Statistical Model for Large-Scale Peptide Identification in Databases from Tandem Mass Spectra Using SEQUEST

ARTICLE *in* ANALYTICAL CHEMISTRY · JANUARY 2005

Impact Factor: 5.64 · DOI: 10.1021/ac049305c · Source: PubMed

Daniel Lopez-Ferrer
Caprion
33 PUBLICATIONS **1,371** CITATIONS
SEE PROFILE

Salvador Martínez-Bartolomé
The Scripps Research Institute
28 PUBLICATIONS **781** CITATIONS
SEE PROFILE

Margarita Villar
University of Castilla-La Mancha
57 PUBLICATIONS **1,144** CITATIONS
SEE PROFILE

Jesús Vázquez
Spanish National Centre for Cardiovascular R...
157 PUBLICATIONS **5,049** CITATIONS
SEE PROFILE

Statistical Model for Large-Scale Peptide Identification in Databases from Tandem Mass Spectra Using SEQUEST

Daniel López-Ferrer,[†] Salvador Martínez-Bartolomé,[†] Margarita Villar,[†] Mónica Campillos,[†,‡] Fernando Martín-Maroto,[§] and Jesús Vázquez*,[†]

Centro de Biología Molecular "Severo Ochoa"-CSIC, 28049 Cantoblanco, Madrid, Spain, and ThermoFinnigan, River Oaks Parkway 355, San Jose, California 95134

Recent technological advances have made multidimensional peptide separation techniques coupled with tandem mass spectrometry the method of choice for high-throughput identification of proteins. Due to these advances, the development of software tools for large-scale, fully automated, unambiguous peptide identification is highly necessary. In this work, we have used as a model the nuclear proteome from Jurkat cells and present a processing algorithm that allows accurate predictions of random matching distributions, based on the two SEQUEST scores Xcorr and ΔCn. Our method permits a very simple and precise calculation of the probabilities associated with individual peptide assignments, as well as of the false discovery rate among the peptides identified in any experiment. A further mathematical analysis demonstrates that the score distributions are highly dependent on database size and precursor mass window and suggests that the probability associated with SEQUEST scores depends on the number of candidate peptide sequences available for the search. Our results highlight the importance of adjusting the filtering criteria to discriminate between correct and incorrect peptide sequences according to the circumstances of each particular experiment.

During the last years, a remarkable development has taken place in the field of proteomics, or the global analysis of gene expression at the protein level. The classical method for proteome analysis is based on the separation of proteins by two-dimensional gel electrophoresis followed by protein identification by mass spectrometry. Over the past few years, many procedures for high-throughput analysis of the proteome components have been developed, most of them based on gel-free approaches. These procedures are based on the digestion in solution of the unseparated protein mixture, followed by large-scale MS/MS analysis of the generated peptides. Because of the enormous complexity of the peptide mixture, this approach needs a rather thorough separation of the mixture prior to MS/MS analysis. Important recent technical advances related to separation techniques such as multidimensional chromatography, mass spectrometry, and data mining approaches have increased sensitivity, reproducibility, and throughput of proteome analysis while establishing the foundations for the development of an integrated technology.

Gel-free proteomics analyses are highly dependent on the software that searches sequence databases using mass spectrometry-derived data. Database searching algorithms try to match the experimental tandem mass spectra of the peptides to the predicted mass spectra of the amino acid sequence contained in the database. The first and one of the most widely used database searching algorithms is SEQUEST;[1−3] this software computes a cross correlation (Xcorr) function to assess the quality of the match between a tandem mass spectrum and amino acid sequence information in a database. This value is a database-independent measure that is dependent on the quality of the tandem mass spectrum and the quality of its fit to the model spectrum.[1] SEQUEST also computes a parameter called ΔCn, which is the normalized score obtained from the difference between the first- and the second-ranked sequences. This value is highly dependent on database size, search parameters, and sequence homologies.[1−3] While these scores have been widely used for peptide identification in a considerable number of research projects, they do not provide a means to interpret the identification results automatically. For an accurate identification of peptides from tandem mass spectra, it is necessary to distinguish correct peptide assignments from false identifications; however, SEQUEST scores are known not to provide a clear separation among these data subsets.[4−6] For small data sets, it is possible to manually inspect the MS/MS spectra to verify peptide assignments; however, this approach is time-consuming and needs expertise in spectra interpretation and, therefore, is not feasible for high-throughput analysis of large data sets, such as those obtained by bidimensional chromatography, which typically contain tens of thousands of spectra.

* To whom correspondence should be addressed. E-mail: jvazquez@cbm.uam.es. Phone: +34 91 497 8276. Fax: +34 91 497 8087.

[†] Centro de Biología Molecular "Severo Ochoa"-CSIC.

[‡] Current address: EMBL, Meyerhofstrasse 1, 69117 Heidelberg, Germany.

[§] ThermoFinnigan.

[1] Eng, J. K.; McCormack, A. L.; Yates, J. R., 3rd. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.

[2] Yates, J. R., 3rd.; Eng, J. K.; McCormack, A. L.; Schieltz, D. M. *Anal. Chem.* **1995**, *67*, 1426−1436.

[3] Yates, J. R., 3rd.; Eng, J. K.; McCormack, A. L. *Anal. Chem.* **1995**, *67*, 3202−3210.

[4] Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383−5392.

[5] Tabb, D. L.; MacCoss, M. J.; Wu, C. C.; Anderson, S. D.; Yates, J. R., 3rd. *Anal. Chem.* **2003**, *75*, 2470−2477.

[6] Sadygov, R. G.; Yates, J. R., 3rd. *Anal. Chem.* **2003**, *75*, 3792−3798.

