A Non-parametric Cutout Index for Robust Evaluation of Identified Proteins*

Oliver Serang†§¶, Joao Paulo§∥, Hanno Steen§**‡‡, and Judith A. Steen‡§‡‡

This paper proposes a novel, automated method for evaluating sets of proteins identified using mass spectrometry. The remaining peptide-spectrum match score distributions of protein sets are compared to an empirical absent peptide-spectrum match score distribution, and a Bayesian non-parametric method reminiscent of the Dirichlet process is presented to accurately perform this comparison. Thus, for a given protein set, the process computes the likelihood that the proteins identified are correctly identified. First, the method is used to evaluate protein sets chosen using different protein-level false discovery rate (FDR) thresholds, assigning each protein set a likelihood. The protein set assigned the highest likelihood is used to choose a non-arbitrary protein-level FDR threshold.

Because the method can be used to evaluate any protein identification strategy (and is not limited to mere comparisons of different FDR thresholds), we subsequently use the method to compare and evaluate multiple simple methods for merging peptide evidence over replicate experiments. The general statistical approach can be applied to other types of data (e.g. RNA sequencing) and generalizes to multivariate problems. Molecular & Cellular Proteomics 12: 10.1074/mcp.O112.022863, 807–812, 2013.

Mass spectrometry is the predominant tool for characterizing complex protein mixtures. Using mass spectrometry, a heterogeneous protein sample is digested into peptides, which are separated by various features (e.g. retention time and mass-to-charge ratio), and fragmented to produce a large collection of spectra; these fragmentation spectra are matched to peptide sequences, and the peptide-spectrum matches (PSMs)† are scored (1). PSM scores from different peptide search engines and replicate experiments can be assembled to produce consensus scores for each peptide (2, 3). These peptide search results are then used to identify proteins (4).

Inferring the protein content from these fragment ion spectra is difficult, and statistical methods have been developed with that goal. Protein identification methods (5–8) rank proteins according to the probability of their being present in the sample. Complementary target-decoy methods evaluate the proteins identified by searching fragmentation spectra against proteins that might be present (targets) and proteins that are absent ( decoys). An identified target protein counts as a correct identification (increasing the estimated sensitivity), whereas each identified decoy protein counts as an incorrect identification (lowering the estimated specificity).

Current target-decoy methods estimate the protein-level false discovery rate (FDR) for a set of identified proteins (9, 10), as well as the sensitivity at a particular arbitrary FDR threshold (11); however, these methods have two main shortcomings.

First, current methods introduce strong statistical biases, which can be conservative (10) or optimistic (12) in different settings. These biases make current approaches unreliable for comparing different identification methods, because they implicitly favor methods that use similar assumptions. Automated evaluation tools that can be run without user-defined parameters are necessary in order to compare and improve existing analysis tools (13).

Second, existing evaluation methods do not produce a single quality measure; instead, they estimate both FDR and sensitivity (which is estimated using the “absolute sensitivity,” which treats all targets as present and counts them as true identifications). For data sets with known protein contents (e.g. the protein standard data set considered), the absolute sensitivity is estimable; however, for more complex data sets with unknown contents, the measurement indicates the relative sensitivity. Even if one ignores statistical biases, there is currently no method for choosing a non-arbitrary FDR threshold, and it is currently not possible to decide which protein set is superior—one with a lower sensitivity and stricter FDR, or another with a higher sensitivity and less stringent FDR. The former is currently favored but might result in significant information loss. Arbitrary thresholds have significant effects: in the yeast data analyzed, 1% and 5% FDR thresholds, respectively, yielded 1289 and 1570 identified protein groups (grouping is discussed in the supplementary “Methods” section).

