

A Review on Common Applications Designed to Optimize the SEQUEST Search Algorithm With Respect to Limiting False Positive Identification

Brandon Williams
Biochem 218

Introduction:

New proteomic technologies have emerged that deal with understanding complex systems and diseases at a global level, thus allowing an expanded experimental view. Before the advent of mass spectrometry, proteomic approaches were initially explored via the use of two-dimensional gel electrophoresis (1) and the application of chromatography techniques (2). Unfortunately these techniques proved to be limited to detecting only small molecular mass. Because of the difficulty in identifying large macromolecular structures such as individual proteins in the high molecular weight range, their modifications or protein complexes, suppressed the broad application of proteomics. This technological hurdle has for the most part been overcome through the application of mass spectrometry (MS) when used in combination with developing proteomic databases. Furthermore, algorithms have been developed that allow the identification of spectra in an automated fashion, reducing the need for manual validation of hits from spectral data. Although, this process has greatly facilitated the development of mass spectrometry with respect to proteomics, the process does come with drawbacks such as false positives that fit search parameter from a theoretical database that is used to perform cross correlation analysis with an experimental spectrum. SEQUEST is an algorithm designed to reduce the false positive identification rates with respect to mass spectrometry. The goal of this review is to enlighten the reader on the more significant improvements made to increase the efficiency of the SEQUEST algorithm.

Principles of Mass Spectrometry:

The basic route taken when measuring the characteristics of a sample using MS requires the application of three processes; ionization of the sample compound, which usually results in the formation of cations or anions, the sorting and separation of the compound by an electric or magnetic field based on the mass to charge ratio (m/z), and finally, detecting the separated ions with a detector apparatus. Since mass spectrometers create and manipulate gas-phase ions, they operate in a high-vacuum system. Data analysis is obviously a requirement if you wish to interpret the data, but this aspect is separate from the actual MS apparatus (3). The overall process is illustrated in figure 2 which shows the basic scheme associated with measuring a compound by MS. The most critical element of MS is the ionization of the compound being analyzed.



Figure 1. Schematic of the general process involved in determining the molecular characteristic of a molecular sample when using mass spectrometry

This process of MS, involves converting the individual molecules of a compound into ions so that they can be moved about and manipulated by external electric and magnetic fields. The actual physical approach involves generating an ionic state of the analyte by shooting electrons at the compound. When an electron collides with the molecule it will often “ionize” it by knocking away one of the molecular electrons (figure 3).

This process will not only affect free electrons (non-bonding) but electrons that are locked in bonds as well. It is also important to

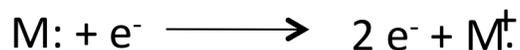


Figure 2. The formation of an ion occurs from the collision between a compound and another electron. This particular diagram depicts a neutral molecule becoming a cation after collision with an electron.

mention here that residual energy from the collision of the compound with an electron can cause the molecules to fractionate into different isoforms (3). Figure 4 illustrates this scenario and depicts a molecule (M) fractionating into either an even-cation and a radical, or an odd cation and another neutral molecule. The individual products of ionization can give different signals within mass spectra data. An example of this can be viewed in figure 5.

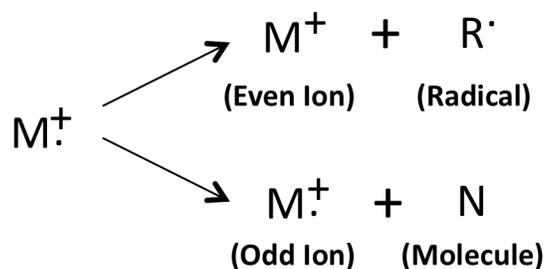


Figure 3. Residual energy from the collision between an electron and a compound will result in different fractionating patterns of the compound.

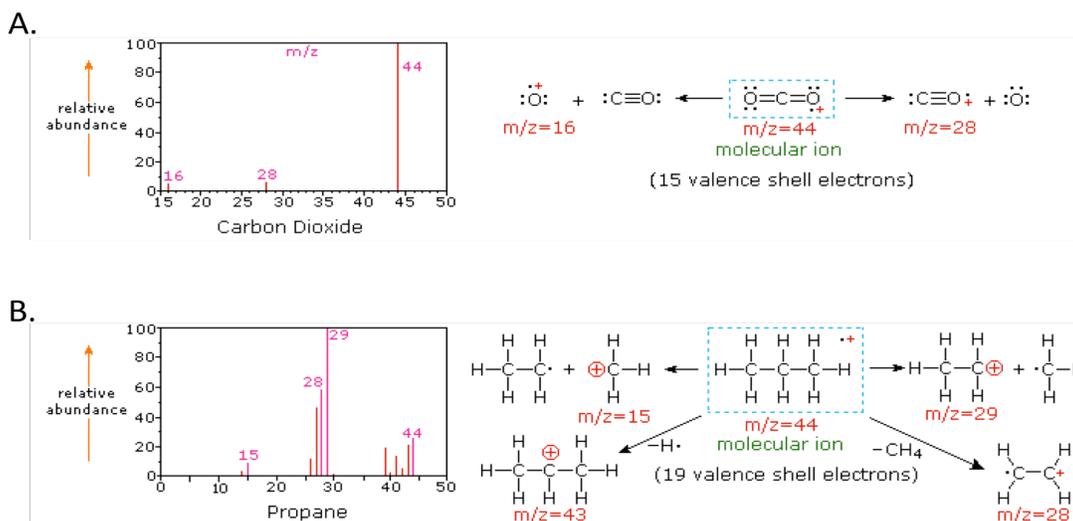


Figure 4. Molecular compounds will fractionate into different products upon ionization, and their relative abundance will be displayed on the mass spectra data. (A) Representative spectra of carbon dioxide. (B) Representative spectra of propane. (The spectra are courtesy of: <http://www.cem.msu.edu/~reusch/VirtualText/Spectrpy/MassSpec/masspec1.html>)

