technical notes Journal of proteome research

Statistical Calibration of the SEQUEST XCorr Function

Aaron A. Klammer,[†] Christopher Y. Park,[†] and William Stafford Noble*[,†,‡]

Department of Genome Sciences, University of Washington, Seattle, Washington 98195, Department of Computer Science and Engineering, University of Washington, Seattle, Washington 98195

Received September 4, 2008

Abstract: Obtaining accurate peptide identifications from shotgun proteomics liquid chromatography tandem mass spectrometry (LC-MS/MS) experiments requires a score function that consistently ranks correct peptide−spectrum matches (PSMs) above incorrect matches. We have observed that, for the Sequest score function *Xcorr*, the inability to discriminate between correct and incorrect PSMs is due in part to spectrum-specific properties of the score distribution. In other words, some spectra score well regardless of which peptides they are scored against, and other spectra score well because they are scored against a large number of peptides. We describe a protocol for calibrating PSM score functions, and we demonstrate its application to *Xcorr* and the preliminary Sequest score function *Sp*. The protocol accounts for spectrum- and peptide-specific effects by calculating *p* values for each spectrum individually, using only that spectrum's score distribution. We demonstrate that these calculated *p* values are uniform under a null distribution and therefore accurately measure significance. These *p* values can be used to estimate the false discovery rate, therefore, eliminating the need for an extra search against a decoy database. In addition, we show that the *p* values are better calibrated than their underlying scores; consequently, when ranking top-scoring PSMs from multiple spectra, *p* values are better at discriminating between correct and incorrect PSMs. The calibration protocol is generally applicable to any PSM score function for which an appopriate parametric family can be identified.

Keywords: calibration • database search • peptide identification • tandem mass spectrometry

1. Introduction

At their core, database search methods for identifying proteins from shotgun proteomics data rely upon a score function that evaluates matches between peptides and spectra. Algorithms such as Sequest[1] first identify a set of candidate peptides whose m/z values are close to the measured m/z of the query spectrum and then select the candidate peptide that maximizes this score function with respect to the query. In this context, an optimal scoring function assigns the highest score to the peptide that actually generated the spectrum.

However, identifying the best peptide for each spectrum only solves half of the peptide identification problem. Once the best-scoring peptide has been identified for every spectrum in a given data set, the mass spectrometrist must determine which of the resulting peptide−spectrum matches (PSMs) are correct. In practice, only 5−30% of the PSMs in a given data set are correct,[2] so this latter phase is very important.

The simplest approach to separating correct from incorrect PSMs is to rank them according to the same scoring function that was used in the initial search. A perfect PSM scoring function will rank all correct PSMs above all incorrect PSMs. Real scoring functions, of course, are imperfect and generally fail to achieve good separation. A common method to improve the separation is to treat spectra with different properties separately, for example, setting different score thresholds for spectra of different charge states.[2] But spectra differ by more than just charge state. A more sophisticated approach would correct for all effects that a particular spectrum has on its corresponding score distribution.

For example, a simple ranking by the SEQUEST score function *Xcorr* ignores two important effects, which are illustrated in Figure 1. The figure shows that the distribution of maximal *Xcorr* scores depends strongly upon (A) the properties of the spectrum and (B) the number of candidate peptides that the spectrum is searched against. Therefore, PSM scores generated from two different spectra or the same spectrum under two different search conditions might not be directly comparable to one another.

In this work, we describe a protocol for correcting for the effects shown in Figure 1. The resulting calibrated scores have well-defined semantics and yield improved discrimination between correct and incorrect PSMs. Our method uses the distribution of PSM scores for a particular spectrum to calculate a conditional *p* value for that score, $P(S > S_{\text{thresh}}|H_0, \text{spectrum})$, or the probability that that we would receive a score S greater than or equal to a threshold score S_{thresh} given the null hypothesis H_0 *and* a particular spectrum.

To explain our method, we first need to describe three kinds of mass spectrometry score distributions:

(1) the distribution of scores for *all* candidate peptides for a particular spectrum under particular search conditions;
(2) the distribution of scores for the *maximum* scoring peptide for a particular spectrum under particular search conditions, a form of extreme value distribution; and

* To whom correspondence should be addressed. William Noble, University of Washington, Box 357730, Seattle, WA 98195. E-mail: noble@gs.washington.edu. Phone: 206 543-8930. Fax: 206 685-7301.
[†] Department of Genome Sciences, University of Washington.
[‡] Department of Computer Science and Engineering, University of Washington.

10.1021/pr8011107 CCC: $40.75
Published on Web 03/10/2009

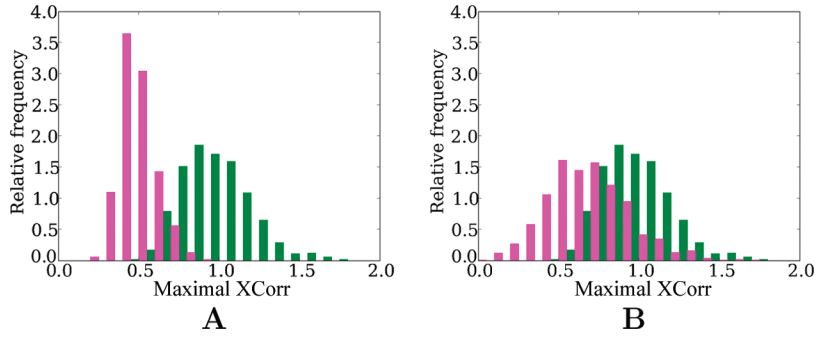

Figure 1. Dependence of score distributions on the query spectrum. (A) Distributions of the maximal *Xcorr* scores produced by searching two different spectra against 100 candidate peptides from 1000 randomly shuffled decoy databases. (B) Distributions of the maximal *Xcorr* scores produced by searching a single spectrum against 100 (green) or 10 (magenta) candidate peptides from 1000 randomly shuffled decoy databases. The histograms are normalized so that the area under them is equal to one.

