Qscore: An Algorithm for Evaluating SEQUEST Database Search Results

Roger E. Moore, Mary K. Young, and Terry D. Lee
Division of Immunology, Beckman Research Institute of the City of Hope, Duarte, California, USA

A scoring procedure is described for measuring the quality of the results for protein identifications obtained from spectral matching of MS/MS data using the Sequest database search program. The scoring system is essentially probabilistic and operates by estimating the probability that a protein identification has come about by chance. The probability is based on the number of identified peptides from the protein, the total number of identified peptides, and the fraction of distinct tryptic peptides from the database that are present in the identified protein. The score is not strictly a probability, as it also incorporates information about the quality of the individual peptide matches. The result of using Qscore on a large test set of data was similar to that achieved using approaches that validate individual spectral matches, with only a narrow overlap in scores between identified proteins and false positive matches. In direct comparison with a published method of evaluating Sequest results, Qscore was able to identify an equivalent number of proteins without any identifiable false positive assignments. Qscore greatly reduces the number of Sequest protein identifications that have to be validated manually. (J Am Soc Mass Spectrom 2002, 13, 378–386) © 2002 American Society for Mass Spectrometry

Proteolytic digestion followed by mass spectrometry and database searching has become the premier approach to sensitive identification of proteins. High throughput approaches to protein identification depend on minimizing human time investment in this analysis. A variety of techniques, including robotic gel band excision and digestion, automated matrix-assisted laser desorption/ionization (MALDI) spotting, autosampled nano-liquid chromatography tandem mass spectrometry (LC/MS/MS) analysis, and automated database searching, have been developed to further this aim. To make automated database searching possible, it is necessary to use a search program that can process spectra without need for human interpretation. In theory, it is also desirable to have an automated scheme to determine the significance and reliability of the database search results.

One of the more popular routines for database matching of peptide MS/MS spectra is Sequest [1]. Sequest can be used to analyze uninterpreted MS/MS spectra and provides a score for each match. However, exact standards for individual peptide matches to be considered significant, or for a group of peptide matches to indicate successful protein identification are still a matter of question. One criterion [2, 3] is what might be termed a golden match standard, in which a single, human validated peptide match is considered to be a conclusive identification of its precursor protein.

While the golden match standard appears valid, it suffers from some difficulties associated with the incompleteness of protein databases. In our experience, it is sometimes the case that the second best match for a peptide would meet numerical and subjective criteria as a golden match if the top scoring peptide were removed from the database. This suggests that the same may be true of some peptides that generate top scores; they are not actually the correct peptide, and generate the top match only because the correct peptide is not in the database. More generally, the golden match criterion is caught in a double bind. In order for a peptide to identify a protein uniquely from the database, it must be sufficiently long that it is unlikely to exist in several unrelated proteins. For a peptide of such length, though, there must be many isobaric peptides that are not present in the database, so the status of the identified peptide as the unique best match cannot be confirmed.

If single matches, no matter how good, are insufficient to ensure a correct protein identification, some approach based on multiplicity of matches must be used. Some authors have proposed standards for identification based on multiple peptide matches [4], but like the golden match standard, these are basically ad hoc criteria. This paper attempts to create a reasonable, statistically based algorithm for determining goodness of protein matches.

Published online February 21, 2002

Address reprint requests to Dr. T. D. Lee, Division of Immunology, Beckman Research Institute of the City of Hope, 1450 E. Duarte Road, Duarte, CA 91010. E-mail: tdlee@coh.org

1044-0305/02/$20.00
PII S1044-0305(02)00352-5

Received July 19, 2001
Revised January 16, 2002
Accepted January 16, 2002

Methods

Computer Programs

The Sequest program was version 26 or 27 (TurboSequest) obtained from ThermoFinnigan (San Jose, CA). Subsidiary programs were written in Perl version 5.005_03, available for download from the Comprehensive Perl Archive Network (http://www.cpan.org). The subsidiary programs are available from the authors' web site (http://www.cityofhope.org/microseq/download.html).

Databases

Protein databases were downloaded from the National Center for Biotechnology Information (NCBI) web site (ftp://ww.ncbi.nlm.nih.gov/blast/db). The NCBI nonredundant (nr) database (as of 3/24/2000 and 4/11/2001) was used for searches against a complete database. The *S. cerevesae* (yeast) database (6298 proteins) was used as an example of a smaller, species specific database. For some experiments, the sequence databases were reversed using the program db_reverse.pl. A database was reversed by reversing the sequence of amino acids in each protein. This generated a database with the same distribution of amino acids, sequence lengths, and sequence homologies as the original database but none of the actual protein sequences. These reversed databases were labeled with the original name spelled backward, i.e., rn and tsaey.

