perspectives Journal of proteome research

Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin

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

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

Received November 14, 2007

A variety of methods have been described in the literature for assigning statistical significance to peptides identified via tandem mass spectrometry. Here, we explain how two types of scores, the *q*-value and the posterior error probability, are related and complementary to one another.

Keywords: *q*-value • posterior error probability • false discovery rate • statistical significance • peptide identification

Introduction

The immediate goal of most tandem mass spectrometry experiments is to identify proteins in a complex biological sample. A perfect experiment would produce a list of all and only the proteins that exist in the sample. In practice, of course, real experiments are imperfect and yield lists that contain proteins that were not actually in the sample—false positives—and leave out proteins that were in the sample—false negatives. Therefore, in interpreting the experimental results and particularly in the context of designing follow-up experiments, a biologist benefits from the availability of statistical scores with well-defined semantics.

The accompanying articles[1–3] describe several methods for associating statistical scores with the results of tandem mass spectrometry experiments. For logistical reasons, the current discussion focuses on peptide-level identifications, rather than protein-level identifications, but similar methods and models can be applied to compute protein-level statistical scores.

Here, we explain why statistical scoring is valuable and attempt to clarify the relationship among a variety of technical terms that have been borrowed from the statistical literature. The take-home message is that two forms of scores, the *q*-value and the posterior error probability (PEP), are valuable and complementary. Therefore, an ideal software package for mass spectrometry analysis should produce both of these scores. Which score the biologist focuses on will depend upon the type of follow-up experiments that are being planned or the type of conclusions being drawn from the results.

False Discovery Rates and *q*-Values

Our accompanying article[3] does not describe new statistical methods; rather, we describe how well-established methods from the statistical literature can be applied to peptide identification from tandem mass spectra. In particular, we show how searching a set of spectra against a decoy protein database, containing reversed, shuffled, or Markov chain-generated amino acid sequences, enables us to associate a particular score, called a *q*-value, with every peptide−spectrum match (PSM). Rather than recapitulate how *q*-values are computed, we focus here on how to interpret these *q*-values. Say that a given spectrum *s* matches a particular peptide EAMRQPK with a *q*-value of 0.01. What does this tell us?

To understand *q*-values, we must first understand the notion of the false-discovery rate (FDR). Say that the goal of our mass spectrometry experiment is to match every observed spectrum to a peptide in the given database, and then separate the list of PSMs into correct and incorrect matches. If we set an FDR threshold of 1%, this means that we are willing to accept a list of PSMs in which 99% of the matches are correct and 1% are not. Clearly, if we increase the FDR threshold to, say, 10%, then we will end up with a much longer list of PSMs. But the tradeoff is that a larger percentage of the PSMs in the list will be incorrect.

Given this definition of FDR, a *q*-value of 0.01 for peptide EAMRQPK matching spectrum s means that, if we try all possible FDR thresholds, then 1% is the minimal FDR threshold at which the PSM of EAMRQPK to s will appear in the output list.

Although the *q*-value is associated with a single PSM, it is important to recognize that the *q*-value depends upon the data set in which the PSM occurs. Say that we search a collection of spectra against the entire nonredundant protein database. We rank the resulting PSMs by score, and we observe (EAMRPK,s) at position 100. If an oracle tells us that (EAMRPK,s) is a correct mapping, but that 10 of the PSMs ranked above (EAMRPK,s) are incorrect, then the true *q*-value associated with (EAMRPK,s) is 0.1.

Now consider an alternative situation. Rather than search the full nonredundant protein database, say that we instead search the tryptic database and that, in the process, we remove 20 of the 99 PSMs that rank above (EAMRPK,s). Among these are 6 incorrect PSMs. In this second scenario, (EAMRPK,s) now

* 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/pr700739d CCC: $40.75
Published on Web 12/04/2007

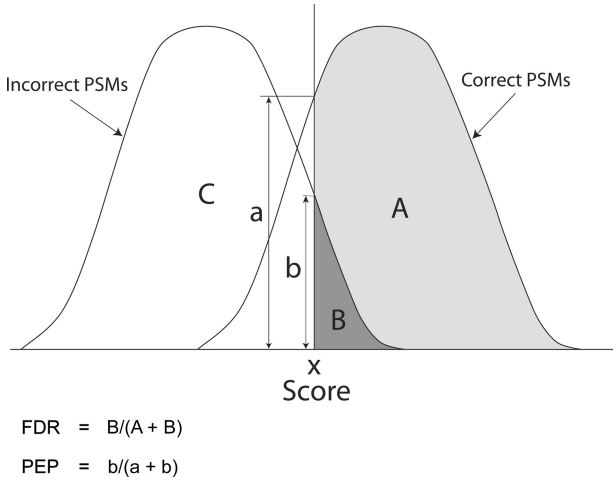

Figure 1. Two complementary methods for assessing statistical signiffcance.

falls at position 80 in the ranked list, and its q-value is $4/80 = 0.05$. Note that the spectrum has not changed, nor has its score, but the associated q-value has changed from 0.1 to 0.05. A similar effect could be observed by switching from the nonredundant protein database to an organism-specific database, or by applying a quality filter to the spectra themselves.

