Estimating the Confidence of Peptide Identifications without Decoy Databases

Bernhard Y. Renard,[†,‡] Wiebke Timm,[‡,§] Marc Kirchner,[‡,§] Judith A. J. Steen,[‡,|] Fred A. Hamprecht,[†,‡] and Hanno Steen*[‡,§]

Interdisciplinary Center for Scientific Computing, University of Heidelberg, Speyerer Strasse 6, 69115 Heidelberg, Germany, Proteomics Center at Children’s Hospital Boston, 320 Longwood Avenue, Boston, Massachusetts 02115, Department of Pathology, Children’s Hospital Boston and Harvard Medical School, 320 Longwood Avenue, Boston, Massachusetts 02115, and Department of Neurobiology, Harvard Medical School and F. M. Kirby Neurobiology Center, Children’s Hospital Boston, 3 Blackfan Circle, Boston, Massachusetts 02115

Using decoy databases to compute the confidence of peptide identifications has become the standard procedure for mass spectrometry driven proteomics. While decoy databases have numerous advantages, they double the run time and are not applicable to all peptide identification problems such as error-tolerant or de novo searches or the large-scale identification of cross-linked peptides. Instead, we propose a fast, simple and robust mixture modeling approach to estimate the confidence of peptide identifications without the need for decoy database searches, which automatically checks whether its underlying assumptions are fulfilled. This approach is then evaluated on 41 LC/MS data sets of varying complexity and origin. The results are very similar to those of the decoy database strategy at a negligible computational cost. Our approach is applicable not only to standard protein identification workflows, but also to proteomics problems for which meaningful decoy databases cannot be constructed.

Peptide identifications must be accompanied with confidence statements such as false discovery rates (FDRs) to ensure the reliability of results. Peptides identified with scores below the corresponding cutoffs have a higher likelihood of being random matches than deemed acceptable and thus are disregarded.[1] The standard procedure of obtaining false discovery rates, as suggested by guidelines of major proteomics journals (see, e.g., ref 1), is through decoy databases searches.[4,16] Databases of nontarget proteins are generated by, e.g., randomizing or reversing the amino acid sequence of the target proteins. The percentage of MS/MS spectra matching these artificial sequences rather than to the original sequence database is regarded as a measure of the (global) FDR. Numerous extensions have been proposed using more elaborate classifiers[9,7,8,2] or semisupervised approaches[11,3] that only need a subset of decoy hits.

While the decoy database-driven approach and its extensions have clear advantages, there are two major limitations: (i) The run time is increased by a factor of 2 since every MS/MS spectrum needs to be evaluated against all candidates in the decoy database,[10] which should have at least the same size as the original database. Thus, the database search can become a bottleneck for large-scale experiments, which is particularly true for data sets acquired on low-resolution/low-accuracy instruments for which database searches take much longer than for high accuracy data sets. Furthermore, decoy database search strategies are suboptimal for specialized applications such as the identification of crosslinked peptides where both, the target as well as the decoy database, are quadratic in size compared to standard databases.[13] (ii) A suitable decoy database cannot be generated for all applications. For instance, for error-tolerant searches, the actual, correct protein sequence is unknown and amino acid substitutions

* Corresponding author. Hanno Steen, Children’s Hospital Boston, Department of Pathology, Enders 1130, 320 Longwood Avenue, Boston, MA 02115. Fax: +1-617-730-0168. E-mail: hanno.steen@childrens.harvard.edu.

[†] University of Heidelberg.

[‡] Proteomics Center at Children’s Hospital Boston. [§] Department of Pathology, Children’s Hospital Boston and Harvard Medical School.

[|] Department of Neurobiology, Harvard Medical School and F. M. Kirby Neurobiology Center, Children’s Hospital Boston.

