Properties of Average Score Distributions of SEQUEST

THE PROBABILITY RATIO METHOD

Salvador Martínez-Bartolomé‡§¶, Pedro Navarro‡§, Fernando Martín-Maroto‡||, Daniel López-Ferrer‡**, Antonio Ramos-Fernández‡¶, Margarita Villar‡ ‡‡, Josefa P. García-Ruiz‡*, and Jesús Vázquez‡‡‡

High throughput identification of peptides in databases from tandem mass spectrometry data is a key technique in modern proteomics. Common approaches to interpret large scale peptide identification results are based on the statistical analysis of average score distributions, which are constructed from the set of best scores produced by large collections of MS/MS spectra by using searching engines such as SEQUEST. Other approaches calculate individual peptide identification probabilities on the basis of theoretical models or from single-spectrum score distributions constructed by the set of scores produced by each MS/MS spectrum. In this work, we study the mathematical properties of average SEQUEST score distributions by introducing the concept of spectrum quality and expressing these average distributions as compositions of single-spectrum distributions. We predict and demonstrate in the practice that average score distributions are dominated by the quality distribution in the spectra collection, except in the low probability region, where it is possible to predict the dependence of average probability on database size. Our analysis leads to a novel indicator, the probability ratio, which takes optimally into account the statistical information provided by the first and second best scores. The probability ratio is a non-parametric and robust indicator that makes spectra classification according to parameters such as charge state unnecessary and allows a peptide identification performance, on the basis of false discovery rates, that is better than that obtained by other empirical statistical approaches. The probability ratio also compares favorably with statistical probability indicators obtained by the construction of single-spectrum SEQUEST score distributions. These results make the robustness, conceptual simplicity, and ease of automation of the probability ratio algorithm a very attractive alternative to determine peptide identification confidences and error rates in high throughput experiments. Molecular & Cellular Proteomics 7:1135–1145, 2008.

Modern proteomics is mainly based on the analysis of proteins by mass spectrometry. With the increasing use of multidimensional peptide separation coupled to tandem mass spectrometry as an alternative to gel-based approaches for protein expression analysis and high throughput protein identification, there is a growing interest in the development of scoring systems for automated, large scale peptide identification.

SEQUEST (1–3), one of the first and most popular scoring schemes, measures the degree of correlation between the experimentally observed and the theoretical MS/MS spectra of peptides present in protein databases and determines the peptide sequence yielding the best correlation score, or Xcorr. Another related SEQUEST parameter is the delta score, or ΔCn, which measures the difference between the best and second best scores (1–3). SEQUEST results must be further processed to determine whether the peptide identification is correct or not. Filtering of SEQUEST outputs has traditionally been done using the parameters Xcorr and ΔCn, by establishing empirically a set of criteria (4–8). Because it has been described that these criteria do not usually provide enough discriminative power to be used universally on any data set (9–11), alternative filtering criteria have been developed based on the distribution of SEQUEST scores and/or machine learning algorithms to achieve better separation between correct and incorrect assignments (9, 12–17). A common approach to evaluate the significance of peptide identification is to construct the distribution of SEQUEST scores after searching large collections of MS/MS spectra against protein sequence databases and subject these distributions to statistical analysis to estimate the probability of finding a true peptide sequence from the scores or the proportion of false positive assignments (9, 12, 13, 16, 17). For this end several statistical models have been used (9, 12, 16, 17). The proportion of false positive assignments among the tentative peptide identifications, also called false discovery rate (FDR)1 (17–19), can be estimated by using decoy databases constructed from the target databases (20). Other scores have also been proposed based on the correlation between observed and theo-

---

1 The abbreviations used are: FDR, false discovery rate; E, expectation.
Properties of Average SEQUEST Score Distributions

retical spectra (21) and on machine learning algorithms, taking into account the intensity pattern of peaks in MS/MS spectra (22).

Other scoring schemes are based on a different conceptual approach. Instead of considering large collections of MS/MS spectra, they consider individually each one of the spectra and try to determine the probability that the best sequence match is produced by a random event (the null hypothesis). Some algorithms are based on theoretical probability models (11, 23–26), whereas others consider the frequency distribution of scores obtained while a search is performed against peptide candidates within a protein sequence database and can hence be applied to determine statistical significance associated to empirical scoring schemes (27). A common measure of statistical significance is given by the so-called p value, which assesses the chance of validly rejecting the null hypothesis (11, 25, 28). A related value is the expectation or E value (11, 25, 27), a parameter that reflects the number of hits in a database expected to happen by chance with a given or better score.

Despite these efforts, the calculation of exact confidence of peptide identification is still considered an open problem. The lack of an appropriate confidence indicator is particularly problematic when large data sets are analyzed (29) because it is impossible to manually validate all the matches between tandem mass spectra and tentative peptide sequences. It has been acknowledged recently that universally accepted and widely available computational tools for validation of the published results are not yet available (29, 30). Besides a significant but undefined number of the proteins being reported as “identified” in proteomics studies are likely to be false positives (13).

In this work we will refer to the (best) score distributions constructed from large collection of different MS/MS spectra as average score distributions, whereas score distributions associated to a particular spectrum, such as these used to calculate E values, will be referred to here as single-spectrum score distributions. Despite their widespread use, the properties and behavior of average score distributions, such as dependence on database size, and particularly those related to SEQUEST scores, are still poorly understood. Besides it is a surprisingly common procedure to extrapolate statistical parameters and distributions obtained using particular data sets and databases to very different conditions where it is not clear whether error rates of peptide identification can be fully predicted using the same criteria. There is also some confusion with the use of the best score and the delta score. The delta score has been used either as a threshold discriminator so that only those spectra having a minimum delta score are considered as potential positives (4–8) or as an additional, independent score parameter that is evaluated together with the best score to determine the statistical matching significance (9, 16, 17). Although they are obviously interrelated, the theoretical relation existing between the best score and the delta score (or the second best score) has not been analyzed yet. Besides the relation between average score distributions and single-spectrum distributions, using the same scoring scheme, and the relative performance of peptide identification in large scale experiments have not been explored.