10.1021/ac049305c CCC: $27.50 © 2004 American Chemical Society
Published on Web 11/02/2004

Several filtering criteria have been proposed to separate correct from incorrect peptide assignments, based on SEQUEST scores and on the properties of identified peptides.[7−10] These criteria were empirically developed and are based on a set of rules involving SEQUEST scores and properties of the assigned peptides; these criteria have been later used in a number of research projects. However, these criteria were originally developed to analyze particular data sets, and it is not possible to predict the rate of correct and false identifications that result from their application to data generated using other sample preparations, in different experimental conditions and against databases of various sizes. More recently, machine learning algorithms[11−13] have been used to separate correct from incorrect peptide identifications, obtaining a better separation than that obtained by using SEQUEST scores alone.

In a recent work, a statistical approach was proposed to estimate the probabilities of correct and incorrect peptide identifications.[4] These authors analyzed four SEQUEST scores to assess their relative contribution to discriminate among true and false positives;[4] this analysis revealed that Xcorr and ΔCn are the scores contributing to most of the discrimination achieved by SEQUEST. These authors then used a discriminant function constructed as a linear combination of these two scores; a training data set composed of a collection of fragmentation spectra from several proteins of known sequence was then used to determine the discriminant score negative and positive distributions, which were modeled by a gamma and a Gaussian distribution, respectively.[4] Subsequently, a curve fitting is carried out using the data set to be analyzed. This method assigns a probability to each particular peptide assignment in the context of the whole experiment, allowing the estimation of the false positive error rate, and has the advantage that it may be adapted to data sets generated in different experimental conditions, with different mass spectrometers, or by searching against databases of different size.

In this work, we propose a statistical model to evaluate SEQUEST results that is also based on the construction of score distributions. In our method, instead of using a training data set, the statistical score distributions of random or incorrect peptide assignments are constructed from the experiment data set itself by using random databases, although we show that it is also possible to determine the distributions using conventional databases. By using appropriate mathematical transformations, we show that the overall distribution of Xcorr and ΔCn scores may be easily modeled by a simple, two-variable Gaussian distribution, from which a simple mathematical formula is deduced to compute the probability that a peptide match is not produced by chance alone. Despite the simplicity of the model, we demonstrate that the method allows a very accurate estimation of the false discovery rate (FDR) and may perform better than other previously published methods at low FDRs. In this work, we also explore the behavior of the random SEQUEST score distributions as a function of database size and precursor mass window and propose a procedure to estimate the behavior of SEQUEST scores as a function of the number of different candidate peptide sequences correlated. The simplicity and general applicability of our model make it a particularly attractive approach for the automated analysis of large MS/MS data sets.

EXPERIMENTAL SECTION

Preparation and Digestion of the Model Proteome. Nuclear proteins from Jurkat cells were extracted as described.[14] The protein concentration in the nuclear extract was quantified by the Bradford procedure. The nuclear protein extract (100 μg of total protein) was subjected to cold acetone precipitation, lyophilized to dryness, dissolved in reducing solution (8M urea, 25 mM ammonium bicarbonate, 10 mM DTT, pH 8), and incubated for 1 h at 37 °C. Iodoacetamide was then added to a final concentration of 50 mM and and the resultant mixture incubated for 45 min at room temperature in darkness. The mixture was diluted 4-fold to reduce urea concentration, and after the addition of 2 μg of trypsin (1:50 protease-to-protein ratio), was incubated at 37 °C overnight.

Two-Dimensional Liquid Chromatography with Tandem Mass Spectrometry. The tryptic peptide pool was lyophilized to dryness, dissolved in solvent A, and loaded onto a 0.18 mm × 150 mm BioBasic SCX column (ThermoHypersil-Keystone) at a flow of 5 μL/min using a Smart microHPLC system (Pharmacia, Uppsala, Sweden) with automatic fraction collection equipped with a flow splitter and working at 40 μL/min. Peptides were fractionated by applying a 80-min gradient from 5 to 35% solvent B (Solvent A: 5 mM phosphate buffer, 25% acetonitrile, pH 3.0. Solvent B: same as (A) with 350 mM KCl). Fourteen peptide fractions were collected, lyophilized to remove acetonitrile by vacuum centrifugation, and then analyzed by RP-HPLC-MS/MS using a 0.18 mm × 150 mm BioBasic 18 RP column (ThermoHypersil-Keystone), operating at ~1.5 μL/min, connected to a Surveyor HPLC system on-line with a LCQ-DECA XP ion trap mass spectrometer (Thermo Finnigan, San Jose, CA). Peptides were eluted using a 90-min gradient from 5 to 60% solvent B (Solvent A: 0.5% acetic acid. Solvent B: 0.5% acetic acid, 80% acetonitrile). Peptides were detected in survey scans from 400 to 1600 amu (8 μscans), followed by three data-dependent MS/MS scans, using an isolation width of 3 amu, a normalized collision energy of 35%, and dynamic exclusion, applied during 3-min periods.

Data Processing. Peptide identification from raw data was performed using the SEQUEST algorithm (Bioworks 3.1 package, Thermo Finnigan). For database searches, the nr.fasta (December 1, 2003) and swissprot.fasta protein databases (November 4, 2003) were used. Equine.fasta, yeast.fasta, rat.fasta, and human.fasta protein databases were built from the nr.fasta database using Fasta Database utilities. The following constraints were used for the searches: tryptic cleavage after Arg and Lys, up to two missed

[7] Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., 3rd. *Nat. Biotechnol.* **1999**, *17*, 676−682.

[8] Washburn, M. P.; Ulaszek, R.; Deciu, C.; Schieltz, D. M.; Yates, J. R., 3rd. *Anal. Chem.* **2002**, *74*, 1650−1657.

[9] Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. *J. Proteome Res.* **2002**, *2*, 43−50.

[10] Florens, L.; Washburn, M. P.; Raine, J. D.; Anthony, R. M.; Grainger, M.; Haynes, J. D.; 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.; Carucci, D. J. *Nature* **2002**, *419*, 520−526.

