Key Computational Problems in Tandem Mass Spectrometry Protein Identification (Postprint)

Authors: Fu Yan, He Simin, Sun Ruixiang, Wang Leheng

Date: 2017-03-09T00:00:00+00:00

Abstract

Protein identification is a fundamental problem in proteomics research, and searching protein sequence databases using tandem mass spectrometry is currently the most successful and widely used method for protein identification. Protein identification software is essentially an information retrieval system that shares common characteristics with other retrieval systems. However, compared with text or multimedia retrieval, it has very distinctive features. For instance, reliability assessment of retrieval results is an indispensable step in protein identification, which is often unnecessary for other retrieval problems. This paper reviews the key computational issues in protein identification search engines and their research progress, including database search matching scoring, statistical reliability assessment of identification results, protein modification identification, etc., and provides a brief introduction to our self-developed protein identification search engine pFind.

Full Text

Preamble

Vol. 8 No. 1 Information Technology Letters Vol.8 No.1
Key Computational Problems in Tandem Mass Spectrometry Protein Identification
Fu Yan, He Simin, Sun Ruixiang, Wang Leheng

Abstract

Protein identification is a fundamental problem in proteomics research, and searching protein sequence databases using tandem mass spectrometry (MS/MS) represents the most successful and widely used approach. Protein identification software is essentially an information retrieval system sharing common characteristics with other retrieval systems, yet it possesses unique features. Notably,

reliability assessment of search results is an indispensable step in protein identification, whereas it is often unnecessary for other retrieval problems. This paper reviews key computational issues in protein identification search engines and their research progress, including database search scoring, statistical evaluation of identification reliability, and protein modification identification. We also provide a brief introduction to pFind, a protein identification search engine developed by our own team.

Keywords: bioinformatics; protein identification; mass spectrometry; information retrieval; pFind

Introduction

In February 2001, the Human Genome Project (HGP) consortium and the American Celera company published the draft human genome sequence and preliminary analysis results in *Nature* and *Science*, respectively. The near-completion of human genome sequencing marked the arrival of the post-genomic era, as life science research sought new frontiers. In April 2001, the Human Proteome Organization (HUPO) was established in the United States to promote international collaborative proteome research. Subsequently, various proteome initiatives were launched, including the human plasma proteome project led by the United States, the human liver proteome project led by China, the human brain proteome project led by Germany, and many others. Concurrently, proteome research on other organisms has been extensively conducted worldwide [1]. The Chinese government has designated protein science as one of four major scientific programs in the "National Medium- and Long-Term Plan for Science and Technology Development," establishing it as a research priority for life sciences from 2006 to 2020.

The term "proteome"was first coined by Wilkins et al. in 1994 to describe the protein complement of the genome. A proteome refers to the entire set of proteins expressed by a biological cell, tissue, or organ at a given time and under given conditions. Proteomics, as the study of the proteome, has the fundamental task of determining the status of all proteins within a specific organism, including their expression, quantification, modifications, and mutations. Proteins are biological macromolecules polymerized from amino acids, and a protein' s amino acid sequence uniquely determines its identity. After translation from DNA via messenger RNA (mRNA), most proteins undergo chemical modifications at specific amino acids to achieve their biological activity. Therefore, identifying protein sequences and characterizing post-translational modifications are crucial for systematically understanding key biological knowledge such as protein structure, function, and evolutionary relationships.

Mass spectrometry is currently the mainstream technology for large-scale protein identification, offering advantages in sensitivity, throughput, and accuracy [2]. In typical bottom-up proteomics strategies, protein samples are enzymatically digested into peptide mixtures, which are then analyzed by liq-

uid chromatography-mass spectrometry to generate tandem mass spectra. Reconstructing peptide sequences from tandem mass spectra represents the core computational problem in protein identification. The most successful and commonly used solution involves searching protein sequence databases with tandem mass spectra: database sequences are theoretically digested and fragmented, predicted spectra are generated, and these are matched against experimental spectra to identify peptide sequences and, consequently, entire proteins. Protein identification based on sequence database searching is essentially an information retrieval problem, with the core computational challenge being the peptide-spectrum match scoring algorithm. Simultaneously, to obtain correct identification results, protein identification systems must statistically evaluate the reliability of search results—a step that is often unnecessary in other retrieval problems. Protein modifications present even greater challenges to the speed and accuracy of protein identification retrieval systems. This review focuses on these key computational issues in protein identification, including database search scoring, reliability assessment, and modification identification, following a brief introduction to the relevant biochemical background.

1.1 Proteins and Peptides

Proteins are the fundamental material basis of life, widely present in various biological tissues and cells, and constitute the most important component of biological cells. As a class of essential biological macromolecules, proteins serve as the primary carriers of structure and function in organisms. The human body contains over 100,000 distinct proteins with diverse structures and functions. However, all proteins are composed of amino acid molecules. The general formula of an amino acid molecule is shown in Figure 1 [Figure 1: see original paper]. An amino acid consists of an -carbon atom attached to a carboxyl group (COOH), an amino group (NH—), a hydrogen atom (H), and a side chain group R. Different amino acids have different side chain groups. The carboxyl group of one amino acid can condense with the amino group of another amino acid to form an amide bond (peptide bond), as shown in Figure 2 [Figure 2: see original paper]. Multiple amino acids connected sequentially by peptide bonds form a chain-like molecule called a peptide, with the amino terminus designated as the N-terminus and the carboxyl terminus as the C-terminus, as shown in Figure 3 [Figure 3: see original paper]. A peptide composed of two amino acids is called a dipeptide; peptides with 2-10 amino acids are called oligopeptides; those with more than 10 amino acids are called polypeptides. Polypeptides with molecular weights above 10 kDa are called proteins.

The vast majority of proteins are composed of 20 common amino acids, though some proteins contain hundreds of uncommon amino acids and non-peptide components (called ligands or prosthetic groups). A protein' s amino acid sequence is termed its primary structure. Proteins can fold to form secondary and tertiary structures. The primary structure—the amino acid sequence—uniquely determines a protein' s identity. The protein identification problem discussed

in this paper refers to the identification of protein sequences.

1.2 Mass Spectrometry Technology

Initially, protein sequence identification primarily employed the manual Edman degradation method, which was highly inefficient. The development of mass spectrometry (MS) opened new avenues for protein sequence identification [3, 4].

The basic principle of mass spectrometry is not particularly complex. In mass spectrometric analysis, analyte particles are first ionized and then passed through appropriate electromagnetic fields. Since ions with different mass-to-charge ratios respond differently to electromagnetic fields, they can be separated and detected based on trajectory and time. Ion intensities are also detected and recorded, producing mass spectral data with mass-to-charge ratio on the x-axis and ion intensity on the y-axis. A mass spectrometer consists of four major components: a sample introduction system, an ion source, a separation system, and a detection system, each with multiple implementation methods. Figure 4 [Figure 4: see original paper] illustrates the basic components of a mass spectrometer.

