perspectives Journal of proteome research

Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases

Lukas Käll,[†] John D. Storey,[†,‡] Michael J. MacCoss,[†] and William Stafford Noble*[,†,§]

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

Received September 18, 2007

Automated methods for assigning peptides to observed tandem mass spectra typically return a list of peptide−spectrum matches, ranked according to an arbitrary score. In this article, we describe methods for converting these arbitrary scores into more useful statistical significance measures. These methods employ a decoy sequence database as a model of the null hypothesis, and use false discovery rate (FDR) analysis to correct for multiple testing. We first describe a simple FDR inference method and then describe how estimating and taking into account the percentage of incorrectly identified spectra in the entire data set can lead to increased statistical power.

Keywords: *q*-value • decoy database • false discovery rate • statistical significance • peptide identification

Introduction

The core problem in the analysis of tandem mass spectra is to identify the peptide that gave rise to an observed fragmentation spectrum. The most commonly used tools for solving this problem, such as SEQUEST,[1] Mascot,[2] or X!Tandem,[3] search a given sequence database for the peptide whose theoretical spectrum best matches the observed spectrum. The output of this stage of the analysis is a collection of peptide−spectrum matches (PSMs), each with an associated score (Table 1). The natural subsequent question is, "Which of these PSMs are correct?"

While these algorithms are very powerful, the problem is that there is substantial overlap between the scores for correct and incorrect peptide identifications. This limits the identification to either eliminating a large number of true positive identifications to minimize the false positives or allowing a large number of false positive identifications to maximize the number of true positive identifications.

The purpose of this article is to describe how well-established statistical methods for significance analysis can be applied to peptides identified via tandem mass spectrometry. We assume that we have a peptide identification method that takes as input a set of spectra and a protein sequence database and produces as output a list of PSMs ranked by some score. The score could be a cross-correlation, a probability, or any arbitrary similarity measure. Our goal is to convert these scores into a more useful set of significance measures.

As an example, we use a collection of 34 499 doubly charged tandem mass spectra generated from a yeast whole cell lysate.[4] These spectra were searched against the predicted yeast open reading frames using SEQUEST, and the resulting PSMs were ranked according to the SEQUEST cross-correlation score (XCorr). Figure 1 shows the number of peptide−spectrum matches exceeding a given XCorr threshold.

Table 1. Terminology

PSM	A peptide−spectrum match, with an associated score
target PSM	A PSM created by searching the original peptide database
decoy database	A shuffled or reversed version of the peptide database
decoy PSM	A PSM created by searching a decoy peptide database
accepted PSM	A PSM whose score is above some user-defined threshold
correct PSM	A PSM whose peptide corresponds to the actual peptide that generated the observed spectrum
PIT	Percentage of target PSMs that are incorrect (also known as π_0[5] or p_0[6]
FDR	False discovery rate—the percentage of accepted PSMs that are incorrect

The Decoy Database As a Model of the Null Hypothesis

The most commonly used significance measure in statistics is the *p*-value. Defining "*p*-value" requires that we first define the notion of a *null hypothesis*. Put simply, the null hypothesis is the condition that we are not interested in. For example, when we are assigning a significance measure to a match between a peptide sequence and a tandem mass spectrum, the null hypothesis is that the peptide was not identified by the mass spectrometer. The *p*-value is then defined as the probability of obtaining a result at least as extreme as the observation at hand, assuming the null hypothesis is correct. Therefore, a low *p*-value means that the probability is small that these data would occur by chance when the null hypothesis is true.

* To whom correspondence should be addressed. E-mail: noble@gs.washington.edu.
[†] Department of Genome Sciences, University of Washington.
[‡] Department of Biostatistics, University of Washington.
[§] Department of Computer Science and Engineering, University of Washington.