Posterior Error Probability

Rather than focusing on computing q-values, the other two accompanying articles describe methods for computing posterior error probabilities. The PEP is, quite simply, the probability that the observed PSM is incorrect. Thus, if the PEP associated with (EAMRPK,s) is 5%, this means that there is a 95% chance that the peptide EAMRPK was in the mass spectrometer when spectrum s was generated.

The PEP can be thought of as a local version of the FDR (indeed, Efron et al.[4] have used the term "local FDR" to refer to the PEP). Whereas the FDR measures the error rate associated with a collection of PSMs, the PEP measures the probability of error for a single PSM. Equivalently, the PEP measures the error rate for PSMs with a given score x. In practice, a given data set will contain only a single PSM with a particular score, so the PEP must be estimated using a model.

The relationship between PEP and FDR can be understood visually from Figure 1. The FDR is the ratio of B (the number of incorrect PSMs with score $> x$) to $(A + B)$ (the total number of PSMs with score $> x$). Note that A and B are areas of the distribution. The PEP, on the other hand, is a ratio of the corresponding heights of the distribution: the number b of incorrect PSMs with score $= x$ divided by the total number $(a + b)$ of PSMs with score $= x$. As pointed out by Choi et al.,[2] the FDR can be computed from the PEPs, because the expected number of incorrect PSMs in a given set is equal to the sum of the PEPs.

A common statistical or machine learning approach to estimating posterior probabilities is to learn the parameters of a probability model from a set of labeled training data, and to use the learned parameters to predict PEPs for all future test data. With this approach, the PEP associated with a PSM with score x will always be the same, regardless of the data set in which the PSM occurs.

One of the appealing features of PeptideProphet is its ability to adjust the PEP estimates on the basis of the current data set. This was true of the original version of PeptideProphet,[5] and the accompanying articles describe several improvements to this component of PeptideProphet.[1,2]

Which Score Is Better?

Say that you have just finished running a mass spectrometry experiment, and you have used a database search program to match each spectrum to a peptide. You now need to choose a piece of software to assign statistical scores to each of these PSMs. Assume that you have two choices: one program that computes accurate q-values and one program that computes accurate PEPs. Which program should you choose?

The answer depends upon what you plan to do with your results. PEPs and q-values are complementary, and are useful in different situations. The q-value estimates the rate of misclassification among a set of PSMs. If you are interested in determining which proteins are expressed in a certain cell type under a certain set of conditions, or if your follow-up analysis will involve looking at groups of PSMs, for example, considering all proteins in a known pathway, evaluating enrichment with respect to Gene Ontology categories, or performing experimental validation on a group of proteins, then the q-value is an appropriate measure.

If the goal of your experiment instead is to determine the presence of a specific peptide or protein, then the PEP is more relevant. For example, imagine that you are interested in determining whether a certain protein is expressed in a certain cell type under a certain set of conditions. In this scenario you should examine the PEPs of your detected PSMs. Likewise, imagine that you have identified a large set of PSMs using a q-value threshold, and among them, you identify a single PSM that is intriguing. Before deciding to dedicate significant resources to investigating a single result, you should examine the PEP associated with that PSM. This is because, although the q-value associated with that PSM may be 0.01, the PEP is always greater than or equal to 0.01. In practice, the PEP values for PSMs near the $q = 0.01$ threshold are likely to be much larger than 1%.

Figure 2 shows the relationship between PEP and q-value for a real data set, a collection of 34 492 2+ fragmentation spectra derived from a yeast whole-cell lysate. Setting a PEP threshold of 1% yields 1029 PSMs, but the estimated FDR of this set of PSMs is only 0.3%. Alternatively, setting a threshold of $q = 0.01$ yields 1978 PSMs. Thus, for this data set, switching from PEP to q-values yields 92% more identifications.