Mass spectrometry has the ability to detect the overall abundance of ionized products (3). The spectra of both carbon dioxide and propane are shown in figure 4. The ionized products are represented in a vertical bar graph X is the mass-to-charge ratio (m/z) and Y is the relative abundance. Within the spectral data, the ion with the highest intensity is always assigned an abundance of 100, and is referred to as the base peak (3). Most of the ions formed from the ionization process will have a single charge, thus the mass-to-charge ratio will be equivalent to the actual mass of the ion. Furthermore, the ion with the highest m/z is generally an ionized compound that has not been fractionated. Spectra with a lower m/z are the fractionated products.

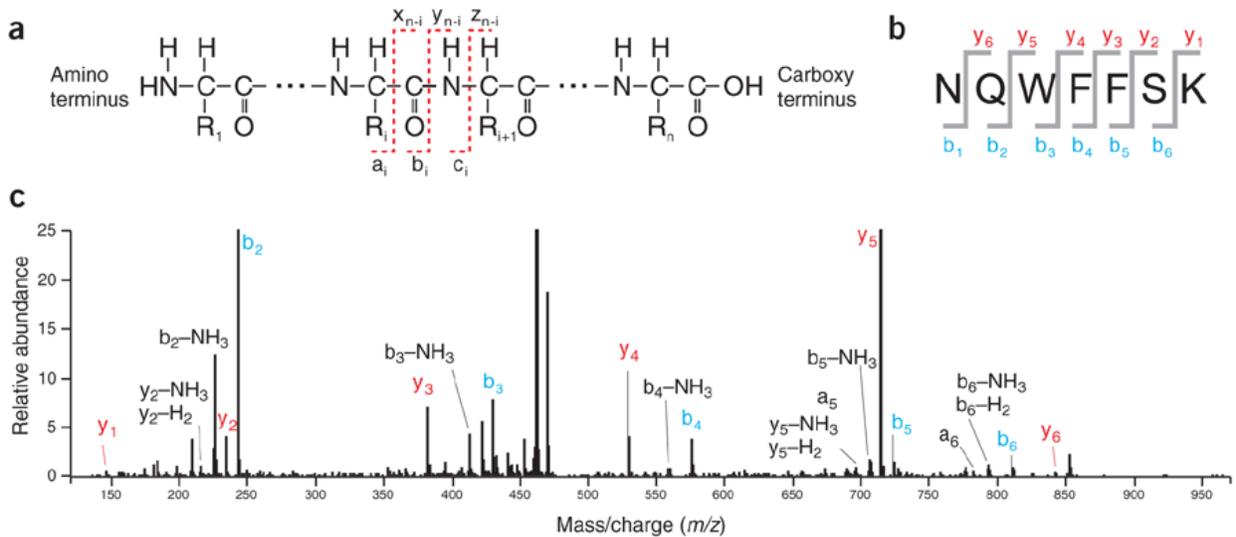


Figure 5. Shotgun proteomics identifies proteins from the fragmentation mass spectra of their respective peptides. (a) Proteins are fragmented into smaller peptides in the mass spectrometer at specific bonds within the amino acid sequence. (b) Computer algorithms focus on the b- and y-ion series as these are usually the most common type of fragmentation patterns obtained from MS/MS. (c) Representational spectra for the amino acid sequence listed in (b) (4).

Shotgun proteomics has proven to be a remarkably powerful technology for identifying proteins. Analysis can be done with individual, purified protein samples, or from samples that contain multiple species of protein such as a cell lysate (4). This technique is often performed using Tandem mass spectrometry (MS/MS). A tandem mass spectrometer can be thought of as two mass spectrometers in series connected by a chamber that can break a series of proteins into fragment ions. This chamber is known as a collision cell. A sample is "sorted" and "weighed" in the first mass spectrometer, then broken into pieces in the collision cell, and a piece or pieces are sorted and weighed in the second mass spectrometer (5). The fragmentation that results from a low energy collision induced dissociation (CID) process usually follows a pattern where the peptide breaks along its backbone between adjacent amino acids. Breakage at the bond between the alpha carbon and the carbonyl carbon produces the a-ion and x-ion series of fragments. Breakage at the carbonyl carbon-amide nitrogen bond produces the b- and y-ion series, and is often the most common bond broken upon fragmentation. Finally, breakage at the amide bond produces the c- and z-ion series (4). The fragmentation of peptides occurs randomly, thus a sequential N-terminal and C-terminal peptide ladder can be constructed due to the detection of different mass/charge ratios for each fragment ion. Identical proteins in a sample that is subjected to low energy (CID) can be used to generate such a fragmentation pattern revealing the sequence of the protein.

