Published in final edited form as:
*J Proteome Res*. 2008 January ; 7(1): 45–46. doi:10.1021/pr700728t.

# What's Driving False Discovery Rates?

**David L. Tabb**†
Departments of Biomedical Informatics and Biochemistry, Vanderbilt University, Nashville, Tennessee 37232-8575

## Abstract

The "Paris Guidelines" have begun the process of standardizing reporting for proteomics. New bioinformatics tools have improved the process for estimating error rates of peptide identifications. This perspective seeks to consider these advances in the context of proteomics' short history. As increasing numbers of proteomics papers come from biologists rather than technologists, developing consensus standards for estimating error will be increasingly necessary. Standardizing this assessment should be welcomed as a reflection of the growing impact of proteomic technologies.

### Keywords

False Discovery Rate; error estimation; peptide identification; MIAPE; standardization

## Perspective

It is because proteomics has become so successful that standardizing its reporting is so important. A decade ago, the quest for the longest list of proteins dominated the laboratories of many proteome technologists, and reported proteins were rarely questioned by reviewers. Since that time, the use of LC/MS/MS has become ubiquitous among those who simply want to apply this technology in answering biological or biochemical questions. This explosion of new users has amplified the importance of assessing proteomic identifications in standard ways. The editors of *Molecular and Cellular Proteomics*, among others, have done the field a great service by starting the conversation about regularized reporting of proteomic data.[1] Nothing could have declared the significance of the automobile as effectively as the passage of the first United States speed limits in 1901; the standardization of proteomic reporting is a similar evolution for our nascent field.

The "Paris Guidelines" (http://www.mcponline.org/misc/ParisReport_Final.shtml) were one of the first documents produced as part of this standardization. They require the following: "For large scale experiments, provide the results of any additional statistical analyses that indicate or establish a measure of identification certainty, or allow a determination of the false-positive rate, e.g., the results of randomized database searches or other computational approaches." Both *Journal of Proteome Research* and *Proteomics* have adopted versions of this requirement as well. How should one compute a false-positive rate, though? Should it be expressed as a probability of correctness[2] or as a false-discovery rate (FDR)?[3] Should it be reported for the set of peptides or for the set of proteins? How can searches employing decoy sequences enable estimation of error?4 What sequences should serve as decoys?5 Should the false-positive rate characterize the data set in aggregate or be associated with each item on the list? These are the questions that this field is attempting to answer.

†To whom correspondence should be addressed. Phone: (615)936-0380. Fax: (615)343-8372. david.l.tabb@vanderbilt.edu.

This issue of *Journal of Proteome Research* contains three articles that address these questions. Käll et al. (Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. *Proteome Res.* **2008**, *7*, 29–34) from the University of Washington begin their paper with an accessible review of estimating aggregate error rates at the peptide level. They continue by demonstrating that FDR estimates may be improved by incorporating the PIT (the percentage of incorrect target matches). One interesting effect of this computation is that PIT-corrected FDR estimates may be less conservative than those that are computed simplistically. The authors note that FDR computations, typically used for estimating aggregate peptide identification errors, can be extended to individual identifications in the form of q-values.

Two papers from the University of Michigan and Penn State University describe advances in the development of PeptideProphet,[2] a widely used algorithm for inferring per-peptide identification accuracy. “Semisupervised Model-Based Validation of Peptide Identifications in Mass Spectrometry-Based Proteomics” (Choi, H.; Nesvizhskii, A. I. *J. Proteome Res.* **2008**, *7*, 254–265) describes modifications to PeptideProphet that use information from target-decoy searches (such as those incorporating protein sequences in both forward and reversed orientations) to improve peptide discrimination through semisupervised machine learning. In essence, PeptideProphet can use the observed scores from decoy matches to learn more accurately how false matches distribute in score. “Statistical Validation of Peptide Identifications in Large-Scale Proteomics Using Target-Decoy Database Search Strategy and Flexible Mixture Modeling” (Choi, H.; Ghosh, D.; Nesvizhskii, A. I. *J. Proteome Res.* **2008**, *7*, 286–292) describes new modes for discerning sets of true and false peptide identifications that forego some of the assumptions of the original PeptideProphet. Instead of modeling true and false scores by manually selected distributions (such as gamma, Gaussian, or Gumbel distributions), this advance enables this software to work in a more distribution-agnostic manner. Taken together, these two articles are likely to improve the extent to which PeptideProphet can be generalized to new peptide identification algorithms, extracting a greater amount of information from noisy peptide identification data.