[11] Anderson, D. C.; Li, W.; Payan, D. G.; Noble, W. S. *J. Proteome Res.* **2003**, *2*, 137−146.

[12] Elias, J. E.; Gibbons, F. D.; King, O. D.; Roth, F. P.; Gygi, S. P. *Nat. Biotechnol.* **2004**, *22*, 214−219.

[13] Razumovskaya, J.; Olman, V.; Xu, D.; Uberbacher, E. C.; VerBerkmoes, N. C.; Hettich, R. L.; Xu, Y. *Proteomics* **2004**, *4*, 961−969.

[14] Armesilla, A. L.; Lorenzo, E.; Gomez del Arco, P.; Martinez-Martinez, S.; Alfranca, A.; Redondo, J. M. *Mol. Cell. Biol.* **1999**, *19*, 2032−2043.

cleavage sites, and tolerances of ±2 Da for precursor ions, unless otherwise indicated in the text, and of ±0.8 Da for MS/MS fragment ions. SEQUEST searches were performed allowing optional Met oxidation and fixed Cys carbamidomethylation. Inverted databases were constructed using a Visual Basic program. Statistical analysis and curve fitting were performed using Microsoft Excel spreadsheets. The mean and the standard deviation from score distributions were determined by least-squares fitting to Gaussian curves; optimal fitting to the right tail were obtained by discarding the points in the left tail that deviate more than one standard deviation from the mean. This procedure was more accurate for making probability predictions than the direct statistical calculation of the mean and standard deviation from the bulk of data.

Estimation of the False Discovery Rate. As a measure of significance, we used the FDR,[15,16] defined as the proportion of false positives among the population of spectra passing a given p-threshold, that is, $\mathrm{FDR}_p = E_p/O_p$, where O_p and E_p are the observed number of spectra and the number of spectra expected to score with probability lower or equal than p, respectively. The number of spectra correctly assigned, C_p, is estimated from the relation $C_p = O_p - E_p$, and the expected number of random or false assignations is calculated as $E_p = p(T - C_p)$, where T is the total number of spectra, and $(T - C_p)$ is used as an estimate of the total number of random or false assignations. Since in our method C_p is unknown, it was replaced in the equation $E_p = p(T - C_p)$ by $O_p - E_p$, obtaining

$$E_p = (T - O_p)\,\frac{p}{1 - p};\quad \mathrm{FDR}_p = \frac{T - O_p}{O_p}\,\frac{p}{1 - p}$$

It should be noted that this method to estimate FDR in no case produces an underestimation of the expected number of false positives. This is exemplified in the Supporting Information, where these formulas are shown to give a very good approximation for the calculation of FDR even when the fraction of true positives in the total population is as high as 50%.

RESULTS AND DISCUSSION

Modeling of the Distribution of False Positive Matches. In this work, the proteome in nuclear extracts from Jurkat cells was used as a model. Since one of the purposes of this work was to analyze how many proteins could be unambiguously identified with a limited amount of material, only 100 μg of protein extract was used; this amount was similar to that needed to run a single, conventional two-dimensional electrophoresis gel. Nuclear protein extracts were digested in solution, and the peptide pool was prefractionated off-line by cation exchange chromatography before LC−MS/MS analysis. A total of 14 peptide fractions were automatically collected and analyzed by ion trap MS/MS analysis using dynamic exclusion, obtaining ~40 000 MS/MS spectra.

To determine the significance of results obtained by database searching using SEQUEST, we first analyzed the distribution of false positive scores. For this purpose, the entire, unfiltered collection of MS/MS spectra were searched against an inverted database, constructed from human.fasta database by reversing the sequence of amino acids in each protein,[9] in order to maintain the amino acid frequency and peptide mass distributions, as well as the sequence homologies of the original database. Spectra were classified according to the assumed charge of the precursor ion, and the distribution of the Xcorr and ΔCn scores of the best matches against the inverted human.fasta database were analyzed. We assumed that all matches obtained against the inverted database were false; therefore, the scores obtained were used as a model to determine the score distribution due to random matching.

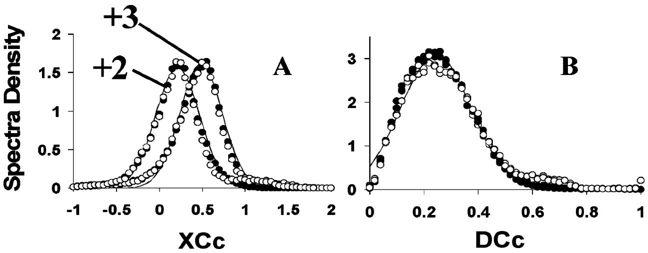

Figure 1. Distribution of corrected SEQUEST scores obtained by searching against human.fasta database. (A) XCc. (B) DCc. Search was performed against nonmodified (empty symbols) or inverted (filled symbols) human.fasta databases, assuming that the spectra were generated from doubly or triply charged precursor ions, as indicated by the numerals. Lines are computer-generated according to eqs 1 and 2, and the best-fit parameters from Table S-1.

We observed that Xcorr scores from spectra having either two or three charges showed a clear asymmetry, having an extended right tail (not shown). Since probability−confidence estimations were particularly critical around the right tail, we searched for models making a good fit to this region of the distribution. Although the Xcorr distribution could be satisfactorily modeled by the gamma function, as observed by others,[4] we noticed that the distribution of a corrected score, given by XCc = ln(Xcorr), for spectra having two or three charges, was symmetric and could be satisfactorily modeled by a Gaussian distribution (Figure 1A), given by

$$g(\mathrm{XCc}) = \frac{1}{\sqrt{2\pi}\sigma_x}\exp\left(-\frac{(\mathrm{XCc} - \mu_x)^2}{2\sigma_x^{\,2}}\right)$$
$$\mathrm{XCc} = \ln(\mathrm{Xcorr}) \qquad (1)$$

where the mean and standard deviations were dependent on the charge of the precursor ion (Figure 1A) and are shown in Table S-1 of the Supporting Information. Due to the resulting analytical simplicity (see below), we preferred to use the Gaussian formula to model the distribution of XCc. We expected that the slight deviation observed around the left tail in the subset of scores corresponding to charge 3+ would be irrelevant for the statistical analysis around the right tail. Similarly, the distribution of ΔCn scores could be described by an exponential function, but for analytical simplicity we preferred to use a corrected score given by $\mathrm{DCc} = \Delta\mathrm{Cn}^{1/2}$, which produced a distribution that could also be satisfactorily modeled by a Gaussian (Figure 1B), according to

[15] Benjamini, Y.; Hochberg, Y. *J. R. Stat. Soc. B* **1995**, *85*, 289−300.