The history of mass spectrometry dates back to the late 19th century. In 1899, Joseph John Thomson invented the first parabolic mass spectrometry device. With technological improvements, mass spectrometers were widely applied to inorganic and organic compound analysis by the late 1950s. The 1950s-1980s represented a prosperous period for mass spectrometry. Revolutionary developments occurred in the 1980s-1990s, primarily involving liquid chromatography-mass spectrometry coupling and two soft ionization techniques: electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI). In the 1980s, John Fenn and Koichi Tanaka independently developed ESI and MALDI mass spectrometry techniques for biomacromolecule analysis, for which they shared the 2002 Nobel Prize in Chemistry with Kurt Wüthrich. Since the 1990s, mass spectrometry has been deeply applied in life sciences.

In mass spectrometry-based protein analysis, protein samples are first hydrolyzed by selected proteases to form peptides. Peptide ions with different mass-to-charge ratios are separated and detected by the mass spectrometer to obtain a primary mass spectrum (MS1). These peptide ions can be further fragmented to produce fragment ions, which are separated and detected to generate tandem mass spectra (MS/MS). Protein identification methods using mass spectrometry therefore fall into two categories.

The first category is based on primary mass spectra and is called peptide mass fingerprinting. These methods search known protein databases, simulate protein hydrolysis in silico to generate theoretical primary mass spectra, and compare theoretical with experimental primary mass spectra, ranking results by match quality. Systems employing this approach include MOWSE [5], Mascot [6], ProFound [7], PeptIdent [8], and MS-Fit [9]. Peptide mass fingerprinting is

suitable for samples containing single proteins or simple mixtures. However, its limitations include significant errors due to protein mixture contamination, incomplete enzymatic digestion, residue modifications (amino acid residues refer to amino acids minus a water molecule), mass accuracy, and other factors, often leading to incorrect search results.

The second category is based on tandem mass spectrometry. These methods first determine peptide amino acid sequences accurately using tandem mass spectrometry (MS/MS), then identify protein sequences through peptide sequences. This approach can be used to identify complex protein mixtures or verify peptide mass fingerprinting results and represents the most commonly used and effective mainstream method today, which we introduce in detail below.

In typical liquid chromatography-tandem mass spectrometry experiments, protein samples are proteolytically digested into peptide mixtures, which are separated by liquid chromatography and ionized. In the mass spectrometer, peptide ions with specific mass-to-charge ratios are selected and filtered, then fragmented under energy bombardment such as collision-induced dissociation (CID) [10] or electron transfer dissociation (ETD) [11]. During fragmentation, cleavage of three types of peptide bonds generates six major fragment series: N-terminal a, b, c fragments and C-terminal x, y, z fragments, as shown in Figure 5 [Figure 5: see original paper]. Fragments may lose neutral water or ammonia molecules [12]. Fragment ions retaining the parent ion charge, unfragmented parent ions, contaminants, and other fragmentation products are detected. Under low-energy fragmentation, generally only one peptide bond breaks per peptide ion, producing mainly a, b, and y-type fragment ions. Measuring the intensities of ions with different mass-to-charge ratios creates peaks in the tandem mass spectrum. Figure 6 [Figure 6: see original paper] shows an example tandem mass spectrum.

Identifying peptide amino acid sequences from tandem mass spectra is the central problem in protein identification. Three computational approaches exist for peptide sequence identification from tandem mass spectra. The most common is the database search method, as described in references [6, 9, 13-18]. In this approach, protein sequences in the database are theoretically hydrolyzed and fragmented to generate theoretical tandem mass spectra, which are compared with experimental spectra to find the peptide sequence that produced the experimental spectrum. This paper focuses on this method.

The second approach is de novo sequencing, which interprets tandem mass spectral data directly to identify peptide sequences without comparing against database sequences, as described in references [19-26]. When target sequences are not present in the database, database searching becomes ineffective, necessitating de novo sequencing. However, this method requires high-quality spectral data and good peptide fragmentation, limiting its widespread practical application. Nevertheless, de novo methods can provide important peptide sequence tags (short peptide fragments of several amino acids) to aid database searching even when complete sequences cannot be determined.

The third approach is sequence tag querying [27-31], which first obtains partial peptide sequence information from tandem mass spectra manually or automatically, then queries the database using these partial sequences to retrieve full peptide sequences. This hybrid approach has attracted increasing attention in recent years.

Several comprehensive review articles on tandem mass spectrometry-based peptide identification have been published recently [32-35].

3 Database Search Scoring Algorithms

In protein identification using tandem mass spectrometry, the problem reduces to the more fundamental peptide identification problem. The database search method is currently the most widely adopted peptide identification approach. Given an experimental tandem mass spectrum, scoring candidate peptides against this spectrum constitutes the core of peptide identification algorithms. Evaluating these scoring results—that is, recognizing correctly identified peptide sequences—is also an essential step.

"Peptide scoring" refers to rating the likelihood that a candidate peptide produced a given experimental tandem mass spectrum, thereby ranking all candidate peptides. In information retrieval terminology, the tandem mass spectrum is the query input, candidate peptides are the objects stored in the database, and the peptide scoring function is essentially the retrieval or ranking function. The scoring function' s purpose is to rank candidate peptides such that the sequence most likely to have produced the experimental spectrum appears at the top. Peptide scoring functions can be categorized into three types based on their construction: those based on spectral vector dot product, those based on probability, and those based on machine learning or pattern classification.

3.1 Spectral Dot Product-Based Peptide Scoring Algorithms

In spectral dot product (SDP) based peptide scoring algorithms, the degree of overlap between theoretical and experimental mass spectra serves as the candidate peptide' s score, and this overlap can be described by vector dot product operations. In SDP, theoretical and experimental spectra are represented as N-dimensional vectors, where N is the number of distinct mass values used. The components i_c and i_t can take 0/1 values or represent the ion intensity at the i-th mass value in the tandem mass spectrum. The SDP between experimental and theoretical tandem mass spectra is defined as the dot product of these vectors. If two spectrum vectors are identical, they should be parallel, and the vector dot product precisely reflects their degree of parallelism, making it suitable as a peptide match score.

Reference [36] used a spectral contrast angle based on SDP as a similarity measure for mass spectra. Reference [37] employed this metric to identify spectra

generated by the same peptide sequence. The early "Shared Peaks Count" (SPC) scoring method represents the simplest form of spectral vector dot product. SPC counts the number of matched fragment ions between theoretical and experimental spectra, corresponding to the case where i_c and i_t take 0/1 values in SDP. The Sonar MS/MS software [16, 38] is a typical representative using SDP as its peptide scoring function, representing spectra as vectors and directly computing the dot product as the score.

