Journal of proteome research

Letter

Cite This: J. Proteome Res. XXXX, XXX, XXX−XXX

pubs.acs.org/jpr

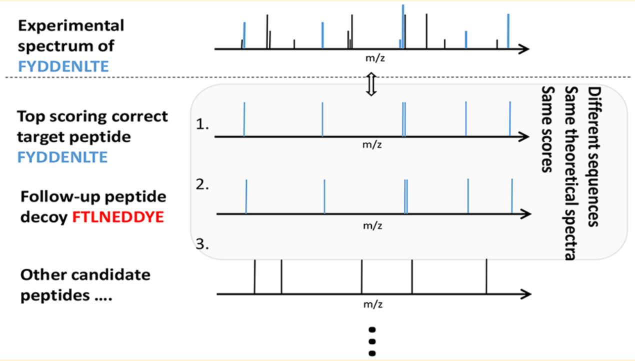
Bias in False Discovery Rate Estimation in Mass-Spectrometry-Based Peptide Identification

Yulia Danilova,[†] Anastasia Voronkova,[†] Pavel Sulimov,[†] and Attila Kertész-Farkas*

Department of Data Analysis and Articial Intelligence, Faculty of Computer Science, National Research University Higher School of Economics (HSE), 3 Kochnovskiy Proezd, Moscow 125319, Russian Federation

Supporting Information

ABSTRACT: Accurate target-decoy-based false discovery rate (FDR) control of peptide identification from tandem mass-spectrometry data relies on an important but often neglected assumption that incorrect spectrum annotations are equally likely to receive either target or decoy peptides. Here we argue that this assumption is often violated in practice, even by popular methods. Preference can be given to target peptides by biased scoring functions, which result in liberal FDR estimations, or to decoy peptides by correlated spectra, which result in conservative estimations.


Experimental spectrum of FYDDENLTE
m/z
Top scoring correct target peptide FYDDENLTE
1.
m/z
Follow-up peptide decoy FTLNEDDYE
2.
m/z
3.
Other candidate peptides ....
m/z
Different sequences
Same theoretical spectra
Same scores

INTRODUCTION

False discovery rate (FDR) control in peptide identification is often carried out using a target-decoy database search.[1] In this approach, protein sequences digested in silico result in a target peptide set, followed by the fabrication of fictitious peptides, called decoy peptides, usually by reversing or shuffling either the whole protein sequence or the nonterminal amino acids of target peptides.[2] The theory of target-decoy-based FDR control in tandem mass spectrometry has been established by He et al.,[3] and a correction to the FDR estimation was introduced by both He et al. and Levitsky et al. independently.[3,4] In brief, a collection of experimental spectra is annotated with the best-scoring peptides via database searching; then, the FDR of the results is controlled based on the amount of decoy peptides found in the search output.[2] Accurate FDR control requires distinct, independent, and roughly equal sized target and decoy peptide collections; if they are not equal, then the FDR calculation must include a correction factor for the size.[5] In this cautionary Letter, we discuss an often neglected requirement to obtain accurate FDR control and estimation, namely, that incorrect spectrum annotations ought to be assigned to either target or decoy peptides with equal likelihood. Standard raw scoring functions, like XCorr,[6] HyperScore,[7] or Andromeda,[8] meet this condition because they do not have the capacity to distinguish between target and decoy peptides. However, machine-learning-based methods that involve target and decoy peptides in training to improve spectrum annotation accuracy can attain preference toward target peptides or annotations matched to target peptides. This violates the requirement and results in a biased FDR estimation. The amount of the bias depends on the capacity of the models and on the software parameter settings.

In this Letter, we inform mass spectrometrist practitioners and developers that (1) certain program parameter combinations in peptide identification pipelines may lead to biased FDR estimation, (2) excessively parametrized score functions may become capable of distinguishing between target and decoy peptides, and (3) traditional methods to reveal bias in FDR estimation, such as null test or decoy−decoy search schemes, may not work.