Generally, there are two techniques used to detect proteins and peptides from an experimental sample; these are MS peptide mass fingerprinting (PMF) and MS/MS peptide fragment fingerprinting (PFF) spectra. In MS (PMF) the mass pattern obtained from measuring the masses of purified or simple protein mixture, enzymatically digested or chemically cleaved, is then compared to theoretical mass patterns generated *in silico* from a protein database (6-8). In MS/MS both intact mass of the peptide (parent ion mass) and fragment ions are recorded. When PMF fails, uninterpreted fragment ion masses can be used in correlative database searching to identify proteins whose peptides would likely yield similar CID spectra under the same fragmentation conditions. Probability-based matching is used here by algorithms such as SEQUEST (9). SEQUEST, in an automated fashion, converts the character based representation of amino acid sequences in a protein database to a fragmentation pattern that is used to match fragment patterns from an experimental spectrum as explained above (9) With the advent of this algorithm and other like it, peptide mass mapping, database searches using tandem mass spectrometry (MS/MS), *de novo* sequencing followed by homology search and validation of protein identifications has almost become entirely hands off allowing for large high throughput studies with respect to proteomics.

The SEQUEST Algorithm:

(Eng et al., *J. Am Soc Mass Specgtrm* 1994, 5, 976-989)

SEQUEST is a method used to correlate the uninterpreted tandem mass spectra of peptides produced under low energy (10-50 eV) collision conditions with a theoretical spectrum consisting of amino acid sequences obtained from protein data banks. The computer algorithm developed by Eng et al., identifies peptides from experimental tandem spectra by first performing an initial reduction of the spectral data to reduce background noise and removing all but 200-500 of the most abundant fragment ions. Additionally, a 10-u (one u is roughly equal to one Dalton) window around the precursor ion is removed to ensure that a predicted fragment with a similar ion-to-mass charge ratio is not matched to the precursor ion. Once background noise from the spectra is removed, an initial scan of protein sequences whose mass falls within the mass-tolerance from a specified database is performed indentifying linear sequence combinations in an N-terminal to C-terminal fashion. Mass-to-charge ratio values of theoretical peptides obtained from a database search are calculated as follows (equation 1). In this equation a_n represents the mass of the amino acid.

$$\text{For type-}b_n \text{ ions: } b_n = \sum a_n + 1$$

$$\text{For type-}y_n \text{ ions: } y_n = MW - \sum a_n$$

Sequence evaluation and scoring of identified peptides is determined by several different criteria. Fragments ions that fall within a +/- 1 u are recorded as well as the sum of their relative abundances (n_i and i_m respectively). The algorithm also takes into account the continuity of consecutive matches for a fragment within the ion series (β) and can be observed below:

$$S_p = (\sum i_m) n_i (1 + \beta) (1 + \rho) / n_t$$

Here, (ρ) represents the presence of ammonium ions and these values are normalized to the total number of predicted sequences (n_t) to generate a score for each hit, S_p . This S_p value is then used to identify the top 500 hits from the experimental spectra. As explained previously, SEQUEST is an algorithm that has the ability to

generate a theoretical library using simple search parameters to perform a cross-correlation analysis. This theoretical database is then used to perform a cross correlation search with the experimental to generate an XCorr value for each hit or protein specie identified. To compute this score, the SEQUEST algorithm first normalizes the experimental spectrum by assigning a magnitude of 50.0 to each predicted mass-to-charge ratio values for type-b and $-y$ ion. Fragment ions within $\pm 1u$ are assigned a magnitude of 25.0 and fragments that exhibit a neutral loss of water and ammonia such as those from a- type ions, are assigned a magnitude of 10.0. This normalized experimental spectrum serves as the X value for the cross correlation function and the theoretical serves as the Y value. The SEQUEST's XCorr value is then computed via the correlation function $Corr(t)$, or the product between the vectors X and Y, and Y shifted with respect to X along the m/z axis by t mass units. This calculation is shown below:

$$(a) \text{Corr}(t) = \sum_i x_i y_{i+t}$$

$$(b) \text{Xcorr} = \text{Corr}(0) - \langle \text{Corr}(t) \rangle$$

Here, the nonzero elements x_i and y_i represent the peaks from the processed (normalized) experimental and theoretical spectrum. A simpler explanation is that the XCorr value is essentially a summation of the number of fragment ions that are common between the X and Y vectors. This can also include small variations such as some mass measurement errors are assumed to occur. Finally, this score is corrected to account for the number of matches that are occurring at random as shown using the formula above (b). Generally, values ≥ 1.9 for +1 ions, \geq for +2 ions, and ≥ 3.75 have been classified by others as significant (Link 1999 and Washburn 2001...Qscore) Finally, SEQUEST computes the relative difference between the first and second XCorr values termed ΔC_n . Generally, high values are regarded as a correct assignment with respect to the top Xcorr value. Although SEQUEST has been shown to be very successful with the identification of protein species from experimental analyte, it does have its faults. Like other algorithms that are based on additive features (for example, the continuity that is taken into account for the S_p value), it becomes dependent on the mass of the peptide, the overall charge state of the ion fragments, and the overall quality of the spectra. Figure 6 (right) displays a graph that plots the correlation between peptide length and Xcorr values. An obvious trend can be observed with respect to peptide length and correct peptide assignment.