[16] Storey, J. D.; Tibshirani, R. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 9440−9445.

$$g(\mathrm{DCc}) = \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{D}}}\exp\left(-\frac{(\mathrm{DCc}-\mu_{\mathrm{D}})^2}{2{\sigma_{\mathrm{D}}}^2}\right)$$

$$\mathrm{DCc} = \sqrt{\Delta\mathrm{Cn}} \qquad (2)$$

In this case, however, the distribution was independent of the assumed charge state of the precursor ion, so that all spectra could be grouped together and used to estimate the best-fit parameters of a unique distribution (Figure 1B). The best-fit parameters were calculated to optimally fit the right tails of the distributions.

We next analyzed the overall distribution of the two corrected parameters, with the implicit assumption that these two scores are independent of one another. As shown in Figure 2A, the 3D distribution had an almost symmetrical Gaussian shape, and the contour lines with identical spectral density were essentially elliptical. This behavior was consistent with a Gaussian distribution with two independent variables:

$$G(\mathrm{XCc}, \mathrm{DCc}) = \frac{1}{2\pi\sigma_x\sigma_{\mathrm{D}}}\exp\left[\frac{(\mathrm{XCc}-\mu_x)^2}{2{\sigma_x}^2}\right]\exp\left[-\frac{(\mathrm{DCc}-\mu_{\mathrm{D}})^2}{2{\sigma_{\mathrm{D}}}^2}\right] \qquad (3)$$

Figure 2B shows a scatterplot of all score pairs, together with some contour lines having identical spectra density according to eq 3; as shown, the contour lines predicted from the model agree satisfactorily with the distribution of the score pairs, delimiting regions of gradually decreasing point density.

Statistical Significance and Testing of the Model. Assuming that eq 3 gives an adequate description of the score distribution of matches against the inverted database and that all matches scored against these database are random, this equation may be used to derive statistical confidence thresholds. All (XCc, DCc) score pairs located in the same elliptical contour line verify the condition

$$t^2 = \frac{(\mathrm{XCc}-\mu_x)^2}{{\sigma_x}^2} + \frac{(\mathrm{DCc}-\mu_{\mathrm{D}})^2}{{\sigma_{\mathrm{D}}}^2} \qquad (4)$$

Therefore, *t* may be considered as an integrated, standardized SEQUEST score. Since the pairs having the same *t*-score have the same spectral density, they are expected to be equiprobable. The contour lines corresponding to different *t* values may be seen in Figure 2B. The significance threshold *p*, that is, the probability that a peptide match is randomly assigned a score equal or less probable than (XCc, DCc) is given by

$$p = \int\int_{(\mathrm{Xcc},\mathrm{DCc})\to\infty} G(\mathrm{XCc},\mathrm{DCc})\ \mathrm{dXCc}\ \mathrm{dDCc} \qquad (5)$$

Applying the transformation in eq 4 and integrating from the elliptical contour line, as defined by *t*, we obtain

$$p = \int_t^\infty (2\pi t)\,\frac{\exp(-t^2/2)}{2\pi}\,\mathrm{d}t = \exp(-t^2/2) \qquad (6)$$

