Direct maximization of protein identifications from tandem mass spectra

Marina Spivak
Department of Machine Learning,
NEC Laboratories America,
4 Independence Way,
Princeton, NJ 08540
marina@nec-labs.com

Jason Weston
Department of Machine Learning,
NEC Laboratories America,
4 Independence Way,
Princeton, NJ 08540
jasonw@nec-labs.com

Michael J. MacCoss
Department of Genome Sciences
1705 NE Pacific Street
Box 355065
University of Washington
Seattle, WA 98195
maccoss@u.washington.edu

William Stafford Noble
Department of Genome Sciences
Department of Computer Science and Engineering
1705 NE Pacific Street
Box 355065
University of Washington
Seattle, WA 98195
william-noble@u.washington.edu

September 24, 2009

Abstract

The goal of many shotgun proteomics experiments is to identify with high confidence as many proteins as possible from a complex biological mixture. Existing solutions to this problem typically subdivide the task into two stages, first identifying a collection of peptides with a low false discovery rate, and then inferring from the peptides a corresponding set of proteins. In contrast, we formulate the protein identification problem as a single optimization problem, which we solve using machine learning methods. The resulting algorithm directly controls the relevant error rate, can incorporate a wide variety of evidence and, for complex samples, provides 24–74% more protein identifications than the current state of the art.

The problem of identifying proteins from a collection of tandem mass spectra involves assigning spectra to peptides, using either a *de novo* or database search strategy, and then inferring the protein set from the resulting collection of peptide-spectrum matches (PSMs). In practice, the goal of such an experiment is to identify as many distinct proteins as possible at a specified false discovery rate (FDR). Thus, an analysis pipeline that identifies a set of 100 proteins, including one incorrect identification, is clearly inferior to a competing pipeline that finds, from the same collection of spectra, a set of 200 distinct proteins with two incorrect identifications.

Much previous work in the context of shotgun proteomics analysis has focused on controlling error rates at the level of PSMs or peptides [Moore et al., 2002, Choi et al., 2008, Käll et al., 2008, Choi and Nesvizhskii, 2007, Keller et al., 2002, Fenyo and Beavis, 2003, Geer et al., 2004, Perkins et al., 1999, Fitzgibbon et al., 2007, Elias and Gygi, 2007, Huttlin et al., 2006], rather than the protein-level FDR. However, especially in the context of a deeply saturated experiment, the PSM-level error rate may be significantly lower than the peptide-level error rate, which is in turn lower than the protein-level error rate [Adamski et al., 2005]. For example, consider a collection of 1000 spectra that map to 100 distinct peptides with a 1% false discovery rate. This 1% false discovery rate corresponds to 10 incorrectly mapped spectra, each of which is likely to map to a different, incorrect peptide. Thus, the PSM error rate of 1% corresponds to a peptide error rate of $10/110 = 9\%$. A similar inflation of error rate will occur if we move to the protein level.

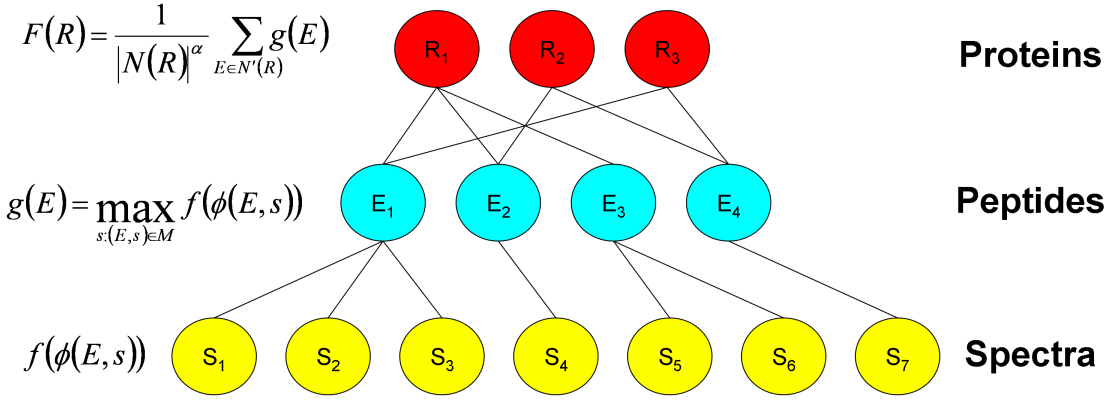

Figure 1: **Barista.** The tripartite graph represents the protein identification problem, with layers corresponding to spectra (gold), peptides (blue) and proteins (red). Barista computes a parameterized nonlinear function $f(\cdot)$ on each PSM feature vector $\phi(E, s)$. The score assigned to a peptide is the maximum PSM score associated with it. The score assigned to a protein is a normalized sum of its peptide scores.

