A face in the crowd: recognizing peptides through database search

Jimmy K. Eng\textsuperscript{1}, Brian C. Searle\textsuperscript{2}, Karl R. Clauser\textsuperscript{3} and David L. Tabb\textsuperscript{4}

\textsuperscript{1} Jimmy K. Eng, University of Washington, Department of Genome Sciences, Box 358050, Seattle, WA 98195, USA.

\textsuperscript{2} Brian C. Searle, Proteome Software, 1340 Bertha Blvd, Suite 10, Portland, OR 97219, USA.

\textsuperscript{3} Karl R. Clauser, Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142, USA.

\textsuperscript{4} David L. Tabb, Vanderbilt University, Department of Biomedical Informatics, 465 21st Ave. S., Learned Lab / MRB III, room U9211, Nashville, TN 37232-8575, USA.

Corresponding author:

Jimmy K. Eng
206.685.6137 (phone)
206.616.0715 (fax)
engj@u.washington.edu
Abstract

Peptide identification via tandem mass spectrometry sequence database searching is a key method in the array of tools available to the proteomics researcher. The ability to rapidly and sensitively acquire tandem mass spectrometry data and perform peptide and protein identifications has become a commonly used proteomics analysis technique due to advances in both instrumentation and software. Although many different tandem mass spectrometry database search tools are currently available from both academic and commercial sources, these algorithms share similar core elements while maintaining distinctive features. This review revisits the mechanism of sequence database searching and discusses how various parameter settings impact the underlying search.
Shotgun proteomics and database search

Innovations in tandem mass spectrometry (MS/MS) have enabled the rapid growth of proteomics for chemists and biologists alike. In addition to the evolution of the instrument hardware to acquire spectra more rapidly and sensitively, improvements to data analysis facilitates the process of identifying and quantifying peptides and proteins for a wide user base. Researchers new to the field can benefit from a review of the fundamentals of tandem mass spectrometry sequence database searching, and even experienced proteomics researchers may profit from considerations of practical strategies to maximize information extraction from data.

In a typical shotgun proteomics experiment, the sample is first denatured to enable proteolysis. An enzyme such as trypsin is applied to cleave proteins to smaller peptide components. This peptide mixture is usually then separated on a liquid chromatography (LC) column; more complex samples may necessitate prior fractionation by strong cation exchange or isoelectric focusing. Peptides are subjected to ionizing voltage as they are electrosprayed from the column, and these ions are introduced into the near-vacuum environment of the mass spectrometer.

In a tandem mass spectrometry experiment, the instrument will first acquire a survey or precursor scan, also referred to as an MS scan, which measures all intact peptide ions eluting into the mass spectrometer at that given time. One or more peptide ions are selected, sequentially isolated, fragmented, and the resulting fragment ions are measured to produce an MS/MS spectrum. This process is repeated to automatically acquire MS/MS spectra on as many different peptide ions as possible throughout the LC gradient. See Fig. 1 for a schematic showing the relationship between precursor ions in the MS scans and the resulting MS/MS scans. While the MS/MS spectrum contains the peptide fragmentation pattern, the experimental peptide mass and charge state are obtained from the precursor ion measure in the MS spectrum.
To clarify terminology, the intact peptide ions in the MS scans are termed precursor ions. Fragment ions are also called product ions and MS/MS spectra are referred to as product ion spectra. MS and MS/MS spectra can also referred to as MS1 and MS2 spectra, respectively. While collision-induced dissociation (CID) is the most common way to generate product ions, others including electron capture dissociation (ECD), electron transfer dissociation (ETD), and infrared multiphoton dissociation (IRMPD) have also been developed.

After data acquisition, the MS/MS spectra are searched against a protein sequence database to identify the underlying peptides and proteins represented in the acquired spectra. A 1994 paper by Mann and Wilm introduced the strategy of searching protein sequence databases using peptide sequence tags interpreted from MS/MS spectra (1). These tags of 3 to 5 consecutive amino acids were inferred by recognizing chains of amino acid mass differences between peaks. The masses flanking the tag along with the inferred sequence were then matched to peptide sequences drawn from a database of protein sequences. This approach was shown to be capable of identifying peptides with post translational modifications (PTMs) by allowing for differences between the measured peptide mass and the calculated mass of peptides in the database. At first, tag inference was done manually so it was a relatively slow process that required some domain expertise to perform. In recent years, automated systems for inferring sequence tags have been joined by new tools to reconcile mass differences between spectra and sequences to yield new capabilities in modification and mutation identification (2-4).

