Ma B. Challenges in computational analysis of mass spectrometry data for proteomics. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 25(1): 1– Jan. 2010

Challenges in Computational Analysis of Mass Spectrometry Data for Proteomics

Bin Ma (马 斌)

Cheriton School of Computer Science, University of Waterloo, Canada
Dingsheng Technologies, Beijing 100085, China

E-mail: binma@cs.uwaterloo.ca

Received September 9, 2009; revised November 21, 2009.

Abstract Mass spectrometry is an analytical technique for determining the composition of a sample. Recently it has become a primary tool for protein identification and quantification, and post translational modification characterization in proteomics research. Both the size and the complexity of the data produced by this experimental technique impose great computational challenges in the data analysis. This article reviews some of these challenges and serves as an entry point for those who want to study the area in general.

Keywords mass spectrometry, proteomics, bioinformatics

1 Introduction

Proteins play the most important role in disease pathways. Modern pharmaceutical researches heavily rely on the identification, quantification and characterization of the proteins in a given sample. Mass spectrometry is an analytical technique that reveals the composition information of a sample by measuring the mass values (in fact, mass to charge ratio) of the molecules in it. Nowadays, mass spectrometry has become the standard technique for protein identification, quantification and characterization in proteomics. Multiple international mass spectrometry instrument vendors exist, and new types of mass spectrometers appear on the market every couple of years. The mass spectrometry hardware is much more advanced than it was a decade ago. On one hand, the new instruments produce more accurate data than before, which supposedly make the data analysis easier. On the other hand, they have much higher throughput and produce much larger data size; and newly invented experimental methods require new data analysis algorithms. These all impose challenges to the data analyses.

Perhaps the largest challenge comes from the new demands periodically raised by the proteomics researchers. In earlier days, researchers used to use mass spectrometry to identify a single purified protein. Today, a single 2D-LC MS/MS experiment is used to identify all proteins in the whole proteome of an organism[1]. On top of that, researchers would also like to know the quantities of the proteins[2]. Even if the research focuses on a few purified proteins, earlier researchers were satisfied by knowing which proteins they are, whereas today's researchers want the complete protein sequences[3], including all the post-translational modifications (PTM)[4-5]. These new demands require new developments in both experimental and computational methods, and constantly provide fruitful research problems to bioinformatics researchers.

Depending on the purposes, there are different wet lab experimental settings and data analysis methods. This article reviews some of these methods and the related computational problems. Throughout this article, we try to highlight the computational challenges by listing them as C1∼C36. This list is not meant to be complete but provides some interesting research problems for those who just started working in this area. Unlike some other areas in bioinformatics and computational biology, the mathematical models for these problems are often not well defined. Researchers in this area tend to model these biological problems in their own ways, and a right model to the problem is equally important to a good algorithm that finds the solution under the model. For this reason, we deliberately avoid giving a definite mathematical model for any challenge listed in the article. Rather, the right mathematical model should be regarded as a part of the challenge. For challenges listed in the paper, we also try to give a few

Survey

This work is supported by the National High-Tech Research and Development 863 Program of China under Grant No. 2008AA02Z313, NSERC RGPIN under Grant No. 238748-2006, and a start up grant at University of Waterloo.

references. These references are selected to be representative rather than comprehensive. They serve as good starting point if a reader is interested in reading more about a particular problem.

The rest of the article is organized as follows. In Section 2 we briefly introduce the mass spectrometers and their limitations. This should help the readers to understand the subtle difficulties in the data analysis, and particularly the errors in the data. In Section 3 we review several applications of mass spectrometry in proteomics and their computational challenges. We note that some challenges reviewed in an earlier application may also occur in a latter application. Thus, readers will see significantly more challenges listed in the first application than others. In Section 4 we study a few research problems that are commonly needed in several different applications introduced in Section 3. Improving the performance of any of these research problems will help multiple applications. Section 5 concludes the paper.

2 Mass Spectrometry Instruments

This section gives a brief introduction to the mass spectrometry instruments. The introduction focuses on the limitation of the technology, the diversity of the instruments, and the errors in the data. Readers will find that these factors are real concerns in the design of an experiment and the development of the data analysis method. For a more thorough introduction on the mass spectrometry technology, readers are referred to textbooks such as [6].

2.1 Mass Spectrometers

A mass spectrometer does not measure the mass of a molecule directly. Rather, the molecules are ionized and the mass to charge (m/z) ratios of the ions (charged molecules) are measured. But very often the charge states of an ion can be determined by examining the isotope ions, and thus the mass value of an ion can be derived from m/z. A mass spectrometer typically contains three components: the ionizer, the mass analyzer, and the detector. A bunch of molecules are first ionized with the ionizer; then the ions are separated in the mass analyzer according to their m/z; finally the ions are detected by the detector and the m/z of the detected ions are calculated and stored in a computer. Each type of ions with the same m/z will form a *peak* in the resulting data (called a *mass spectrum*). Fig.1(a) shows an example. The intensity of a peak indicates the ion counts detected by the detector at the m/z, which is related to the abundance of the corresponding type of molecules in the original sample. However, because different molecules have different ionization efficiencies, the abundances of two different molecules cannot be compared solely by their peak intensities.

The primary function of a mass spectrometer is to measure the m/z and intensities of many ions simultaneously. This very basic function has been exploited by instrument developers and biochemistry researchers to perform many different tasks in proteomics. Before reviewing the computational challenges in this field, it is necessary to first examine a few limitations of mass

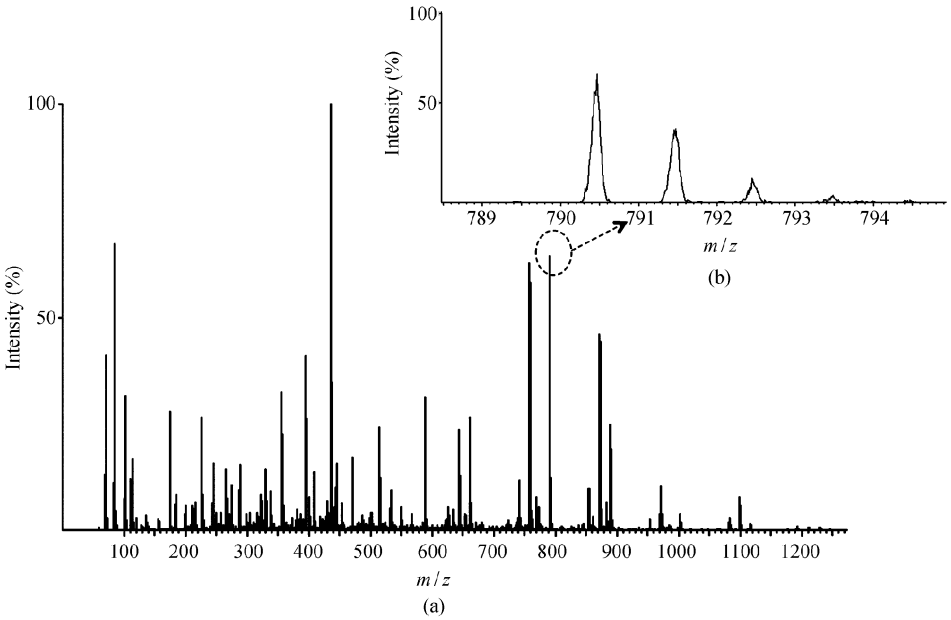


Fig.1. (a) Exemplary mass spectrum. (b) Zooming in a peak shows more details. In particular, each peak spans a width on the m/z direction.

spectrometers. As we will see later, these limitations heavily affect the experiment design and drastically increase the complexity of the data analysis.

The first limitation is that the m/z values measured by a mass spectrometer have small errors. Due to the measurement variation, each peak spans a width in the m/z direction (Fig.1(b)). The width of the peak affects the *resolution* of the instrument, since two adjacent peaks may become indistinguishable when they overlap each other. Before any data analysis is done, a process of "*centroiding*" is often needed to assign each peak a single m/z value, which is usually the centroid of the peak shape. Many publicly available data are already centroided so they often do not show the peak shape as seen in Fig.1(b). Centroiding is not a trivial process because of the possible overlaps of adjacent peaks. The m/z of the centroided peak may still have a small error compared to the real m/z of the ion. Different mass spectrometers have different *mass error tolerances*. The mass error tolerance is one of the most important parameters in the data analysis.

Secondly, each instrument setting has a limited detection range of m/z values. Only ions that fall in this range can be measured. For example, a typical ion trap instrument has a significant low m/z cut-off. Ions below the cut-off m/z value will not form peaks in the spectrum. But a time-of-flight (TOF) instrument does not have this significant low mass cut-off. Some mass spectrometers can be configured to have a rather large m/z range with the price of reduced resolution and mass accuracy. The m/z range used for proteomics analysis is typically from below 100 Da to a few thousand Da. This mass range enables an optimized combination of sensitivity, resolution and accuracy. Because of this limitation, a protein is not usually analyzed directly due to its large size. Instead, it is enzymatically digested into peptides that fall into this preferred m/z ranges. Then each peptide is measured and analyzed separately before their information is put together to characterize the protein. This approach is called the *bottom-up* analysis.

The third limitation is that not all molecules in the sample are measured with the same efficiency. Due to the reasons, such as *charge competition*[7], some molecules may produce much lower intensity peaks than other molecules with the same abundance in the sample. In many cases, the peaks of some molecules may become indistinguishable from noise peaks in the spectrum, causing the absence of the expected signal peaks. The missing of peaks greatly increases the complexity of data analysis and is the largest obstacle in the development of data analysis algorithms.