SEQUEST [14], one of the most widely used commercial peptide identification software packages, employs cross-correlation analysis to compare mass spectra, which is also fundamentally based on spectral vector dot product. SEQUEST first predicts a theoretical spectrum for a matching amino acid sequence according to specific rules, then processes the experimental spectrum appropriately to enable cross-correlation analysis that reflects fragment ion similarity. The cross-correlation between discrete signals of the experimental spectrum x(t) and theoretical spectrum y(t) is computed, where represents the displacement between two signals. The correlation function measures the similarity between two signals; if they are identical, the correlation function reaches its maximum at = 0. The SEQUEST scoring formula is defined based on this principle, where the Xcorr score is essentially the SDP minus the mean of SDPs at a series of displacements.

3.2 Probability-Based Peptide Scoring Algorithms

Another class of peptide scoring algorithms is probability-based, including Mascot [6], SCOPE [13], ProbID [18], PepSearch [39], and others [40]. Mascot is another widely adopted commercial protein identification software besides SEQUEST. However, the literature on Mascot does not specify its peptide scoring algorithm in detail. Generally, Mascot attempts to calculate the probability p that an experimental tandem mass spectrum is randomly generated by a candidate peptide, with the candidate peptide' s score being -log(p). Mascot' s probability scoring algorithm comprehensively considers peptide length distribution, missed cleavage probability, mass error distribution, and ion intensity information.

SCOPE is a scoring algorithm designed by Celera that uses a Bayesian model to calculate the posterior probability of each sequence given a spectrum. SCOPE simulates tandem mass spectrum generation through a two-step stochastic process: (1) generating peptide fragments according to probability distributions, and (2) generating spectra from fragments based on instrument measurement errors.

ProbID attempts to calculate the Bayesian posterior probability that an experimental tandem mass spectrum is randomly generated by a candidate peptide. However, ProbID' s calculated probability cannot be considered a true probability but rather a simple product of several factors, including the presence of immonium ions, whether peptide sequence cleavage sites meet enzyme speci-

ficity, matched and unmatched spectral peaks, and the matching of consecutive and complementary ions.

Although SCOPE and ProbID establish probabilistic models at different levels, they share the common feature that conditional probabilities used in calculations —such as probabilities of different ion types, error distributions, and ion intensity distributions—are specified or assumed based on expert experience and are therefore inaccurate.

Havilio et al. [40] and Dancik et al. [20] attempted to learn these probabilities from mass spectral data. Reference [20] learned the probability of detecting certain fragment types from data rather than assuming them a priori. The method in [40] generalizes the algorithm from [20], designing a series of scoring functions that can incorporate various experimental observations and prior theoretical knowledge about peptide fragmentation, including intensity correlations, fragment types, fragment masses, ratios of fragment mass to peptide mass, important ion types such as isotopes and multiply-charged fragments, which are often ignored by tandem mass spectrometry analysis software. The parameter learning process is automatic: the mass axis of the spectrum is divided into equal-width bins, and the probability of observed intensities is calculated for all ions falling within each bin. If fragments are assumed independent, all probabilities are multiplied; if correlated fragment pairs exist, joint probabilities of correlated ion pairs are calculated. The drawback of this approach is that mass spectral data are not annotated, and statistics are only roughly collected for all ions matching a particular spectral peak, making the statistical results inevitably inaccurate. The same problem exists in other mass spectral data mining work [41].

The probability-based peptide scoring algorithms described above model the probability of peptide fragmentation generating mass spectra. Another class of probability-based peptide scoring algorithms does not model the peptide fragmentation process but instead models the probability of matches between predicted ions and spectral peaks. For example, Sadygov and Yates used the hypergeometric distribution [42], Fridman et al. employed a more complex form of hypergeometric distribution [43], and Geer et al. adopted the Poisson distribution [44]. The advantage of these probability-based peptide scoring algorithms is their ability to provide probabilities of correct or random matches between candidate peptides and experimental spectra, but they insufficiently utilize theoretical spectrum prediction and spectral peak intensity information.

3.3 Machine Learning-Based Peptide Scoring Methods

Peptide identification can essentially be viewed as a two-class classification problem that categorizes candidate peptides as "correct" or "incorrect." In machine learning-based peptide scoring functions, various matching information between candidate peptides and experimental spectra is represented as feature vectors, and machine learning methods are applied to learn a scoring function from

training data with known sequences. Although machine learning methods for retrieval functions have long existed in information retrieval, they have only recently been applied to peptide identification. Various machine learning algorithms have been applied to classify SEQUEST search results, including support vector machines (SVM) [45], neural networks [46, 47], logistic regression [48], and ensemble methods such as boosting and random forest [49]. The benefit of using machine learning methods for peptide identification scoring is the ability to integrate many matching metrics, with each metric serving as a dimension of the pattern. How to fuse these metrics into a peptide scoring function becomes the task of the machine learning method, eliminating the need for user intervention. In fact, integrating numerous matching metrics into a single peptide scoring function has long been a challenge in peptide scoring design. Given the flexibility of machine learning methods, researchers later began using them to directly construct independent peptide scoring functions rather than merely post-processing existing software results [50, 51]. Essentially, peptide identification via database searching is a specific information retrieval problem, so directly constructing independent peptide scoring functions using machine learning methods is fundamentally a learning problem for retrieval or ranking functions in information retrieval. Such retrieval functions can be discriminative or serve only a ranking function.

4 Statistical Assessment of Identification Reliability

After database searching, each spectrum will have a top-scoring candidate peptide, but this peptide may not necessarily be correct. Many factors can cause this outcome: the peptide scoring algorithm is imperfect and makes errors, failing to rank the correct peptide first; the searched protein sequence database is incomplete and does not contain the target peptide sequence; the input mass spectral data consist entirely of noise without useful information; or the target peptide contains unexpected modifications or results from abnormal enzymatic cleavage. Therefore, after peptide scoring, it is necessary to determine whether the highest-scoring peptide is correct—that is, to assess the reliability of peptide scoring results and identify correct peptide identifications.