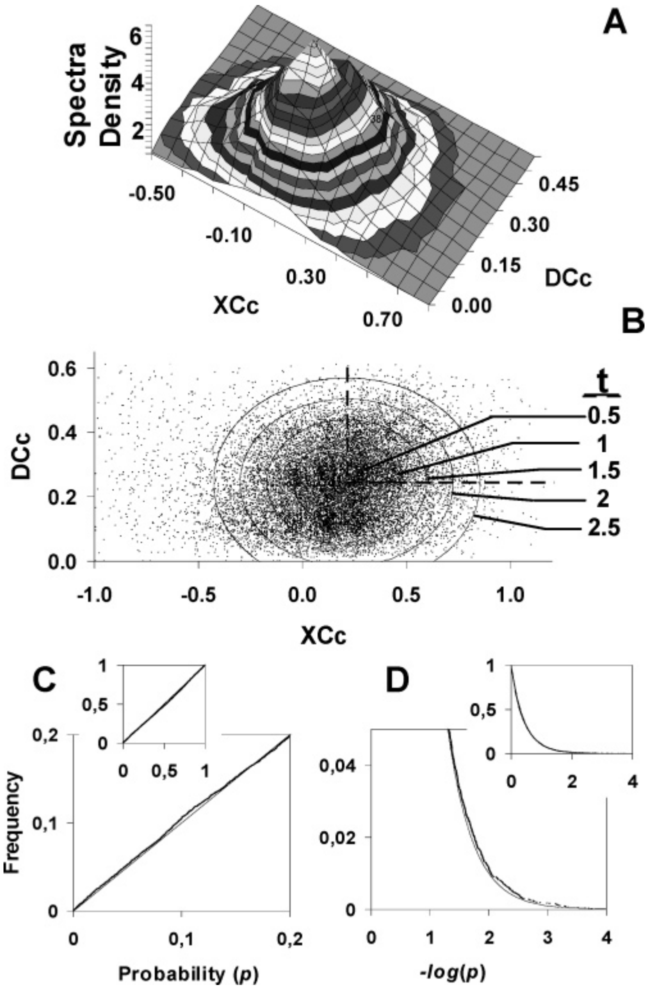

Figure 2. Modeling of corrected SEQUEST scores obtained by searching against inverted human.fasta database from doubly charged ions by a two-variable Gaussian model. (A) 3-D histogram showing the frequency distribution of scores. Regions having the same spectral density are indicated by different gray shades. (B) 2-D scatterplot of scores. Equiprobable elliptical contour lines are drawn according to its *t*-value, as indicated by the numerals, and eqs 3 and 4, using the same parameters as in Figure 1. Broken lines delimitate the upper right or “positive quadrant”. (C, D) Dependence of the normalized cumulative frequency of scores as a function of the probability predicted by the model, according to eqs 4 and 6. Frequency was plotted against probability, *p*, in (C) or against -log(*p*) in (D). Insets show the same graphs at a different scale. The observed values (thick lines) are compared with the values expected for an exact probability model (thin lines).

Equations 4 and 6 allow a very simple calculation of the *p*-value associated to any (XCc, DCc) score pair. The simplicity of eq 6 is the reason for which we have chosen the transformations of Xcorr and DCc that generate Gaussian distributions.

To analyze the validity of the statistical model, the *t*-score was calculated from each of the (XCc, DCc) score pairs obtained from the search of the data against the inverted human.fasta database using eq 4. The *p*-value was then calculated from the *t*-score using eq 6. The matches were then filtered according to predefined *p*-confidence levels, and the proportion of matches having *p*-values below the threshold were evaluated. Since this procedure is a one-tailed test with respect to the two variables, that is, any score pair having a *XCc* or a *DCc* value below the corresponding mean, μ_X and μ_D, respectively, must be considered as negative or purely

Table 1. Predictions According to the Statistical Model. Search against Inverted human.fasta Database[a]

threshold (p)	observed	expected false	observed (%)	expected true	FDR (%)
0,75	5816	6396	73%	0	100
0,5	3970	3978	50%	0	100
0,25	1941	2002	24%	0	100
0,1	835	790	11%	45	95
0,01	91	79	1,1%	12	87
0,001	13	8	0,2%	5	61
0,0001	0	0,8	0%	0	
0,00001	0	0,1	0%	0	

[a] The total number of spectra in the positive quadrant was 7.948. Calculations were performed as described in the Experimental Section.

random match, the analysis was restricted to the score pairs belonging to the "positive quadrant" in the XCc versus DCc plot (see Figure 2B). This restriction also diminished 4-fold the fraction of expected false positives. As summarized in Table 1, the proportion of spectra passing the p-threshold was essentially identical to that expected for a wide range of p-values. For instance, 11 and 1.1% of the spectra gave a match at the $p < 0.1$ and $p < 0.01$ confidence levels, respectively, as expected. In a further analysis, the fraction of predicted false positives having a score equal or better was plotted against the p-value for all spectra in the positive quadrant; as shown in Figure 2C, the predictions of the model (thick lines) were in perfect agreement with the expected values (thin lines) along the entire range of p-thresholds. The model was also very accurate in the range of very low p-values, which are particularly critical for confidence estimations, as shown in the semilog plots in Figure 2D. Taken together, these results demonstrate that, despite its mathematical simplicity, the statistical model predicted with very good accuracy the random matching distribution of the two SEQUEST scores.

Application of the Model to the Analysis of the Nuclear Proteome. The entire collection of spectra from the analysis of the nuclear proteome from Jurkat cells was then searched against the human.fasta database. The t-score and statistical significances were obtained as explained above, and the matches were filtered according to the same statistical criteria. The distributions of scores, shown in Figure 1, are essentially similar to those obtained with the inverted database, except around the right tail, where a population of spectra, corresponding to the correct matches, is observed. In this common situation, where the majority of spectra produced random matches, we found it possible to fit the Gaussian distributions directly to the score distributions obtained using the noninverted database, obtaining essentially the same parameters, since the mean and the standard deviation are deduced from the bulk of spectral data and are not much influenced by the distribution of correct matches. In Figure 3, a plot of the score pairs for precursors having two and three charges is shown together with several p-confidence threshold lines; although the distribution of the vast majority of score pairs follow the same trend as that observed using the inverted human.fasta database (compare Figure 2B with Figure 3B), the scores that do not follow the expected random behavior, corresponding to the correct matches, fall clearly beyond the limits established by the confidence levels.

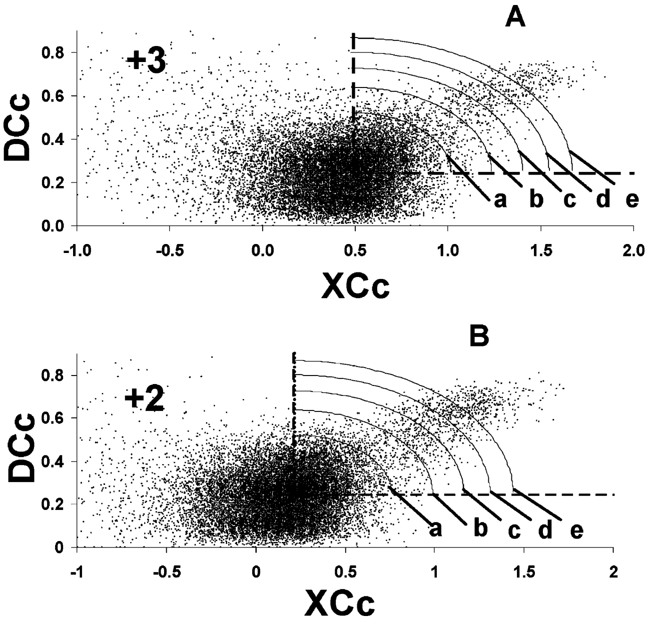

Figure 3. Discrimination among true and false peptide matches according to predictions of the probability model. Shown are 2-D scatterplots of scores obtained by searching against human.fasta database, from doubly (A) or triply charged precursor ions (B). The threshold probability lines are drawn according to eqs 4 and 6 and (a) $p = 0.1$, (b) $p = 0.01$, (c) $p = 0.001$, (d) $p = 0.0001$, and (e) $p = 0.000\,01$. Broken lines delimit the allowed region in the first quadrant.

