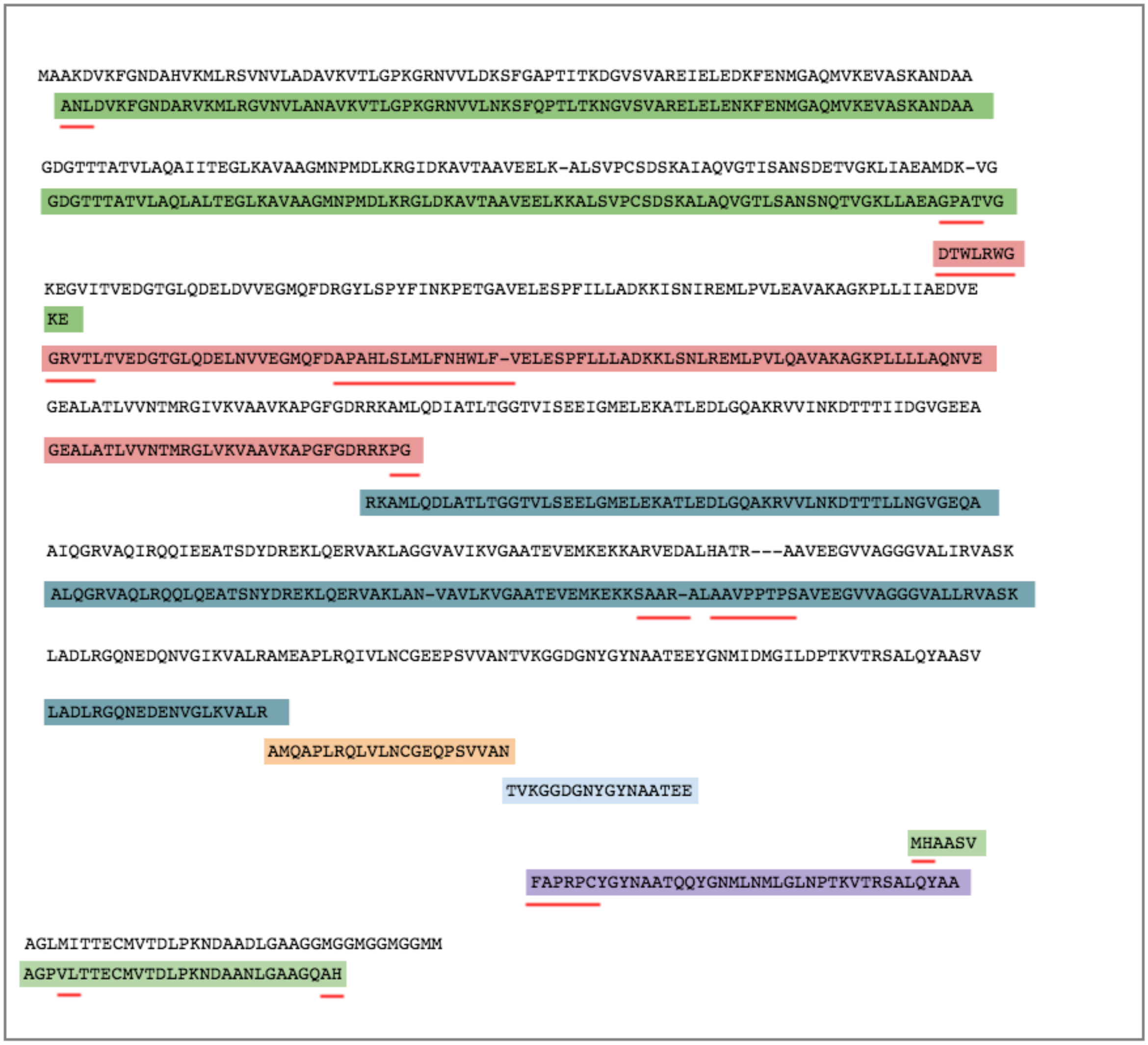
[VAAQKSAPAELSELGKVD]MNPTQQPALGTTCYASGWGSLEPQEFLTPKKLQCVDLHVLSNDVCAQVHPQKVTKFMLCAGRAMTGGKSTCSGDSGGPLVCNGVLQGLTSWGSQPCALPERPSLYTKVV[VKAQTSDGV]LKNTLVANP

