Improving Peptide and Protein Identification Rates Using a Novel Semi-Supervised Approach in Scaffold

Brian C. Searle, Caleb J. Emmons and Bryan Head

Proteome Software Inc., 1340 SW Bertha Blvd, Suite 10, Portland, OR, 97219-2039, United States
E-mail: Brian.Searle@ProteomeSoftware.com, Telephone: (503) 244-6027, Fax: (503) 245-4910

Abstract

Optimally validating peptide identifications made by database search engines remains an open problem in the field of proteomics. Current algorithms rely on target/decoy strategies to validate peptide hits and determine a given false discovery rate (FDR) threshold. Probabilistic algorithms, such as PeptideProphet, improve upon this by assessing a probability of assignment rather than fixed thresholds, allowing marginal but valid identifications to contribute to protein hits. Percolator is a probabilistic algorithm that goes further by generating new classifiers for each experiment using a variety of statistics about target and decoy identifications. We present a new approach to validate peptide identifications with discriminant scoring using a naïve Bayes classifier generated through iterative rounds of training and validation to optimize training data set choices. Peptide probabilities are assessed using a Bayesian approach to local FDR (LFDR) estimation. Rather than simply using mass accuracy as a term in discriminant score training, peptide probabilities are modified by likelihoods calculated from parent ion delta masses. Finally, we propose a strategy for protein-level FDR optimization by considering a multi-dimensional FDR "landscape", rather than just a single score metric. Improvements over current methods are demonstrated using Mascot 2.4 analysis of the public ABRF iPRG2009 orbitrap data set [1] in Figure 1a and 1b to the right.

(Improvement A) Naïve Bayes Classification

Many database search engines report multiple scores (such as Sequest's XCorr, Sp, and DeltaCn) and converting these scores into a single discriminant is central to any successful peptide validation tool. Trained classifiers such as Linear Discriminant Analysis (LDA) used by PeptideProphet [2] or Percolator's [3] Support Vector Machines (SVMs) are powerful algorithms for creating near optimal separation between "correct" and "incorrect" peptide identifications. While PeptideProphet and Percolator's impact upon the field of MS/MS based proteomics is undisputed, we've found that in certain circumstances these algorithms can be overly aggressive or yield misleading results. Instead of PeptideProphet's LDA or Percolator's SVM classifier, we use log-likelihood ratios generated by naïve Bayes classifiers to discriminate between target and decoy hits. Naïve Bayes utilizes a simple approach that calculates the score distributions of targets and decoys in the training set. Peptide matches are considered on the ratio of likelihoods that their scores fit closer to the target distributions over the decoy distributions. While naïve Bayes classifiers are mathematically oversimplified in comparison to LDA and SVMs, they have been found to perform remarkably well in real world situations specifically due to their robustness to over-fitting. Our analysis suggests that this wisdom holds true with peptide identification data.

Selection of training data is critical to the success of any classifier. With target/decoy analysis we have a perfect set of "incorrect" assignments by considering only scores for decoy matches. However, a significant percentage of target matches are "incorrect" as well. It is imperative that we train our classifier using as many "correct" target matches as possible to improve generalization, while simultaneously including as few "incorrect" target matches to improve specificity. We iteratively test training data set sizes (and resulting classifiers) to hone in on the optimal number of spectra to include to avoid training with incorrect identifications assigned to target proteins. Preferred classifiers have the same number of target PSMs used for training to the number of target matches found at a 2% peptide FDR level.

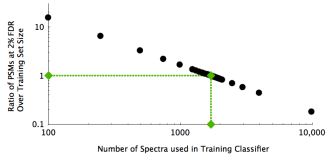
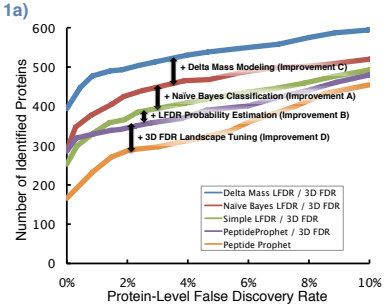


Figure 2:
Choosing the right training data is very important. We iteratively test the effect of training set size on the number of identified spectra to find the optimal 1:1 ratio of "correct" training spectra to confidently classified PSMs. We match this number with an equal number of "incorrect" decoy PSMs.

1a)



1b)

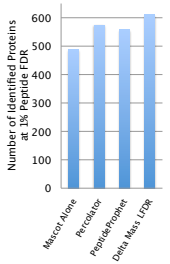


Figure 1a and 1b:

a) Receiver Operator Characteristic (ROC) curves demonstrating the iterative improvement of the four different algorithms described in this poster over PeptideProphet (as implemented in Scaffold 3) alone. Scaffold 4 combines all of these improvements (described in A, B, C and D), showing over a 50% improvement at the 1% protein FDR level in the iPRG2009 data set.

b) Comparisons of PeptideProphet (Scaffold 3) and Delta Mass LFDR (Scaffold 4) with Mascot and Percolator at a 1% peptide FDR. These comparisons can only be done using peptide FDR due to constraints in Mascot 2.4 and Percolator.

