Nonlinear Fitting Method for Determining Local False Discovery Rates from Decoy Database Searches

Wilfred H. Tang, Ignat V. Shilov, and Sean L. Seymour*

Applied Biosystems| MDS Analytical Technologies, 850 Lincoln Centre Drive, Foster City, California 94404

Received August 2, 2007

False discovery rate (FDR) analyses of protein and peptide identification results using decoy database searching conventionally report aggregate or global FDRs for a whole set of identifications, which are often not very informative about the error rates of individual members in the set. We describe a nonlinear curve fitting method for calculating the local FDR, which estimates the chance that an individual protein (or peptide) is incorrect, and present a simple tool that implements this analysis. The goal of this method is to offer a simple extension to the now commonplace decoy database searching, providing additional valuable information.

Keywords: decoy database • false discovery rate • instantaneous error rate • nonlinear fitting • local false discovery rate

Introduction

Tandem mass spectrometry (MS/MS) has attained considerable importance as a tool for high-throughput protein identification. From MS/MS data, peptide identifications (and by inference, protein identifications) are typically made using database search software.[1] A significant challenge in effectively using database search software is figuring out the correct identifications while maintaining control over false positive identifications.[2,3] Many database search engines provide an estimate of the probability of correctness of putative identifications. It is desirable, however, to also have a method for independently assessing the correctness of results. This type of independent assessment is typically accomplished using a technique known as decoy database searching, in which the database search engine is presented with potential answers that are known to be incorrect, typically generated by some variant of reversing or randomizing true sequences. The rate of occurrence of these known-incorrect answers in the output given by the search software is then used to estimate the false discovery rate (FDR).[4]

To date, the FDRs calculated by decoy database searching are aggregate or global FDRs, in which the FDR applies to the entire set of proteins (or peptides). In this paper, we describe a method for calculating the instantaneous or local FDR,[5,6]which measures the FDR of an individual protein (or peptide) — in other words, how likely a specific protein or peptide is incorrect, rather than the overall error rate for the set of proteins or peptides it is a member of (typically a set defined by some threshold value). For many experiments, the local FDR is the more useful and desirable way of describing error. The local FDR calculation described here can be performed by the average proteomics scientist using general data analysis software such as Excel or KaleidaGraph. The method is simple, transparent, and independent of the search engine, all factors that contribute to the current popularity of decoy database searching for determining global FDRs.

Although the method can be done without special software, we also present here a software tool that provides an implementation of the method. Proteomics System Performance Evaluation Pipeline (PSPEP) software is an add-on to ProteinPilot Software and automatically runs this independent error analysis after the search has completed. The Materials and Methods section describes our preferred methodology, as implemented by PSPEP, whereas the Results and Discussion section explains the reasons for our choices as well as potential generalizations.

Materials and Methods

False Discovery Rate Analysis. The decoy database is constructed by reversing all the protein sequences in the database of interest (e.g., UniProtKB/Swiss-Prot) and appending these reversed sequences to the original database. The list of identifications obtained by searching of the concatenated target + decoy database is ordered from best to worst (e.g., by sorting from highest confidence to lowest confidence or sorting from highest score to lowest score). As the list is traversed from best to worst, the number of hits to the decoy (reverse) portion is accumulated in the variable D, while the total number of identifications traversed so far is accumulated in the variable N. Note that for each item on the list, only one answer (the best answer) is considered. If there are multiple best answers—for example, multiple proteins tied for best in the same protein group, or multiple peptide hypotheses tied for best for the same MS/MS spectrum—*one* best answer is chosen arbitrarily. This prevents "cheating" by a search engine by designating multiple possibilities as answers without justification. To consider an absurd extreme, if the search engine designates everything in the FastA as equivalent best answers, one of those answers is certainly correct (assuming that the answer is actually in the

* To whom correspondence should be addressed. E-mail seymousl@appliedbiosystems.com. Phone 510-708-9483. Fax 650-638-6223.

10.1021/pr070492f CCC: $40.75

Published on Web 08/14/2008

FastA) but it would be ridiculous to give the search engine credit for getting the correct answer in such a case. For peptide identifications, an additional complication must be considered because it is possible for a peptide sequence to occur in both forward and reversed proteins. This is particularly prevalent for short peptides. Our approach is to split the classification, incrementing the decoy count D by 1/2 and the total count N by 1.