(3) the distribution of the *maximum* scores for the maximum scoring peptides for a *group* of spectra under particular search conditions.

Our method for calculating *p* values corrects for spectrum-specific variation in (1) and (2) and is outlined here. First, for a given spectrum, we compute a score (e.g., *Xcorr*) for all candidate peptides, and we use the scores to learn the parameters of a parametric score distribution (i.e., we fit to distribution (1)). Second, we use the parameters of this score distribution to compute a *p* value for the maximal score (i.e., we calculate a *p* value for the score under distribution (2)), correcting for the difference in expected score due to variation in the number of candidate peptides the spectrum is compared to. In an optional third step, these *p* values can then be used as input to standard multiple-hypothesis testing calculations (e.g., see ref 3) to correct for the effects of distribution (3).

This method differs from other published probability calculations for tandem mass spectrometry scores, which either calculate a *p* value without conditioning on the spectrum (i.e., $P(S > S_{\text{thresh}}|H_0)$)[4–8] or calculate a probability score that is not a *p* value.[9–11] These methods therefore do not control for the effects shown in Figure 1.

A handful of methods do use spectrum-specific score distributions, either to compare different score functions or to calculate an *E*-value. For example, OMSSA[12] uses the number of candidate peptides to convert from *p* values (score (1) above) to *E*-values. The method outlined in ref 13 uses a survival function for (1) to calculate an *E*-value for a score taken from (2). We do not claim any theoretical superiority for our method relative to that of Fenyo et al.,[13] rather, our method yields *p* values with desirable statistical properties that, as demonstrated in section 3.1, the Fenyo et al.[13] method lacks. Recently, Alves et al.[14] described a methodology for calibrating PSM scores with respect to a particular score function and laboratory. Unlike previously reported *E*-values, the *E*-values computed by Alves et al. incorporate all three of the effects mentioned above. Their method uses a collection of reference spectra, searched against a decoy database, to fit a calibration function for any given scoring scheme. In contrast, our approach involves fitting the full observed score distribution to a parametric distribution for each particular spectrum. Alves et al. noted the desirability of fitting to spectrum-specific distributions but were unable to do so because most search algorithms do not report the full score distribution. This obstacle is overcome for Sequest by our reimplementation of Xcorr in the software package Crux.[15]

In this work, we demonstrate how to compute *p* values for *Xcorr* and the preliminary Sequest score *Sp*. We show that these *p* values are valid; that is, the *p* values are uniform under the null hypothesis, and we show that the *p* values improve discrimination between correct and incorrect PSMs relative to their underlying scores. Contingent upon identifying an appropriate parametric family of distributions, the proposed statistical calibration protocol can be applied to any PSM score function.

1.1. Approach. Our primary goal is to calculate accurate *p* values for PSMs. We divide our approach into two stages, each controlling for one of the effects shown in Figure 1. The first stage controls for the difference in maximal score distribution due to the particular spectrum being searched (Figure 1A and distribution (1) from the Introduction) and is described in section 1.1.1. The second stage controls for the difference in score distribution due to the number of peptides being searched against (Figure 1B and distribution (2) from the Introduction) and is described in section 1.1.2.

1.1.1. Calculating *p* Values. Calculating *p* values from a distribution of scores can be approached in one of two ways: using a nonparametric method, where there are no assumptions made about the underlying distribution, or using a parametric method, where the distribution is assumed to follow a particular form. We are concerned with calculating *p* values for the most extreme peptide in a score distribution. Thus, we choose to estimate *p* values parametrically, because parametric methods can extrapolate to areas of a distribution (such as the tail) where there is little or no data.

We have found empirically that both the *Sp* and *Xcorr* score distributions for a particular spectrum can be approximated using a Weibull distribution.[16] The three-parameter Weibull distribution has the following form:

$$f(x;\beta,\eta,\mu) = \frac{\beta}{\eta}\left(\frac{x-\mu}{\eta}\right)^{\beta-1} e^{-[(x-\mu)/\eta]^{\beta}}$$

$$F(x;\beta,\eta,\mu) = 1 - e^{-\left(\frac{x-\mu}{\eta}\right)^{\beta}}$$

where $f(x;\beta,\eta,\mu)$ is the probability density function, $F(x;\beta,\eta,\mu)$ is the cumulative distribution function (CDF), β is the shape parameter, η is the scale parameter (analogous to the standard deviation of the normal distribution), and μ is the location parameter (analogous to the mean of the normal distribution). An example of a fit to an *Xcorr* distribution is seen in Figure 3. To avoid fitting scores from correct PSMs, we remove the maximal score before fitting the Weibull distribution. Unlike in sequence comparison, in mass spectrometry, sequence homology does not introduce a substantial number of high-

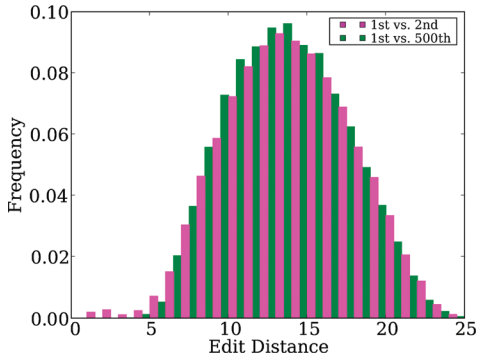

Figure 2. Distribution of edit distances between the top ranked peptide and the second (magenta) and 500th (green) ranked peptides There is little difference in the distributions, except for a small fraction of second ranked peptides of edit distance less than four (0.8%), of which a third are identical if leucine and isoleucine are treated as identical (41/153).

scoring matches that are neither completely correct nor completely incorrect (Figure 2).

The Weibull distribution has the advantage of being easy to fit to truncated data where one is only interested in modeling the behavior of the most extreme high-scoring region of the distribution. This distribution has an additional advantage in that it has a closed form CDF and thus can be readily used to calculate p values without resorting to precomputed tables.