In this work we perform a mathematical analysis of SEQUEST average score distributions and analyze the behavior and properties of these distributions. For this end, we introduce the concept of spectrum quality and express the average score distributions as quality-weighted averages of single-spectrum distributions. Using this mathematical framework we make inferences about the dependence of these distributions on database size and use the concept of probability ratio as a statistical indicator to take into account the information provided by the best and second best scores. All the properties predicted by the mathematical analysis are tested in the practice by using large collections of MS/MS spectra. Finally the performance of peptide identification by SEQUEST using the probability ratio indicator is compared with that obtained using other approaches based on empirical modeling of average score distributions or the construction of single-spectrum score distributions. Our results facilitate understanding the properties and behavior of average score distributions as well as their practical limitations, provide a theoretical framework to relate them with single-spectrum distributions, and demonstrate the utility in the practice of the probability ratio concept as a very simple, robust, and accurate method to evaluate peptide identification confidence using SEQUEST in large scale experiments.

EXPERIMENTAL PROCEDURES

Two MS/MS spectra data sets were taken from a previous work (17); they were obtained by analyzing trypsin-digested whole cell extracts by ion exchange chromatography followed by reverse phase chromatography on line with ion trap mass spectrometry using either an LCQ Deca XP or an LTQ machine (Thermo Fisher) as described previously (17). One of the proteomes, analyzed using the Deca XP machine, was composed of nuclear proteins from Jurkat cells (17), and its analysis produced a collection of more than 40,000 MS/MS spectra. The other proteome, analyzed using the LTQ machine, was composed of total cell extracts from mesenchymal stem cells prepared from human bone marrow samples as described previously (31) and produced a data set containing more than 150,000 MS/MS spectra. In addition, a third collection, containing 13,000 MS/MS spectra, obtained from the analysis of an Escherichia coli proteome by using a hybrid LTQ-Orbitrap machine (Thermo Fisher) and kindly provided by Michaela Sciigelova, was used to test the performance of the method in conditions of high accuracy mass determination of precursor ions.

Database searches using SEQUEST were performed as described previously (17). Single-spectrum probability score distributions were constructed by performing extensive searches against large collections of random trypptic peptide sequences generated by Monte Carlo, taking into account the natural amino acid frequency in the nr.fasta database, or taken from reversed databases. Average score distributions for the first, second, . . . jth best SEQUEST scores were constructed from the whole collection of spectra by considering these scores separately and independently from the spectra from which
they were derived; this was accomplished by ranking each one of the
sets in decreasing order of score and by expressing the average prob-
ability associated to the score as its relative rank position (rank/$E$ where
$E$ is the total number of spectra). Scores obtained from the same
spectra assuming different charge states were analyzed as if they
were produced from different spectra. Data were produced by reject-
ing for fragmentation precursor ions with charge lower than 2 or
higher than 3; hence only doubly and triply charged peptides were
analyzed in this work. Probability ratios were calculated as illustrated
in Fig. 1 and explained in the legend using a program written in C#
(freely available upon request). This program uses as input the result
files (in srf format) obtained from the separate search of raw files from
one or more proteome fractions against the target and decoy data-
bases; decoy databases were constructed by reversing the amino
acid sequence of each one of the proteins. The probability ratio
program analyzes as a whole all the results obtained from a large
scale experiment, allowing the simultaneous analysis of hundreds of
srf files. The output is made in PepXML format. Further details about
the program and raw file collections to test the method can be
obtained upon request. Probability ratios were corrected by charge
and length by a simple numerical correction of SEQUEST Xcorr
scores. Charge is adjusted by considering that the amount of frag-
ment information in an MS/MS spectrum that can randomly match
doubly and triply charged theoretical fragments is approximately half
and a third, respectively, of the information that can match singly
charged fragments. Therefore, assuming a spectrum to come from a
triply charged precursor ion increases the potential for random
matching by a factor of 11/9 = 1.22 with respect to the doubly
doubly charged precursor hypothesis; hence correlation scores for triply
corrected were just divided by this factor, and no correction was
accounted for to doubly corrected precursor ions. Peptide length was ac-
counted for by using the formula $\log(x)/\log(2L)$, where $x$ is the charge-
corrected first, second, . . . , or $j$th best SEQUEST score and $L$ is the
peptide length, in a manner similar to those proposed by other au-
thors (9, 32). FDRs (17, 18, 19), defined as the proportion of expected
false positives within the population of identified peptides, were cal-
culated by counting up the number of peptides identified in the decoy
and in the target databases as illustrated in Fig. 1.

RESULTS

Properties of SEQUEST Average Score Distributions: the
Probability Ratio Concept—In this work we analyze two dif-
ferent kinds of distributions of SEQUEST scores. The first are
distributions that are characteristic of each one of the MS/MS
spectra. The single-spectrum distribution for a best score $x$
after searching $N$ sequence candidates is a function that
measures the probability that the MS/MS spectrum produces
a best score equal to or better than $x$ by chance. The math-
ematical properties of this kind of distributions are analyzed in
the supplemental information where these properties are also
tested in the practice using some examples obtained with real
MS/MS spectra. The second kind of score distributions, char-
acteristic of a compilation of spectra, are the average distri-
butions of the best score $I_N(x)$ that are obtained when a large
collection of MS/MS spectra is searched against a decoy
sequence database. We also analyze the distributions that are
obtained by considering the second best scores instead of
the best ones, the average probability distributions of the
second best score, or $H_N(x)$, and, in general, those of the $j$th
best score.