(1) Bradshaw, R. A.; Burlingame, A. L.; Carr, S.; Aebersold, R. *Mol. Cell. Proteomics* **2006**, *5*, 787–788.
(2) Choi, H.; Ghosh, D.; Nesvizhskii, A. I. *J. Proteome Res.* **2008**, *7*, 286–292.
(3) Choi, H.; Nesvizhskii, A. I. *J. Proteome Res.* **2008**, *7*, 254–265.
(4) Elias, J. E.; Gygi, S. P. *Nat. Methods* **2007**, *4*, 207–214.
(5) Frank, A.; Pevzner, P. *Anal. Chem.* **2005**, *77*, 964–973.
(6) Goloborodko, A. A.; Mayerhofer, C.; Zubarev, A. R.; Tarasova, I. A.; Gorshkov, A. V.; Zubarev, R. A.; Gorshkov, M. V. *Rapid Commun. Mass Spectrom.* **2010**, *24*, 454–462.
(7) Higgs, R. E.; Knierman, M. D.; Freeman, A. B.; Gelbert, L. M.; Patil, S. T.; Hale, J. E. *J. Proteome Res.* **2007**, *6*, 1758–1767.
(8) Jiang, X.; Jiang, X.; Han, G.; Ye, M.; Zou, H. *BMC Bioinf.* **2007**, *8*, 323.
(9) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aaebersold, R. *Anal. Chem.* **2002**, *74*, 5383–5392.
(10) Kim, S.; Gupta, N.; Pevzner, P. A. *J. Proteome Res.* **2008**, *7*, 3354–3363.
(11) Ka¨ll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; Maccoss, M. J. *Nat. Methods* **2007**, *4*, 923–925.
(12) Korn, E. L.; Troendle, J. F.; Mcshane, L. M.; Simon, R. *J. Stat. Plann. Inference* **2004**, *124* (2), 379–398.
(13) Maiolica, A.; Cittaro, D.; Borsotti, D.; Sennels, L.; Ciferri, C.; Tarricone, C.; Musacchio, A.; Rappsilber, J. *Mol. Cell. Proteomics* **2007**, *6*, 2200–2211.

are allowed. Thus, it cannot be assured that the decoy database contains no potential target peptides since it cannot be filtered a priori against the actual correct sequences. As a consequence, the decoy FDR might be suboptimal since correct identifications could be counted as decoy hits (see the Supporting Information for a simulation). Similarly, for *de novo* identification strategies, a sequence database is not even required and thus a decoy database cannot be constructed.

Several approaches have been proposed, which do not require a decoy database to estimate a FDR. Kim et al.[10] propose the development of scoring schemes with a realistic probabilistic output for each single spectrum. Current search engines, however, give an unreliable probabilistic output for an individual spectrum since they only have a small number of data points available for the estimate.[10] Other approaches focus on estimating the FDR from all spectra: PeptideProphet[9] applies a parametric twocomponent mixture modeling approach, which is optimized to incorporate decoy database search results but does not require them. While very powerful when at least a subset of a decoy database is available, its reliability suffers without decoy information: The estimate is based on parametric assumptions for the distributions of the incorrect and correct identifications that are not always fulfilled. Further, PeptideProphet is currently only applicable to Mascot, Sequest, and X!Tandem search results and automatically checks whether the sample size is sufficient to fit a model but does not test whether the parametric model assumptions are fulfilled. An extension proposed by Choi et al.[2] utilizes a mixture model with an unlimited number of components which can be fitted to an arbitrary distribution using a reversible jump Markov chain Monte Carlo approach. While the model is very powerful and has valuable theoretical properties, the authors themselves state that important weaknesses from a practical point of view are its major computational requirements as well as its dependency on the proper setting of priors, which can be data set dependent.

The goal of this contribution is to bridge the gap between the two mixture models proposed by Nesvizhskii et al.[9] and Choi et al.[2] We present a method which combines the high reliability of the latter model[2] achieved by allowing a mixture of distributions for the incorrect identifications with the simplicity and quick run time,[9] making it applicable to a wider range of proteomic applications without the need for decoy databases. We outline the methodology and show that the results of our approach applied to 41 proteomics data sets randomly chosen from ongoing experiments closely follow the results obtained by decoy searches and are closer to the decoy results than other methods not based on decoy database searches. Our proposed procedure has the additional advantage that it is not limited to specific search engines, but generally applicable to various scoring schemes, and it automatically tests whether the model assumptions are fulfilled.