In this work, we describe a machine learning method to directly optimize the desired quantity, the total number of proteins identified by the experiment. We use a target-decoy search strategy, searching each spectrum against a database of real (target) peptides and shuffled (decoy) peptides. We then train a supervised learning algorithm to induce a ranking on the combined set of target and decoy proteins, learning the parameters of the model so that the top of the ranked list is enriched with target proteins. Compared to existing methods [Nesvizhskii et al., 2003, Alves et al., 2007, Zhang et al., 2007, Li et al., 2008], the direct approach offers two primary advantages. First, by formulating an optimization problem that operates at the protein level, our approach correctly controls the relevant error rate. Second, our approach does not filter any PSMs at any stage of the analysis, with the motivation that even low scoring PSMs can carry information about a protein's presence when considered in the context of other PSMs belonging to this protein.

The protein identification problem can be represented as a tripartite graph, with layers corresponding to spectra, peptides and proteins (Figure 1). An edge from a spectrum to a peptide indicates that the database search procedure assigned a high score to the peptide. In general, more than one spectrum may be assigned to a single peptide. An edge from a peptide to a protein implies that the peptide occurs in the protein. This peptide-to-protein mapping is many-to-many because each protein contains multiple peptides and each peptide may appear in more than one protein. The input to the problem is the tripartite graph, with a fixed set of features assigned to each peptide-spectrum match. In this work, we represent each PSM using 17 features (Supplementary Table 1) that collectively describe properties of the spectrum and of the peptide, as well as the quality of the match between the observed and theoretical spectra. The desired output is a ranking on proteins, with proteins that are present in the sample appearing near the top of the ranked list.

We solve the protein identification problem using a target-decoy training strategy. Decoy databases have been used in shotgun proteomics for two complementary purposes: (1) to provide false discovery rate estimates for peptide identifications [Moore et al., 2002, Peng et al., 2003, Käll et al., 2008], and (2) to learn to discriminate between correct and incorrect PSMs produced by a database search algorithm [Käll et al., 2007, Spivak et al., 2009, Choi and Nesvizhskii, 2008]. In the current work, we produce a decoy database by randomly permuting the amino acids in each target protein. We then merge the target and decoy databases, and we search each spectrum against the combined target-decoy database, retaining a fixed number of top-scoring peptides for each spectrum. For the purposes of training our ranking function, the target proteins are labeled as positive examples, whereas decoy proteins are labeled as negative examples.

In any supervised learning procedure, we must ensure that the data used to train the model is kept

apart from the data used to test the model. Therefore, to produce a protein ranking for a given data set, we use a cross-validation procedure to train and test a collection of models. First, we identify connected components in the given tripartite graph, and we subdivide the graph into n approximately equal-sized tripartite graphs, ensuring that no edges are eliminated in the process. We then train a model using $n-1$ of the subgraphs as a training set and one subgraph as the test set, and we repeat this train/test procedure using each subgraph as one test set. In the end, we merge the scored proteins from the various test sets, yielding a ranking on the entire set of proteins. We refer to this cross-validated train/test procedure as Barista. Software implementing this procedure is available as part of the Crux software toolkit (`http://noble.gs.washington.edu/proj/crux`).

The Barista model consists of three score functions, defined with respect to PSMs, peptides and proteins (see Figure 1). The PSM score function is a nonlinear function of the 17 input features, defined by a two-layer neural network with three hidden units. The score function of a protein is a normalized sum of its peptide scores, and the score function of a peptide is the maximum PSM score associated with it. We then learn a protein score function that performs well on the target-decoy training task using a simple iterative update procedure (Supplementary Algorithm 1). During training, the weights of the neural network that define the PSM score function are optimized, because the PSM score is part of the protein score calculation. These weights are the only adjustable parameters of the learning task.