In the case of *Sp* and *Xcorr*, we found the Weibull fit yielded uniform p values when the underlying distributions were truncated to include only a fraction of the top scoring PSMs (Figure 3B). We describe the method for selecting what fraction of the tail to fit for the two scores in section 3.3 and Figure 8. Once the distribution is fit, a p value for a particular score X can be readily calculated using the Weibull CDF, where $p(x) = 1 - F(x; \beta, \eta, \mu) = e^{-[(x-\mu)/\eta]^{\beta}}$. Spectra with fewer than 20 peptides in their mass window did not produce reasonable fits, so we do not attempt to fit score distributions for these spectra. These spectra are rare under realistic search conditions (Figure 4).

1.1.2. Extreme Value Distribution Correction. The p value calculated in section 1.1.1 is for a particular score taken from the score distribution from a *single* search (the distribution shown in Figure 3 and distribution (1) from the Introduction). But we want a p value for the maximum score taken from this distribution (the distribution shown in Figure 1 and distribution (2) from the Introduction). In other words, if we sample n times from a particular score distribution (corresponding to searching a spectrum against n peptides), we need to compute the probability that the maximum value among these n scores is greater than or equal to S. These maximum scores should follow an extreme value distribution (EVD), regardless of the distribution of the underlying scoring function. We can calculate *the form of this EVD using the Weibull distribution from section 1.1.1.

This idea is expressed below in more concrete terms. $X_1, X_2, \ldots, X_n$ are n independent and identically distributed continuous random variables drawn from a (known) probability density function f, with a corresponding cumulative distribution function F. Define

$$X^* = \max\{X_1, X_2, \cdots, X_n\}$$

We would like to calculate a p value for any particular value of $X^* = x$, assuming a fixed value of n. For this, we need to know

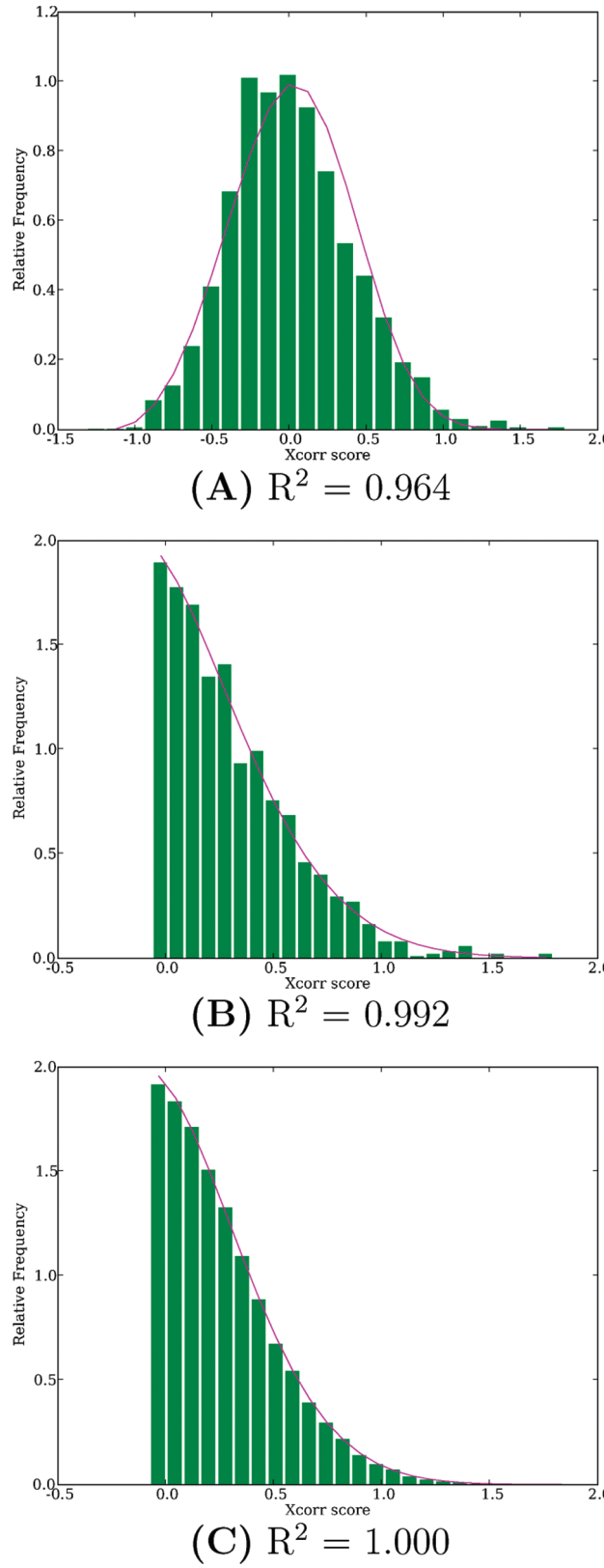

Figure 3. Weibull fits. The figure plots, for the same spectrum, a fit to the *Xcorr* score distribution for candidate peptides from a single search against a shuffled yeast protein sequence database using (A) all peptides from the distribution or (B) the top 0.55 fraction of peptides. The poor fit to the full distribution is improved when only the tail is fit. Also shown is a deeper (approximately 30-fold larger) sampling of the distribution (C).