The phenomenon *bias* scrutinized here is commonly discussed under the term *fairness* in machine learning and is often referred to as *machine bias* among data scientists. In general, machine bias emerges from the fact that statistical models and machine-learning methods lack common sense and logic; thus they require constant calibration and adjustment to prevent negative discrimination against legally recognized protected groups, certain social classes, races, or, in our case, decoy peptides.[9]

BIASED FEATURES

Percolator,[10] for instance, is able to discriminate certain types of decoy peptides from target ones. In a fully tryptic peptide data set (cleave after amino acids K or R if not followed by P), target peptides contain zero missed internal enzymatic cleavage sites; however, decoy peptides, generated with the peptide-reverse approach, can acquire new missed sites that are counted by a feature, called enzInt or mc, in Percolator. For instance, the enzInt feature value is 0 for peptide "FLAYRPK" because trypsin would not cleave at RP; however, it is 1 for the reversed peptide "FPRYALK" because trypsin should cleave

Received: December 28, 2018
Published: April 15, 2019

DOI: 10.1021/acs.jproteome.8b00991

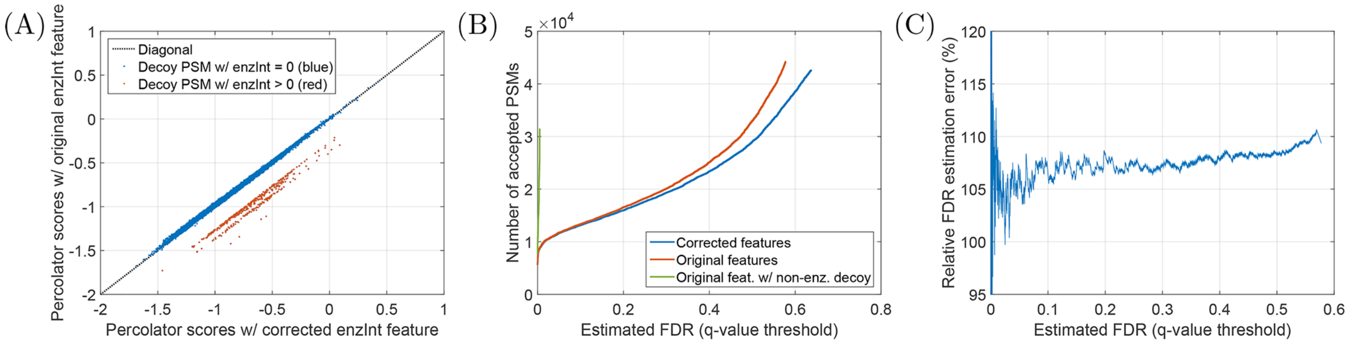


Figure 1. Bias in Percolator's features. (A) Percolator scores of decoy PSMs obtained with original features versus corrected enzInt features. (B) Percolator's results using default settings (red), using corrected enzInt features (blue), and using the nonenzymatic decoy peptide database (green). (C) Estimated error, induced by enzInt feature, in FDR estimation at various levels, calculated based on the horizontal difference between the red and blue curves from panel B.