Uninterpreted MS/MS database search algorithms, also introduced in 1994, rapidly became the standard method for protein identification. Eng et al. described the SEQUEST algorithm to match MS/MS spectra to peptides drawn from protein sequence databases (5). This approach gained the ability to identify post-translational modifications the following year (6). With no requirement of manual interpretation of spectra prior running searches, the ability to automate the search process increases throughput and opened the door for non-experts to perform the analyses. Subsequent years saw the publication of both commercial and open-source tools that implemented this approach. For a detailed review of specific search tools and
a collection of search engine comparisons, please see (7-14). The common elements of these tools and practical considerations guiding their use are the specific focus of this review.

Structure of database search algorithms

From a simplified view, most MS/MS database search tools perform the same basic functions. They all read a collection of MS/MS spectra, query a sequence database to select peptides of the right mass, score these peptides against the experimental spectra, and return putative peptide identifications. A schematic of process is depicted in Fig. 2. For each MS/MS spectrum, an experimental peptide mass can be derived from the precursor mass-to-charge ratio (m/z) and assumed or measured precursor charge state. A database search tool will select peptides from the sequence database that are of the same approximate mass as the experimental peptide mass for the query spectrum. The set of potential peptides that get scored against each spectrum are hereby referred to as candidate peptides. The choice of the precursor mass tolerance setting is influenced directly by the accuracy of the mass analyzer being used to measure the precursor spectra. The set of candidate peptides is also influenced by other factors such as the enzyme specificity setting and any post translational modifications being considered in the search. A set of expected fragment ion masses are then calculated for each candidate peptide sequence. The activation method used to fragment the peptides defines which types of fragment ion masses are considered, e.g. b- and y-ions for CID and c- and z•-ions for ETD. These fragment ions are compared against those in the experimental spectrum and a similarity score or closeness-of-fit metric is calculated. Each candidate peptide is scored, and the highest-rated peptides for each spectrum query are reported.

Sequence databases, usually in FASTA format, are the source of protein and peptide sequences for database search algorithms. Rather than consider all possible sequences (the province of de novo algorithms), these tools limit their search to the sequences present in the database. Typically, researchers will select a sequence database that contains all proteins known for the organism of interest from genome annotation efforts. This database can be augmented by sequences of common
contaminants. Many researchers also augment these databases by adding sets of “decoy” sequences to enable false discovery rate (FDR) estimation (15).

Because tandem mass spectra are collected for peptides rather than proteins, database search algorithms must emulate the specificity of digestion enzymes in generating potential peptides from supplied protein sequences. In the case of trypsin, this entails cutting after arginine and lysine residues, but not if those residues are followed by a proline. The digestion enzyme parameter setting directly affects the list of candidate peptides that are generated and scored against each experimental spectrum. The impact of this setting is discussed below.

Database search engines need not compare every possible candidate peptide to every MS/MS spectrum. Only the peptide sequences that have masses near the observed precursor ion must be compared to a given MS/MS spectrum. This precursor mass tolerance may be specified in the m/z, mass, or parts-per-million (ppm) domain, depending on the mass analyzer used to generate the data and tolerance options available in a search engine. Like the enzyme parameter above, this precursor mass tolerance parameter acts to filter the set of candidate peptides that get compared to each experimental spectrum.

Given a candidate peptide and an experimental MS/MS spectrum, the database search algorithm must evaluate how well a sequence corresponds to the spectrum. Recognizing the correct peptide sequence from a crowd of thousands is a significant challenge, especially for spectra that contain noise and are missing peaks. The resulting search scores are used for two distinct purposes. First, the software must rank all the candidate peptides compared to a single MS/MS spectrum such that the correct peptide appears at the top of the list. Second, these scores will be used after the search to determine the set of correctly identified peptides from the thousands of spectra subjected to the search. As a result, these scoring strategies have attracted considerable interest from a broad variety of researchers.