Even with such a simple data set, this subtle change results in 281 more target identifications, of which unknown subsets of 66 (0.05 × 1570 − 0.01 × 1289 ≈ 66) are expected to be
A Non-parametric Cutout Index for Robust Evaluation of Identified Proteins

Fig. 1. Schematic for non-parametric probabilistic evaluation of identified proteins. Under the supposition that the identified protein set (blue) is present, all peptides matching those proteins (also blue) might be present and have an unknown score distribution. When the correct set of proteins is identified, the remaining peptides (i.e., those not matching any shaded proteins in this figure) have a score distribution resembling that of absent peptides. Thus, the similarity of the remaining peptide score distribution (red dashed line) to the absent peptide score distribution (black solid line) determines the quality of the identified proteins.

false identifications and 215 are expected to be true identifications (281 – 66 = 215).

Here we introduce the non-parametric cutout index (npCI), a novel, automated target-decoy method that can be used to compute a single robust and parameter-free quality measure for protein identifications. Our method does not require prior expertise in order for the user to select parameters or run the computation. The npCI employs target-decoy analysis at the PSM level, where its assumptions are more applicable (4). Rather than use assumptions to model PSM scores matching present proteins, our method remains agnostic to the characteristics of present proteins and analyzes PSMs not explained by the identified proteins. If the correct present set of proteins is known, then the distribution of remaining, unexplained PSM scores resembles the decoy distribution (14). We extend this idea and present a general graphical framework to evaluate a set of protein identifications by computing the likelihood that the remaining PSMs and decoy PSMs are drawn from the same distribution (Fig. 1).

Existing non-parametric statistical tests evaluating the similarity between two collections of samples (e.g., Kolmogorov-Smirnov test, used in Ref. 14, and the Wilcoxon signed rank test) were inadequate because infrequent but significant outliers (e.g., high-scoring PSMs) are largely ignored by these methods. Likewise, information-theoretic measures, such as the Kullback-Leibler divergence, were inadequate because they require a prior on the smoothing parameter that weighs more smoothing and higher similarity against less smoothing and lower similarity (a problem reminiscent of the original compromise between sensitivity and FDR); without the application of such a prior, the optimal Kullback-Leibler divergence occurs with infinite smoothing, which will make any distributions equal, rendering them completely uninformative and thus making it impossible to distinguish one identified protein set from another. For these reasons, we derived a novel, Bayesian, non-parametric process to compute the likelihood that two continuous collections are drawn from the same distribution. It can be used to provide a robust and efficient evaluation of discoveries.

**EXPERIMENTAL PROCEDURES**

We applied the npCI to two problems for which there is no current automated and parameter-free solution. First, we estimated the number of proteins present in two previously published data sets analyzed using PeptideProphet: the UPS1 protein standard (15), and unfractionated yeast lysate (2). Protein standards like the UPS1, which contains 48 human proteins and common contaminants, often produce a crisp “elbow” (or localized point of diminishing returns) in the sensitivity-FDR relationship; we hypothesized that the npCI would select a protein set close to the sensitivity-FDR elbow. The yeast data were selected because we had no prior knowledge of the sample content. Complex mixtures produce a diffuse elbow in the sensitivity-FDR curve; thus, it is almost impossible to manually identify the point with the best sensitivity-to-FDR ratio.

Second, we evaluated simple “peptide rule” methods for merging peptide evidence from replicate experiments on HeLa lysate using Mascot.

**Sample Preparation and Peptide Search**—

**Existing Data**—Both UPS1 (15) and yeast (2) data were searched using SEQUEST (16), scored using PeptideProphet (17) (as described elsewhere (7)), and ranked according to the highest-scoring associated peptide. The UPS1 data set used a database including human target proteins and reversed decoy proteins. The yeast data set used a yeast ORF target database and a shuffled decoy database. The minimum allowed peptide score was 0; peptides reported to have lower scores in the original publication were thresholded to 0.