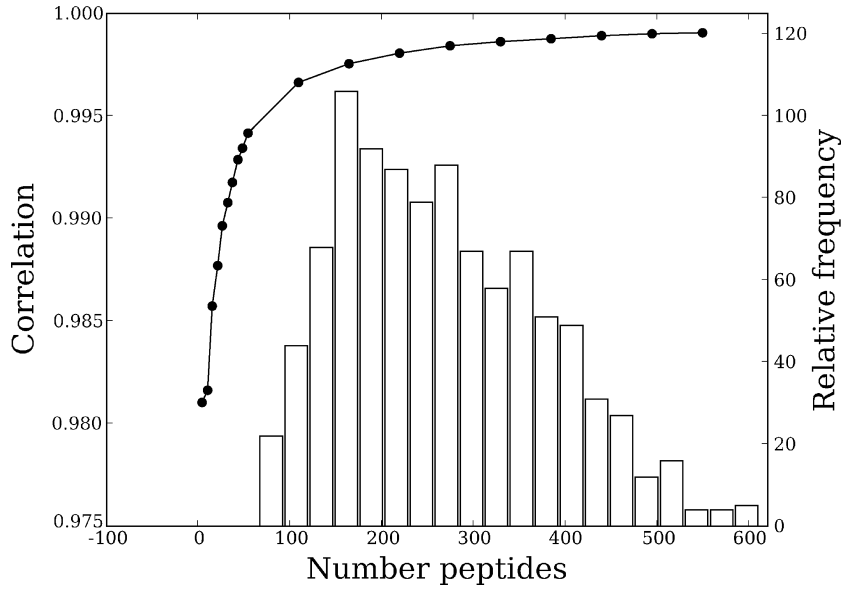

Figure 4. Weibull fit as a function of number of peptides fit. Shown are average correlation values across 1000 spectra (left axis) for the Weibull fits as a function of the number of fit peptides. Also shown is the distribution of the number of peptides (right axis) in a ± 3.0 *m*/*z* window for the same 1000 spectra.

the cumulative distribution function for X^*, F^*. We exploit the known cumulative distribution F to get the probability that any single X_i is less than x. The desired cumulative density is then

$$\begin{aligned} F^*(x) &= P(X_1 \leq x, X_2 \leq x, \cdots, X_n \leq x) \\ &= P(X_1 \leq x)P(X_2 \leq x)\cdots P(X_n \leq x) \\ &= F(x)^n \end{aligned}$$

The corresponding p value is just $1 - F(x)^n$. If $F(x)$ is close to 1, then we can approximate the above equation as follows. Defining $p(x) = 1 - F(x)$, and exploiting the rule $(1 - x)^n \approx 1 - nx$, we get

$$\begin{aligned} p^*(x) &= 1 - F(x)^n \\ &= 1 - (1 - p(x))^n \\ &\approx 1 - (1 - np(x)) \\ &\approx np(x) \\ &\approx ne^{-[(x-\mu)/\eta]^\beta} \end{aligned}$$

The last step above substitutes the formula for the Weibull p value given in section 1.1.1. In our experiments, we use either the exact or approximate formula, depending on the size of $p(x)$.

2. Materials and Methods

Supplementary data is available at http://noble.gs.washington.edu/proj/msms, and source code and binaries at http://noble.gs.washington.edu/proj/crux.

2.1. Mass Spectrometry. All tandem mass spectra used in this paper were generated as described previously[17] using a standard shotgun proteomics LC-MS/MS preparation protocol. Briefly, a complex yeast lysate was prepared by growing *Saccharomyces cerevisiae* strain S288c to an optical density of 1.2. The cells were lysed, centrifugated, and resuspended in 1% PPS (Protein Discovery, Knoxville, TN). The resulting protein mixture was reduced with dithiothreitol, alkylated with iodoacetic acid, digested to peptides with trypsin for 4 h, quenched by acidification with HCl, and finally centrifuged, with the resulting supernatant stored at −80 °C.

The peptide sample was analyzed by data-dependent tandem mass spectrometry using a 4 μm, 90 Å-pore size Jupiter Proteo reverse phase material (Phenomenex, Ventura, CA) packed into a 60 cm column, which was placed inline with an Agilent 1100 Binary HPLC and Autosampler (Palo Alto, CA). Peptides were eluted from the microcapillary columns with a 4-h organic gradient, and emitted into an LTQ mass spectrometer (ThermoFisher Scientific, San Jose, CA). Application of mass spectrometer scan functions and HPLC solvent gradients were controlled by the Xcalibur data system (ThermoFisher Scientific). Inorganic buffer was 95% water−5% acetonitrile−0.1% formic acid (buffer A), and organic buffer was 5% water−95% acetonitrile−0.1% formic acid (buffer B). The final data set contained 18 149 individual tandem mass spectra.

2.2. Database Search. Tandem mass spectra were searched with the sequence database search algorithm Sequest (v. 27), InsPecT (v. 20070523) and X!Tandem (v. 2007.07.01.2) as well as our own in-house software package known as Crux.[15] We have implemented in Crux near-exact replicas of the Sequest functions *Xcorr* and *Sp*. False discovery rates were estimated by searching against shuffled sequence databases, produced by taking as input a fasta file and randomly shuffling the sequences of each protein sequence individually, to maintain protein length distribution and amino acid composition. All searches were performed against the yeast protein sequence database downloaded on November 16, 2006. For all algorithms, we searched tryptic peptides with length between 7 and 50, and mass between 200 and 7200, allowing three charge states (1, + 2 and + 3) and one missed tryptic cleavage site.

The data for Figure 1 were generated by searching two spectra (scans 4769 and 14 246 searched as + 3 and + 1, respectively) against 1000 randomly shuffled databases using