Meta-SPS

ILLGRHSLFHPEDTGQVFQVSHSFPHPLYDMSLLKNR[FL]RPGDDSSHDLMLLRLSEPAELTDAVKVMDL[PT]QE[PA]LGTTCYASGWGSIE[PE]E[FL]TPKKLQ[CV]D[LH]VISNDVCAQVH[PQ]KVTKFML[CA][GR][WT][GG]K

Figure 4.13: Sequence coverage of kallikrein related peptidase

As shown in Figure 4.13, the longest sequence generated for kallikrein is 136 AA at 82.7% accuracy. Errors in the sequence are mainly caused by the mass replacement happened in the de novo peptide sequencing process.


MAAKDVKFGNDAHVKMLRSVNVLADAVKVTLGPKGRNVVLDKSFGAPTITKDGVSVAREIELEDKFENMGAQMVKEVASKANDAA
ANLDVKFGNDARVKMLRGVNVLANAVKVTLGPKGRNVVLNKSFQPTLTKNGVSVARELELENKFENMGAQMVKEVASKANDAA
GDGTTTATVLAQAIITEGLKAVAAGMNPMDLKRGIDKAVTAAVEELK-ALSVPCSDSKAIAQVGTISANSDETVGKLIAEAMDK-VG
GDGTTTATVLAQLALTEGLKAVAAGMNPMDLKRGLDKAVTAAVEELKKALSVPCSDSKALAQVGTLSANSNQTVGKLLAEAGPATVG
DTWLRWG
KEGVITVEDGTGLQDELDVVEGMQFDRGYLSPYFINKPETGAVELESPFILLADKKISNIREMLPVLEAVAKAGKPLLIIAEDVE
KE
GRVTLTVEDGTGLQDELNVVEGMQFDAPAHLSLMLFNHWLF-VELESPFLLLADKKLSNLREMLPVLQAVAKAGKPLLLLAQNVE
GEALATLVVNTMRGIVKVAAVKAPGFGDRRKAMLQDIATLTGGTVISEEIGMELEKATLEDLGQAKRVVINKDTTTIIDGVGEEA
GEALATLVVNTMRGLVKVAAVKAPGFGDRRKPG
RKAMLQDLATLTGGTVLSEELGMELEKATLEDLGQAKRVVLNKDTTTLLNGVGEQA
AIQGRVAQIRQQIEEATSDYDREKLQERVAKLAGGVAVIKVGAATEVEMKEKKARVEDALHATR---AAVEEGVVAGGGVALIRVASK
ALQGRVAQLRQQLQEATSNYDREKLQERVAKLAN-VAVLKVGAATEVEMKEKKSAAR-ALAAVPPTPSAVEEGVVAGGGVALLRVASK
LADLRGQNEDQNVGIKVALRAMEAPLRQIVLNCGEEPSVVANTVKGGDGNYGYNAATEEYGNMIDMGILDPTKVTRSALQYAASV
LADLRGQNEDENVGLKVALR
AMQAPLRQLVLNCGEQPSVVAN
TVKGGDGNYGYNAATEE
MHAASV
FAPRPCYGYNAATQQYGNMLNMLGLNPTKVTRSALQYAA
AGLMITTECMVTDLPKNDAADLGAAGGMGGMGGMGGMM
AGPVLTTECMVTDLPKNDAANLGAAGQAH