The mathematical properties of SEQUEST average distri-
butions are analyzed in the supplemental information; these
distributions were mathematically treated as the superimpo-
sition of the single-spectrum best score distributions from all
MS/MS spectra in the data set. Spectra were conceptually
classified according to a parameter called quality, whose
interpretation in the practice is that spectra with higher quality
tend to give better best scores by chance alone (see “Discus-
sion” and supplemental information for further details). Two
main properties with relevant implications in the practice were
derived from the analysis. One is that the values taken by
average probability distributions tend to be proportional to the
number of sequence candidates when the probabilities are
sufficiently low, i.e. when the best scores take high values,
and positive peptide identifications are expected to occur.
This means that the average probabilities become propor-
tional to database size. The second property is that the value
taken by the average distribution at the best score produced
by a given spectrum is a very accurate estimation of the fraction
of spectra in the collection whose quality is equal or
higher. The average distribution of the second best score has
the same properties. Because the second best score ob-
tained when a spectrum data set is searched against a target
database may be reasonably assumed to be random (a point
that is further discussed below), the value taken by the aver-
age score distribution at the second best score may be used
as an accurate estimator of the relative ranking position of the
spectrum, according to its quality, within the collection.

On the basis of this mathematical framework, we have
analyzed how the information provided by the first and sec-
ond best scores should be treated to make statistical infer-
ences about peptide identification in a large scale experiment.
We arrived to the conclusion that the conditional probability of
getting a first score equal to or better than $x_F$ when a second
best score equal to or better than $x_S$ is obtained may be
estimated by the ratio of the average probability of the first
score to the average probability of the second best score, that
is by $I_N(x_F)/I_N(x_S)$ (Equation 14 from the supplemental
information). The conditional probability may also be estimated by
using the average distribution of the second best score in the
denominator, i.e. $H_N(x_S)$, instead (see Equation 13 from the
supplemental information).

Properties of the Probability Ratio—The probability ratio is
easily computed in the practice as schematized in Fig. 1.
Briefly the spectra collection is searched against a decoy
database, and the set of best (Xcor) scores is used to con-
struct the average probability distribution for the best score
$I_N(x)$; this is done by ranking the scores in decreasing order
and plotting the normalized ranking position as a function of
the score. The curve is used to determine, by direct interpo-
lation, the average probability values for the best and second
best scores, $I_N(x_F)$ and $I_N(x_S)$, which are used to compute the
probability ratio. The same curve is also used to compute the
probability ratio from the scores determined after searching
the target database. When the best score from the target database search is better than the highest score in the distribution (as it happens when there is a very clear match between the mass spectrum and the correct peptide sequence), the probability was not computed but just assumed to be lower than that of the best decoy score (Fig. 1, arrow in left panel). The probability ratios obtained with the decoy and target databases are then used to compute the FDR (18, 19) at any probability ratio threshold as the proportion of peptides identified in the decoy database among the population of peptides identified in the target database (Fig. 1, right panel).

As explained in the preceding section, the average probability of the second best score is a good estimator of the relative ranking position of the spectrum in the collection according to its quality. Hence the probability ratio may be conceived as a score with an internal correction that, taking into account spectrum quality, compensates for the fact that the collection of MS/MS spectra is heterogeneous and that there are large differences from one spectrum to another in terms of the scores obtained against random peptide sequences.

The probability ratio may also be computed as the ratio \( \frac{I_n(x_i)}{H_n(x_i)} \) by constructing separately the curves for the first and the second best scores from the decoy database search. As exemplified in Fig. 2A, showing results obtained from the analysis of a proteome from Jurkat nuclei, there is no appreciable difference between this indicator and that calculated as depicted in Fig. 1 regarding the performance of peptide identification in terms of FDR. Similarly the probability ratio may, in general, be calculated by using the \( j \)th best score and the average probability distribution corresponding to
this jth best score; the FDR curves are also indistinguishable from those obtained using the original method (not shown). This behavior of the probability ratio is justified by the properties of these score distributions as explained in the supplemental information.

Because the probability ratio introduces an intrinsic correction for spectrum quality, factors known to influence distribution of SEQUEST scores have little effect, if any, and may be ignored. These include the uncertainty in charge state assignment by which spectra are usually searched two or more times assuming different charge states, i.e. as if they were in the practice two or more different spectra; this may affect the correct determination of FDRs by increasing the search space. This effect was tested by comparing the probability ratios obtained when the same spectra in different charge states are considered as if they were different spectra with the probability ratios obtained by selecting only the charge state associated with lower (i.e. more significant) values. As exemplified in Fig. 2A, selection of the best charge state was found not to alter the FDR curve at values lower than 0.05 and was only appreciable at FDR values above 0.2 (not shown). Hence a correct peptide charge assignment is not a critical factor for peptide identification when the probability ratio is used as indicator.