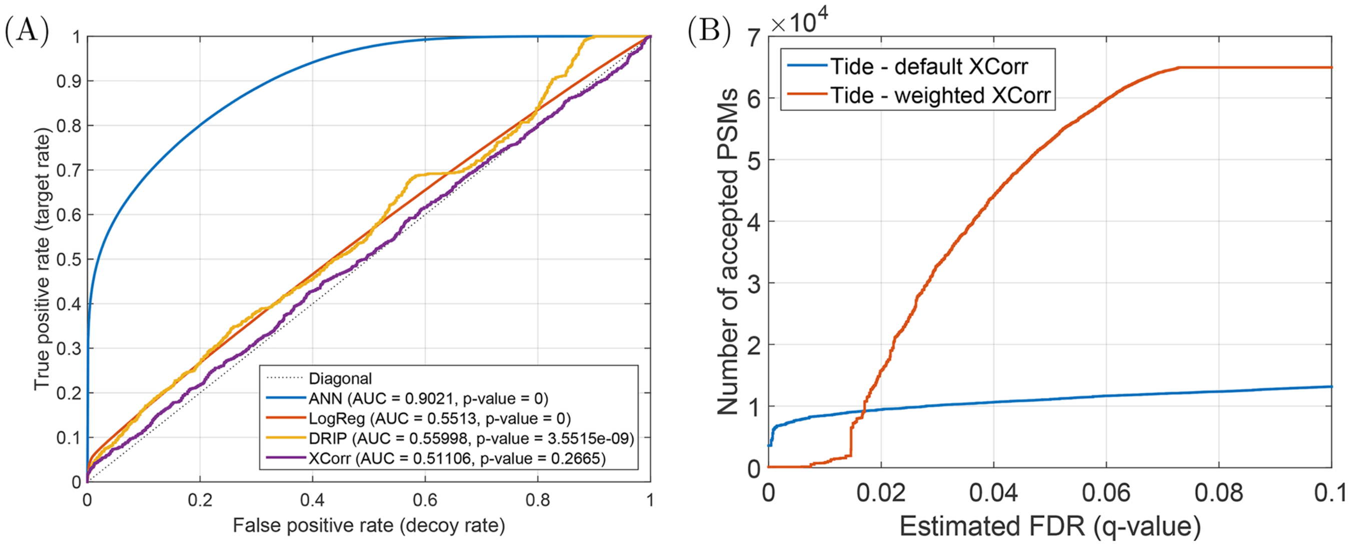


Figure 2. Discrimination of target and decoy peptides. (A) Discrimination of target against decoy peptides with ANN (blue), logistic regression (LogReg, red), DRIP (orange), and XCorr scoring function (purple) evaluated by ROC analysis. The diagonal line (dashed line) indicates an unbiased scoring function and identical distributions of the target and decoy PSM scores. *p* values of ROC analyses were obtained with the two-sided Mann–Whitney U-test. (B) Number of annotated spectra at various FDR levels using default (blue) and weighted (red) XCorr scoring functions on yeast data.

after PR. Note that PR or PK can also appear in decoy peptides generated by shuffling. Simply put, the decoy generation can introduce new types of decoy peptides, which can be indicated by a positive enzInt feature value, resulting in different distributions for the decoy peptides and the incorrect target peptides in the feature space. Percolator can capture this information and penalize many decoy peptide-spectrum matches (PSMs). Taking the yeast data set from the Percolator article, we performed an experiment with Crux toolkit v3.0[11] and Percolator using default parameter settings. (See Supplementary Note S1.1.) We generated a fully tryptic peptide data set, and we found 5226 PSMs (out of 69 705) annotated with decoy peptides having positive enzInt feature values (up to three), whereas all target PSMs (PSMs that are matched to target peptides) had zero enzInt feature values. Only these 5226 decoy PSMs (PSMs annotated with decoy peptides) received 0.3 to 0.4 lower Percolator scores than they receive when their enzInt feature is set to zero, as shown in Figure 1A. In our opinion, Percolator's enzInt feature extraction is improper because decoy PSMs receive smaller scores than target PSMs among incorrect annotations. In this experiment, that produces a 6–8% increase in errors in FDR estimation at various levels (as shown in Figure 1B,C). For further discussion, see Supplementary Note S1.2. This leads to liberal FDR estimation and thus to more accepted PSMs at various FDR (q value) thresholds. Removing proline suppression from digestion rules would be a plausible[12] and a quick fix for this problem; however, we note that other features can leak peptide-type-related information depending on parameter settings. For instance, if a decoy peptide generation procedure modified the terminal amino acids of the target peptides, then the procedure can create nonenzymatic decoy peptides for a fully enzymatic (tryptic) target peptide data set. This information is then extracted by the EnzC and EnzN features in Percolator, leading to erroneous results. (See the green curve in Figure 1B.)

■ BIASED SCORE FUNCTIONS

Scoring functions can give preference to target (or decoy) peptides during the spectrum identification search without even considering peptide labels. Contrary to expectations, the distribution of the theoretical target and decoy spectra (i.e.,

DOI: 10.1021/acs.jproteome.8b00991

spectra generated from peptide sequences in silico) is slightly different in the spectrum vector space, and a simple linear model can exploit this information. For instance, a logistic regression (LogReg) achieved a 0.551 AUC score on classifying the theoretical target and decoy peptides. The positive data comprised the theoretical spectrum vectors of target peptides, and the negative data comprised the theoretical spectrum vectors of the decoy peptides. The peptides were generated from semitryptic digestion of yeast protein sequences in which decoy peptides were produced by reversing. For more details, see Supplementary Note S2. The result means that scoring functions that take into account peak location specific weights can induce bias, whether the weights are tuned manually or are learned by a particular machine learning algorithm. The situation with vanilla artificial neural networks (ANNs) is even more dismal (or astonishing). In the same data set, an ANN achieved a spectacular AUC score as high as 0.902 in peptide classification. (See the blue and red ROC lines in Figure 2A, and see Supplementary Note S2 for details of the training and the results.) This means that scoring functions that account for peak-pair- and peak-location-specific weights can induce large biases. Perhaps deep learning methods may achieve better discrimination between target and decoy peptides. However, when the target peptides are split randomly into positive and negative sets, the ANN achieves an AUC score of only 0.543. (See Figures S3−S5 for detailed results, and see Supplementary Note S2.4 for overfitting tests.) In our opinion, this shows that the distributions of the target and decoy spectra are indeed different, and ANN does not achieve a high AUC score due to data memorization. (For more details, see Supplementary Note S2.4.) We analyzed the weights of the neural network to see the origin of the differences between the target and decoy peptides. The weight matrix of the ANN is shown in Figure S3E. We suspect that the difference may be included as a result of changing the amino acid distributions around the terminal amino acids.

