research articles Journal of proteome research

Enhancing Peptide Identification Confidence by Combining Search Methods

Gelio Alves,[†] Wells W. Wu,[‡] Guanghui Wang,[‡] Rong-Fong Shen,[‡] and Yi-Kuo Yu*[,†]

National Center for Biotechnology Information, Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, and Proteomics Core Facility, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892

Received November 29, 2007

Confident peptide identification is one of the most important components in mass-spectrometry-based proteomics. We propose a method to properly combine the results from different database search methods to enhance the accuracy of peptide identifications. The database search methods included in our analysis are SEQUEST (v27 rev12), ProbID (v1.0), InsPecT (v20060505), Mascot (v2.1), X! Tandem (v2007.07.01.2), OMSSA (v2.0) and RAId_DbS. Using two data sets, one collected in profile mode and one collected in centroid mode, we tested the search performance of all 21 combinations of two search methods as well as all 35 possible combinations of three search methods. The results obtained from our study suggest that properly combining search methods does improve retrieval accuracy. In addition to performance results, we also describe the theoretical framework which in principle allows one to combine many independent scoring methods including *de novo* sequencing and spectral library searches. The correlations among different methods are also investigated in terms of common true positives, common false positives, and a global analysis. We find that the average correlation strength, between any pairwise combination of the seven methods studied, is usually smaller than the associated standard error. This indicates only weak correlation may be present among different methods and validates our approach in combining the search results. The usefulness of our approach is further confirmed by showing that the average cumulative number of false positive peptides agrees reasonably well with the combined *E*-value. The data related to this study are freely available upon request.

Introduction

Confident peptide identification through tandem mass spectrometry (MS) is one of the most important components in MS-based proteomics. For this reason, a great amount of effort has been invested to develop automated data analysis tools to identify peptides through tandem MS (MS^2) spectra. Among available data analysis tools, methods based on database searches are most frequently used. Because each search method uses a different algorithm and proceeds from a different view of what spectrum components contain the most critical information for identification, the search results for one spectrum from various search engines may differ significantly. However, it is also well-recognized that such difference may be turned into positive use: it would be useful to find complementary engines and combine the results in an effective way to enhance peptide identification.[1]

Combining search results from different methods, if feasible, definitely bears the possibility to improve the peptide identification confidence via reducing noise and utilizing complementary strengths. The difficulty in combining results from different search methods largely comes from the lack of a common statistical standard.[2] The importance of having a community standard has been stressed, and efforts in reaching such community standard have been invested. Using iterative expectation-maximization (EM), Keller et al. [3] proposed a statistical model to estimate the probability, determined through a global analysis of MS^2 spectra from an experiment, for a given spectrum to have correct peptide identification. In principle, results from different search methods may go through the same analysis and thus compared. However, if after the statistical analysis two different methods report different confident identifications for the same spectrum, one ends up needing to invent an *ad hoc* rule to decide which identification should be kept. Furthermore, to use the EM approach, one needs to *assume* or *guess*, without theoretical/statistical foundation, the forms of the score distributions for true positives and false positives. This, unfortunately, must weaken the validity of any statistical significance assignment obtained from such type of analyses.

In our recent work,[4] it was shown possible to calibrate the statistics (*E*-value) of various search methods to reach a universal standard that is in agreement with the fundamental definition of *E*-value. For a given query spectrum and quality score cutoff *S*, *E*-value is defined as the expected number of hits, in a random database, with quality score being the same as or larger than the cutoff. A realistic *E*-value assignment thus provides the user with the number of false positives to

* To whom correspondence should be addressed. E-mai: yyu@ncbi.nlm.nih.gov.
[†] National Center for Biotechnology Information, NIH.
[‡] Proteomics Core Facility, National Heart, Lung, and Blood Institute, NIH.

 10.1021/pr700798h
Published on Web 06/18/2008

anticipate when setting a quality score threshold. Most importantly, this peptide-centric statistical calibration allows one to combine search results even if the top hits from various methods disagree. Another possible approach to establish common statistical standard is through equating the false discovery rate (FDR)[5] of various methods considered. This approach, however, does not provide statistical significance for each peptide hit and thus is not directly applicable to peptide-centric combination of different search results. To be explicit, one may refer from FDR the *E*-value of a peptide hit with score identical to the first false positive, but any peptide hits with score better than the first false positive cannot have their *E*-value assigned. Furthermore, for peptides with scores falling in the range $[S_{k+1}, S_k]$, where S_k represents the kth best score of false positives, one cannot distinguish them statistically well except by using some *ad hoc* interpolations. There also exist other issues concerning misleading inference using FDR, but this is not the focus of the current paper and we refer the readers to a few relevant literatures.[6,7]

In this paper, in addition to providing a universal protocol to combine search results, we also carry out the performance assessment for all possible combinations, among seven database search methods, of two and three search methods using the Receiver Operating Characteristic (ROC) curves. The database search methods employed in our analysis are SEQUEST[8] (v27 rev12), ProbID[9] (v1.0), InsPecT[10] (v20060505), Mascot[11] (v2.1), X! Tandem[12] (v2007.07.01.2), OMSSA[13] (v2.0) and RAId_DbS.[14]

To better illustrate the main points of this paper, we have relegated to the Supporting Information a large number of ROC curves that convey similar information of that exhibited in the plots of the main text. Since the centroid mode seems to be the dominant mode in MS^2 database searches today, we present in the main text only results from centroid mode data and results from profile mode data are shown in the Appendix A. Throughout the paper, we use Dalton (Da) as the unit for molecular weight. In the following, we first describe the theoretical foundation for combining the search results. We will then describe briefly in Materials the implementation, followed by our main results: best combinations within search methods tested. We conclude with a brief summary, remarks and future directions.

Theory

In this section, we will start with the definitions of *P*-value and *E*-value, which will be frequently used for the rest of this paper. We then describe the mathematical underpinnings of how to combine the *P*-values of different database search methods to result in a final *E*-value. We should note that the mathematical formulation employed here was first introduced by Fisher,[15] and its extensions and applications to other research areas also exist.[16–18] To the proteomics community, however, this is still relatively new. Therefore, for the sake of completeness, we will provide sufficient mathematical details.

***P*-Value and *E*-Value.** Let us define the *P*-value and *E*-value in the context of peptide identification in database searches. For a given spectrum σ and a score cutoff S_c, one may ask what is the probability for a *qualified* (with molecular weight in the allowed range) random peptide to reach a score larger than or equal to S_c. This probability $P(S_c)$, a function of S_c, is called the *P*-value. For spectrum σ, if a database contains N_σ qualified, unrelated random peptides, one will expect to have $E(S_c) = N_\sigma P(S_c)$ number of random peptides to have quality score larger than or equal to S_c. This expectation value $E(S_c)$ is by definition the *E*-value associated with score cutoff S_c.

If one further assumes that the occurrence of a high-scoring random hit is a rare event and thus can be modeled by a Poisson process with expected number of occurrence $E(S_c)$, one may then define another *P*-value, which is called the database *P*-value, via

$$P_{\mathrm{db}}(S_c) = 1 - e^{-E(S_c)} \tag{1}$$

The database *P*-value $P_{\mathrm{db}}(S_c)$ represents the probability of seeing at least one hit in a given random database with quality score larger than or equal to S_c. Note that, at the level of P_{db}, one may compare the statistics from different search methods using different sizes of random databases. Because of the differences in the choices of optimal search parameters, it is likely that different search methods, for the same query spectrum, may search over different number of qualified peptides, that is, having different effective database sizes. Therefore, combining the database *P*-values is the natural choice if one were to merge results from different search methods.

Suppose that one wishes to combine the search results from L different search methods, each peptide candidate will have in principle L different *P*-values reported by the L search methods. The formula in eq 6 of the next subsection provides us the final combined *P*-value P_{comb} from the list. Once P_{comb} is obtained, we may invert the formula in eq 1 to get a combined *E*-value E_{comb} via

$$E_{\mathrm{comb}} = \ln\left(\frac{1}{1 - P_{\mathrm{comb}}}\right) \tag{2}$$

Having outlined how to obtain the final quantity of interest, E_{comb}, we now turn to the mathematical underpinnings of how to combine a list of, ideally independent, *P*-values reported by different database search methods.

Combine Independent *P*-Values. Consider an event labeled by a list of L independent quantities $s_1, s_2, \ldots, s_L$. Each quantity s_i may have an associated *P*-value p_i depending on the statistics of the variable s_i. An important issue to address is how one should combine all the p_i values to obtain an overall *P*-value. In the context of combining search results of different methods to assign statistical significance to a certain candidate peptide π, s_i represents the quality score assigned to π by the ith search method.

The question then reduces to the following. Given L random variables $(p_1, p_2, \ldots, p_L)$ uniformly distributed in the interval $(0, 1)$], what is the probability of finding their product to be smaller than a certain threshold τ. To put it in a more concrete framework, one may consider a unit hypercube whose interior points having coordinates $(x_1, x_2, \ldots, x_L)$ with $0 \le x_i \le 1$ for all $1 \le i \le L$. One then asks what is the volume bounded by the hypersurfaces $x_i \ge 0$ and $(\prod_{i=1}^{L} x_i) \le \tau$ with $\tau = \prod_{i=1}^{L} p_i$.