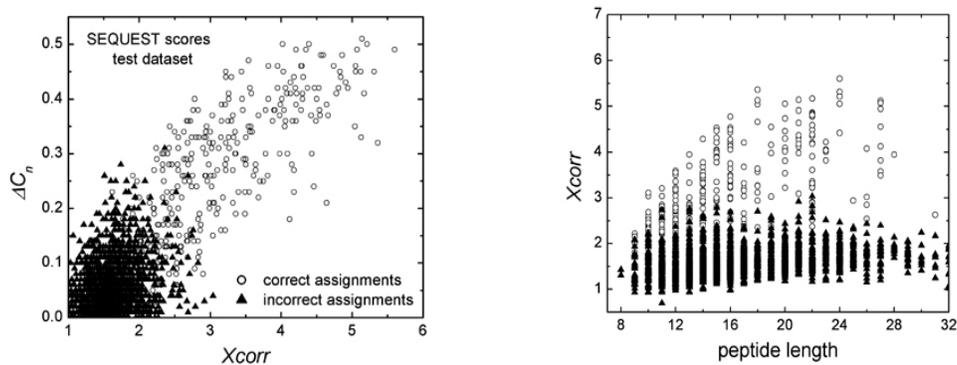


Figure 6. The graph on the left displays the correlation of ΔC_n and Xcorr. As mentioned previously, the higher the difference between the first and second Xcorr values, the higher the probability that the first Xcorr is correct. The graph on the right shows the correlation between Xcorr values and peptide fragment length with respect to the correct assignment of a peptide from experimental spectra. This plot suggests that the Xcorr value is dependent on peptide fragment length generated from a low energy CID (10).

Current Dilemma in Database Search Algorithms, “The false positive”:

Although, search algorithms have made the process of protein identification extremely efficient, problems still persist that often require some manual assistance. One of the most significant problems that are encountered when using programs such as SEQUEST that are designed to identify protein species from mass spectral data, is failing to assign the correct peptide sequences from a search (false positives). It is totally plausible that the best matched returned by the database search tool is not the correct one (10-12). In fact, one can expect that a large fraction of the experimental spectra generated by MS/MS, are not matched to a sequence database. This issue is the result of many reasons that include deficiencies in the scoring scheme, low MS/MS spectrum quality, fragmentation of multiple peptide ions, presence of homologous peptides, incorrectly determined charge state of peptide mass, restricted database search and sequence variants of novel peptides. Deficiencies within the scoring scheme, if not designed well enough, hamper the process of true peptide identification and often elicit the identification of the incorrect peptide as a result of a higher score generated than the correct peptide or protein. Most databases are designed to generate a simplified representation of the peptide ion fragmentation process. Early aspects of spectra search algorithms when generating a theoretical spectrum, often assume that all ion fragment peaks are present in equal intensities. But, anyone who has experience using mass spectrometry knows that the fragmentation patterns vary greatly with respect to the amino acid sequence. For example, the presence of a proline in a peptide sequence can result in a high fragmentation pattern corresponding to the breakage of the N-terminal side of the proline, thus skewing the relative frequency of fragmented peptides (13, 14). Low MS/MS spectrum quality can also hinder the identification process for the correct peptide sequence present from an experimental analyte. A spectrum of low quality will result in noise peaks, low signal-to-noise ratios or even have missing fragment peaks owing to incomplete peptide fragments. Finally, when using shotgun proteomics, the number of low-quality spectra in a dataset can be high as a result of various contaminants that may be present in the experimental analyte, such as non-specific proteases (15).

Another common reason for incorrect protein identification is the variation in the charge state of the fragment ions or as well as the mass of the peptide. If the charge state of an ion cannot be determined, then a search is again performed against the database but with the assumption that the fragment has a +2 or +3 charge state (10). But if the charge state is actually +4, or an incorrect classification of charge state is made by the analyzing software, then the correct peptide will not be found and possibly an incorrect peptide is considered a good match. Finally, one of the most significant reasons for the generation of false positives is an design of the search parameters used that in effect restricts a database search. For example, database searches that take into account many type of post-translational modifications of often ignored, excluding modifications that are considered common, due to the fact that such a search can be so time consuming. Furthermore, databases are often searched with the specifications towards one type of digestive product from a particular protease (as mentioned above). A spectrum will not be identified if it is produced by a peptide containing an unspecified modification, resulting from an unexpected protein cleavage, or containing more than the allowed number of missed cleavage (16).

For many MS/MS spectra, the best scoring peptide assignment returned by a database search algorithm is incorrect. Thus, manual validation of the search results has to be performed in order to filter out the false positive (9). Although very time-consuming, manual validation can be performed on small scale levels, such as a single species or a highly purified sample of protein. Even then contaminants can be present, but such spectra can be teased out by a trained mass spectrometry specialist. But, Manual validation of a large datasets containing tens of thousands of spectra though, is simply not feasible. Instead, manual validation is replaced

with an automated or semi-automated fashion that generates a database search score representing “hits” by programs such as SEQUEST (9). The next section takes a look at some of the methods that were previously used to optimize the SEQUEST search algorithm.