Scoring algorithms for comparing calculated fragment ion masses against experimental MS/MS spectra have come from the fields of engineering, computer science, mathematics, and statistics. The score used
in SEQUEST measures the extent to which experimental and theoretical spectra align. Its cross-correlation sums together the products of intensities between the observed spectrum and a spectrum predicted from the sequence. The robustness of this correlation can be enhanced by a correction calculated by shifting one spectrum relative to the other in m/z. The “xcorr” score is thus the cross-correlation with no shift minus the average cross-correlation calculated from a range of shifts. This absolute score for the spectrum is accompanied by a relative difference score between the top two scoring peptides termed the deltaCn score; it is calculated by dividing the difference between the top two xcorr scores by the top xcorr.

Mascot (16) introduced a statistical appraisal of match quality. Its Ions Score incorporates the number of fragment ions sought in the MS/MS, the number matched, the number of peaks observed within the spectrum above a threshold intensity, and the number of peptide sequences compared to the spectrum. Several counting distributions have since been used to model the likelihood of a false identification to compute a theoretical probability based on the number of fragment matches that occur between model and observed spectra. These include the Poisson distribution used by OMSSA (17), the hypergeometric distribution in PEP_PROBE (18) and MyriMatch (19), and the binomial distribution in Andromeda (20). Although estimating a probability directly from the count of identified peaks is extremely useful, these methods do not to explicitly consider any predicted intensities of calculated fragment ions. As a result, despite using similar strategies, scoring algorithms can differ strongly from each other based on how peaks are chosen for counting consideration. In many cases, these tools report the negative logarithm of the probability that this match would occur by random chance; a higher score in these tools implies a lower probability of random match rather than a higher probability of correctness.

Phenyx (21) compares two models to describe each match. Under a foreground model, the software computes the probability of the match given that it reflects the correct peptide for this spectrum. Simultaneously, it computes the probability for the match given that the peptide is incorrectly associated with the spectrum. The logarithm of the ratio between these probabilities will be positive if the peptide-spectrum match is more probable under the assumption the sequence is correct. ProteinLynx Global
Server (Waters Corporation) computes a likelihood ratio based on similar considerations (22). In contrast, Spectrum Mill (Agilent Technologies Inc.) employs a heuristic scoring approach that seeks to capture the frequency and incremental sequence-specific information content of each fragment ion series.

**Database search options and their implications**

There is no single set of search parameters that is optimal for all types of analysis across all search engines. The disparate views and opinions on MS/MS search that exist in the community are influenced by the software tools and analysis strategies that each research lab uses. Different data sets, scientific questions, search methods, search tools and downstream analyses all play a role in defining what is optimal for each query. Any parameter setting or search methodology can be optimal in one case and suboptimal for another. With this caveat, our goal here is to discuss considerations for common search parameters and highlight some issues that impact sequence database search performance for researchers new to the field.

*Sequence Database*

The first choice a user has in performing an MS/MS database search is what sequence database to search against. Most MS/MS search tools either read a FASTA-formatted sequence database directly or convert from FASTA format (23). Protein sequence databases of many organisms are readily available, though one may instead select a particular taxonomy to search from a database containing many taxa, such as UniProt (24). When the sample of interest has unknown origin, these multi-species databases are necessary. If protein sequence databases are not available for the organism of interest, searching nucleic acid databases (DNA, mRNA, or expressed sequence tags) or six-frame translations of these databases is feasible.

Only the peptides present in the sequence database may be identified by the search. Even a single amino acid difference will generally change the mass of the peptide sufficiently to prevent the base sequence from being compared to a spectrum. Researchers investigating samples that contain likely
mutations may need to use protein databases that incorporate variant sequences. Appending common contaminants to the sequence database is helpful not only to identify the contaminants present in the sample but also to explain as many MS/MS spectra as possible. When proteins have been produced by cloning their genes into other species, the proteins for the vector organism are frequently observable, and commercially available protein samples frequently contain other proteins from the same species due to incomplete purification. All of these are scenarios in which sequences may be present above and beyond the commonly considered set for a species.

The use of target/decoy database searches has become increasingly common (15). In brief, the inclusion of a large set of protein sequences known to be false enables post-identification analysis to determine the threshold necessary to achieve a target false discovery rate. Publication guidelines for most proteomic journals now require researchers to characterize error rates of identification. A separate review will address these post-identification steps in detail.