2.2 Tandem Mass Spectrometers

A tandem mass spectrometer has two mass analyzers (or two sequential analyses in the same analyzer). The first analyzer selects ions at a certain m/z window (usually a very small window so that only copies of the same ion are selected). This is called the *precursor ion* or the *parent ion*. Then the ion is fragmented into *fragment ions* by some fragmentation methods. And finally the fragment ions are measured as usual to form a *tandem mass spectrum* (or *MS/MS spectrum*). Fig.2(a) illustrates the possible fragmentation sites of a peptide. For example, when the fragmentation occurs at the peptide bond (between C and N atoms), the left component forms a *b*-ion and the right component forms a *y*-ion. The subscript k of the y_k ion indicates the number of residues in the fragment. Fig.2(b) shows an annotated MS/MS spectrum. Because amino acid residues have different mass values (except for Leucine and Isoleucine), the m/z values of the fragment ions can be used to identify peptides.

In proteomics, the tandem mass spectrometry analysis provides much more information about a peptide than the mass spectrometry analysis. Thus, most of today's proteomics analyses are done with MS/MS. And most new mass spectrometers on the market support

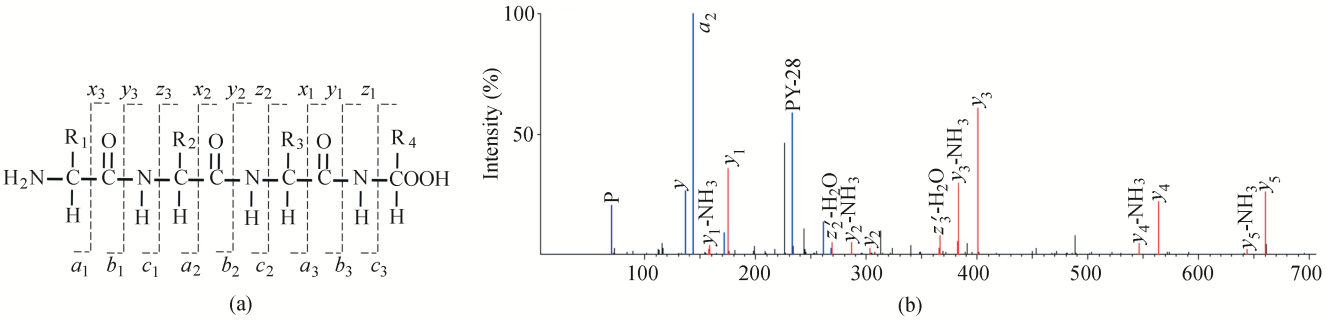


Fig.2. (a) Fragmentation of a four-residue peptide in MS/MS. The fragmentation can happen at each bond on the peptide backbone, resulting in different fragment ion types. (b) Annotated CID MS/MS spectrum of a peptide GLPYPQR. CID produces mostly y and b ions.

the MS/MS function. For this reason, unless otherwise specified, we use mass spectrometry (or mass spectrometer) to also refer to the tandem mass spectrometry (or tandem mass spectrometer) in the rest of this article.

2.3 Mass Spectrometer Configurations

Each of the three aforementioned components (the ionizer, the mass analyzer and the detector) of a mass spectrometer can be made with different technologies, causing different properties of the data. In proteomics MS/MS data analysis, one cares mostly about the ionizer type, the mass analyzer type, and the peptide fragmentation method.

Two types of ionizers are commonly used in proteomics. These are MALDI (matrix-assisted laser desorption/ionization) and ESI (electrospray ionization). The main difference is that MALDI produces singly charged ions ($z = 1$) and ESI produces singly and multiply charged ions ($z \geqslant 1$). The advantage of ESI is that a relatively large molecule can still fall into the m/z range of a mass spectrometer when $z > 1$. However, the existence of multiply charged ions increases the complexity of the spectrum because 1) a single type of molecule may produce multiple peaks due to different charge states, and 2) the charge state of a peak needs to be determined by other means in order to convert the m/z value back to the mass value.

Common mass analyzers used in proteomics include: ion trap, quadrupole, TOF (time-of-flight), FTICR (Fourier transform ion cyclotron resonance), and orbitrap. The difference in mass analyzers mostly affects the resolution and the mass accuracy of the data. Normally the order of performance in terms of resolution and accuracy is iontrap $\approx$ quadrupole $<$ TOF $<$ FTICR $\approx$ orbitrap.

A few fragmentation methods exist for tandem mass spectrometers: CID (collision induced dissociation), CAD (collision activated dissociation), IRMPD (infrared multiphoton dissociation), SORI-CID (sustained off resonance irradiation collision induced dissociation), ECD (electron capture dissociation), ETD (electron transfer dissociation), and HCD (higher-energy C-trap dissociation). These methods tend to fragment at different sites of a peptide, and often generate different types of fragment ions and therefore significantly different spectra for the same peptide. Thus, most commercial data analysis software provides different parameter settings for different instruments, and allows the users to choose before the data analysis.

3 Applications of Mass Spectrometry

Mass spectrometry has been used for many applications in proteomics. This section reviews the most common ones. Some of these applications are related to each other. For example, the protein quantification (Subsection 3.8) relies on the successful protein identification (Subsection 3.1) as the first step of the analysis. Some techniques developed in one application may also be useful in other applications.

3.1 Protein Identification with a Database

Protein identification is by far the most popular application of mass spectrometry in proteomics today. In this application, the mass spectrometry data are used to identify the proteins in the sample with the assistance of a protein sequence database.

Although the wet-lab procedures for protein identification may be slightly different from each other, a typical procedure consists of four steps and is illustrated in Fig.3. The mixture of proteins is first digested into peptides, which are then separated with liquid chromatography (LC) before the mass spectrometry measurement. Both MS and MS/MS spectra are measured in the experiment.

In the so-called *data dependent acquisition* (DDA) mode, each MS scan may be followed by a few MS/MS scans. Each MS/MS scan fragments a different peak in the MS scan. This typical LC-MS/MS experiment is often modified in the lab depending on the instrument type and the complexity of the protein mixture. For example, for a MALDI instrument, the LC has to be done offline. For very complex protein samples, another separation with 1D or 2D gel[8], or 2D LC[1] may be necessary. But what is common is that all these experiments result in thousands (even hundreds of thousands) of MS/MS spectra for one sample. There are two complications for the MS/MS data. First, the MS/MS spectra can correspond to either peptides or chemical noises. Secondly, many peptides in the sample do not produce MS/MS spectra because of their low concentration and the competition from other peptides.

In addition to the MS/MS data, we are usually given a protein sequence database that supposedly contains all the target proteins. Therefore, the computational task is to select the correct proteins from the database. Many software packages exist for this task. The popular ones include Mascot[9], PEAKS[10], Sequest[11], Tandem[12], Ommsa[13], and Phenyx[14]. Different packages use slightly modified procedure for the data analyses. But all of them include two major steps: first, each MS/MS spectrum is used to identify a peptide sequence from the database; secondly, the peptides are grouped together to identify the proteins from the database.

For the peptide identification step, a scoring function is needed to measure the quality of the matching

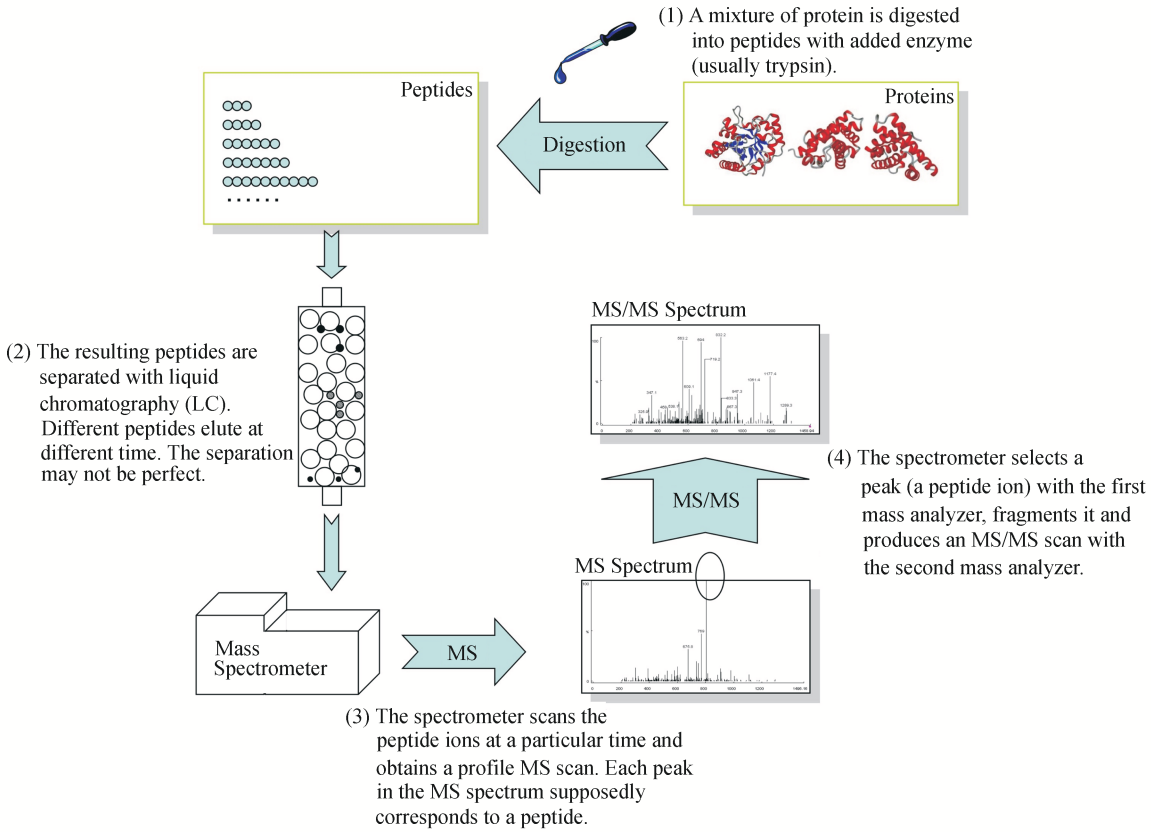


Fig.3. Typical LC MS/MS experiment procedure.