**HeLa Sample Preparation and LC-MS/MS**—

**Cell Lysis and Protein Extraction**—One milliliter of TBSp (50 mM Tris, 150 mM NaCl, pH 7.4, supplemented with 1X Roche Complete protease inhibitors), 1% Triton X-100, and 0.5% SDS were added to each cell pellet. Cells were homogenized with 12 passages through a 27-gauge needle (1.25 inches long) and incubated on ice with gentle agitation for one hour. The homogenate was sedimented via ultracentrifugation at 3000 × g for 5 min at 4°C, after which the supernatant was removed.

**Cell Growth and Harvesting of HeLa Cells**—In brief, HeLa cells were propagated in Dulbecco’s modified Eagle’s-F12 medium supplemented with 10% fetal bovine serum. Upon achieving >90% confluency, the growth media was aspirated from the 10-cm dish and the cells were washed three times with ice-cold phosphate-buffered saline (PBS). The cells were dislodged with non-enzymatic CellStripper, harvested following the addition of 10 ml PBS, and pelleted via centrifugation at 3000 × g for 5 min at 4°C, after which the supernatant was removed.
was reduced in 20 mM TCEP for 30 min at 56°C. After cooling to room temperature, proteins were alkylated with 1% acrylamide for 30 min.

**Acetone Precipitation**—Ice-cold 100% acetone (four sample volumes; 500 μl total) was added to 125 μl of sample (–250 μg of protein), vortexed briefly, and incubated at –20°C for three hours. The samples were then centrifuged at 20,000 × g for 30 min at 4°C. The supernatants were carefully aspirated, and the pellets were allowed to air dry at room temperature.

**Tryptic Digestion**—For tryptic digestion, the sample was resuspended in 50 mM ammonium bicarbonate, pH 8.1. The sample was incubated with 2.5 g of trypsin for 16 h. Following the incubation, the reaction was acidified with formic acid to a final concentration of 0.1% and evaporated via vacuum centrifugation. To remove interfering substances prior to mass spectrometry, peptides were isolated with C18 spin columns per the manufacturer’s instructions. Samples were again vacuum centrifuged to dryness and stored at –80°C until analysis. Prior to mass spectrometric analysis, the peptides were resuspended in sample loading buffer (5% formic acid, 5% acetonitrile, 90% water).

**Mass Spectrometry**—The peptides were subjected to fractionation using reversed-phase high performance liquid chromatography (Thermo Scientific, Waltham, MA, and Eksigent, Dublin, CA), and the gradient-eluted peptides were analyzed using a hyphenated linear trap quadrupole (LTQ) or Orbitrap Classic hybrid mass spectrometer (Thermo Scientific, Waltham, MA). The liquid chromatography columns (15 cm × 100 μm inner diameter) were packed in-house (Magic C18, 5 μm, 100 Å beads, Michrom BioResources, Auburn, CA) into PicoTips (New Objective, Woburn, MA). Samples were analyzed with a 120-min linear gradient (0%-35% acetonitrile with 0.2% formic acid), and data were acquired in a data-dependent manner in which MS/MS fragmentation was performed using the six most intense peaks of every full MS scan. Ten technical replicates were run on each of the instruments (LTQ and Orbitrap Classic).

**HeLa Mascot Search**—MGF files were converted by selecting the 200 most intense peaks with msconvert. Spectra were searched against Human UniProt (Swiss-Prot). The database was appended with common contaminants (using the common Repository of Adven-
titious Proteins, or cRAP proteins, listed in the supplementary materia-

tal). The database also included decoy entries either reversed using in-house software or shuffled with Mascot 2.3.02 (i.e. spectra were searched using target-decoy competition). These databases (regardless of reversal or shuffling) contained 71,614 proteins (including contaminants and decoys). Tryptic peptides with up to one missed cleavage were allowed. A ±1.0 Da precursor mass tolerance was used for the LTQ, and a 10 ppm mass tolerance was used for the Orbitrap (using monoisotopic masses). Both instruments used a ±0.8 Da fragment mass error tolerance. A search was performed using a reversed human decoy database and was repeated on the shuffled human decoy database in supplemental Figs. S1 and S2. Three variable modifications—deamidation (NO), propionamide (C), and oxidation (M)—were permitted.