The global and local FDRs are calculated as follows:

$$\mathrm{FDR}_{\mathrm{Global}} = 2\frac{D}{N} \tag{1}$$

$$\mathrm{FDR}_{\mathrm{Local}} = 2\frac{dD}{dN} \tag{2}$$

The assumption underlying these equations is that for each decoy (reverse) hit, there exists (on average) a corresponding forward hit that is also incorrect − in other words, the number of incorrect identifications is twice the number of decoy hits (on average).[7,8]

The calculation of the global FDR given by eq 1 is straightforward, but the calculation of the local FDR given by eq 2 requires an additional step. The plot of D vs N is approximated using the following model:

$$D = c\left(\frac{\ln(e^{b(N-a)}+1) - \ln(e^{-ba}+1)}{b}\right) \tag{3}$$

There is no particular physical significance to this equation, other than empirical trial and error found this function to be a reasonably good model of the observed traces of D vs N, thereby providing data smoothing as well as an analytical approximation for calculating the derivative. The parameters a, b, and c are determined using least-squares fitting—that is, a, b, and c are optimized to minimize the χ^2

$$\chi^2 = \sum_i \frac{(D_{i,\mathrm{observed}} - D_{i,\mathrm{model}})^2}{\sigma_i^2} \tag{4}$$

where, for each data point i, $D_{i,\mathrm{observed}}$ is the observed number of decoy hits, $D_{i,\mathrm{model}}$ is calculated from eq 3, and σ_i is the measurement uncertainty, which we estimate as $\sqrt{D_{i,\mathrm{observed}}}$, based on counting statistics if $D_{i,\mathrm{observed}} > 0$, or set to 0.2 if $D_{i,\mathrm{observed}} = 0$ (an empirical value, necessary to avoid division by zero). We determine values for the parameters a, b, and c using the Levenberg−Marquardt method, which is the standard method for least-squares fitting of nonlinear equations.[9] The fitting is restricted to the region from $N = 0$ to the point where the global FDR reaches 10% and D must exceed 10. The latter condition ensures that there at least *some* decoy hits in the fit, while the former condition ensures that the fit focuses on the region of greatest interest − one rarely cares about FDRs above 10%. From eqs 2 and 3, the local FDR is calculated as:

$$\mathrm{FDR}_{\mathrm{Local}} = 2\frac{dD}{dN} = 2c\left(\frac{e^{b(N-a)}}{e^{b(N-a)+1}}\right) \tag{5}$$

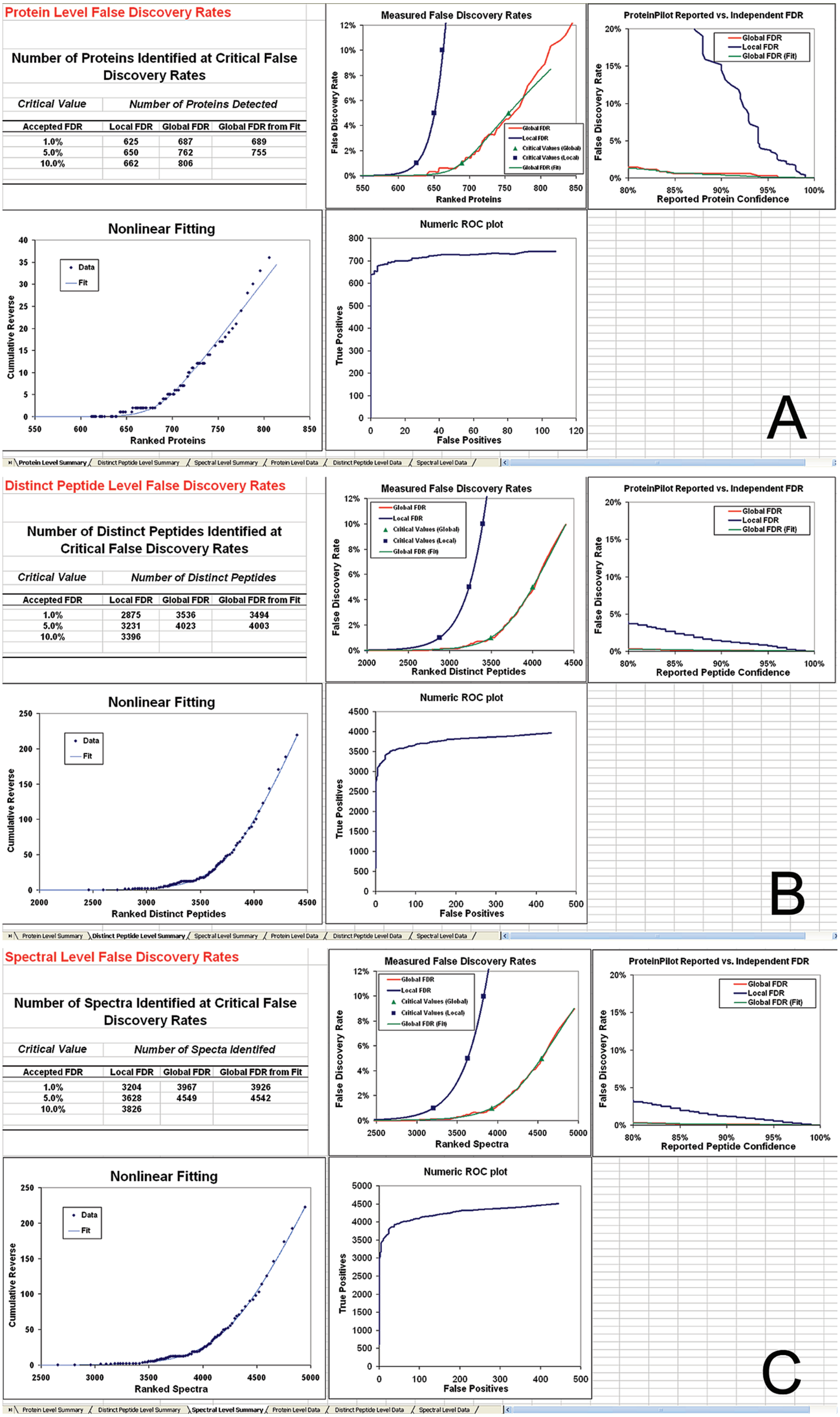
These calculations are implemented by the PSPEP add-on to ProteinPilot software. PSPEP automatically performs the following operations:

1. The concatenated target + decoy database is created from the database of interest and is submitted for use by the search engine, if it does not already exist.
2. After the search engine finishes, false positive analysis is performed on three different levels using the methods described above: the protein level, the spectral level, and the distinct peptide level.
3. The results of the false discovery analysis are placed in a Microsoft Excel template, which displays the results graphically and provides access to the source data for the graphs and tables.

LC−MS/MS Experiment. Twenty-five milliliter cultures of *Escherichia coli* K12 (MG1655) were grown in M9 minimal media supplemented with 2 mM $MgSO_4$, 200 μM $CaCl_2$, 0.002 mg/mL $FeSO_4$, 1 g/L casamino acids, and glucose to a final concentration of 0.2%. Cells were grown to late-exponential phase ($OD_{600} = 0.60$) and harvested by centrifugation at 5000× g for 15 min. The cell pellets were washed with 10 mL 60 mM Tris pH 8.4, 10 mM NaCl, and 1 mM EDTA and repelleted. The cell pellets were resuspended in 750 μL of identical buffer and lysed by 3 × 30 s passes through a bead-beater. The beads were clarified by low speed spin 1000 xg × 5 min, and the cell lysates were clarified by centrifugation at 15,000 xg for 12 min. Total protein was quantified using a MicroBCA protein assay kit against a standard curve of BSA (Pierce). Two microliters of sample was diluted with 18 μL of lysis buffer prior to BCA analysis (Molecular Devices plate reader). The R^2 of the calibration curve was 1.0 and the concentration of the final lysate (approximately 700 μL) was 2.90 mg/mL ($\pm 1\sigma = 0.3$, CV = 10%). 500 μg of total protein was resuspended in 50% 2,2,2-Trifuoroethanol (TFE) and 5 mM DTT and allowed to sit for 45 min at 60 °C. The sample was then alkylated using 20 mM iodoacetamide (IAM) for 1 h at room temperature (RT) in the dark. Excess IAM was quenched using an additional 5 mM DTT for 1 h at RT in the dark. The sample was prepared for digestion by diluting 10-fold with 25 mM ammonium bicarbonate buffer (which brings the TFE concentration down to 5% and adjusts the pH to ~8). The sample was digested by adding 5 μg of trypsin and allowing the reaction to go overnight (~15 h). The digestion reaction was quenched by adding 4 μL of neat formic acid. The sample was analyzed on a QSTAR Elite LC−MS/MS System (LC Packings PepMap 100, 3 μm, 75 μm × 150 mm, 90 min gradient; IDA settings: 1 survey scan of 0.25 s followed by 6 product ion scans, MS/MS threshold = 20 counts, Smart Exit = 2, Smart CE on). The MS/MS data were analyzed using the Paragon Algorithm[10] of ProteinPilot Software version 2.0.1 with settings: *Sample type*: Identification; *Cys Alkylation*: Iodoacetamide; *Digestion*: Trypsin; *Instrument*: QSTAR ESI; *Species*: (no filter applied); *Search Effort*: Thorough; *ID focus*: biological modifications and amino acid substitutions; *Database*: UniprotKB/Swiss-Prot of July 12, 2007; and the *Detected Protein Threshold* was set very low (10% confidence) in order to include a sufficient number of wrong answers to enable the curve fitting.