Charge state and peptide length are also known to affect Xcorr scores; as shown in Fig. 3, A and D, there was an appreciable shift toward higher values as peptide length or charge state increased as observed by other authors (9, 32). In clear contrast, when spectra were classified according to these factors, no appreciable shift was observed in the distribution of probability ratios (Fig. 3, B and E). This behavior is expected for random matches where the same quality (estimated by ranking position) is determined from the first and from the second best scores in their respective distributions. It should be noted that as peptide length or charge state increased, the ranking positions of SEQUEST scores tended to decrease, and hence their ratios, computed from smaller numbers, showed a significantly higher dispersion (Fig. 3, B and E). This numerical effect altered differentially the tail of the probability ratio distributions; as a consequence of this asymmetry a better performance of peptide identification was observed when FDRs were calculated separately in spectra subgroups classified by charge state (Fig. 2B). Because this effect is purely numerical, it was easily accounted for by a simple transformation of SEQUEST scores that corrected the bias imposed by charge and length (see “Experimental Procedures”). The distributions of “corrected” probability ratios for different charge states and peptide lengths were almost in-

**Fig. 3. Effect of charge and length on the distribution of Xcorr and probability ratio scores.** The collection of MS/MS spectra from Fig. 2 were searched against a decoy database, and the results were used to construct histograms of the total number of peptides as a function of their respective Xcorr scores (A and D) or as a function of their probability ratio scores, computed as $\frac{\text{xS}}{\text{H}}$ (B, C, E, and F). Spectra in the upper panels (A–C) were classified according to peptide mass (M) in three categories: 800–1300 Da (gray lines), 1300–1800 Da (black lines), and 1800–2300 Da (discontinuous lines). Spectra in the lower panels (D–F) were classified according to charge state: $z = 2$ (black lines) and $z = 3$ (discontinuous lines). The probability ratio was uncorrected (B and E) or corrected for charge and length as described under “Experimental Procedures” (C and F).
distinguishable (Fig. 3, C and F), and the FDR curves obtained without any kind of spectra classification showed an even better performance (Fig. 2B). As expected, separating the spectra according to charge state did not result in a significant improvement in comparison with the results obtained analyzing all the spectra together when the corrected probability ratio was used as indicator (Fig. 2B). Similar comparative results were obtained when the proteome from mesenchymal stem cells was analyzed (not shown).

To determine whether the particular quality distribution of spectra in the collection used to compute the probability ratio may affect the statistical significance, the probability ratios of spectra belonging to a collection were calculated using the average distribution of the best score obtained from the analysis of a different spectra collection of larger size, and vice versa the probability ratios of the latter were determined from the average distribution of the former. As shown in Fig. 4, A and B, the number of identified peptides did not change appreciably as a function of the FDR when a different average distribution was used to determine the probability ratio. The best score distributions of these two spectra collections are compared in Fig. 4, C and inset. These results further illustrate the robustness of the quality correction introduced by the probability ratio, which makes the final results tolerant towards the distribution of the MS/MS data set used to construct the average score distribution. Note that although different score distributions may be used to calculate the probability ratio, the error rates must always be calculated using the probability ratios obtained by searching the same data set against the corresponding decoy database and interpolating into the same score distribution.

Comparative Performance of the Probability Ratio Indicator—The results obtained using the probability ratio as indicator were compared with those obtained by a previously published empirical approach describing the behavior of the best score and the delta score on the basis of a two-variable Gaussian model (17). This served as an appropriate test of the method because the performance of that empirical method in comparison with other known methods, such as PeptideProphet, using the same data sets used in this work has already been analyzed, and it showed a better performance at FDRs lower than 0.05 (17). As shown in Fig. 5A, even without any correction for charge and length the probability ratio had a performance similar to that of the two-variable Gaussian model (which uses separate distributions for spectra with different charge states), whereas the performance of the corrected probability ratio indicator was clearly superior. Hence the results obtained by the probability ratio method, developed on the basis of purely analytical considerations, despite its conceptual and computational simplicity and the absence of adjustable functions, fitting parameters, and spectra classifications, are superior to those obtained by empirical methods specifically designed to deal with SEQUEST score parameters.

**Fig. 4.** Effect of the average score distribution used to calculate the probability ratio on the performance of the method. The black lines represent the FDR curves obtained from the analysis of tryptic peptides from the human Jurkat nuclei proteome (A) and from the human mesenchymal stem cells proteome (B) by the conventional probability ratio method. The gray lines are the FDR curves obtained when the probability ratios from the peptides of the former proteome are calculated by interpolating in the average score distribution of best decoy scores obtained from the analysis of the latter proteome (A) and vice versa (B). The score distributions from the nuclear proteome (gray points) and the stem cell proteome (black points) are shown using a normalized frequency plot (C) or using a normalized rank plot in a logarithmic scale (C, inset).
Although statistical methods used to validate peptide identifications using SEQUEST scores are based on average score distributions constructed with the results obtained from a collection of spectra, there are other scoring methods based on the analysis of single-spectrum score distributions, which may hence be applied to the MS/MS spectrum considered individually. For instance, survival functions constructed from frequency histograms of peptide scores obtained during the database search have been used to calculate the expectation or E value of peptide identification (27). This approach has been applied to scores, like Sonar (21), for which a stochastic probability distribution is not available in a general, parameterized form (24). As noted before, these survival functions are equivalent to what we call here single-spectrum score distributions. A question that remains unanswered is whether by using single-spectrum distributions with SEQUEST, instead of average score distributions, a better performance of peptide identification may be achieved. To address this point, we made a computationally exhaustive determination of the single-spectrum score distributions of each one of the spectra derived from the two model proteomes used in this study as in supplemental Fig. 1. The probabilities associated to the observed SEQUEST Xcorr scores were then directly computed from these score distributions and the number \( N \) of sequence candidates. Although this approach was too time-consuming for routine use in the practice, the results were very informative in the context of this work. The comparative performance of peptide identification obtained by this approach was compared with that of the probability ratio and the two-variable empirical model in Fig. 5A. As shown, the FDR curves obtained using this method are similar but not better than those obtained by the empirical model and the uncorrected probability ratio method but clearly worse than that obtained by using the corrected probability ratio. This finding indicates that, although the method based on the empirical calculation of single-spectrum distributions allows the determination of peptide identification confidence (i.e., the probability associated to the best score) in spectra considered individually, it does not produce significant improvements in comparison with the probability ratio when it comes to the validation of positive peptide identification on the basis of the error rate of the experiment considered as a whole.