METHODS

Overview. We propose an adaptive two-or three-component Gaussian mixture model for the score distribution. With the use of an automated heuristic, the more appropriate model is chosen and the null distribution stemming from random hits to the database is extracted. The confidence of peptide identifications is then computed based on this distribution. In a final quality control step, a ^{2}-test is used to check the model assumptions.

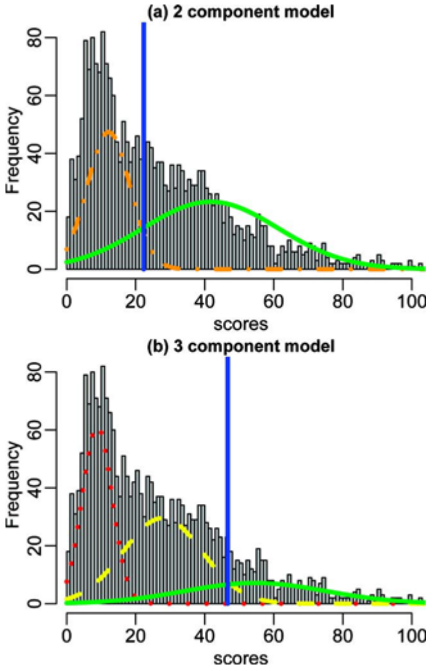

Figure 1. Schematic for the two-(a) and three-components (b) model. The orange, dashed-dotted distribution corresponds to the incorrect peptides in the two-component model, which is separated into the red dotted distribution of low-quality spectra and the yellow dashed distribution of random hits in the three-component model. The green solid distribution corresponds to the correct peptides. The blue line shows the 5% confidence cutoff as indicated by our approach.

A flowchart of the approach is given in the Supporting Information.

Mixture Model. We follow the idea of Keller et al.[9] which begins with a mixture model of two distributions. One normal distribution represents the random matches to the database, i.e., incorrect peptide identifications. A second normal distribution represents correct identifications (Figure 1a). To enhance the model flexibility and to account for low-quality spectra which cannot reliably be matched to any peptide sequence and thus show lower scores, we also fit a three component model (Figure 1b).

Number of Component Decision Criterion. Let μ) (μ_1,μ_2) and σ^2) (σ_1^2,σ_2^2) denote the parameters of the two-component mixture model and η) (η_1,η_2,η_3) and τ^2) $(\tau_1^2,\tau_2^2,\tau_3^2)$ denote the parameters of the three-component mixture model, sorted by location, with μ_1 and η_1 corresponding to the lowest average score. Our approach then applies a simple heuristic to decide between the models: When application of the three-component model leads to the distribution of the low-quality spectra being used to explain random hits, and the random hits distribution itself is shifted to partially explain correct hits, the threecomponent model does not fulfill its purpose. Since there is no need to precisely model the correct scores with more than one distribution, the two-component model is then used. In other words, the two component model is selected if $\mu_2 < \eta_2 + 2.57\tau_2$ where μ_2 is the mean of the high score component of the two component approach and η_2 and τ_2 correspond to the mean and standard deviation of the central component of the three component mixture model. The value of 2.57 corresponds to the 99.5% quantile of the normal distribution.

False Discovery Proportion. We let f_0) $N(\nu_{0,0})$ denote the normal distribution of the incorrect hits and f_1) $N(\nu_{1,1})$