The kth-order Markov chain models could be used for decoy peptide generation,[13,14] alternatively. In the case of $k = 0$, the Markov chain would be equivalent to the shuffling approach. On the one hand, increasing k would decrease the difference between the distribution of the decoy and target theoretical spectra; therefore, it would be more difficult for a machine-learning-based score function to distinguish between the types of peptide. On the other hand, higher-order Markov chains would generate decoy peptides that are more correlated to target ones.

In practice, for instance, the DRIP scoring function[15] can give preference to target peptides. In our opinion, the underlying problem is conceptual. DRIP uses a set of spectra labeled by correct target peptides as a training set to learn the parameters of a dynamic Bayesian network and to model correct spectrum-peptide alignments. However, the training procedure also learns a bit of the distribution of the target peptides, resulting in a preference for them. We showed this by testing the DRIP program in the decoy−decoy search scenario with the UPS1 spectrum data set, which contained 3368 real experimental spectra.[16] We increased the precursor mass value of every experimental spectrum by +19.0 and searched it against the yeast protein fasta file. Therefore, the true target peptides did not show up among the candidate peptides of the experimental spectra in the database searching step, resulting in only incorrect PSMs. We ran a standard database search using the XCorr scoring function, which achieved a fairly random assignment with a 0.51 AUC score. However, DRIP, trained with the original data the authors used, showed a slight preference toward target peptides when the top-20 most intense peaks were retained in the experimental spectra. The subsequent ROC analysis yielded an AUC of 0.56, which had an associated p value of 3.55×10^{-9}. The p value, obtained with the two-sided Mann−Whitney U-test, indicates this bias. (See purple lines for XCorr and orange lines for DRIP ROC in Figure 2A.)

It is actually effortless to create biased scoring functions. Consider the well-known XCorr[6] scoring function XCorr(s,t) $= s^T t$ for discretized theoretical t and experimental spectra s, which already incorporates a cross-correlation correction (defined as $s: = s - \frac{1}{151}\sum_{\tau=-75}^{75} s_\tau$, in which all vector elements shifted by τ steps in s_τ).[17] Using a weight matrix W, XCorr can be parametrized as $s^T W t$ to maximize its score for any target and minimize its score for any decoy peptides. We trained W to do so using the maximum likelihood estimation method (for more details, see Supplementary Note S3) on the yeast data set. The number of spectrum annotations in the results (possibly all garbage) skyrockets as the FDR level increases. (See Figure 2B.) In fact, the whole yeast spectrum data set (69 705 spectra) is annotated by target peptides with ~7% FDR. These results would hopefully seem doubtful to mass spectrometrists. The improved performance arises from the preference given to target peptides. They receive higher matching scores than their decoy counterparts, which results a biased FDR estimation. Note that the weighted XCorr function is equivalent to the default XCorr function using with transformed spectra $\tilde{s} = s^T W$. Therefore, spectrum preprocessing methods can give preference to target peptides, thus yielding biased FDR estimations. These examples illustrate how easily even simple models can introduce bias.

■ CORRELATED TARGET-DECOY PEPTIDES

Lastly, we discuss the possible consequences of correlated theoretical target and decoy peptides. For instance, the target FYDDENLTE and its reversed FTLNEDDYE peptides generate exactly the same theoretical spectra and the same matching scores against the peptide's experimental spectrum, resulting in a high-scoring decoy PSM. This is illustrated in Figure 3. Consequently, we argue that statistics involving high-scoring correlated decoy peptides either implicitly, as in the ratio of the top two scores (Δc_n), or explicitly, as in separated target-decoy search, can result in a conservative FDR estimation yielding fewer spectrum annotations at various FDR levels (q-value thresholds). Furthermore, the effect of the correlated target-decoy peptides is more enriched in small proteome data sets with high precursor information, because (a) correlated decoy peptides are always present among candidate peptides and (b) they face less competition from fewer others due to data sparsity.