10.1021/pr700600n CCC: $40.75

Published on Web 12/08/2007

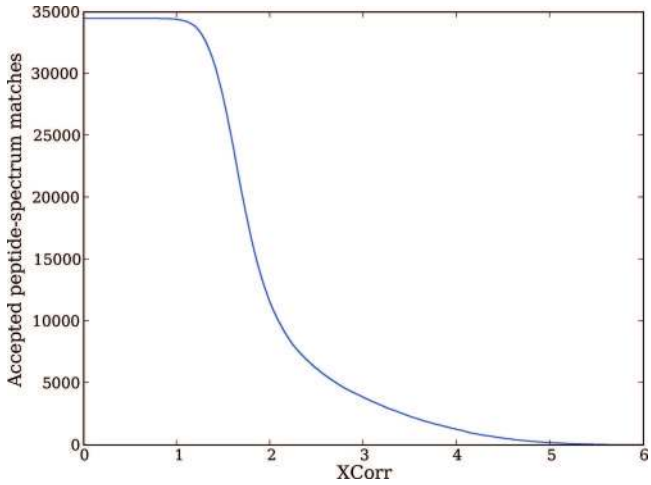

Figure 1. Identifying peptides using SEQUEST's XCorr. The number of peptide−spectrum matches exceeding the XCorr threshold. A collection of 34 499 2+ charged tandem mass spectra derived from a yeast whole-cell lysate was searched against the predict yeast open reading frames using SEQUEST.

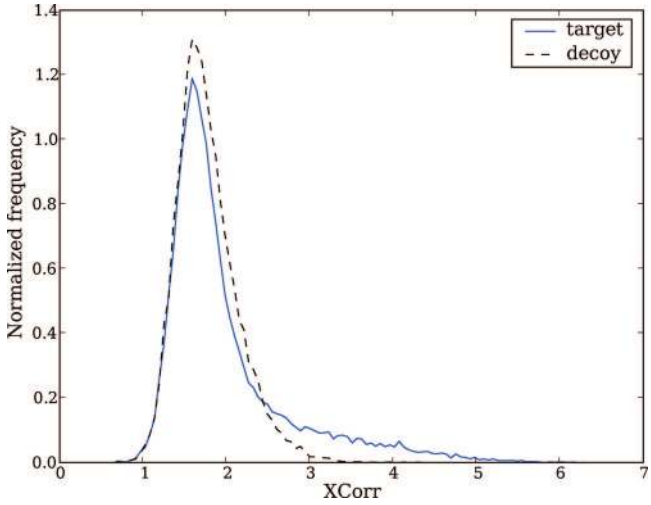

Figure 2. Distribution of XCorr values for target and decoy PSMs. The distribution of XCorr values for the target PSMs (solid line) and decoy PSMs (dashed line).

Now consider that we want to compute *p*-values for our collection of 34 499 ranked PSMs. A commonly used model of the null hypothesis is to search the original set of spectra against a *decoy database*. A decoy database is a database of amino acid sequences that is derived from the original protein database (called the *target database*) by reversing the target sequences,[7] shuffling the target sequences,[8] or generating the decoy sequences at random using a Markov model with parameters derived from the target sequences.[9] There is no clear consensus in the literature as to which method for generating a decoy database is best. For our purposes, what matters is that the decoy database contains peptide-like amino acid sequences that are not in the target database. Therefore, if we search against the decoy database, we can be quite sure that the resulting protein identification is incorrect; that is, the identification is an instance of the null hypothesis being correct. Figure 2 shows the distributions of XCorr values for PSMs derived from the target and decoy databases. The distributions are similar in shape, except that the distribution of target PSM scores has a heavier tail to the right. If the distribution of scores generated by searching the decoy database is an accurate representation of the null distribution, then this tail to the right should reflect the scores corresponding to correct PSMs.

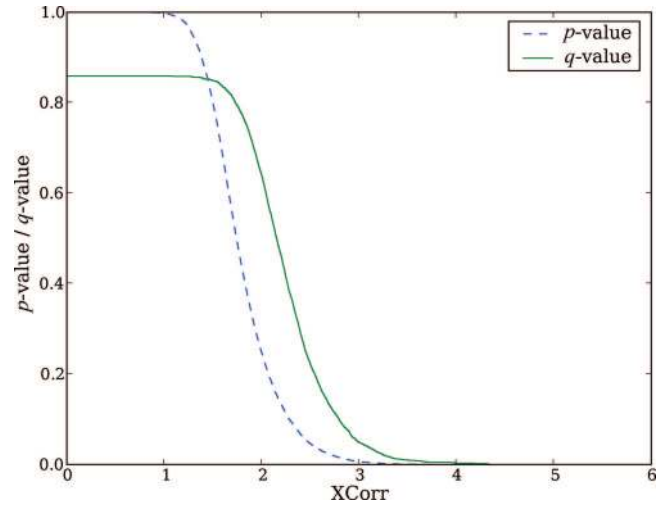

Figure 3. Mapping from XCorr to *p*- and *q*-values. A plot of the *p*-value (dashed line) and *q*-value (solid line) as a function of XCorr.