Results and Discussion

Results. Figure 1 shows an example of the output generated by PSPEP. At each level of false positive analysis (protein (Figure 1A), distinct peptide (Figure 1B), and spectral level (Figure 1C)), the following five elements are displayed. (1) The *Summary Table* provides the yield of identifications at several critical FDRs, and this is done using both the global and local method for calculating FDRs. The global FDRs are calculated two different ways: the values reported in the *Global FDR* column are based on the observed number of decoy hits, whereas the values reported in the *Global FDR from Fit* column are based on the fitted model. [Strictly speaking, the observed decoy counts are used to calculate a global FDR profile. This profile



Protein Level False Discovery Rates

Number of Proteins Identified at Critical False Discovery Rates

Critical Value	*Number of Proteins Detected*		
Accepted FDR	Local FDR	Global FDR	Global FDR from Fit
1.0%	625	687	689
5.0%	650	762	755
10.0%	662	806	

Distinct Peptide Level False Discovery Rates

Number of Distinct Peptides Identified at Critical False Discovery Rates

Critical Value	*Number of Distinct Peptides*		
Accepted FDR	Local FDR	Global FDR	Global FDR from Fit
1.0%	2875	3536	3494
5.0%	3231	4023	4003
10.0%	3396		

Spectral Level False Discovery Rates

Number of Spectra Identified at Critical False Discovery Rates

Critical Value	*Number of Specta Identifed*		
Accepted FDR	Local FDR	Global FDR	Global FDR from Fit
1.0%	3204	3967	3926
5.0%	3628	4549	4542
10.0%	3826		

Figure 1. Example Excel output generated by PSPEP. (A–C) Example results of the three levels of FDR analysis as output into Excel by PSPEP: (A) analysis at the protein level, (B) analysis at the distinct peptides level where only the highest confidence instance of each distinct peptide is considered, and (C) FDR analysis at the spectral level. The full Excel file is included in Supporting Information.

Table 1. Characteristics of the Function Fitted to Derive FDRs

model characteristic (see eqs 3 and 5)	corresponding expected FDR behavior
For $N = 0$, $D = 0$	Starting point.
As N increases, dD/dN increases monotonically	The local FDR should increase as the list of identifications is traversed from best to worst.
For small N, $dD/dN \rightarrow 0$ (if $ab \gg 0$)	At the beginning of the list, where the best identifications are located, the local FDR should be low (practically no incorrect answers).
As $N \rightarrow \infty$, $dD/dN \rightarrow c$	At the end of the list, where the worst identifications are located, the local FDR should approach 100% (all answers are incorrect). This implies that c should be equal to 1/2 for equal sized target and decoy components. Note, however, that if there are a nontrivial number of very poorly fragmenting peptides with very poor rankings, the local FDR may never actually reach 100%.
Parameter a controls where the slope dD/dN changes from $\sim$0 toward $\sim c$. Initial value for a in Levenberg–Marquardt fitting can be estimated either as the x-axis value in the curving region or as the approximate expected number of true positives in the result set.	
Parameter b controls how quickly this change occurs. Initial value for b in Levenberg–Marquardt fitting can be estimated as 0.001–0.01.	

has an overall trend of increasing global FDR as N increases but, counterintuitively, can have slightly decreasing global FDR over short scales. Thus, we perform a minor transformation in order to make the global FDR profile monotonically increasing (see the q-value method described by Storey and Tibshirani[11]), and the values reported in the *Global FDR* column are obtained from this transformed global FDR profile.] (2) The *Measured False Discovery Rates Graph* shows the yield of identifications at all FDRs, and the square or triangle symbols represent the data displayed numerically in the *Summary Table*. (3) The *ProteinPilot Reported vs Independent FDR Graph* compares the confidences reported by the Paragon algorithm to the FDRs determined independently by PSPEP. The confidences reported by the Paragon algorithm are intended to be local measures (as opposed to global measures). Thus, ideally, the blue local line on this graph would be a straight line from the bottom right to the top left. For example, in Figure 1A, the blue line conforms to this optimal diagonal quite well, indicating the reported proteins confidences were very accurate, while this same plot for the spectral level in Figure 1C shows the blue line well below the center diagonal, meaning that the peptide confidences reported by the Paragon algorithm were too conservative. This demonstrates the value of doing the independent assessment using PSPEP. (4) The *Nonlinear Fitting* graph shows how well the fit using eq 3 actually works. By inspecting this graph, the scientist can assess whether or not the nonlinear curve fit to the actual data is a good fit. If the fit does not model the data well, FDR rates derived from it should not be considered reliable. This includes both the *Local FDR* and the *Global FDR from Fit* columns. (5) The *Numeric ROC Plot* (Receiver Operating Characteristic) shows the tradeoff between the number of true positive (correct) identifications that can be obtained versus the number of false positives that must be tolerated. The ROC curve is calculated using the assumption that one forward answer is incorrect for each observed reversed answer.

