Database Searching in Mass Spectrometry Based Proteomics

Attila Kertész-Farkas*[,1], Beáta Reiz[1,2,3], Michael P. Myers[1] and Sándor Pongor[1,2]

[1]International Centre for Genetic Engineering and Biotechnology, 34012 Trieste, Italy; [2]Szeged Biological Center, Temesvári krt 67, Szeged 6720, Hungary; [3]Institute of Informatics, University of Szeged, Aradi vértanúk tere 1, Szeged 6720, Hungary.

Abstract: Bottom-up proteomics (mass spectrometry analysis of peptides obtained by proteolysis and separated by liquid chromatography, (LC-MS/MS)) is one of the most frequently used techniques for identifying and characterizing proteins in biological samples. A key element of the analysis is database searching when the mass spectra of the peptides are compared with a database of theoretically computed (or experimental) peptide spectra. Here we discuss the main computational approaches to spectrum database searching and the statistical analysis of the results.

Keywords: Database search, Mass spectrometry, Protein identification, Proteomics.

1. INTRODUCTION

Liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) is one of the most frequent techniques used for identifying proteins in biological samples. This is a complex technology which gave rise to highly specialized computational approaches. The goal of this article is to provide an introduction to an important sub-problem, database search-based protein identification used in LC-MS/MS, for students and bioinformaticians who are new to this field. We note that there are various methods for fragmenting molecules for tandem mass spectrometry but for simplicity we limit this discussion to collision-induced dissociation (CID) methods that are perhaps the most widely used for analyzing complex biological samples.

In a typical proteomics experiment, a sample containing a mixture of proteins is first digested with a protease, such as trypsin, and the resulting mixture of peptides is subjected to automated analysis by Liquid Chromatography coupled to tandem mass spectrometry (LC-MS/MS.) The LC functions to remove contaminants, concentrate the analytes, and give the mass spectrometer more time to analyze the sample, as the constituent peptides will elute at different times. The mass spectrometer initiates an analysis cycle by measuring the mass to charge ratio of the precursors (originating from intact peptides) which are eluting from the LC. It is important to note that a mass spectrometer can only measure charged species and the mass and the charge determine the behavior of a particular ion in the mass spectrometer. Thus the unit of measurement for a mass spectrometer is the mass to charge ratio, which is abbreviated as m/z. The mass spectrometer then attempts to isolate a single precursor from all other co-eluting species, and then subjects this precursor to fragmentation within the mass spectrometer. The mass to charge ratio (m/z) of the resulting fragment ions is then measured. The output is a distribution of the observed m/z values, called a fragment ion mass spectrum, a product ion mass spectrum, or a tandem mass spectrum. Whatever it is called, the fragment ion spectra are typically associated with the precursor m/z, the charge state (z), and sometimes the chromatographic elution time. This data is then interpreted by one of three strategies:

a. *De novo* sequencing, either manually or computationally, attempts to infer the complete amino acid sequence of the fragmented peptide without the use of a sequence database. This approach is not discussed in this review, see chapter 11 of [1] for an excellent introduction.

b. In sequence-tagging, runs of amino acids are inferred from the spacing of the fragmentation peaks and these peptide-words are then used to identify proteins in a sequence database [2].

c. Finally, the most common strategy is database-searching where the mass spectrum is used without interpretation to query a database of theoretical spectra [3].

A variant of this final strategy uses libraries of previously identified spectra rather than a database of theoretically derived spectra. The use of spectral libraries has certain advantages over the use of theoretical spectra, however, both strategies rely on the correct peptide sequence residing in the database. LC-MS/MS produces large amounts of data that require specific preprocessing solutions [for an overview see accompanying paper]. Namely it is not uncommon to collect thousands of spectra from a single sample, but a large percentage (sometimes up to 80%) of the detected peaks can be noise [4].

In a broad sense, database searching using MS/MS spectra is largely an exercise in pattern recognition and entails finding those entries in a database that are the most similar to a query spectrum, typically in terms of some similarity criterion like correlation coefficients [3]. This is a fundamental task in bioinformatics (e.g. genome annotation, 3D structure analysis) but also in searching music or image databases. The terminology and the approaches vary depending on the field, nevertheless the algorithms share several commonalities:

*Address correspondence to this author at the Protein Structure and Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology, 34012 Trieste, Italy; Tel: +39-040 375 7354; Fax: +39-040 226 555; E-mail: kfattila@icgeb.org

2212-392X/12 $58.00+.00

a. The representation of the data and the choice of the similarity measure are critical, there are no universal solutions even within the same application area.

b. The data formats used for storage and for database searching are often different, and the choice between on-the-fly reformatting and database preformatting may depend on many factors (time and memory space requirements, updating frequency etc.). In most cases, databases are just series of records, but pre-grouping and/or indexing of the data is used in many areas in order to decrease the time-requirement of the searches [5].

c. There are multiple search schemes where one first uses inexpensive screening methods in order to identify a smaller number of potential positives which will then be further analyzed by more time-consuming, accurate scoring methods [6].

d. The results of the scoring procedure are often transformed into probabilistic quantities, such as statistical significance, using estimates of randomly occurring scores [7].

e. Ranking methodologies, top-lists and related statistical concepts such as ROC analysis [8] are frequently used in the evaluation.

In most bioinformatics applications, the goal of database searching is classification or annotation, i.e. we usually want to assign an object (sequence, structure, article, etc.) to a known class of objects. A class is always represented by some kind of a model, and we compare a query either with a database of annotated objects, or with a database of class descriptions. In a slightly different approach, we can compare our search results with an abstract description of an uncharacterized class, and decide if the query is the seed of a new class [9].

The goal of this review is to place proteomics database searching into this general framework. The proteomics workflow has additional challenges that are not frequently seen in other bioinformatics workflows. For example, one of the distinctive properties of the proteomics approach is the use of enzyme digestion that facilitates collection of experimental data. Unfortunately, this step multiplies the number of necessary database searches and simultaneously breaks the relationship between the measured peptides and the original protein. So the proteins present need to be deduced from their peptides identified by database search, in a process termed “protein inference” or “peptide grouping”. Furthermore, the proteomics data pipeline is beset with problems, such as the errors associated with the m/z measurements, as well as, missing and extraneous data point. These create particular challenges that must be overcome.