Notably, the protein score function includes a per-protein normalization factor $|N(R)|$, which is defined as the number of peptides that occur in protein R, assuming enzymatic cleavages. Barista uses the theoretical number of peptides, rather than the number of observed peptides, because the theoretical peptide number implicitly supplies an additional piece of information: how many peptides appear in the protein but have not been matched by any spectrum. This information allows Barista to penalize longer proteins, which are more likely to receive random matches during the database search procedure.

When reporting the set of proteins identified by Barista, we eliminate all redundant proteins as described in [Zhang et al., 2007]. Specifically, we merge all proteins that contain a common set of identified peptides into a single meta-protein, and count it as a single protein in all the reported results. Likewise, we identify proteins whose peptides are completely contained in another protein, and we report only the larger protein. For *degenerate peptides*—peptides that appear in several proteins—Barista produces a non-parsimonious solution, assigning these peptides equally to each matching protein (see Supplementary Materials for details). We chose this route because the mass spectrometry experiment does not provide any information allowing one to assign a peptide to one of its proteins, nor does it rule out the possiblity that both of these proteins were present in the original mixture. If we consider a protein score to be a measure of belief that the protein is correctly identified, then assigning a peptide to a higher scoring protein (either fully or using weights) simply boosts this belief arbitrarily.

We compared ProteinProphet [Nesvizhskii et al., 2003], IDPicker 2.0 [Zhang et al., 2007, Ma et al., 2009] and Barista using four previously described data sets [Käll et al., 2007]. The first set consists of spectra acquired from a tryptic digest of an unfractionated yeast lysate and analyzed using a four-hour reverse phase separation. In this data set, peptides were assigned to spectra by using the Crux implementation of the SEQUEST algorithm [Park et al., 2008], with tryptic enzyme specificity and with no amino acid modifications enabled. The search was performed against a concatenated target-decoy database composed of open reading frames of yeast and their randomly shuffled versions. The top three PSMs were retained for each spectrum. The next two data sets were derived in a similar fashion from the same yeast lysate, but treated by using different proteolytic enzymes, elastase and chymotrypsin. The database search was performed using no enzyme specificity and with no amino acid modifications enabled. The fourth data set is derived from a *C. elegans* lysate digested by trypsin and processed analogously to the tryptic yeast data set. The PSMs obtained from the search were subsequently analyzed by Barista, ProteinProphet and IDPicker (see supplement).

Figure 2A–D demonstrates that Barista successfully identifies more target proteins than ProteinProphet and IDPicker across a wide range of false discovery rates and across all four data sets. Note that, for simplicity, we report results with respect to absolute numbers of false positives; however, the results are very similar if we use an FDR-based threshold (Supplementary Figure 1). At an FDR threshold of 1%, Barista identifies 25% more proteins than ProteinProphet (1351 compared to 1079) and 24% more than IDPicker (1351 compared to 1084) for the "yeast trypsin" data set. For worm, the corresponding improvements are 74% and 45%, respectively (see Supplementary Table 2 for details). ProteinProphet does not support training