Mass tolerance

There are typically two sets of mass tolerances that are inputs to a database search: peptide mass tolerance and fragment mass tolerance. The peptide mass tolerance setting is based on the mass accuracy of the analyzer used to measure the precursor ions masses in the MS scans. In various search engines, this may be referred to as the peptide tolerance, peptide mass error, or precursor tolerance. The fragment mass tolerance is based on the mass accuracy of the analyzer used to measure the fragment ions in the MS/MS scans. This parameter can be referred to as MS/MS tolerance, fragment mass error, or product ion tolerance. Although both are mass tolerance settings, they are applied in a database search quite differently.

The peptide mass tolerance plays a direct role in determining which peptides are compared to each tandem mass spectrum. Only peptides within the specified mass tolerance are subjected to fragment prediction and comparison. The peptide mass tolerance parameter is generally contingent on the mass accuracy of the MS1 spectra, e.g. 10 ppm for Fourier Transform (FT) or time of flight (TOF) instruments.
and 2 daltons (Da) for ion traps (this low mass accuracy reflects the ion trap’s inability to resolve isotopes for sparse or highly charged peptides, see next paragraph). Because this parameter acts as a strict peptide filter, setting the peptide mass tolerance too narrow for a search, e.g. applying a 5 ppm tolerance to FT data when the actual mass error ranges to 10 ppm, will result in the correct peptide sequence being omitted from comparison to a tandem mass spectrum. In contrast, setting the mass tolerance to a value larger than deemed necessary will result in longer search times as more candidate peptides are analyzed. Larger mass tolerances may result in decreased search sensitivity because each putative peptide match must now compete against a larger pool of candidate peptides, each with a chance to randomly score higher than the correct peptide.

It is important to note that most search engines apply this parameter in mass space, that is to the neutral or singly charged peptide mass, whereas instrument mass errors are usually measured in m/z. So a ~7 ppm error for a 4+ precursor ion m/z will correspond to a ~28 ppm error in peptide mass. Similarly, a ~0.6 Da error for a 3+ precursor ion m/z will correspond to a ~1.8 Da error in peptide mass. Unless your search engine allows specification of the precursor tolerance in the m/z domain, you will need to account for the charge state scaling in the measured mass error and set the peptide mass tolerance accordingly.

For high-resolution instruments, the observed mass may be more than a few ppm away from the true monoisotopic mass of the precursor. It is not uncommon for peak detection algorithms, either stand-alone or part of instrument control software, to mistakenly assign the 13C isotope peak as the monoisotopic peak for a precursor ion, causing the precursor mass to be off by a 13C isotope mass (1.00335 Da). To address this problem, one might specify a precursor tolerance greater than 1.00335 Da to overcome the isotope mass error. However, many search engines include an isotope offset search option that effectively selects candidate peptides from multiple, narrow windows of mass around each isotope offset. This allows for an accurate mass search while also addressing possible isotope peak detection errors. In Mascot, this option is selected using the “#13C” dialogue box; in X!Tandem, it is the “parent monoisotopic mass isotope error” parameter; and in OMSSA, the “multiisotope” option for the precursor ion search type.
Fragment ion mass tolerance has generally been a less frequent target for optimization. This is a parameter that directly affects how peaks in each MS/MS spectrum contribute to its search engine score. For a predicted fragment to be matched, an observed ion in the MS/MS spectrum must be found within this tolerance of its expected location in m/z. For MS/MS scans collected in an ion trap, a fragment tolerance of 0.4 or 0.5 m/z in either direction might be applied, with spectra from more accurate mass analyzers using tighter tolerances. Search results are relatively insensitive to small changes in the fragment tolerance setting because MS/MS fragmentation patterns are fairly specific to each peptide sequence. Most reasonable fragment tolerance settings will work fine so caution in using too narrow a fragment ion tolerance value as genuinely observed fragments will fail to match, thus diminishing scores.

Enzymatic constraint

Trypsin is the most frequently employed protease for proteomic LC-MS/MS. It cleaves proteins with high specificity, creating peptides of approximately 6-35 amino acids in length by cleaving the peptide bond to the C-terminal side of lysine and arginine residues. If proline is the next residue of the sequence, digestion is far less probable (25). However, trypsin does not cut every potential site with equal probability, yielding some peptides that contain “missed cleavages” in the form of internal arginine or lysine residues. Thus a typical enzyme search setting will specify trypsin and allow for one or two missed cleavage sites. Enzymes such as chymotrypsin, AspN, and LysC are also used in proteomics experiments; these alternative proteases generate peptides that terminate at different residues than those from trypsin, improving sequence coverage when multiple digestion strategies are used.