between a given peptide and the MS/MS spectrum. Each peptide in the database with proper mass value is scored using the spectrum and the scoring function, and the highest scoring peptide is output as the answer. A good scoring function is of primary importance for the peptide identification accuracy. Given a peptide and a spectrum, most software computes the theoretical m/z values of the fragment ions of the peptides, and matches the peaks of the spectrum with the m/z values. Usually the intensities and the numbers of the matched peaks, as well as the mass errors of the matching are taken into account in the scoring function. The fragment ion types are also important because a certain type of mass spectrometer usually produces higher intensity peaks for certain ion types. Readers are referred to [9-16] for some examples of the scoring functions in use. Some of these scoring functions can also be used in the *de novo* sequencing application introduced later in Subsection 3.2. To develop a good scoring function, Zhang also predicted the theoretical MS/MS spectrum of the input peptide using complex fragmentation pathways and compares it directly with the experimental MS/MS spectrum[17].

The peptide identification scores of different search engines cannot be directly comparable to each other. Fenyo and Beavis suggested a method to "normalize" different scores by using the significance of the matching[18]. During the database searching, the suboptimal matches to the input are used to train a "survival function", which is then used to convert the matching score to a significance value. Their paper claimed that this significance value can be compared across different database search algorithms.

Even with a good scoring function, false discoveries still exist. For high throughput data analysis, proteomics researchers very much want to know the false discovery rate at certain score threshold. This will help them to determine which analysis results are trustworthy and the others discarded. Currently, this result validation step is commonly done with the so-called decoy database method[19]. In such a method, a random database (the decoy) is generated with similar statistical properties as the target database. The peptide identification algorithm is done on both the target database and the decoy database. The false discovery rate at a given score threshold is estimated by the number of matches in the decoy database with scores above the threshold. There have been several reports on how to generate a good decoy database[19-20]. Notably, in the original proposal for using decoy database[21], the

target database sequences are reversed to form the decoy. There is still no consensus in this community on the optimal way of using decoy database. For example, it is recommended in [19] that the target and the decoy should be concatenated and searched together, while in [20] the suggestion is that they should be searched separately.

The following three related problems are very useful to increase the peptide identification performance and are still not completely solved.

C1. Accurate prediction of the MS/MS spectrum of a given peptide.

C2. Better scoring function to assess the matching between a peptide and an MS/MS spectrum.

C3. Result validation method to estimate the false discovery rate of the peptide identification.

After all peptides are identified, the protein identification is still a challenging problem. The first reason is that not all peptides of a protein can be identified. When there are two or more peptides of a protein are identified with high confidence, the protein is usually true. However, a protein may have only one peptide identified, making it very difficult to judge whether it is a false discovery. This is commonly known as the "one hit wonders" in protein identification. A method of combining the MS and MS/MS spectra is proposed in [22] to help improve the situation. The second reason is that each identified peptide may be shared by a few proteins in the database. This is often caused by the existence of homologous proteins in the database. It is difficult to determine which of the proteins sharing the same peptides is real. Or perhaps all of the homologous proteins are present in the sample. One can imagine that the relationship between the identified peptides and the proteins in the database is a bipartite graph (Fig.4). Inferring the correct proteins from this bipartite graph is a difficult problem and researches aiming to deal with this situation include [23]. These aforementioned difficulties raise the following four problems.

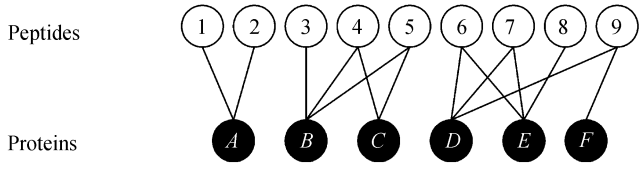


Fig.4. Each peptide may be contained by multiple proteins, resulting in a bipartite graph that is hard to resolve.

C4. Solving the "one hit wonders".

C5. Result validation for protein identification.

C6. Use both MS and MS/MS spectra for protein identification.

C7. Accurate protein inference from peptide assignments.

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. Result validation is urgently needed for both protein and peptide identification. In fact, as an effort to minimize the errors in published results, a working group has started to develop guidelines in publications of peptide and protein identification data[24].

Several other complications in the data also give difficulties to the software currently in use. It is reported that for some instruments only 5∼50% of the MS/MS spectra can be confidently mapped to the peptides in the database[25]. A few reasons have been reported for this low utilization of the data. The largest reason is perhaps due to noise spectra, which are caused by either poor fragmentations of the peptide ions, or the fact that the selected parent ion is not a peptide ion at all. The inclusion of these noise spectra not only increase the computational complexities, but also increase the false discovery rates of the results. Therefore, the following computational task is useful for most mass spectrometry applications discussed in this paper, including protein identification. Researches on removing the noise spectra can be found in [25] and its references.

C8. Scoring function to evaluate the quality of each input MS/MS spectrum.

Another reason for the aforementioned low utilization of the data is the PTMs in the peptides. Usually, a significant portion of peptides in the digested samples are modified, which causes mass changes to some residues. A protein identification software package usually allows both variable PTMs and fixed PTMs. A fixed PTM of a residue means that every occurrence of the residue in the target proteins is modified with the PTM, whereas for a variable PTM a residue may or may not be modified. The dealing with fixed PTM is rather simple — the software can simply substitute the mass of the residue by the mass of the modified residue during the database search. However, for each variable PTM on a residue, the software has to try both cases that the PTM is on and off. This causes exponential growth of the searching space when there are multiple variable PTMs provided to the software. PTMs significantly increase the complexity of the protein identification and deserve a separate section (Subsection 3.5) to review it.

C9. Efficient algorithm to allow multiple variable PTMs in database searching.

The throughput and scale of the mass spectrometry experiments have grown rapidly in the past ten years. Today, one experiment dealing with the whole proteome of an organism may involve tens of hours of LC-MS/MS

runs. If the mass spectrometer produces one MS or MS/MS scan per second, this can give up to hundreds of thousands of MS and MS/MS spectra for a single data analysis task. The size of the data has exceeded the scalability of most software on the market. Researchers often need to divide the dataset themselves, run the data in batches with existing software, and then merge the results together at the end. However, some steps of the analysis may require dealing with the dataset as a whole. In order to fully utilize the information in the data, it is better to incorporate this data division into the data analysis algorithm of the software more carefully. Some existing software such as PEAKS has started doing this[26]. Although the handling of large data may not require fancy algorithms, it is a very practical concern in this field. Therefore, we would like to list the data size problem as one of the computational challenges too.

C10. Handling extremely large mass spectrometry data size.

3.2 Peptide *De Novo* Sequencing

The database search approach for protein identification requires the target proteins and peptides to be in the database. However, this prerequisite is often not satisfied due to many reasons such as incomplete genome sequencing, inferior gene prediction from the genome, alternatively spliced genes, sequence variations between two individuals of the same species, and non ribosomal peptides. When this happens, *de novo* sequencing is the only choice for identifying the peptides. A *de novo* sequencing algorithm takes an MS/MS spectrum as input, and outputs a peptide sequence that best matches the spectrum. The computation does not require any protein database. Rather, the peptide sequence is constructed by the algorithm from the MS/MS spectrum.

Recall that the most important component of the database search approach is a scoring function to assess the matching between a spectrum and a peptide. However, the algorithm component is rather simple — as simple as enumerating every peptide in the database with proper mass values. This is not the case any more for *de novo* sequencing. Enumerating every possible amino acid combinations with a given total mass value will take exponential time. Therefore, it is also important to design efficient algorithm to construct the optimal peptide sequence for *de novo* sequencing. For this reason, *de novo* sequencing has gained more interests among the computer science researchers in bioinformatics. Some commonly used software packages for *de novo* sequencing include PEAKS[10], PepNovo[27] and Lutefisk[28]. One of the earliest mathematical models for *de novo* sequencing was given in [29]. In such a model a spectrum is converted to its *spectrum graph* representation and the finding of a solution is then solely done on the graph. This model has been polished by later researchers. Notably, a completely different model was used in the algorithm of PEAKS software[30]. The readers are also referred to the review articles[31-33] for more complete introductions of *de novo* sequencing and its algorithms. A comparison of several commonly used *de novo* sequencing packages can be found in [34].

When the *de novo* sequencing is done manually, a human interprets the spectrum by examining the ion ladders. A series of high intensity peaks are called ladders if the m/z difference between every adjacent pair of peaks is approximately equal to the mass of an amino acid residue. In a CID MS/MS spectrum, if the peptide fragmentation is ideal, all of the y-ions (or b-ions) can be observed and their peaks should form a complete series of y-ion (or b-ion) ladders. For other types of fragmentation methods, ladders of other ion types may be observed. The mass differences between adjacent peaks in the ladders can be used to derive the amino acid sequence of the peptide.