We searched the *Plasmodium falciparum* spectrum data set[18] (for more details about the data, see Supplementary Note S4) with the standard tide-search program from the Crux toolkit v3.0, calculated the p values with the XCorr p-value (XPV) program,[19] and adjusted them with the Šidák correction. Among 12 086 matched spectra, 2639 spectra were identified at an estimated 1% FDR level. Of these, 221 (8.4%) spectra received a target and its corresponding reversed decoy peptide in their top-two scoring places. We note that only 6 out of the

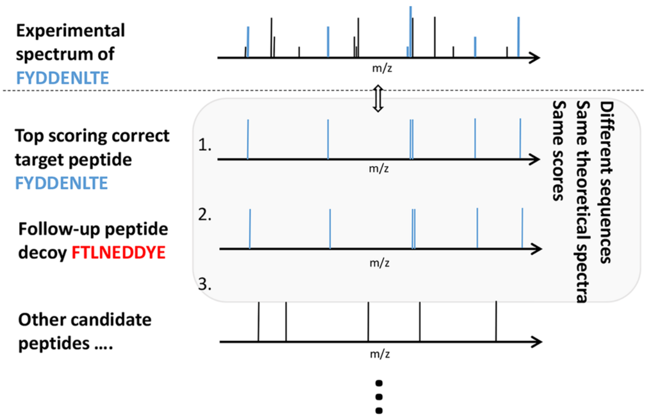

Experimental spectrum of FYDDENLTE
m/z
Top scoring correct target peptide FYDDENLTE
1.
m/z
Follow-up peptide decoy FTLNEDDYE
2.
m/z
3.
Other candidate peptides ....
m/z
Different sequences
Same theoretical spectra
Same scores

Figure 3. Illustration of the target-decoy peptide correlation. The target FYDDENLTE and the decoy FTLNEDDYE peptides produce exactly the same theoretical spectrum, and they correlate with the experimental spectrum generated by the FYDDENLTE peptide's molecule.

221 spectra had two candidate peptides or fewer. For more details of similar experiments with synthetic data, see Supplementary Note S5. The p values of the decoy PSMs confirm on a quantile−quantile plot (see in Figure 4A) that peptide-level decoy generation approaches produce more significant p values than they would be if they were coming from the true null distribution; however, p values from the protein shuffle scenario remain uniform. The distribution of the p values produced with the protein-reverse scenario act roughly uniform; however, some p values deviate. For a discussion of p-value calibration, see Supplementary Note S6. Therefore, peptide-shuffling and reverse approaches in a separated target-decoy search can produce more significant p values for the decoy peptides, yielding a conservative FDR control and fewer spectrum annotations than were obtained with the protein-shuffling method. (See Figure 4B.) We are aware of the fact that protein-shuffling decoy generation results in a conservative FDR estimation;[20−22] however, an appropriate correction for the sizes of the target and decoy peptide sets can circumvent this bias. See Supplementary Note S7 and Figure S7 for more discussion on this topic. Nevertheless, had protein shuffling resulted in a conservative FDR estimation, then peptide-level decoy generation would have resulted in an even more conservative FDR estimation. Consequently, we recommend using target-decoy competition and ignoring Δc_n features in scoring or using the averaging strategy[23] for small proteome data.

CONCLUSIONS

The increasing complexity of the peptide identification pipelines and the advancement in the mass spectrometric instrumentation yield new challenges for maintaining fair FDR estimation, and old approaches may fail to reveal bias in modern methods. For instance, the null test[24] and the decoy−decoy search approaches are ineffective at pinpointing the bias we showed in this Letter because these methods do not utilize the target peptide set (see Supplementary Note S1.2), and thus they do not consider the differences between the distribution of or the correlation between the theoretical spectra of the target and decoy peptides. Granholm et al. proposed a so-called semilabeled method to demonstrate biased features in Percolator using so-called entrapment sequences.[25] Unfortunately, the semilabeled method is not able to identify the bias in the enzInt feature because it does not involve the peptide-level decoy generation procedure. Consequently, the fairness of the FDR estimation in new methods might need to be reassessed with new validation techniques.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.8b00991.