Search engines may be configured to require one or both of the peptide termini to conform to the specificity of the protease. A “fully” tryptic search only evaluates peptides with two conforming ends, while a “semi-tryptic” search considers peptides that have either one or two conforming ends. “No enzyme” or “unconstrained” searches consider all possible peptides from a database. Possible motivating factors for performing semi-tryptic or no-enzyme searches are outlined below in the sections titled “Need for sufficient search space” and “Applying accurate mass and enzyme constraints during vs. post-search”. For now, we will simply note that there are exceptions to the tryptic digestion rules listed above resulting
in peptides exhibiting unspecific cleavages (i.e. semi-tryptic). Digestion protocols do vary, either intentionally or not, and this does alter the distribution of resulting peptides with respect to expected cleavage specificity. For example, Strader \textit{et al.} identified a correlation between digestion specificity and the solvent in which the digestion was performed (26). In-source fragmentation can result in acquisition of MS/MS spectra of semi-tryptic peptides. Lastly, other unexpected proteolytic activities in the sample may be at work. So if there is a goal of identifying every possible peptide, semi-tryptic or even no enzyme searches might be employed.

\textit{Modifications}

The ability to identify peptides containing post-translational modifications (PTMs) is a powerful tool for studying complex regulatory mechanisms in cells. Even routine searches typically allow for modifications resulting from common sample handling chemistry. In the most common implementations of modification support in MS/MS search algorithms, accounting for PTMs simply involves a mass change to one or more amino acid residues. There are two classes of modifications that a user can specify which have different effects on searches.

The first type of modification is where the mass of a given amino acid residue is simply changed to a different mass. This modification is termed static or fixed. One common static modification that is applied in many proteomics searches is the addition of 57 Da to cysteine residues to account for cysteine alkylation. Because static modifications simply involve an amino acid mass change, there’s little impact on the database search itself in terms of run times and the actual search mechanism. The second type of modification is where a mass change sometimes occurs on a particular residue. This modification is termed a variable or a dynamic modification and an example of such a modification would be the addition of 16 Da to methionine due to oxidation. Searching for serine or threonine phosphorylation would be another common application of a variable modification search. Because a search engine must test all combinations of modified and unmodified residues in a variable modification search, the effective search space of candidate peptides is increased, leading to increased search times. Applying multiple variable modifications compounds the problem significantly and should be avoided if possible.
When a peptide contains more than one residue capable of bearing a modification, such as a single phosphorylation present in a peptide with multiple serine and threonine residues, the confidence with which the modification can be localized to a particular residue is dependent on the presence of fragment ions in the MS/MS spectrum derived from fragmentation between the possible sites of modification. Mainstream search engines currently pay little or no attention to representing the certainty or ambiguity of modification localization in their top-scoring result. In a 2010 phosphopeptide identification study conducted by the Proteome Informatics Research Group (iPRG) of the Association of Biomolecular Resource Facilities (27), nearly all of the studies’ 35 participants employed a post-search, auxiliary tool such as Ascore (28) for measuring the certainty of phosphorylation site localization for each peptide spectrum match. It is foreseeable that some auxiliary score for modification localization will be incorporated into forthcoming versions of most search engines (29). Two good resources for PTMs are Unimod (30) and RESID (31).

**Need for sufficient search space**

When performing a sequence database search with a given tandem mass spectrum, most software frameworks assume that either a single, correct peptide matches the spectrum or that no correct match to a spectrum can be found in the database for a particular search configuration. Given a set of candidate peptides being analyzed in a search, the premise is that all candidate peptides are wrong except for possibly one correct match. This population of wrong candidate peptides is important because the scores for these peptides are frequently used to evaluate the significance of the top scoring hit. Determining a correct identification can boil down to the question, “Does the top ranked peptide score significantly better than all of the other candidate peptides in the search?”. If so, then that peptide is an outlier of the wrong population and thus is likely a correct match. The deltaCn score in SEQUEST compares the top two peptides matches to a spectrum. Tandem’s E-value is based on the score distribution for all matches to the spectrum (32). For these and many other search engines, the total set of candidate peptides and their score distributions play a key role in determining whether the top hit from any given query is likely wrong.
(with a score that is similar to other candidate peptide scores) or likely right (with a score well separated from the other scores).