**A Non-parametric Bayesian Approach to Evaluation**—By design, absent target PSMs should closely resemble decoy PSMs (decoy PSMs are constructed to emulate absent PSMs); we exploited this fact to propose a graphical method for evaluating protein identifications by computing the likelihood that the remaining PSMs and decoy PSMs will be drawn from the same distribution (Fig. 1). By supposing that a certain protein set is present, the method marks PSMs that might depend on those proteins and which thus might be drawn from the uncharacterized distribution of potentially present PSM scores; however, if the correct set of proteins is identified, then the unmarked PSM scores (i.e. those not dependent upon the present protein set) should be drawn from the distribution of absent PSM scores, which resemble decoy scores. Thus, the identified set of proteins can be evaluated by computing the probability that the remaining (i.e. supposed absent) PSM scores are drawn from the same distribution as the decoy PSM scores.

As described above, existing measures for computing the probabilistic or information-theoretic similarity between the remaining and decoy PSM scores were inadequate for these data. For this reason, we derived a novel Bayesian measure of probabilistic divergence. This novel Bayesian method computes a likelihood that follows directly from the supposition that the remaining target PSM scores (denoted by $D_{\text{target}}$) are drawn from the decoy score distribution (denoted by $pdf_{\text{decoy}}$) and that the decoy PSM scores (denoted by $D_{\text{decoy}}$) are drawn from the remaining target score distribution (denoted by $pdf_{\text{target}}$).

$$Pr(D_{\text{target}} | pdf_{\text{decoy}}) \times Pr(D_{\text{decoy}} | pdf_{\text{decoy}})$$

where $pdf$ denotes the probability density function. If PSM scores were restricted to two categories (low-scoring and high-scoring), then $Pr(D_{\text{target}} | pdf_{\text{decoy}})$ and $Pr(D_{\text{decoy}} | pdf_{\text{decoy}})$ could each be computed with a simple binomial; however, because there are an infinite number of possible PSM scores, we computed these probabilities using a novel method reminiscent of a Dirichlet process that can be thought of as the generalization of the binomial from two categories to an infinite number of categories. Our method requires no prior knowledge of the distribution families and has no free parameters; even the parameter specifying the degree of smoothing is chosen internally in an automated, probabilistic manner, without the use of a prior. This Bayesian method is described in greater detail in the supplementary “Methods” section.

As implemented in the python prototype, the method took less than a minute to process each data set from this manuscript (replicate data sets, such as the HeLa data, can be processed in parallel when enough cores/processors are available). The most computationally expensive task performed was the smoothing of each proposed set of remaining PSM scores (i.e. $D_{\text{target}}$) into its corresponding probability density (i.e. $pdf_{\text{decoy}}$). Although the method is currently efficient in practice, it can be substantially sped up by the use of a compiled language for numerical processing and by exploiting linearity in the smoothing process to smooth in only the new PSM scores when recomputing $pdf_{\text{decoy}}$. The increased speed will permit more efficient computation of likelihood curves more highly resolved than those shown in Fig. 2.

**RESULTS**

Figs. 2A–2F apply the npCI to the UPS1 standard (first row) and yeast (second row). The first column (Figs. 2A and 2D) compares the distribution of peptide scores from empirically absent peptides to the peptides that remain when too few, too many, and the most likely set of proteins are identified. Too few protein identifications yield remaining peptides enriched for high scores; conversely, too many protein identifications yield too few high-scoring remaining peptides. Both are implausible because they differ from the empirical distribution of absent peptide scores. The most likely protein set has remaining target peptides with a score distribution highly similar to that of decoys. The second column (Figs. 2B and 2E) shows the relationship between the likelihood and the number of identified protein groups. The third column (Figs. 2C and 2F) shows the relationship between the protein-level FDR and the
number of identified target protein groups (which approximates the sensitivity).