denote the normal distribution of correct hits with $(\nu_0,\nu_{1,0,1})$ $)(\mu_1,\mu_2,\sigma_1,\sigma_2)or(\nu_0,\nu_{1,0,1})$ $)(\eta_2,\eta_3,\tau_2,\tau_3)$ as indicated by the number of component decision criterion. We estimate the score significance cutoff inspired by the idea of the false discovery proportion (FDP),[12] which is closely related to the concept of the global false discovery rate (FDR). The FDR is interpreted as the expected value of the FDP: FDR)IE(FDP).[14] While the FDR estimate depends on reliable estimates of both distributions, i.e., f_0 and f_1,[9] we can estimate the FDP from f_0 alone[12] and do not have to rely on accurately modeling f_1.A FDP of R % can be understood as having at maximum a probability of Rof a false discovery under the null distribution (f_0). Since we have obtained f_0, which is the null distribution, we can use it to estimate the likelihood that any score stems from the null distribution:

$$P(x \,\mathrm{g}\, s)\,)\ \int_{s}^{\infty} f_0(x)\mathrm{d}x\ )1 - \int_{-\infty}^{s} f_0(x)\mathrm{d}x$$

$$)1 - \int_{-\infty}^{s} \frac{1}{\sqrt{2\pi}_0} e^{(x-\nu_0)^2/_0^2}\,\mathrm{d}x$$

Accordingly, for a given confidence level of (100 -R)%,we can estimate the first score to be rejected using the appropriate quantile. Consequently, the FDP can also be used to estimate the expected number of false positives above the cutoff.

�^{2}-Test. The proposed procedure assumes normal distributions for the mixture model. As a final quality control step, we test whether this assumption is supported by the data with a ^{2}test and obtain a *p*-value (see the Supporting Information).

Implementation. Our approach is implemented in R using the mixtools package[17] and is available as open source code from http://software.steenlab.org/curveFDP and http://hci.iwr.uni-heidelberg.de/software.php. Analysis CPU time on a data set of approximately 5000 spectra was3sona2GHz AMD Opteron machine.

Experiments. We conducted four sets of experiments. In the first set, we compare the performance of our approach on four previously described data sets[15] to PeptideProphet (with and without decoy information) and to the original decoy approach (see the supplementary data in the Supporting Information). The data sets "Yeast1" and "Human 1" were analyzed on an LTQ-Orbitrap (Thermo Scientific) equipped with a nanoflow HPLC system (Eksigent). For the analysis of "Mouse1" and "Human 2", an LTQ equipped with a split-based microscale capillary HPLC system was used (both Thermo Scientific). Mascot (version 2.2.04) was used for database searching and scoring. A list of the data acquisition details and search parameters is given in the Supporting Information.

In a second set of experiments, we used Mascot search results to evaluate the performance of our approach and the original decoy approach on 37 additional data sets. "Human3" to "Human21" are low-accuracy data acquired on an LTQ, whereas "Human22" and "Rat1" to "Rat17" are high accuracy data acquired on an LTQ-Orbitrap (see the Supporting Information for details and data availability). All data sets were searched using Mascot and the confidence of peptide identifications was estimated using the original decoy FDR approach, the Mascot E-value FDR,[6] and our procedure, which did not use the decoy database information.

In a third experiment, we study the influence of the sample size on the confidence level estimation as well as on its variance. Using the "Human 2" data set of the first experiment, we randomly sampled between 100 and 5000 spectra from the original data set and computed the decoy FDR as well as the curve FDP.

The fourth experiment investigates the influence of the good/ bad spectra ratio on the confidence level estimation. Using the largest data set from the second experiment, "Rat 1", we applied our curve FDP approach to estimate for each spectrum the probability of belonging to the distribution of random matches or correct identifications. We then sampled spectra from both distributions according to these probabilities and with ratios of correct identifications ranging from 0% to 100%. The resulting sets of spectra were used to estimate confidence levels for the decoy FDR and our curve FDP approach.

To analyze the variation of each estimate, we used a bootstrap resampling procedure to obtain 90% empirical confidence intervals (see the Supporting Information for details).

RESULTS

Comparison with PeptideProphet without Decoy Information. While the existing nondecoy database search strategy of PeptideProphet is very conservative when evaluated against the original decoy approach and identifies on average 500 peptides or a quarter less than decoy FDRs (original decoy FDR or PeptideProphet with decoy information), our curve based approach closely follows the original decoy FDR (see Figure 2). While our curve FDP approach does show significantly more variation in the bootstrap estimates for some of the data sets ("Mouse" and "Human2"), the confidence intervals overlap for all data sets and hence indicate similar performance.