Early reliability assessment of peptide identification results used empirical threshold methods. As the name suggests, empirical threshold methods apply a threshold to raw scoring results based on experience, recognizing only peptides scoring above the threshold as correct identifications. A typical example is the SEQUEST software [14]. SEQUEST outputs two main scores: Xcorr and DeltCn. In the past, these two scores were widely used to filter SEQUEST peptide scoring results. For instance, a common filtering criterion required DeltCn greater than 0.1 and Xcorr greater than 1.9, 2.2, and 3.75 for singly, doubly, and triply charged peptides, respectively [52]. For SEQUEST, methods have also been developed to compute better scores in post-processing steps for filtering, such as RScore [53]. In probability-based peptide scoring algorithms, although the goal is to calculate the probability of true or random matches,

it has been noted that such probabilities cannot be accurately calculated objectively for various reasons. Therefore, probability-based peptide scoring methods typically still require specifying a threshold or using additional evaluation methods. Empirical threshold methods are simple and intuitive but have obvious drawbacks: threshold specification relies solely on experience without theoretical justification. As database size increases, the highest scores of incorrect candidate peptides also rise. Moreover, the reliability of results filtered by thresholds cannot be quantitatively estimated. Using empirical thresholds is an arbitrary practice; in reality, regardless of how high the peptide identification score, some uncertainty always remains. To effectively estimate identification reliability, statistical methods must be employed. Currently, the most widely used statistical reliability metrics are expectation values for single-spectrum identification and false discovery rates for multiple-spectrum identification.

4.1 Expectation Value Methods

The expectation value (commonly abbreviated as E-value) refers to the mean of a random variable. In bioinformatics sequence alignment problems [54-56], expectation values were first successfully applied; the BLAST sequence alignment program initially used E-values to measure the likelihood of sequence alignment scores occurring by chance (http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html). In peptide identification, given a spectrum and n random candidate peptides, the random variable of interest is the number of incorrect candidate peptides with scores at least x. In meaning, the expectation value for score x is the expected number of candidate peptides that can achieve scores equal to or exceeding x by random chance. For example, if a candidate peptide' s score has an expectation value of 10 in a database search, this means that on average, one hundred thousand such searches would be required to randomly obtain a score equal to or greater than this value by chance. Therefore, theoretically, if a candidate peptide' s expectation value exceeds 1, that candidate peptide can be excluded because even in a completely random scenario, on average one candidate peptide would achieve the same or higher score.

The probability distribution of random scores determines the calculation method for expectation values. Based on different sources of probability distributions, expectation value calculation methods can be divided into three categories: empirical fitting, theoretical modeling, and exhaustive calculation. Empirical fitting estimates probability distributions by fitting actual score distribution data. Theoretical modeling derives score random distributions theoretically by assuming certain random probability models. Exhaustive calculation obtains the true score distribution by enumerating all possible candidate peptides. Search engines that calculate expectation values through empirical fitting include Sonar [38], X!Tandem [57], pFind [58-60], and RAId_DbS [61]. Those using theoretical modeling include Mascot [6] and OMSSA [44]. Exhaustive calculation was

recently proposed by Kim et al. [62]. Each method has advantages and disadvantages: empirical fitting works for any scoring function but requires sufficient candidate peptide scale to fit score distributions and appropriate distribution form assumptions; theoretical modeling can calculate expectation values for any given peptide sequence but only applies to probability-based scoring functions, with accuracy depending on the probability model' s accuracy; exhaustive calculation can directly compute the true score distribution but only applies to additive scoring functions and has high computational complexity. Empirical fitting is currently the most commonly used and successful expectation value calculation method, briefly introduced below.

In peptide identification database searching, given a spectrum, at most one candidate peptide can be correct, so almost all candidate peptides can be considered incorrect. The empirical expectation value calculation method directly uses all candidate peptide scores from a single database search to fit the random score distribution [38]. Assuming incorrect candidate peptide scores follow a certain probability distribution and estimating parameters from empirical data, the probability of achieving a score not less than x can be inferred, and multiplying by the number of candidate peptides participating in scoring yields the expectation value.

Reference [38] assumes that random scores x follow an extreme value distribution. Based on this assumption, P(x) and x have an approximate log-linear relationship in the high-score region, where coefficients c and c can be estimated from the high-end portion of the candidate peptide score distribution in a single search. Once P(x) is estimated, expectation values can be calculated for any candidate peptide score. The X!Tandem software employs this method to calculate expectation values [57].

However, not all scoring algorithms conform to this distribution assumption. In fact, recent discussions have emerged regarding the applicability of the above expectation value calculation method to peptide identification and how to fit score distributions [63, 64]. Although expectation values have clear definitions, the values output by various software lack strict absolute meaning because assumptions and approximations are always made in calculations. Indeed, different software employs different assumptions and implementation methods, making their output expectation values not directly comparable [65]. This does not diminish the superiority of expectation values. Compared with raw scores, expectation values consider score distribution and database size, thus serving as normalized relative scores. Our task is to make expectation value estimates as close to true values as possible.

4.2 False Discovery Rate Methods

The expectation value method described above enables reliability assessment of individual peptide identification results. However, in proteomics experiments, thousands of spectra are typically identified simultaneously. For large numbers

of peptide identification results filtered by a given peptide scoring (or expectation value) threshold, overall reliability assessment is required. Currently, false discovery rate (FDR) calculation is commonly used to evaluate the reliability of peptide identification result populations [66]. FDR calculation can be divided into two categories: one estimates posterior error probabilities by fitting certain score distribution models; the other introduces decoy sequence libraries as controls. From a machine learning perspective, the former is unsupervised, while the latter is supervised.

PeptideProphet is the most representative and successful model-based false discovery rate estimation method [67]. Each proteomics experiment generates numerous tandem mass spectra, and through database searching, each spectrum is assigned a top-scoring candidate peptide. The PeptideProphet method is based on analyzing the distribution of these top scores. In PeptideProphet, several scores provided by SEQUEST are first linearly combined into a discriminant score, with incorrect matches assumed to follow a gamma distribution and correct matches assumed to follow a Gaussian distribution. For each specific experiment, PeptideProphet uses the Expectation-Maximization (EM) algorithm to estimate parameters of the discriminant score distribution, thereby finding a score threshold that optimally separates correct from incorrect matches while estimating the error rate.

The more widely used false discovery rate calculation method today is based on decoy sequence library searching. Decoy sequences are defined as sequences that definitely do not contain target proteins; searching these sequences necessarily yields incorrect results and thus can serve as negative samples for estimating false discovery rates. Decoy sequences are typically reversed, randomly generated, or generated according to certain probability models from target sequences. Decoy sequences must not contain target sequences while possessing "characteristics" of target sequences so that estimated false discovery rates are accurate. Currently, the method of using reversed sequence libraries to estimate false positive rates is simple, practical, and has been widely adopted by the proteomics community, becoming a standard for proteomics data false positive analysis [68, 69]. The steps for estimating false discovery rates using the reversed library method are as follows:

1. Reverse sequences in the target protein database to obtain reverse sequences, then combine reverse and forward sequences to form a target-decoy database;
2. Search the target-decoy database using the same peptide scoring and filtering methods as used for the target database;
3. Estimate the false discovery rate of positive peptide identification results: Let N_f denote the number of positive peptide identifications with sequences from forward protein sequences, and N_r denote the number of positive peptide identifications with sequences from reverse protein sequences (if peptide scoring and filtering methods are effective, N_r should generally be small). The false discovery rate of peptide identification re-

sults is:

The above false discovery rate calculation formula is based on the assumption that correctly identified peptides must come from forward protein sequences, while incorrectly identified peptides have equal probability of coming from forward or reverse protein sequences. Since forward and reverse sequences have the same length, we can assume that forward sequence peptide identification results contain the same number of false positive identifications as the number of reverse sequence peptide identifications.