Mass Spectra

Two sets of experimental spectra were used to test the searching approaches used. All spectra were derived from LC/MS/MS experiments using a ThermoFinnigan (San Jose, CA) LCQ Classic ion trap mass spectrometer and a custom built nano-flow HPLC and nano-ESI interface [5].

The first data set consisted of spectra from 8 LC/MS/MS analyses of *S. cerevesae* proteins. The spectra were subjected to a series of screening steps to ensure that they were capable of generating positive Sequest search results. All spectra were screened using the winnow.pl program [6], then searched against the NCBI nr database. All spectra corresponding to yeast proteins that generated two or more matches were used. The resulting data set consisted of 511 spectra.

The second data set consisted of spectra from nine LC/MS/MS analyses of human proteins. No attempt was made to ensure that the spectra were of high quality; the only criterion for inclusion was that the spectra should produce matches when analyzed using Sequest. This resulted in a collection of 4268 spectra. In a small number of cases the charge of the precursor ion could not be determined, so the data was searched twice, once assuming a doubly charged precursor and once assuming a triply charged precursor. This resulted in a total of 4316 individual searches.

Additional sets of spectra used to test the Qscore algorithm were the product of routine analyses carried out by the Mass Spectrometry Core Facility at the City of Hope and were derived from a variety of organisms.

Sequest Searching

Most Sequest searches were carried out using moderately restrictive parameters. The only modification considered was a quantitative carbamidomethylation of the peptides, which had been reduced and alkylated with iodoacetamide, and tryptic cleavage was assumed. Parent ion and fragment ion tolerances were set to 2.5 Da. For direct comparison with published approaches, less restrictive parameters were used. In addition to cysteine carbamidomethylation, non-quantitative methionine oxidation was considered, and no enzymatic constraints were used.

In some cases, Sequest search results were checked by an analyst. Criteria for manual validation were similar to those described by others [2]. Matched sequences were validated if all major peaks in the spectrum were explained by the candidate sequence and the spectrum contained enough peaks to confirm most of the peptide's sequence. Sequest Sp and Xcorr scores were not used as a criterion except where specifically mentioned.

Predicted Spectra

Predicted spectra contained the *b* and *y* ion series as well as ions representing water loss from the *b* and *y* ions. Multiply charged ions were predicted with a maximum charge equal to the lower of the precursor ion charge or 1 plus the number of basic residues (lysine, arginine, and histidine). The *b*- and *y*- type ions were given an intensity of 1 divided by their charge and water loss ions were given 1/4 the intensity of their parent *b* or *y* ion. If two ions had the exact same mass to charge ratio, their intensities were added.

Results and Discussion

False Positive Matches

The crucial factor to consider in deciding if a match is correct is whether there is a reasonable chance that such a match could have come about by chance. Sequest will always return a best match peptide as long as at least one peptide from the database falls within the peptide mass tolerance. When using a reasonable peptide mass tolerance, this condition is met even for databases smaller than the smallest non-viral proteomes. It is therefore reasonable to adopt as a null hypothesis that all matches are essentially random and accept as identified only those proteins that generate more matches than would be expected by this null hypothesis.

Table 1. Results of searching different data sets against sequence reversed databases. Data shown are the actual and predicted number of matches generating at least 2 or 3 unique peptide sequences, and the largest number of unique peptide sequences for any match

Peptides required for a protein match		Number of protein matches data set 1 (511 searches)		Number of protein matches data set 2 (4316 searches)	
		Tsaey database[b]	Rn database[c]	Tsaey database[b]	Rn database[c]
2 or more[a]	Actual	27	2	1034	118
	Predicted	20.7	0.27	1479	19.2
3 or more	Actual	1	0	402	4
	Predicted	0.56	9.4×10^{-5}	338	5.7×10^{-2}
Largest	Actual	3	2	10	3
	Predicted	3	1	6	2

[a] The predicted values for 2 or more peptides/protein are actually overestimates, as each group of 3 matches is treated as 3 independent groups of 2 matches.
[b] The tsaey database contains 6298 proteins.
[c] The rn database contains 483730 proteins.

The expected number of matches can be established by a number of approaches. The simplest approach is to derive a prediction analytically for the generic case. Given: N = number of individual searches; M = number of matches against a specific protein; P = number of proteins in the database.