Recent Methods Used to Decrease False Positive Hits from a SEQUEST Search:

(MacCoss *et al.*, *Anal. Chem.*, 2002 74 (21), 5593-5599)

Earlier work that focused on increasing confidence with respect to hits generated by SEQUEST addressed the problem of non-specific cleavage of sample peptides by non-specific proteases. The XCorr values produced by SEQUEST are not only database size-dependent, but are also dependent of the quality of the spectra and the quality of the fit to the model spectrum as well and non-specificity of fragments can increase false-positive hits with respect to these dependencies. An effective method to decrease the false positives identifications relative to true positive hits was employed by Yates *et al.*, which provided a probably-based validation for peptides hits generated by the SEQUEST algorithm. Before the advent of algorithms that focused on enriching true positives from a SEQUEST search, scores were essentially normalized using the difference between the first- and second-ranked sequences (ΔC_n). This proved inefficient since an XCorr value for one peptide may not reflect a similar closeness of fit for another peptide with the same score. Therefore a scoring parameter was needed that functioned specifically to maintain consistency for peptides regardless of how the ion fragments are generated (i.e. variable proteases or different peptide separation techniques), or how large the peptide is. Like SEQUEST, SEQUEST-NORM (normalized cross-correlation scoring algorithm) identifies the “best” 500 amino acid sequences from a protein database and performs a cross-correlation between the theoretical spectrum and the experimental tandem mass spectrum data. The major difference between SEQUEST and SEQUEST-NORM is that the cross-correlation scores are then normalized to the cross-correlation of the input spectrum against itself (autocorrelation), effectively identifying the best possible match between the experimental spectrum and a theoretical spectrum from a protein database. Validation of this autocorrelation was performed using the following equation:

$$\text{Pept}_{\text{Prob}}^{\text{XCorr}} = \sum_{i=0}^{\text{XCorr}} f_i$$

At variable intervals, XCorr values will result in either a correct match or incorrect match which is represented by f . The sum of all the incorrect matches is then used to evaluate the confidence of the peptide sequence identified by SEQUEST-NORM. What this essentially means is that as the XCorr value increases, the probability that the identified peptide is correct increases and consequently the likelihood of a given peptide being a false positive, decreases. In order to validate the normalization of the SEQUEST-NORM algorithm for individual peptide matches with respect to the identification of a particular protein in the experimental spectra, a test to determine how multiple peptides that are normalized would affect the identification of a protein from experimental spectra as a whole. For this particular goal, the equation below was used.

$$\text{Prot}_{\text{prob}} = 1 - [(1 - \text{Pept1}_{\text{Prob}})(1 - \text{Pept2}_{\text{Prob}})(1 - \text{Pept3}_{\text{Prob}})(1 - \text{Pept}n_{\text{Prob}})\dots]$$

The final product of a SEQUEST-NORM search will result in more true positive hits as a result of removing the dependence on peptide length, which can occur with a SEQUEST search. One major advantage of SEQUEST-NORM over SEQUEST as a result of the normalizing process in conjunction with a probability based calculation, is that it does not rely on cleavage specificity as an input parameter. This effectively

decreases the trend of peptide mass and XCorr values that often occurs with SEQUEST. A comparison of XCorr values generated by peptides that were subjected to multiple proteases in order to simulate non-specific cleavage, was made between SEQUEST and SEQUEST-NORM. A linear plot of SEQUEST and SEQUEST-NORM showed that the normalized SEQUEST-NORM search results did not result in a correlation of peptide mass and XCorr values as SEQUEST search often does (XCorr values tend to increase as peptide size increases). Furthermore, a search of spectra with SEQUEST-NORM of peptides generated by multiple proteases showed no increase in mean XCorr values of ions with a charge state of +2 and +3 with respect to false positives. Finally, search was performed on a total of 215, 419 tandem mass spectra to determine a frequency of correctly identified peptides and incorrectly identified peptides versus XCorr values after applying a second-ordered Savitsky-Golay filter. This showed that the correctly matched XCorr values have a Gaussian distribution (where $r = 0.972$), versus incorrectly matched XCorr values that are predominantly skewed toward higher XCorr values suggesting that the normalizing SEQUEST algorithm can indeed enrich true positive hits from a tandem mass spectral search.

(Moore et al, J. Am Soc Mass Spectrom 2002, 13 378-386)

Another earlier development that focused on enriching true positive hits from a SEQUEST search was based on taking advantage of the multiplicity of matches from a tandem mass spectrum and determining the quality of the match. The algorithm, Qscore, takes into account that SEQUEST will always return a best match peptide as long as at least one peptide from the database falls within the peptide mass tolerance. Up to this point the quality of this match was still not able to be determined. Furthermore, the quality of the match could be so low that the overall score for the protein is worse when it is included, further reducing the reliability of SEQUEST hits. Therefore Moore et al developed a statistically based algorithm used to determine the goodness of a protein match that is able to ignore SEQUEST matches that effectively decrease the scores of true positives. As mentioned before, it is important to determine if there is a reasonable chance that a match could have come about by chance. Here, Moore explains that even the smallest non-viral database can generate a hit from a single peptide match from a SEQUEST search. Therefore, he postulates that “a null hypothesis can be adopted which specifies that all matches are essentially random and accept as identified on those proteins that generate more matches than would be expected by this null hypothesis”. The first aspect of the algorithm determines the expected number of matches for a particular peptide is as follows:

$$N_{\text{match}}(M, N, P) = P^{(1-M)}N!(N-M)!M!$$

Here, N is equal to the number of individual searches, M is equal to a number of matches against a specific protein, and P is equal to the number of proteins in a database. Although this formula is designed to determine the numbers of expected matches from a database search with the number of matches increasing as the number of spectra are searched, several factors can generate more expected results than assumed. For example, a single search of a database can often result in more than one protein match. One particularly difficult trend that must be accounted for is the representation of peptide fragments in very large proteins. This can happen multiple times in several unrelated species due to the simple fact that there are a limited number of possible sequences. Large proteins could potentially contain hundreds of similar peptides and depending on the database size, can effectively skew the number of predicted matches. To get around this dilemma, Moore et al, devised a second component of the Qscore algorithm that could account for the discrepancies of the database size. By performing a theoretical digestion of the entire database and counting the number of distinguishable peptides derived, a comparison can be made between the number of identified peptides in the database and the

number of peptides fragments generated from a theoretical database. Accounting for database size discrepancy is determined as follows:

$$N_{\text{match}}(M, N, p, d) = (d/p)^{(1-M)} * N! / (N-M)! M!$$

Here, (d) is equal to the number of peptides identified in the database and (p) is equal to the number of peptides that are generated from a theoretical digestion. The final aspect of the Qscore algorithm takes into consideration the actual quality of peptide matches. Because a strong match is less likely to be produced from a false positive, it is possible to include this information into the algorithm and essentially generate a formula that excises out such scores. Thus we get the Qscore, an algorithm that can greatly reduce the number of SEQUEST protein identifications that have to be validated manually, and is computed as follows:

$$\text{Qscore}(M, N, p, d) = -\log(d/p) + \sum_{i=m}^M \log[(i * q(i) * (d/p)) / (N+1-i)]$$

Here (i) or q(i) is equal to the match quality information for the i^{th} peptide.

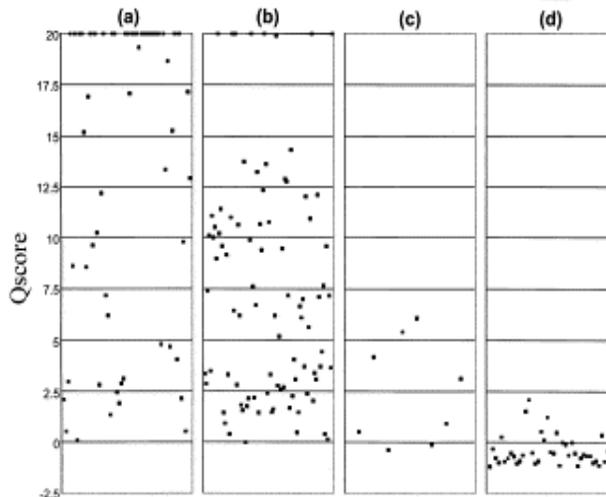


Figure 7. This plot shows the distribution of Qscores for proteins identified manually (a), those that are considered contaminants (b), those that are considered present in the sample but were missed even with manual validation (c) and finally, false positives (d).

In order to evaluate the Qscore algorithm, Moore et al performed a controlled search with roughly 1800 spectra in conjunction with SEQUEST. This search resulted in the identification of 58 matches, 88 contaminants, 8 misses, and 42 false positives. The Qscores for the matches ranged from a minimum of 0.14 to 467.01 and false positives ranged from -1.1 to 2.12. These values and categories can be observed in figure 7 showing the effectiveness of the Qscore algorithm.

(Razumovskaya et al., Proteomics 2004, 4, 961-969)

Separation of false positives from true identifications is a difficult task when using peptide identification algorithms. SEQUEST primarily uses the Xcorr and the number of charges to filter out said incorrect profile matches. Unfortunately this can be misleading as an X correlation score for a particular peptide may have a different meaning for another peptide that differs in length and charge as mentioned previously. Furthermore,

the distribution of the X correlation scores for correct and incorrect hits, often do not have a clear separation. Therefore manual interpretation of the top hits derived from the SEQUEST scoring scheme is required to identify any false positives. In work described by Razumovskaya et al., a new scheme has been developed that focuses on assessing the reliability of peptides identified via SEQUEST by combining a neural network and a statistical model that normalizes SEQUEST scores. This method has been shown to provide a reliability estimate for each SEQUEST hit and in general, is accomplished by first normalizing peptide scores and statistically estimating the probability that the peptides identified by SEQUEST are correct. These statistical validations are then applied to the identified peptides by SEQUEST, limiting false positives in an automated manner.

The effectiveness of this method stems from the use of a neural network. An artificial neural network (ANN), often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase (17). The neural network described by Razumovskaya et al., uses the following six input parameters to identify peptides from spectral data: the SEQUEST Xcorr, peptide charge state, ΔC_n , SpRank (S_p), ion coverage and the length of the peptide to determine true positives of a data set. The first aspect of the algorithm begins with the training of the neural net. The neural net was given a training data set that included 2,784 correct hits and 34,287 incorrect SEQUEST hits. 70% of this data set was used to actually train the network and the other 30% was used to test it. The scoring output from the neural network ranges from 0.0 to 1.0, where the output represents an estimate of assignments correctness, which can then be used to generate a histogram. This improvement can be seen in figure 8. One advantage of this approach is that it provides improved resolution of assignments for peptides that SEQUEST scores. This can be particularly true for those hits that reside in what is considered a "gray area" of separation. Once the neural network has been trained, Razumovskaya describes a simple Bayesian formula that estimates the conditional probability, P, of true peptide identification. Once the probability of a peptide being correct has been determined, this information can be used to determine what peptides would identify a false protein by chance. This calculation can be observed below:

$$P_{score} = \frac{1 - P_i}{1 - \sum_{i=1}^M P_i}$$

Here, the probability of (ai) to be a true hit is defined as (P_i), and M_i is the number of proteins that were found to contain peptide ai . Thus, P_{score} is equal to the probability that protein A is not in the analyte. In simple terms, this formula is essentially determining the probability of a protein being identified by chance with respect to the peptides originally identified by the neural network. Furthermore, Razumovskaya have shown that the smaller the P_{score} , the more likely that a proteins identification was not by chance.