We may express this volume $F(\tau)$ mathematically as an integral:

$$F(\tau) = \int_0^1 \cdots \int_0^1 \theta\left(\tau - \prod_{i=1}^{L} x_i\right) \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_L \tag{3}$$

where $\theta(x)$ is a step function with $\theta(x > 0) = 1$ and $\theta(x < 0) = 0$. Apparently we have $F(\tau = 0) = 0$. We now evaluate the function $F(\tau)$ by first taking its derivative. Let $f(\tau) \equiv \partial F(\tau)/\partial\tau$, we have

$$f(\tau) = \int_0^1 \cdots \int_0^1 \delta\left(\tau - \prod_{i=1}^{L} x_i\right) \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_L \quad (4)$$

with $\delta(x)$ being the Dirac delta function that takes zero value everywhere except when $x = 0$ where its value approaches infinity.

For the ease of computation, we make the following changes of variables: $\tau \equiv e^{-t}$ and $x_i \equiv e^{-u_i}$. After this change, all the new variables t and u_i are in the range $(0,\infty)$. Equation 4 now becomes (with $\tau = e^{-t}$ understood)

$$f(\tau) = \int_0^\infty \cdots \int_0^\infty e^t e^{-\Sigma_i u_i} \delta\left(t - \sum_{i=1}^{L} u_i\right) \mathrm{d}u_1 \mathrm{d}u_2 \ldots \mathrm{d}u_L$$

where the identity $\delta(e^{-t} - e^{-c}) = e^t \delta(t - c)$ is used. Using the integral expression of the delta function,

$$\delta(t-c) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ik(t-c)} \mathrm{d}k$$

we may rewrite eq 4 as (with $\tau = e^{-t}$ understood)

$$\begin{aligned} f(\tau) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}k\, e^{t - ikt} \left[\int_0^\infty e^{-u_i(1-ik)}\right]^L \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}k\, e^{-it(k+i)} \left[\frac{i}{k+i}\right]^L = \frac{t^{L-1}}{(L-1)!} \end{aligned} \quad (5)$$

where the last equality results from choosing the integration path to enclose the lower half of the complex k plane. We may now go back to $F(\tau)$ by integrating $f(\tau)$.

$$\begin{aligned} F(\tau) &= \int_0^\tau f(\tau')\mathrm{d}\tau' = \int_\infty^{\ln(1/\tau)} e^{-t} \frac{t^{L-1}}{(L-1)!} (-\mathrm{d}t) \\ &= \int_{\ln(1/\tau)}^{\infty} e^{-t} \frac{t^{L-1}}{(L-1)!}\, \mathrm{d}t \\ &= \tau \sum_{n=0}^{L-1} \frac{[\ln(1/\tau)]^n}{n!} \end{aligned} \quad (6)$$

with $\tau = \prod_{i=1}^{L} p_i$ while combining the L P-values p_1, p_2,..., p_L. As specific examples, when $L = 2$, we have $F(p_1p_2) = p_1p_2[1 - \ln(p_1p_2)]$, and when $L = 3$, we have $F(p_1p_2p_3) = p_1p_2p_3[1 - \ln(p_1p_2p_3) + \frac{1}{2}\ln^2(p_1p_2p_3)]$. We will provide in the Appendix B more examples to elucidate the consequence of the formula provided in eq 6.

Materials

Two data sets were used in this study: one consists of centroid spectra, while the other consists of profile spectra. The centroid data set was generated by the Institute for Systems Biology.[19] This data set contains 12 subsets, each subset consists of spectra from a run of the experiment. Since the statistical calibration is not yet done for centroid data, we used some subsets (A5−A8) for statistical calibration following the protocol described earlier[4] while used other subsets (A1−A4 and A9−A12) for performance test. Each group of four subsets contains about 7000 raw spectra, a large number arising from not setting threshold during extraction. We chose not to set any threshold to avoid bias toward any method tested. However, inevitably, spectra of low parent ion count are all included, and thus, correct peptide identification from them is expected to be difficult. The profile data set was generated in house with procedures described in ref 14. To conform with the centroid data, no threshold was set during the extraction of the profile data. For profile data, since the statistical calibration was already done,[4] we adopted the results there and used them for analyzing our in-house data set as described in the next section.

Analyses and Results

For each spectrum, database search is performed for each of the seven methods studied using their respective default parameters; each method returns a list of candidate peptides. The candidate peptides in the reported lists are then compared against the target proteins. A candidate peptide is called a true positive if it is a partial sequence of any of the target proteins, and is called a false positive otherwise. Although we have obtained results from using both the centroid mode data as well as the profile mode data, in the figures and tables of the main text, we only present the centroid results. The corresponding results from using profile data are shown in the Appendix A for interested readers.

Individual performance of all seven methods and the performances of combining multiple methods are shown using the ROC curves which we detail below. For a given search method, the reported peptides may be pooled together to form two groups: one contains true hits and the other contains only false hits. Given a cutoff, which can be either score or E-value, one may further classify the group of true hits into *true positives* (TP) with score/E-value larger/smaller than the cutoff and *false negatives* (FN) with score/E-value smaller/larger than the cutoff. Similarly, with a cutoff given, one may also further classify the false hits into *false positives* (FP) with score/E-value larger/smaller than the cutoff and *true negative* (TN) with score/E-value smaller/larger than the cutoff. At a given cutoff, the *sensitivity* is expressed as TP/(TP + FN) while the *specificity* is given by TN/(FP + TN).

There are two types of ROC curves that one may use. The first kind plots sensitivity versus 1 − specificity by varying the cutoff. For this type of ROC, the area under curve (AUC), with maximum value 1, is also termed *accuracy* and the quantity 2 × AUC − 1 may be viewed as the *discriminating power*. The second kind of ROC curve plots directly TP (true hits with score/E-value larger/smaller than the cutoff) versus FP (false hits with score/E-value larger/smaller than the cutoff) by varying the cutoff. A ROC curve is therefore a parametric plot of either score, E-value, or other chosen internal parameter. The first type of ROC curve, although popularly used, does not reflect the *total number* of true hits found within a given cutoff. Furthermore, it is likely that for a given spectrum different search methods report different number of true/false hits, and thus, the trend of AUC may not agree with the second type of ROC curve. Because the AUC derived from the first type of ROC curve seems most common, we used it as a reference to sort different search method combinations but by no means suggest using it as the *only* measure of the merit of a search method or any combination of search methods. To be more complete, we find it informative to provide both types of ROC curves in our analysis.

To produce a ROC curve for a single search method, we used the E-value as the internal parameter for methods reporting E-values. For other methods, the internal parameter is chosen to be some sort of quality scores. For example, we use X-correlation for SEQUEST, posterior probability for ProbId, and MQ_score for InSpecT. When combining search results from multiple search methods to form a single ROC curve, we use the protocol described earlier[4] to calibrate E-values first, we then convert the E-value into a database P-value (see eq 1), use formula 6 to obtain the final P-value, and eq 2 to obtain the final combined E-value. It should be noted that, for different search methods, the best database retrieval may be achieved by using parameters other than those chosen here. For

example, it may be desirable to introduce some sort of discriminant function that is aimed to incorporate more information than the parameters chosen here. However, it is up to the software developers/experienced users to choose their best discriminant function and then calibrate the *E*-values using their best discriminant function. Therefore, one may not wish to view our ROC curves as complete performance comparisons. The key point of this paper is to show that there exists a theoretically and statistically sound approach to combine search results from different search methods.

In a ROC curve associated with a search method or a combination of a number of methods, the abscissa plots the number of false positives (or 1 − specificity) and the ordinate plots the number of true positives (or sensitivity). Therefore, the more toward the upper-left corner a ROC curve is, the better the corresponding method is performing. Another popular way to assess the performance is to count the number of true positives at a *fixed* false positive number threshold. One may, for example, fix the false positive number threshold to be 500 and see how many true positives are found by each method (allowing up to 500 false positives) to evaluate various methods.

Standardized *E*-Values and Single Search Method. We will start this section with a brief description of how one may standardize statistics via converting quality scores reported by different search methods to *E*-values. Readers interested in details are encouraged to read ref 4. The basic idea is to adhere to the textbook definition of the *E*-value: expected number of (false positive) hits from a random database. For a peptide hit with quality score *S*, its *E*-value *E*(*S*) indicates the expected number of hits from a random database with quality score larger or equal to *S*.

One starts by constructing randomized protein databases of various sizes and making sure that true peptides (partial sequences of target proteins used for the calibration purpose) are absent from those random databases. Querying the spectra produced from target protein mixture against the random databases, one obtains for each spectrum only false positive hits. To calibrate the statistics, one needs an internal parameter which may vary from method to method. For a method that reports *E*-value, we simply use the reported *E*-value as the internal parameter. Otherwise, one may use e^{-S} as the internal parameter, where *S* represents the reported quality score. Basically, a more confident hit is associated with a smaller internal parameter. After this step is done, for a given method, one may pool the reported peptide hits from all spectra and compute for a given internal parameter cutoff the total number of false positives with their internal parameters less than the cutoff. Upon dividing this number by the total number of query spectra, one obtains the "expected number of false positives" within a given internal parameter cutoff. For every search method, this procedure allows one to identify the corresponding "expected number of false hits" within any given internal parameter cutoff, which in turn is related to the statistical significance or the quality scores reported. Since the calibrated *E*-values will depend on the database size, one also needs to identify the necessary database size corrections, which is demonstrated explicitly in ref 4. The accuracy of this standardization procedure may be obtained via repeating the procedure several times, each time with a different random database, for every database size of interest.