Comparison with Original Decoy FDR. Results for experiment 2 where we study 37 data sets are given in Figure 3. Overall, these results confirm the observations from the first experiment: The numbers of peptides identified by our curve FDR approach closely follow those of the decoy-based FDR. We see a slightly less conservative estimate than observed with the decoy approach for the LTQ-data sets and a slightly more conservative approach for the LTQ-Orbitrap experiments. The Mascot *E*-value FDR is by far the most conservative approach identifying on average more than 1000 peptides or 40% less than the other two approaches. The respective cutoff scores and the average precursor charge values are available in the Supporting Information. With the exception of two outliers for the "Human4" and the "Human8" data set which show unusually large variation, the variation of our curve FDP estimate is larger, but in the same overall range as the variation of the decoy FDR.

Influence of Sample Size. For the decoy FDR, we see a decrease in the variance for an increasing sample sizes whereas our approach shows constant variation (see Figure 4). For small sample sizes, the two approaches show similar variation, for larger ones, our approach has more variation. Overall, it becomes evident

[14] Pawitan, Y.; Calza, S.; Ploner, A. *Bioinformatics* **2006**, *22*, 3025–3031.

[15] Renard, B. Y.; Kirchner, M.; Monigatti, F.; Ivanov, A. R.; Rappsilber, J.; Winter, D.; Steen, J. A. J.; Hamprecht, F. A.; Steen, H. *Proteomics* **2009**, *9*, 4979–4984

[16] Weatherly, D. B.; Atwood, J. A.; Minning, T. A.; Cavola, C.; Tarleton, R. L.; Orlando, R. *Mol. Cell. Proteomics* **2005**, *4*, 762–772.

[17] Young, D.; Benaglia, T.; Chauveau, D.; Elmore, R.; Hettmansperger, T.; Hunter, D.; Thomas, H.; Xuan, F. *mixtools: Tools for analyzing finite mixture models*, R package, version 0.3.2; 2008.

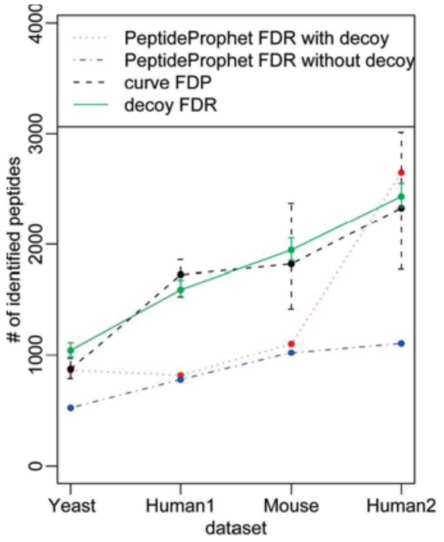

Figure 2. Comparison of the number of identified peptides at the 5% confidence level using PeptideProphet with decoy information (dotted line) and without decoy information (dashed-dotted line), our curve FDP approach (dashed line), and the decoy FDR (straight line); connecting lines are drawn for visual clarity only. Our approach and the decoy FDR show very similar behavior on all four data sets,[15] whereas PeptideProphet without decoy information is significantly more conservative.