Although the peptide fragmentation pattern can be instrument dependent [10], most instruments are set up so that the peptides are fragmented in so that one particular molecule typically breaks at only one position somewhere along the peptide backbone. Of the possible ion types (Fig. **1** inset) the frequently seen *b*- and *y*- ions are most commonly used for peptide identification (Fig. **1**). In addition, there are characteristic derivative ions (loss of ammonia, loss of water oxidation, etc.), which are also used by some of the search engines (for the nomenclature of the ions see [11]). The fragmentation of peptides has a few unusual consequences: a) reconstruction of peptide sequence from MS spectra can be relatively straightforward, but is computationally expensive; b) Theoretical MS/MS spectra can be constructed from sequence using a few straightforward rules. This property allows one to produce theoretical MS/MS spectra collections for spectral searching methods. c) Each protein is characterized by a highly informative set of MS/MS spectra, which makes it possible to identify proteins from just a few MS/MS spectra, even in complex mixtures which contain thousands of proteins, such as clinical samples or biofluids. b) and c) are the properties that make MS/MS analysis the method of choice for identifying proteins in many current biological and medical applications. The *in silico* production of theoretical spectra is relatively straightforward as long as no attempt is made at predicting the intensity of the fragment ions, as the understanding of the sequence dependent effects of fragment ion intensity are not completely clear [12, 13].

2. DATA AND DATA FORMATS (EXPERIMENTAL AND THEORETICAL SPECTRA)

A mass spectrum is a plot of signal intensity vs. mass in which peaks correspond to fragment ions (Fig. **1**). Importantly, not all of the theoretically possible peptides are actually observed. For a review of fragmentation rules, ion nomenclature, type of ions formed see [10, 14]. The peak intensity (*y* axis) is a measure of the abundance of a particular fragment. The *x* axis is the mass-to-charge ratio (m/z) of the fragment, sometimes simply (but always incorrectly) referred to as the mass. As previously noted, in a typical MS/MS experiment the fragmentation ions falling into an instrument and precursor mass dependent range, are collected and their intensities and mass values are recorded and are associated with the precursor mass (the mass of the proteolytic peptide the fragments derive from), the charge state of the precursor, the intensity of the precursor and the chromatographic elution time. Fig. (**1**) shows an experimental MS/MS spectrum in a graphic form. One LC-MS/MS experiment generates many thousand MS/MS spectra that are recorded as a series of peak-lists.

Unfortunately, there is not a uniform format for these peak-lists and formats are usually generated by instrument manufacturers and bioinformatic groups [15, 16]. Standardization of these formats is still underway. There are currently a number of commonly used formats like *.mgf* (mascot generic format), *.dta* (Sequest), *.pkl*(Micromass), *.mzxml, .mzml, .mzdata,* etc. (for an in depth description of these formats see http://www.matrixscience.com/help/data_file_help.html). Recently an open file format *.mzXML* has been designed in order to facilitate the data exchange between programs in the computational pipeline [17]. Although the lack of a standardized file format is problematic, most search engines accept a variety of formats and a number of file conversion tools have been developed [18, 19].

Theoretical MS spectra are those calculated from peptide sequences. There are a variety of programs that can take a protein sequence, digest it *in silico* into fragments corresponding to the specificity of a user-selected proteolytic enzyme and output ions masses according to the theoretic fragmentation rules as indicated in Fig. (**1**) (Table 1). Such

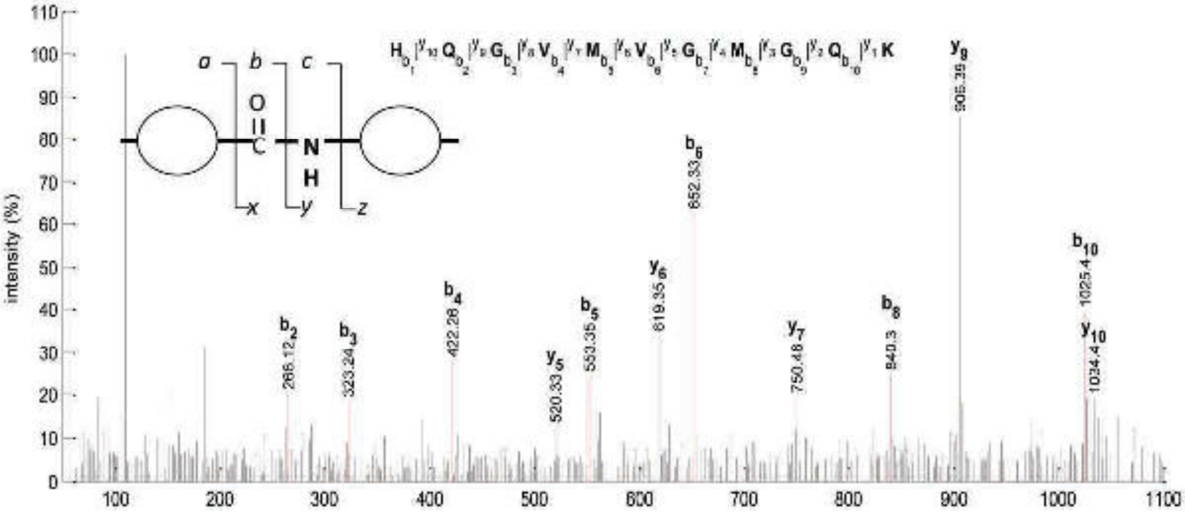


Fig. (1). An experimental spectrum of the peptide HQGVMVGMGQK (precursor m/z 1172.4, CID detection). In the mass spectrometer, a peptide breaks into two parts at various points of the peptide backbone, which gives rise to a series of ions (*a,b,c,x,y,*z, see inset). The annotated peaks correspond to the *b-y* fragment ions which are the most frequently used for peptide identification. Some *b-y* ions e.g. b_1, y_9 are not observed in this example. Other peaks can be considered noise. Note that the theoretically possible *a,c,x,*z ions are not detected in the most commonly used CID spectra, and b_1 ions form only under special conditions.

theoretical spectra contain a perfect ladder of *b*- and *y*- ions and sometimes their derivatives. Inconveniently, any of these theoretical ion may not be observed in the associated experimental spectra, due to limitations of the instrumentation, the physiochemical properties of the peptide or other factors. To help this problem, Protein Prospector generates ionization and instrument specific fragmentation tables (http://prospector.ucsf.edu/). Efforts to deal with the peptide sequence dependency of fragmentation have been more difficult and have only met with limited success [12, 13, 20]. There are currently no universally accepted or widely used methods to calculate theoretical peak intensities. Therefore, the intensities of theoretical spectra are usually taken as unity (binary vector description).