This method has provided significant improvement with respect to increasing the resolution between a true positive and a false positive. Furthermore the neural network makes the use of additional parameter when identifying the likelihood that a peptide fragment was correctly identified by SEQUEST. But this method does have its drawbacks. First it does not address the possibility that the first score given in a SEQUEST search may not be the correct protein, and since the neural network is trained on only top score hits, there is the possibility that will not be able to distinguish the top score of an incorrect peptide, and a second score being the correct identification.

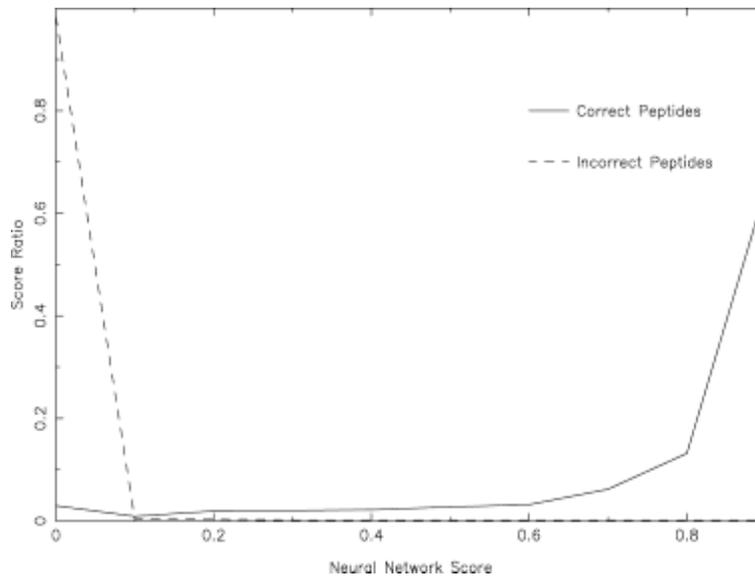


Figure 8. The neural network has the ability to finely separate the correct peptide hits from incorrect peptide hits from a SEQUEST search.

(Sun et al., *J Mol & Cell Proteomics* 2004, 3.12, 1194-1199)

Another program that allows one to validate SEQUEST results in an automated fashion is described by Sun et al. Advanced mass spectrum screener or AMASS, takes advantage of the fact that highly abundant fragment ions are more likely to be signals and that the MS/MS spectrum of an optimally fragmented peptide should theoretically contain continuous fragments. Based on these two observations, two functions were incorporated into the algorithm. These are Match Percentage and Continuity. Match Percentage is a function that serves to reduce the background noise from spectra and is calculated as follows: MatchPct = [number of matched daughter ions with relative abundance higher than RACutoff/number of total daughter ions with relative abundance higher than RACutoff] x 100%. Here the RACutoff (Relative Abundance Cutoff) is a number used to serve as a relative abundance cutoff point in MS/MS spectra and is a value that ranges from 0 to 100. Generally, the higher the value, the less ion fragments were selected. Consequently, a higher MatchPct value is representative of more fragment ions above a certain RACutoff are matched and that usually the higher the MatchPct value, the better the quality of the identified peptide fragments.

The second function used to identify peptides from spectral data is the Continuity function. This is determined by the equation displayed below:

$$\text{Cont} = \left[\sum_{i=1}^I (b(i) + y(i) + f(i)) \right] / [(I-1)^2 + I] * 100,$$

Here, $f(i) = 1$ if the i^{th} b or y series ion is matched, 0 otherwise; $b(i) = n^2$ if the $(i + 1)^{\text{th}}$ b series is not matched and $n =$ the number of continuously matched b ions immediately before the i^{th} (including the i^{th} ion), 0 otherwise; $y(i) = n^2$ if the $(i + 1)^{\text{th}}$ y series ion is not matched and $n =$ the number of continuously matched y ions immediately before the i^{th} (including the i^{th} ion), 0 otherwise; and $I =$ the amino acid number of the peptides.

The *Cont* algorithm will function specifically by adding the number of continuously matched b and y series ions to the second degree in addition to the total number of matched ions. This value is then normalized by dividing the maximum possible value of the addition and multiplying by 100. Generally a higher *Cont* value suggests more continuous matching fragments from a SEQUEST search.