Table 2. Predictions According to the Statistical Model. Search against Inverted human.fasta Database[a]

threshold (p)	obsd	expected false	obsd (%)	expected true	FDR (%)	expected true[b]
0,5	4988	3949	56%	1039	79	1749
0,25	3240	1899	36%	1341	59	1734
0,1	2246	743	25%	1503	33	1697
0,01	1454	76	16%	1378	5	1476
0,001	1059	8	12%	1051	0,7	969
0,0001	751	0,8	8%	750	0,11	397
0,00001	475	0,1	5%	475	0,02	137

[a] The total number of spectra in the positive quadrant was 8,937. [b] Determined by using PeptideProphet.[4]

As shown in Table 2, 1.454 peptide matches scored with $p < 0.01$ against the human.fasta database. At the individual peptide level and according to our random score distribution model, this statistical significance could have been considered as a good indicator that these sequences are likely correct, since the probability that they are due to random matching is lower than 1 in 100. However, these results are the consequence of the analysis of more than 30.000 spectra, and from this large collection of cases, 76 of the positives are expected to be due to random matching (Table 2); this means that ~5% of the matches are expected to be false positives (Table 2, FDR column). The FDR drops to 0.7 and 0.11% at the $p < 0.001$ and $p < 0.0001$ confidence levels, respectively. Since less than one false positive was expected from a total of 751 peptide sequences at the $p < 0.0001$ confidence level, we concluded that essentially all these peptide matches could be considered true positives, while the 308 additional peptide matches with p between 0.0001 and 0.001 belonged to a group of likely

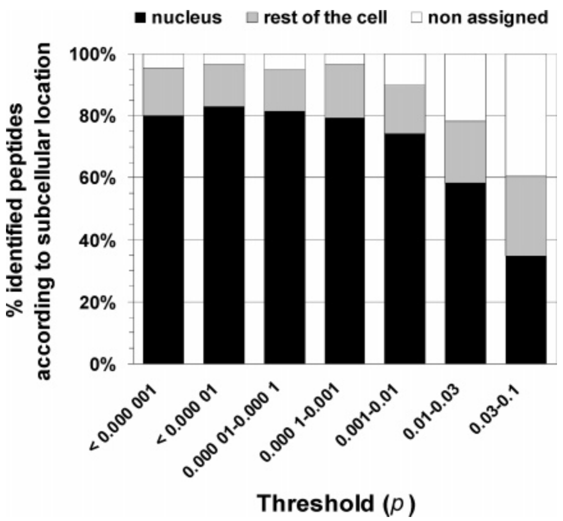

Figure 4. Classification of peptide matches according to the subcellular distribution of the corresponding proteins, as a function of the discriminant *p*-threshold value. The proportion of peptides is plotted as a percentage.

positives among which 8 are expected to be false, corresponding to a 3% fraction of false positives. From the collection of 751 peptide matches obtained with a false discovery rate of 0.11%, we could identify 260 proteins in the nuclear proteome (Table S-3 of the Supporting Information); while at a FDR of 0.7%, 320 proteins were identified.

Our results suggest that only 10% of the fragmentations yielding a score in the positive quadrant resulted in positive peptide identification. Increasing the amount of starting protein material would have increased the rate of "useful" fragmentations, and therefore, we could have used a less restrictive value of the confidence level p to obtain the same false discovery rate. Similarly, if the amount of protein were more limited, we would have needed to apply a lower confidence level to obtain the same fraction of false positives. For these reasons, the criterion used to ascertain whether a peptide sequence match is correct should not be based on the match score alone, but should take into account the total number of spectra analyzed and the actual rate of fragmentations yielding peptide information. In this context, the FDR gives a more appropriate statistical significance than the p-threshold, as has been observed in genome-wide studies.[15,16]

Analysis of the subcellular distribution of the identified proteins (Table S-3) allowed us to classify the identified peptides according to the subcellular distribution of the proteins from which they were derived. As shown in Figure 4, ~80% of the peptides with $p < 0.001$ corresponded to nuclear proteins; the fraction of nuclear peptides remained invariable when p was lowered to 10^{-6}. Since at these confidence levels all the peptides were expected to correspond to true positives, we assumed that the actual fraction of nuclear proteins in the sample is ~80% and that the remaining proteins were contaminants present in the preparation; this is consistent with the expected purity of the nuclear sample preparation. The fraction of nuclear peptides started to decrease when the threshold level was increased above 0.001, due to the inclusion of a significant fraction of incorrectly assigned peptides, in good agreement with the expected false discovery rates estimated in Table 2.

Table 3. Predictions According to Several Criteria

	observed			
criterion[a]	inverted human.fasta	human.fasta	expected true	FDR (%)
A	102	1164	1062	8,8
B	1493	2565	1072	58,2
C	2	592	590	0,3
D	49	1026	977	4,8

[a] (A) ΔCn > 0,1; Xc > 2,2 (2+), Xc > 3,75 (3+). (B) ΔCn > 0,08; Xc > 1,5 (2+), Xc > 3,3 (3+). (C) ΔCn > 0,08; Xc > 3,0 (2+), Xc > 4,0 (3+). (D) ΔCn > 0,08; Xc > 2,5 (2+), Xc > 3,5 (3+).