Finally the performance of peptide identification obtained with the probability ratio was also compared with that obtained by using fixed thresholds for both Xcorr and \( \Delta C_n \) that 0.07 without previous charge classification (without charge sep; discontinuous thin line); lines were drawn by fixing the \( \Delta C_n \) threshold optimized in each case and varying the Xcorr threshold. In C the FDR values obtained by using separate database searches were compared with those obtained by searching against a concatenated database for either the fixed, optimized SEQUEST score thresholds method (FT; black points) or the probability ratio method (PR; white points). The numbers indicate the slope of the lines.
are dynamically adjusted to obtain a desired FDR (20). This was accomplished by searching the same data set against a concatenated database constructed with the target and decoy databases and calculating the rate of false positives by doubling up the number of decoy matches above the threshold as suggested previously (20). The two SEQUEST scores were iteratively varied until an optimum performance was obtained at FDRs of 0.005, 0.01, and 0.07; this was performed separately in spectra subsets classified according to charge state, adding up the number of identified peptides at each FDR. Fig. 5B shows the curves obtained by fixing optimum ΔCn thresholds for each charge state and FDR value and varying the Xcorr threshold. As shown, although the curves approached that obtained by using the probability ratio at the selected FDRs, the performance of peptide identification was worse in the rest of the curves. Note that with this procedure spectra classification into different categories such as charge state is critical to obtain the best performance (compare for instance the FDR curves obtained with and without spectra classification according to charge state in Fig. 5). In clear contrast, only one parameter without spectra classification was needed to construct the entire FDR curve using the probability ratio method, and it allowed a superior performance of peptide identification. In addition, by sorting the identified peptides in increasing order of probability ratio, as schematized in Fig. 1, it is possible to assign an FDR value to each one of the tentative peptide identifications, indicating the error rate in the population of peptides that have an equal or better probability ratio score. This allows very practical outputs where FDR cutoffs can be relaxed to inspect lower confidence identifications without losing the error information associated to good peptide identifications.

In a recent work, it was suggested that estimations of FDRs using decoy and target databases should be performed by a single search against a concatenated database and not by searching separately target and decoy databases (20). The concatenated approach allows a direct competition between decoy and target sequences for the best scores, and thus the relatively high scores that are typically produced in decoy databases by MS/MS spectra corresponding to true peptides are not taken into account (20). In other words, the concatenated database strategy avoids the quality effect produced by this subset of MS/MS spectra. In the case of the probability ratio, although searches are performed separately against the target and decoy databases (Fig. 1, right panel), spectrum quality is intrinsically taken into account by the indicator, and hence the quality effect becomes negligible. To illustrate this property, FDRs obtained by using the competition strategy in a composite database were compared with those obtained by searching separately against the target and decoy databases with both the conventional method based on discrete Xcorr/ΔCn thresholds and the probability ratio method. As shown in Fig. 5C, when adjustable thresholds of SEQUEST parameters are used according to the conventional method, FDRs increased by 40% when the separate search method was compared with the concatenated search strategy in good agreement with results published previously (20). In clear contrast, no differences in FDRs were appreciable when the two search methods were compared using the probability ratio. Hence this indicator also displays a robust behavior in relation to the strategy used to determine error rates using decoy databases.

Similar results have been consistently obtained when analyzing other factors that influence search space, like variable modifications, maximum number of missed cleavages, and precursor mass tolerance. In general, factors that increase the number of sequence candidates make more reliable the probability ratio as indicator. The performance of the indicator was also tested in situations where the search space is very small. For this end, we applied the corrected probability ratio method to a collection of more than 10,000 MS/MS spectra obtained from an E. coli protein extract using an LTQ-Orbitrap machine; the data set was searched against a protein database containing only entries from this species with a precursor mass tolerance of 10 ppm, no variable modifications, and no missing cleavages. The performance of peptide identification in terms of FDR curves remained superior to the other empirical methods tested here (not shown).

DISCUSSION

Almost all methods described to make statistical inferences for peptide identification using SEQUEST are based on the analysis of the set of best scores (Xcorr scores) obtained after searching sequence databases with a large collection of MS/MS spectra (9, 12, 13, 16, 17). The best scores are grouped together, ranked, and used to construct a cumulative distribution of scores, which is used to determine the statistical significance of peptide identification and also the FDR. These cumulative distributions may be conceived as statistical estimates of the average probability distribution of the best score, a function that evaluates the probability of obtaining a spectrum yielding an equal or better best score in the collection. In this work we perform an analytical study of these average score distributions. For this end, we analyzed the properties of the single-spectrum distributions, which are functions characteristic of each one of the spectra, and expressed the average distribution as a combination of single-spectrum distributions. There is a conceptual relation between the single-spectrum distributions and the p and E values, which are common parameters used by database searching algorithms (11, 25, 27, 28); this relation is explained in the supplemental information. The fundamental reason to express the average distributions as a function of the underlying single-spectrum distributions is that the latter may be markedly dissimilar in different spectra, and hence large collections of spectra cannot be expected to have a homogeneous statistical behavior.