We have applied the protocol mentioned above to calibrate profile data. Therefore, no further calibration is done when analyzing profile data search results. For a given peptide hit, the respective calibration formula presented earlier[4] was employed to generate the standardized *E*-value for each search method. On the other hand, the complete *E*-value calibration protocol is performed on the A5−A8 subsets of the centroid data to obtain the calibration formulas for each of the seven methods tested.

In Figure 1, we demonstrate the statistical calibrations for centroid data. Using formulas found for profile data,[4] we first transform the respective quality scores (or *E*-values) of each method to profile *E*-values. Panel A of Figure 1 shows the profile *E*-value versus the average of the cumulative number of false positives when tested using the centroid data (A5−A8 subsets of ref 19). It is seen that the profile calibration, when used on centroid data, still provides the *relative E*-values reasonably accurately. To bring the *E*-value to agree with its fundamental definition, all one needs is an additional method-specific constant factor to be multiplied to the calibration formula developed for profile data. Specifically, for a given method, we have $E_{\text{cent.}} = a_{\text{method}} E_{\text{prof.}}$, where $E_{\text{prof.}}$ represents the *E*-value obtained through employing calibration formulas[4] given earlier and $E_{\text{cent.}}$ represents the calibrated *E*-value when data is collected using centroid mode.

As an example, for SEQUEST, we need an additional factor of ¼ to bring the profile calibration to the centroid calibration. For a peptide hit with an *X*-correlation value 3.5, our profile mode calibration[4] converts this *X*-correlation value into an *E*-value of 0.00413 when searching a database of size 100 Mega residues. This means that for a hit, from querying a centroid spectrum, with the same *X*-correlation value, the corresponding *E*-value will be 0.00103, a more significant value than before. This, however, is not surprising because, when the spectrum is more sparse (centroid mode), it is less likely to have a strong *X*-correlation, and thus, a smaller *E*-value should be assigned to a centroid hit than to a profile hit when the same *X*-correlation value is obtained for both.

Table 1 documents the overall factor needed for each method. Except for Mascot, every method has a conversion factor less than or equal to one, similar to the scenario of SEQUEST. Since the scoring detail of Mascot is not available, we cannot make comments regarding why when transforming from $E_{\text{prof.}}$ to $E_{\text{cent.}}$ it has a conversion factor larger than 1. It is worth noting that for RAId_DbS and X! Tandem the mode conversion factor a_{method} is 1; that is, these two methods require no further adjustments in statistical significance assignment for data acquired in centroid mode.

Similar to panel A, panel B of Figure 1 also shows the statistical calibration results but after (manually) removing highly homologous peptides from the hit list and after the method-specific factor (Table. 1) has been applied. In panels C and D of Figure 1, we apply the calibrated formula (along with the method-specific factor found from the first calibration) to the subsets A1−A4 and A9−A12 of ref 19. We find that the calibration done using subset A5−A8 (when applied to other centroid data subsets) provides us with realistic statistics, supporting the universality of statistical calibration.

For performance baselines on peptide identification, we show the ROC curves in Figure 2 for the seven methods tested. Panels A and B display ROC curves of the first and the second type for the methods tested when spectral data are acquired in centroid mode (subsets A1−A4 of ref 19). Note that the number of false positives in panel B may seem high at the level of say 100 true positives. This may be because a peptide is

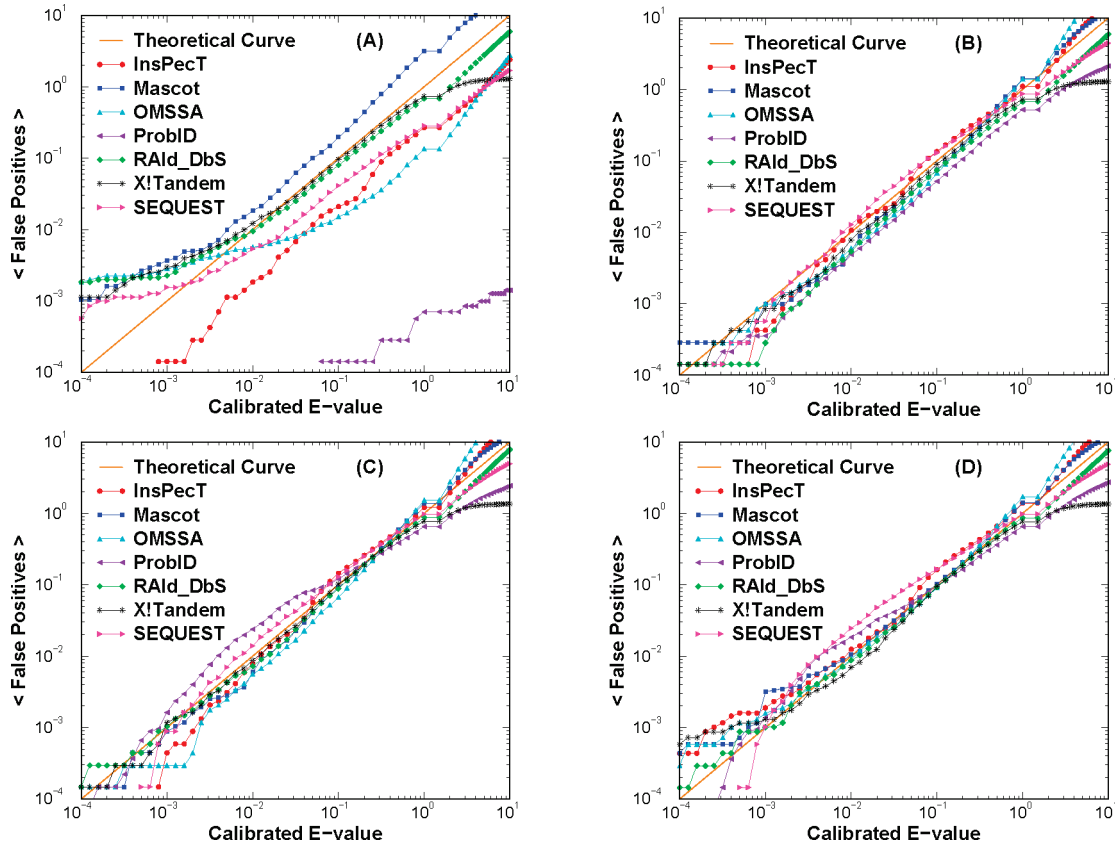

Figure 1. The statistical calibrations for centroid data. (A) The profile *E*-value (using calibration formulas[4]) versus the average of the cumulative number of false positives when tested using the centroid data (A5−A8 subsets of ref 19). (B)The statistical calibration results but after (manually) removing highly homologous peptides from the hit list and after the method-specific factor a_{method} (see text) has been applied. (C and D) We apply the calibrated formula (along with the method-specific factor found from the first calibration) to the subsets A1−A4 and A9−A12 of ref 19. We find that the calibration done using subset A5−A8 (when applied to other centroid data subsets) provides us with realistic statistics, supporting the universality of statistical calibration. It is worth noting that the lowest *E*-value in those calibration plot can only go to roughly one over the total number of spectra used for calibration. Since we used about 10 000 spectra, the lowest *E*-value that can be shown is of order 10^{-4}. In real database searches, a really significant hit probably may have an *E*-value much smaller than 10^{-4} and many users may not wish to consider hits with *E*-values larger than 10^{-1}.

Table 1. The Numerical Factor Needed to Obtain Calibrated *E*-Values for Centroid Data from Calibrated *E*-Values for Profile Data

method	a_{method}
RAId_DbS	1
X! Tandem	1
Mascot	2
OMSSA	1/6
ProbID	6.1×10^{-5}
SEQUEST	1/4
InsPecT	1/5

counted as a false positive if it is not a partial sequence of any of the target proteins, even if it is *very homologous* to the true hit or if it is a partial sequence of proteins that are *very homologous* to the target proteins. As a matter of fact, if one were to introduce a *decoy* database, one will not find any hits from the decoy database within the cutoff yet. This will suggest that one is looking at a region of zero FDR, while if one were to calculate the FDR from the ROC curve, one will get a large value. Therefore, it is a region where FDR exhibits a considerable uncertainty and we do not advise the readers to infer FDR from the ROC curves provided here.

Prior to presenting the results from combining search results from multiple search methods, let us first outline how *P*-values are combined. For a given spectrum σ, to combine search results from m search methods (say method A_1,..., A_m), we first construct a union peptide list $L(\sigma) \equiv L_{A_1}(\sigma) \cup ... \cup L_{A_m}(\sigma)$, where $L_{A_i}(\sigma)$ is the reported list of peptide hits by method A_i for spectrum σ. A peptide in the union list has at least one, and may have up to m calibrated *E*-values, depending on how many search methods reported that specific peptide in their candidate lists. Each of the calibrated *E*-values associated with a peptide will be first transformed into a database *P*-value. For a given peptide π, for method(s) that did not report π as a candidate, the associated database *P*-value(s) of π from that (those) method(s) is (are) set to 1. After this procedure, each peptide in the list $L(\sigma)$ have m database *P*-values and eq 6 is applied to obtain the final *P*-value associated with π. The final *P*-value $P_{\text{comb}}(\pi)$ will then be transformed into a final *E*-value $E_{\text{comb}}(\pi)$ via eq 2. We then use $E_{\text{comb}}(\pi)$ as the final *E*-value to determine the statistical significance of peptide candidate π, similar to what is used in ref 18.