It can be shown[8] that thresholding FDR or PEP are actually two equivalent ways of implicitly balancing the tradeoff between false positives and false negatives. This tradeoff also depends upon the prior probabilities of being in one class or the other, how different the distributions are between the null and alternative hypotheses, and so forth. From this perspective, it should be clear why both PEP and FDR are important. While a PEP cutoff allows one to easily determine the tradeoff between false positives and false negatives (see Storey[8] for the specific formula to do this), the FDR allows one to quantify the overall quality of the discrimination procedure, particularly focusing on those PSMs that we call significant.

Pathological scenarios can be constructed where it appears that it is necessary to employ a PEP measure. One simple example is the case where 100 PSMs are called significant. At the same time, PEP = 0 for the 99 most significant PSMs and PEP = 1 for the 100th most significant PSM. Calling all 100 PSMs significant leads to FDR = 1%, which is perfectly

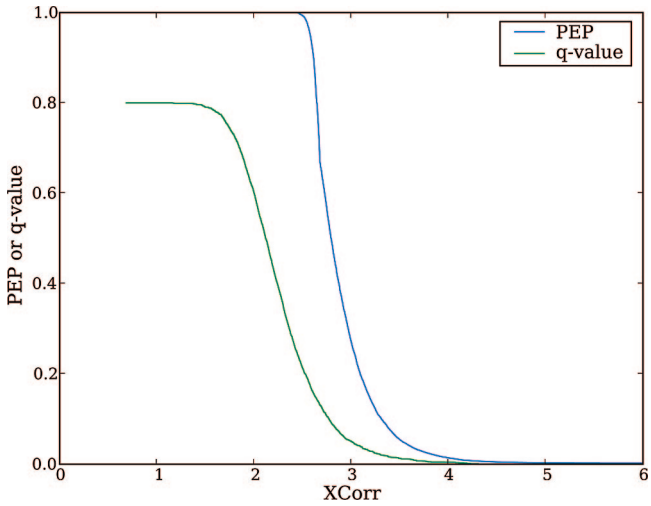

Figure 2. Relationship between *q*-value and posterior error probability. The figure plots the estimated *q*-value (green curve) and the estimated posterior error probability (blue curve) as a function of the score threshold. A set of 34 492 2+ fragmentation spectra were searched using Sequest[6] against a database of yeast predicted open reading frames and separately against a shuffled version of the same database. The *q*-values were estimated using standard methods.[3] Inspired by Storey et al.,[7] PEPs were estimated by fitting a piecewise linear curve to a histogram of logit-transformed empirical error rates.

reasonable. However, if we had also employed PEP, we would know that the 100th most significant PSM should not be called significant.

Although compelling, when one considers how the researcher utilizes FDR and *q*-values, this counter-example falls apart. A standard technique is to plot the number of expected false positives versus the number of PSMs called significant. This simultaneous use of *q*-values (i.e., considering all possible FDR levels without having to decide on one beforehand) has been statistically justified.[9] Clearly, in the above case, we would see that zero false positives are expected when calling the top 99 PSMs significant and one false positive is expected when calling the top 100 PSMs significant. From this, it is easily deduced that the 100th most significant PSM is most likely a false positive.

Alternative scenarios can also be constructed where it does not suffice to only employ the PEP measure. For example, suppose that a PEP threshold of 5% is employed. If all PEP values meeting this threshold are equal to 5%, then it can be shown that the FDR of the set of significant PSMs is also exactly 5%. However, if a range of PEP values exist, all the way from 0% to 5%, then the FDR will actually be much smaller than 5%. Thus, as pointed out by Choi et al.,[2] setting a threshold on the PEP also bounds the false-discovery rate. However, setting the threshold in this way is extremely conservative: selecting PSMs such that the PEP is less than 5% will generally produce a much smaller list of PSMs than setting a *q*-value threshold of 0.05. If the goal is to understand properties of a set of PSMs, then the *q*-value is the correct metric to use.

In general, it takes much additional work to verify all of the conditions required to calculate a PEP, and it is arguable that the necessary assumptions may never be true.[10] The amazing property of FDR is that it can easily be calculated in a nonparametric fashion based on standard *p*-values without ever having to invoke the more sophisticated Bayesian classification theory approach.[11]

Shotgun proteomics with data-dependent acquisition is a high-throughput technology. The volume of data produced in such an experiment is arguably more amenable to analysis of groups of results, rather than single measurements. For example, a common goal is to obtain sets of significant PSMs that show a biological signal of interest. In this context, it is important to obtain high-quality sets of PSMs; the FDR simply gives the level of noise in any set. As such, we believe that FDR-based metrics such as the *q*-value are likely to be the most widely applicable significance scores for this type of data.