Once the spectra have been searched against the target and the decoy databases, computing *p*-values is straightforward. For a given target PSM with score *s*, we simply compute the percentage of decoy PSMs that receive score *s* or higher. In our example, a target PSM with XCorr 3.0 is assigned a *p*-value of 0.0063 because 219 out of 34 492 decoy PSMs receive scores greater than or equal to 3.0. The dashed line in Figure 3 plots the *p*-value as a function of XCorr. At a *p*-value threshold of 0.01, we accept 4190 PSMs. When a *p*-value threshold of 0.01 is used, there is a 1% chance that a null peptide−spectrum match will be called correct. In practice, the decoy database is usually the same size as the target database; however, this is not necessary. Using a larger decoy database leads to more accurate *p*-value estimates at the expense of more computation.

Using a good model of the null distribution is extremely important. In general, if the null is inaccurate, for example, if the data contains dependencies that are not taken into account in the null, then all of the resulting significance estimates will be inaccurate. Furthermore, if a perfect null model is unavailable, then it is preferable to use a conservative null model. Such a model will yield *conservative* estimates of significance, meaning that a *p*-value of 0.01 may actually mean that there is a *less than* 1% chance that a null PSM will be called correct. We discuss below a technique for evaluating the accuracy of a given null model, and we provide evidence that the simple null we use here (a shuffled protein database) is conservative.

Multiple Testing Correction Using the False Discovery Rate

Unfortunately, the preceding analysis is incomplete. Using a *p*-value threshold is inadequate because we have performed our statistical test so many times. Our ranked list contains 34 499 PSMs, and so we expect to observe $0.01 \times 34\,499 = 345$ PSMs with *p*-values less than 0.01 simply by chance. We need to perform what statisticians call *multiple testing correction*.

In a study such as ours, in which we perform many statistical tests and in which we expect many of the tests to be positive

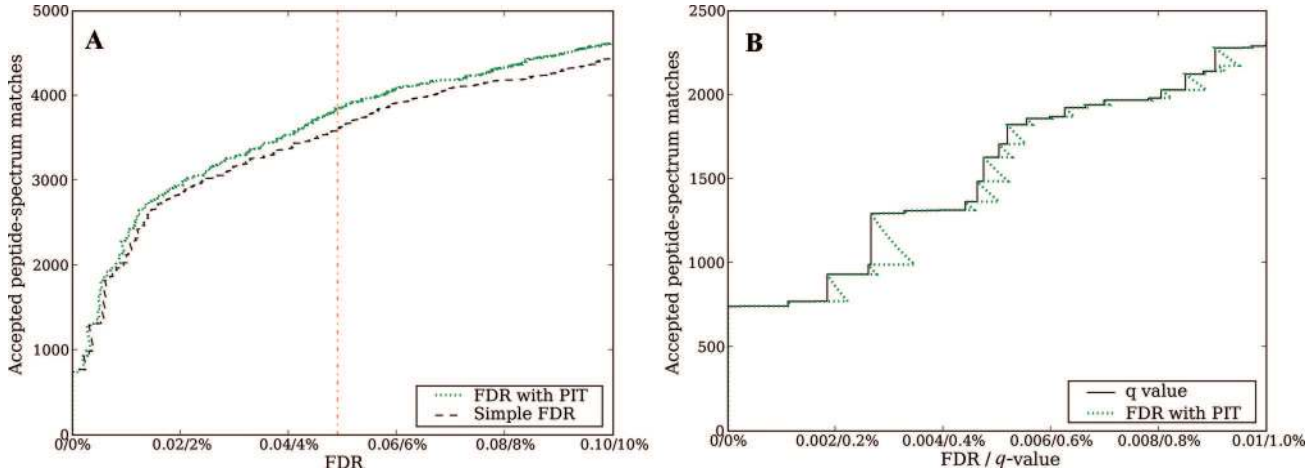

Figure 4. Mapping from the number of identified PSMs to the estimated false discovery rate. (A) The figure plots the number of PSMs above the threshold as a function of the estimated false discovery rate. Two different methods for computing the FDR are plotted, with and without an estimate of the percentage of incorrect target PSMs (PIT). The vertical line corresponds to an XCorr of 3.0. (B) A zoomed-in version of panel A, with the estimated FDR shown as a dotted line and the *q*-value shown as a solid line.