5 Protein Modification Identification

After translation from mRNA, proteins may have functional groups added to certain amino acids, or other proteins or peptides attached, or the chemical properties or structures of amino acids may be altered. This process is called chemical modification. Since it occurs after translation, it is termed post-translational modification (PTM). PTMs can change amino acid chemical properties, cause protein structural changes, and regulate protein activity and function. PTMs are ubiquitous in organisms, with the vast majority of proteins containing one or more modifications. Studying PTMs is crucial for elucidating protein functions and explaining mechanisms of major diseases [70, 71]. Human proteome research indicates that for tryptic peptides expressed at relatively high levels (>1%), nearly every amino acid has some modified form [72]. In addition to in vivo modifications, many modifications are inevitably introduced during sample processing [73]. Protein modification types are numerous; as of June 24, 2009, the Unimod modification database contained 590 entries. Mass spectrometry-based proteomics provides effective analytical tools for large-scale PTM research [74-76]. Currently, identifying modified proteins using tandem mass spectrometry data has become one of the core and frontier issues in proteomics research.

A common tandem mass spectrometry-based method for identifying modified proteins specifies variable modification types during database searching, considering both modified and unmodified states when generating candidate peptides and evaluating all possible combinations when multiple potential modification sites exist in a candidate peptide. This approach accounts for the dynamic nature of protein PTMs (the same amino acid site may or may not be modified), but it is impractical to consider too many modification types during database searching because the search space would explode combinatorially, drastically reducing search speed and increasing false positive results. Existing search engines such as SEQUEST and Mascot typically allow no more than about 10 variable modification types, which clearly cannot meet practical needs. In most cases, researchers have limited knowledge about modification types present in protein samples and rely mainly on empirical speculation. Most often, oxidation on methionine is the only variable modification specified in database searching, potentially missing other modifications present in the sample. Consequently, many spectra from modified peptides remain uninterpreted. This approach of specifying a limited number of modification types is called restricted modifi-

cation identification, which suffers from serious problems including blindness, combinatorial explosion of search space, and inability to discover novel modification types.

The large protein sequence database and numerous variable modification types together cause combinatorial explosion in the candidate peptide space. If searching is limited to smaller protein databases, more variable modification types can be considered. A common approach is two-stage refined database searching [77-80]. In the first search, the entire protein database is searched with minimal variable modification types and the strictest enzymatic cleavage rules. In the second search, proteins identified in the first search are used to construct a small database for refined searching with more variable modification types, relaxed cleavage rules, and sequence mutations. This method was first used in the MASCOT software [77]. Its basic assumption is that each protein present in the experimental sample can be identified by at least one peptide in the first search [78]. This coarse-to-fine search strategy can significantly improve search speed and the number of variable modification types considered. Simultaneously, by considering more modification types, cleavage rules, and sequence mutations, more peptides and their variants can be identified, achieving two goals at once. However, if the above assumption is not satisfied, some peptides and proteins and their variants may be missed. Moreover, this two-stage search still requires users to specify a list of variable modification types and remains a restricted modification identification approach, unable to detect unknown or unlisted modification types.

To enable database searching for unknown or unexpected modification types, the most direct approach is to relax the peptide-spectrum match mass constraint, allowing correct candidate peptide sequences to enter the search space for matching with experimental spectra. This undoubtedly increases computational load, which we discuss later. More importantly, the challenge is how to match experimental spectra containing modifications with unmodified candidate peptide sequences so that correct candidate peptides can be discovered and modification masses and sites determined. MS-Alignment is the earliest and most famous method of this type [17, 81, 82]. MS-Alignment aligns theoretical and experimental mass spectra in a manner similar to sequence alignment in genomics, allowing any modifications to occur.

However, the MS-Alignment algorithm has several limitations: (1) The computational complexity of finding optimal matches between experimental spectra and peptide sequences is very high, making data analysis extremely slow and limiting practical applications to searching very small protein sequences; (2) To enable dynamic programming alignment of theoretical and experimental spectra, simple scoring functions must be used, reducing spectral alignment accuracy; (3) The reliability of search results is low—a recent study by Chen et al. [83] showed that MS-Alignment severely underestimates the false discovery rate of results.

Spectra generated by modified and unmodified forms of the same peptide sequence represent a typical type of related spectra. In reality, due to the dynamic

nature of modifications, both modified and unmodified forms of the same peptide often coexist simultaneously. This provides another clue for unrestricted PTM detection—identifying related spectrum pairs generated by modified and unmodified peptides. The spectral network algorithm is based on this principle, detecting PTMs and mutations by identifying related spectrum pairs [84]. However, the spectral network algorithm uses MS-Alignment to calculate spectral similarity and thus faces the same enormous computational challenges. Additionally, if a peptide undergoes a modification but its unmodified form is absent or undetectable by the mass spectrometer, or if the modified-unmodified spectrum pair similarity is insufficient, methods based on spectrum pairs become inapplicable.

Unrestricted PTM detection, as the forefront of proteomics research, has attracted increasing attention from researchers. Some have proposed strategies that first use de novo sequencing to obtain peptide sequence fragments, then locate proteins through sequence matching to further determine modification masses and sites [31, 85]. However, this strategy heavily depends on spectral signal quality, and de novo sequencing itself remains an unsolved problem [33]. Savitski et al. [86] proposed a modification detection method combining two peptide fragmentation modes (ECD and CAD), but it is only applicable to this special mass spectrometry operation mode. In summary, research on unrestricted PTM detection is still in its infancy, and no mature solution currently exists.

6 The pFind Protein Identification System Developed by Our Institute

Since 2002, the Bioinformatics Research Group at the Institute of Computing Technology, Chinese Academy of Sciences, has been researching protein identification algorithms and software based on mass spectrometry data. We have proposed a series of innovative algorithms and technologies in mass spectral data signal processing, theoretical spectrum prediction, spectral similarity measurement, protein PTM detection, database indexing, and search engine design. Based on these developments, we independently developed pFind (http://pfind.ict.ac.cn), China' s first and only software system for large-scale identification of proteins and their post-translational modifications. pFind uses Kernel Spectral Dot Product (KSDP) as its core matching scoring algorithm, representing a nonlinear extension of traditional dot product scoring [58]; it accelerates database searching through database indexing, search workflow optimization, and parallel computing techniques [87, 88]; and it uses spectral clustering methods to rapidly detect potential modification types from mass spectral data [89]. In addition to the core search engine, pFind includes various supporting software for result analysis, spectrum annotation, and database processing [59, 60], totaling over 210,000 lines of code. A parallel version of pFind has also been deployed. Currently, the pFind system has reached the level of international mainstream commercial software such as SEQUEST and Mascot in terms of accuracy and

speed. The pFind system has published over 10 academic papers in renowned domestic and international journals and conferences [51, 58-60, 87-99], receiving recognition and citations from international peers; it has applied for 8 patents, with 3 already granted; and it has applied for 12 software copyrights.