Our method

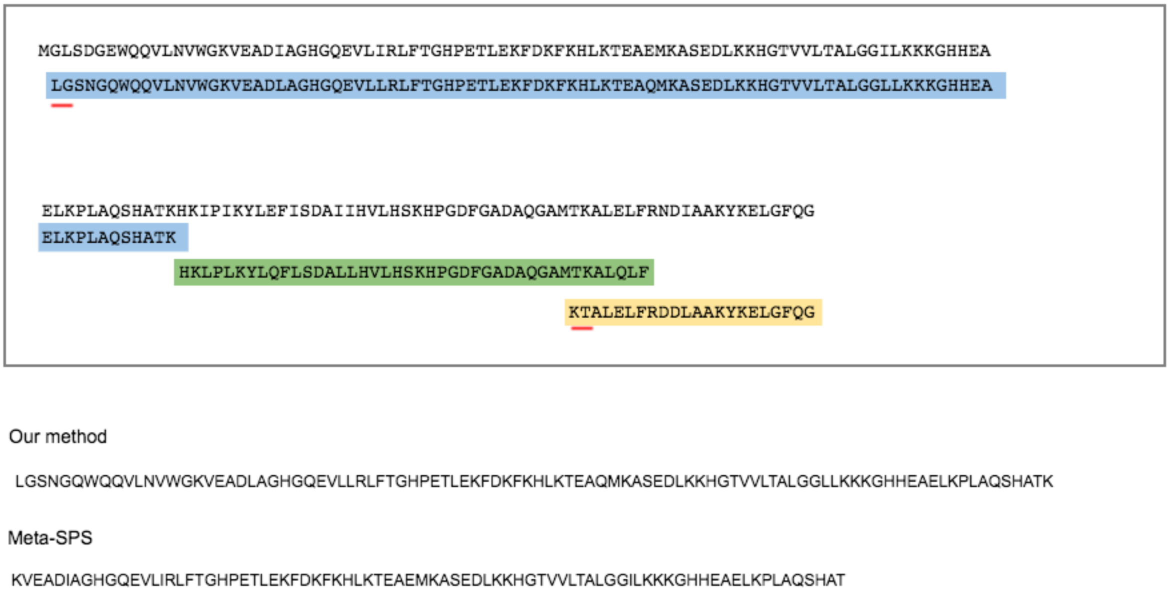
[ANL]DVKFGNDARVKMLRGVNVLANAVKVTLGPKGRNVVLNKSFQPTLTKNGVSVARELELENKFENMGAQMVKEVASKANDAAGDGTTTATVLA
QLALTEGLKAVAAGMNPMDLKRGLDKAVTAAVEELKKALSVPCSDSKALAQVGTLSANSNQTVGKLLAEA[GPATVG]KE

Meta-SPS

AKDVKFGNDA[H,19]VKMLR[SV,-
30]NVLADAVKVTLGPKGRNVVL[DK]SFGAPTITKDGVSVAREIELEDKFENMGAQMVKEVASKANDAAGDGTTTATV[LA]QAIITEGLKAVAAGMNPM
DLKR[GI]DKAVTAAVEELKALSVPCSDSKAIAQVGTISANSDETVGKLIAEAMDKVGKEGVITVED[GTG]LQDELDVVEGMQFD

Figure 4.14: Sequence coverage of groEL

As shown in Figure 4.14, the longest sequence generated for groEL is 170 AA at 95.7% accuracy. Our method generated three long-length sequences for groEL. Errors in the middle part of the sequence are introduced by merging contigs with unmatched heads or tails.



```
MGLSDGEWQQVLNVWGKVEADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASEDLKKHGTVVLTALGGILKKKGHHEA
 LGSNGQWQQVLNVWGKVEADLAGHGQEVLLRLFTGHPETLEKFDKFKHLKTEAQMKASEDLKKHGTVVLTALGGLLKKKGHHEA

ELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKHPGDFGADAQGAMTKALELFRNDIAAKYKELGFQG
ELKPLAQSHATK
            HKLPLKYLQFLSDALLHVLHSKHPGDFGADAQGAMTKALQLF
                                             KTALELFRDDLAAKYKELGFQG
```