At this time, the proteomics field seems to be divided between groups that compute error estimates for collections of peptide identifications and groups that estimate error for individual peptides. Arguments for either can certainly be made. The amount of information available for each peptide match is limited to that reported by the database search algorithm. In the case of Sequest,[6] for example, XCorr and DeltCN are the primary metrics, with additional information coming from the preliminary score, the precursor mass error, and the number of tryptic termini observed. Since DeltCN is computed from the XCorrs of the top two matches, these two metrics clearly contain mutual information. PeptideProphet modifies and combines these subscores into a single discriminant score for each identified peptide. Discriminant scores are then mapped to error probabilities. Figures in the new papers from Nesvizhskii’s group argue convincingly that PeptideProphet’s error probabilities are surprisingly accurate. Strategies for estimating aggregate errors for peptide collections, however, can be considerably simpler than the PeptideProphet approach. This simplicity is a compelling argument in its own right. It is possible that aggregate error rates are more accurate if computed directly rather than summing over large numbers of individual peptide error rates. Directly comparing these two chief error estimation strategies is a challenge that has not yet been surmounted.

A similar discussion took place for DNA sequencing during the 1990s. DNA sequencers produced electropherograms representing the fluorescence traces observed while separating dye-terminated DNA sequence ladders. Was it sufficient to truncate the error-prone sequences interpreted from the beginnings and ends of these electropherograms, or should error be estimated for individual basecalls? Two papers from Ewing et al.[7,8] introduced “Phred,” which became the de facto standard for sequence quality assessment. Individual basecalls were associated with quality scores reflecting the probability that each base was in error. These

statistics were then found to be essential in combining sequence “reads” into larger “contig” sequences in the process of genome sequence assembly.[9] It may be that proteomics will come to a similar resolution of its debate on error rate estimation. If protein inference is substantially improved through individual peptide probabilities, this may become the dominant means by which peptide identification data is evaluated.

## Conclusion

As the field of proteomics standardizes, it seems clear that the rules for identification reporting will change. Different groups compute FDR in different ways; for example, Käll et al. prefer **Decoy/Target** to the form **(2 * Decoy)/(Decoy + Target)**.[10] Publishing in this transitional time will be aided by authors reporting their evaluation strategies as explicitly as space permits. This need not be limited to reporting algorithms and formulas. Authors should also consider reporting protein lists generated at multiple levels of statistical confidence. Groups willing to publish their raw data sets will make it possible to determine the impact of different bioinformatic pipelines in protein identification. The next several years should see the formation of a field consensus on proper proteomic reporting.

As automobiles became commonplace, other technologies evolved to support them, from gas stations to superhighways. Proteomics has already been supported by the emergence of high-resolution mass spectrometry, improved separations strategies, and isotopic labeling technologies in its short lifetime. The chaos of the first years of automobiles soon gave way to a more orderly world of traffic lights, speed limits, and parking lots. Standardization will bring this same sense of order and reliability to proteomics. Mass production made automobiles available at accessible prices, making cars and trucks ubiquitous. Mass spectrometry has only begun the market penetration it will one day achieve. As the proteomics community develops a common language to characterize identification data, it is helping to set a course to this future.

## Acknowledgments

This work was supported by NIH/NCI 1 R01 CA126218-01 and NIH/NCI 1 U24 CA126479-01.

## References

1. Bradshaw RA, Burlingame AL, Carr S, Aebersold R. Reporting protein identification data: the next generation of guidelines. Mol Cell Proteomics 2006;5(5):787–788. [PubMed: 16670253]
2. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 2002;74(20):5383–5392. [PubMed: 12403597]
3. Moore RE, Young MK, Lee TD. Qscore: an algorithm for evaluating SEQUEST database search results. J Am Soc Mass Spectrom 2002;13(4):378–386. [PubMed: 11951976]
4. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 2007;4(3):207–214. [PubMed: 17327847]
5. Higdon R, Hogan JM, Van Belle G, Kolker E. Randomized sequence databases for tandem mass spectrometry peptide and protein identification. OMICS 2005;9(4):364–379. [PubMed: 16402894]
6. Eng JK, McCormack AL, Yates JR 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. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 1998;8(3):186–194. [PubMed: 9521922]
8. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 1998;8(3):175–185. [PubMed: 9521921]
9. Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES. ARACHNE: a whole-genome shotgun assembler. Genome Res 2002;12(1):177–189. [PubMed: 11779843]

10. Zhang B, Chambers MC, Tabb DL. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. J Proteome Res 2007;6(9):3549–3557. [PubMed: 17676885]