We now comment on the effect of assigning *P*-value of 1 for the missing *P*-values. Prior to combining *P*-values, setting the unreported peptide's *P*-value to be 1, the largest *P*-value (or least statistically significant) possible, makes the final *P*-value larger than it should be, that is, more conservative. However, Figure 3 shows that even with this drastic choice

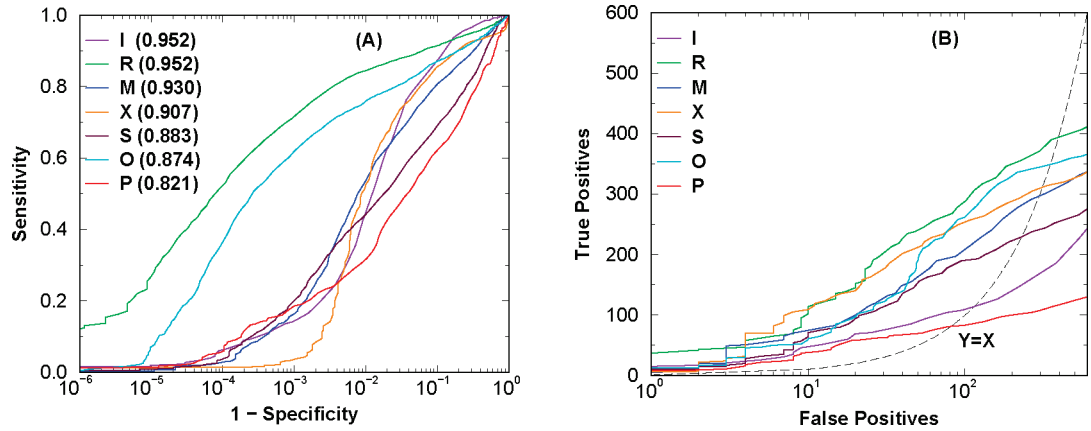

Figure 2. ROC curves for the seven database search methods tested when using the centroid data (A1−A4 of ISB data set). Each search method is abbreviated by its first letter in the figure legend. ROC curves of the first type are displayed in panel A, while the ROC curves of the second type are displayed in panel B. Since the total number of spectra in this subset is about 7000, in panel B, the displayed highest number of false positives, 600, corresponds approximately to *E*-value 0.1. We did not show ROC curves of the second type to larger FP value because users probably will not be too interested in the large *E*-value regime.

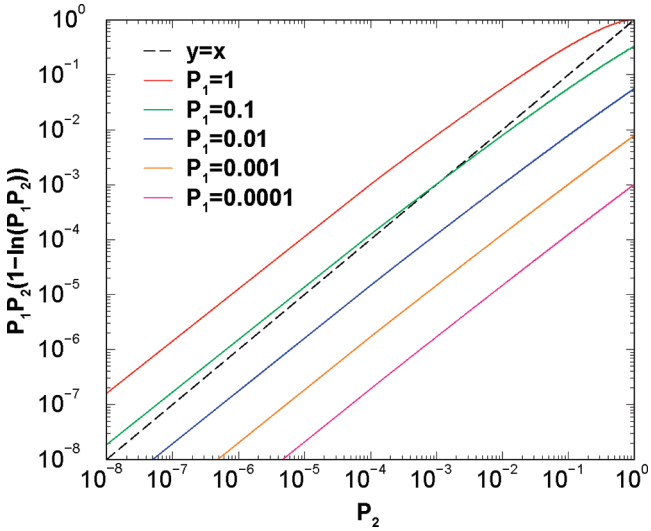

Figure 3. Final *P*-value from combining a reported *P*-value P_2 and a fixed *P*-value P_1. The fixed *P*-value P_1 is chosen to be either 1, 10^{-1}, 10^{-2}, 10^{-3} or 10^{-4}. As one may see, although the relationship between P_2 and the final *P*-value is still reasonably linear in the log − log plot, the slope has deviated from 1.

the increase in the final *P*-value, from combining the reported *P*-value and 1, is not drastically larger than the reported *P*-value. We also display in Figure 3 the final *P*-value versus a starting *P*-value when it is combined with a *P*-value of 0.1, 0.01, 0.001, or 0.0001.

Pairwise Combinations. Using E_{comb} as the internal parameter, the cumulative number of TP, FN, FP and TN may be expressed as

$$\text{TP}(E_{\text{comb}} \le E_c) = \sum_{\sigma} \sum_{\pi \in L(\sigma)} \theta(E_c - E_{\text{comb}}(\pi) | \pi \in \{tp\}) \quad (7)$$

$$\text{FN}(E_{\text{comb}} \le E_c) = \sum_{\sigma} \sum_{\pi \in L(\sigma)} \theta(E_{\text{comb}}(\pi) - E_c | \pi \in \{tp\}) \quad (8)$$

$$\text{FP}(E_{\text{comb}} \le E_c) = \sum_{\sigma} \sum_{\pi \in L(\sigma)} \theta(E_c - E_{\text{comb}}(\pi) | \pi \notin \{tp\}) \quad (9)$$

$$\text{TN}(E_{\text{comb}} \le E_c) = \sum_{\sigma} \sum_{\pi \in L(\sigma)} \theta(E_{\text{comb}}(\pi) - E_c | \pi \notin \{tp\}) \quad (10)$$

where $\{tp\}$ represents the set of all partial sequences of target proteins and $\theta(x|G)$ is a conditional step function that takes value 1 if both $x \geq 0$ and condition G holds true, and takes value 0 if $x < 0$ or condition G is false.

Seven pairwise combinations of search methods with *best* AUC, measured by the first type of ROC, are shown in Figure 4. Panel A shows ROC curves of the first kind resulting from the centroid data (A1−A4 subsets of ref 19), while panel B documents the ROC curves of the second kind from the same data. It is apparent that many of these pairwise combinations of search methods outperform, in terms of AUC and TP(FP = 500), each individual search methods shown in Figure 2. This provides a proof of principle that properly combining search results does enhance peptide identification accuracy.

Triplet Combinations. Using E_{comb} as the internal parameter, the cumulative number of TP, FN, FP and TN are obtained via eqs 7–10. The ROC curves of seven combinations of three search methods, giving rise to the seven best AUC, are shown in Figure 5. Panel A shows ROC curves of the first kind resulting from the centroid data (A1−A4 subsets of ref 19), while panel B documents ROC curves of the second kind from the same data. It is quite visible that most of these combinations of three search methods outperform individual search methods and the pairwise combinations of search methods shown in Figures 2 and 4. Nevertheless, it is also obvious that the improvement from combining two search methods to combining three search methods is weaker than from single search method to combining two search methods. A preliminary test, combining four best individual performers, seems to be in accordance with the trend of diminishing improvement size.

The trend of diminishing improvement size—when going from single search method, combining two search methods, to combining three search methods and more—suggests that there may exist non-negligible correlations among the search methods examined. In some way, this is intuitively plausible. Because most methods agree on the idea that, for MS^2 spectra produced by collision induced dissociation, *b* and *y* fragment series are most prominent, the scoring will somehow emphasize more these fragment series thus introducing some correlations among various methods. To quantify the correlations among methods and the impact of correlations on the effectiveness of combining search

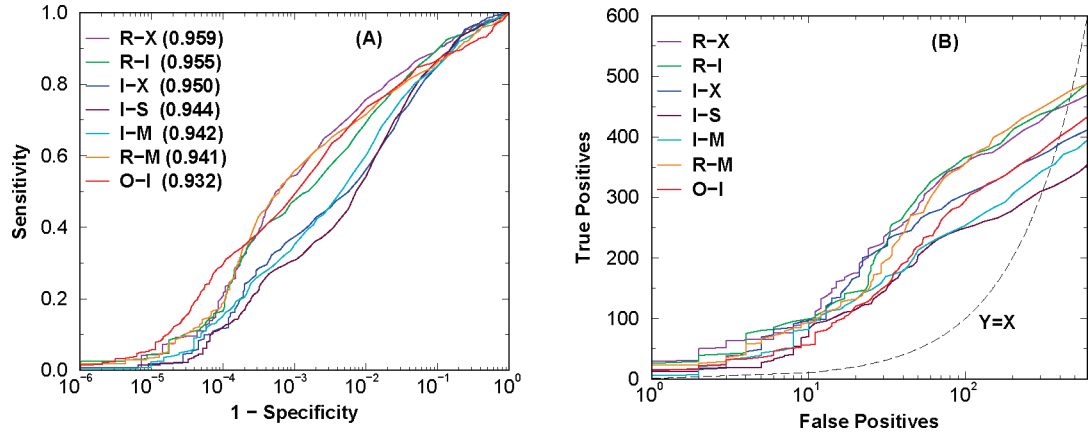

Figure 4. ROC curves for the seven pairwise combinations giving rise to seven largest AUC, values shown in panel A, when using the centroid data (A1−A4 subsets of the ISB data). Each search method is abbreviated by its first letter in the figure legend. ROC curves of the first type are displayed in panel A. Panel B shows ROC curves of the second type. Since the total number of spectra in this subset is about 7000, in panel B, the displayed highest number of false positives, 600, corresponds approximately to *E*-value of 0.1. We did not show ROC curves of the second type to larger FP value because users probably will not be too interested in the large *E*-value regime.