The difficulties of *de novo* sequencing are mostly due to the imperfect data. First, when the ion ladders are incomplete, only partial sequence information can be derived. Most algorithms examine both the N-terminal ion (e.g., y-ion) and C-terminal ion (e.g., b-ion) ladders to improve the sequencing accuracy and coverage. The PEAKS algorithm additionally utilizes some of the internal fragment ions to further improve the accuracy[10]. Secondly, there are a lot more peaks than just the N-terminal and C-terminal ion ladders. Many of these peaks are from other fragmentations of the peptide. Some of the other peaks can be misinterpreted by the algorithm as the peaks in the N-terminal or C-terminal ion ladders, causing errors in the result. There are efforts in determining whether a peak is a y-ion or a b-ion peak by examining the other related peaks in the spectrum[35].

Peptide *de novo* sequencing is a significantly harder problem than the peptide identification with a database. It requires much higher quality data in order to derive the complete peptide sequence. When the complete peptide sequencing is not possible, it is desirable to derive a partial sequence tag. Many *de novo* sequencing packages such as Lutefisk[28] and Sherenga[36] output partial sequence tags when they are unsure about some amino acids. PEAKS software computes a "local confidence score" for each amino acid in its *de novo* sequencing result[37]. By removing the amino acids below a confidence threshold, the remaining amino acids form a sequence tag.

To overcome some of the data quality problem,

researchers have tried to produce more than one spectra of the same peptide with different fragmentation modes, and perform *de novo* sequencing by using the multiple spectra together[38-39]. This approach was first proposed in [38] by combining CAD and ECD. In [39] CID and ETD spectra are combined. CID and CAD produce more b- and y-ions, whereas ECD and ETD produce more c- and z'-ions. The b- and c-ions (or y- and z'-ions) in the two spectra, respectively, can confirm each other. This will significantly increase the chance of observing a complete ion ladder, and reduces the chance of misinterpretation of other peaks as ion ladder peaks. As a result, significant improvements on accuracy were observed in [38-39].

The database searching approach is good at "re-identifying" proteins and peptides. To make new discovery, peptide *de novo* sequencing is necessary. Both the mass spectrometer instrument quality and *de novo* sequencing software have greatly improved in the past ten years, resulting in much improved *de novo* sequencing accuracy and coverage. But there is still a lot of room to improve the *de novo* sequencing performance, by developing new scoring functions, new algorithms, and new experimental methods. Comparing the database search approaches, the trouble for *de novo* sequencing is that the scoring function must be designed together with the efficient algorithm that computes the solution. Therefore, we list *de novo* sequencing as a whole challenge, instead of dividing it into smaller problems. In short term, the combination of multiple fragmentation modes show great promise in the improvement of *de novo* sequencing.

C11. Better peptide *de novo* sequencing methods and algorithms.

3.3 Peptide/Protein Identification with a Homologous Database

Both the database search method and the *de novo* sequencing method have their limitations. The former requires the protein or peptide sequence to be in the database and the latter requires higher quality spectrometry data. In this subsection we review some existing efforts to combine the strengths of both methods.

The genomes of more than 180 organisms have been sequenced as of today[40]. For any commonly studied organism, there is a good chance that it has a close relative whose genome has been sequenced. Once the genome is sequenced, gene prediction software can be used to predict the genes and obtain the protein sequence database. Consequently, even if our target protein does not exist in any databases, there may be a database protein sequence that is closely homologous to the target protein. The homologous sequence can provide useful (although not completely accurate) information towards the identification of the target protein.

In addition, because the small genome variation between different individuals of the same species, even if the protein database exists, the target protein sequence may be slightly different from the one in the database. Hence, the peptide/protein identification with a homologous database is also useful even if the genome of the studied organism has been sequenced.

Earlier utilization of homologous databases was conducted by *de novo* sequencing followed with a standard homology search. This requires some fine tunes in the searching parameters because the sequence tags are usually short. Three general purpose homology search programs, FASTA, Shotgun and BLAST have been modified to sequence tag search programs: FASTS[41], MS-Shotgun[42] and MS-BLAST[43]. Given a list of *de novo* sequencing tags, these programs often can find a protein that is homologous to the target protein.

However, there are often errors in the *de novo* sequencing tags. The most frequent *de novo* sequencing errors are one segment of amino acids is replaced by another segment with the same total mass value. For example, a peptide sequence *LSCFAV* is mistakenly sequenced as *EACFAV*. Notice that the mass values of *LS* and *EA* are both approximately 200.1 Da. When the fragmentation between the two residues L and S does not form high peaks in the spectrum, this error can be easily made by the *de novo* sequencing software. This type of errors has very different properties than those statistical models developed for homology mutations, which makes the general purpose homology search inappropriate for searching *de novo* sequencing tags.

The SPIDER[44] and OpenSea[45] programs are developed for the homology search with the *de novo* sequencing errors in mind. Both programs match partially correct sequence tags with a database to identify the homologous or modified proteins. The difference is that SPIDER's algorithm allows the homology mutations and the *de novo* sequencing errors to occur at the same site.

It is noteworthy that these sequence tag searching program can be used to search against the exact database (instead of a homologous database) as well. SPIDER has a special option to support this type of search. Also, another sequence tag searching program, GutenTag[46], can be used to do sequence tag searching on exact protein databases. It was reported in [46] that this type of sequence tag search found different set of peptides and has a lower false discovery rate than the database searching approach for peptide identification.

Another approach of using homologous database is to try to mutate the amino acids of each database peptide and match the mutated peptides with the input spectrum. This can be regarded as if the regular database search is done on an expanded protein database. The hope is that the target peptides are included in the expanded database. However, this approach inevitably increases the searching complexity.

C12. Peptide and protein identification with a homologous database.

3.4 Complete Protein Sequencing

As reviewed in Subsection 3.3, the protein database is often not available. Even it is, the complete sequence of the protein may be slightly different from the one in the database. Consequently, merely identifying the protein does not always tell us the accurate sequence of the protein. When the complete protein sequence is wanted, new experimental and computational methods are needed.

Traditionally, the sequencing of novel or mutated proteins is done by the time-consuming Edman degradation. Recently, the possibility of using MS/MS to sequence novel proteins has drawn researchers' attention. Protein sequencing with MS/MS has been previously done manually in proteomics as follows. First, the target protein is digested with multiple enzymes. Because different enzymes digest at different sites, the multiple digests result in overlapping peptides. Then each digest is measured with MS/MS and *de novo* sequencing is used to derive the sequence of each peptide. At last, an assembly step is performed to put all the overlapping peptides together.

By using the above approach, a few groups have successfully sequenced complete proteins[3,47] with MS/MS. Automated software tools were also developed for analyzing this type of data[48-51]. Among these works, Bandeira's algorithm[48-50] is slightly different from the manual analysis procedure. Instead of using *de novo* sequencing to get the peptide sequences, their algorithm produces an intermediate *prefix residue mass spectrum* from each MS/MS spectrum, and then assembles the prefix residue mass spectra together. The Champs algorithm in [51] is very similar to the aforementioned manual analysis procedure. However, it utilizes a homologous protein database to assist the assembly. The homologous protein in the database allows the algorithm to use the SPIDER algorithm[44] to correct the *de novo* sequencing errors and serves as a template for the assembly. This allowed the algorithm to achieve almost full accuracy and coverage on two standard proteins in [44].

The research in protein sequencing with MS/MS has just started in the bioinformatics community. This is only because the mass spectrometry instruments and peptide *de novo* sequencing have been improved to a level such that automated complete protein sequencing becomes possible. In fact, all the current works in the literature for complete protein sequencing required carefully controlled wet lab experiments on purified proteins. Once this can be done accurately in an automated and high-throughput fashion, there will be great needs for this method in proteomics. We feel this is a very promising direction and a lot more research in this problem will appear.

C13. Complete protein sequencing with MS/MS.

3.5 PTM Characterization

There are a few hundreds of known PTMs with the most common being phosphorylation, glycosylation, methylation, acetylation and acylation[4,52]. Many of these modifications have significant influence on the activity and specificity of proteins, and may even play a role in stabilizing a protein's structure and in regulating enzymatic activity. In many cases, the proteins are modified at several sites by a number of added functionalities. PTMs have also been reported to be involved in various diseases including cancers (see e.g., [53]). Clearly, it is important to report all PTMs in the target protein. A PTM on an amino acid residue usually changes the mass of the residue, which is reflected by the change in the peptide's MS/MS spectrum. Thus, the characterization of PTM is possible with MS/MS, but with a few difficulties.

First, since PTMs are added after the translation from mRNA to protein, there is no simple rule to determine the modification sites from the genome. Thus, even an organism's genome is sequenced, the PTMs are unknown. As a result, protein databases usually do not contain the PTM information. The protein identification algorithm has to expand the protein database by adding the possible PTMs, resulting in high computational complexity. There have also been researches on predicting the PTM sites from the protein sequences. For example, Blom *et al.* used the protein sequence and structure to predict the phosphorylation sites[54]. At this moment we are not aware of researches on combining such predictions with the mass spectrometry. But this combination can potentially be very useful.

The second difficulty is the low coverage of the protein in the data. Many peptides of the target protein do not produce spectra for reasons such as low concentration and competitions among peptides. Consequently, there will be no information in the data about the PTM on those uncovered regions of the protein.

Thirdly, peptides with certain modifications can

produce more complex spectra than a simple peptide. For example, when a phosphorylated peptide is subject to CID, the β-elimination mechanism can cause a neutral loss of -96 Da or -80 Da on the phosphorylated serine (S) or threonine (T). This results in altered MS/MS spectrum than without β-elimination. The scoring function trained for unmodified peptide may not be suitable for the peptides with these complex modifications. The more involved PTM is the glycosylation, which is reviewed separately in Subsection 3.6.