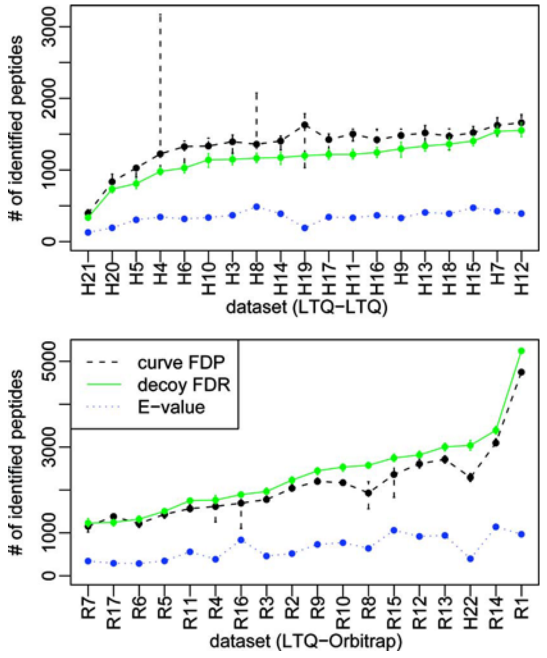

Figure 3. The performance with regard to the number of identified peptides at the 5% confidence level of our curve FDP approach (dashed line) is compared to the decoy FDR approach (solid line) and the Mascot *E*-value FDR (dotted line). Data sets are ordered by the number of identified peptides using the decoy FDR; connecting lines are drawn for visual clarity only. 90% confidence intervals based on bootstrap sampling are drawn but too narrow to see in some cases. Our approach closely follows the decoy FDR; it is less conservative with regard to the low-accuracy LTQ-data (upper plot) and more conservative with regard to the high-accuracy Orbitrap data (lower plot), while the *E*-value FDR is much more conservative.

that the estimation of confidence levels contains uncertainty which is commonly ignored at the moment. This uncertainty should thus be accounted for in subsequent analysis steps, e.g., for the inference of identified proteins.

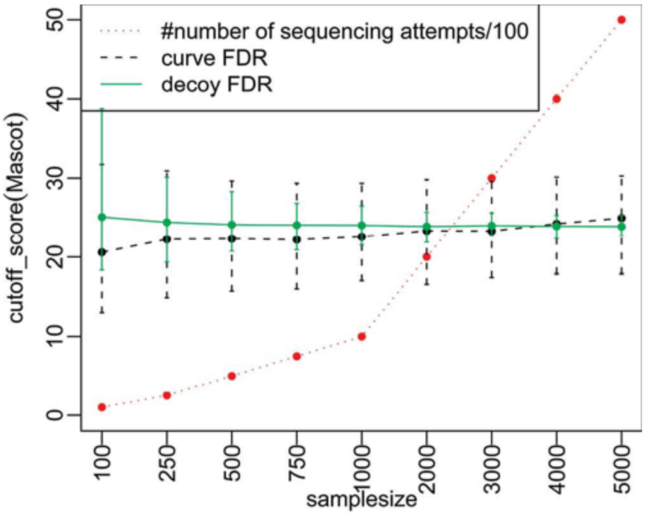

Figure 4. Analysis of the influence of the sample size on the cutoff score estimation and its variation. At the 5% confidence level, our curve FDP approach (dashed line) is compared to the decoy FDR approach (solid line). The overall number of spectra (divided by 100) is shown by the dotted line. Both approaches show a rather constant behavior with regard to the sample size, a slight downward movement can be observed for the curve FDP for smaller sample sizes and a slight upward movement for the decoy FDR. While the variation of our approach is rather constant, we see an increase of the variation for the decoy FDR for smaller sample sizes where it shows larger variation than the curve FDP approach.

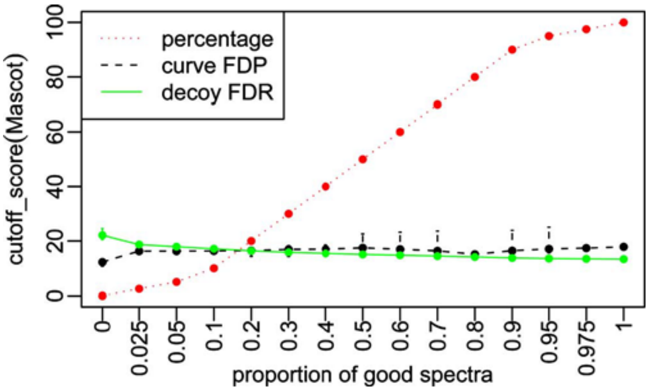

Figure 5. Influence of the ratio of good to bad spectra. Both the curve FDP (dashed line) as well as the decoy FDR (solid line) show a rather stable behavior over varying ratios of good spectra. The curve FDP estimates show more overall variation, while the decoy FDR shows a downward trend with less strict estimates for higher ratios of good spectra. Both approaches show the most extreme behavior for the case of 0% good spectra with an increase in the cutoff score for the decoy FDR and a decrease for the curve FDP.