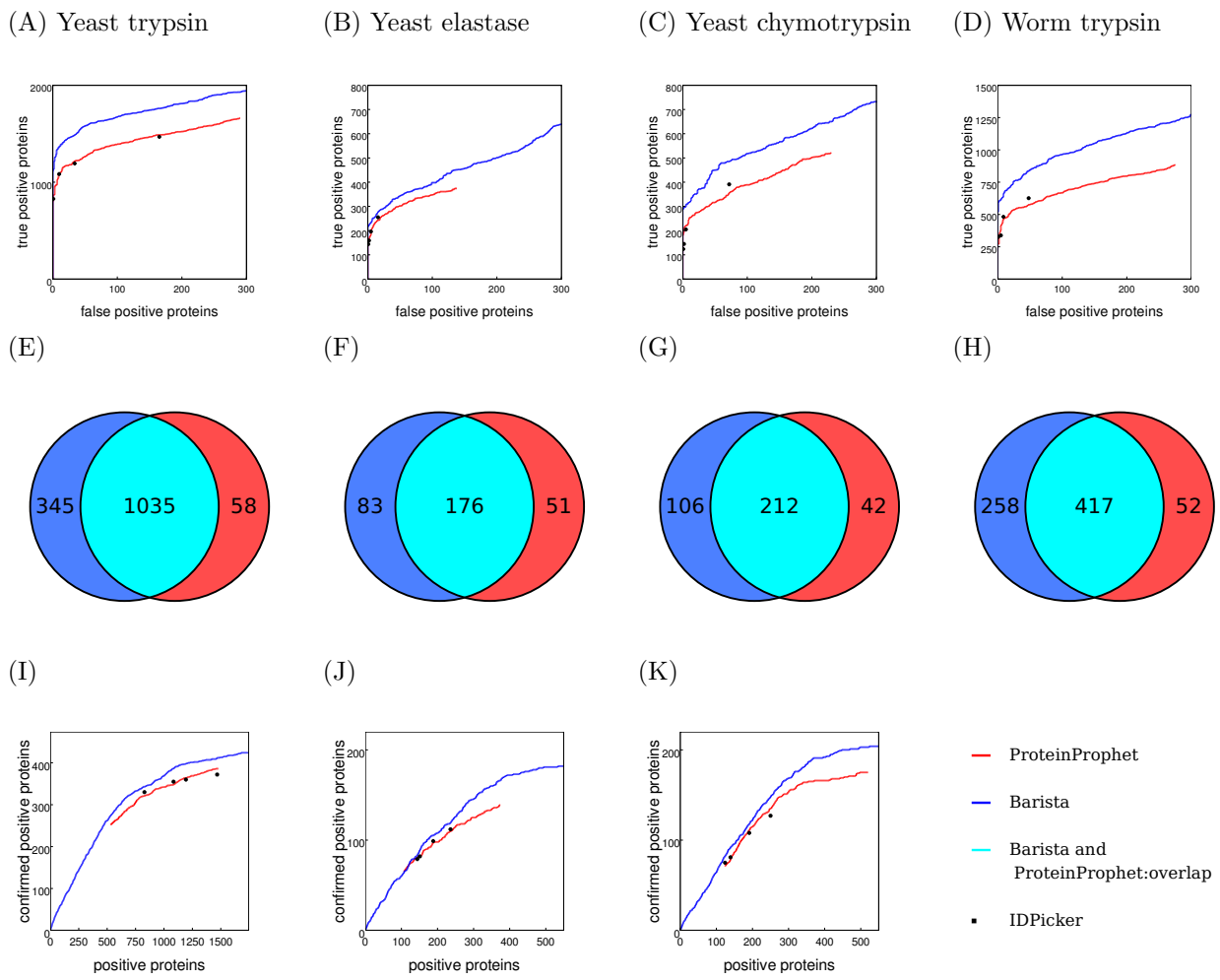

Figure 2: **Comparison of ProteinProphet, IDPicker and Barista.** Panels (A)-(D) panel show receiver operating characteristic (ROC) curves, plotting the number of true positive protein identifications as a function of the number of false positive identifications as we vary the threshold. Panels (E)-(G) show the overlap between proteins identified by Barista and ProteinProphet when 10 false positives are allowed. Panels (I)-(K) plot the number of externally validated yeast proteins identified by Barista and ProteinProphet as a function of the total number of proteins identified by the method. Note that, in each case, ProteinProphet only induces a partial ranking because many proteins receive probabilities of 1.0 or 0.0.

a model on one data set and then applying the trained model to a separate data set; therefore, to allow a fair comparison of algorithms, the results in Figure 2 are based on training and testing on the entire data set. However, Supplementary Figure 2 demonstrates that, even when we split the data into four equal parts and train on only 3/4 of the data, Barista still performs better on the held-out test set than ProteinProphet in nearly every case. Furthermore, Supplementary Figures 2 and 3 provide evidence that Barista is not overfitting the training set, because the performance on the test set is similar to the performance on the training set.

Because Barista does not arbitrarily break peptide-level ties when the data does not provide sufficient information to do so, a direct comparison to a parsimonious method such as IDPicker may give Barista an unfair advantage. Therefore, we implemented a more parsimonious version of Barista (*p-Barista*, see Supplement for details), in which each peptide is mapped to a single protein. Supplementary Figure 4 shows that p-Barista consistently outperforms both IDPicker and ProteinProphet. This result is particularly impressive, given that we enforced parsimony only as a post-processing step.