Lastly, researchers often do not know which PTMs exist in their sample. Letting the software try all of the few hundred known PTMs makes the computation infeasible. Turning on too many PTMs in the searching also increases the false discoveries significantly because the growth of the searching space. The most attractive solution is to let the software identify the possible PTMs automatically from the data. Tsur *et al.* developed a "blind search" strategy to identify the PTMs by aligning the spectrum (with PTM) with the database peptides directly[55]. Moreover, for a variable PTM, both the modified and unmodified copies of the same peptide may appear in the sample. By comparing the spectra of the modified and unmodified peptides, one can also possibly identify the PTM. MacCoss *et al.* realized this potential[56] and Bandeira *et al.* exploited this technique with a spectral network concept[57].

Here we list the following three challenges related to PTM discovery. Readers will also find the review article[58] and its references a useful resource.

C14. Combining sequence-based PTM prediction and MS/MS for PTM discovery.

C15. Better scoring function for matching MS/MS spectrum with modified peptides.

C16. Discovery of unknown (unspecified) PTMs.

3.6 Glycan Structure Determination

Glycosylation is the most common PTM in mammal proteins. It is estimated that over 50% of all mammalian proteins in eukaryotic systems are glycosylated at some point during their existence[59]. Glycoproteins are known to be involved in a long list of diseases including rheumatoid arthritis[60] and cancer[61]. Glycosylation adds a glycan structure to the peptide (Fig.5). Unlike other simple PTMs that have a fixed mass change, the glycan may have variable structures and variable mass values at different modification sites. This makes the characterization of glycosylation significantly more involved than other PTMs.

A glycan has a tree structure consists of many monosaccharides (sugar units) connected with glycosidic linkages (Fig.5(b)). There are a few common sugar units, most of which having different mass values. The glycosidic linkages between the sugar units breaks with lower energy than the peptide bonds when glycopeptide ions are fragmented in MS/MS experiments (Fig.5(a)). The resulting fragment ions will form characteristic peaks of the glycan in the MS/MS spectrum. Therefore, structural information of the glycan can in principle be deduced from the spectrum. However, the glycan structure determination is more difficult than the peptide sequencing because the target structure is now a tree instead of a linear sequence.

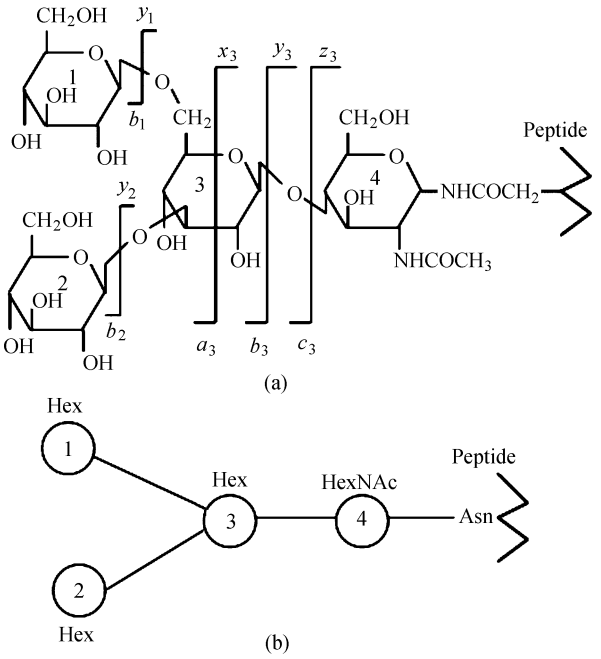


Fig.5. (a) Glycan structure fragments and produces different types of fragment ions in MS/MS. (b) Tree abstraction of the glycan structure.

There have been attempts to solve glycan structure problem by using MS/MS. The classical methods for the characterization of glycoproteins by mass spectrometry were to cleave glycans with enzymes and then analyze the structures of the released glycans. Therefore, most of reported algorithms focus on interpreting MS/MS spectra of released glycans (see [62] and its references). Recently, biochemists began to analyze glycopeptides derived from trypsin digestion of glycoproteins directly[63-64], and algorithms have been developed for analyzing this type of data[65-66].

Although the models in [62, 65] are slightly different, both of their algorithms construct good solutions for smaller sized trees and then assemble the constructed tree structures into larger and larger ones. The algorithm in [62] uses dynamic programming. But in each step, it keeps the 200 best solutions under a simple scoring function. Then a post-processing step re-evaluates

these solutions with a more accurate scoring function at the end of the dynamic programming. The algorithm in [65] uses a heuristic strategy that is similar to dynamic programming. For each tree size, the best 1000 structures constructed so far are kept. The re-evaluation with an accurate scoring function happens immediately after each larger structure is assembled. Shan *et al.* also proved that under some models, glycan structure determination using MS/MS is an NP-hard problem[65]. This justifies the need for either a heuristic algorithm or an exponential time algorithm.

C17. Glycan structure determination with MS/MS.

3.7 Spectrum Library Searching

There are more and more mass spectrometry data becoming publicly available. Some of the popular data repositories are Open Proteomics Database[67] and Peptide Atlas[68]. There are also efforts to produce annotated libraries of experimental MS/MS spectra with known peptide sequences such as the NIST Peptide Mass Spectral Libraries[69].

The MS/MS spectrum of a peptide is very hard to be predicted from the sequence, causing the difficulty in developing an accurate scoring function for peptide identification. However, the spectrum of a peptide is fairly reproducible if the same type of mass spectrometer is used under the same condition. Therefore, if a peptide's MS/MS spectrum has been previously included in an annotated spectrum library, the best way to identify this peptide is to match the experimental spectrum with the library spectrum directly. Algorithms and software systems have been developed to use this annotated spectrum library search approach to solving the peptide identification problem[69-72].

The difficulty of building an annotated spectrum library is the quality control of the annotation. For high quality spectrum, a standard database searching approach for peptide identification can already identify the peptide with high confidence. The benefit of using the annotated spectrum library is greatly reduced in this case. However, for lower quality spectra, it is not obvious how to guarantee the correctness of the annotation, since the annotation is also produced by some sort of peptide identification software. The searching in the library also requires efficient algorithm especially if the library is large. In addition, it will be very useful if one can compare the experimental spectrum of a modified peptide with the library spectrum of the unmodified or differently modified peptide. In this way, the application of the annotated library searching will be greatly expanded. We note that if efficiency is not a concern, Bandeira *et al.*'s work on spectral network[57] can be readily used here to compare spectra of differently modified peptides.

C18. Construction of a quality annotated spectrum library.

C19. Efficient searching for similar spectra in annotated spectrum library.

C20. Efficient matching between the spectra of modified peptides with the spectra of unmodified or differently modified peptides in the library.

3.8 Protein Quantification

Scientists are not satisfied by only knowing the identities of the proteins. The expression levels of the proteins in the sample reveals a lot more information about the protein's participation in a particular function or malfunction of the cells. Protein quantification (also known as quantitation) could provide a comprehensive description of the expression level changes of the proteins under the influence of various perturbations, including stress, infection, or disease. Drug administration and therapeutic effects could also be determined though protein quantitation. Quantitative proteomics can help identify biomarkers of a particular disease and aid in an early diagnosis and intervention.

Several different wet-lab experimental methods have been developed for quantification. The popular ones are ICAT (Isotope-Coded Affinity Tags)[73], SILAC (Stable Isotope Labeling by Amino Acids in Cell Culture)[74], iTRAQ (Isobaric Tag for Relative and Absolute Quantitation)[75], and label-free quantification[76-77]. The first three methods require some sort of isotope labeling of the samples. Multiple samples are labeled with different isotope labels with the same composition but different masses. Then the samples are mixed together and analyzed in the same LC MS/MS experiment. The same peptide from different samples would still elute from the LC at the same time but produces different peaks in the same MS (for ICAT and SILAC) or MS/MS spectrum (for iTRAQ). The mass difference between these peaks is equal to the mass difference of the different labeling reagents. So these characteristic peaks can be easily recognized by the instrument and the software. The relative intensities between the characteristic peaks can be used to compute the quantity ratio of the peptide in the given samples. The protein quantity ratio can be computed from the ratios of the peptides assigned to the protein. The readers are referred to the above references for more details about the labeling methods for quantification.

Recently, the label-free quantification method is gaining more and more attention. In this method, multiple (for example, two) samples are analyzed in separate LC MS/MS experiments under the same condition. Each MS scan of the data has a *retention time*,

indicating at which time of the LC experiment the MS scan is taken. A peptide common in both samples will form peaks in the MS scans of each sample. These characteristic peaks (or *peptide features*) will have the same m/z value; but the MS scans containing them may have slightly different retention times due to the LC variation. The analytical algorithm needs to correct the retention time variation and correctly map the peptide features. Then the peak intensities of the peptide features can be used to compute the quantity ratios of the peptides. The protein quantity ratio is calculated by averaging the peptide ratios together.

The label-free method does not require the costly labeling reagents and avoids problems such as sample loss and unwanted side reactions common with ICAT and iTRAQ. In addition, the label-free method would, in principle, allow the comparison of datasets of current samples with datasets of samples that do not exist anymore, and potentially would allow for the comparison of datasets obtained from separate labs. The labs would need very similar experimental protocols, but would not need to exchange samples. All these make the label-free method the most promising method for large scale comparison of hundreds of samples that are required by biomarker discovery. Moreover, label-free method imposes more computational challenges for bioinformatics researchers. In the rest of this section we focus on these challenges.