Our analysis takes into account the fact that the number of sequence candidates against which MS/MS spectra are
searched in databases is usually a large number. This has relevant consequences on the properties of average score distributions, depending on the magnitude of the probability. In the low probability region where positive peptide identifications are expected to occur, the average probability is proportional to the number of sequence candidates and therefore to database size. This property was demonstrated in the practice by using model average SEQUEST score distributions constructed by single-spectrum distributions from real MS/MS spectra. Another fundamental property, also validated in the practice, is that the value taken by the average distribution for a given score reflects very accurately the fraction of spectra in the collection that have higher quality, i.e., that tend to produce a better score by chance alone. The term quality has been applied previously to MS/MS spectra referring to the spectra features that make them more likely to be derived from the fragmentation of a peptide (33, 34). In our study quality is introduced as a mathematical concept whose practical interpretation is that spectra having higher quality are those that tend to produce higher scores when they are searched against a random sequence database. The relation between quality and score ranking position may appear quite obvious because the average distribution is constructed by ranking the spectra according to their best score, and hence the relative ranking position is just the fraction of spectra that have a higher score. However, it should be noted that the cumulative score distribution constructed in the practice from a decoy database search is but an estimation of the underlying average distribution of the collection of MS/MS spectra (which is unknown) and that the ranking position of a spectrum may only be used to estimate its also unknown quality in the same sense that a finite number of determinations of a normally distributed variable may only be used to estimate its hidden normal distribution. Because the number of spectra used in large scale identification experiments is usually very large, the cumulative score distributions obtained from a database search may be considered reliable estimations of the underlying average score distributions. This is supported in the practice by the fact that when the same data set is searched against different decoy databases having equivalent distributions of random peptide sequences (for instance, reversed and pseudoreversed databases (20)), the same number of random matches are obtained above discrete score thresholds (20). However, the same conclusion concerning estimation of quality from the ranking position of a given spectrum cannot be reached unless the behavior (i.e., the dispersion) of the random best scores produced by this particular spectrum is analyzed; this information is contained within the single-spectrum score distributions. Our analysis shows that the single-spectrum distributions for the best score rapidly become steep functions as the number N of sequence candidates increases, making the density distributions of the best score very narrow, hence having a very small dispersion. It is only after these properties are analyzed that we can conclude that the quality of a spectrum may be reliably estimated from the ranking position of the spectrum, according to its score, within the collection.

Because the average score distributions reflect mainly the quality distribution, they should be different when different samples are analyzed and would be affected only by factors altering the quality distribution of the sample. For instance, increasing peptide concentration in the sample is expected to increase the proportion of spectra having a good fragmentation and hence the proportion of spectra tending to produce high scores by chance alone; the quality distribution among the spectra observed would then be shifted toward higher values. Changing database size, on the other hand, by altering the number of sequence candidates presented to each one of the spectra will also affect the distribution of scores; this would produce some displacement in the bulk of the distribution but not significantly changing their shape as described in other works (17). However, this effect will be particularly important in the tail corresponding to the low probability region where probabilities will tend to be proportional to database size. To our knowledge these effects have not been treated before from a purely analytical viewpoint.

Our results once more highlight the fact that the relation between average probabilities or score thresholds and error rates is strongly dependent on the experiment, database size, and searching conditions used and cannot be directly extrapolated from data obtained with different spectra data sets or different databases. Therefore, it is impossible to establish fixed peptide identification criteria of universal validity. Surprisingly using fixed SEQUEST peptide identification thresholds, estimating error rates using databases with a size very different than that used to analyze the data, and extrapolating results obtained previously using a different data set are still common practices. Another important concern is the extrapolation of curves, such as Gaussian or γ functions, used to fit average score distributions to estimate peptide identification probability. As shown in this work, average score distributions are in a completely different regime in the low probability region, and therefore no extrapolations of any kind from the bulk of the distribution should be performed.

When SEQUEST scores are statistically analyzed in large scale experiments, the two parameters usually considered are Xcorr (or the best score) and the delta score ΔCn (or the relative difference between the best and second best scores) (1–3); these two parameters are assumed to contain the most relevant information about the tentative peptide identification (9, 12, 16, 17). These two parameters are usually treated as independent indicators, and different approaches have been empirically proposed to take them together in the form of a single statistical score (9, 12, 16, 17). However, it is still unclear whether or not these algorithms are optimal, and this issue remains largely unexplored from the analytical viewpoint. In this work we have analyzed how to treat the information obtained from a database search of a large collection.
of MS/MS spectra when only the information provided by the best and the second best scores from each hit is available. The mathematical analysis led us to derive a new indicator, the probability ratio, which efficiently takes into account the information provided by SEQUEST database searches using large collections of MS/MS spectra from which only the best and second best scores are considered. Although the derivation of the probability ratio expression is not trivial, its meaning is easy to understand on the basis of the properties of average distributions noted in the preceding paragraph. The average probability of the first score \( I_m(x_r) \) measures how likely it is to obtain a best score equal to or better than \( x_r \) by chance in the collection of MS/MS spectra; however, a low average probability does not necessarily indicate that the identification is correct (a non-random match) because it may also indicate that spectrum quality is very high. For instance, if the collection contains 100,000 MS/MS spectra, an average probability of \( 10^{-5} \) for a best score \( x_r \) may occur just by chance if this score is reached by the spectrum with highest quality in the data set. Hence using the best score as the only confidence indicator may underestimate the confidence of identification in the population of spectra having high quality; this may produce false positives. This effect is corrected by taking into account the quality of the spectrum, which, as noted in preceding sections, may be estimated from the average probability of the second best score. Hence the denominator in the probability ratio may be viewed as a factor that exactly compensates for the quality effect. Following the example above, the average probability of the second best score would also be in the order of \( 10^{-5} \); hence the probability ratio would yield a value near 1, indicating that this match could have been produced by chance. Although this effect may be evaluated, in a certain sense, by calculating the relative difference between the first and the second best scores, i.e. the delta score \( \Delta C_m \), until now it has remained unclear how to analyze statistically this parameter in conjunction with the best score; the mathematical analysis of this work gives for the first time a conceptual framework to analyze statistically the information provided by these two parameters.