(i.e, many of the PSMs are correct), the accepted method for multiple testing correction is to estimate the *false discovery rate* (FDR).[10,11] Storey and Tibshirani[12] provide a description of FDR methods that is accessible to nonstatisticians and that includes more recent developments. In our case, the FDR associated with a particular score threshold is defined as the expected percentage of accepted PSMs that are incorrect, where an "accepted PSM" is one that scores above the threshold (Many proteomics papers incorrectly refer to this quantity as the "false positive rate.") However, other scientific fields define the false positive rate as the fraction of true null tests that are called significant,[13–17] whereas the false discovery rate is defined as the fraction of true null tests among all of those that are called significant). For example, at an FDR of 1%, if we accept 500 PSMs, then we expect five of those matches to be incorrect.

The simplest way to calculate the FDR is analogous to the calculation of *p*-values, above. For a given score threshold, we count the number of decoy PSMs above the threshold and the number of target PSMs above the threshold. We can now estimate the FDR by simply computing the ratio of these two values. For example, at a score threshold of 3.0, we observe 3849 accepted target PSMs and 219 accepted decoy PSMs, yielding an estimated FDR of 5.7%. Figure 4 plots the number of accepted PSMs as a function of the estimated FDR, and the series labeled "Simple FDR" was computed using the ratio of accepted decoys versus accepted targets.

Estimating the Percentage of Incorrect Target PSMs

A slightly more sophisticated method for calculating the FDR takes into account the observation that, whereas all decoy PSMs are incorrect by construction, not all target PSMs are correct. Ideally, the presence of these incorrect target PSMs should be factored into the FDR calculation. For example, suppose that among 10 000 target PSMs, 8000 are incorrect and 2000 are correct. We would like to know the 8000 quantity so that we can adjust our FDR estimates.

Figure 2 shows that the distributions of scores assigned to target and decoy PSMs are similar, except that the target PSM score distribution has a heavier tail to the right. This tail arises because the set of target PSMs is comprised of a mixture of correct and incorrect PSMs. Figure 5 shows simulated distributions that illustrate the underlying phenomenon. For this simulation, we assume that our PSM score function follows a normal distribution, and we set the standard deviation to 0.7 (The assumption of normality is for the purposes of illustration only; the methods we describe here do not require any particular form of distribution, nor do we assume that XCorr is normally distributed). For incorrect PSMs, we set the mean of the distribution to 1.0, and for correct PSMs, we change the mean to 3.0. Our simulated data set contains 10 000 decoy PSMs, 8000 incorrect target PSMs, and 2000 correct target PSMs. The figure shows the resulting decoy score distribution (black line), the target score distribution (blue line), and its two component distributions (dotted and dashed blue lines). In this simulated data set, the percentage of incorrect targets (PIT) is 80%. This PIT is equivalent to the ratio of the area under the dotted black line (the incorrect target PSMs) to the area under the solid black line (the decoy PSMs).

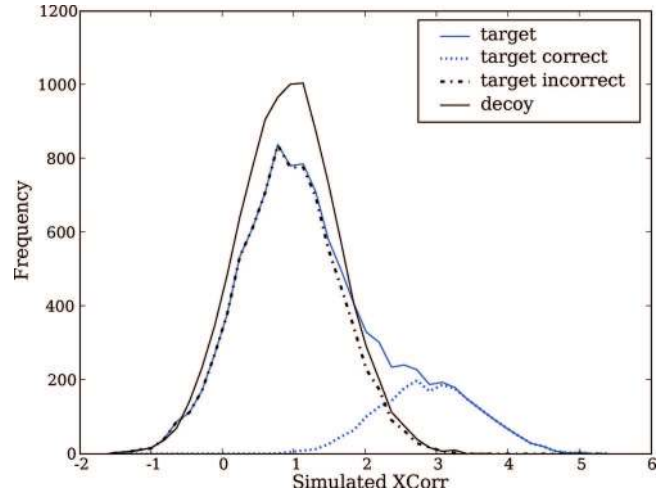

Figure 5. Simulated target and decoy PSM score distributions.