Because various scores and metrics are calculated based on the population of candidate peptides being analyzed, it is important to include a sufficient number of candidate peptides to prevent a skew in the calculated statistics. This is of particular concern when using tight precursor mass tolerances that limit the number of candidate peptide comparisons. Using narrow ppm tolerances with a full enzyme search can result in matching single digit or a few tens of candidate peptides for each spectral query and has been shown to result in dubious identifications in published proteomics analysis (33). This is not to imply that all such searches are full of dubious identifications but rather to inform of the potential problem.

**Applying accurate mass and enzyme constraints during vs. post-search**

There are two primary ways to take advantage of precursor mass accuracy in MS/MS data analysis. One method practiced by many users is to set the peptide mass tolerance to a narrow value thus only considering candidate peptides that fall in the accurate mass range. The other method, also practiced by many, is to apply the accurate precursor mass measurements during post-search analysis of the search results. This implies the primary search is performed using a relaxed or wider peptide mass tolerance. There are advantages and disadvantages to both methods. Using accurate mass during the primary search may lead to issues discussed in the previous section regarding suspect statistics due to sparse peptide search space. Searches will be faster because the search space is small. And one might identify correct peptides for spectra that might otherwise match a higher scoring incorrect peptide when querying a wider search space. Using a wider tolerance during the primary search will make searches run longer and there’s potential to miss identifications because correct peptides are competing against many more candidates. But a wide mass tolerance search allows one to filter search results not only by search scores but also by mass accuracy. Positive identifications should be enriched at the accurate mass measurements while incorrect identifications should be evenly distributed across the entire mass tolerance range.
Zubarev and Mann propose methods to take advantage of mass accuracy, including more intelligently assigning mass error based on individual signal quality and incorporating mass deviation in the database search scores (34). Hass et al. noted that high mass accuracy search benefits identification for low quality spectra, especially for phosphorylated peptides where fragment ion are low in intensity compared to phosphoric acid and water neutral loss peaks (35). Brosch et al. evaluated Mascot and X!Tandem and show that Mascot is more sensitive to changes in the peptide mass tolerance parameter compared to X!Tandem and show benefit to applying accurate mass as a post search filter to a relaxed tolerance search (36). Hsieh et al. also demonstrate that post search accurate mass filtering yields a greater number of identifications at a given false discovery rate using SEQUEST (37). Cottrell warns against setting too narrow a precursor tolerance in order to test a sufficient search space but recommends avoiding relaxing the enzyme constraint (38).

From a data analysis perspective, relaxing the enzymatic cleavage constraint has similar implications to relaxing the precursor mass tolerance. The candidate peptide search space is increased greatly in semi-tryptic and no-enzyme searches so there is a significant, real cost in terms of increased search times associated with such analyses. Additionally, with the large increase in candidate peptides considered for each spectrum, there is a potential loss in sensitivity when an incorrect semi-tryptic sequence randomly out-competes the correct tryptic sequence for a spectrum. One reason to consider relaxing the enzyme constraint is that semi-tryptic peptides can now be identified whereas they would not even be considered in a fully tryptic search (39). And as noted, applying the enzyme constraint as a post search analysis filter, either explicitly or as part of tools like Percolator (40, 41) and PeptideProphet (42), could possibly identify more peptides compared to using the enzyme constraint during the primary search (43).

As there is clearly no consensus on what search strategies and analysis methods are optimal, each user is encouraged to understand the consequences of each option. Targeting the identification of a single, low abundance protein requires different analysis criteria than surveying thousands of proteins. Ideally one performs an unbiased evaluation of analysis strategies to discern what method works best for a
particular query. For a larger scale, shotgun analysis, making use of target/decoy database searches to estimate the number of correct identifications at a fixed false discovery rate is one nice option perform this evaluation.

**Multi-pass search considerations**

One mode of MS/MS analysis that has gained some prominence recently is the ability to more thoroughly query a small subset of identified proteins looking for PTMs or peptides with relaxed enzyme constraint. This is attractive as it allows typically expensive calculations to be computed quickly because of the greatly reduced search space. This mode of searching is implemented as “error tolerant” mode in Mascot (44), “refinement” mode in Tandem (45), “unassigned single mass gap” mode coupled with searching previous hits in Spectrum Mill, and OMSSA’s interactive search. In a related strategy, the Paragon algorithm in ProteinPilot (46) doesn’t focus on subset proteins but rather determines which sequence regions to evaluate more thoroughly during a search using a combination of statistics (using sequence tags to compute Sequence Temperature Values) and search feature probabilities.