The chance that a group of M searches will all match the same protein is then simply:

$$P_{\text{match}}(M, P) = P^{(1-M)} \tag{1}$$

And the chance that they will not all match is:

$$P_{\text{no match}}(M, P) = 1 - P^{(1-M)} \tag{2}$$

The number of groups of M searches chosen from N searches is:

$$N_{\text{groups}}(M, N) = N!(N - M)!M! \tag{3}$$

The chance that no group of M matches will all match to the same protein is then:

$$P_{\text{no match}}(M, N, P) = (1 - P^{(1-M)})^{N!/(N-M)!M!} \tag{4}$$

Making the chance that there is such a match:

$$P_{\text{match}}(M, N, P) = 1 - (1 - P^{(1-M)})^{N!/(N-M)!M!} \tag{5}$$

The expected number of matches can also be estimated as the number of groups that can generate a match times the chance that each will generate a match:

$$N_{\text{match}}(M, N, P) = P^{(1-M)}N!(N - M)!M! \tag{6}$$

It is important to note that for $P_{\text{match}}(M, N, P) \ll 1$ (i.e., when the chance of a false positive is low), $P_{\text{match}}(M, N, P) \approx N_{\text{match}}(M, N, P)$. $N_{\text{match}}(M, N, P)$ will tend to overestimate the number of matches if $N_{\text{match}}(M + 1, N, P) > 1$, as each match to M + 1 spectra is treated as M + 1 matches of M spectra. The results of these formulas can be experimentally tested by searching a group of real spectra against a deliberately falsified database. This was carried out for the two data sets described against both the rn and tsaey (sequence reversed) databases.

Effective Database Size

The results of database searching (Table 1) show the trends expected from the formulas. The number of matches tends to go up as the number of spectra searched increases and down as the number of proteins in the database increases. The actual numbers of proteins matched were not exactly as expected. The search of the tsaey database resulted in a few more matches than predicted, while that for the rn database gave many more than predicted. In effect, the tsaey database behaves as though it is somewhat smaller than its actual size and the rn database behaves as though it is much smaller than its actual size. Several factors appear to account for the discrepancy.

A single search can result in matches to more than one protein. Short peptides can exist in several unrelated proteins simply because there are a limited number of possible sequences, while longer peptides may exist in several homologous proteins. Furthermore, some peptides while of different sequence are not distinguishable by mass spectrometry. Frequently this occurs because they contain isobaric amino acid substitutions, such as leucine for isoleucine or glutamine for lysine. Occasionally peptides with greater sequence differences will not be distinguished because fragments that could distinguish them are not observed. The presence of multiple identical or indistinguishable peptides in the database reduces its effective size.

Another critical factor is that peptides are not evenly distributed among all proteins. Large proteins may contain hundreds of peptides that may be matched in a

Table 2. Characteristic statistics are given for the databases used in the study, the yeast genomic database, the NCBI non-redundant (nr) database at the listed dates, and two sequence reversed databases, tsaey (reversed yeast) and rn (reversed nr as of 3/24/00)

				NCBI nr	
	Tsaey	Yeast	Rn (3/24/00)	3/24/00	4/11/01
Proteins	6298	6298	483730	483730	666948
AAs	2974038	2974038	151518084	151518084	210073119
Peptides[a]	1142678	1138430	51743728	51562434	70959541
Distinguishable[b]	1061951	1058350	30809348	30685370	40600343
% Unique	92.9	93.0	59.5	59.5	57.2

[a]Peptides counted were tryptic peptides with lengths between 5 and 30 amino acids.
[b]Peptides were considered distinguishable if they differed by more than I/L or K/Q substitutions.

search, while small proteins may contain fewer than ten. Proteins with more peptides are naturally more likely to generate multiple matches than those with fewer peptides.

Distinguishable Peptides

A simple and reasonable approach to correct for the discrepancies in database size is to perform a theoretical digestion of the entire database and count the number of distinguishable peptides generated. A program, db_stats.pl, was written to do this. It is important to note that the number of distinguishable peptides is necessarily somewhat inexact and dependent on experimental conditions. For the experiments described here, the pairs leucine/isoleucine and glutamine/lysine were considered indistinguishable. This is an approximation. Even with low mass-accuracy ion trap spectra, Sequest will occasionally generate slightly different scores for glutamine versus lysine containing peptides. For the yeast databases used in this study, 93% of the tryptic sequences are unique. For the NCBI databases however, more than 40% of the tryptic peptide sequences are redundant (Table 2).