The PIT is important because it allows us to reduce the estimated FDR associated with a given set of accepted target PSMs. In our simulation, if we accept X decoy PSMs with scores above a certain threshold, then we expect to find $0.8X$ incorrect target PSMs above the same theshold. A more accurate estimate of the FDR, therefore, is to multiply the previous estimate—the

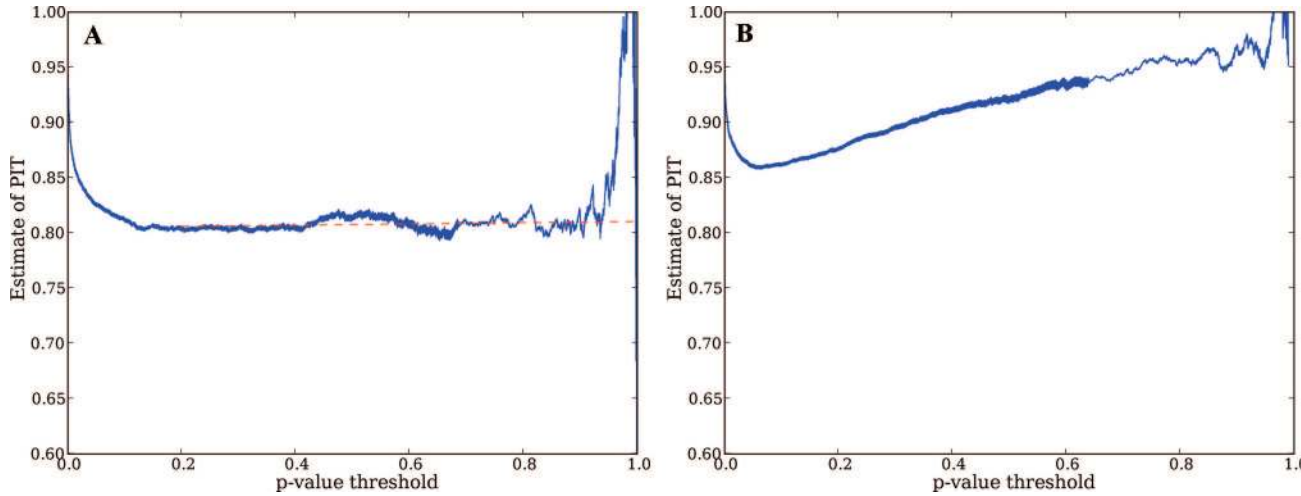

Figure 6. Estimation of percentage of incorrect target PSMs. Each panel plots the estimated PIT as a function of the *p*-value threshold for (A) simulated data and (B) yeast data. For each PIT estimate, only those PSMs with *p*-values greater than the given threshold are considered.

ratio of the number of decoy PSMs versus target PSMs above a threshold—by the PIT.

Employing this more accurate FDR estimate requires that we estimate the PIT from the observed score distributions. If we accept all PSMs with XCorr ≥ 0 (or, equivalently, accept all PSMs with *p*-value ≤ 1), then the PIT is defined as the ratio of the number of false discoveries to the total number of PSMs. The numerator in this ratio is unknown but can be conservatively estimated by considering PSMs with scores close to zero and calculating the ratio of the number of target PSMs to the number of decoy PSMs within that set.[5] The rationale behind this estimate is that the majority of target PSMs with small values should correspond to incorrect target PSMs. For example, if 8000 out of 10 000 target PSMs are incorrect, then for PSMs with scores close to zero, we expect most target PSMs to come from this 8000. The decoy PSMs are constructed under the scenario that all 10 000 are incorrect. Therefore, the ratio of target PSMs to decoy PSMs should be about 80%. In expectation, the ratio is slightly greater than 80% because there will be a few correct target PSMs with low scores.