The pFind system has been demonstrated and applied in over 10 proteomics research institutions in China, including the Shanghai Institute of Biochemistry and Cell Biology, Institute of Biophysics, Institute of Genomics, Institute of Zoology, and Institute of Chemistry of the Chinese Academy of Sciences, as well as the Beijing Proteome Research Center, National Institute of Biological Sciences, Beijing North Center for Human Genome Research, Institute of Basic Medical Sciences and Cancer Institute of Peking Union Medical College, Shanghai Bioinformatics Center, and Fudan University, with a total of 62 pFind system installations. In 2008, pFind participated in the international protein identification data analysis evaluation organized by ABRF (Association of Biomolecular Resource Facilities), demonstrating strong competitiveness in identification accuracy and false positive rate control, and began to gain international recognition. The Beijing Proteome Research Center used the pFind system to identify core fucosylation modifications, successfully identifying over 100 core fucosylation sites from human liver cancer plasma samples—the largest number reported to date—representing significant progress for subsequent cancer early marker discovery research. This collaborative achievement was published in 2009 in the internationally renowned proteomics journal *Molecular & Cellular Proteomics* [100], marking the first time pFind software was successfully applied to a real biological problem and recognized by a top-tier international academic journal.

7 Conclusion

Proteomics research is flourishing, and protein identification and modification identification based on mass spectrometry data are among its most important problems. This paper introduced key computational challenges facing protein identification search engines from several perspectives: database retrieval scoring, reliability assessment of search results, and modification identification. These problems currently lack satisfactory solutions, and some have become bottlenecks in proteomics data analysis. This situation presents both major challenges for the computing field and opportunities for computational technology to solve problems in life sciences. Few research teams in China work in this area, but they have already reached the forefront of the field. With enhanced confidence and redoubled efforts, the pFind protein identification system will undoubtedly play a greater role in future proteomics research.

References

[1] Qian X, He F: *Proteomics: Theory and Methods.* Beijing: Science Press; 2003.

[2] Aebersold R, Mann M: Mass spectrometry-based proteomics. *Nature* 2003, 422:198-207.
[3] Yang P, Qian X, Sheng L: *Mass Spectrometry Technology and Methods for Biological Analysis*. Beijing: Science Press; 2003.
[4] Xia Q, Zeng R: *Protein Chemistry and Proteomics*. Beijing: Science Press; 2004.
[5] Pappin DJ, Hojrup P, Bleasby AJ: Rapid identification of proteins by peptide-mass fingerprinting. *Curr Biol* 1993, 3(6):327-332.
[6] Perkins DN, Pappin DJ, Creasy DM, Cottrell JS: Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999, 20:3551-3567.
[7] Zhang W, Chait BT: ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal Chem* 2000, 72(11):2482-2489.
[8] Wilkins MR, Williams KL: Cross-species protein identification using amino acid composition, peptide mass fingerprinting, isoelectric point and molecular mass: a theoretical evaluation. *J Theor Biol* 1997, 186(1):7-15.
[9] Clauser KR, Baker P, Burlingame AL: Role of accurate mass measurement (± 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal Chem* 1999, 71:2871-2882.
[10] Wells JM, McLuckey SA: Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol* 2005, 402:148-185.
[11] Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF: Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A* 2004, 101(26):9528-9533.
[12] Sun S, Yu C, Qiao Y, Lin Y, Dong G, Liu C, Zhang J, Zhang Z, Cai J, Zhang H et al: Deriving the probabilities of water loss and ammonia loss for amino acids from tandem mass spectra. *J Proteome Res* 2008, 7(1):202-208.
[13] Bafna V, Edwards N: SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* 2001, 17:S13-S21.
[14] Eng JK, McCormack AL, Yates JR, III: An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 1994, 5:976-989.
[15] Fenyö D, Qin J, Chait BT: Protein identification using mass spectromic information. *Electrophoresis* 1998, 19:998-1005.
[16] Field HI, Fenyo D, Beavis RC: RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics* 2002, 2:36-47.
[17] Pevzner PA, Dancik V, Tang CL: Mutation-tolerant protein identification by mass-spectrometry. *J Comput Biol* 2000, 7:777-787.
[18] Zhang N, Aebersold R, Schwikowski B: ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* 2002, 2:1406-1412.
[19] Taylor JA, Johnson RS: Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry* 1997, 11:1067-1075.

[20] Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA: De novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* 1999, 6:327-342.
[21] Ma B, Zhang KZ, Hendrie C, Liang CZ, Li M, Doherty-Kirby A, Lajoie G: PEAKS: powerful software for peptide de novo sequencing by MS/MS. *Rapid Commun Mass Spectrom* 2003, 17:2337-2342.
[22] Frank A, Pevzner P: PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* 2005, 77:964-973.
[23] Bern M, Goldberg D: EigenMS: de novo analysis of peptide tandem mass spectra by spectral graph partitioning. In: *Ninth Annual International Conference on Research in Computational Molecular Biology*: 2005; 2005: 357-372.
[24] Baginsky S, Cieliebak M, Gruissem W, Klemann T, Liptak Z, Muller M, Penna P: AUDENS: a tool for automated peptide de novo sequencing. *Journal of Proteome Research* 2005, 10:1768-1774.
[25] Chen T, Kao MY, Tepel M, Rush J, Church J: A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* 2001, 8:325-337.
[26] Zhang Z: De novo peptide sequencing based on a divide-and-conquer algorithm and peptide tandem spectrum simulation. *Anal Chem* 2004, 76:6374-6383.
[27] Mann M, Wilm M: Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem* 1994, 66:4390-4399.
[28] Sunyaev S, Liska AJ, Golod A, Shevchenko A, Shevchenko A: MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. *Anal Chem* 2003, 75:1307-1315.
[29] Frank A, Tanner S, Pevzner P: Peptide sequence tags for fast database search in mass-spectrometry. In: *Ninth Annual International Conference on Research in Computational Molecular Biology*: 2005; 2005:
[30] Tabb DL, Saraf A, Yates JR, III: GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal Chem* 2003, 75:6415-6421.
[31] Han Y, Ma B, Zhang K: SPIDER: software for protein identification from sequence tags containing de Novo sequencing error. In: *IEEE 2004 Computational Systems Bioinformatics Conference*: 2004; 2004:
[32] Johnson RS, Davis MT, Taylor JA, Patterson SD: Informatics for protein identification by mass spectrometry. *Methods* 2005, 35:223-236.
[33] Lu B, Chen T: Algorithms for de novo peptide sequencing via tandem mass spectrometry. *Drug Discovery Today: BioSilico* 2004, 2:85-90.
[34] Sadygov RG, Cociorva D, Yates JR, III: Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat Methods* 2004, 1:195-202.
[35] Nesvizhskii AI, Vitek O, Aebersold R: Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods* 2007, 4(10):787-797.
[36] Wan KX, Vidavsky I, Gross ML: Comparing similar spectra: from similarity index to spectral contrast angle. *J Am Soc Mass Spectrom* 2002,