Once available, the number of distinguishable peptides in the database (d) can be compared to the number of distinguishable peptides generated by a theoretical digestion of a protein from the database (p) to determine the probability of a chance match:

$$P_{\mathrm{match}}(M, p, d) = (d/p)^{(1-M)} \tag{7}$$

which gives the predicted number of matches:

$$N_{\mathrm{match}}(M, N, p, d) = (d/p)^{(1-M)}*N!/(N-M)!M! \tag{8}$$

If a different enzyme from that used to generate the database profile is used, the profile data can still be used to approximate the match probability. In this case the data used is the number of distinct peptides in the database (d), the total number of peptides in the database (t), the number of amino acids in the protein (l), and the number of amino acids in the database (a):

$$\begin{aligned} P_{\mathrm{match}}(M, d, t, l, a) &= [(a/l)*(d/t)]^{(1-M)} \\ &= (a*d/l*t)^{(1-M)} \end{aligned} \tag{7a}$$

$$N_{\mathrm{match}}(M, N, d, t, l, a) = (a*d/l*t)^{(1-M)}N!/(N-M)!M! \tag{8a}$$

In effect, the chance of finding a peptide from a protein is treated as the fraction of amino acids in the database that are present in that protein, with the size of the database corrected for the presence of indistinguishable peptides. In a few cases (e.g., a database consisting of just one protein that includes repeated sequences) this may result $P_{\mathrm{match}} > 1$, but this is unlikely to be true of any database used for general searching. This is not an issue when using the actual number of distinct peptides in the protein and database, as every distinct peptide in a given protein must also be present in the database.

A factor that counterbalances peptide duplication in the database is that the same peptide may generate multiple matches in an analysis. While data dependent approaches [7] can minimize re-analysis of the same peptide, they cannot completely prevent it. Combining multiple matches is comparatively straightforward; each distinct sequence identified counts as a single match. This is important because it was observed that Sequest returns the same incorrect peptide for related spectra when the correct peptide is not present in the database. Spectra derived from the trypsin autolytic fragment LGEHNIDVLEGNEQFINAAK, for instance, gave the sequence SLSHVAKGLVKVNGGTILCKM when searched against the rn database and NFFYAFNKSTIIHLQLVR when searched against tsaey. To account for peptide duplication, the total number of matches should be adjusted to equal the number of distinguishable peptides identified, rather than the number of searches carried out.

Sequest Cross-Correlation Scores

Many of the matches of MS/MS spectra to the rn database (which has no correct protein sequences) gave Sequest Xcorr scores in the range (≥1.9 for +1 ions, ≥2.2 for +2 ions, and ≥3.75 for +3 ions) classified as significant by others [2, 3] (Table 3). Manual examina-

Table 3. Observed distribution of Sequest Xcorr scores for searches against the rn (sequence reversed) database. With the exception of very small peptides, none of these identifications should be correct, as the database used does not contain any real protein sequences

	Data set 1 (511 searches)[a]			Data set 2 (4316 searches)[b]		
Charge state	1+	2+	>2+	1+	2+	>2+
Count	189	308	14	2411	1438	467
Mean Xcorr	1.57	2.24	2.66	1.21	1.74	1.93
Median Xcorr	1.57	2.25	2.53	1.27	1.69	1.93
Maximum Xcorr	2.70	4.04	3.48	2.72	3.48	3.78

[a] The spectra in data set 1 were selected for high quality.
[b] The spectra in data set 2 were not.

tion of the matches found for the smaller data set resulted in 30 that fulfilled all of our criteria for a valid match. Thus, the proportion of false positives that escape detection by manual validation is estimated to be 5%.

Accounting for Match Quality

The only factor not yet considered is the quality of individual peptide matches. While a single peptide match may not be sufficient to conclusively identify a protein, it is quite clear that not all matches are created equal. It is plausible to assume that a strong match is less likely to result from a false positive than a weak match is. By rearranging eq 8, it is possible to derive an expression where some information about the quality of match of each peptide can logically be included:

$$N_{\mathrm{match}}(M, N, p, d) = (d/p)*_{i=1}\Pi^{M}(N + 1 - i)/i*(d/p)$$

$$q(i) = \text{match quality information for the } i^{\mathrm{th}} \text{ peptide} \qquad (9)$$

$$\mathrm{Score}(M, N, p, d) = (d/p)*_{i=1}\Pi^{M}(N + 1 - i)/[i*q(i)*(d/p)] \qquad (10)$$

In practice, it may happen that the quality of a match is so low that the overall score for the protein is worse when it is included. Rearranging the formula as a product allows the score to be calculated iteratively, making it simple to ignore matches that worsen the final score. It is also convenient to convert the score into its negative logarithm, giving a Quality based score (Qscore):

$$\mathrm{Qscore}(M, N, p, d) = -\log(d/p) + {}_{i=1}\Sigma^{M} \log[(i*q(i)*(d/p))/(N + 1 - i)] \qquad (11)$$