Local FDR Calculation. Calculation of the local FDR requires calculation of the derivative dD/dN (see eq 2). Calculating this derivative is challenging because the variable D effectively has a stochastic component. On average, over a large scale, the expected value of D is one-half the number of false positives, and D appears to be smooth. Over small scales, however, there can be significant deviations from this average behavior. Consequently, estimating the derivative dD/dN using $\Delta D/\Delta N$ (where ΔD and ΔN must be calculated over very short distances in order to get meaningful results) does not work well. This difficulty is overcome by fitting the entire set of all (D, N) data pairs to a smooth, analytical curve (eq 3). An examination of the functional form given by eq 3 reveals that this model has suitable characteristics, as listed in Table 1. In addition, we have tested this model on a variety of data sets and have found empirically that this model appears to work reasonably well on a diversity of data types. The *Nonlinear Fitting* graph in the PSPEP output allows the scientist to examine how well this actually works on specific data sets.

In examining the fitting quality on many data sets, we observed that a naïve fit over the entire D vs N trace sometimes resulted in poor fitting of the key region where the yields for critical error rates, like 1 and 5%, are derived. Since the purpose of the fitting function is to provide smoothing of the observed data, thereby enabling reliable slope estimation, our implementation evolved to emphasize optimal smoothing of the key bending region. This was accomplished by restricting the fit to a subset of the D vs N trace, as described in further detail in the Materials and Methods section above.

In our experience, outright failures in fitting usually arise from one of two possible reasons. (1) The number of identifications is small, meaning that there are too few data points to allow for an adequate fit. An example of this case is the protein level false positive analysis for a sample containing few proteins. (2) The number of decoy hits found by the decoy database search is zero (or small)—if there are no decoy hits, it is simply impossible to infer meaningful error rates. This problem can usually be solved by repeating the decoy database search with a lower threshold for reporting identifications. Like most techniques for FDR assessment, this nonlinear fitting approach is generally intended for large sets of results—typically sets containing greater than 100 true proteins, peptides, or spectra. However, we have seen cases where it performs reasonably at the protein level for files known to contain only 20−50 proteins. It is critical to visually inspect the fit or compute fit quality metrics such as R^2 to assess whether error rates are likely to be reasonable.

Decoy Database Construction. We favor using reversed protein sequences rather than randomized protein sequences as the decoys, simply because the latter brings with it an additional complication. Because of both homology (biological redundancy) and duplicate or near duplicate entries (informatic redundancy) among the proteins in databases, peptide sequences are frequently found multiple times in the database. Reversing two identical sequences produces the same peptides in both cases, and this preserves the equal size of the decoy portion and the target portion. By contrast, randomizing two identical sequence sections produces two *different* peptides, thereby making the decoy portion of the composite database larger than the target portion, which would require assessment and correction to produce accurate FDRs.

Generalization and Limitations of the Model. With a small extension, it is possible to compensate for different sizes of the target vs decoy portions of the database. Instead of making the assumption that there are two incorrect answers (one target, one decoy) for each decoy answer observed, the FDR calculations in eqs 1 and 2 can be generalized by replacing the fixed scaling factor of 2 in eqs 1 and 2 with the scaling factor *s* to yield the generalized equations:

$$\mathrm{FDR}_{\mathrm{Global}} = s\frac{D}{N} \tag{6}$$

$$\mathrm{FDR}_{\mathrm{Local}} = s\frac{dD}{dN} = s\left(\frac{s^{b(N-a)}}{e^{b(N-a)}+1}\right)c \tag{7}$$

where *s* is the number of incorrect answers (on average) implied by each decoy answer observed. This generalization enables the FDR calculations to be applied to arbitrary decoy strategies. For example, the decoy portion of the database can be constructed to be much larger in size than the target portion of the database; such a strategy would yield more accurate estimates for low FDR values. Another potential use of this generalization would be to compensate for the difference in effective size between the target and decoy portions of the database when the decoy proteins are generated by randomization.

In some cases, it may be possible to estimate the scaling factor *s* from the fit to eq 3. As *N* becomes large, the slope of the function asymptotically approaches *c*, and it may be possible to estimate *s* as 1/*c*, thus yielding:

$$\mathrm{FDR}_{\mathrm{Local}} = s\frac{dD}{dN} = \frac{1}{c}\frac{dD}{dN} = \left(\frac{1}{c}\right)\left(\frac{e^{b(N-a)}}{e^{b(N-a)}+1}\right)c = \frac{e^{b(N-a)}}{e^{b(N-a)}+1} \tag{8}$$

In practice, the assumptions underlying eq 8 may not always hold true. Estimating *s* as 1/*c* is valid only if *all* of the lowest-ranking answers (corresponding to the region where *N* becomes large) are truly wrong answers. It is not uncommon, however, to have a nontrivial number of peptides with very poor fragmentation, resulting in a nontrivial number of correct answers interspersed among the lowest-ranking wrong answers. In such cases, estimating *s* as 1/*c* is clearly not valid. Another potential problem is that frequently there is insufficient data in the asymptotic linear region to accurately determine the fit parameter *c*. Because of these risks and because our preference for using reversed decoys over random decoys allows a safe assumption of equal target and decoy components, our preferred implementation in PSPEP assumes equal target and decoy components and uses eqs 1, 2, 3, 4 and 5. Thus, the functional fitting is currently leveraged only as a smoothing mechanism that enables robust estimation of slopes in the key region where the function bends most sharply.