We also used the npCI to evaluate methods for aggregating evidence from 10 replicates of HeLa lysate processed on two instruments (LTQ and Orbitrap Classic). We considered all protein rankings requiring an identified protein match of at least $k$ peptides in at least $n$ replicate experiments, where $k$ and $n$ are free parameters. We chose the most likely protein set from the most likely ranking (Fig. 3).

**DISCUSSION**

As hypothesized, the npCI chose a protein set on the elbow of the UPS1 FDR–sensitivity curve (the most likely point has a protein FDR of $\sim 0\%$). This is remarkable, because the npCI does not use the protein-level target-decoy information in Fig. 2C. The yeast data had a less distinct elbow; nevertheless, the npCI found a distinct maximum likelihood (protein FDR of $\sim 7\%$).

Fig. 2G demonstrates the npCI’s improved reproducibility relative to the two-peptide rule, with peptide $q$-values below $1\%$. For this data set, evaluation with npCI revealed that the ranking produced by using the one-peptide rule with a variable $q$-value ($k = 1, n = 1$) is superior to the rankings obtained with the other peptide rules, including the two-peptide rule (Fig. 3). Because the samples are identical, the protein inference results should be the same for both instruments; high-mass-accuracy instruments simply have a better trade-off between sensitivity and specificity. Rather than dramati-
the npCI models the score distributions nonparametrically by the target-decoy approach only at the peptide level. Second, the npCI is very different from the scores computed and optimized by protein identification methods. This complementary approach is more robust because it makes fewer assumptions. More important, it avoids protein-level biases by using the target-decoy approach only at the peptide level. Second, the npCI models the score distributions nonparametrically by simply smoothing the density of observed examples, which allows the user to remain agnostic about the shape and family of the score distributions. The algorithm is completely automated and has no free parameters; it requires only the peptide search engine output for the experiment (or replicate experiments) and either ranked proteins or specific sets of proteins to evaluate. These ranked proteins can come from different identification methods. This manuscript evaluates proteins identified using different protein-level FDR thresholds and proteins identified using different families of peptide rules; however, the countless methods used to rank proteins could be similarly evaluated using our approach. The method can be applied as is to evaluate families of rankings produced by peptide- or PSM-level FDR cutoffs, and it even can be used to choose free parameters from parametric methods (7, 19).

Robust metrics eschewing \textit{ad hoc} parameters are crucial for the advancement of proteomics, as well as for general statistical analysis. We have described an automated, parameter-free method for evaluating discoveries; the method can be used with arbitrary peptide scores and rankings from any identification algorithm, and it can be run by a user without prior knowledge or expertise. Furthermore, the general statistical method behind the npCI can be trivially modified for application to other data, such as RNA sequencing and paired case-control analyses. The npCI method can easily be adapted to multivariate problems and applied to various (possibly correlated) types of evidence in a fully automated manner (\textit{e.g.} RNA sequencing and LC-MS/MS on the same sample).

**Acknowledgments**—Thanks to Alex Kentsis, Marc Kirchner, Peter Warren, and Mathias Wilhelm for their comments; David Sichau for helping create the web server interface; and other colleagues who gave helpful suggestions. O.S. proposed the graphical, non-parametric method for evaluating proteins using the remaining PSM scores. J.P. prepared the HeLa lysate samples and performed the LC-MS/MS analysis. O.S., J.S., and H.S. proposed the experiments in the paper.
Our free, web-based prototype implementation of npCI is available online. The UPS1 standard data, yeast data, and HeLa data (including RAW files, FASTA databases, and Mascot search parameters and raw results) are available via the Proteome Commons tranche using the following hash: asb+K4xwG/XVlc4hR8kuZ46pdR3LClQa/RgOl4/FJ9TsxHge/m97AhzBhh1c1Vbn9KdsNg+/gmovOp0AF2EHc3jUMkAAAAAAABu4Q = =.

* This work is funded in part by NIH Grant Nos. NS007473, NS006973 (O.S.), DK085835 (J.P.), NS006973 (J.S.), and GM094844 (H.S.).