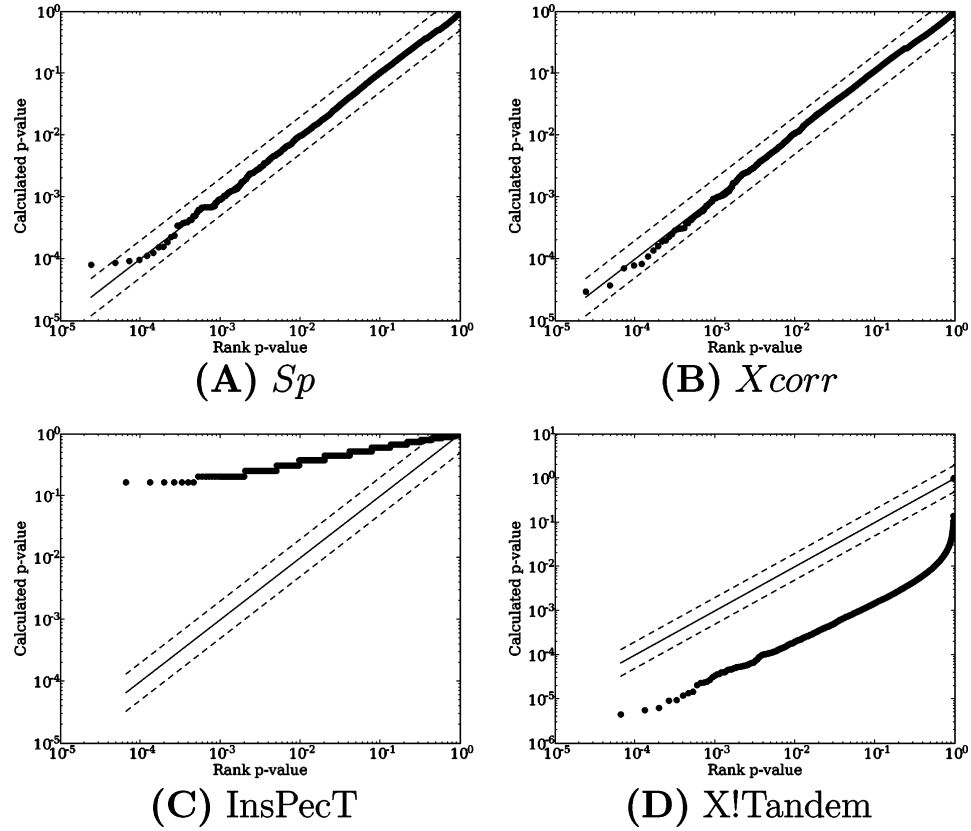

Figure 5. Calculated *p* values for null PSMs are uniformly distributed. The figure plots, for each of 15 149 spectra searched against a shuffled database, the computed *p* value as a function of the empirical (rank-based) *p* value, computed using (A) *Sp* and (B) *Xcorr*. For comparison, we show *p* values from the database search algorithm (C) InsPecT[5] and (D) X!Tandem.[18] The distribution of *p* values are compared against $y = x$ (solid black line) and $y = (x/2)$ and $y = 2x$ (dotted lines).

Crux and selecting the maximum *Xcorr* score for each of these searches. The shuffled databases were generated from 444 randomly selected proteins from the yeast protein sequence database to decrease analysis time.

3. Results

3.1. Uniformity of *p* Values. To evaluate our *p* value calculation method, we first test that *p* values for PSMs generated from the null distribution are uniformly distributed and therefore valid. We use a collection of 15 149 spectra from our data set and score each spectrum against a shuffled version of the yeast protein sequence database using *Xcorr* and *Sp*. The resulting PSMs are, presumably, all incorrect. For each scored PSM, a *p* value is calculated as described in section 1.1 for *Sp* or *Xcorr*. These 15 149 *p* values are then sorted in ascending order and plotted in a log−log plot against the empirical *p* value implied by their position in the sorted list. For perfectly calculated *p* values, the first value in a sorted list of 1000 spectra would be 0.001, the second 0.002, and so on up to 1.000, tracing a line of $y = x$. Deviation from this line indicates deviation from perfectly calculated *p* values.

Figure 5 plots our calculated *p* values as a function of the empirically determined rank-based *p* value for *Xcorr* (panel A) and *Sp* (panel B). The plots demonstrate that the calculated *p* values are accurate to within a factor of 2.

For comparison, Figure 5C,D show *p* values computed by the database search algorithms InsPecT and X!Tandem on the same data set. The *p* values for X!Tandem were calculated by dividing the *E*-value returned by the number of candidate peptides in the mass window searched. The *p* values from these algorithms are quite far from uniform: the smallest *p* value among 15 149 PSMs is 0.13 for InsPecT, which is more than 3 orders of magnitude larger than would be expected from a uniform model, and 4.0×10^{-6} for X!Tandem, more than an order of magnitude smaller that would be expected by chance. In approximately 600 cases, X!Tandem reports *E*-values that are larger than the number of peptides in the mass window; in these cases, we set the *p* value to 1.

We also show a similar plot but for a *single* spectrum searched against *multiple* (1000) shuffled databases (Figure 6), with similar results for the four algorithms.

3.2. Discrimination between Correct and Incorrect PSMs. To test the ability of our *p* values to discriminate between correct and incorrect PSMs, we implemented a database search algorithm that is modeled on Sequest but that incorporates our *p* value calculation. We then searched 15 149 spectra in the data set described in section 2.1 against a target database of the yeast protein sequence database and a decoy database comprised of shuffled sequences from the target database.

For each score, we compute a *q* value, which is defined as the minimum false discovery rate at which that particular score is called significant. The underlying FDR estimates are computed by using the decoy score distribution as an approximation of the null distribution.[19] Thus, we rank the target scores $S_1,\ldots, S_n$ and estimate the FDR associated with score S_i as In

$$\widehat{FDR}(S_i) = \frac{\#S_{decoy} \geq S_i}{\#S_{target} \geq S_i} \quad (1)$$