Our method

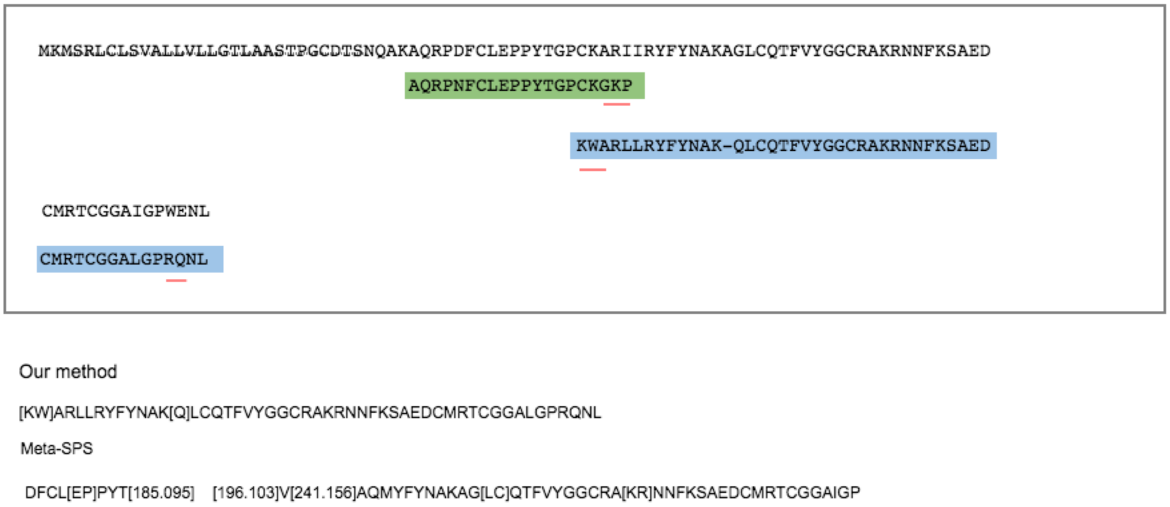
LGSNGQWQQVLNVWGKVEADLAGHGQEVLLRLFTGHPETLEKFDKFKHLKTEAQMKASEDLKKHGTVVLTALGGLLKKKGHHEAELKPLAQSHATK

Meta-SPS

KVEADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASEDLKKHGTVVLTALGGILKKKGHHEAELKPLAQSHAT

Figure 4.15: Sequence coverage of myoglobin

As shown in Figure 4.15, compared with Meta-SPS, we achieve better sequence coverage and accuracy. The longest sequence generated for myoglobin is 96 AA at nearly 100% accuracy. Errors in the sequence are mainly caused by the amino acid reversals.



```
MKMSRLCLSVALLVLLGTLAASTPGCDTSNQAKAQRPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAED
                                        AQRPNFCLEPPYTGPCKGKP

                                                              KWARLLRYFYNAK-QLCQTFVYGGCRAKRNNFKSAED

CMRTCGGAIGPWENL

CMRTCGGALGPRQNL
```

Our method

[KW]ARLLRYFYNAK[Q]LCQTFVYGGCRAKRNNFKSAEDCMRTCGGALGPRQNL

Meta-SPS

DFCL[EP]PYT[185.095]   [196.103]V[241.156]AQMYFYNAKAG[LC]QTFVYGGCRA[KR]NNFKSAEDCMRTCGGAIGP

Figure 4.16: Sequence coverage of aprotinin

As shown in Figure 4.16, compared with Meta-SPS, we achieve better sequence coverage and accuracy. The longest sequence generated for aprotinin is 52 AA at 96.1% accuracy. The sequence generated by Meta-SPS contains many mass gaps.