The retention time correction is the first and the key step of the label-free analysis[78], and is typically done by a multiple alignment of the sequences of MS scans of all samples. Similarity scores are calculated for each pair of scans and the alignment is computed in a similar fashion of the Smith-Waterman algorithm for sequence alignment[79]. Additionally, if two MS/MS spectra from different samples correspond to the same identified peptide, their corresponding MS scans can be used as an anchor of the alignment[80]. This helps to improve the accuracy and speed of the alignment algorithm.

C21. Retention time alignment for label-free quantification.

Peptide features can be recognized before or after retention time correction. The features from the same peptide of different samples are mapped together. The intensity of the feature can be the total peak intensities for the feature, or the area size under the peak profile. These calculations are nontrivial due to the noisy data and the overlaps of different peptide features in the spectra.

C22. Peptide feature detection in LC-MS spectral data.

C23. Peptide feature mapping for label-free quantification.

At last, the peptide ratios are averaged together to calculate the protein ratios. This last step is much harder than it appears to be. In the database searching approach for protein identification, we mentioned that a peptide can be shared by multiple proteins, causing difficulties to assign the identified peptides to proteins. This is a bigger problem in quantification. When a peptide feature is shared by two proteins, the intensity should be split to the two proteins before the ratio is calculated. This intensity splitting is a difficult problem. Another problem in protein ratio calculation is that some of its peptide ratios contain large errors and are outliers. This can be caused by reasons such as overlapping peptide features, shared peptides, and wrong peptide identification. An outlier removal step is needed before the ratios are averaged together.

C24. Accurate calculation of protein ratios from peptide ratios.

Since many proteins are modified with variable PTMs, researchers in protein quantification are also interested in knowing what percentage of the proteins are modified by a certain variable PTM. The percentage changes across samples can potentially be used as biomarkers too.

C25. PTM quantification.

3.9 Sequencing Non-Standard Peptides

In all the above reviewed applications, we assume the peptide to be a linear sequence of amino acid residues. However, in some other applications a peptide can have a more complex structure. Mass spectrometry has been used in sequencing non-standard peptides. These include two peptides bound with disulphide bonds[81] and non-ribosomal peptides[82].

C26. Determining the structure of peptides with disulphide bonds with MS/MS.

C27. Identification and sequencing of non-ribosomal peptides with MS/MS.

Recall that a peptide ion is selected by the first mass analyzer of the tandem mass spectrometer based on the m/z value. In an LC MS/MS experiment, there is a small chance that two different peptides with the same m/z get fragmented together and form a single MS/MS spectrum. The resulting spectrum contains the fragment ions from both peptides. When the mixture is dominated by one peptide, the standard peptide identification methods can still identify the dominating peptide. But the identification of both peptides is difficult. An initial research for this problem can be found in [83].

C28. Identifying both peptides from their mixed MS/MS spectrum.

3.10 Top-Down Protein Identification

All the previous reviewed methods for protein identification belong to the so called bottom-up approach. That is, a protein needs to be digested into shorter peptides before the mass spectrometry analysis, and the identification of the protein requires the identification of the shorter peptides first. Recently, with the assistance of high-end mass spectrometers (such as FTMS and Orbitrap), there have been attempts to analyze the intact protein directly (see [84] and its references). The intact protein is highly charged to give a proper m/z value for the measurement in a mass spectrometer. The highly charged protein ion is fragmented and an MS/MS spectrum is produced for the fragments of the whole protein. The protein can be identified by comparing the theoretical fragment ions with the observed peaks in the spectrum. This method is called top-down protein identification (or top-down protein sequencing). Since the fragmentation pathway for the much longer protein is more complicated than a shorter peptide, the scoring functions for peptide identification cannot be used here. New scoring functions need to be developed.

C29. Top-down protein identification.

4 Related Research Topics

In this section we briefly review some related research topics in mass spectrometry data analysis. These research problems do not belong to any single application in Section 3. However, solving these problems will help all of the above applications in general.

4.1 Peptide Detectability

It is known that some peptides of a particular protein are easier to be detected in a mass spectrometry experiment than the others. Reasons affecting the detectability of a peptide include the following. 1) Peptides may not be correctly digested in the protein digestion step. 2) There are PTMs on the peptides. 3) The peptides are lost in the LC column. 4) Peptides do not ionize well and therefore the resulting low intensity peaks in the MS scans are ignored by the data dependent acquisition. 5) Peptides fragment poorly and produce low quality MS/MS spectrum. Efforts have been made to predict the detectability of peptides from the protein sequence[85-86]. If this detectability can be predicted rather accurately, it can help improve all applications in the bottom-up analysis. For example, it can help solve the protein inference problem[86] and is especially useful for protein quantification[85].

C30. Accurate prediction of the peptide's detectability.

4.2 Peptide Identification with Multiple Spectra

The inherent difficulty of peptide identification is the low-quality data due to poor fragmentation of the peptide. There exist two approaches to improving the fragmentation. One is to fragment the peptide with two different techniques such as CID and ETD, collectively or respectively. The other is to use multistage MS, which selects a fragment peak from an MS/MS spectrum and fragment it again to form an MS^3 spectrum. In theory this process can be continued to form the MS^n spectrum. The technical note[87] even used two fragmentation techniques to do multistage MS.

In Subsection 3.2 we reviewed some efforts on combining two MS/MS spectra of the same peptide with different fragmentation techniques to improve the peptide *de novo* sequencing. There are also researches[88] to use multistage MS data to do *de novo* sequencing. Apparently this type of data should also be helpful in other applications such as protein identification with database searching. Hence we raise the following general problems as challenges.

C31. Peptide/protein identification or sequencing with multistage MS.

C32. Peptide/protein identification or sequencing with multiple fragmentation techniques.

4.3 Data Compression

The raw mass spectrometry data are getting larger and larger. A typical LC MS/MS experiment on a Q-T of instrument produces 1 G bytes of data per hour. Data compression becomes a very useful technique. The general file compression programs do not give the optimal compression ratio and do not allow the direct access to individual spectrum of the compressed data. There have been efforts on developing better compression tools specifically for mass spectrometry data[89-90]. Some tools support both lossy and lossless compression. In general, the m/z information of a peak is more important than the intensity information for the data analysis. Therefore, for the lossy compression, one can give up more on the intensity information to gain better compression ratio.

C33. Mass spectrometry data compression.

4.4 Retention Time Prediction

For LC MS/MS experiments, each MS/MS spectrum is associated with a retention time, which is the time the peptide elutes from the LC column. This time is fairly reproducible under the same LC condition, and is predictable from the sequence of the peptide[91]. The retention time can be used to validate whether

the peptide identification result is correct, or included in the scoring function to increase the identification accuracy[92]. Also, the traditional *mass fingerprint* method uses the peptides' m/z values as characteristics to identify the protein[93]. This now seemingly obsolete method clearly can be improved by combining the retention time. If this is successful, then all the peaks in the MS scans can be utilized in the data analysis. These scans are currently being ignored in the protein identification analysis.

C34. Accurate peptide retention time prediction, and its applications in protein identification and quantification.

4.5 Better Spectrum Preprocessing

A spectrum often contains a lot of noisy small peaks that cannot be utilized by the peptide identification algorithms. It is a difficult task to select the signal peaks from a noisy spectrum. Good peak picking algorithms have been reported to improve the accuracies of the protein identification and quantification[94-95]. A particular difficulty for peak picking is that some peaks overlap each other.

ESI ionization produces multiply charged ions. Many analysis algorithms are based on the mass (instead of m/z) of the fragments. Thus, they require the conversion of the multiply charged ions to the singly charged ions. This process is called *deconvolution*. Deconvolution is typically done by examining the m/z difference between the monoisotopic peak and the isotopic peaks of the same ion. If the difference is Δ, the charge state of the peak is the integer rounding of the value $1/\Delta$. However, this is not a trivial task when the overlap of two ions causes difficulty in recognizing the isotopic peaks.

C35. Better peak picking and deconvolution algorithms.

4.6 Biomarker Discovery

A major motivation for proteomics using mass spectrometry is to identify protein or PTM biomarkers. This is typically a post-analysis after the protein identification or quantification analysis. For example, novel protein biomarkers for Down syndrome disease were identified from the pregnant women's blood samples using MS/MS and protein identification software[96]. There are also efforts to directly identify biomarkers from the mass spectrometry data first before the protein identification[97]. Biomarker discovery is a much broader topic than can be possibly covered in this article. We list it as a challenge here without detailed discussion.

C36. Biomarker discovery from mass spectrometry data.

5 Discussion

The applications of mass spectrometry in proteomics have drastically changed the study of proteins from a labour-intensive protein-by-protein style to a computation-intensive high-throughput fashion. This is also why the aforementioned bottom-up approach is also called the shotgun proteomics by many researchers. The need for bioinformatics in this area is inevitable. And there are too many research topics in this area to be reviewed in details in this article. Nevertheless, this article is the author's effort to provide an entry point for those who are interested in this area. Readers are encouraged to continue their reading with the references of this article. In addition to the references cited before, readers may find the review articles[98-99] and books[100-101] very useful.

References

[1] Peng J, Elias J E, Thoreen C C, Licklider L J, Gygi S P. Evaluation of multidimensional chromatography coupled with Tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: The yeast proteome. *Journal of Proteome Research*, 2003, 2(1): 43-50.

[2] Mann M. Quantitative proteomics? *Nature Biotechnology*, 1999, 17(10): 954-955.

[3] Martin-Visscher L A, van Belkum M J, Garneau-Tsodikova S, Whittal R M, Zheng J, McMullen L M, Vederas J C. Isolation and characterization of carnocyclin A, a novel circular bacteriocin produced by Carnobacterium maltaromaticum UAL307. *Applied and Environmental Microbiology*, 2008, 74(15): 4756-4763.