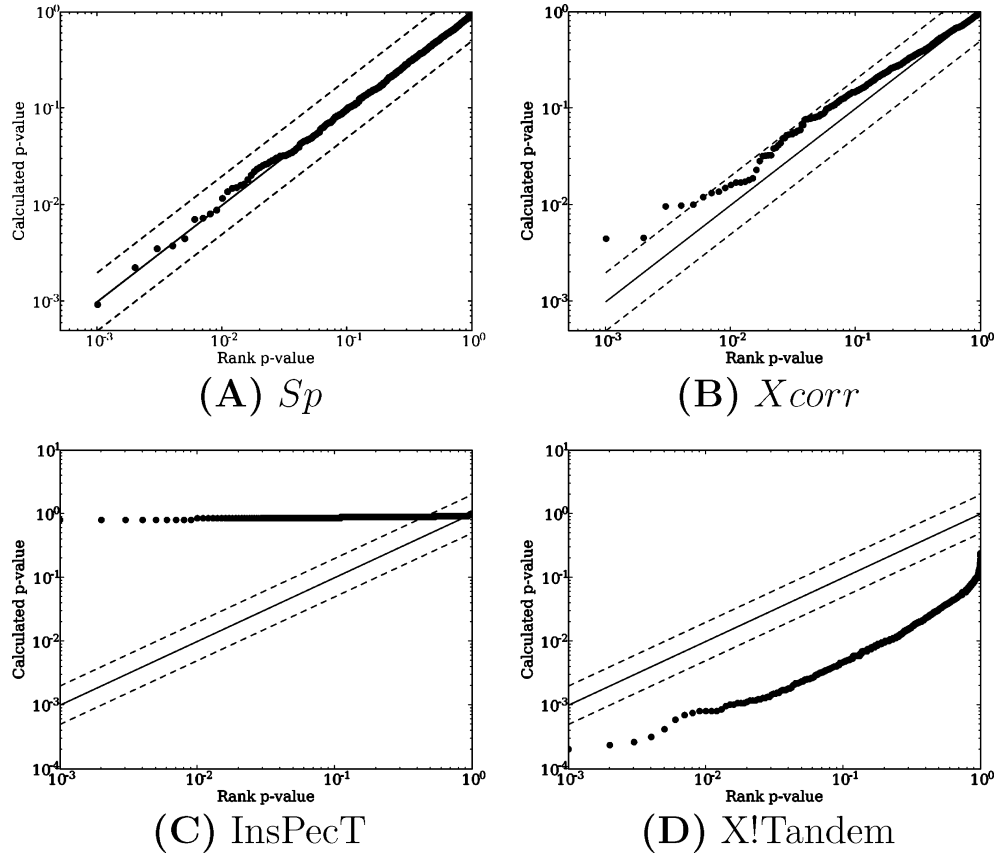

Figure 6. Calculated *p* values for null PSMs are uniformly distributed. The figure plots, for a *single* spectrum searched against 1000 shuffled databases, the computed *p* value as a function of the empirical (rank-based) *p* value, computed using (A) *Sp* and (B) *Xcorr*. Again, for comparison, we show *p* values from the database search algorithm (C) InsPecT[5] and (D) X!Tandem.[18]

addition, for *p* values calculated using our method, we can obtain a second FDR estimate using the method of Benjamini and Hochberg,[20] as follows:

$$P_\alpha = \max\left\{j : p_j \leq \frac{j}{m}\alpha\right\} \quad (2)$$

where P_α is the number of positives at a desired false discovery rate of α.

We can draw three conclusions from Figure 7. First, the close correspondence between the lines for the Sequest and Crux implementations of *Xcorr* and *Sp* suggest that our reimplementation of these scores is accurate. Second, the relative improvement when we switch from raw scores to *p* values shows the value of calibrating our scores on a per-spectrum basis. For example, at a *q* value threshold of 0.05, our *p* value calculation increases the number of identified PSMs for *Xcorr* and *Sp* by 31% and 112%, respectively. Finally, the relatively close correspondence between the two *q* value estimates with and without using the decoy database show that the extra step of searching a shuffled database to estimate false discovery rate can be eliminated, allowing a 2-fold speed improvement. Note that, although these two curves disagree slightly with one another, it is not obvious which is correct, because the decoy database may not be a perfect model of the null hypothesis.

3.3. Selection of Fraction of Distribution Tail to Fit. In section 1.1.1, we described how we calculate *p* values from the tail of the score distribution. Here, we describe our exact method for selecting how much of the tail to fit. We enumerated all distribution fractions between 0.1 and 1.0, incremented by 0.1, and we evaluated each fraction using two figures of merit. The first is the slope of the QQ plot shown in Figure 5; the closer the *p* value slope error is to unity, the more uniform the calculated *p* values are, and the better our fraction of fit peptides is considered to be.[21] The second is the number of positive PSMs at 5% false discovery rate; for this figure of merit, higher is better. We selected the best fraction separately for *Sp* and *Xcorr* using 1000 held out spectra (Figure 8), and validated this selection on two other held out sets of 1000 spectra (see Supporting Information). The held-out data sets were not used in subsequent analyses. We visually inspected the plots and selected tail fractions of 0.40 and 0.55, for *Sp* and *Xcorr*, respectively.

4. Discussion

We have demonstrated that accurate *p* values can be computed for the Sequest PSM score functions *Xcorr* and *Sp* by fitting a Weibull distribution to the observed score distribution on a spectrum-specific basis. By correcting for spectrum-dependent properties of these score functions, and by correcting for the number of candidate peptides within a given m/z window, our method significantly improves the discriminative power of the overall algorithm. Furthermore, by producing valid *p* values, the method eliminates the need to search each spectrum against a decoy database, thereby decreasing the overall running time by a factor of 2 when the target and decoy databases are of equal size.

In practice, database search procedures are often followed by a postprocessing procedure that ranks PSMs using additional