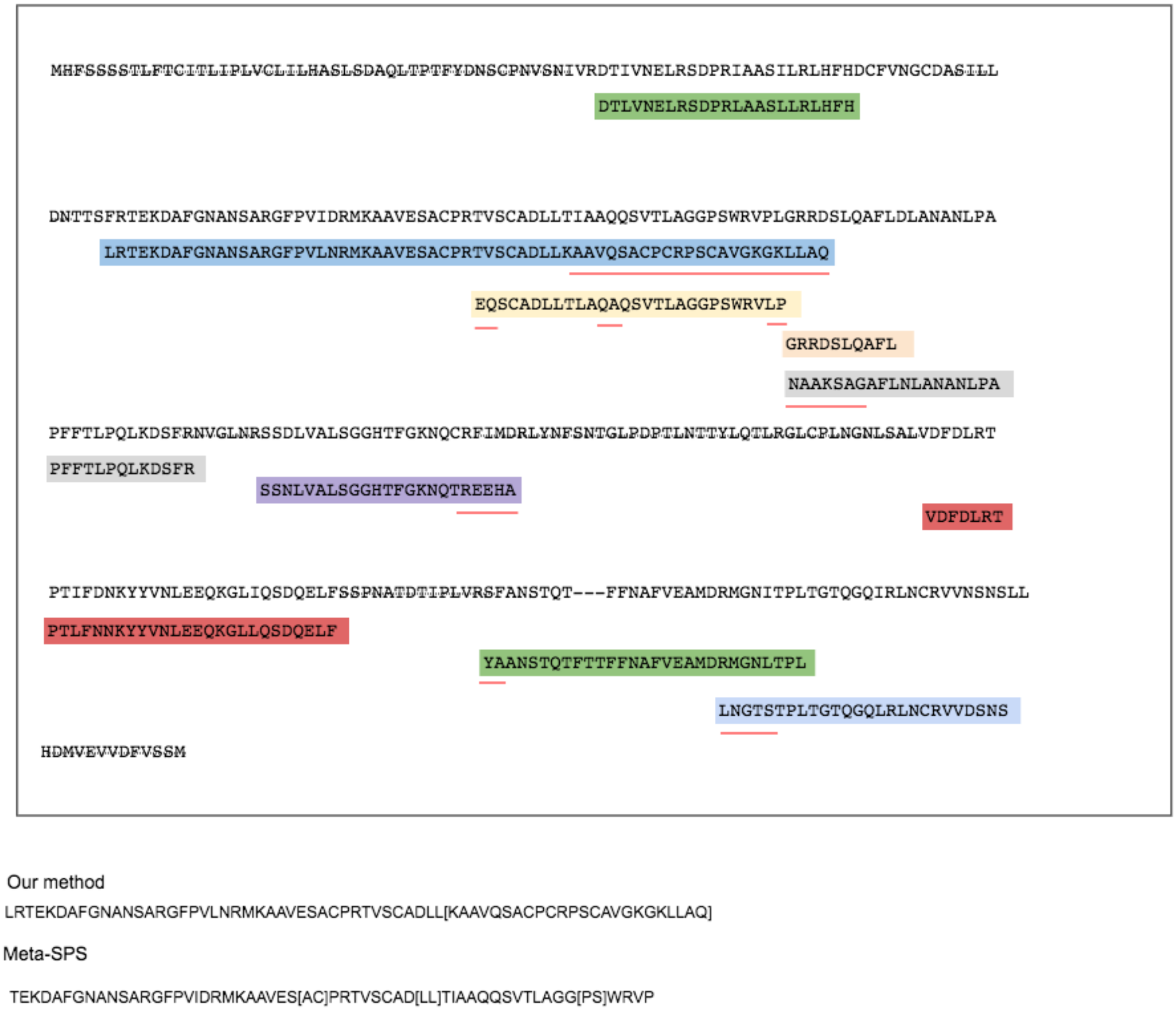

MHFSSSSTLFTCITLIPLVCLILHASLSDAQLTPTFYDNSCPNVSNIVRDTIVNELRSDPRIAASILRLHFHDCFVNGCDASILL
DTLVNELRSDPRLAASLLRLHFH
DNTTSFRTEKDAFGNANSARGFPVIDRMKAAVESACPRTVSCADLLTIAAQQSVTLAGGPSWRVPLGRRDSLQAFLDLANANLPA
LRTEKDAFGNANSARGFPVLNRMKAAVESACPRTVSCADLLKAAVQSACPCRPSCAVGKGKLLAQ
EQSCADLLTLAQAQSVTLAGGPSWRVLP
GRRDSLQAFL
NAAKSAGAFLNLANANLPA
PFFTLPQLKDSFRNVGLNRSSDLVALSGGHTFGKNQCRFIMDRLYNFSNTGLPDPTLNTTYLQTLRGLCPLNGNLSALVDFDLRT
PFFTLPQLKDSFR
SSNLVALSGGHTFGKNQTREEHA
VDFDLRT
PTIFDNKYYVNLEEQKGLIQSDQELFSSPNATDTIPLVRSFANSTQT---FFNAFVEAMDRMGNITPLTGTQGQIRLNCRVVNSNSLL
PTLFNNKYYVNLEEQKGLLQSDQELF
YAANSTQTFTTFFNAFVEAMDRMGNLTPL
LNGTSTPLTGTQGQLRLNCRVVDSNS
HDMVEVVDFVSSM
Our method
LRTEKDAFGNANSARGFPVLNRMKAAVESACPRTVSCADLL[KAAVQSACPCRPSCAVGKGKLLAQ]
Meta-SPS
TEKDAFGNANSARGFPVIDRMKAAVES[AC]PRTVSCAD[LL]TIAAQQSVTLAGG[PS]WRVP

Figure 4.17: Sequence coverage of peroxidase

As shown in Figure 4.17, the longest sequence generated for peroxidase is 65 AA at only 61.5% accuracy. The reason is that, in de novo sequencing, the whole sequence reversal sometimes happens. The two halves of the peptide are reversed. Our method is not perfect for this situation.

4.3 Discussion

Overlap information from PEAKS de novo peptides enable long-length automated de novo sequencing of protein mixtures at high accuracy without reference. Minor contaminants would not heavily affect our results. In the experiment, the six target proteins are actually mixed together, along with some common contaminants in the laboratory environment, such as human keratin, Lys-C, trypsin precursor and so on. When it comes to a large dataset, to achieve a better result, we could filter some contaminants by using a contaminant library. Although this approach is still unable to reconstruct a complete protein, the sequence length approached 170 AA at the maximum. While related methods for de novo protein sequencing are either reference-based or spectrum-based, our method explores a new research direction in the de novo protein sequencing field by assembling de novo peptide tags. Free from being troubled by frequent upgrades in the mass spectrometry technology, our methods are able to generate results comparable to the results from spectrum-based