3. DATABASES

In other spectroscopic techniques, such as MS of organic molecules, data interpretation is based on large collections of experimental spectra. In MS/MS proteomics there are ongoing efforts to collect experimental spectra into spectral libraries [21-24]. However the task is formidable: the roughly 60 thousand proteins human proteins listed in the IPI database (v. 3.71) give rise to over 700 thousand unique tryptic peptides, and the experimenter may use different enzymes, combinations of enzymes, or is looking for post translational modifications. So the current method of choice is to use a protein sequence database to produce a collection of peptides *in silico*, and calculate their theoretical spectra. In principle, virtually any protein database can be used to calculate theoretical spectra. However, the non-redundant and curated databases (Table 2) are more suited for this approach as they can be searched more quickly. The *storage format* of the databases is a sequence collection, and for the production of the *search format* (the theoretical spectra), one can employ either preset variables or one can employ variable, experiment-dependent rules, such as user-defined enzyme cleavages, posttranslational modifications etc. In the case of preset variables, the theoretical spectra can be precalculated and stored in an indexed format. However, this approach limits the experimenter to preprogrammed datasets or requires that all possible variables be pre-calculated. Therefore, many theoretical databases perform these operations on-the-fly -of course indexing can also be done with on-the-fly production, which of course increases the time requirement.

Spectral libraries are collections of experimental spectra that have already been confidently assigned to peptide sequences [22]. In principle, these collections can be used instead of the theoretical peptide and protein sequence databases [21-24]. On the other hand we note that the most frequent problem for the database searches if a peptide is not in the database. This is especially true for current spectral libraries, even though frequently detected peptides can be quite readily identified from them. Additionally, most of current spectral libraries are dominated by ion trap data, so may not be universally applicable. In most cases the same algorithms can be used for searching both spectral libraries and sequence databases.

Theoretical spectra can also be created from decoy sequences. These decoy datasets are designed for statistical purposes, i.e. to model randomly occurring similarities between spectra [27-30]. A decoy database should fulfill the following criteria: i) the amino acid distribution should be the same as the original (target) database ii) the distribution of protein and peptide sizes should also be the same iii) the production of the decoy database should be reproducible by other labs. They can be produced from random shuffling [31] or Markov-chain generated amino acid sequences [32], or more typically, by simply reversing the sequence of proteins in the database [27] and then processing the database identically to the target database, which is often called a "reverse" dataset. The decoy dataset can either be appended to the target dataset or the two datasets can be searched separately and there are various advantages to each approach [33].

Table 1. Programs Available for Calculating Theoretical MS Spectra from Protein Sequence

Name	Web address
MS-product	http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msproduct
Peptide Fragmentation Modeller	http://omics.pnl.gov/software/PeptideFragmentationModeller.php
Theospec [25]	http://sourceforge.net/projects/protms/files/theospec/
InSilicoSpectro/FRagmentator [26]	http://insilicospectro.vital-it.ch/

Last accesses of web pages were on Feb 28th of 2011.

Table 2. List of Protein Databases Used in MS/MS Proteomics

Protein Sequence Databases	Web Address
Entrez Protein DB[a]	http://www.ncbi.nlm.nih.gov/sites/entrez?db=protein
Reference Sequence (RefSeq)[a]	http://www.ncbi.nlm.nih.gov/RefSeq/
UniProt, consisting of Swiss-Prot and TrEMBL[b]	http://www.uniprot.org/
International Protein Index (IPI)[c]	http://www.ebi.ac.uk/IPI/IPIhelp.html
Annotated Spectrum Libraries	
PeptideAtlas[d]	http://www.peptideatlas.org/
SBEAMS[d]	http://www.sbeams.org/
PRIDE[c]	http://www.ebi.ac.uk/pride/

Last accesses of web pages were on Feb 28th of 2011. [a] NCBI National Center of Bioinformatics), [b]SIB(Swiss Institute of Bioinformatics) and EBI(European Bioinformatics Institute), [c]EBI and [d]ISB(Institute for System Biology)

4. SPECTRUM COMPARISON

Mass spectra are typically compared as vectors, and in order to be comparable, the spectra (peak-lists given at high precision) have to be mapped to the same mass (m/z) coordinates. This is largely due to mass errors in the experimental spectra. This process can be pictured as binning the mass range; for instance, binning mass range of 0-4000 Daltons into 0.5 Dalton bins gives 8000 dimensional vectors. The bin width depends on the precision of the measurement, the less the mass resolution, the larger the bin must be. The intensity assigned to a bin is an aggregate value (maximum, average, sum) of the peak intensities that fall in the bin. Such vectors can then be numerically compared using a one of the many scoring functions available for calculating the similarity of vectors. The one which is most widely used to compare MS spectra is the so-called inner product (or often referred as dot product).

$$I(q,t) = \sum_i q_i t_i \ , \qquad (1)$$

where the experimental (query) and database spectra are denoted by q and t, respectively and a_i denotes the i^{th} component of the vector a. Since the query is an experimental spectrum with peak intensities, and the target is a binary vector from the database, the inner product in (eqn. 1) is equal to the sum of matching intensities. This is a good measure of spectrum similarity in itself, and it is used in the so-called *Xcorr* similarity measure used by the SEQUEST search engine [34]:

$$Xcorr(q,t) = I(q,t) - \frac{1}{151}\sum_{i=-75}^{75} I(q,t[i]) \qquad (2)$$

The second term in the formula (eqn. 2) is called cross-correlation and it represents the average back-ground noise, calculated by shifting query and the theoretical peptide spectrum with respect to each other, from -75 to +75 steps around the original position. Neither *I*, nor *Xcorr* are bounded, higher values indicate higher similarities.

The similarity of two spectra can also be measured by simply counting the matched non-zero peaks in the vectors, i.e. the peaks that are in the same position in both the experimental and the theoretical spectrum. Formally,

$$M(q,t) = \sum_i sign(q_i t_i)\ , \qquad (3)$$

where the function *sign*(.) is the *signum* function. This is the simplest similarity measure, it is discrete and its value falls between 0 and the total number of peaks in the sparser spectrum.

One way to convert the number of matching peaks into a bounded scoring function is to convert it into a probability value [35]. Geer and associates used the Poisson distribution to define a probability that the n matching peaks occur at random [36]. The resulting *OMSSA* score is defined as

$$P(q,t) = \frac{\mu^n}{n!} e^{-\mu} \qquad (4)$$

In the formula $\mu = 2Thv/m$, $n = M(q,t)$, where '*T*' is the resolution of the instrument, *h* and *v* are the number of the peaks in the experimental and the theoretical spectrum, respectively, and *m* denotes the neutral mass of the precursor. The division by *m* is a frequently used strategy to normalize the score to the mass of the peptide, as larger peptides will typically produce more fragment ions than smaller peptides. The value of P is between 0 and 1, lower values indicate better similarity.