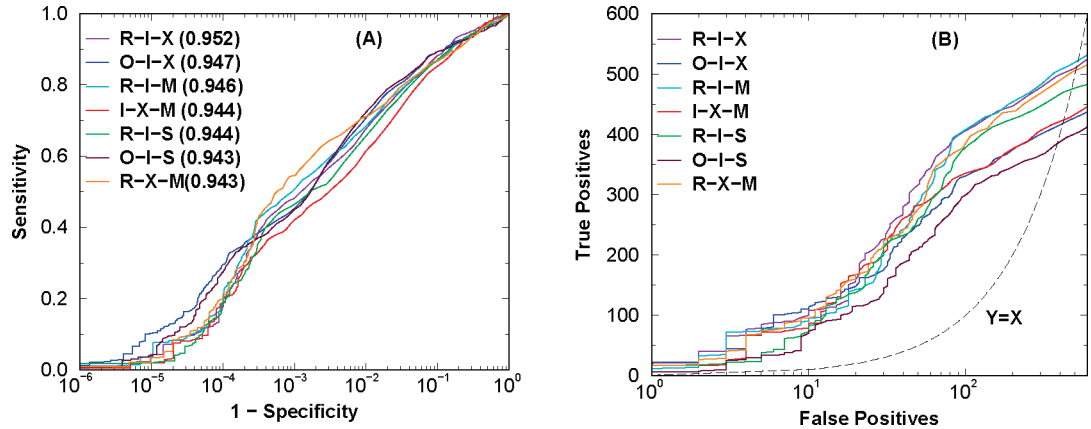

Figure 5. ROC curves for the seven triplets giving rise to seven largest AUC, values shown in panel A, when using the centroid data (A1−A4 subsets of the ISB data). Each search method is abbreviated by its first letter in the figure legend. ROC curves of the first type are displayed in panel A. Panel B shows ROC curves of the second type. Since the total number of spectra in this subset is about 7000, in panel B, the displayed highest number of false positives, 600, corresponds approximately to *E*-value of 0.1. We did not show ROC curves of the second type to larger FP value because users probably will not be too interested in the large *E*-value regime.

methods, the pairwise method−method correlations are investigated whose details we now turn to.

Method Correlations and Combined Statistics. For a pair of search methods, say method A and method B, the correlation between them is assessed in two ways. First, we consider the number of common false positives and the number of common true positives up to a given specified *E*-value threshold. Second, we consider the correlation of reported *E*-values. Since not every peptide will be reported by both methods, we need to simulate the missing *E*-values in order to perform the second investigation. In addition to method correlation, we also examine here how well the final *E*-value, after combining the results from two methods, agrees with the theoretical definition.

Consider a given spectrum σ. Let us again denote by $L_A(\sigma)$ ($L_B(\sigma)$) the candidate peptide list returned by method A (B) when using σ as the query spectrum. In addition to the union list $L(\sigma) \equiv L_A(\sigma) \cup L_B(\sigma)$, let us also define the intersection list $I(\sigma) \equiv L_A(\sigma) \cap L_B(\sigma)$. Each peptide in the list $I(\sigma)$ is thus either a common true positive or a common false positive.

As before, each peptide in the list $L(\sigma)$ has its final *E*-value. We denote by $\mathrm{TTP}_\sigma(E \le E_c)$ the total number of true positives in the list $L(\sigma)$ with *E*-value less than or equal to E_c. We further denote by $\mathrm{CTP}_\sigma(E \le E_c)/\mathrm{CFP}_\sigma(E \le E_c)$ the number of common true/false positives in the list $I(\sigma)$ with *E*-value less than or equal to E_c. Ideally, methods that complement each other well should have $\sum_\sigma \mathrm{CTP}_\sigma(E \le E_c)$ large and $\sum_\sigma \mathrm{CFP}_\sigma(E \le E_c)$ small. Furthermore, one will mostly be interested in only the region of *E*-value cutoff where $\sum_\sigma \mathrm{CFP}_\sigma(E \le E_c) \le \sum_\sigma \mathrm{TTP}_\sigma(E \le E_c)$.

The following ratios may serve as measures of degree of correlation between two search methods at various *E*-value cutoffs:

$$\mathrm{RC}(E \le E_c) = \frac{\sum_\sigma \mathrm{CTP}_\sigma(E \le E_c)}{\sum_\sigma \mathrm{CFP}_\sigma(E \le E_c)} \tag{11}$$

$$P_{F\to T}(E \le E_c) = \frac{2\sum_\sigma \mathrm{CFP}_\sigma(E \le E_c)}{\sum_\sigma \mathrm{CFP}_\sigma(E \le E_c) + \sum_\sigma \mathrm{TTP}_\sigma(E \le E_c)} \tag{12}$$

Basically, one may regard $P_{F\to T}(E \le E_c)$ as the probability of mistaking a false hit reported by both methods as a positive hit upon combining the search result. To be more specific, up

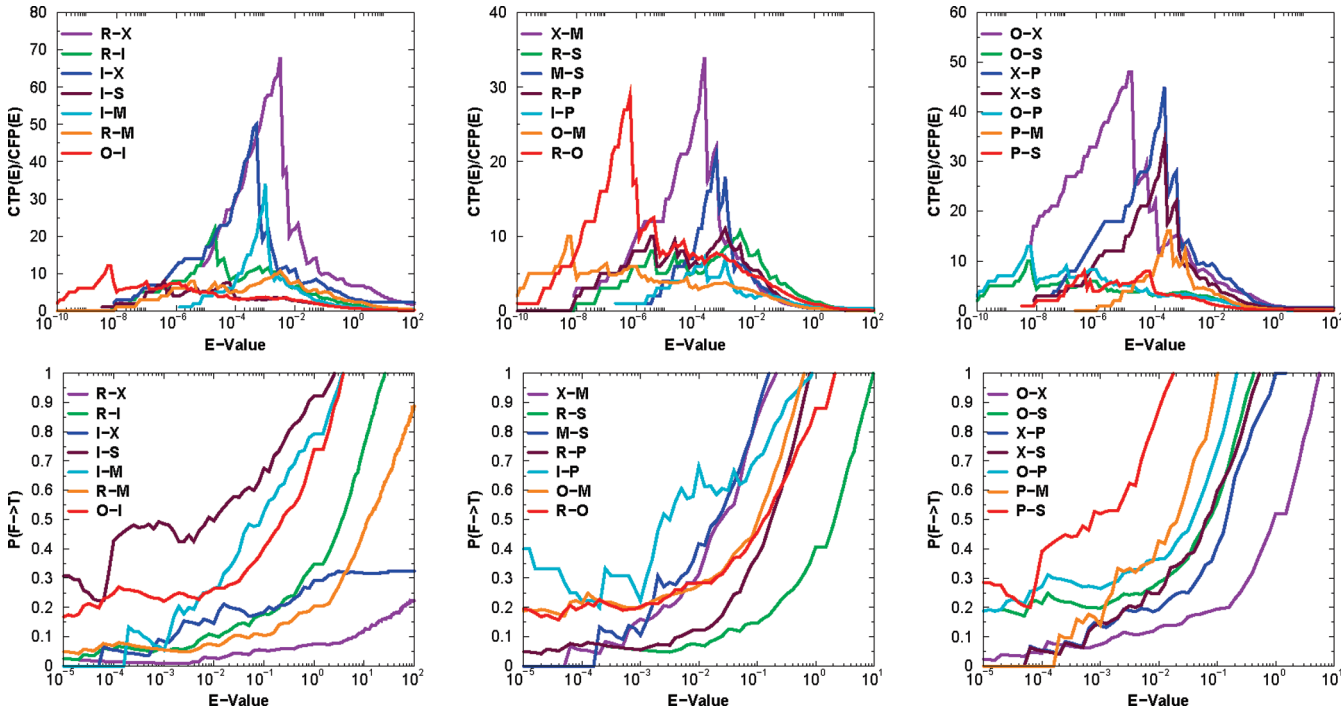

Figure 6. Method correlations evaluated using the centroid data (A1−A4 subsets of the ISB data). Each search method is abbreviated by its first letter in the figure legend. The panels on the first row display the RC ratio, CTP($E \leq E_c$)/CFP($E \leq E_c$), described in eq 11 as a function of the cutoff *E*-value. The panels on the second row display the likelihood of mistaking a common false hit as a significant hit, see eq 12.

to a given *E*-value cutoff, provided the best *E*-values reported by both search methods are smaller than the cutoff, combining the search results will have approximately the probability $P_{F\rightarrow T}(E \leq E_c)$ for counting a common false hit as a significant positive hit.

The two ratios introduced above as a function of the cutoff *E*-value E_c are plotted for every possible pairwise combinations of search methods in Figure 6 (for the centroid data). These plots provide information regarding the range of combined *E*-values that is fruitful to use. For example, if one were to look at the merged results from RAId_DbS and Mascot, one probably will consider peptides with combined *E*-values as high as 0.1 where the probability of incorporating a false positive into a true positive drastically increases and the magnitude of RC drops significantly.

As an illustration of how we calculate the global correlation between method A and B, let us again consider a given spectrum σ, the hit lists $L_A(\sigma)$, $L_B(\sigma)$ and the union hit list $L(\sigma) \equiv L_A(\sigma) \cup L_B(\sigma)$. Each peptide in the union list will constitute a data point. The overall analysis of all the points in $L(\sigma)$ will generate a spectrum-dependent correlation between the two methods. One may obtain the mean correlation by averaging over a large number of spectra.