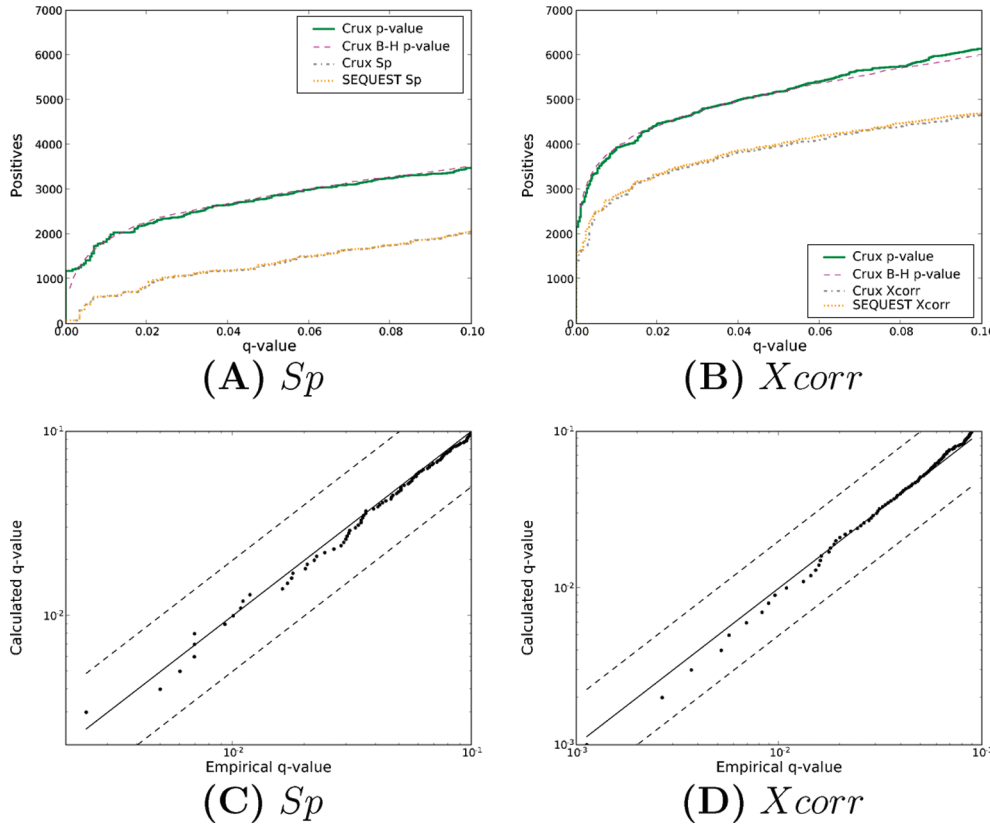

Figure 7. Discrimination between correct and incorrect PSMs for *p* values and their underlying score functions. The top two panels plot, for two different score functions (Sequest *Sp* (A) and *Xcorr* (B)), the number of positive peptide identifications as a function of *q* value (a measure of false discovery rate) before and after the *p* value calculation (dashed gray and yellow, and solid green lines, respectively). Also shown is the Benjamini-Hochberg false discovery rate estimate calculated *without* a shuffled database (dashed magenta line). Panels (C) and (D) explicitly compare the estimated *q* values computed using our protocol and using a decoy database. Each point in the figure corresponds to a fixed number of positive PSMs.

information. For example, PeptideProphet[4] and Percolotor[22] make use of multiple score functions (Xcorr, Sp and ΔCn), the difference between the observed and theoretical precursor mass, the assumed charge state, the length and number of tryptic termini of the peptide, as well as features of the entire collection of PSMs such as the number of times that the given peptide was observed. Not surprisingly, incorporating this additional information yields larger improvements in the overall number of PSMs at a fixed FDR than is offered by the *p* value computation procedures that we describe, which are only based on a single score. For example, for the data set shown in Figure 7, the *p* value calculation increases the number of positive PSMs at a 5% false discovery rate by 31% from 3963 to 5182, whereas PeptideProphet applied to the same data set yields an improvement of 109% (8297 PSMs). Therefore, a clear direction for future work is the incorporaton of our statistically calibrated scores into a postprocessing procedure such as PeptideProphet or Percolator. It is likely that the improvements offered by the

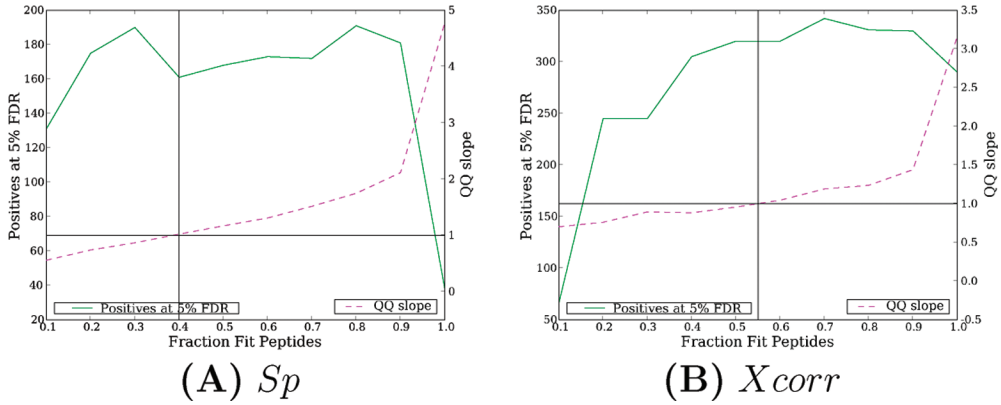

Figure 8. Evaluation of different fractions of the distribution tail to fit. Varying the portion of the peptide score distribution tail to fit affects the number of positive peptides at a 5% FDR (green lines) and the slope of the QQ plot (dashed magenta lines) for *Sp* (A) and *Xcorr* (B). A perfect QQ plot has a slope of 1.0 (horizontal black lines). After visual inspection, we selected a fraction of 0.40 for *Sp* and 0.55 for *Xcorr* (vertical black lines).

two types of methods would be complementary: the statistical calibration procedure takes into account properties of the spectrum-specific score distribution, whereas, with the exception of the ΔCn score, this type of distributional information is not used by PeptideProphet or Percolator.

The general calibration approach that we employ here has been used successfully in other areas of bioinformatics, most notably in sequence analysis. Some pairwise sequence comparison algorithms[23] and hidden Markov model-based scoring systems[24,25] estimate p values by fitting extreme value distributions to observed score distributions. In this work, we found it necessary to switch from the extreme value distribution to the more general Weibull distribution. In addition, the mass spectrometry setting of this problem is more complex than the sequence analysis setting, because we must separately consider the distribution of candidate peptide scores for a single spectrum and the distribution of maximal PSM scores across multiple spectra.