The probability ratio can be calculated by a simple and non-parametric approach, which avoids using adjustable fitting parameters or performing any kind of assumptions or extrapolations in the tail of the random average distribution. Hence no inaccuracies of any kind are introduced in the calculation of average probabilities, and the FDRs can be straightforwardly calculated for each one of the potentially identified peptides. The probability ratio concept is completely new, is derived on the basis of purely analytical considerations, and is much simpler to apply in practice than other empirical approaches. Despite its simplicity, the probability ratio yields a performance superior to other previously published empirical statistical methods, and it has additional advantages. By introducing an internal quality correction for each one of the spectra, this method makes it unnecessary to classify spectra according to parameters such as peptide length or charge state (9, 17), and predefined mathematical functions and adjustable parameters do not need to be chosen to fit the data (9, 17, 24, 35). Besides the quality correction introduced in the method makes it very robust with respect to the particular quality distribution used and with respect to the method used to estimate FDR using decoy databases.

Although in the most general case the second best score may be considered the consequence of a random match, in some situations it can be expected that a significant proportion of second best matches may be derived from sequences having high homology with correct peptide sequences. This is the case for searches against unfiltered, whole protein databases where the peptides identified may belong to proteins from a variety of different organisms. In these cases it may be useful to take advantage of the theoretical prediction that the quality fraction may also be calculated from average score distributions constructed from the third, fourth, . . . best scores. Hence it is possible to calculate a probability ratio using for instance the best and fourth best scores instead of the second or even an averaged quality obtained from the distributions of the second, third, and fourth best scores. This option is included in the program.

A result with relevant implications in the practice is that determining statistical significance of peptide hits by constructing the single-spectrum best score distributions for each one of the MS/MS spectra did not improve the results obtained using the probability ratio and the first and second best scores only in terms of performance of peptide identification as a function of the error rate in a large scale identification experiment. This suggest that the information provided by the first and second best SEQUEST scores may in the practice be sufficient to establish near optimal confidence values, and no practical improvements will be obtained from the analysis of large collections of scores produced by random matching. The comparative performance of the probability ratio method in comparison with other published methods together with the simplicity and robustness of the conceptual approach and the non-parametric nature of the probability ratio makes this method particularly attractive for automatic, unattended calculation of error rates in large scale peptide identification experiments.

* This work was supported in part by Grants BIO2003-01805 and BIO2006-10085 from the Ministry of Education and Science of Spain and Grants GR/SAL/0141/2004 and BIO/0194/2006 from the Comunidad Autónoma de Madrid (CAM), the Fondo de Investigaciones Sanitarias (Ministerio de Sanidad y Consumo, Instituto Salud Carlos III, RECAVA); and an institutional grant from Fundación Ramón Areces (to Centro de Biología Molecular Severo Ochoa). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.
§ These authors contributed equally to this work.
¶ Present address: Centro Nacional de Biotecnologia, Universidad Autónoma de Madrid, 28049 Cantoblanco, Madrid, Spain.
** Present address: Biological Sciences Division and Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA 99352.
†† A postdoctoral fellow from Junta de Comunidades de Castilla-La Mancha (JCCM). Present address: Inst. de Tecnologías Química y Medioambiental (ITQUMIA), Campus Universitario s/n, 13071 Ciudad Real, Spain.
§§ To whom correspondence should be addressed: Centro de Biología Molecular Severo Ochoa, Universidad Autónoma de Madrid, 28049 Cantoblanco, Madrid, Spain. Tel.: 34-91-497-8276; Fax: 34-91-497-8087; E-mail: yavezque@cbm.uan.es.