[4] Mann M, Jensen O N. Proteomic analysis of post-translational modifications. *Nature Biotechnology*, 2003, 21(3): 255-261.

[5] Keykhosravani M, Doherty-Kirby A, Zhang C, Brewer D, Goldberg H A, Hunter G K, Lajoie G. Comprehensive identification of post-translational modifications of rat bone osteopontin by mass spectrometry. *Biochemistry*, 2005, 44(18): 6990-7003.

[6] Hoffmann E, Stroobant V. Mass Spectrometry: Principles and Applications. John Wiley & Sons Ltd., 2007.

[7] Tang K, Page J S, Smith R D. Charge Competition and the linear dynamic range of detection in electrospray ionization mass spectrometry. *Journal of American Society of Mass Spectrometry*, 2004, 15(10): 1416-1423.

[8] Gygi S P, Corthals G L, Zhang Y, Rochon Y, Aebersold R. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *PNAS*, 2000, 97(17): 9390-9395.

[9] Perkins D N, Pappin D J, Creasy D M, Cottrell J S. Probability-based protein identification by searching sequence database using mass spectrometry data. *Electrophoresis*, 1999, 20(18): 3551-3567.

[10] Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G. PEAKS: Powerful software for MS/MS peptide *de novo* sequencing. *Rapid Communications in Mass Spectrometry*, 2003, 17(20): 2337-2342.

[11] Eng J K, McCormack A L, Yates III J R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Amer. Soc. Mass Spectrom.*, 1994, 5(11): 976-989.

[12] Craig R, Beavis R C. TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics*, 2004, 20(9): 1466-1467.

[13] Geer L Y, Markey S P, Kowalak J A, Wagner L, Xu M, Maynard D M, Yang X, Shi W, Bryant S H. Open mass spectrometry search algorithm. *J. Proteome Research*, 2004, 3(5): 958-964.

[14] Colinge J, Masselot A, Giron M, Dessingy T, Magnin J. OLAV: Towards high-throughput tandem mass spectrometry data identification. *Proteomics*, 2003, 3(8): 1454-1463.

[15] Bafna V, Edwards N. SCOPE: A probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics*, 2001, 17(Supplement 1): S13-S21.

[16] Wan Y *et al.* PepHMM: A hidden Markov model based scoring function for mass spectrometry database search. In *Proc. RECOMB 2005*, Standford, USA, May 21-22, 2005, pp.342-356.

[17] Zhang Z. Prediction of low-energy collision-induced dissociation spectra of peptides. *Analytical Chemistry*, 2004, 76(14): 3908-3922.

[18] Fenyo D, Beavis R C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical Chemistry*, 2003, 75(4): 768-774.

[19] Elias J E, Gygi S P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 2007, 4(3): 207-214.

[20] Bianco L, Mead J A, Bessant C. Comparison of novel decoy database designs for optimizing protein identification searches using ABRF sPRG2006 standard MS/MS data sets. *Journal of Proteome Research*, 2009, 8(4): 1782-1791.

[21] Moore R E, Young M K, Lee T D. Qscore: An algorithm for evaluating SEQUEST database search results. *Journal of the American Society for Mass Spectrometry*, 2002, 13(4): 378-386.

[22] Lu B, Motoyama A, Ruse C, Venable J, Yates J R III. Improving protein identification sensitivity by combining MS and MS/MS information for shotgun proteomics using LTQ-Orbitrap high mass accuracy data. *Analytical Chemistry*, 2008, 80(6): 2018-2025.

[23] Nesvizhskii A I, Aebersold R. Interpretation of shotgun proteomic data — The protein inference problem. *Molecular & Cellular Proteomics*, 2005, 4(10): 1419-1440.

[24] Carr S, Aebersold R, Baldwin M, Burlingame A, Clauser K, Nesvizhskii A. The need for guidelines in publication of peptide and protein identification data. *Molecular and Cellular Proteomics*, 2004, 3(6): 531-533.

[25] Junqueira M *et al.* Separating the wheat from the chaff: Unbiased filtering of background tandem mass spectra improves protein identification. *J. Proteome Research*, 2008, 7(8): 3382-3395.

[26] Hughes C, Doble B, Xin L, Chen C, Shan B, Ma B, Lajoie G. SILAC quantitation with PEAKS to a depth of 3000 proteins from a double knockout GSK-3 of mouse embryonic stem cells. In *ASMS* 2009, Philadelphia, USA, May 31-June 4, 2009, Session Bioinformatics: Quantification, Poster, No. 056.

[27] Frank A, Pevzner P. Pepnovo: *De novo* peptide sequencing via probabilistic network modeling. *Analytical Chemistry*, 2005, 77(4): 964-973.

[28] Taylor J A, Johnson R S. Implementation and uses of automated *de novo* peptide sequencing by tandem mass spectrometry. *Analytical Chemistry*, 2001, 73(11): 2594-2604.

[29] Bartels C. Fast algorithm for peptide sequencing by mass spectroscopy. *Biomed. Environ. Mass Spectrom.*, 1990, 19(6): 363-368.

[30] Ma B, Zhang K, Liang C. An effective algorithm for the peptide *de novo* sequencing from MS/MS spectrum. *Journal of Computer and System Sciences*, 2005, 70(3): 418-430.

[31] Lu B, Chen T. Algorithms for *de novo* peptide sequencing via tandem mass spectrometry. *Drug Discovery Today: BioSilico*, 2004, 2(2): 85-90.

[32] Xu C, Ma B. Review of software for computational peptide identification from MS/MS data. *Drug Discovery Today*, 2006, 11(13/14): 595-600.

[33] Hughes C, Ma B, Lajoie G. De Novo Sequencing Methods in Proteomics. Methods in Molecular Biology, Series, Springer. (to appear)

[34] Pevtsov S, Fedulova I, Mirzaei H, Buck C, Zhang X. Performance evaluation of existing de novo sequencing algorithms. *Journal of Proteome Research*, 2006, 5(11): 3018-3028.

[35] Yan B, Qu Y, Mao F, Olman V, Xu Y. PRIME: A mass spectrum data mining tool for de novo sequencing and PTMs identification. *Journal of Computer Science and Technology*, 2005, 20(4): 483-490.

[36] Dancik V *et al. De novo* peptide sequencing via tandem mass-spectrometry. *J. Comp. Biology*, 1999, 6(3/4): 327-342.

[37] Xin L, Lajoie G, Ma B. New method for the validation of de novo sequencing results. In *ASMS* 2008, Denver, USA, Jun. 1-5, Session: Bioinformatics III, Poster, No. 645.

[38] Savitski M M, Nielsen M L, Kjeldsen F, Zubarev R A. Proteomics-Grade de Novo Sequencing Approach. *J. Proteome Research*, 2005, 4: 2348-2354.

[39] Datta R, Bern M. Spectrum fusion: Using multiple mass spectra for de novo peptide sequencing. In *Proc. RECOMB*, 2008, pp.140-153.

[40] Genome News Network. http://www.genomenewsnetwork.org/.

[41] Mackey A J, Haystead T A J, Pearson W R. Getting more for less: Algorithms for rapid protein identification with multiple short peptide sequences. *Mol. Cell. Proteomics*, 2002, 1(2): 139-147.

[42] Huang L, Jacob R J, Pegg S C H, Baldwin M A, Wang C C, Burlingame A L, Babbitt P C. Functional assignment of the 20 S proteasome from Trypanosoma Brucei using mass spectrometry and new bioinformatics approaches. *J. Biol. Chem.*, 2001, 276(30): 28327-28339.

[43] Shevchenko A, Sunyaev S, Loboda A, Shevchenko A, Bork P, Ens W, Standing K G. Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole timeof-flight mass spectrometry and BLAST homology searching, *Anal. Chem.*, 2001, 73(9): 1917-1926.

[44] Han Y, Ma B, Zhang K. SPIDER: Software for protein identification from sequence tags containing de novo sequencing error. *Journal of Bioinformatics and Computational Biology*, 2005, 3(3): 697-716.

[45] Searle B C *et al.* High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. *Anal. Chem.*, 2004, 76(8): 2220-2230.

[46] Tabb D L, Saraf A, Yates J R III. GutenTag: High-throughput sequence tagging via an empirically derived fragmentation model. *Anal. Chem.*, 2003, 75(23): 6415-6421.

[47] Hopper S, Johnson R S, Vath J E, Biemann K. Glutaredoxin from rabbit bone marrow. Purification, characterization, and amino acid sequence determined by tandem mass spectrometry. *J. Biol. Chem.*, 1989, 264(34): 20438-20447.

[48] Bandeira N, Tang H, Bafna V, Pevzner P. Shotgun protein sequencing by tandem mass spectra assembly. *Analytical Chemistry*, 2004, 76(24): 7221-7233.

[49] Bandeira N, Clauser K R, Pevzner P. Shotgun protein sequencing: Assembly of peptide tandem mass spectra from mixtures of modified proteins. *Mol. Cell Proteomics*, 2007, 6(7): 1123-1134.

[50] Bandeira N, Pham V, Pevzner P, Arnott D, Lill J R. Automated de novo protein sequencing of monoclonal antibodies.

Nature Biotechnology, 2008, 26(12): 1336-1338.

[51] Liu X, Han Y, Yuen D, Ma B. Automated protein (re)sequencing with MS/MS and a homologous database yields almost full coverage and accuracy. *Bioinformatics*, 2009, 25(17): 2174-2180.

[52] Unimod database. http://www.unimod.org.