This method enables the thorough interrogation of data by addressing the computational expense of PTM database searching. However, users should show caution in interpreting such search results. When searching unmatched spectra from the first pass against a small population of proteins, the statistics of the searches can be skewed by too few candidates, just as when using tight precursor tolerances against small databases (47). Each modification allowed for a candidate peptide may induce a fit in its predicted fragments that allows it to be matched to a spectrum of an unrelated peptide. As a result, false hits are likely to be augmented for modified peptides. Normal database search and a multi-pass or refinement mode search differ in that the false positives in the former are distributed across all database entries; refinement searches force false positives to occur for the subset of plausible proteins, creating a sense of false confidence in otherwise dubious matches. Everett et al. discusses this problem and proposes a computational solution with respect to decoy database searching and X!Tandem’s refinement mode(48); Bern and Kil propose a follow-up improvement (49).
Peptide identifications are just a start

Tandem mass spectrometry sequence database searching is an early step in the proteomics analysis pipeline. Peptide identifications and subsequent protein inferences, either performed in a single tool or as separate analysis, are not necessarily the end product of most proteomics experiments. Many of the following topics will be addressed in separate review articles, but they are worth mentioning here. The statistics for assessing identification data, whether evaluating individual peptide-spectrum matches (32, 42, 50), stratifying confident identifications from likely errors across an LC-MS/MS file (40, 51, 52), or computing probabilities associated with proteins and isoforms (53-55), are key to discerning the value of these findings. Combining identifications from multiple search engines to more broadly interrogate the MS/MS data is commonly performed by software tools such as PEAKS inChorus (Bioinformatics Solutions Inc.), Phenyx (GeneBio), Proteome Software’s Scaffold (56), and the Trans-Proteomic Pipeline’s iProphet (57). A wide variety of tools adds more biological context to the list of identified proteins; a small sample of these include ProteinCenter (Thermo Fischer Scientific), Kyoto Encyclopedia of Genes and Genomes (58), PANTHER Classification System (59), Cytoscape (60), and WebGestalt (61).

Conclusion

Database search is the mainstay of proteome informatics, linking experimental tandem mass spectra to biological sequences. Appropriate sequence databases and search configurations prevent reduced identification rates and excessive search times. While simple sets of PTMs may be identified through these algorithms, researchers may need dedicated software tools for data sets in which broader palettes of modifications are present (62). Database search tools are most powerful when integrated into pipelines with other tools, such as those that perform quantitation and sophisticated protein inference. Translating collections of identified peptides into confident protein lists requires appropriate filtering and protein assembly tools. Even with an up-to-date software pipeline, however, a large proportion of collected
MS/MS scans may remain unidentified. As the field of proteome informatics evolves, identification rates will continue to climb.

Acknowledgements

DLT was supported by NCI CPTAC cooperative agreements U24 CA126479, U24 CA0126477, and U24 CA126476. This work is supported in part by the University of Washington's Proteomics Resource (UWPR95794).
REFERENCES

Figure 1. A snapshot of MS/MS data acquisition. A schematic showing the relationship between MS scans (blue) and MS/MS scans (red) in a typical tandem MS experiment. In this example, a precursor ion in each MS scan is isolated and fragmented, resulting in the MS/MS spectra that alternate every scan with the MS spectra. Typical data acquisition would acquire multiple MS/MS scans for each MS scan. In addition to the fragmentation pattern from each MS/MS spectrum used in a database search, the precursor ion’s m/z, charge state, and mass accuracy of the measured precursor m/z are obtained from the MS spectrum.
Figure 2. Data acquisition and database search. For a given experimental MS/MS spectrum, protein sequences from a database are in silico digested and peptides of the right mass are selected. Theoretical fragment ions from each of candidate peptide are calculated and used to generate a similarity or probability score by comparing the theoretical fragment ion masses against the experimental spectrum. Each candidate peptide is scored against the experimental spectrum and the best matching peptides and their scores are reported.