It is also worth noting that, although the local method is presented here in an implementation that uses the concatenated approach (where the target and decoy databases are searched together), it could certainly be extended to alternate approaches. In the most common alternate approach, the decoy database portion is searched separately, in which case the global and local FDRs are calculated as *D*/*T* and *dD*/*dT* respectively, where *D* is the number of hits above a threshold in a search conducted against decoy proteins only and *T* is the number of hits above the same threshold in a separate search against target proteins only. We expect our fitting function to work equally well for smoothing the *D* vs *T* trace, and the parameters obtained by fitting can be substituted into the revised global and FDR expressions to obtain revised equations for this alternate approach. A second alternate approach that we have been exploring is “intraprotein concatenation”, in which each target protein in the database is transformed into a single composite target + decoy protein by concatenating the decoy (reversed) portion to the target. The conventional decoy approaches all implicitly assume that search effort is uniform throughout the database, but this assumption breaks down for search engines that search high-scoring proteins more extensively (via a second pass, for example). Connecting the corresponding target and decoy portions in a single, inseparable protein entry by intraprotein concatenation enables the calculation of accurate FDRs even when the search effort is not uniform. In the intraprotein concatenation approach, the global and local FDRs are unchanged from eqs 1 and 2, the fitting function applies similarly, and thus all the calculations remain unchanged.

Local FDR vs Global FDR. The advantage of the local FDR method over the global method is particularly obvious when considering the protein level. The local FDR measures the error rate for individual proteins, while the global FDR measures the error rate for the entire set of proteins. Knowledge of the FDR enables the scientist to decide which set of proteins to accept for further analysis while maintaining the error rate at a reasonable (ideally, predetermined) level, that is, all proteins above threshold are “accepted” for results interpretation, whereas all proteins below threshold are discarded. We believe that thresholding based on the local FDR is often the better approach because it assures that all reported identifications in our accepted set of proteins have at least some minimal quality. By contrast, if thresholding is based on a global FDR, the quality of the poorest identification in our set of proteins is unknown and can in fact be quite bad. For example, in Figure 1A, thresholding at a global FDR of 1% would lead us to accept proteins 1−687, but note that the local FDR for our poorest identification in our accepted set (protein 687) is about 30%! For the scientist who cares about protein 687 or a protein in this vicinity, it is obvious which type of error estimation is more useful. There is little reassurance in knowing the FDR of the whole set is 1% when the protein you care about has a 30% chance of being wrong. For numerous purposes, for example, controlling what results are published in journals, placed in repositories, or used for subsequent analyses in any experiment, controlling the minimal quality of the worst member in any set would seem preferable to controlling the global error rate of the whole set if it could be done well.

While it has been suggested that global and local error measures can be complementary[12] or even serve roles in different stages of a workflow, we do not find these arguments very compelling. If one wishes to have greater sensitivity (by being more permissive) during the early stages of a study, one could set higher acceptable FDR thresholds with either type of error assessment. The local FDR has the advantage of allowing one to ensure that no accepted results are less likely to be correct than the specified value, and this gives much more control in how one adjusts the balance of sensitivity and specificity concerns at a given stage of work. Any kind of local error rate (regardless of the method presented in this paper) directly indicates a "pain ratio"—the cost of each new additional right answer in terms of wrong answers that will come with it and require manual inspection or simply cause errors. Thus, it is easier to find a point of diminishing returns. By contrast, the global FDR is a cruder tool for assessing the sensitivity vs specificity tradeoff. For example, increasing the global FDR threshold may not bring *any* additional correct answers—it is possible that the global error rate is increasing entirely because of additional wrong answers *only*. Increasing the local FDR threshold, on the other hand, assures that there are some right answers among the additional answers obtained from the more permissive threshold, so long as the local FDR has not yet reached 100%.

Regardless of any of these arguments for which error type is better for setting thresholds, it cannot be argued that having a good estimation of the probability of correctness for each individual identification is not a valuable thing. So long as it can be calculated well, the local error rate estimates seem generally more informative and empowering and are certainly valuable to have, regardless of whether the global or local FDR is used to set the threshold.