Note S1: Additional information on Percolator search. Figure S1: Spectrum identification with Percolator and Tide in decoy−decoy search scenario. Note S2: Information on classification of target-decoy theoretical spectra. Figure S2: Model architectures of the logistic regression and the vanilla neural network models. Figure S3: ROC results on the classification of target against reversed decoy peptides. Figure S4: ROC results on the classification of target against shuffled decoy peptides. Figure S5: ROC results on the classification of target-

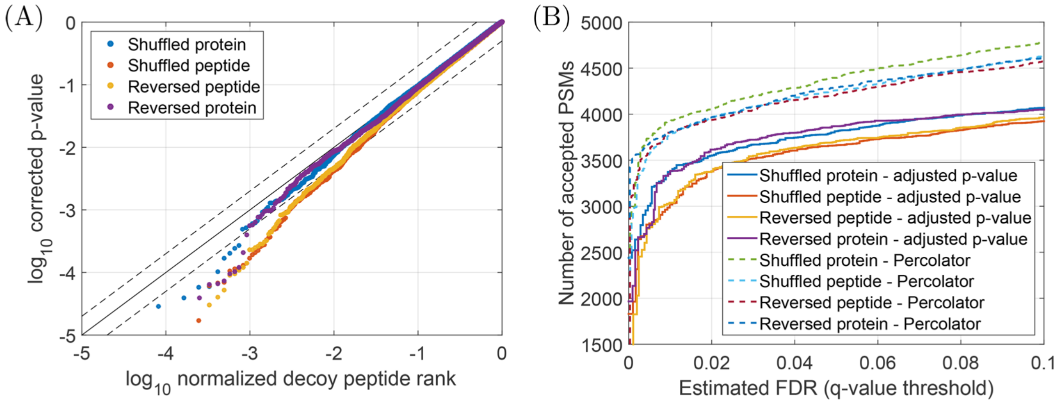


Figure 4. Correlation between target and decoy peptides obtained with peptide-reverse (yellow), peptide-shuffle (red), protein-shuffle (blue), and protein-reverse (purple) decoy generation methods in the *Plasmodium falciparum* data set. (A) Uniformity tests of p values of decoy peptides obtained with XCorr p-value program from Crux. p values were corrected with the Šidák correction. The solid diagonal line corresponds to $y = x$; dashed lines correspond to $y = 2x$ and $y = x/2$. The diagonal line indicates perfectly unbiased p values. (B) Number of accepted PSMs in separated target-decoy search using adjusted p values (XPV) (solid lines) and Percolator (dashed lines).

against-target and decoy-against-decoy peptides. Note S3: Information about training weighted XCorr score function with logistic regression. Note S4: Information about the *Plasmodium falciparum* data. Table S1: Statistics of the peptide sets using various peptide generation methods. Note S5: Notes and details of the target-decoy peptide correlation using synthetic data. Figure S6: Correlation between target and decoy peptides obtained with various peptide generation methods. Note S6: Notes on *p*-value validation. Note S7: Notes on the correction for protein-level shuffling. Figure S7: ROC analyses of protein-level shuffling (PDF)

Yeast spectrum and protein sequence data. *Plasmodium falciparum* spectrum and protein sequence data. USP1 spectrum data (ZIP)

AUTHOR INFORMATION

Corresponding Author

*E-mail: akerteszfarkas@hse.ru. Tel: +7 (499) 152-07-41.

ORCID

Attila Kertész-Farkas: 0000-0001-8110-7253

Author Contributions

[†]Y.D., A.V., and P.S. contributed equally.

Author Contributions

A.K.-F. designed the experiments, performed data analysis, and wrote the manuscript. Y.D., A.V., and P.S. developed scripts and programs and performed data analysis. All authors read and approved the final manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank William Stafford Noble for the useful discussion about Percolator's features and John T. Halloran for help with the DRIP program. We acknowledge the support of NVIDIA Corporation for its donation of the GTX Titan X GPU used for model parameter training in this research.

REFERENCES

(1) 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.

(2) Kertész-Farkas, A.; Reiz, B.; Myers, M. P.; Pongor, S. Database searching in mass spectrometry based proteomics. *Curr. Bioinf.* **2012**, *7*, 221−230.

(3) He, K.; Fu, Y.; Zeng, W.-F.; Luo, L.; Chi, H.; Liu, C.; Qing, L.-Y.; Sun, R.-X.; He, S.-M. A Theoretical Foundation of the Target-Decoy Search Strategy for False Discovery Rate Control in Proteomics. 2015, arXiv:1501.00537. arXiv.org e-Print archive. https://arxiv.org/abs/1501.00537 (accessed Dec 2018).