In addition to target-decoy validation, we compared the ability of ProteinProphet, IDPicker and Barista to recover proteins that had been identified in yeast cells using alternative experimental methods. For this purpose, we gathered a set of 1295 proteins whose presence in yeast cells during log-phase growth is supported by three independent assays: (1) mRNA counts established by microarray analysis [Holstege et al., 1998], (2) incorporating antigen specific tags into the yeast ORFs and detecting the expression of the resulting protein with an antigen, and (3) incorporating the sequence of green fluorescent protein into the yeast ORFs and detecting the resulting fluorescence [Ghaemmaghami et al., 2003]. Figure 2I–K shows that, across the three yeast data sets, Barista's ranked list of proteins is more highly enriched with these externally validated proteins than ProteinProphet's and IDPicker's ranked lists.

To better understand the relationship between the proteins identified by ProteinProphet and Barista, we computed the overlap between the sets of proteins identified as true positives by the two methods at a fixed number of false positives (Figure 2E-H). For all four data sets, ProteinProphet and Barista identify many of the same proteins. We further investigated the composition of the non-overlapping sets in the yeast datasets identified by ProteinProphet and Barista by checking them against the proteins established by the alternative experimental methods described above. For trypsin-digested yeast, the percent of non-overlapping proteins also identified by the alternative experimental methods was 18% for Barista and 10% for ProteinProphet. For elastase, these percentages were 32% and 21%, respectively, and for chymotrypsin, 49% and 37%. Thus, in each case, the external validation more strongly supports the Barista identifications than the ProteinProphet identifications.

Next we investigated in detail the properties of the proteins identified by one method and not the other for the tryptic yeast data set. Figure 3 shows a schematic of the 59 ProteinProphet-only and 331 Barista-only proteins identified at 10 false positives in the tryptic yeast data set. Note that, to allow direct comparison of the two sequence sets, we have colored all the identified peptides according to their PeptideProphet probabilities, even though Barista does not use these probabilities. We have also included in the figure all PeptideProphet probabilities, even though ProteinProphet was set to only consider probabilities > 0.05.

A general feature of protein identification algorithms is that they are more likely to successfully identify longer proteins, simply because such proteins contain more peptides. Barista is less biased against short proteins compared to ProteinProphet. The average length of the 1077 proteins identified by both methods is 577 amino acids, which is substantially longer than the average length of 451 amino acids across all proteins in the database. As shown in Figure 3, the Barista-only proteins are much shorter (average 403 amino acids) than ProteinProphet-only proteins (1073 amino acids). Similar trends are shown in Supplementary Figure 5 and Supplementary Table 3 across all four data sets and varying numbers of false positives.

Some proteins are identified only by Barista because it successfully exploits multiple weak matches. For example, protein YIL066C is correctly identified by Barista on the basis of 22 distinct peptides, none of which receives a PeptideProphet probability > 0.5. Indeed, many Barista-only proteins receive additional support in the form of multiple peptides with PeptideProphet probabilities < 0.05. For YIL066C, for example, only 7 of the 22 peptides receive probabilities > 0.05. The remaining 15 peptides are ignored by ProteinProphet. However, simply lowering the PeptideProphet threshold, in order to include these additional peptides, significantly hurts ProteinProphet's overall performance (Supplementary Figure 6).

Barista identifies a larger proportion of "one-hit wonders"—proteins whose only evidence consists of a single strong peptide identification. Using ProteinProphet's threshold of 0.05 to define such proteins, we find