What is particularly effective about this algorithm is that it can be combined with the RScore algorithm developed by Wang et al, which has been shown to increase the number of true positive peptides identified from spectra (18). The RScore criterion is a measurement of randomness and can be used to evaluate tandem mass spectra data in conjunction with SEQUEST. This is generally accomplished by comparing hits from a SEQUEST cross-correlation search with other hits from the same SEQUEST search. For peptide identification numbered from 1 to n for the same spectrum, RScore is calculated as followed:

$$RScore(i) = \max_{\substack{1 \leq j \leq n \\ j \neq i}} [XCorr(j)/XCorr(i) + k * IntenPct(j)/IntenPct(i)]$$

Here, *IntenPct(i)* is the percentage of intensity that was matched in identification *i*, and *k* is a parameter that depends on the instrument behavior. Since RScore takes into account the quality of the spectra taking into account the relative intensity of fragment ions from the spectra. Thus, the SEQUEST threshold for determining true positives with respect to peptide identification can be lowered to accept more true positives. Using spectra obtained from a mixture of 18 proteins, a SEQUEST search was performed against a database that contained 88,374 proteins which included the Swiss-Prot human protein database and the original 18 proteins from the experimental mixture. In order to assess the accuracy of AMASS, four searches were performed using the different filter combinations described above (figure 9). These were the Xcorr, the MatchPct, the Continuity function (*Cont*) and AMASS. This search using each of the filtering parameters showed that alone, AMASS could lower the false positive rate without significantly affecting the number of true positives generated from a search, providing a reliable means to increase the identifications by reducing the false positive detection from a search.

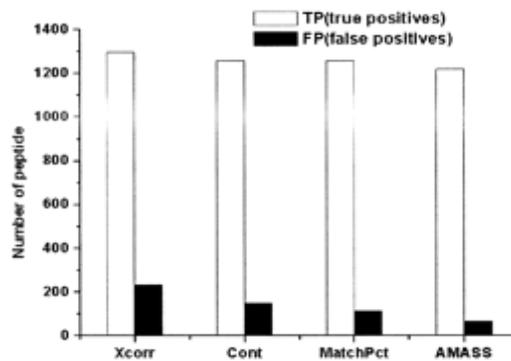


Figure 9. True positive (TP) and false positive (FP) hits from the different filter used to determine the effectiveness of the AMASS filter. As observed, AMASS can reduce the FP identification rate without affecting the identification of true positives.

Conclusion

As shotgun proteomics becomes more defined, the efficiency of methods created for SEQUEST, makes the technology much more applicable towards modern day high throughput proteomics studies. As described previously, a better search can be performed by using increasingly large databases as well as the number of database searched for experimental spectrum comparisons and construction of theoretical spectrum for peptide and protein identification. As database sizes increase and proteomics studies become more ambitious, the need to increase the speed of the search becomes increasingly important. One such method has been developed by Eng et al., which uses an alternative method to the cross correlation score that does not require the Fast Fourier Transforms that is used to compute the original SEQUEST correlation function. In a comparison of this new method and the old saw an increase ~ 75 times greater than the older, original method. Because of the way that the Xcorr value is computed with this faster algorithm, it is now possible to determine an Xcorr value for every peptide from the sequence database that is being searched. Thus new, faster, more sensitive cross correlations can be performed relative to older methods that were computationally expensive. This should enable more sensitive and sophisticated search methods to develop that in the end will be able to take into account multiple variables such as non-specific proteolysis, incorrectly determined charge state as well as chemical/post-translational modifications. Eventually the creation of new algorithms that are ever increasingly more sophisticated will help to refine the identification of true positives, the field of proteomics, and open the door for amazing new discoveries.

References:

1. Celis JE, Gromov P, *Curr Opin Biotechnol.*, 1999 Feb;10(1):16-21
2. Labugger R, Simpson JA, Quick M, Brown HA, Collier CE, Neverova I, Van Eyk JE, *Clin Chem.*, 2003. 873-9
3. Hernandez P, Muller M, Appel R, *Mass Spec Reviews*, 2006, 25, 235-254
4. Marcotte, EM, *Nat Biotech.*, 2007, vol. 25, pp 755-757
5. Hoffmann, E, *J Mass Spectrom.*, 1998, vol. 31:2, pp 129-137
6. Pappin DJ, Hojrup P, Bleasby AJ., *Curr. Biol.* 1993, 3:6, pp. 327-32
7. Yates JR, Speicher S, Griffin PR, Hunkapiller T., *Anal. Biochem.* 1993 214 (2): 397-408.
8. Clauser KR, Baker P, Burlingame AL., *Anal. Chem.* 1999 **71** (14): 2871-82
9. Yates III, Eng et al., *Anal. Chem.*, 1995, vol. 67, pp 1426-36
10. Keller A, Nesvizhskii A, Kolker E, Aebersold R, *Anal. Chem.* 2002, 74 (20), 5383-5392
11. Nevizhskii A, Aebersold R, *Drug Discov Today*, 2004 9(4), 173-81
12. Patterson S, *Nat. Biotechnol.*, 2003, 21, 221-222
13. Tabb D, Smith L, Brechi L, Wysocki V, Lin D, Yates J., 2003, *Anal. Chem.* 75, 1155-1163
14. Kapp E, Schutz F, Reid G, *Anal. Chem.*, 2003, 75, 6251-6254
15. Johnson R, Davis M, Taylor A, Patterson S, *Methods*, 2005, vol. 35, 223-236
16. Nesvizhskii A, *Methods Mol Biol.* 2007, 367, pp 87-119
17. Bhadeshia H, *ISIJ International* 1999, 39: 966-979
18. Li F, Sun W, Gao Y, Wang J, *Rapid Commun. Mass Spectrom* 2004, 18: 1655-1659