(4) Levitsky, L. I.; Ivanov, M. V.; Lobas, A. A.; Gorshkov, M. V. Unbiased false discovery rate estimation for shotgun proteomics based on the target-decoy approach. *J. Proteome Res.* **2017**, *16*, 393−397.

(5) Gupta, N.; Bandeira, N.; Keich, U.; Pevzner, P. A. Target-decoy approach and false discovery rate: when things may go wrong. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 1111−1120.

(6) Yates, J. R.; Eng, J. K.; McCormack, A. L.; Schieltz, D. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.* **1995**, *67*, 1426−1436.

(7) Fenyö, D.; Beavis, R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **2003**, *75*, 768−774.

(8) Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **2011**, *10*, 1794−1805.

(9) O'Neil, C. *Weapons of Math Destruction*; Crown Random House, 2016.

(10) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4*, 923.

(11) McIlwain, S.; Tamura, K.; Kertesz-Farkas, A.; Grant, C. E.; Diament, B.; Frewen, B.; Howbert, J. J.; Hoopmann, M. R.; Käll, L.; Eng, J. K.; et al. Crux: rapid open source protein tandem mass spectrometry analysis. *J. Proteome Res.* **2014**, *13*, 4488−4491.

(12) Rodriguez, J.; Gupta, N.; Smith, R. D.; Pevzner, P. A. Does trypsin cut before proline? *J. Proteome Res.* **2008**, *7*, 300−305.

(13) Colinge, J.; Masselot, A.; Giron, M.; Dessingy, T.; Magnin, J. OLAV: Towards high-throughput tandem mass spectrometry data identification. *Proteomics* **2003**, *3*, 1454−1463.

(14) Feng, J.; Naiman, D. Q.; Cooper, B. Probability-based pattern recognition and statistical framework for randomization: modeling tandem mass spectrum/peptide sequence false match frequencies. *Bioinformatics* **2007**, *23*, 2210−2217.

(15) Halloran, J. T.; Bilmes, J. A.; Noble, W. S. Learning peptide-spectrum alignment models for tandem mass spectrometry. *Uncertainty Artif. Intell.* **2014**, *30*, 320.

(16) Bish, R. A.; Fregoso, O. I.; Piccini, A.; Myers, M. P. Conjugation of complex polyubiquitin chains to WRNIP1. *J. Proteome Res.* **2008**, *7*, 3481−3489.

(17) Diament, B. J.; Noble, W. S. Faster SEQUEST searching for peptide identification from tandem mass spectra. *J. Proteome Res.* **2011**, *10*, 3871−3879.

(18) Pease, B. N.; Huttlin, E. L.; Jedrychowski, M. P.; Talevich, E.; Harmon, J.; Dillman, T.; Kannan, N.; Doerig, C.; Chakrabarti, R.; Gygi, S. P.; Chakrabarti, D. Global analysis of protein expression and phosphorylation of three stages of Plasmodium falciparum intra-erythrocytic development. *J. Proteome Res.* **2013**, *12*, 4028−4045.

(19) Howbert, J. J.; Noble, W. S. Computing exact p-values for a cross-correlation shotgun proteomics score function. *Mol. Cell. Proteomics* **2014**, *13*, 2467−2479.

(20) Wang, G.; Wu, W. W.; Zhang, Z.; Masilamani, S.; Shen, R.-F. Decoy methods for assessing false positives and false discovery rates in shotgun proteomics. *Anal. Chem.* **2009**, *81*, 146−159.

(21) Granholm, V.; Käll, L. Quality assessments of peptide−spectrum matches in shotgun proteomics. *Proteomics* **2011**, *11*, 1086−1093.

(22) 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.

(23) Keich, U.; Tamura, K.; Noble, W. S. Averaging Strategy To Reduce Variability in Target-Decoy Estimates of False Discovery Rate. *J. Proteome Res.* **2019**, *18*, 585−593.

(24) Zhang, S.-R.; Shan, Y.-C.; Jiang, H.; Liu, J.-H.; Zhou, Y.; Zhang, L.-H.; Zhang, Y.-K. The Null-Test for peptide identification algorithm in Shotgun proteomics. *J. Proteomics* **2017**, *163*, 118−125.

(25) Granholm, V.; Noble, W. S.; Kall, L. 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.* **2011**, *10*, 2671−2678.