method, such as Meta-SPS. For reference-based method, generally speaking, they can achieve better accuracy, since they have the reference sequence to guide the whole process. However, reference-based is still limited when we meet a real unknown sequence. For our method, results could be possibly improved from the following aspects:

- High quality input

  The majority of experimental mass spectra are of poor quality. They are not good enough to be interpreted by de novo methods. For de novo peptide sequencing, only around 50% of the input mass spectra can be identified. Even for the most advanced de novo sequencing software, the error rate is around 40% to 50%. Multi spectrum acquisition of high resolution CID, HCD, and ETD gives us a relatively better data quality and PEAKS 7.0 software provides us with de novo peptide tags with relatively higher quality. On top of that, we have tried very hard to avoid error accumulation in our method by considering the confidence score, constructing the consensus part and evaluating the merging candidate. We believe that with the development of mass spectrometry and de novo sequencing technologies, the quality of mass spectra and the accuracy of de novo peptide sequencing will be improved. With these improvements, our method can achieve better assembly results.

- Peptide fragmentation solutions

  Our method is based on the overlap between de novo peptides. A good sample digestion that create more peptide overlaps is crucial to our method. With more overlap information, our method can obtain much longer and more accurate sequences. To increase the overlap information, the experimental data is digested using several enzymes including Arg-C, Asp-N, CNBr, Glu-C, Lys-C, trypsin and chymotrypsin. Still, increasing the overlap information by trying different enzyme combinations is a worthy problem to be explored.

- Post-translational modifications

  The way we deal with post-translational modifications (PTM) is simple. After the de novo sequencing step, we just keep the plain amino acid sequence for analysis. However, if the PTM can be correctly identified by the de novo sequencing step, we can use the de novo tag directly and assemble them more accurately.