It is important to note that the inclusion of data about the quality of the peptide matches may prevent the Qscore from being a true probability. Unless the scores accurately reflect the probability that the matches are true positives, the overall Qscore will be based on probability theory but not reflect a true probability. Unfortunately, it is not possible to recover a probability that a match is a true positive from the data available in a Sequest search, so this weakness is unavoidable in any approach that uses Sequest data. Treating the Qscore as a true probability can still be useful in understanding its meaning, even though such an interpretation needs to be viewed cautiously. In such an interpretation Qscore is -log(expected matches of this quality). A Qscore of 0 thus corresponds to the quality of match that could be expected to exist based on chance alone.

Our approach to the quality issue is to measure the agreement between the experimental spectrum and a spectrum predicted based on the matched sequence. This can be done using a spectral product algorithm in which the experimental spectrum (A) and predicted spectrum (B) are compared to generate a product spectrum (C). Each peak in spectrum A is translated into a peak in C that has intensity equal to the intensity of the peak in A times the summed intensity of all peaks in B within a specified mass tolerance. The product score, s(A, B) is the summed intensity of the peaks in C. To ensure that the agreement between the spectra is a value between 0 and 1, each spectrum is first normalized so that its product with itself is 1. The degree of agreement is then:

$$a(A, B) = s(A, B)/(s(A, A)*s(B, B))^{1/2} \qquad (12)$$

The degree of agreement between the actual and predicted spectra can then be treated as the probability that the match is a true positive. The qualitative scoring information is then the inverse of the false positive chance:

$$q = 1/[1 - a(\text{actual spectrum, predicted spectrum})] \qquad (13)$$

This suggests that as peptide match quality approaches perfection, the degree of confidence in the protein match also approaches perfection, which is essentially a re-statement of the golden match principle. In practice, there is a limit to the degree of confidence in any

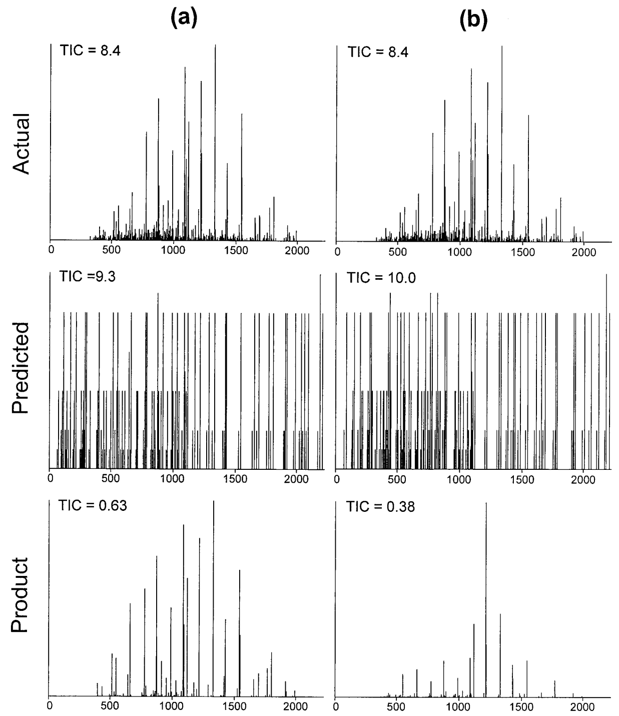

Figure 1. An illustration of the peptide match scoring procedure applied to the MS/MS spectrum for the doubly charge form of the trypsin autolytic peptide LGEHNIDVLEGNEQFINAAK matched to (**a**) the correct sequence and (**b**) the sequence SLSHVAKGLVKVNGGTILCKM from the rn (sequence reversed NCBI nr) database. Actual spectrum (upper panel) and predicted spectra (middle panel) are multiplied peak by peak to generate a product spectrum (lower panel), whose total ion count (TIC) is the match score. Actual and predicted spectra are first normalized so that the product with themselves has a TIC value of 1.

peptide match, which depends on the quality of the mass spectral data used. This uncertainty should be a natural consequence of the matching algorithm.

The process of comparing an actual spectrum with a predicted spectrum is illustrated using a spectrum derived from a trypsin autolytic peptide (Figure 1). When the actual spectrum (Figure 1, top panel a and b) is compared to the predicted spectrum for the correct sequence (Figure 1, middle panel a), the product spectrum is very similar in appearance (Figure 1, bottom panel a) to the actual spectrum. When compared to the predicted spectrum for the incorrect sequence matched in the rn database (Figure 1, middle panel b), the product ion spectrum (Figure 1, lower panel b) differs significantly from the actual spectrum. This obvious qualitative difference is reflected in the TIC scores of the two product ion spectra.