[53] Oki M, Aihara H, Ito T. Role of histone phosphorylation in chromatin dynamics and its implications in diseases. *Subcellular Biochemistry*, 2007, 41: 319-336.

[54] Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *Journal of Molecular Biology*, 1999, 294(5): 1351-1362.

[55] Tsur D, Tanner S, Zandi E, Bafna V, Pevzner PA. Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol.*, 2005, 23(12): 1562-1567.

[56] MacCoss M J *et al.* Shotgun identification of protein modifications from protein complexes and lens tissue. *Proc. Natl. Acad. Sci. USA*, 2002, 99(12): 7900-7905.

[57] Bandeira N, Tsur D, Frank A, Pevzner P. Protein identification by spectral networks analysis. *Proc. Natl. Acad. Sci. USA*, 2007, 104(15): 6140-6145.

[58] Witze E S, Old W M, Resing K A, Ahn N G. Mapping protein post-translational modifications with mass spectrometry. *Nature Methods*, 2007, 4(10): 798-806.

[59] Dwek R A, Butters TD , Platt F M, Zitzmann N. Targeting glycosylation as a therapeutic approach. *Nature Reviews Drug Discoveries*, 2002, 1(1): 65-75.

[60] Parekh R B *et al.* Association of rheumatoid arthritis and primary osteoarthritis with changes in the glycosylation pattern of total serum IgG. *Nature*, 1985, 316(6027): 452-457.

[61] Dennisa J W, Granovskya M, Warrena C E. Glycoprotein glycosylation and cancer progression. *Biochimica et Biophysica Acta* (*BBA*) — *General Subjects*, 1999, 1473(1): 21-34.

[62] Tang H, Mechref Y, Novotny M V. Automated interpretation of MS/MS spectra of oligosaccharides. *Bioinformatics*, 2005, 21(Suppl. 1): i431-i439.

[63] Zala J. Mass spectrometry of oligosaccharides. *Mass Spectrometry Reviews*, 2004, 23(3): 161-227.

[64] Zhang C, Doherty-Kirby A, Lajoie G. Investigation of cationic peanut peroxidase glycans by electrospray ionization mass spectrometry. *Phytochemistry*, 2004, 65(11): 1575-1588.

[65] Shan B, Lajoie G, Ma B, Zhang K. Complexities and algorithms for glycan structure sequencing using tandem mass spectrometry. *Journal of Bioinformatics and Computational Biology*, 2008, 6(1): 77-91.

[66] An H J, Tillinghast J S, Woodruff D L, Rocke D M, Lebrilla C B. A new computer program (GlycoX) to determine simultaneously the glycosylation sites and oligosaccharide heterogeneity of glycoproteins. *Journal of Proteome Research*, 2006, 5(10): 2800-2808.

[67] Prince J T, Carlson M W, Wang R, Lu P, Marcotte E M. The need for a public proteomics repository. *Nature Biotechnology*, 2004, 22(4): 471-472.

[68] Desiere F *et al.* The PeptideAtlas project. *Nucleic Acids Research*, 2006, 34(Database Issue): D655-D658.

[69] Rudnick P *et al.* NIST reference libraries of peptide fragmentation spectra: 2008. In *ASMS* 2008, Denver, USA, Jun. 1-5, Session: Bioinformatics III, Poster, No. 2008.

[70] Craig R, Cortens J, Fenyo D, Beavis R. Using annotated peptide mass spectrum libraries for protein identification. *J. Proteome Res.*, 2006, 5(8): 1843-1849.

[71] Dutta D, Chen T. Speeding up tandem mass spectrometry database search: Metric embeddings and fast near neighbor search. *Bioinformatics*, 2007, 23(5): 612-618.

[72] Wu Z, Lajoie G, Ma B. MSDash: Mass spectrometry database and search. In *Proc. the 7th Int. Conf. Computational System Bioinformatics*, Stanford, USA, Aug. 26-29, 2008, pp.63-71.

[73] Gygi S P, Rist B, Gerber S A, Turecek F, Gelb M H, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, 1999, 17(10): 994-999.

[74] Ong S E, Blagoev B, Kratchmarova I, Kristensen D B, Steen H, Pandey A, Mann M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics*, 2002, 1(5): 376-386.

[75] Wiese S, Reidegeld K A, Meyer H E, Warscheid B. Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research. *Proteomics*, 2007, 7(3): 340-350.

[76] Wang *et al.* Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Analytical Chemistry*, 2003, 75(18): 4818-4826.

[77] Old W M *et al.* Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell Proteomics*, 2005, 4(10): 1487-1502.

[78] Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.*, 2006, 78(3): 779-787.

[79] Chen W W *et al.* New algorithm for label-free protein quantification. In *ASMS*, Philadelphia, USA, May 31-June 4, 2009, Session MPB: Bioinformatics: Quantification, Poster, No. 043.

[80] Andreev V P, Li L, Cao L, Gu Y, Rejtar T, Wu S L, Karger B L. A new algorithm using cross-assignment for label-free quantitation with LC/LTQ-FT MS. *Journal of Proteome Research*, 2007, 6(6): 2186-2194.

[81] Lee T, Singh R, Yen TY, Macher B. An algorithmic approach to automated high-throughput identification of disulfide connectivity in proteins using tandem mass spectrometry. In *Proc. Computational System Bioinformatics Conference*, San Diego, USA, Aug. 13-17, 2007, pp.41-51.

[82] Ng J, Bandeira N, Liu W T, Ghassemian M, Simmons T L, Gerwick W H, Linington R, Dorrestein P C, Pevzner P A. Dereplication and de novo sequencing of nonribosomal peptides. *Nature Methods*, 2009, 6(8): 596-599.

[83] Zhang N *et al.* ProbIDtree: An automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics* 2005, 5(16): 4096-4106.

[84] Kelleher N L, Lin H Y, Valaskovic G A, Aaserud D J, Fridriksson E K, McLafferty F W. Top down versus bottom up protein characterization by tandem high-resolution mass spectrometry. *Journal of the American Chemistry Society*, 1999, 121(4): 806-812.

[85] Tang H *et al.* A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics*, 2006, 22(14): e481-e488.

[86] Alves P, Arnold R J, Novotny M V, Radivojac P, Reilly J P, Tang H. Advancement in protein inference from shotgun proteomics using peptide detectability. In *Proc. Pac. Symp. Biocomput.*, Maui, USA, Jan. 3-7, 2007, pp.409-20.

[87] Håkansson K *et al.* Combined electron capture and infrared multiphoton dissociation for multistage MS/MS in a Fourier transform ion cyclotron resonance mass spectrometer. *Anal. Chem.*, 2003, 75(13): 3256-3262.

[88] Nuno Bandeira, Jesper V Olsen, Matthias Mann, Pavel A Pevzner. Multi-spectra peptide sequencing and its applications to multistage mass spectrometry. *Bioinformatics*, 2008, 24(13): i416-i423.

[89] Xie M, Ma B. MSPack — Mass spectrometry data compression software. In *Proc. the 54th ASMS Conf. Mass Spectrometry*, Seattle, USA, May 28-June 1, 2006, Session: Computer Applications, Poster, No. 071.

[90] Miguel A C, Kearney-Fischer M, Keane J F, Whiteaker J, Feng L C, Paulovich A. Near-lossless compression of mass spectra for proteomics. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Honolulu, USA, April 15-20, 2007, pp.I369-I372.

[91] Meek J L. Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. *Proc. Natl. Acad. Sci. USA*, 77(3): 1632-1636.

[92] Strittmatter E F *et al.* Application of peptide LC retention time information in a discriminant function for peptide identification by tandem mass spectrometry. *Journal of Proteome Research*, 2004, 3(4): 760-769.

[93] Henzel W J, Billeci T M, Stults J T, Wong S C, Grimley C, Watanabe C. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. USA*, 1993, 90(11): 5011-5015.

[94] Du P, Kibbe W A, Lin S M. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 2006, 22(17): 2059-2065.

[95] Katajamaa M, Orešič M. Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics*, 2005, 6: 179.

[96] Nagalla S R *et al.* Proteomic analysis of maternal serum in down syndrome: Identification of novel protein biomarkers. *Journal of Proteome Research*, 2007, 6(4): 1245-1257.

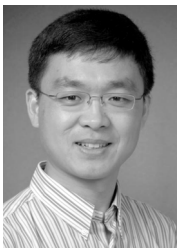
[97] Issaq H J, Veenstra T D, Conrads T P, Felschow D. The SELDI-TOF MS approach to proteomics: Protein profiling and biomarker identification. *Biochemical and Biophysical Research Communications*, 2002, 292(3): 587-592.

[98] Hancock W S, Wu S L, Shieh P. The challenges of developing a sound proteomics strategy. *Proteomics*, 2002, 2(4): 352-359.

[99] Steen H, Mann M. The ABC's (and XYZ's) of peptide sequencing. *Nature Reviews Molecular Cell Biology*, 2004, 5(9): 699-711.

[100] Snyder A P. Interpreting Protein Mass Spectra: A Comprehensive Resource. The American Chemical Society and Oxford University Press, 2000.

[101] Kinter M, Sherman N E. Protein Sequencing and Identification Using Tandem Mass Spectrometry. John Wiley & Sons Inc., 2000.



Bin Ma is an associate professor and university research chair in David R. Cheriton School of Computer Science at University of Waterloo. He received his Ph.D. degree from Beijing University in 1999. During 2000∼2008 he worked at University of Western Ontario as assistant professor, associate professor, and Canada research chair. He received the Ontario PREA Award in 2003 and Ontario Premier's Catalyst Award for Best Young Innovator in 2009.