13:85-88.
[37] Tabb DL, MacCoss MJ, Wu CC, Anderson SD, Yates JR, III: Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal Chem* 2003, 75:2470-2477.
[38] Fenyö D, Beavis RC: A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal Chem* 2003, 75:768-774.
[39] Sheng Q, Tang H, Xie T, Wang L, Ding D: A statistical method for peptide identification using tandem mass spectrometry. *Acta Biochim Biophys Sin* 2003, 35(8):734-740.
[40] Havilio M, Haddad Y, Smilansky Z: Intensity-based statistical scorer for tandem mass spectrometry. *Anal Chem* 2003, 75:435-444.
[41] Kapp EA, Schutz F, Reid GE, Eddes JS, Moritz RL, O' Hair RAJ, Speed TP, Simpson RJ: Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal Chem* 2003, 75:6251-6264.
[42] Sadygov RG, Yates JR, III: A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal Chem* 2003, 75:3792-3798.
[43] Fridman T, Razumovskaya J, Verberkmoes N, Hurst G, Protopopescu V, Xu Y: The probability distribution for a random match between an experimental-theoretical spectral pair in tandem mass spectrometry. *J Bioinform Comput Biol* 2005, 3:455-476.
[44] Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH: Open mass spectrometry search algorithm. *J Proteome Res* 2004, 3:958-964.
[45] Anderson DC, Li W, Payan DG, Noble WS: A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J Proteome Res* 2003, 2:137-146.
[46] Baczek T, Bucinski A, Ivanov AR, Kaliszan R: Artificial neural network analysis for evaluation of peptide MS/MS spectra in proteomics. *Anal Chem* 2004, 76 (6):1726-1732.
[47] Razumovskaya J, Olman V, Xu D, Uberbacher EC, VerBerkmoes NC, Hettich RL, Xu Y: A computational method for assessing peptide identification reliability in tandem mass spectrometry analysis with SEQUEST. *Proteomics* 2004, 4:961-969.
[48] Higdon R, Kolker N, Picone A, van Belle G, Kolker E: LIP index for peptide classification using MS/MS and SEQUEST search via logistic regression. *OMICS* 2004, 8(4):357-369.
[49] Ulintz PJ, Zhu J, Qin ZS, Andrews PC: Improved Classification of Mass Spectrometry Database Search Results Using Newer Machine Learning Approaches. *Molecular & Cellular Proteomics* 2006, 5:497-509.
[50] Liu J, Ma B, Li M: PRIMA: peptide robust identification from MS/MS spectra. In: *Third Asia-Pacific Bioinformatics Conference*: 2005; 2005: 181-190.

[51] Wang H, Fu Y, Sun R, He S, Zeng R, Gao W: An SVM Scorer for More Sensitive and Reliable Peptide Identification via Tandem Mass Spectrometry. In: *11th Pacific Symposium on Biocomputing*: 2006; 2006:
[52] Washburn MP, Wolters D, Yates JR, III: Large-scale analysis of the yeast proteome via multidimensional protein identification technology. *Nat Biotech* 2001, 19:242-247.
[53] Li F, Sun W, Gao Y, Wang J: RScore: A Peptide Randomicity Score For Evaluating MS/MS Spectra. *Rapid Communications in Mass Spectrometry* 2004, 18(14):1655-1659.
[54] Karlin S, Altschul S: Methods for assessing the statistical significance of molecular sequence features using general scoring schemes. *Proc Natl Acad Sci USA* 1990, 87:2264-2268.
[55] Karlin S, Altschul S: Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc Natl Acad Sci USA* 1993, 90:5873-5877.
[56] Pearson WR: Empirical statistical estimates for sequence similarity searches. *J Mol Biol* 1998, 276(1):71-84.
[57] Craig R, Beavis RC: TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004, 20:1466-1467.
[58] Fu Y, Yang Q, Sun R, Li D, Zeng R, Ling CX, Gao W: Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. *Bioinformatics* 2004, 20:1948-1954.
[59] Li D, Fu Y, Sun R, Ling C, Wei Y, Zhou H, Zeng R, Yang Q, He S, Gao W: pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry. *Bioinformatics* 2005, 21(13):3049-3050.
[60] Wang LH, Li DQ, Fu Y, Wang HP, Zhang JF, Yuan ZF, Sun RX, Zeng R, He SM, Gao W: pFind 2.0: a software package for peptide and protein identification via tandem mass spectrometry. *Rapid Commun Mass Spectrom* 2007, 21(18):2985-2991.
[61] Alves G, Ogurtsov AY, Yu YK: RAId_DbS: Peptide Identification using Database Searches with Realistic Statistics. *Biol Direct* 2007, 2:25.
[62] Kim S, Gupta N, Pevzner PA: Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J Proteome Res* 2008, 7(8):3354-3363.
[63] Segal MR: On E-values for tandem MS scoring schemes. *Bioinformatics* 2008, 24(14):1652-1653; author reply 1654.
[64] Giddings JKaM: In response to 'On E-value for tandem MS scoring schemes' . *Bioinformatics* 2008, 24(14):1654.
[65] Alves G, Ogurtsov AY, Wu WW, Wang G, Shen RF, Yu YK: Calibrating E-values for MS2 database search methods. *Biol Direct* 2007, 2:26.
[66] Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)* 1995, 57(1):289-300.
[67] Keller A, Nesvizhskii AI, Kolker E, Aebersold R: Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database Search. *Anal Chem* 2002, 74:5383-5392.