It is also necessary to account for peptides or sets of indistinguishable peptides that are matched more than once. Matching a peptide several times does increase the confidence that the match is correct. This can be accomplished by taking the product of the individual match scores to arrive at a score for the group of matches. In this case it is important to apply a maximum allowable score, so that multiple incorrect matches do not overly skew the results.

Applying the Qscore Algorithm

A program, qscore.pl, was written to apply the Qscore algorithm to real Sequest search results. The qualitative score information, q, for each distinguishable peptide was capped at 20 (equivalent to a false positive chance of 5%). This chance is in agreement with the observed false positive rate for manual validation of high quality spectra previously discussed. Proteins that generated only one distinguishable peptide sequence match were ignored. The qscore.pl program was then applied to 34 LC/MS/MS analyses of trypsin digested gel bands that had previously been subjected to manual validation. The criterion for a positive match was two manually validated peptides. The analyses varied considerably, with a range of 20 to 1784 spectra searched. Proteins matched using the Qscore algorithm were grouped into four categories:

1. Matches: Those proteins that were identified by manual validation
2. Contaminants: Common contaminant proteins such as keratin, trypsin, etc., or known contaminants from the specific preparation that were ignored in the manual analysis
3. Misses: Proteins that are probably present in the sample (from the correct organism and of correct approximate molecular weight) but not identified by manual validation
4. False positives: All other matches

A total of 58 matches, 88 contaminants (not all unique), 8 misses, and 42 false positives were found (Figure 2). The algorithm also eliminated one apparent match that the analyst classified as a probable artifact. Qscores for the matches ranged from a minimum of 0.14 to a maximum of 467.01, and false positives ranged from −1.1 to 2.12. The overlap in ranges was minimal, with only 10% of matches having scores lower than the highest scoring false positive and 17% of false positives having scores higher than the lowest scoring match.

The sharp differentiation between true and false positives allows for significant automation of analyses. As we now use the program, protein identifications with Qscores higher than 2.5 are automatically accepted, those with Qscores lower than 0 are automatically rejected. Those with Qscores falling in between are manually validated. For the data set shown, there was an automatic acceptance of four of the eight proteins previously missed by manual validation.

Comparison with Published Approach

A subset of the data used for the evaluation of the Qscore algorithm was searched again using criteria similar to those published by others [3]. Briefly, the

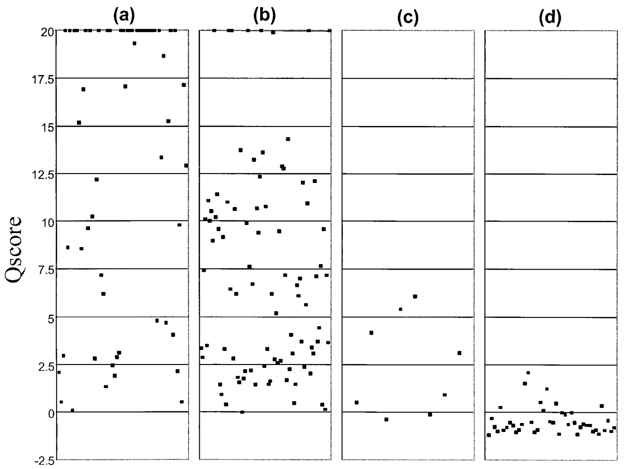

Figure 2. Distribution of Qscores for a set of Sequest matched peptide MS/MS spectra. Protein matches were divided into four groups after manual validation: (**a**) Proteins that were identified, (**b**) known or anticipated contaminant proteins, (**c**) proteins that might have been identified (from the right species and of the correct approximate molecular weight) but weren't, and (**d**) all other proteins. Scores higher than 20 were capped at that value. Points are scattered on the x-axis to make visualization of individual data points easier.