For a concrete example of this estimation procedure, consider all PSMs with XCorr ≤ 3, corresponding to the vertical red line in Figure 4A. This interval contains 34 499 − 3849 = 30 650 target PSMs and 34 499 − 219 = 34 280 decoy PSMs, yielding a ratio of 89%. A similar estimate can be formed for all XCorr intervals of the form [0, *c*], not just [0, 3]. As the interval is made larger, the PIT estimate becomes more conservative and the variance decreases.[12] Therefore, a variety of methods exist for averaging information across the various choices of intervals to balance this bias-variance tradeoff.[5,12,18–21]

Figure 6 illustrates how the estimated PIT varies as we change the XCorr interval (or, equivalently, the *p*-value threshold). In Figure 6A, we use the simulated data described previously. As the *p*-value threshold increases, that is, as we restrict our attention to PSMs with lower and lower scores, the estimated PIT decreases. One simple estimation procedure is to fit a straight line (shown in red) to these estimates, which yields an estimated PIT of 0.81. This estimate is slightly conservative with respect to the true PIT of 0.80.

Figure 6B shows what happens when we repeat the estimation procedure using real data. The increasing trend in the plot is evidence of a conservative null model. Apparently, there is an enrichment of target PSMs with very low scores, which likely correspond to poor quality spectra. A significant avenue for future research is finding a better null model that does not yield this type of artifact. In this particular case, using the method of Storey,[5] we estimate the PIT at 0.86; that is, we estimate that 14% of our target PSMs are correct. The series labeled "FDR with PIT" in Figure 4A shows the results of applying the PIT as a multiplicative factor. At the threshold considered previously, our estimated FDR is 0.86 × 219/3849 = 4.9%, which is lower than the previous estimate of 5.2%.

From the mass spectrometrist's perspective, incorporating an estimate of the PIT adds significant value because it leads to a much larger number of peptide identifications for a given FDR. For example, at an FDR of 1%, the simple estimation procedure yields 2123 accepted target PSMs. After estimating the PIT, the number of accepted target PSMs increases by 9.3% to 2320. A similar effect has been observed in several genomics applications.

q-Values

Unfortunately, as shown in Figure 4B, the FDR has the somewhat counterintuitive property that it is not a function of the underlying score: two different scores can lead to the same FDR. In our case, a score threshold of 4.14 yields 4 decoy PSMs and 919 target PSMs, implying an FDR of 0.35%, whereas a threshold of 3.98 yields a larger set of accepted PSMs (4 decoys and 1294 targets) but a smaller estimated FDR (0.27%). This property makes it difficult to apply an FDR threshold to a given data set.

To address this problem, Storey and Tibshirani[12] propose a new metric, the *q*-value, which in our case is defined as the minimal FDR threshold at which a given PSM is accepted (Note that the *q*-value is not related to the Qscore[7]). The solid line in Figure 4B shows the *q*-value as a function of score. In the above example, a *q*-value threshold of 0.27% unambiguously yields 1294 identifications. The primary distinction between the FDR and the *q*-value is that the former is a property of a set PSMs, whereas the latter is a property of a single PSM. We can therefore associate a unique *q*-value with every target PSM in our data set.

The *q*-value is intended to be analogous to the *p*-value, but taking into account multiple testing correction. Figure 3

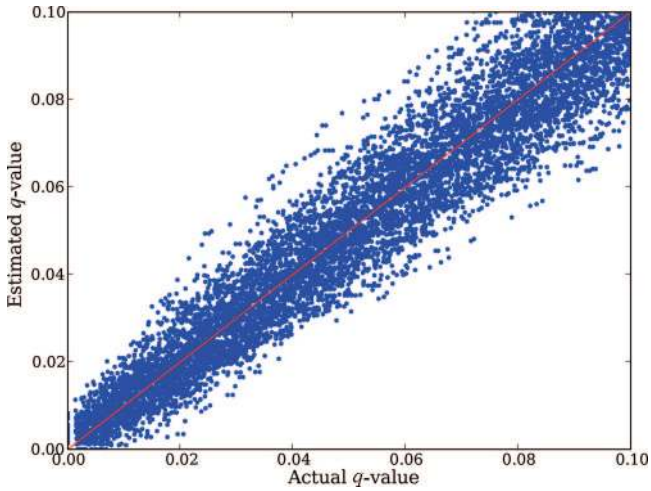

Figure 7. Accuracy of estimated *q*-values. The figure plots the estimated *q*-value as a function of the actual *q*-value for a simulated data set, as described in the text.

illustrates the relationship between *p*-values and *q*-values for our example data set. In general, the *q*-value associated with a given XCorr threshold is higher than the corresponding *p*-value, except when the XCorr threshold is quite low.