[68] Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP: Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res* 2003, 2(1):43-50.
[69] Elias JE, Gygi SP: Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 2007, 4(3):207-214.
[70] Uy R, Wold F: Posttranslational covalent modification of proteins. *Science* 1977, 198(4320):890-896.
[71] Walsh CT: *Posttranslational Modification of Proteins: Expanding Nature's Inventory*. Englewood (Colorado): Roberts & Company Publishers; 2005.
[72] Nielsen ML, Savitski MM, Zubarev RA: Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics. *Mol Cell Proteomics* 2006, 5(12):2384-2391.
[73] Hunyadi-Gulyás É, Medzihradszky KF: Factors that contribute to the complexity of protein digests. *Drug Discovery Today: TARGETS* 2004, 3(2, S1):3-10.
[74] Wilkins MR, Gasteiger E, Gooley AA, Herbert BR, Molloy MP, Binz PA, Ou K, Sanchez JC, Bairoch A, Williams KL et al: High-throughput mass spectrometric discovery of protein post-translational modifications. *J Mol Biol* 1999, 289(3):645-657.
[75] Mann M, Jensen ON: Proteomic analysis of post-translational modifications. *Nat Biotechnol* 2003, 21(3):255-261.
[76] Witze ES, Old WM, Resing KA, Ahn NG: Mapping protein post-translational modifications with mass spectrometry. *Nat Methods* 2007, 4(10):798-806.
[77] Creasy DM, Cottrell JS: Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics* 2002, 2(10):1426-1434.
[78] Craig R, Beavis RC: A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun Mass Spectrom* 2003, 17(20):2310-2316.
[79] Matthiesen R, Bunkenborg J, Stensballe A, Jensen ON, Welinder KG, Bauw G: Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V2.0. *Proteomics* 2004, 4(9):2583-2593.
[80] Chamrad DC, Korting G, Schafer H, Stephan C, Thiele H, Apweiler R, Meyer HE, Marcus K, Bluggel M: Gaining knowledge from previously unexplained spectra-application of the PTM-Explorer software to detect PTM in HUPO BPP MS/MS data. *Proteomics* 2006, 6(18):5048-5058.
[81] Pevzner PA, Mulyukov Z, Dancik V, Tang CL: Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Genome Res* 2001, 11:290-299.
[82] 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.
[83] Chen Y, Chen W, Cobb MH, Zhao Y: PTMap–a sequence alignment software for unrestricted, accurate, and full-spectrum identification of post-

translational modification sites. *Proc Natl Acad Sci U S A* 2009, 106(3):761-766.
[84] Bandeira N, Tsur D, Frank A, Pevzner PA: Protein identification by spectral networks analysis. *Proc Natl Acad Sci U S A* 2007, 104(15):6140-6145.
[85] Na S, Jeong J, Park H, Lee KJ, Paek E: Unrestrictive identification of multiple post-translational modifications from tandem mass spectrometry using an error-tolerant algorithm based on an extended sequence tag approach. *Mol Cell Proteomics* 2008, 7(12):2452-2463.
[86] Savitski MM, Nielsen ML, Zubarev RA: ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Mol Cell Proteomics* 2006, 5(5):935-948.
[87] Li D, Gao W, Ling CX, Wang X, Sun R, He S: IndexToolkit: an open source toolbox to index protein databases for high-throughput proteomics. *Bioinformatics* 2006, 22(20):2572-2573.
[88] Li Y, Chi H, Wang L-H, Wang H-P, Fu Y, Yuan Z-F, Li S-J, Liu Y-S, Sun R-X, Zeng R et al: Speeding up Tandem Mass Spectrometry Database Searching by Peptide and Spectrum Indexing. *Accepted by Rapid Communications in Mass Spectrometry* 2010.
[89] Fu Y, Jia W, Lu Z, Wang H, Yuan Z, Chi H, Li Y, Xiu L, Wang W, Liu C et al: Efficient discovery of abundant post-translational modifications and spectral pairs using peptide mass and retention time differences. *BMC Bioinformatics* 2009, 10 Suppl 1:S50.
[90] Fu Y, Sun R, Yang Q, He S, Wang C, Wang H, Shan S, Liu J, Gao W: A Block-Based Support Vector Machine Approach to the Protein Homology Prediction Task in KDD Cup 2004. *SIGKDD Explorations* 2004, 6:120-124.
[91] Fu Y, Yang Q, Ling CX, Wang H-P, Li D-Q, Sun R-X, Zhou H, Zeng R, Chen Y, He S-M et al: A Kernel-based Case Retrieval Algorithm with Application to Bioinformatics. *LNAI* 3157 2004:544-553.
[92] Zhang J, Gao W, Cai J, He S, Zeng R, Chen R: Predicting molecular formulas of fragment ions with isotope patterns in tandem mass spectra. *IEEE/ACM T Comp Biol Bioinfo* 2005, 2(3):217-230.
[93] Zhang JF, He SM, Cai JJ, Cao XJ, Sun RX, Fu Y, Zeng R, Gao W: Preprocessing of tandem mass spectrometric data based on decision tree classification. *Genomics Proteomics Bioinformatics* 2005, 3(4):231-237.
[94] Wang H, Fu Y, Sun R, He S, Zeng R, Gao W: pepReap: a support vector machine-based peptide identification algorithm. *Journal of Computer Research and Development* 2005, 42(9):1511-1518.
[95] Sun R, Fu Y, Li D, Zhang J, Wang X, Sheng Q, Zeng R, Chen Y, He S, Gao W: Computational proteomics research based on mass spectrometry technology. *Science in China Series E: Information Sciences* 2006, 36(2):222-234.
[96] Fu Y, Gao W, He S, Sun R, Zhou H, Zeng R: Mining tandem mass spectral data for more accurate mass error model for peptide identification. In: *12th Pacific Symposium on Biocomputing*: 2007; 2007:
[97] Zhang J, He S, Ling CX, Cao X, Zeng R, Gao W: PeakSelect: preprocessing tandem mass spectra for better peptide identification. *Rapid Commun Mass Spectrom* 2008, 22(8):1203-1212.

[98] Zhang J, Xu D, Gao W, Lin G, He S: Isotope pattern vector based tandem mass spectral data calibration for improved peptide and protein identification. *Rapid Commun Mass Spectrom* 2009, 23(21):3448-3456.
[99] Sun R, Dong M, Chi H, Yang B, Xiu L, Wang L, Fu Y, He S: Proteomics research based on electron capture dissociation/electron transfer dissociation tandem mass spectrometry. *Progress in Biochemistry and Biophysics* 2010, 37(1).
[100] Jia W, Lu Z, Fu Y, Wang HP, Wang LH, Chi H, Yuan ZF, Zheng ZB, Song LN, Han HH et al: A strategy for precise and large scale identification of core fucosylated glycoproteins. *Mol Cell Proteomics* 2009, 8(5):913-923.

Author Biographies:

Fu Yan, Associate Professor, Advanced Research Laboratory, Institute of Computing Technology, Chinese Academy of Sciences; yfu@ict.ac.cn
He Simin, Professor, Advanced Research Laboratory, Institute of Computing Technology, Chinese Academy of Sciences
Sun Ruixiang, Associate Professor, Advanced Research Laboratory, Institute of Computing Technology, Chinese Academy of Sciences
Wang Leheng, Engineer, Advanced Research Laboratory, Institute of Computing Technology, Chinese Academy of Sciences

Note: Figure translations are in progress. See original paper for figures.