The b- and y-ions can be distinguished in the theoretical spectra, we denote them by t_b and t_y, respectively and the inner products for b and y ions can be calculated separately, which will increase the specificity of the comparison. Fenyö and coworkers developed a hyperscore [35] on this basis:

$$hs(q,t) = [I(q,t_b) + I(q,t_y)][M(q,t_b)!M(q,t_y)!], \qquad (5)$$

where the two [.] terms are the intensity and the match count term, respectively. The peak intensities are usually scaled between 0 and 100. The match count term employs a factorial function, so this term will dominate the hyperscore. The application of binning to produce discrete vectors has limitations due to discretization error. Many search engines have evolved ways to avoid the discretization errors invoked by binning though the actual calculations follow the principles outlined in this section.

5. SIMILARITY SEARCH AND PEPTIDE ASSIGNMENT

In principle, an experimental query spectrum should be compared with all members of a database. In MS/MS proteomics, the search space is prohibitively large. Most search engines limit the spectral comparisons to experimental and theoretical spectra that are derived from precursors with similar masses. Indexed databases can then be used to provide a significant speed up [5, 37]. While this speeds up the search itself, storing an indexed database can be problematic, as many of the experimental variations need to be considered prior to database construction. Therefore, many search engines produce the database on-the-fly, using experimenter adjustable conditions and these on-the-fly databases may or may not be indexed. A heuristic solution, also used in other fields, is to carry out multiple searches, using a fast screening search as the first step followed by a more thorough search in the second step. The X!tandem search engine [38, 39] takes advantage of this strategy even farther. It first calculates the hyperscore (eqn. 4) on the limited set, then retrieves all proteins that had at least one top-scoring peptide in the first round, and carries out a more exhaustive search on this smaller dataset, using the same scoring algorithm, but extending the database with post translational modifications (to be discussed later). Another strategy is to use an inexpensive scoring function first, retain a set number of top-scoring entries and re-score them with a more computationally expensive algorithm in order to find the top-seated theoretical peptides. SEQUEST [34] calculates a so-called S_p score in the first round, defined as:

$$S_p(q,t) = \frac{I(q,t)M(q,t)(1+0.0075R(q,t))}{M(t,t)}, \qquad (6)$$

where $R(q,t)$ is the maximum number of consecutive b- or y- ions from the theoretical spectrum that appear in the spectrum q, and $M(t,t)$ is the number of the peaks in the theo-retical peptide spectrum. SEQUEST selects the top 500 scoring peptides and rescores them by *Xcorr* (eqn. 2), which is used as the final ranking function. Other approaches use more sophisticated filtering methods [40, 41]. The top seated peptide is assumed to be correctly identified and as long as the score of the top-seated peptide is above a preselected threshold, it is called "hit".

Finally, experimenters themselves usually optimize search time by selecting the sequence database on an empirical basis. One school limits the database to a single target species, trying to get the best annotated and least redundant dataset, refine it (e.g by adding isoforms, splice variants, removing signal and propeptides, N-terminal Met residues etc.) before using it for generating the theoretical spectra. This approach is efficient, but problems arise if the dataset is too small for calculating statistical significance (discussed in the next section), especially if common contaminants, such as keratin, are not accounted for. Another empirical approach is to use a comprehensive (multispecies or all species) dataset in the beginning, and limit the search to species that produce recurrent hits. This approach is more comprehensive and more time-consuming, especially if the sample contains evolutionarily conserved peptides. In other words, the search space could be limited based on the available experimental information considering the origin of the sample, the mass accuracy featured by the instrument, the ion types that particular activation will produce etc.

Calculation of Significance of the Peptide Assignment

The notion of significance used in MS/MS proteomics follows the same principles as used in various other fields, for instance in BLAST statistics [42]. The significance of a peptide hit with a similarity score h is the probability of observing a random score x that is higher or equal than the hit h, that is $P(h \le x)$. This probability is the so-called p-value, its definition follows the Fisherian principles of statistical hypothesis testing, and requires knowledge of the distribution of chance or random similarity scores. In this context the random score means the similarity score between an experimental and a decoy database spectrum.

The significance of peptide hits is estimated by various methods:

a. Explicit calculation of the random probabilities based on theoretical considerations; the probability distribution of the random scores depends on the type of the scoring function. In the case of the matching peak count, this distribution can be modeled with hypergeometric [43] or Poisson distribution [36], for matched intensities (eqn. (1)) it is modeled with a normal distribution [44].

b. Fitting a distribution to a sample of random scores. These can be either extracted from data obtained by comparison of the query spectrum with the theoretical database (and leaving out the highest similarity scores as non-random) or determined by comparing the query spectrum to a separate decoy dataset [35]. The drawback of the latter is that it doubles the database search time. The random hyperscores (eqn. 4) are modeled with fitting an extreme value distribution [35].

c. PeptideProphet [45] calculates a probability that the hit is correct via building distributions on scores and peptide properties (the number of termini compatible with enzymatic cleavage (for unconstrained searches), the mass difference with respect to the precursor ion, the presence of an N-glycosylation motif (for N-glycosylation capture experiments), etc.) of correct and incorrect hits using an Expectation-Maximization (EM) method.

d. When the distribution type of the random scores are not known, the *p*-values can be approximated in non-parametric way by simply calculating the percentage of the scores on the decoy dataset, equal or higher score than the actual hit itself [29].

e. *DeltaCN* of SEQUEST calculates the approximate significance of a query hit [3] as the ratio of the best score value to the second best score value obtained in the database search. Note that this value is not a statistical significance in the rigid mathematical sense, it relates to likelihood ratio [46].

After calculating the significance of the query hits, one can choose a threshold to keep the top significant hits and calculate the False Discovery Rate (FDR) either on a probabilistic basis or using the hits obtained on the decoy dataset [29]. For instance, one can adjust the threshold for truncating the top list of hits so as to have a given percent of decoy hits above the threshold, typically 1% or 5%, corresponding to FDR= 0.01 or 0.05. The quantities used for significance analysis In LC-MS/MS are not very different from those used in other fields (Box 1.). What is different is the frequent (and, perhaps, not always critical) use of decoy databases. In reality, a) decoy databases can contain high scoring peptides that are similar to real sequences and b) not all the hits against a target library are true positives. Both of these can bias the evaluation, nevertheless, decoy databases remain useful tools for roughly pinpointing the range of important similarities.