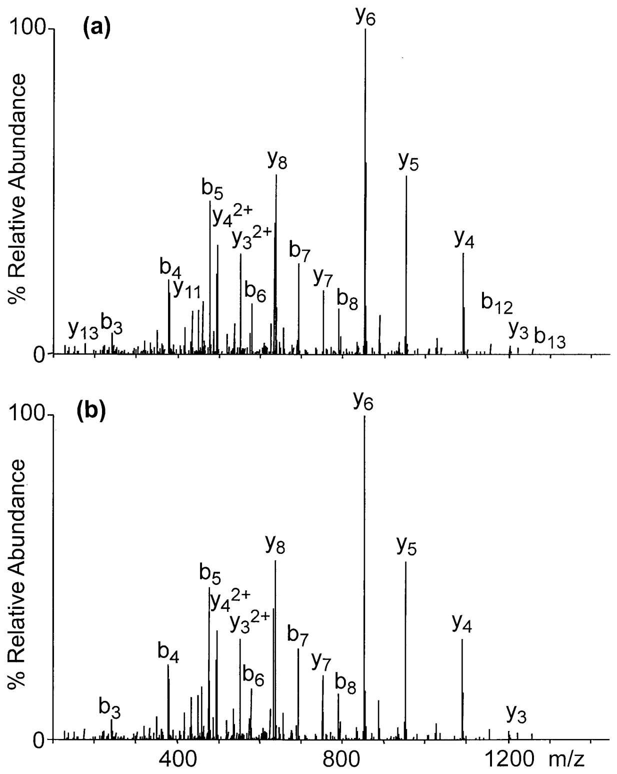

Figure 3. Example of a spectrum for which the second best match would meet acceptance criteria if the best match peptide were not in the database. The same spectrum is shown with peak assignments for (**a**) the correct match, NLLHVTDTGVGM*TR (M* = oxidized methionine, Xcorr = 3.98, deltaCn = 0.157) from human endoplasmin and (**b**) the second best match, KVLHVTDT-NKFDN (Xcorr = 3.37, deltaCn = 0.196) from Vaccinia virus protein Hind III. The identified peptides exhibit substantial sequence identity even though their parent proteins are completely unrelated.

spectra were searched without enzymatic constraint and considering possible methionine oxidation, and matches were only considered if their Xcorr and DeltaCn scores exceeded specified thresholds. The Xcorr thresholds were based on published values [3] and are summarized in Table 4. Spectra were also required to have a DeltaCn score of 0.1 or higher. An exception to the DeltaCn threshold was made if the first and second scoring peptides were very similar, defined as having only leucine/isoleucine, aparagine/aspartic acid, or lysine/glutamine/glutamic acid substitutions. Spectra that passed the Xcorr and DeltaCn thresholds were then manually validated. Figure 3 shows a specific example of a spectrum accepted under these criteria in which an erroneous second best match would have been accepted had the best match peptide been absent from the database.

The same results were then analyzed using Qscore and the results compared (Table 5, row A. The two methods agreed on 30 of the identified proteins. For five of those, the threshold method identification was based on a single peptide. Other peptides matched to those peptides did not have sufficiently high Sequest scores. The threshold approach identified 14 proteins not found using Qscore, each based on a single peptide. Fully half of the additional matches were clearly false positives as they came from organisms that could not plausibly have been present in the samples. The Qscore approach identified two proteins not found by the threshold method, generated no identifiable false positives, and required less manual validation. Ignoring single peptide matches when using the threshold approach would eliminate all of the false positive matches, but would result in the identification of seven fewer proteins than the Qscore approach. It is important to note that the false positive matches were only identified because the data was searched against a non-redundant database, so the species of origin could be used as an independent check on the reasonableness of the match. With a species specific database this would not be possible.

Table 4. Summary of Xcorr score and tryptic cleavage criteria used to automatically judge Sequest matches. Spectra were only considered if their Xcorr exceeded the specified value

	Precursor charge			
	+1	+2	+3	>+3
Non-tryptic	—[a]	3.0	—	—
Partially tryptic	—	2.2	3.75	—
Fully tryptic	1.9	2.2	3.75	—

[a]Categories marked with — were never considered.

Even better results were obtained with the Qscore approach using our standard search criteria that examine only tryptic cleavage and do not consider methionine oxidation (Table 5 row B). Qscore and the thresh-

Table 5. Comparison of results using Qscore to approach using Xcorr and DeltaCn thresholds

	Both[c]		Threshold only[d]		Qscore only[e]	
	Multiple	Single	Positive	False positive	Positive	False positive
A[a]	25	5*[f]	7*	7*	2	0
B[b]	25	6*	6*	7*	5	0

[a]The Qscore results include exactly the same Sequest search results as the threshold approach.
[b]The Qscore results obtained considering only tryptic cleavage and no methionine oxidation.
[c]Results were categorized by whether the identified protein was found using both approaches.
[d]Only the Threshold approach.
[e]Only the Qscore approach. Proteins found using both approaches are categorized by whether the Threshold approach found a single peptide or more than one peptide. Proteins found using only one approach are categorized as true or false positives depending on whether the protein came from a species that could plausibly have been present in the sample. Known contaminant proteins such as keratin were not considered.
[f]Matches marked with a * depend on a single peptide match when using the Threshold approach.

old method agreed on 31 of the identified proteins. The threshold method identified an additional six proteins and the above mentioned seven false positives. Qscore identified five additional proteins with no identifiable false positives. Ignoring single peptide matches when using the threshold method would eliminate all of the false positives, but would identify 11 fewer proteins than the Qscore approach.