REFERENCES

1. Eng, J. K., McCormack, A. L., and Yates, J. R., III (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 5, 976–989
2. Yates, J. R., III, Eng, J. K., McCormack, A. L., and Schiltz, D. (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal. Chem. 67, 1426–1436
3. Yates, J. R., III, Eng, J. K., and McCormack, A. L. (1999) Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. Anal. Chem. 71, 3202–3210
4. Link, A. J., Eng, J., Schiltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., and Yates, J. R., III (1999) Direct analysis of protein complexes using mass spectrometry. Nat. Biotechnol. 17, 676–682
5. Washburn, M. P., Wolters, D., and Yates, J. R., III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat. Biotechnol. 19, 242–247
6. Peng, J., Elias, J. E., Thoreen, C. G., Licklider, L. J., and Gygi, S. P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J. Proteome Res. 2, 43–50
7. Florens, L., Washburn, M. P., Raine, J. D., Anthony, R. M., Grainger, M., Haynes, D. J., Moch, J. K., Muster, N., Sacci, J. B., Tabb, D. L., Witney, A. A., Wolters, D., Wu, Y., Gardner, M. J., Holder, A. A., Sinden, R. E., Yates, J. R., and Carucci, D. J. (2002) A proteomic view of the Plasmodium falciparum life cycle. Nature 419, 520–526
8. Qian, W. J., Liu, T., Monroe, M. E., Strittmatter, E. F., Jacobs, J. M., Kangas, L. J., Petritis, K., Camp, D. G., II, and Smith, R. D. (2005) Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: the human proteome. J. Proteome Res. 4, 53–62
9. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal. Chem. 74, 5383–5392
10. Tabb, D. L., MacCoss, M. J., Wu, C. C., Anderson, S. D., and Yates, J. R., III (2003) Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. Anal. Chem. 75, 2470–2477
11. Sadygov, R. G., and Yates, J. R., III (2003) A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. Anal. Chem. 75, 3792–3798
12. Moore, R. E., Young, M. K., and Lee, T. D. (2002) Qscore: an algorithm for validating protein identifications using a modified SEQUEST algorithm. Anal. Chem. 74, 5593–5599
13. Anderson, D. C., Li, W., Payan, D. G., and Noble, W. S. (2003) A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. J. Proteome Res. 2, 137–146
14. Razumovskaya, J., Olman, V., Xu, D., Uberbacher, E. C., Ver Berkmoes, N. C., Hettich, R. L., and Xu, Y. (2004) A computational method for assessing peptide-identification reliability in tandem mass spectrometry analysis with SEQUEST. Proteomics 4, 961–969
15. Kislinger, T., Rahman, K., Radulovic, D., Cox, B., Rossant, J., and Emili, A. (2003) PRISM, a generic large scale proteomic investigation strategy for mammals. Mol. Cell. Proteomics 2, 96–106
16. Lopez-Ferrer, D., Martinez-Bartolome, S., Villar, M., Campillos, M., Maroto, F., and Vazquez, J. (2004) Statistical model for large-scale peptide identification in databases from tandem mass spectra using Pro看待。Anal. Chem. 76, 6853–6860
17. Benjamini, Y., Elmer, G., Kafkafi, N., and Golani, I. (2001) Controlling the false discovery rate in behavior genetics research. Behav. Brain Res. 125, 279–284
18. Storey, J. D., and Tibshirani, R. (2003) Statistical significance for genome-wide studies. Proc. Natl. Acad. Sci. U. S. A. 100, 9440–9445
19. Elias, J. E., and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods 4, 207–214
20. Field, H. I., Fenyo, D., and Beavis, R. C. (2002) RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. Proteomics 2, 36–47
21. Elias, J. E., Gibbons, F. D., King, O. D., Roth, F. P., and Gygi, S. P. (2004) Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nat. Biotechnol. 22, 214–219
22. Bafna, V., and Edwards, N. (2001) SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. Bioinformatics 17, Suppl. 1, S13–S21
23. Colinge, J., Masselot, A., Giron, M., Dessingy, T., and Maguin, J. (2003) OLAV: towards high-throughput tandem mass spectrometry data identification. Proteomics 3, 1454–1463
24. Geer, L. Y., Markey, S. R., Kowalak, J. A., Wagner, L. X., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004) Open mass spectrometry search algorithm. J. Proteome Res. 3, 958–964
25. Colinge, J., Masselot, A., Cusin, I., Mahe, E., Niknejad, A., Argoud-Puy, G., Reffas, S., Bederr, N., Gleizes, A., Rey, P. A., and Bougueneret, L. (2004) High-performance peptide identification by tandem mass spectrometry allows reliable automatic data processing in proteomics. Proteomics 4, 1977–1984
26. Fenyo, D., and Beavis, R. C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. Anal. Chem. 75, 768–774
27. Altshul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. J. Mol. Biol. 215, 403–410
28. Cari, S., Aebersold, R., Baldwin, M., Burlingame, A., Clauer, K., and Nesvizhskii, A. (2004) The need for guidelines in publication of peptide and protein identification data. Working Group on Publication Guidelines for Peptide and Protein Identification Data. Mol. Cell. Proteomics 3, 531–533
29. Baldwin, M. A. (2004) Protein identification by mass spectrometry. Issues to be considered. Mol. Cell. Proteomics 3, 1–9
30. Ogueda, S., Munoz, J., Obregón, E., Delgado-Baeza, E., and Garcia-Ruiz, J. P. (2002) Prolactin is a component of the human synovial liquid and modulates the growth and chondrogenic differentiation of bone marrow-derived mesenchymal stem cells, Mol. Cell. Endocrinol. 190, 51–63
31. Nesvizhskii, A. I., Roos, F. F., Grossmann, J., Vogelzang, M., Eddes, J. S., Cruissem, W., Baginsky, S., and Aebersold, R. (2006) Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data. Toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. Mol. Cell. Proteomics 5, 652–660
32. Bern, M., Goldberg, D., McDonald, W. H., and Yates, J. R., III (2004) Automatic quality assessment of peptide tandem mass spectra. Bioinformatics 20, Suppl. 1, i49–i54
33. Strittmatter, E. F., Kangas, L. J., Petritis, K., Mottaz, H. M., Anderson, G. A., Shen, Y., Jacobs, J. M., Camp, D. G., II, and Smith, R. D. (2004) Application of peptide LC retention time information in a discriminant function for peptide identification by tandem mass spectrometry. J. Proteome Res. 3, 760–769
34. Sadygov, R. G., Liu, H., and Yates, J. R. (2004) Statistical models for peptide validation using tandem mass spectral data and protein amino acid sequence databases. Anal. Chem. 76, 1664–1671