6. PROTEIN INFERENCE

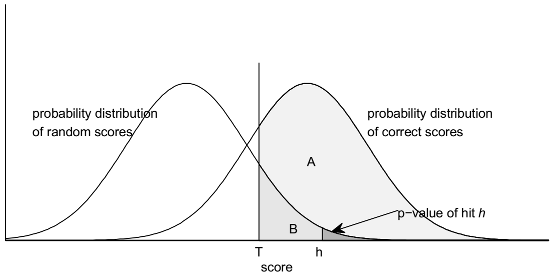
Once we have a list of significant peptide hits, the next task is to infer a list of proteins which are present in the sample. This can be done by assigning a score to proteins via some aggregation method applied to the peptide scores (or *p*-values). This task is not trivial, different methods can result different solutions. The main reasons are that i) some true peptides may not generate significant p-values, ii) some peptides (called degenerate peptides) may be present in more than one protein, and iii) very large proteins in the theoretical database may be selected, even though none of the individual peptide hits would be considered statistically significant. A practical approach, sometimes referred to as the parsimony principle, is to select the simplest group of proteins that is sufficient to explain all the observed peptides [47, 48]. However, it should be noted that this procedure can provide more than one solution depending on the parsimony principle applied. In the example shown in Fig. (**2**), the 5 peptides have been identified that are parts of 4 different proteins. If the principle is to select the minimal set of proteins that explain the results, then *c* and *d* are sufficient and *a* and *b* are discarded. However, if the parsimony principle is altered to pick the minimal set of proteins that also have the most peptide evidence, then proteins *a* and *c* are selected, and proteins *b* and *d* are discarded.

In addition to using simple parsimony rules, the protein inference problem can be solved by converting the peptide scores (or significance measures) to protein based measures. The X!tandem software calculates an *E*-value for proteins by simply forming the product of the peptide *E*-values

$$E(protein) = \sqrt{M}\prod_{i=1}^{n} e(p_i)\,, \tag{7}$$

where $e(p_i)$ denotes the *E*-value of a peptide p_i, peptide p_i is part of the protein and *n* is the number of the peptides of the protein of interest [35], where *M* is the number of the significant query hits. This would bias the reporting to the proteins with the best peptide matches, rather than towards the protein with the most peptide matches.

Box 1. Statistics Terminology



A hit *h* is labeled significant hit if it is greater than or equal to the threshold τ. Random score is a similarity score between a randomly chosen experimental and theoretical decoy peptide spectra.

Hit	A hit of a query q is the maximum of the similarity scores calculated between the query and the theoretical peptide spectra of the database, formally $hit_D(q) = \max_{t \in D}\{s(q,t)\}$, where *D* denotes the database.
p-value	The *p*-value of a query hit *h* is the probability of observing a random score x greater or equal to the score of the hit i.e $P(h \le x)$. The smaller the *p*-value, the more significant the observed hit. The *p*-value depends on how the distribution of the random scores is modelled.
E-value	The *E*-value of a query is the expected number for finding a database element with random score greater than or equal to the query hit *h* on a database of *n*, i.e. $E(q) = nP(h \le x)$. For instance, an *E*-value of 10^{-2} means that the score *h* is expected to occur by chance only once in 100 independent similarity searches over the database. If the E-value is 10, then ten random hits with score greater or equal to *h* are expected within a single similarity search.
FPR	False Positive Rate, the probability of labelling a random score significant (area B in the figure). A FPR of 0.01 means that 1% of the random scores are labelled significant.
FDR	False Discovery Rate, the ratio of random scores within significant scores, formally $FDR = A/(A+B)$ (A, and B are defined in inset). The FDR = 0.01 means the 1% of the scores labelled significant are actually observed by chance. FDR is often used to control the ratio of the false positives. The threshold τ can be set to keep the *FDR* under a certain level, typical levels are 0.01 or 0.05, i.e experimenters set thresholds to allow 1% or 5% of false positives. The lower the *FDR* the more true (non-random) similarity hits are lost.
q-value	The *q*-value of the query is the minimum *FDR* when its hit is called significant. A *q*-value of 0.01 means that trying all possible values for threshold τ, the lowest FDR is 0.01 when the hit is labelled significant. Note that the *q*-value also depends on the database [49].

	Proteins			
	a	b	c	d
Peptide 1			+	
Peptide 2	+	+	+	+
Peptide 3	+	+	+	+
Peptide 4	+			+
Peptide 5	+	+	+	

Fig. (2). A protein/peptide composition map. A '+' denotes that a peptide is present in the protein.

Protein Prophet [47] computes the probability of a protein being present in a sample with the following statistical model:

$$P(prot) = 1 - \prod_{i=1}^{N}\left(1 - \frac{w_i^n p(+\,|\,D_i, NS_i^n)}{p(-\,|\,D_i, NS_i^n) + p(+\,|\,D_i, NS_i^n)}\right) \tag{8}$$

where $w_i^n = P_n / \sum_{s=1}^{m} P_s$ is the weight of peptide i being assigned to the protein, NS_i^n is the number of siblings of peptide i in the proteins of interest, + and - denote the 'correct' and 'incorrect' assignment resp. Finally $p(+|D_i, NS_i^n) = p(+\,|\,D_i)p(NS_i^n\,|+)$, where $p(NS_i^n\,|+)$ is the probability of having a particular NS value of correct or incorrect peptide assignment and $p(+\,|\,D_i)$ is the probability of the correct assignment having information D_i. For parameter learning, ProteinProphet uses an EM algorithm. We note that all solutions presented here are mathematical, but in the practice, any of the proteins that contain a particular peptide can be present. For a review paper about the protein inference problem see [50].