† This article contains supplemental material.

‡ To whom correspondence should be addressed: Oliver Serang, Boston Children’s Hospital Enders Research Building Suite 1154 320 Longwood Avenue Boston, MA 02115. E-mail: Oliver.Serang@Childrens.Harvard.edu.

DOUBLE-SPACED REFERENCES

1. Nesvizhskii, A. I. (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J. Proteome Res 9, 5346–5357
2. Käll, L., Canterbury, J., Weston, J., Noble, W. S., and MacCoss, M. J. (2007) A semi-supervised machine learning technique for peptide identification from shotgun proteomics datasets. Nat. Methods 4, 923–928
3. Shteynberg, D., Deutsch, E. W., Lam, H., Eng, J. K., Sun, Z., Tasman, N., Mendoza, L., Moritz, R. L., Aebersold, R., and Nesvizhskii, A. I. (2011) iProphet: multi-level integrative analysis of shotgun proteome data improves peptide and protein identification rates and error estimates. Mol. Cell. Proteomics 10, page 1–15
4. Serang, O., and Noble, W. S. (2012) A review of statistical methods for protein identification using tandem mass spectrometry. Stat. Interface 5, 3–20
5. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal. Chem. 75, 4646–4658
6. Li, Y. F., Arnold, R. J., Li, Y., Radivojac, P., Sheng, Q., and Tang, H. (2008) A Bayesian approach to protein inference problem in shotgun proteomics. In Proceedings of the Twelfth Annual International Conference on Computational Molecular Biology, (Vingron, M., and Wong, L., eds.), Lecture Notes in Bioinformatics, Vol. 12, pp. 167–180, Springer, Berlin, Germany
7. Serang, O., MacCoss, M. J., and Noble, W. S. (2010) Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. J. Proteome Res 9, 5346–5357
8. Gupta, N., and Pevzner, P. (2009) False discovery rates of protein identifications: a strike against the two-peptide rule. J. Proteome Res 8, 4173–4181
9. Elias, J. E., and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods 4, 207–214
10. Reiter, L., Claassen, M., Schrimpf, S. P., Jovanovic, M., Schmidt, A., Buhmann, J. M., Hengartner, M. O., and Aebersold, R. (2009) Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. Mol. Cell. Proteomics 8, 2405–2417
11. Claassen, M., Reiter, L., Hengartner, M. O., Buhmann, J. M., and Aebersold, R. (2011) Generic comparison of protein inference engines. Mol. Cell. Proteomics 11, O110.007088
12. Barboza, R., Cociorva, D., Xu, T., Barbosa, V. C., Perales, J., Valente, R. H., França, F. M. G., Yates, J. R., and Carvalho, P. C. (2011) Can the false-discovery rate be misleading? Proteomics 11, 4105–4108
13. Yates, J. R., Park, S. K., Delahunty, C. M., Xu, T., Savas, J. N., Cociorva, D., and Carvalho, P. C. (2012) Toward objective evaluation of proteomic algorithms. Nat. Methods 9, 455–456
14. Granholm, V., Noble, W. S., and Käll, L. (2011) On using samples of known protein content to assess the statistical calibration of scores assigned to peptide-spectrum matches in shotgun proteomics. J. Proteome Res 10, 2671–2678
15. Tabb, D. L., Fernando, C. G., and Chambers, M. C. (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J. Proteome Res 6, 654–661
16. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 5, 976–989
17. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identification made by MS/MS and database search. Anal. Chem. 74, 5383–5392
18. Serang, O., Moruz, L., Hoopmann, M. R., and Käll, L. (2012) Recognizing uncertainty increases robustness and reproducibility of mass spectrometry-based protein inferences. J. Proteome Res 11, 5586–5591
19. Serang, O. (2013) Concerning the accuracy of fido and parameter choice. Bioinformatics (in press)
20. Storey, J. D. (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. Ann. Stat. 31, 2013–2035