What about *p*-Values and *E*-Values?

The *q*-value is a relatively new significance measure. More familiar to many people is the *p*-value, and some readers may be wondering how the two concepts are related. It is possible, given a set of decoy PSM scores, to compute a *p*-value for each target PSM, and we describe a procedure for doing this in the accompanying article. However, the *p*-value is not corrected for multiple tests. Intuitively, if we search a very large collection of spectra against a given sequence database, we should expect some of the resulting *p*-values to be small, simply by chance. If we do not perform multiple testing correction, then our significance scores will be anticonservative, meaning that we will erroneously assign statistical significance to some examples that are not actually significant.

The simplest form of multiple testing correction is the Bonferroni correction. This correction says that, if you are aiming for a significance threshold of 0.05 but you repeat your test 1000 times, then you should adjust your threshold to $0.05/1000 = 0.00005$. In a typical peptide identification experiment, the effective number of tests is very large. If we can compute the distribution of our PSM score, then we can compute the *p*-value associated with a single PSM. However, we have to correct for the number of candidate peptides that the spectrum was compared to, that is, the total number of database peptides whose mass is within a specified range around the inferred precursor mass of the spectrum. Furthermore, if we are searching a large collection of spectra against the same database, then we also have to correct for the total number of spectra in our data set. A Bonferroni correction that takes into account both of these factors, number of candidate peptides and number of spectra, will be extremely conservative, and we will end up identifying very few peptides.

The *E*-value is an alternative method for multiple testing correction. *E*-values are computed by X!Tandem,[12] OMSSA,[13] and Mascot.[14] In these programs, the *E*-value calculation is essentially the converse of the Bonferroni correction. Rather than dividing the target significance threshold by the number of tests performed, the *E*-value is the product of the *p*-value and the number of tests. The *E*-value can be interpreted as the expected number of times that you would expect to observe a PSM with a score x by chance. If the significance threshold is kept the same, then using *E*-values is exactly equivalent to the Bonferroni correction. Note, however, that the *E*-values reported by X!Tandem, OMSSA, and Mascot only correct for the number of candidate peptides, not the number of spectra in the data set. In the context of a large collection of mass spectra; therefore, these *E*-values are anticonservative.

Conclusion

Figure 3 summarizes the relationship among various methods for assigning significance to a collection of PSMs. Using

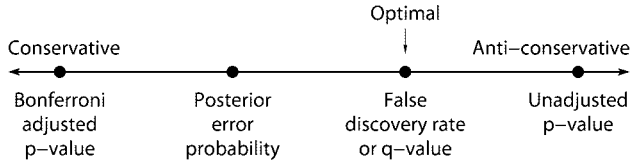

Figure 3. Methods for assigning statistical significance

unadjusted *p*-values is invalid. The other three methods, thresholding by *q*-value, PEP, or using a Bonferroni correction, are increasingly conservative. If the goal is to identify as many peptides/proteins as possible in a statistically valid fashion, then the *q*-value is the metric of choice. However, the *q*-value is, fundamentally, a measure of error rate within a collection of PSMs. It is therefore complementary to the PEP, which measures the probability of error for a single PSM. Depending upon whether you are looking at groups of PSMs or individual PSMs, you should choose the appropriate significance score.

Rejoinder. Several of the accompanying commentaries distinguish between searching a concatenated target-decoy database versus searching the two databases separately.[15,16] This distinction is somewhat misleading. As pointed out by Fitzgibbon et al., after performing separate searches, it is trivial to perfom the target-decoy competition a posteriori. Conversely, if we use database search software that allows reporting all PSMs, rather than just a few top PSMs, then the top-scoring target and decoy scores could be extracted from the output of a concatenated search.

The important distinction, in this context, is not how the search is conducted, but how the statistical significance is computed from the resulting PSM scores. We use as a model of the null distribution the complete set of decoy PSM scores. In contrast, the protocol of Elias and Gygi[17] uses the distribution of decoy PSM scores that win the target-decoy competition as a null model for the target PSM scores that win the target-decoy competition. While the target-decoy competition has a number of promising attributes that are beyond the scope of this rejoinder, we have concerns that these data may not accurately reflect the significance of a database search result. In the target-decoy competition protocol, the null distribution is computed with respect to a set of spectra that is completely disjoint from the spectra in the target distribution. It is not clear to us whether this is an accurate null model.