The strongest argument we can see to favor using global error estimation over local estimates for setting thresholds is that the uncertainty in determining the former is generally lower. Regardless of the specific method of how one estimates the local error rate, the basic counting statistics involved dictate higher uncertainty compared to determining cumulative totals. For purposes requiring comparison, alignment, or consistency across result sets, global FDRs may be a better choice of threshold method simply because it is possible to more accurately determine the global FDR. A journal or a repository could feel more confident of the consistency in the quality of submitted data if a threshold global FDR for submissions were specified rather than a local FDR threshold.

Fitting Can Improve the Global FDR Estimation Quality. If the global FDR is preferred for a given purpose, the method presented in this paper still has a potential benefit to offer. While the primary purpose of the nonlinear fitting is to enable computation of the local FDR, the fitted model can also be used to improve the estimate of the global FDR by reducing noise. The output of PSPEP includes two types of global FDR estimates in the summary table − both the conventional approach to the global FDR and also the smoothed estimation derived from the nonlinear fitting (using the function value directly, rather than its derivative as with local FDR). The *Measured False Discovery Rates* graph shows the full traces using both of these global FDR calculation methods. In addition to assessing the fitting quality, one can also inspect this graph to decide if the smoothing is beneficial enough to favor this calculation rather than the conventional global calculation approach. This method for estimating the global FDR via fitting may prove particularly useful for estimating low global FDR values where low counts make the uncertainty high.

Three Levels of Analysis. The spectral level analysis (Figure 1C) is included because it is the most conventional measure currently used. Analysis on this level can be useful for comparison of different identification algorithms on the same exact data set. However, spectral level analysis is not useful for comparison across different data sets, such as different experiments or different instruments, because redundant acquisition is rewarded by this measure. That is, replicate acquisitions of a peptide that fragments very well yielding an easy identification all count toward this measure, even though the redundant spectra bring no new information. A more useful analysis for many comparative purposes is the distinct peptide level analysis (Figure 1B). This analysis only considers the highest confidence instance for each distinct peptide molecule. Because repeated acquisitions of the same molecule are not considered, this analysis is well suited for measuring the information yield in any acquired data set. When trying to optimize acquisition methods or compare across different experiments or instruments, this analysis is the key measure. The protein level analysis (Figure 1A) is based on the list of detected proteins, that is, the first representative winner protein from each of the protein groups. Although this level of analysis is included in the PSPEP output, the use of this information for comparative purposes to other methodologies can be problematic.

Is the Protein Level Analysis Valid? In general, interpreting the protein level analysis obtained from decoy database searching requires caution. Decoy database searching is fundamentally incapable of analyzing a list of protein identifications for inappropriate redundancy. A detailed discussion of this topic is beyond the scope of this paper,[13,14] but consider a simple example. Suppose we have a sample containing a single protein, say, bovine serum albumin. A good search engine with appropriate protein inference analysis should return a protein list consisting of a single protein, namely bovine serum albumin. However, many search engines will return a long protein list that contains many proteins homologous to the bovine serum albumin because the reliance on common spectral evidence of these additional proteins is not recognized. In decoy database searching, all of these proteins are in the target portion of the composite target + decoy database and are thus deemed correct despite the fact that this list of homologous proteins is clearly much larger than the true number of species that have been detected.

All Paragon database searches use the Pro Group Algorithm[14,15] to perform rigorous protein inference analysis. This prevents inappropriate protein redundancy from creeping into the list of proteins, and, therefore, the PSPEP protein level analysis of Paragon results is legitimate, and protein level comparison across data sets analyzed this way is valid. Comparison across protein level FDR analyses derived from searches using different identification software that may vary in their degree of rigor in protein inference should be avoided. It is much better to rely on the distinct peptide level when comparing result sets of unknown or differing protein inference rigor.

Where to Set the Threshold? The minimal quality that is acceptable in any type of results is clearly the choice of the scientist, journal, or repository. However, it is worth noting one important aspect of process that is enabled by the method and tool presented here. The number of hits in a set of results is commonly determined prior to decoy analysis by settings in the search engine. Subsequent decoy analysis then yields an arbitrary FDR value. This is counter to the more conventional statistical approach where the acceptable error rate (typically a critical value like 1 or 5%) is decided before experimentation and then the set of results that satisfy this requirement is determined. Because PSPEP calculates FDRs for all thresholds, it is simple to follow this more conventional protocol and report the yield of proteins, spectra, or distinct peptides that correspond to the predetermined critical error rate, as presented in the summary table and graphs on each tab (see Figure 1). This is especially important for comparative assessments of experiments or methodologies where this is very difficult to do rigorously without fixing on a common FDR.