We performed a simple simulation to illustrate the accuracy of the *q*-values estimated via the procedure described above. We used the simulated data shown in Figure 5: 10 000 target PSMs and 10 000 decoy PSMs, with the PIT set at 80%. Figure 7 plots the estimated *q*-value as a function of the true *q*-value when repeating the experiment 200 times. Reassuringly, the points lie symmetrically close to the line $y = x$: 96% of the estimated *q*-values in the range of 0 – 0.1 are within a factor of 2 of the actual *q*-value. Thus, the plot demonstrates that this method of estimating *q*-values is accurate.

Discussion

We have described a method for assigning significance measures—in particular, *q*-values—to a ranked list of PSMs by exploiting a null model derived from a decoy protein database. Although we have used the SEQUEST XCorr score throughout, the method described here is quite general and can be used for essentially any PSM scoring routine.

Many research groups routinely use FDR calculations based on a target-decoy search strategy,[7,22–25] though most of these approaches use the "simple FDR" estimation procedure shown in Figure 4A. We have demonstrated that taking into account the percentage of incorrect target PSMs increases the number of accepted target PSMs at a fixed FDR. In the future, as mass spectrometry technology and PSM scoring schemes improve, the PIT is likely to decrease, making this type of FDR calculation even more valuable.

Methods for estimating FDRs and *q*-values, similar to the methods we propose, have been described and validated extensively in the statistical literature.[5,11,12] However, these techniques all require that the true null distribution has been used, or at the very least a conservative version of the true null distribution. Effects from hidden covariates, when not taken into account, have been shown to warp the null distribution in multiple testing situations.[26–30] Also, it should be noted that each spectrum results in the formation of a target PSM and a decoy PSM. However, the decoy PSMs are pooled together and used to evaluate the significance of the set of target PSMs. It is not yet well-understood what assumptions are required to treat the set of decoy PSMs in this exchangeable fashion, and it may be the case that the decoy PSM derived from one spectrum is not representative of the null PSM distribution of a different spectrum. The important point is that several issues affect our ability to obtain the correct null distribution, and these have to be considered carefully.

The procedure that we have described involves searching spectra against target sequences and decoy sequences separately. The methods could be extended to estimate significance for peptide identifications obtained by searching the spectra once against a merged database containing both target and decoy sequences. However, computing valid significance estimates when using this strategy is difficult, because one must ensure that the distribution of decoy PSM scores accurately represents the target null distribution. For example, it is important to compensate for the fact that target-decoy competition yields more target PSMs than decoy PSMs. Failure to compensate for this effect could lead to a general underestimation of FDRs.

Note that, ideally, significance measures should also be assigned to proteins as well as PSMs. The methods that we describe here could be applied at the level of protein identifications, but doing so requires an appropriate protein-level scoring scheme.

Acknowledgment. This work was supported by NIH awards R01 EB007057 and P41 RR11823.

References

(1) Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.
(2) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probabilitybased protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551–3567.
(3) Craig, R.; Beavis, R. C. Tandem: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466–1467.
(4) Käll, L.; Canterbury, J.; Weston, J.; Noble, W. S.; MacCoss, M. J. A semi-supervised machine learning technique for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4*, 923–925.
(5) Storey, J. D. A direct approach to false discovery rates. *J. R. Stat. Soc.* **2002**, *64*, 479–498.
(6) Efron, B.; Tibshirani, R.; Storey, J.; Tusher, V. Empirical bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* **2001**, *96*, 1151–1161.
(7) Moore, R. E.; Young, M. K.; Lee, T. D. Qscore: An algorithm for evaluating sequest database search results. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 378–386.
(8) Klammer, A. A.; MacCoss, M. J. Effects of modified digestion schemes on the identification of proteins from complex mixtures. *J. Proteome Res.* **2006**, *5*, 695–700.
(9) Colinge, J.; Masselot, A.; Giron, M.; Dessingy, T.; Magnin, J. OLAV: Towards high-throughput tandem mass spectrometry data identification. *Proteomics* **2003**, *3*, 1454–1463.
(10) Soric, B. Statistical discoveries and effect-size estimation. *J. Am. Stat. Assoc.* **1989**, *84*, 608–610.
(11) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc., Ser. B* **1995**, *57*, 289–300.
(12) Storey, J. D.; Tibshirani, R. Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 9440–9445.
(13) Rice, J. A. *Mathematical Statistics and Data Analysis*, 2nd ed.; Duxbury PressBelmont, CA, 1995.
(14) Baldi, P.; Brunak, S. *Bioinformatics: The Machine Learning Approach*; MIT Press: Cambridge, MA, 1998.
(15) Finkelstein, M. O.; Levin, B. A. *Statistics for Lawyers*. Springer: New York, 2001.