- Advanced score function

  The experimental dataset is not easy to get and the lack of large dataset for training limited the use of an advanced scoring function. We have tried very best to avoid overfitting by putting robustness as a main consideration in developing the scoring function. For example, in our experiment, we observed that the signal of most low accuracy de novo peptide tags, which is because of ambiguous interpretations of MS/MS fragmentation, tends to be low in the LC-MS (Liquid chromatography-mass spectra). However, in de novo peptide sequencing, the low accuracy de novo peptide tags may still have high confidence score. We tried to simply include the intensity feature in our score function, however, We did not observe significant improvement in our results. We tried to tune the coefficient of this feature slightly, but the gain is marginal. To avoid overfitting, we just discarded this feature.

Chapter 5

Conclusion

5.1 Conclusion

Knowing the sequence of proteins is of great importance because of the many important functions that proteins may have. The specific contributions and conclusions of this study can be summarized as follows:

We propose a novel protein de novo sequencing approach to generate de novo sequences. The input of our method is the mixed protein samples. The key idea is to use the overlap information among de novo peptide tags to merge those pairwise peptides and choose the most promising one to update our dataset. We designed an effective score function to evaluate different merging scheme. Experimental results show that our approach can yield de novo sequences up to 170 AA at around 96% sequencing accuracy.

5.2 Future work

The objective of this thesis is to design a protein de novo sequencing method to realize de novo sequencing of unknown proteins without the support from reference sequences. There remain some problems in our research:

In our work, we design a score function to evaluate the candidates, which considers the overlap information, support information, confidence scores and cleavage sites. However, due to the lack of large data set for training, we only included the limited features to avoid overfitting. The score function can be further improved once we have enough training data.

APPENDICES

Appendix A

List of Software and Hardware Used

This section lists the software programs and hardware used. Software List

- PEAKS Studio 7.0
  - de novo sequencing and database search
- `C++`

Hardware List

- Personal Computer: OS X Yosemite Intel(R) Core(TM) i5 CPU @ 2.60 GHz

References

[1] Jacques U Baenziger. A major step on the road to understanding a unique posttranslational modification and its role in a genetic disease. *Cell*, 113(4):421–422, 2003.

[2] Nuno Bandeira, Victoria Pham, Pavel Pevzner, David Arnott, and Jennie R Lill. Automated de novo protein sequencing of monoclonal antibodies. *Nature biotechnology*, 26(12):1336–1338, 2008.

[3] Nuno Bandeira, Haixu Tang, Vineet Bafna, and Pavel Pevzner. Shotgun protein sequencing by tandem mass spectra assembly. *Analytical chemistry*, 76(24):7221–7233, 2004.

[4] Hao Chi, Rui-Xiang Sun, Bing Yang, Chun-Qing Song, Le-Heng Wang, Chao Liu, Yan Fu, Zuo-Fei Yuan, Hai-Peng Wang, Si-Min He, et al. pnovo: de novo peptide sequencing and identification using hcd spectra. *Journal of proteome research*, 9(5):2713–2724, 2010.

[5] John B Fenn, Matthias Mann, Chin Kai Meng, Shek Fu Wong, and Craig M Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, 1989.

[6] Jorge Fernández-de Cossio, Javier Gonzalez, and Vladimir Besada. A computer program to aid the sequencing of peptides in collision-activated decomposition experiments. *Computer applications in the biosciences: CABIOS*, 11(4):427–434, 1995.

[7] Ari Frank and Pavel Pevzner. Pepnovo: de novo peptide sequencing via probabilistic network modeling. *Analytical chemistry*, 77(4):964–973, 2005.

[8] Christian K Frese, AF Maarten Altelaar, Marco L Hennrich, Dirk Nolting, Martin Zeller, Jens Griep-Raming, Albert JR Heck, and Shabaz Mohammed. Improved peptide identification by targeted fragmentation using cid, hcd and etd on an ltq-orbitrap velos. *Journal of proteome research*, 10(5):2377–2388, 2011.