Conclusions

To date, most proteomics researchers have only been able to measure global FDRs using the popular decoy database strategy. The fitting method presented in this paper is a straightforward extension that enables calculation of the local FDR as well. This method can easily be done manually with common data analysis software packages or implemented directly in other tools. Our PSPEP tool, which calculates both global and local FDRs at three levels of analysis, is a simple implementation of this method.

Regardless of the arguments for the relative merits of global vs local FDR, the local FDR provides valuable information. Most importantly, the local FDR is a direct measure of the cost of each new additional right answer in terms of wrong answers that will come with it.

In the future, a finalized HUPO Proteomics Standards Initiative AnalysisXML results format[16] would enable us to generalize PSPEP to work on any database search engine that writes results into this standardized format. We note, however, that the protein level analysis given by PSPEP would be invalid for results from software or analysis workflows that do not include rigorous protein inference. Similarly, because the spectral level rewards redundant acquisition, it should only be used for comparison of different analysis tools on the same data set. The distinct peptide level analysis is the best assessor of information yield and the safest and most robust measure for comparative purposes.

Acknowledgment. We thank Matt Champion and Jose Meza for providing the *E. coli* data.

Supporting Information Available: The PSPEP output example Excel file shown in Figure 1 is included in Supporting Information, along with the input data for the search in both raw (.wiff) format and in Mascot Generic Format (.mgf) peak list representation generated using ProteinPilot software. Also included is a study that explores the validity of the nonlinear fitting method. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Marcotte, E. M. *Nat. Biotechnol.* **2007**, *25*, 755–757.
(2) Carr, S.; Aebersold, R.; Baldwin, M.; Burlingame, A.; Clauser, K.; Nesvizhskii, A. The need for guidelines in publication of peptide and protein identification data. *Mol. Cell. Proteomics* **2004**, *3*, 531–533.
(3) Bradshaw, R. A.; Burlingame, A. L.; Carr, S.; Aebersold, R. Reporting protein identification data: the next generation of guidelines. *Mol. Cell. Proteomics* **2006**, *5*, 787–788.
(4) Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. Evaluation of Multidimensional Chromatography Coupled with Tandem Mass Spectrometry (LC/LC-MS/MS) for Large-Scale Protein Analysis: The Yeast Proteome. *J. Proteome Res.* **2003**, *2* (1), 43–50.
(5) Efron, B. Size, power and false discovery rates. *Ann. Statist.* **2007**, *35* (4), 1351–1377.
(6) Aubert, J.; Bar-Hen, A.; Daudin, J-J.; Robin, S. Determination of the differentially expressed genes in microarray experiments using local FDR. *BMC Bioinformatics* **2004**, *5*, 125.
(7) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4*, 207–214.
(8) Beausoleil, S. A.; Villén, J.; Gerber, S. A. J. R.; Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **2006**, *24*, 1285–1292.
(9) Press, W. H.; Teukolsky, S. A. *Numerical Recipes in C: The Art of Scientific Computing*; Cambridge University Press: New York, 1992.
(10) Shilov, I. V.; Seymour, S. L.; Patel, A. A.; Loboda, A.; Tang, W. H.; Keating, S. P.; Hunter, C. L.; Nuwaysir, L. M.; Schaeffer, D. A. The Paragon Algorithm: A Next Generation Search Engine that Uses Sequence Temperature Values and Feature Probabilities to Identify Peptides from Tandem Mass Spectra. *Mol. Cell. Proteomics* **2007**, *6*, 1638–1655.
(11) Storey, J. D.; Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (16), 9440–9445.
(12) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin. *J. Proteome Res.* **2008**, *7*, 40–44.
(13) Tang, W. H.; Seymour, S. L.; Patel, A. A. False positive assessment in database search. In *54th ASMS Conference on Mass Spectrometry Proceedings, May 28–June 1, 2006, Seattle, WA*; American Society for Mass Spectrometry: Santa Fe, NM, 2006; Poster TP647.
(14) Seymour, S. L.; Loboda, A.; Tang, W. H.; Patel, A. A.; Shilov, I. V.; Schaeffer, D. A. The Pro Group Algorithm: A rigorous approach to protein inference and quantitation, in preparation.
(15) Seymour, S. L.; Loboda, A.; Tang, W. H.; Nimkar, S.; Schaeffer, D. A. A new protein identification software analysis tool to group proteins and assemble and view results. In *Proceedings of the 52nd ASMS Conference on Mass Spectrometry and Allied Topics, Nashville, TN, May 23–27, 2004*, American Society for Mass Spectrometry: Santa Fe, NM, 2006; Poster ThPA 002.
(16) Proteomics working group of the HUPO Proteomics Standards Initiative (PSI). http://www.psidev.info/index.php?q=node/105#analysisXML.

PR070492F