(Improvement B) LFDR Probability Estimation

We derive posterior peptide probabilities using LFDR estimates in a Bayesian framework similar to that described in [3,4] using the following model:

$$p(+ \mid D_i) = \left(1 - \frac{p(r,D_i)}{p(r)}\right) \Big/ \left(1 - p(r,D_i)\right)$$

where $p(r,D_i)$ is the percentage of decoys in a score bin D_i, and $p(r)$ is the percentage of decoys in the database. We use variable width bins rather than discrete sizes to the number of values in each bin constant. This gives more refined assessments of probability in score areas with more values, while simultaneously ensuring that LFDR estimates stay reasonable with fewer.

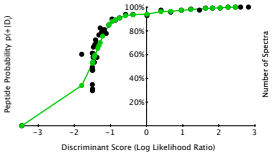


Figure 3:
Local FDR estimates for the peptide probability are plotted in black. Curve fitting using a lowpass filter (green) insures monotonicity and a continuous fit.

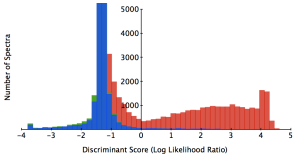


Figure 4:
Histograms of 1% protein FDR "correct" PSMs (red), "incorrect" PSMs (blue) and decoy PSMs (green). Note that decoys track perfectly with "incorrect" PSMs.

(Improvement C) Delta Mass Modeling

While many peptide validation models (including PeptideProphet and Percolator) use mass accuracy as a score classifier term, we incorporate it into the peptide probability as a Bayesian term. We define the delta mass score for peptide identification as the neutron adjusted mass accuracy in parts per million (Δppm):

$$\Delta ppm_{prec} = \frac{1{,}000{,}000 * (\Delta m_{prec} - round[\Delta m_{prec}/1.00273] * 1.00273)}{m_{prec}}$$

This calculation eliminates inconsistencies due to inaccurate parent ion triggering. Δppm values are binned using variable width bins similar to that used in the basic LFDR probability model. The likelihood that a peptide identification within a particular Δppm bin (b) is correct is notated as $p(\Delta ppm_b \mid +)$, and can be computed as the ratio:

$$p(\Delta ppm_b \mid +) = \frac{\sum_{\Delta ppm_{bin}} p(+ \mid \Delta ppm_{bin})}{\sum_{\Delta ppm_{bin}} p(+ \mid \Delta ppm_{bin})}$$

The likelihood for an incorrect identification within a particular Δppm bin can be computed in an analogous way.

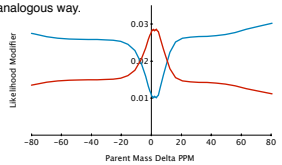


Figure 5:
Likelihood values for "correct" (red) and "incorrect" PSMs (blue). If the mass spectrometer is calibrated well then "correct" PSMs should cluster around a Δppm of 0, while "incorrect" PSMs should be scattered randomly across all of the bins.

Bayes's law can be sequentially extended to use $p(\Delta ppm_b \mid +)$ to modify $p(+ \mid D_i)$:

$$p(+ \mid D_i, \Delta ppm_b) = \frac{p(\Delta ppm_b \mid +)p(+ \mid D_i)}{p(\Delta ppm_b \mid +)p(+ \mid D_i) \; + \; p(\Delta ppm_b \mid -)p(- \mid D_i)}$$

This modified $p(+ \mid D_i, \Delta ppm_b)$ should be more accurate than the original $p(+ \mid D_i)$ because it incorporates not only the search engine score, but the parent ion mass accuracy as well.

(Improvement D) 3D FDR Landscape Tuning

Peptide probabilities are combined into protein assignments using the ProteinProphet algorithm followed by hierarchical protein clustering. Scaffold employs three independent filters: minimum protein probability, peptide probability, and minimum count of assigned peptides. We map the number of target and decoy protein identifications assigned across the entire landscape of these three filters in 1% increments, enabling us to determine an optimal list of protein identifications at a given protein FDR threshold. Often there is a balance where an optimal list may be achieved using a stringent peptide probability threshold combined with a weak protein probability threshold, a stringent protein and a weak peptide threshold, or a combination of the two. Scaffold 4 will automatically choose the best strategy, but some experimental designs may dictate that one filtering method is most appropriate.

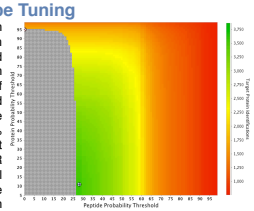


Figure 6:
The number of protein identifications at various peptide and protein probability thresholds. Thresholds that provide lists below a given FDR cutoff are suppressed in gray.

[1] https://proteomecommons.org/dataset.jsp?i=66137
[2] Keller, A.; Nesvizhskii, A. I.; Kolker, E.; and Aebersold, R. *Anal Chem.* 2002, *74*, 5383-92
[3] Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; and MacCoss, M. J. *Nat Methods.* 2007, *4*, 923-5
[4] Hather, G.; Higdon, R.; Bauman, A.; von Haller, P. D.; and Kolker, E. *Proteomics.* 2010, *10*, 2369-76