7. POST TRANSLATIONAL MODIFICATIONS AND PARTIAL DIGESTION

Until now we considered an ideal case where all peptides found in the sample perfectly match a theoretical peptide. However this is not the case in practice, and posttranslational modifications (PTMs) are one of the main reasons of discrepancy. On average, human proteins are thought to carry 3 PTMs on one of their amino acid side chains. However less than 1% of the proteins in the UniProtKB/Swiss-Prot are annotated with PTMs [51]. In additional to biologically relevant modifications, modifications can also occur during sample preparation and these also have to be accounted for during the database search. The current list of these two kinds of peptide modifications is over 500, they are listed in lookup tables according to their specificity and the mass change they induce (see www.unimod.org or http://www.ebi.ac.uk/RESID/). PTMs that are stable during the fragmentation will change both the precursor mass and some peaks within the fragmentation spectrum: the *b*- ions that lie C-terminal to the modified site and the *y*- ions N-terminal to it will be increased by the mass of the PTM. The rest of the spectrum may remain unchanged with respect to a spectrum from an unmodified peptide. The overlap between modified and unmodified spectra from the same peptide is often sufficient for clustering the modified and unmodified spectra together but may not be sufficient for getting a significant score from a database search engine. Therefore the search engine needs to be capable of predicting the changes in precursor mass and the resulting fragmentation spectrum.

The calculation of theoretical spectra for all possible PTMs is not feasible because each PTM under consideration leads to a combinatorial explosion of the spectra to be tested, especially when more than one PTM per peptide is allowed. In essence, the inclusion of PTMs results in the rapid expansion of the search space.

For PTMs identifications two main approaches have been developed. In the first approach (called restricted modification searches) the experimenter has to provide the kind of PTMs that are expected and then the program will generate all the possible combinations of modified theoretical spectra. In the second approach (called unrestricted modification searches) the programs try to identify PTMs without a priori assumptions on modifications that might be present and this approach can use PTMs annotated in databases or previously unseen ones [52, 53]. For a review paper on various strategies for finding PTMs see [54].

Partial digestion refers to the fact that the digestive enzyme used for sample preparation (e.g. trypsin) may not cleave at all the expected sites. So in addition to the expected peptides, the sample may contain peptides with one or more uncleaved sites (missed cleavages). The number of missed cleavages causes a steep increase to the number of peptides. A protein that yields 10 peptides upon perfect cleavage, yields 19 or 27 if one or two partially cleaved bonds are allowed. Similarly, the enzyme may also cut in unexpected places (unanticipated cleavages), which can also confound search engines. A popular strategy is to require at least one end of the peptide to be a "true" cleavage and the other end is allowed to vary [36, 55].

In contrast to spectral matching, discussed above, the sequence tagging approach is closely related to *de novo* sequencing, inasmuch as runs of amino acids ("sequence tags" or "amino acid words") are directly inferred from the spacing of the fragmentation peaks [2]. Importantly, a few short runs of at least 2-3 amino acids (called amino acid words) are often enough to identify the peptide sequence, via querying a sequence database for similar sequences [56]. The sequence tagging approach is particularly useful for finding post translational modifications (PTMs), as many of these short amino acid words can be determined that are not influenced by the PTM. This is because the sequence tagging approach looks at the spacing between peaks and not at their absolute masses. Similarly, missing peaks, missed and unanticipated cleavages are not so problematic in the sequence tagging approach. From the point of view of database searching, sequence tagging is not radically different from spectral matching. The difference lies in the fact that the query spectrum is transformed to amino acid word sequences, and the database is not converted into spectra. Word based searching is one of the oldest, though not the most sensitive, methods in sequence comparison. In this particular application, the selectivity of the search is increased by the fact that we identify words in peptides

Table 3. List of Database Search-Based Protein Identification Software's

Name	License	Distributor	Web Address	Ref.
Database Search				
Sequest	Proprietary	ThermoFinnigan	http://fields.scripps.edu/sequest/	[34]
Mascot	Proprietary	Matrix science Inc.	http://www.matrixscience.com/	[57]
X!tandem	Open source	TheGPM	http://www.thegpm.org/	[55]
Phenyx	Free on-line version,	GenBio	http://www.genebio.com/products/phenyx/	[32]
OMSSA	Open Source	NCBI	http://pubchem.ncbi.nlm.nih.gov/omssa/	[36]
MyriMatch	Open Source	Vanderbilt Medical Center	https://www.mc.vanderbilt.edu/msrc/bioinformatics	[58]
Graylag	Open Source	Stowers Institute for Medical Research	http://greylag.org/	
MassWiz	Open Source	Institute of Genomics and Integrative Biology	http://masswiz.igib.res.in/	
Protein Prospector	Proprietary	Thermo Fisher	http://www.thermoscientific.com/	
Andromeda	Freeware	Max-Planck Inst.	http://www.maxquant.org/	[59]
Spectral Matching				
Spectra ST		ISB	http://www.peptideatlas.org/spectrast/	[60]
X! Hunter	Open Source	TheGPM	http://www.theGPM.org	[61]
Sequence tag/hybrid Approaches				
Inspect		Univ. California	http://proteomics.ucsd.edu/Software/Inspect.html	[62]
GutenTag		Yates Lab.	http://fields.scripps.edu	[63]
Paragon	Proprietary	Applied Biosystems	http://proteomics.ucsd.edu/Software/Inspect.html	[64]

Last accesses of web pages were on Feb 28th of 2011.

rather than in entire protein sequences. However it has to be noted that the direction of short amino acid runs is not necessarily known, so a naïve word search algorithm may not be efficient in this case. Practical implementations of this strategy use a structure for the peptide database which speeds up the searching procedure [56, 62].

8. PROGRAMS

The field of proteomics database searching is still emerging and there is an ever growing number of database searching engines (many of which are listed in the Table 3.) The search engines can be divided into commercial programs, such as MASCOT and SEQUEST, and freely available/open source programs such as X!tandem and OMSSA. Direct comparison of these applications is often difficult, because each search engine uses different spectral properties for finding a match - and this can be affected by a particular instrument or experimenter. In other words, a particular search engine may be the best performer with one dataset and the worst performer with a different dataset. For example, a search engine that scores only the b- and y- ions will do well on datasets where these are the dominant signals, in datasets with additional ion series, a search engine that includes more series will tend to do better. In general terms, direct comparisons of different search engines indicate that there will be an overlapping set of results, that many search engines are capable of finding, as well as, a set of results that are unique to a particular search engine [65]. In order to increase the number, as well as, confidence in the results, it has become increasingly popular to combine the results from multiple search engines [5, 65, 66]. However, this strategy dramatically increases the analysis time, and multipass algorithms, which use a succession of different algorithms that are guided to limit redundant searches are also being developed [51]. Some searching parameters with their default values are listed in Table 4.