The results obtained by using our method were compared with those obtained by using other statistical approaches based on the analysis of Xcorr and ΔCn. As shown in Table 2, rightmost column, our method yielded results similar to those obtained by using the method of Keller et al. (PeptideProphet),[4] although some differences were appreciable. While PeptideProphet allowed the identification of a greater number of peptides above a FDR of ~1%, our method performed clearly better at lower FDRs; for instance, at a FDR of 0.1%, the number of positives identified by our method was almost 2-fold greater (Table 2). This comparative behavior was not related to the number of spectra analyzed but was consistently observed when the two methods were applied to the analysis of other proteomes (see Figure S-3 in Supporting Information).

In Table 3, we describe the results we would have obtained in the same experiment using several conventional filtering criteria. The number of observed and expected peptide sequences were determined from the searches against human.fasta and inverted human.fasta databases, respectively. The expected number of correct peptide matches was determined as the difference between the observed and expected number of spectra. As shown, some scores produced a higher number of correct sequences, at the expense of a higher FDR, whereas other filtering criteria were more conservative. In all cases, these criteria generated a lower number of expected positives than predicted by the model exposed here at identical false discovery rates, indicating that the sensitivity of this model is superior.

Since this work concentrated on the analysis of the two most relevant SEQUEST parameters, the number of peptides matching the same protein was not included in the identification criteria. Calculating the probability associated with the multiple peptide assignments on the same proteins would increase the number of identified peptides and the statistical confidence associated with each protein (the "protein probability") and would allow the identification of proteins by several peptides that did not score individually with enough confidence. In a preliminary report, we have demonstrated that high confident protein and peptide identification are sometimes possible on the basis of this factor alone.[17]

Effect of Database Size and Precursor Mass Window. To analyze whether the distribution of scores depended on the size of the database used, we first repeated the search against the inverted equine.fasta database using different precursor mass

[17] Rudd, C. J.; Martín-Maroto, F.; Scigelova, M.; Huhmer, A.; Biringer, R.; Vázquez, J. *J. Am. Soc. Mass Spectrom.* **2003**, (Suppl.) 14, 35S.

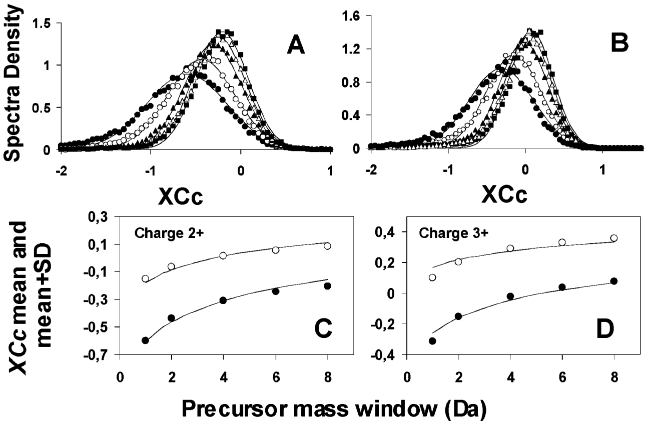

Figure 5. Dependence of XCc distribution parameters on the precursor mass window used for database search. (A, B) Distribution of XCc from doubly (A) or triply charged spectra (B) obtained by searching against inverted equine.fasta database using a precursor window of 1 (filled circles), 2 (empty circles), 4 (filled triangles), 6 (empty triangles), and 8 Da (filled squares). Curves are computed-generated according to eq 1 and the best-fit parameters from Table S-1. (C, D) Dependence of distribution parameters for doubly (C) and triply charged ions (D) on the precursor mass window: mean (filled circles), mean + standard deviation (empty circles). Curves are computer-generated according to eq 7, taking as a reference the parameters obtained for doubly charged ions using a 4-Da precursor mass window. In the case of triply charged ions (D), the amount of information was assumed to be 3-fold the information used for doubly charged ions.

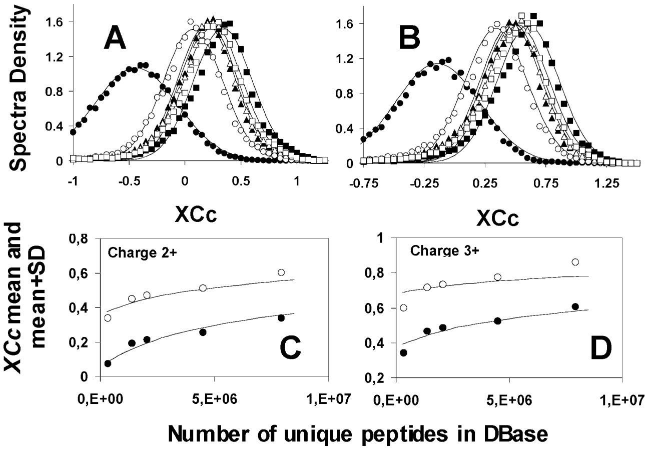

Figure 6. Dependence of XCc distribution parameters on database size. (A, B) Distribution of XCc from doubly (A) or triply charged spectra (B) obtained by searching against inverted databases of different sizes: equine.fasta (filled circles), yeast.fasta (empty circles), rat.fasta (filled triangles), human.fasta (empty triangles), swissprot.fasta (empty squares), and nr.fasta (filled squares). Curves are computed-generated according to eq 1 and the best-fit parameters from Table S-2. (C, D) Dependence of distribution parameters for doubly (C) and triply charged ions (D) on the number of unique peptides contained in each database: mean (filled circles), mean + standard deviation (empty circles). Curves are computer-generated according to eq 7, taking as a reference the parameters obtained for doubly charged ions and the human.fasta database. In the case of triply charged ions (D), the amount of information was assumed to be 4.5-fold the amount of information used for doubly charged ions.

windows. We hypothesized that increasing the mass window would increase the number of peptides available for random matching, thus increasing the average expected score. For this purpose, and since peptides are known to distribute in clusters separated by ~1 Da, the mass window was set to increase in integer steps, so that the number of peptides would increase proportionally. Figure 5A and B and Table S-1 confirmed that our assumption was correct; the mean value of XCc increased gradually with the precursor mass window.