For a candidate peptide that is reported by both methods, say methods A and B, we plot, respectively, the logarithm of the reported *E*-values by methods A and B along the *x*-axis and the *y* -axis. For a candidate peptide that is reported by only one method, say method A, one needs to estimate its corresponding *E*-value if it were to be reported by method B. As one may have expected, such *E*-value will be larger than or equal to the maximum *E*-value, $E_{\max}(B, \sigma)$, reported by method B for the spectrum considered. To tackle this problem, we employ the approximation method elaborated earlier.[4] Following our earlier elaboration,[4] the probability of having at least *r* peptide hits all with *E*-value smaller than or equal to E_c is given by $e^{-E_c}[\sum_{l=k}^{\infty} E_c^l/l!] = 1 - e^{-E_c}[\sum_{l=0}^{k-1} E_c^l/l!]$. Consequently, the probability of having at least one hit, out of $k+1$ reported hits, with *E*-value larger than E'_c is given by $e^{-E'_c}[\sum_{l=0}^{k} E'^l_c/l!]$. The associated probability density is found to be $e^{-E'_c}E'^k_c/k!$. We may then sample using this pdf but restricted to the region $E \geq E_{\max}(B, \sigma)$. Similarly, for a peptide that is not reported by method A, we sample its *E*-value from the same pdf but restricted to the region $E \geq E_{\max}(A, \sigma)$. The *simulated* Pearson correlation between two method for a given spectrum σ is then easily obtained.

The *simulated* correlation between any two methods is computed for each spectrum. The average correlation and the associated standard error over all available spectra from a given data type are also computed. Table 2 documents the results obtained. Most pairwise correlation has mean close to 0 and size much smaller than its associated standard error. This implies that the correlation strength between any pair of method is at most weak, which was also observed in ref 2 and justifies our use of the Fisher's formula in combining the statistical significance. However, one should note that we are presenting the correlations averaged over a large number of spectra. For each individual spectrum, there might still exist non-negligible correlations, positive or negative, among various methods which eventually results in a diminishing improvement size upon combining more and more search methods.

Finally, we examine the accuracy of the final combined *E*-value. For a given spectrum σ, let $L(\sigma) \equiv L_A(\sigma) \cup L_B(\sigma)$ denote the union hit list from considering both methods A and B, with $L_A(\sigma)$ ($L_B(\sigma)$) representing the hit list from method A(B). Within $L(\sigma)$, we may separate the peptide hits into two groups: one

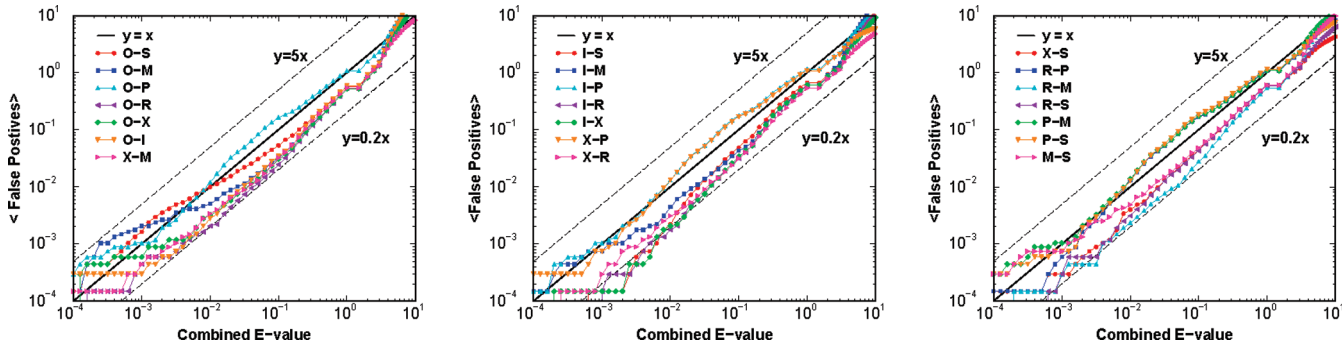

Figure 7. Examination of the combined *E*-value when using the centroid data (A1−A4 subsets of the ISB data). In every panel, the average cumulative number of false hits is plotted against the combined *E*-value. Within the *E*-value range investigated, the final combined *E*-value is mostly within a factor of 5 of the theoretical value, represented by $y = x$ lines. As before, each method is represented by its first letter in the figure legend.

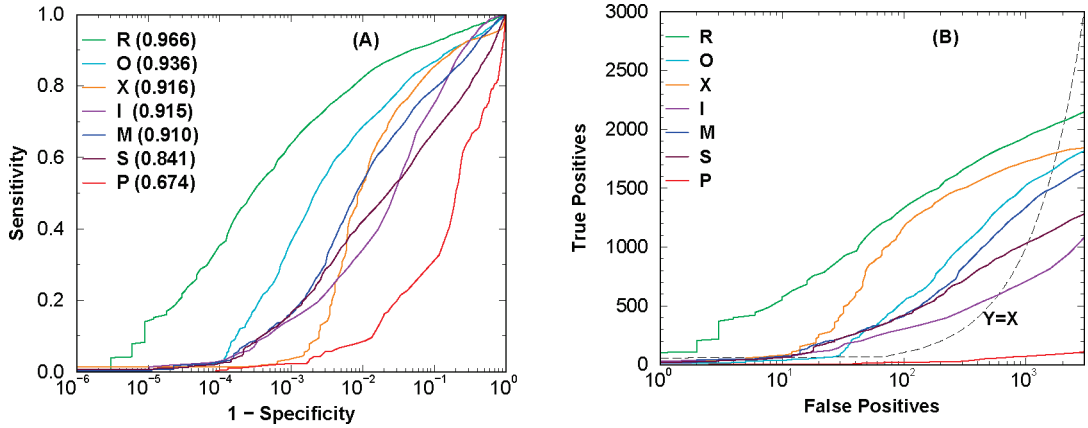

Figure 8. ROC curves for the seven database search methods tested when using the profile data. Each search method is abbreviated by its first letter in the figure legend. ROC curves of the first type are displayed in panel A, while the ROC curves of the second type are displayed in panel B. Since the total number of spectra in this subset is about 7000, in panel B, the displayed highest number of false positives, 3000, corresponds approximately to *E*-value of 0.4. We did not show ROC curves of the second type to larger FP value because users probably will not be too interested in the large *E*-value regime.

contains the true hits and the other contains only false hits. Using the group containing only the false hits, we may find the cumulative number of false positives as a function of the *combined E*-value cutoff. Ideally, the average number of false positive should be identical to the *E*-value cutoff and thus follow the $y = x$ straight line. In Figure 7, we plot the combined *E*-value along the abscissa and the average number of false positives along the ordinate. The curves plotted mostly band together within the 5-fold range of the theoretical line; that is, they mostly fall between the two straight lines parametrized by $y = 5x$ and $y = x/5$.

Concluding Summary and Outlook

In this paper, we propose a procedure suitable for combining search results for different database search methods. The method proposed is generic and can, in principle, be applied to any identification methods (*de novo* sequencing, spectral library or database searches) or any independent information one wishes to combine. The key factor that makes combining search methods possible is to have common statistics standard among methods/information of interest. Such a common statistics standard may be obtained by a universal protocol for statistical calibration developed earlier[4] or by other means that can assign meaningful database *P*-values for each candidate peptides reported by each search method.

When comparing the results in Figure 2 to their counterparts using profile data (see Figure 8 in Appendix A), it seems consistently true that more true positive hits were found from profile data compared to centroid data for every search method even though most search methods are designed for searching using centroid data. This identification rate increase should be further checked by more tests. If it turns out to be generally true, it might be attributed to the fact that the profile data contains more information than the centroid data and may motivate software developers to emphasize profile mode database searching in their future development.

Accurate peptide identification may benefit significantly protein identification, one of the most important problems in proteomics. As far as how one may maximize protein identification through having accurate statistics at the level of peptide identification, it remains an open and interesting problem that deserves a separate, thorough investigation.

In our method correlation studies, we had focused on studying the average correlation between a pair of methods over a large number of spectra. However, it has not escaped our attention that it may also be fruitful to calculate for each spectrum the correlation between a pair of methods and use it to modify the strategy of combining search results. This idea, although interesting, is beyond the scope of the current paper and should be investigated in the near future.

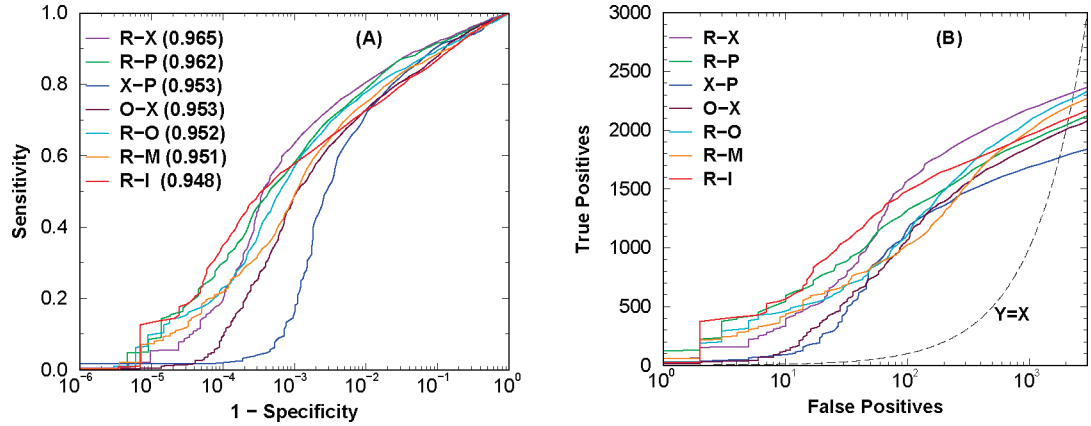

Figure 9. ROC curves for the seven pairwise combinations giving rise to seven largest AUC, values shown in panel A, when using the profile data. Each search method is abbreviated by its first letter in the figure legend. ROC curves of the first type are displayed in panel A for the profile data. For the same data, panel B shows ROC curves of the second type. Since the total number of spectra in this subset is about 7000, in panel B, the displayed highest number of false positives, 3000, corresponds approximately to *E*-value of 0.4. We did not show ROC curves of the second type to larger FP value because users probably will not be too interested in the large *E*-value regime.