CONCLUSIONS AND PERSPECTIVES

As the proteomics field advances, statistic based search algorithms have become increasingly vital for the proper interpretation of these results [35]. In fact, it is virtually impossible to publish proteomics results without a statistical analysis of the results [67, 68]. A number of strategies have been developed to determine the statistical quality of a match and these can either be an integral part of the search engine or can be performed on the results of the search engine. It is important to note that not every strategy is compatible, for example the scoring algorithm of X!tandem needed to be altered to remove the hypergeometric distribution to make it more compatible with ProteinProphet [69].

Although classically used measures of statistical confidence, such as *E*-values and *p*-values, are commonly used to assess results from search engines, it is becoming increasingly common to use target-decoy strategies to accurately determine the false discovery rate (FDR) [28, 33, 69]. The future of database searching will be driven, in part, by a better handling of the false discovery rate. In fact a number of recent reports indicate the danger of blindly

Table 4. List of the Most Common Parameters of the Database Search Software's

Parameter name	Default Values	Description
Taxonomy	Human, prokaryotes, viruses, etc.	Determines organism where the sample is from
Mass tolerance	<1.0 Dalton	The precursor mass difference for database filtering.
Match tolerance	0.4 Dalton	The difference within two peak positions considered equal. i.e. the resolution of the spectra.
List of modifications (chemical, PTM)	Oxidation, Alkylation of cysteine	A fixed list of modification. Using this modification. The number of the modification exponentially increases the searching time.
Number of miscleavages	1,2	Number of the missed cleavages during the enzymatic digestions.
Digestion enzyme	Trypsin,	The enzyme that is used to digest the proteins.
Filtering threshold (log(E-value))	-1	Threshold to label a hit to be significant.

adhering to particular statistical models [70-72]. These discrepancies indicate the need for independent means of validating the results from proteomics search engines. This validation can take the form of using rescoring algorithms to further validate the initial analysis [73, 74], or can be more complex and take into account information from other databases, such as Gene Ontologies or Protein Interactions [75]. This latter category of validation has the potential for revolutionizing the field, as a large amount of effort is spent on trying to understand which parts of a large list of results are the most biologically relevant.

CONFLICT OF INTEREST

None declared.

ACKNOWLEDGEMENT

None declared.

REFERENCES

[1] Pevzner PA. Computational Molecular Biology: An Algorithmic Approach: MIT Press 2000.

[2] Mortz E, O'Connor PB, Roepstorff P, *et al.* Sequence tag identification of intact proteins by matching tanden mass spectral data against sequence data bases. Proce Nat Acad Sci USA 1996; 93(16): 8264-7.

[3] Yates JR, 3rd, Eng JK, McCormack AL, Schieltz D. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal Chem 1995; 67(8): 1426-36.

[4] Renard BY, Kirchner M, Monigatti F, *et al.* When less can yield more - Computational preprocessing of MS/MS spectra for peptide identification. Proteomics 2009; 9(21): 4978-84.

[5] Li Y, Chi H, Wang LH, *et al.* Speeding up tandem mass spectrometry based database searching by peptide and spectrum indexing. Rapid Commun Mass Spectrom 2010; 24(6): 807-14.

[6] 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): 2310-6.

[7] Fenyo D, Beavis RC. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. Anal Chem 2003; 75(4): 768-74.

[8] Sonego P, Kocsor A, Pongor S. ROC analysis: applications to the classification of biological sequences and 3D structures. Briefings in bioinformatics. 2008; 9(3): 198-209.

[9] Duda R, Hart P, Stork D. Pattern Classification (2nd Edition): Wiley-Interscience 2001.

[10] Wells JM, McLuckey SA, Burlingame AL. Collision-induced dissociation (CID) of peptides and proteins. Methods Enzymol 2005; 402: 148-85.

[11] Roepstorff P, Fohlman J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. Biomed Mass Spectrom 1984; 11(11).

[12] Huang Y, Triscari JM, Pasa-Tolic L, *et al.* Dissociation Behavior of Doubly-Charged Tryptic Peptides: Correlation of Gas-Phase Cleavage Abundance with Ramachandran Plots. J Am Chem Soc 2004; 126(10): 3034-5.

[13] Paizs B, Suhai S. Fragmentation pathways of protonated peptides. Mass Spectrom Rev 2005; 24(4): 508-48.

[14] Roepstorff P, Fohlman J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. Biomed Mass Spectrom 1984; 11(11): 601.

[15] Deutsch EW, Mendoza L, Shteynberg D, *et al.* A guided tour of the Trans-Proteomic Pipeline. Proteomics 2010; 10(6): 1150-9.

[16] Shah AR, Davidson J, Monroe ME, *et al.* An efficient data format for mass spectrometry-based proteomics. J Am Soc Mass Spectrom. 2010; 21(10): 1784-8.

[17] Pedrioli PG, Eng JK, Hubley R, *et al.* A common open representation of mass spectrometry data and its application to proteomics research. Nat Biotechnol 2004; 22(11): 1459-66.

[18] Falkner JA, Falkner JW, Andrews PC. ProteomeCommons.org IO Framework: reading and writing multiple proteomics data formats. Bioinformatics 2007; 23(2): 262-3.

[19] Kohlbacher O, Reinert K, Gropl C, *et al.* TOPP--the OpenMS proteomics pipeline. Bioinformatics 2007; 23(2): e191-7.

[20] Yu C, Lin Y, Sun S, *et al.* An iterative algorithm to quantify factors influencing peptide fragmentation during tandem mass spectrometry. J Bioinform Comput Biol 2007; 5(2a): 297-311.

[21] Barsnes H, Eidhammer I, Martens L. A global analysis of peptide fragmentation variability. Proteomics 2011.

[22] Craig R, Cortens JC, Fenyo D, Beavis RC. Using annotated peptide mass spectrum libraries for protein identification. J Proteome Res 2006; 5: 1843-9.

[23] Deutsch EW. Tandem mass spectrometry spectral libraries and library searching. Methods Mol Biol (Clifton NJ) 2011; 696: 225-32.

[24] Lam H, Aebersold R. Building and searching tandem mass (MS/MS) spectral libraries for peptide identification in proteomics. Methods 2011.

[25] Boehm AM, Grosse-Coosmann F, Sickmann A. Command line tool for calculating theoretical MS spectra for given sequences. Bioinformatics 2004; 20(16): 2889-91.

[26] Colinge J, Masselot A, Carbonell P, Appel RD. InSilicoSpectro: an open-source proteomics library. J Proteome Res 2006; 5(3): 619-24.

[27] Moore RE, Young MK, Lee TD. Qscore: an algorithm for evaluating SEQUEST database search results. J Am Soc Mass Spectrom 2002; 13: 378-86.