We then tried to analyze how the statistical parameters could be influenced by the number of peptides in the search. We rationalized that if $p(x)$ is the probability that a given spectrum and a given random sequence match with score equal or better than x, the probability that the same or better score is obtained at least once when the experiment is repeated N times is given by $p_N(x) = 1 - (1 - p(x))^N$. Therefore, the relation existing among the probabilities $p_N(x)$ and $p_M(x)$ that we would obtain when the experiment is repeated N or M times, respectively, depends only on the ratio M/N, according to

$$p_M(x) = 1 - (1 - p_N(x))^{M/N} \quad \text{if } p_N(x) \ll 1 \text{ and } p_N(x)\frac{M}{N} \ll 1 \quad (7)$$

where M/N is the relative amount of different and independent candidate peptide sequences used in one experiment in relation with the other. As indicated, this equation is only valid when both $p_N(x)$ and $p_N(x)M/N$ are much lower than the unity. We then analyzed whether eq 7 could be used to relate the probabilities of getting a score x when the database search is performed in two different conditions, if we know the relative proportion of candidate sequences available for random matching in the two cases. Assuming that the mass window is proportional to the number of peptides contained within the window, we found that the distribution of XCc scores obtained using doubly charged spectra could be satisfactorily predicted as a function of the precursor mass window, from the statistical parameters obtained using a mass window of 4 Da (Figure 5C). Interestingly, the statistical parameters obtained using triply charged spectra could also be derived from those obtained from doubly charged spectra, assuming that the number of peptides increased exactly 3-fold at each mass window (Figure 5D).

The entire population of spectra was then searched against inverted databases of different sizes. In this analysis, the actual database size was expressed as the number of unique peptides contained in the database. As shown in Figure 6A and B and Table S-2, the distribution of XCc scores was gradually displaced toward higher values as the size of the database increased. We also observed that eq 7 could be used to make a satisfactory prediction of the shift of the distributions from the parameters of inverted human.fasta database, assuming that the number of candidate peptide sequences available for random matching was proportional to the total number of unique peptides contained in each of the databases (Figure 6C). In this analysis, we did not attempt to estimate the parameters corresponding to the equine.fasta database, since the conditions of validity of eq 7 were not fulfilled in this case. Using the same trend, we could even predict the shift on distribution parameters corresponding to triply charged precursors from those obtained with the doubly charged precursors and inverted human.fasta database, assuming that the number of available peptide sequences increased 4.5-fold (Figure 6D).

Taken together, all these observations supported the notion that a direct relationship exists among the peptide probability due to random matching and the number of candidate peptide sequences made available for the peptide search. Although this finding deserves a more detailed mathematical study, which is clearly beyond the scope of this paper, from all these results we can conclude that any change in the search conditions that affect the number of peptides against which the spectra are searched is expected to alter the statistical distribution parameters. This observation leads to the important conclusion that the statistical analysis cannot be performed using a unique set of distribution parameters, but must be performed taking into account the parameters from the particular score distribution obtained in each case. Since the absolute Xcorr scores of the correct sequences do not change in the different databases (provided that they contain the correct peptide), searching against a large database yields more conservative confidence values; while a smaller database may be more sensitive to identify peptides that do not yield a sufficiently high score. Whether these confidence values are correct depends on the experimenter's prior knowledge and the costs associated with type I and type II errors.

During the course of the analysis of other proteomes, we have also observed that the score distributions are affected by other parameters such as mass range, ionization conditions, and even the method used to generate the dta files used by SEQUEST (not shown). We have found that the distribution of random SEQUEST scores, in practice, must be determined for each particular experiment. Although in this work, for mathematical simplicity, we have tried to keep the number of possible factors at a minimum, and have only considered the charge of the precursor ions, we have observed that other factors such as peptide length or mass range may also have an appreciable effect in the random distribution of SEQUEST parameters (not shown). Therefore, a rigorous mathematical analysis of probabilities would require the division of the spectra into subsets, according to these factors, and the analysis of the random matching distribution in each of these categories. This procedure would increase the method sensitivity at the expense of increasing the complexity of analysis.

CONCLUSIONS

In this report, we present a statistical method to evaluate the results obtained after database search of fragmentation spectra when SEQUEST is applied to large-scale identification of proteins. In comparison with other published methods, our statistical approach poses the advantage of being based in an extremely simple formulation, which does not come at the cost of decreased accuracy, particularly when low false discovery rates are needed. Our method can be applied to searches against databases of any size or performed using different searching parameters. The behavior of the scores is calculated using normal distributions, which are straightforwardly described by their mean and standard deviation. In addition, the assumed independent random behavior of XCc and DCc allows their combined distribution to be described by a two-dimensional Gaussian function; this allows an exact integration of the probability, which can be determined, even with a hand calculator, using a simple formula. Despite its simplicity, this method makes a very accurate estimation of the random matching probability. The conceptual simplicity and mathematical accuracy of the approach makes the model particularly attractive when dealing with large amounts of fragmentation spectra, allowing a high degree of automation.

Our results illustrate how the database searching conditions, and not only the database size, affect critically the random score distribution and, hence, the probabilities associated with all the spectra in the data set. We conclude that the criteria based on predefined rules using SEQUEST scores for the identification of proteins are not universally valid but need to be established in each particular experiment.

ACKNOWLEDGMENT

This work was supported by Grants SAF2000-0178, CICYT BIO2003-01805, and CAM 08.5/0065.1/2001 and by an institutional grant by Fundación Ramón Areces to CBMSO. M.V. is recipient of a postdoctoral fellowship from Comunidad Autónoma de Madrid.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review May 11, 2004. Accepted September 16, 2004.

AC049305C