Using a constrained search also allows full advantage to be taken of the speed enhancements available in later versions of Sequest (TurboSequest version 27). Assuming tryptic cleavage, only quantitative modifications, and using a predigested/indexed database reduced the time to search each spectrum about 100-fold compared to a search using no enzymatic constraint and considering possible oxidation of methionine. This eliminates the problem of slow search speed that is the most obvious objection to using such a large database. The use of tryptic constraints was reasonable for these samples because they were digested with trypsin that had been modified to minimize chymotryptic activity. Under these conditions, peptides containing non-tryptic cleavages or oxidized methionine were generally observed only in the presence of unoxidized tryptic peptides from the same protein, and therefore were not needed to identify the protein.

The additional protein matches obtained by Qscore using the trypsin constraint result from peptides that do not fare well in the preliminary scoring step of Sequest but yield high Xcorr scores. When the list of possible peptides is expanded by searching without enzymatic constraint, these peptides are no longer in the top 500 by preliminary score, so they are never considered by Xcorr.

Conclusions

Adopting a probabilistic scoring scheme such as Qscore has a number of advantages. It can minimize the need for either extensive manual match validation or ad hoc criteria for goodness-of-match. While these factors are not completely eliminated, as some criterion for the difference between a good and bad match will always be needed, Qscore provides an objective measure of goodness-of-match that includes all relevant data about the match. Qscore factors in the number and quality of peptide matches, the total number of searches that were carried out, the size of the matched protein, and the size and characteristics of the database searched. The Qscore algorithm does not use any information specific to Sequest, so it should be applicable to other matching approaches such as the sequence tag approach [8].

Most importantly, Qscore makes explicit a factor of database searching that manual validation and ad hoc cutoff approaches tend to ignore, that search results are not a binary yes or no answer. There are peptide matches that are either clearly right or clearly wrong, but there are also very many that are of intermediate quality. Using a score cutoff and/or analyst's judgment to force intermediate quality results into positive or negative categories actually interferes with the goal of maximizing the data extracted from the system. Multiple matches to a single protein, each of which falls just below the threshold for classification as a positive match, will be ignored even if they collectively indicate a strong positive match. Only by incorporating all of the data from the search, good and bad, is it possible to extract all of the available information.

Acknowledgments

This work was funded in part by grants from the National Institutes of Health (RR06217 and CA33572). The authors thank Nikita Dinh, Helen Ge, and Denise Keen of the City of Hope Mass Spectrometry Core Facility for their assistance in acquiring the data used in this study.

References

1. Eng, J. K.; McCormack, A. L.; Yates, J. R., III. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom.* **1994,** *5,* 976–989.
2. Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., 3rd. Direct Analysis of Protein Complexes Using Mass Spectrometry. *Nat. Biotechnol.* **1999,** *17,* 676–682.

3. Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd. Large-Scale Analysis of the Yeast Proteome by Multidimensional Protein Identification Technology. *Nat. Biotechnol.* **2001,** *19,* 242–247.
4. Haynes, P. A.; Fripp, N.; Aebersold, R. Identification of Gel-Separated Proteins by Liquid Chromatography-Electrospray Tandem Mass Spectrometry: Comparison of Methods and Their Limitations. *Electrophoresis* **1998,** *19,* 939–945.
5. Davis, M. T.; Lee, T. D. Rapid Protein Identification Using a Microscale Electrospray LC/MS System on an Ion Trap Mass Spectrometer. *J. Am. Soc. Mass Spectrom.* **1998,** *9,* 194–201.
6. Moore, R. E.; Young, M. K.; Lee, T. D. Method for Screening Peptide Fragment Ion Mass Spectra Prior to Database Searching. *J. Am. Soc. Mass Spectrom.* **2000,** *11,* 422–426.
7. Stahl, D. C.; Swiderek, K. M.; Davis, M. T.; Lee, T. D. Data-Controlled Automation of Liquid Chromatography Tandem Mass Spectrometry Analysis of Peptide Mixtures. *J. Am. Soc. Mass Spectrom.* **1996,** *7,* 532–540.
8. Mann, M.; Wilm, M. Error Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. *Anal. Chem.* **1994,** *66,* 4390–4399.