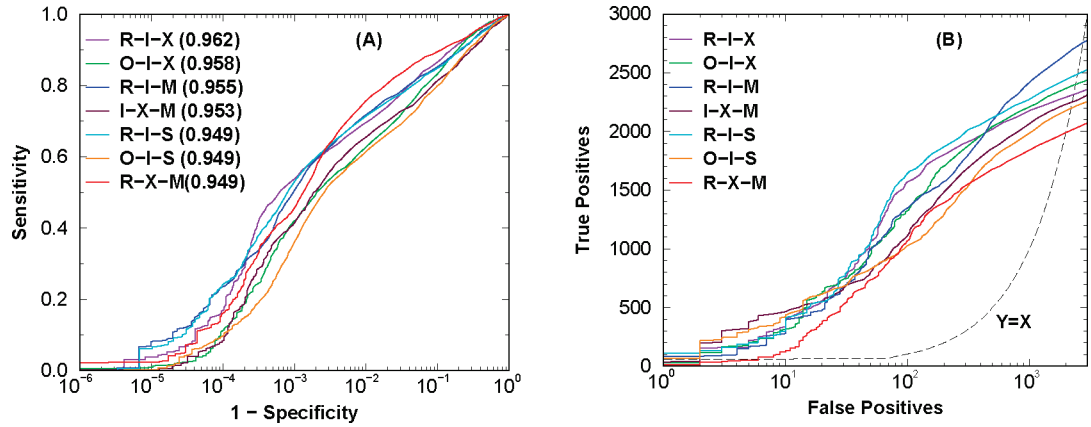

Figure 10. ROC curves for the seven triplets giving rise to seven largest AUC, values shown in panel A, when using the profile data. Each search method is abbreviated by its first letter in the figure legend. ROC curves of the first type are displayed in panel A. Panel B shows ROC curves of the second type. Since the total number of spectra in this subset is about 7000, in panel B, the displayed highest number of false positives, 3000, corresponds approximately to *E*-value of 0.4. We did not show ROC curves of the second type to larger FP value because users probably will not be too interested in the large *E*-value regime.

Appendix A

In the appendix, we show the results from using the profile data.

Panels A and B of Figure 8 show, respectively, ROC curves of the first and the second type when spectral data are collected in profile mode. In Figure 9, panel A displays the ROC curves of the first kind resulting from the profile data, while panel B documents the ROC curves of the second kind resulting from the same data. In Figure 10, panel A displays the ROC curves of the first kind resulting from the profile data, while panel B documents ROC curves of the second kind resulting from the same data.

The two ratios, RC and $P_{F\to T}$, introduced in the main text as a function of the cutoff *E*-value E_c are plotted for every possible pairwise combinations of search methods in Figure 11 when using the profile data.

Analogous to Table 2, Table 3 documents the average correlation and its standard error between any two methods when using the profile data. Most pairwise correlation has mean close to 0 and size much smaller than its associated standard error. This implies that the *average* correlation strength between any pair of method are at most weak, which was also observed in ref 2.

In Figure 12, we plot the combined *E*-value along the abscissa and the average number of false positives along the ordinate when using the profile data. The curves plotted mostly band together within the 5-fold range of the theoretical line; that is, they mostly fall between the two straight lines parametrized by $y = 5x$ and $y = x/5$.

Appendix B

To elucidate the behavior of the combined *P*-value formula expressed in eq 6, we consider three cases.

Case A: First, $p_1 = p < 1$ and $p_2 = 1$, we obtain the combined *P*-value to be $p[1 + \ln(1/p)]$; second, $p_1 = p < 1$, $p_2 = 1$, and $p_3 = 1$, when combining these three *P*-values together, we obtain a larger value, $p[1 + \ln(1/p) + (\ln(1/p))^2/2]$, than the first case. This is intuitively reasonable: the latter case has more evidence for the hit to be insignificant and thus should be assigned a larger *P*-value. To avoid potential confusion, let us comment on the large *L* limit that constitutes our cases B and C.

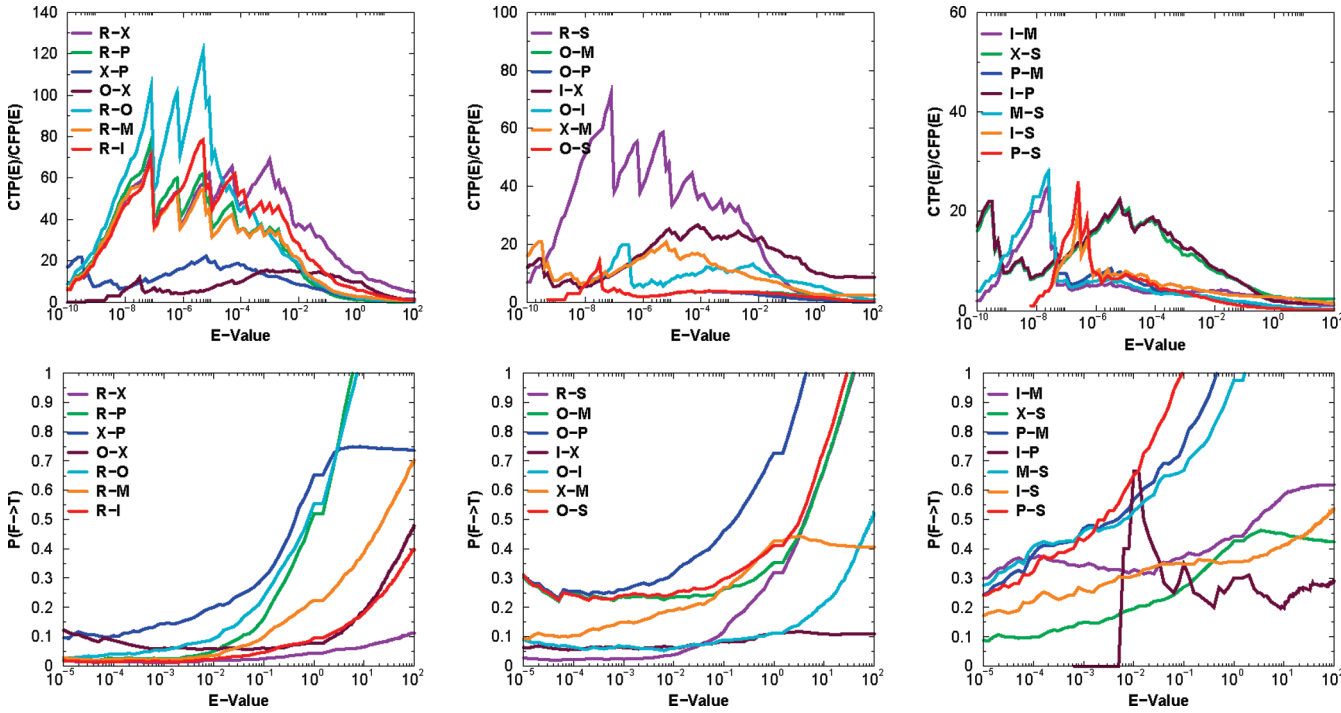

Figure 11. Method correlations evaluated using profile data. Each search method is abbreviated by its first letter in the figure legend. The panels on the first row display the RC ratio, CTP($E \leq E_c$)/CFP($E \leq E_c$), described in eq 11 as a function of the cutoff E-value. The panels on the second row display the likelihood of mistaking a common false hit as a significant hit, see eq 12.

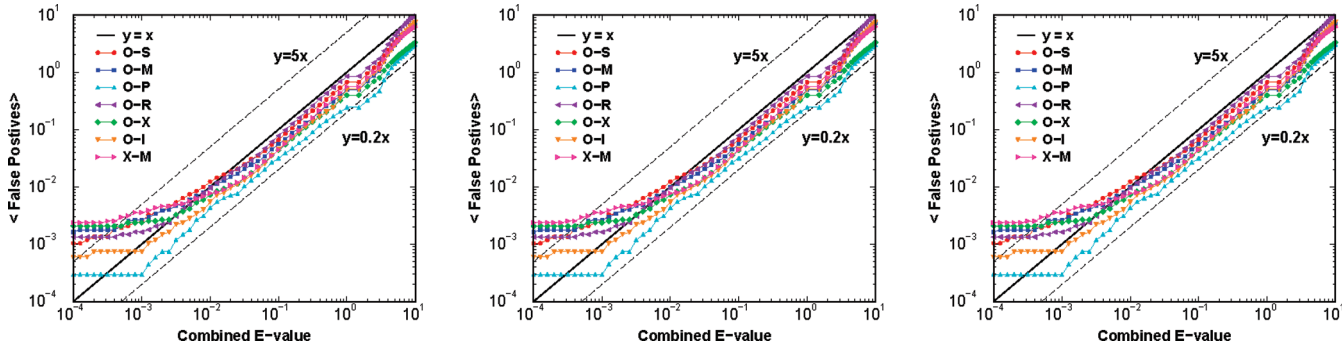

Figure 12. Examination of the combined E-value when using the profile data. In every panel, the average cumulative number of false hits is plotted against the combined E-value. Within the E-value range investigated, the final combined E-value is mostly within a factor of 5 of the theoretical value, represented by $y = x$ lines. As before, each method is represented by its first letter in the figure legend.