We chose to compute p values from *Xcorr* and *Sp* because of the popularity and good performance of the Sequest algorithm. Generalizing this work to other scoring functions may or may not be straightforward, depending upon whether a good parametric approximation can be found for the empirical score distribution. In searching for such a distribution, the Weibull distribution is a good place to start, both because of the logistical considerations mentioned in section 1.1.1, and because the shape parameter of the Weibull allows it to fit many empirical distributions well.

There is some evidence that the empirical score distributions are not perfectly Weibull. As illustrated in Figure 8, we achieve the most accurate p values and the best discrimination, respectively, by fitting different fractions of the tail of the empirical distribution. This observation implies that the results reported here could be improved even further if we could achieve a better fit to the observed score distributions.

Our procedure requires setting one hyperparameter, which is the fraction of the empirical distribution to which the Weibull is fit. We have shown in Figure 8 how to set this parameter using held-out data. In other tests, this parameter setting generalizes well (see Supporting Information), indicating that it will not usually be necessary to redo this parameter selection procedure.

The most fruitful extensions to this work would probably involve refining our estimates of false discovery rate. We use the method of Benjamini and Hochberg[20] to calculate FDRs from our p values; there are, however, other more accurate methods that take into account the fraction of target PSMs that are incorrect.[3,19] These methods would likely push the number of positives at fixed FDR even higher. Also, while our method of fitting the Weibull distribution is quite robust, it does fail on spectra with very few (less than 20) peptides. This problem could be addressed by generating additional decoy peptides on the fly if the candidate database contains too few peptides to properly fit the distribution.

Abbreviations: LC-MS/MS, liquid chromatography tandem mass spectrometry; PSM, peptide−spectrum match.

Acknowledgment. The authors thank the anonymous referees for numerous valuable suggestions. This work was funded by NIH awards R01 EB007057 and P41 RR11823.

Supporting Information Available: Further procedures for selection of fraction of distribution tail to fit; evaluation of different fractions of the distribution tail to fit. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.
(2) Washburn, M. P.; Wolters, D.; Yates, J. R., III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **2001**, *19*, 242–247.
(3) Storey, J. D.; Tibshirani, R. Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 9440–9445.
(4) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identification made by MS/MS and database search. *Anal. Chem.* **2002**, *74*, 5383–5392.
(5) Tanner, S.; Shu, H.; Frank, A.; Ling-Chi, Wang; Zandi, E.; Mumby, M.; Pevzner, P. A.; Bafna, V. InsPecT: Identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* **2005**, *77*, 4626–4639.
(6) Higgs, R. E.; Knierman, M. D.; Freeman, A. B.; Gelbert, L. M.; Patil, S. T.; Hale, J. E. Estimating the statistical signficance of peptide identifications from shotgun proteomics experiments. *J. Proteome Res.* **2007**, *6*, 1758–1767.
(7) López-Ferrer, D.; Martínez-Bartolomé, S.; Villar, M.; Campillos, M.; Martín-Maroto, F.; Vázquez, J. Statistical model for large-scale peptide identification in databases from tandem mass spectra using sequest. *Anal. Chem.* **2004**, *76*, 6853–6860.
(8) Sadygov, R. G.; Yates, J. R., III. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal. Chem.* **2003**, *75*, 3792–3798.
(9) Zhang, N.; Aebersold, R.; Schwikowski, B. ProbID: A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* **2002**, *2*, 1406–1412.
(10) Bafna, V.; Edwards, N. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* **2001**, *17*, S13–S21.
(11) Cannon, W. R.; Jarman, K. H.; Webb-Robertson, B.-J. M.; Baxter, D. J.; Oehmen, C. S.; Jarman, K. D.; Heredia-Langner, A.; Auberry, K. J.; Anderson, G. A. Comparison of probability and likelihood models for peptide identification from tandem mass spectrometry data. *J. Proteome Res.* **2005**, *4*, 1687–1698.
(12) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, *3*, 958–964.
(13) Fenyo, D.; Beavis, R. C. A method for assessing the statistical significance of mass spectrometry-based protein identification using general scoring schemes. *Anal. Chem.* **2003**, *75*, 768–774.
(14) Alves, G.; Ogurtsov, A. Y.; Wu, W. W.; Wang, G.; Shen, R. F.; Yu, Y. K. Calibrating E-values for MS2 database search methods. *Biology Direct* **2007**, *5* (2), 26.
(15) Park, C. Y.; Klammer, A. A.; Käll, L.; MacCoss, M. P.; Noble, W. S. Rapid and accurate peptide identification from tandem mass spectra. *J. Proteome Res.* **2008**, *7* (7), 3022–3027.
(16) Weibull, W. A statistical distribution function of wide applicability. *J. Appl. Mech.* **1951**, *18* (3), 293–297.
(17) Klammer, A. A.; Yi, X.; MacCoss, M. J.; Noble, W. S. Improving tandem mass spectrum identification using peptide retention time prediction across diverse chromatography conditions. *Anal. Chem.* **2007**, *79* (16), 6111–6118.
(18) Craig, R.; Beavis, R. C. Tandem: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466–1467.
(19) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.* **2008**, *7* (1), 29–34.
(20) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **1995**, *57*, 289–300.
(21) Bailey, T.; Gribskov, M. Estimating and evaluating the statistics of gapped localalignment scores. *J. Comput. Biol.* **2002**, *9* (3), 575–593.
(22) Käll, L.; Canterbury, J.; Weston, J.; Noble, W. S.; MacCoss, M. J. A semi-supervised machine learning technique for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4*, 923–925.
(23) Pearson, W. R. Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **1998**, *276*, 71–84.
(24) Eddy, S. Maximum likelihood fitting of extreme value distributions. ftp.genetics.wustl.edu/pub/eddy/papers/evd.pdf, November 1997.
(25) Durbin, R.; Eddy, S.; Krogh, A.; Mitchison, G. *Biological Sequence Analysis*; Cambridge University Press: Cambridge, 1998.

PR8011107