[9] V Gold, KL Loening, AD McNaught, and P Shemi. Iupac compendium of chemical terminology. *Blackwell Science, Oxford*, 1997.

[10] Adrian Guthals, Karl R Clauser, Ari M Frank, and Nuno Bandeira. Sequencing-grade de novo analysis of ms/ms triplets (cid/hcd/etd) from overlapping peptides. *Journal of proteome research*, 12(6):2846–2857, 2013.

[11] Yonghua Han, Bin Ma, and Kaizhong Zhang. Spider: software for protein identification from sequence tags with de novo sequencing error. *Journal of bioinformatics and computational biology*, 3(03):697–716, 2005.

[12] Teng-Yi Huang and Scott A McLuckey. Gas-phase chemistry of multiply charged bioions in analytical mass spectrometry. *Annual review of analytical chemistry (Palo Alto, Calif.)*, 3:365, 2010.

[13] Richard S Johnson and Klaus Biemann. The primary structure of thioredoxin from chromatium vinosum determined by high-performance tandem mass spectrometry. *Biochemistry*, 26(5):1209–1214, 1987.

[14] Michael Karas, Doris Bachmann, U el Bahr, and F Hillenkamp. Matrix-assisted ultraviolet laser desorption of non-volatile compounds. *International journal of mass spectrometry and ion processes*, 78:53–68, 1987.

[15] Richard J Lewis and Maria L Garcia. Therapeutic potential of venom peptides. *Nature Reviews Drug Discovery*, 2(10):790–802, 2003.

[16] Xiaowen Liu, Lennard JM Dekker, Si Wu, Martijn M Vanduijn, Theo M Luider, Nikola Tolić, Qiang Kou, Mikhail Dvorkin, Sonya Alexandrova, Kira Vyatkina, et al. De novo protein sequencing by combining top-down and bottom-up tandem mass spectra. *Journal of proteome research*, 13(7):3241–3248, 2014.

[17] Xiaowen Liu, Yonghua Han, Denis Yuen, and Bin Ma. Automated protein (re) sequencing with ms/ms and a homologous database yields almost full coverage and accuracy. *Bioinformatics*, 25(17):2174–2180, 2009.

[18] Bin Ma. Challenges in computational analysis of mass spectrometry data for proteomics. *Journal of Computer Science and Technology*, 25(1):107–123, 2010.

[19] Bin Ma. Computational proteomics. Course material for advanced topic in bioinformatics, 2014.

[20] 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.

[21] Lingdong Quan and Miao Liu. Cid, etd and hcd fragmentation to study protein post-translational modifications. *Modern Chemistry &amp; Applications*, 2013, 2013.

[22] Kanti Rai and Michael Hallek. Future prospects for alemtuzumab (mabcampath$^{\text{TM}}$). *Medical Oncology*, 19(2):S57–S63, 2002.

[23] J Alex Taylor and Richard S Johnson. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry*, 11(9):1067–1075, 1997.

[24] J Alex Taylor and Richard S Johnson. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Analytical chemistry*, 73(11):2594–2604, 2001.

[25] thermofisher. Save on hram orbitrap ms for quantitation and more. `https://www.thermofisher.com/ca/en/home/products-and-services/promotions/industrial/save-hram-orbitrap-ms.html`, 2016.

[26] Wikipedia. Mass spectrometry — Wikipedia, The Free Encyclopedia. `https://en.wikipedia.org/w/index.php?title=Mass_spectrometry&oldid=712212233`, 2016.

[27] Eric S Witze, William M Old, Katheryn A Resing, and Natalie G Ahn. Mapping protein post-translational modifications with mass spectrometry. *Nature methods*, 4(10):798–806, 2007.

[28] Changjiang Xu and Bin Ma. Software for computational peptide identification from ms–ms data. *Drug Discovery Today*, 11(13):595–600, 2006.

[29] 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 &amp; Cellular Proteomics*, 11(4):M111–010587, 2012.