Table 2. The Correlations among Different Search Methods When Using the Centroid Data[a]

	O	I	X	R	P	M	S
O		−0.13	0.02	0.02	−0.45	0.01	0.06
I	±0.16		−0.08	0.03	−0.26	−0.12	−0.01
X	±0.12	±0.28		0.03	−0.01	0.03	−0.04
R	±0.29	±0.19	±0.11		0.04	0.02	0.02
P	±0.12	±0.13	±0.07	±0.42		−0.07	−0.04
M	±0.17	±0.27	±0.27	±0.14	±0.12		−0.01
S	±0.18	±0.25	±0.22	±0.16	±0.17	±0.28	

[a] The upper right triangle of the matrix documents the average pairwise method correlations, while the lower triangle of the matrix documents the standard error associated with each method pair. As before, each method is represented by its first letter in the figure legend.

Table 3. The Correlations among Different Search Methods When Using the Profile Data[a]

	O	I	X	R	P	M	S
O		−0.06	0.08	−0.04	−0.22	0.07	0.08
I	±0.29		−0.01	0.01	−0.12	−0.18	−0.09
X	±0.30	±0.42		0.08	0.03	0.10	0.13
R	±0.25	±0.22	±0.19		−0.20	0.06	0.07
P	±0.26	±0.19	±0.1	±0.22		−0.04	−0.01
M	±0.35	±0.41	±0.45	±0.28	±0.20		0.04
S	±0.31	±0.37	±0.37	±0.28	±0.20	±0.42	

[a] The upper right triangle of the matrix documents the average pairwise method correlations, while the lower triangle of the matrix documents the standard error associated with each method pair. As before, each method is represented by its first letter in the figure legend. Note that the size of the average correlations are in agreement within the standard errors across different data types, see Table 2.

Case B: Apparently, when $L \rightarrow \infty$, the series $\sum_{n=0}^{L-1} \{[\ln(1/\tau)]^n/n!\}$ becomes $\exp[\ln(1/\tau)] = 1/\tau$. That is, if one were to hold $\tau \equiv \prod_{i=1}^{L} p_i$ constant while letting L approach infinity, our formula will render the final combined P-value to be $\tau \times 1/\tau = 1$. This result, however, is exactly what one would have anticipated. Recall that each P-value is in the range (0, 1]. Holding τ constant while letting $L \rightarrow \infty$ is possible only if the following

two conditions hold true: (a) only finite number of methods report P-values smaller than 1, and (b) infinitely many methods report P-values to be 1. Violation of these two conditions will result in diminishing τ as L increases to infinity. In the context of peptide identifications, this corresponds to the scenario that we have infinitely many methods reporting a peptide to be totally insignificant (with P-value 1) while only a finite number of methods reporting that peptide to be potentially significant (with P-value less than 1). Therefore, the final ruling that the peptide considered is totally insignificant (with P-value 1) is natural and intuitively correct. In case C, we address a more subtle large L limit.

Case C: For a more realistic case, let us assum that the geometric average of the P-values reported stays a constant less than 1 in the limit $L \rightarrow \infty$. That is, we assume that the limit

$$p \equiv \lim_{L\rightarrow\infty}\left[\prod_{i=1}^{L} p_i\right]^{1/L}$$

exists and with $p < 1$. Note that in case B, one would have obtained $p = 1$ as the limit.

We then have $\tau = p^L$ and $\ln(1/\tau) = L\ln(1/p)$. Because $\tau \rightarrow 0$ as $L \rightarrow \infty$ here, one should not use the series expansion to investigate the asymptotic behavior, instead we use the integral in eq 6

$$\begin{aligned} F(\tau) &= \int_{\ln(1/\tau)}^{\infty} e^{-t}\frac{t^{L-1}}{(L-1)!}\,\mathrm{d}t \\ &= \int_0^{\infty} e^{-t}\frac{t^{L-1}}{(L-1)!}\,\mathrm{d}t - \int_0^{\ln(1/\tau)} e^{-t}\frac{t^{L-1}}{(L-1)!}\,\mathrm{d}t \\ &= 1 - \frac{1}{(L-1)!}\int_0^{L\ln(1/p)} e^{-t}t^{L-1}\,\mathrm{d}t \end{aligned}$$

Note that as long as $\ln(1/p) > 1$, the saddle point of the integrand is enclosed in the integral range, and thus, the remaining integral will have value close to $(L-1)!/(L-1)! = 1$ making the final $F(\tau) \rightarrow 0$. In other words, if there are infinitely many methods reporting P-values whose geometric mean is smaller than $1/e \approx 0.36788$, then the combined asymptotic P-value will become 0 in our formulation. However, if the geometric mean of the P-values remains larger than $1/e$, the integral in the above expression diminishes as $L \rightarrow \infty$ and $F(\tau) \rightarrow 1$ similar to case B. That is, as $L \rightarrow \infty$, depending on the geometric mean of the P-values, the combined P-value reaches either 1 or 0 in the manner of a step function. In other words, the probability of obtaining a combined P-value other than 0 or 1 diminishes as L increases, implying that one should expect a crispier combined P-value (or a better separation of true from false positives) if for each candidate peptide there are more and more *independent* and *accurate* P-values available.

Acknowledgment. We thank Dr. Ning Zhang for providing the data in ref 19. We thank the administrative group of the NIH biowulf clusters, where all the computational tasks were carried out. G.A. thanks Susan Chacko for her help with running Mascot. This work was supported by the Intramural Research Program of the National Library of Medicine and the National Heart, Lung, and Blood Institute at the National Institutes of Health. Funding to pay the Open Access publication charges for this article was provided by the NIH.

Supporting Information Available: A large number of ROC curves that convey similar information of that exhibited in the plots of the main text are shown in the Supporting Information. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Kapp, E. A.; Schütz, F.; Connolly, L. M.; Chakel, J. A.; Meza, J. E.; Miller, C. A.; Fenyo, D.; Eng, J. K.; Adkins, J. N.; Omenn, G. S.; Simpson, R. J. An evaluation, comparison, and accurate benchmarking of several publicly available ms/ms search algorithms: sensitivity and specificity analysis. *Proteomics* **2005**, *5*, 3475–3490.

(2) Boutilier, K.; Ross, M.; Podtelejnikov, A. V.; Orsi, C.; Taylor, R.; Taylor, P.; Figeys, D. Comparison of different search engines using validated MS/MS test datasets. *Anal. Chim. Acta* **2005**, *534*, 11–20.

(3) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; R., A. Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search. *Anal. Chem.* **2002**, *74*, 5383–5392.

(4) Alves, G.; Ogurtsov, A. Y.; Wu, W. W.; Wang, G.; Shen, R.-F.; Yu, Y.-K. Calibrating E-values for MS^2 library search methods. *Biol. Direct* **2007**, *2*, 26.

(5) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **1995**, *B57*, 289–300.

(6) Finner, H.; Roters, M. Multiple hypotheses testing and expected number of type I errors. *Ann. Stat.* **2002**, *30*, 220–238.

(7) Ge, Y. C.; Dudoit, S.; Speed, T. P. Resampling-based multiple testing for microarray data analysis. *Test* **2003**, *12*, 1–77.

(8) Eng, J. K.; McCormack, A. L.; Yates III, 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.

(9) Zhang, N.; Aebersold, R.; Schwikowski, B. A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* **2002**, *2*, 1406–1412.

(10) Tanner, S.; Shu, H.; Frank, A.; Wang, L.-C.; Zandi, E.; Mumby, M.; A, P. P.; Bafna, V. Inspect: Identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* **2005**, *77*, 4629–4639.

(11) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence database using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551–3567.

(12) Craig, R.; Beavis, R. C. Tandem: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466–1467.

(13) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; W, S.; Bryant, S. H. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, *3*, 958–964.

(14) Alves, G.; Ogurtsov, A. Y.; Yu, Y.-K. RAId_DbS: Peptide identification using database searches with realistic statistics. *Biol. Direct* **2007**, *2*, 25.

(15) Fisher, R. A. *Statistical Methods for Research Workers*, 2nd ed.; Hafner: New York, NY, 1958.

(16) Elston, R. C. On fisher's method of combining *p* -values. *Biom. J.* **1991**, *33*, 339–345.

(17) Bailey, T. L.; Gribskov, M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **1998**, *14*, 48–54.

(18) Yu, Y.-K.; Gertz, E.; Agarwala, R.; Schäffer, A.; Altschul, S. Retrieval accuracy, statistical significance and compositional similarity in protein sequence database searches. *Nucleic Acids Res.* **2006**, *34*, 5966–5973.

(19) Keller, A.; Samuel, P.; Nesvizhskii, A. I.; Stolyar, S.; Goodlett, D. R.; Kolker, E. Experimental protein mixture for validating tandem mass spectral analysis. *OMICS* **2002**, *6*, 207–212.

PR700798H