Fitzgibbon et al.[15] describe the possibility of using a decoy database that is smaller or larger than the target database. An alternative strategy is to compute decoy PSMs for only a subset of the spectra. In this case, the relative number of decoy PSMs would be included as a multiplicative factor in the FDR calculation. Fitzgibbon et al. then argue convincingly that a smaller decoy database should be used when doing so would yield suffciently accurate FDR estimates. The converse is also true. In particular, if a researcher is interested in identifying a set of identifications with an extremely low FDR, then a very large collection of decoy PSMs could be constructed by searching each spectrum against multiple decoy databases. The choice of the number of decoy PSMs comes down to a tradeoff between desired accuracy of the significance estimates versus computational expense. Note that standard methods do exist for estimating the error on FDR estimates,[18] allowing the researcher to make this choice in a principled fashion.

Choi and Nesvizhskii[16] argue that the distinction between FDR and *q*-values may not be crucial. The *q*-value serves the same purpose for FDR as the *p*-value does for the false-positive rate (or type I error rate). To put this into a historical context, in the early days of hypothesis testing, Neyman and Pearson suggested that all hypothesis testing be done with a predetermined false-positive rate and the only result one should report is whether the test is significant or not. However, R. A. Fisher argued that a *p*-value is more informative and that the *p*-value should be reported instead. In looking at the scientific literature, it is clear that the reporting of *p*-values has won this argument. Indeed, the *p*-value is the simplest implementation of monitoring the false-positive rate in a data-adaptive manner. The *q*-value serves this exact same purpose for FDR; it also allows the researcher to evaluate a list of significant tests in an unbiased manner. We therefore believe that it is relevant for a researcher to understand the distinction between FDR and *q*-values, in the same manner that we currently make a distinction between false-positive rate and *p*-values.

Finally, we agree wholeheartedly with Choi and Nesvizhskii that, ultimately, mass spectrometrists should be interested in computing statistical significance at the level of protein identifications, rather than at the level of individual spectra. The methods that we describe can be applied at the protein level, modulo the considerations mentioned by Choi and Nesvizhskii.

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

References

(1) Nesvizhskii, A.; Choi, H. Semi-supervised model-based validation of peptide identification in mass spectrometry-based proteomics. *J. Proteome Res.* **2008**, *7*, 254–265.
(2) Choi, H.; Ghosh, D.; Nesvizhskii, A. Statistical validation of peptide identifications in large-scale proteomics using target-decoy database search strategy and flexible mixture modeling. *J. Proteome Res.* **2008**, *7*, 286–292.
(3) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.* **2008**, *7*, 29–34.
(4) Efron, B.; Tibshirani, R.; Storey, J.; Tusher, V. Empirical bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* **2001**, *96*, 1151–1161.
(5) Keller, A.; Nezvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identification made by MS/MS and database search. *Anal. Chem.* **2002**, *74*, 5383–5392.
(6) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. 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.
(7) Storey, J. D.; Akey, J. M.; Kruglyak, L. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol.* **2005**, *3*, 1380–1390.
(8) Storey, J. D. A direct approach to false discovery rates. *J. R. Stat. Soc.* **2002**, *64*, 479–498.
(9) Storey, J. D.; Taylor, J. E.; Siegmund, D. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Stat. Soc., Ser. B* **2004**, *66*, 187–205.
(10) Dudoit, S.; Shaffer, J. P.; Boldrick, J. C. Multiple hypothesis testing in microarray experiments. *Stat. Sci.* **2003**, *18*, 71–103.
(11) Storey, J. D.; Tibshirani, R. Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 9440–9445.
(12) Fenyo, D.; Beavis, R. C. A method for assessing the statistical significance of mass spectrometry-based protein identification using general scoring schemes. *Anal. Chem.* **2003**, *75*, 768–774.
(13) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, *3*, 958–964.
(14) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551–3567.

(15) Fitzgibbon, M.; Li, Q.; McIntosh, M. Modes of inference for evaluating the confidence of peptide identifications. *J. Proteome Res.* **2008**, *7*, 35–39.
(16) Choi, H.; Nesvizhskii, A. I. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J. Proteome Res.* **2008**, *7*, 47–50.
(17) 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.
(18) Huttlin, E. L.; Hegeman, A. D.; Harms, A. C.; Sussman, M. R. Prediction of error associated with false-positive rate determination for peptide identification in large-scale proteomics experiments using a combined reverse and forward peptide sequence database strategy. *J. Proteome Res.* **2007**, *7*, 392–398.

PR700739D