(16) Lang, T. A.; Secic, M. *How To Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers*; ACP Press: Philadelphia, PA, 2006.
(17) Schroeder, C. S.; Ollendick, T. H. *Encyclopedia of Clinical Child and Pediatric Psychology*; Kluwer Academic: New York, 2003.
(18) Langaas, M.; Lindqvist, B.; Ferkingstad, E. Estimating the proportion of true null hypotheses, with application to dna microarray data. *J. R. Stat. Soc., Ser. B* **2005**, *67*, 555–572.
(19) Meinshausen, N.; Rice, J. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.* **2006**, *34*, 373–393.
(20) Hsueh, H.; Chen, J. J.; Kodell, R. L. Comparison of methods for estimating the number of true null hypotheses in multiplicity testing. *J. Biopharm. Stat.* **2003**, *13*, 675–689.
(21) Nguyen, D. V. On estimating the proportion of true null hypotheses for false discovery rate controlling procedures in exploratory DNA microarray studies. *Comput. Stat. Data Anal.* **2004**, *47*, 611–637.
(22) Weatherly, D. B.; Astwood, J. A.; Minning, T. A.; Cavola, C.; Tarleton, R. L.; Orlando, R. A Heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results. *Mol. Cell. Proteomics* **2005**, *4*, 762–772.
(23) Olsen, J.; Ong, S.; Mann, M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **2004**, *3*, 608–614.
(24) Nielsen, M.; Savitski, M.; Zubarev, R. Improving protein identification using complementary fragmentation techniques in fourier transform mass spectrometry. *Mol. Cell. Proteomics* **2005**, *4*, 835–845.
(25) Higgs, R. E.; Knierman, M. D.; Freeman, A. B.; Gelbert, L. M.; Patil, S. T.; Hale, J. E. Estimating the statistical signficance of peptide identifications from shotgun proteomics experiments. *J. Proteome Res.* **2007**, *6*, 1758–1767.
(26) Devlin, B.; Roeder, K. Genomic control for association studies. *Biometrics* **1999**, *55*, 997–1004.
(27) Qiu, X.; Xiao, Y.; Gordon, A.; Yakovlev, A. Assessing stability of gene selection in microarray data analysis. *BMC Bioinf.* **2006**, *7*, 50.
(28) Klebanov, L.; Yakovlev, A. Treating expression levels of different genes as a sample in microarray data analysis: Is it worth a risk? *Stat. Appl. Genet. Mol. Biol.* **2006**, *5*, 9.
(29) Efron, B. Correlation and large-scale simultaneous significance testing. *J. Am. Stat. Assoc.* **2007**, *102*, 93–103.
(30) Leek, J. T.; Storey, J. D. Capturing heterogeneity in gene expression studies by "surrogate variable analysis". *PLoS Genet.* **2007**, *3*, e161.
(31) Price, T. S.; Lucitt, M. B.; Wu, W.; Austin, D. J.; Pizarro, A.; Yokum, A. K.; Blair, I. A.; FitzGerald, G. A.; Grosser, T. EBP, a program for protein identification using multiple tandem mass spectrometry datasets. *Mol. Cell. Proteomics* **2007**, *6*, 527–536.
(32) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75*, 4646–4658.
(33) Sadygov, R. G.; Yates, J. R. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal. Chem.* **2003**, *75*, 3792–3798.

PR700600N