Influence of the Ratio of Good to Bad Spectra. Both the decoy FDR and our curve FDP approach show an overall similar stability over varying ratios of good spectra. The curve FDP estimate exhibits more overall variation whereas the decoy FDR shows more bias with respect to the percentage of good spectra. Only for the extreme case of 0% good spectra, we see large deviations: the decoy FDR overestimates the cutoff score and the curve FDP underestimates it (see Figure 5), when comparing the data set with the original ratio of good and bad spectra.

DISCUSSION

Our curve FDP estimation method is built on two key assumptions. The first one is the decomposition of the score values into an appropriate number of components and the second one is assumption of normal distributions.

Determination of the Number of Components. As il-lustrated in Figure 8 in the Supporting Information, we see a strong difference in the final number of fitted components between score distributions derived from LTQ and LTQ-Orbitrap data: for the former, there is a pronounced tendency toward three-distribution-fits, whereas the latter can mostly be approximated with two-component fits. While we generally note differing characteristics of the score distributions for the LTQ and LTQ-Orbitrap data (e.g., indicated by the differing cutoff scores in Figure 7 in the Supporting Information), a possible explanation for the need of an additional component might be that fragmentation on the LTQ is not limited to peptides since no charge-state screening is performed resulting in a higher number of fragmentation spectra of nonpeptide origin equivalent to low-quality spectra.

Assumptio**n** o**f** Norma**l** Distributions**.** For the majority of data sets in this analysis, there was no evidence of a departure from normality. However, departure from this normality assumption led to stronger differences between our approach and the decoy FDR. This is underscored by the sample set which shows the largest difference between our approach and the decoy FDR approach (“Human22”): the normality assumption was rejected in 91 of 100 bootstrap resamples. However, such unreliable results are detected by the incorporated automated check of the underlying assumptions. In contrast to existing approaches, this allows specialized treatment for these cases. From our experience, results usually tend to become more conservative when the assumptions are not met. Since our FDP-based approach only relies on the distribution of the incorrect hits to be estimated correctly, it is particularly robust to departures from the normal distribution for the correct identifications. For low scores, the mixture of two normal distributions, separating overall bad spectra and the false identifications, increase the flexibility and robustness of the model. The flexibility of the proposed model is also illustrated by application to other scoring schemes (see the Supporting Information, Figure 10, for an application to PepNovo[5] scores).

CONCLUSIONS

We describe a fast, simple, and robust alternative to estimating confidence levels of peptide identifications from any protein identification algorithm. Our approach combines the simplicity of the model of PeptideProphet[9] with the flexibility provided by Choi’s approach.[2] The main advantage of our approach is that confidence levels can be computed without the need for a decoy database, thus resulting (i) in a time advantage for standard protein identification workflows and (ii) applicability to proteomics problems for which no meaningful decoy database search strategies can be formulated. An automated test for the appropriateness of the model assumptions ensures the quality of the results. Using a bootstrap approach, we demonstrated that confidence estimates for peptide identification contain significant variation, which should be accounted for in the subsequent analysis steps

ACKNOWLEDGMENT

The authors would like to thank Michael Hanselmann and Anna Kreshuk (Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Germany) for comments, suggestions, and fruitful discussions. We gratefully acknowledge financial support by the DFG under Grant No. HA4364/2-1 (B.Y.R., F.A.H.), the Alexander von Humboldt-Foundation (Grant 3.1-DEU/ 1134241 to M.K.), Robert Bosch GmbH (F.A.H.), as well as the Helmholtz Initiative for Systems Biology (F.A.H.).

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.