[28] Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 2007; 4: 207-14.

[29] Kall L, Storey JD, MacCoss MJ, Noble WS. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. J Proteome Res 2008; 7(1): 29-34.

[30] Choi H, Ghosh D, Nesvizhskii AI. Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. J Proteome Res 2008; 7(1): 286-92.

[31] Klammer AA, MacCoss MJ. Effects of modified digestion schemes on the identification of proteins from complex mixtures. J Proteome Res 2006; 5(3): 695-700.

[32] Colinge J, Masselot A, Giron M, Dessingy T, Magnin J. OLAV: Towards high-throughput tandem mass spectrometry data identification. Proteomics 2003; 3: 1454-63.

[33] Blanco L, Mead JA, Bessant C. Comparison of novel decoy database designs for optimizing protein identification searches using ABRF sPRG2006 standard MS/MS data sets. J Proteome Res 2009; 8(4): 1782-91.

[34] Eng JK, McCormack AL, Yates Iii JR. 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(11): 976-89.

[35] Fenyo D, Beavis RC. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. Anal Chem 2003; 75: 768-74.

[36] Geer LY, Markey SP, Kowalak JA, *et al.* Open mass spectrometry search algorithm. J Proteome Res 2004; 3(5): 958-64.

[37] Roos FF, Jacob R, Grossmann J, *et al.* PepSplice: cache-efficient search algorithms for comprehensive identification of tandem mass spectra. Bioinformatics 2007; 23(22): 3016-23.

[38] Craig R, Beavis RC. Rapid Commun Mass Spectrom 2003; 17(20): 2310.

[39] Craig R, Cortens JP, Beavis RC. Open source system for analyzing, validating, and storing protein identification data. J Proteome Res 2004; 3(6): 1234-42.

[40] Ramakrishnan SR, Mao R, Nakorchevskiy AA, *et al.* A fast coarse filtering method for peptide identification by mass spectrometry. Bioinformatics 2006; 22(12): 1524-31.

[41] Dutta D, Chen T. Speeding up tandem mass spectrometry database search: metric embeddings and fast near neighbor search. Bioinformatics 2007; 23(5): 612-8.

[42] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990; 215(3): 403-10.

[43] Sadygov RG, Yates JR. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. Anal Chem 2003; 75: 3792-8.

[44] Sadygov R, Wohlschlegel J, Park SK, Xu T, Yates JR, 3rd. Central limit theorem as an approximation for intensity-based scoring function. Anal Chem 2006; 78(1): 89-95.

[45] 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: 5383-92.

[46] Kajan L, Kertesz-Farkas A, Franklin D, Ivanova N, Kocsor A, Pongor S. Application of a simple likelihood ratio approximant to protein sequence classification. Bioinformatics 2006; 22(23): 2865-9.

[47] Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 2003; 75: 4646-58.

[48] Zhang B, Chambers MC, Tabb DL. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. J Proteome Res 2007; 6(9): 3549-57.

[49] Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA 2003; 100: 9440-5.

[50] Shi J, Wu F-X. Protein Inference by Assembling Peptides Identified from Tandem Mass Spectra. Curr Bioinform 2009; 4: 226-33.

[51] Tharakan R, Edwards N, Graham DR. Data maximization by multipass analysis of protein mass spectra. Proteomics 2010; 10(6): 1160-71.

[52] 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): 935-48.

[53] Bern M, Cai YH, Goldberg D. Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. Anal Chem 2007; 79: 1393-400.

[54] Ahrne E, Muller M, Lisacek F. Unrestricted identification of modified proteins using MS/MS. Proteomics. 2010; 10(4): 671-86.

[55] Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics 2004; 20: 1466-7.

[56] Shilov IV, Seymour SL, Patel AA, *et al.* The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. Mol Cell Proteomics 2007; 6(9): 1638-55.

[57] Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 1999; 20: 3551-67.

[58] Tabb DL, Fernando CG, Chambers MC. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J Proteome Res 2007; 6 654-61.

[59] Cox Jr, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. J Proteome Res 2011: null-null.

[60] Lam H. Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics 2007; 7: 655-67.

[61] Craig R, Cortens JP, Beavis RC. The use of proteotypic peptide libraries for protein identification. Rapid Commun Mass Spectrom 2005; 19: 1844-50.

[62] Tanner S. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem 2005; 77: 4626-39.

[63] Tabb DL, Saraf A, Yates JR. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. Anal Chem 2003; 75: 6415-21.

[64] Shilov IV, Seymour SL, Patel AA, *et al.* The Paragon Algorithm, a Next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem. Mass Spectra 2007: 1638-55.

[65] Searle BC, Turner M, Nesvizhskii AI. Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. J Proteome Res 2008; 7(1): 245-53.

[66] Jones AR, Siepen JA, Hubbard SJ, Paton NW. Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. Proteomics 2009; 9(5): 1220-9.

[67] Binz PA, Barkovich R, Beavis RC, *et al.* Guidelines for reporting the use of mass spectrometry informatics in proteomics. Nat Biotechnol 2008; 26(8): 862.

[68] Wilkins MR, Appel RD, Van Eyk JE, *et al.* Guidelines for the next 10 years of proteomics. Proteomics 2006; 6(1): 4-8.

[69] Choi H, Nesvizhskii AI. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. J Proteome Res 2008; 7(1): 254-65.

[70] Bern M, Kil YJ. Comment on "Unbiased Statistical Analysis for Multi-Stage Proteomic Search Strategies". J Proteome Res 2011.

[71] Cooper B. The Problem with Peptide Presumption and Low Mascot Scoring. J Proteome Res 2011.

[72] Everett LJ, Bierl C, Master SR. Unbiased statistical analysis for multi-stage proteomic search strategies. J Proteome Res 2010; 9(2): 700-7.

[73] Brosch M, Yu L, Hubbard T, Choudhary J. Accurate and sensitive peptide identification with Mascot Percolator. J Proteome Res 2009; 8(6): 3176-81.

[74] Nesvizhskii AI, Aebersold R. Analysis, statistical validation and dissemination of large-scale proteomics data sets generated by tandem MS. Drug Discov Today 2004; 9: 173-81.

[75] Mayburd AL, Martlinez A, Sackett D, *et al.* Ingenuity network-assisted transcription profiling: Identification of a new pharmacologic mechanism for MK886. Clin Cancer Res 2006; 12(6): 1820-7.

Received: April 06, 2011 Revised: July 15, 2011 Accepted: July 18, 2011