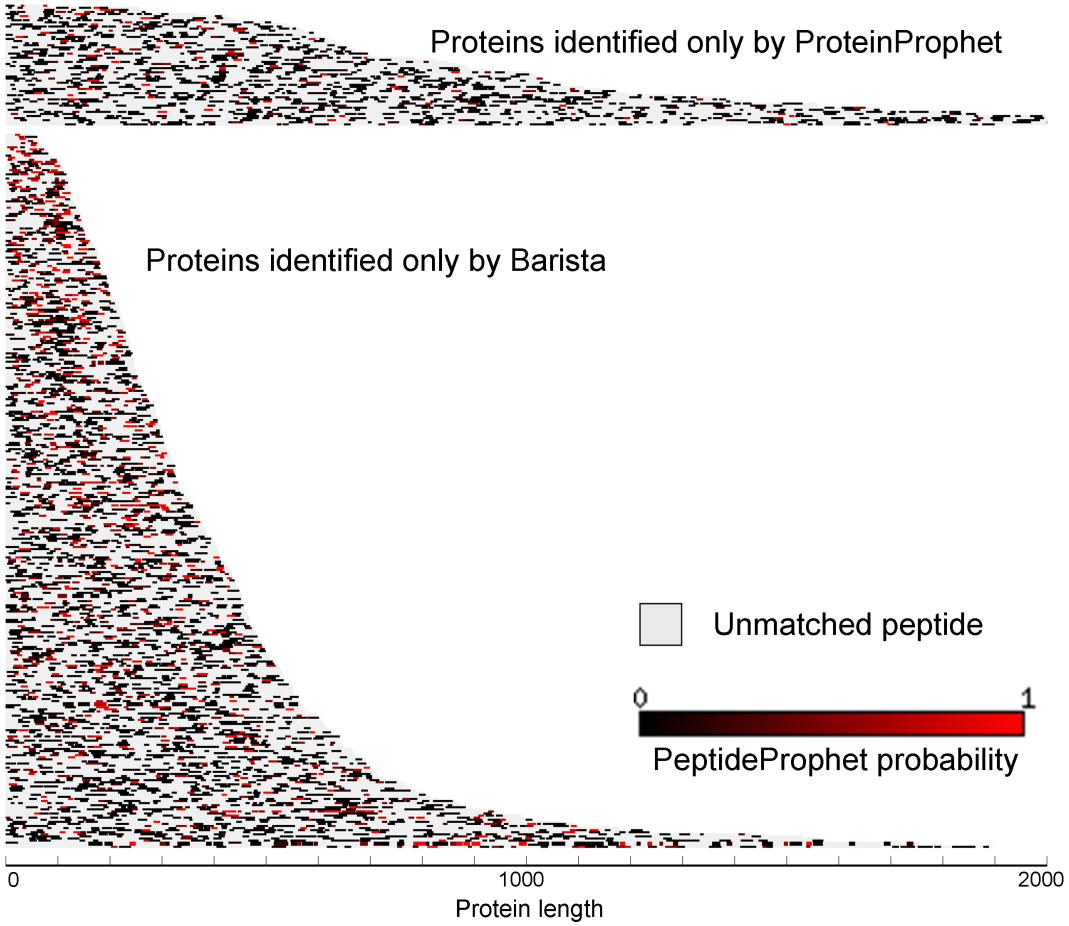

Figure 3: **Proteins identified only by ProteinProphet and only by Barista.** Each protein is represented as a horizontal bar. The color scale from black to red indicates the PeptideProphet probability assigned to each peptide, with unmatched peptides indicated in light gray. Five very long proteins identified by ProteinProphet (lengths 2039, 2123, 2143, 3093 and 3268 amino acids) were truncated for the figure.

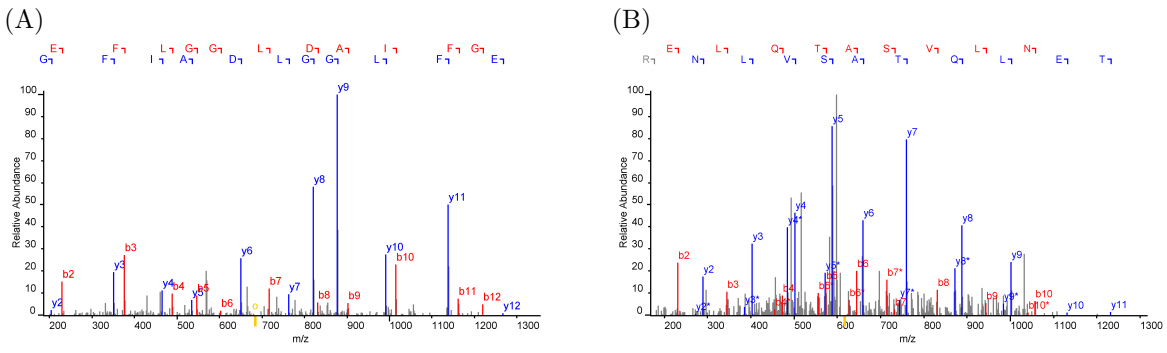

Figure 4: **Two proteins identified via single PSMs.** (A) The annotated spectrum that matched the +2 charged peptide VEFLGGLDAIFGK, one of only two observable tryptic peptides in the 100 amino acid protein URM1. Colored peaks represent b-ions (red) and y-ions (blue), with the precursor m/z indicated in yellow. (B) The annotated spectrum that matched the +2 charged peptide TELQTASVLNR in protein ATP15. This 63 amino acid protein contains a total of four observable tryptic peptides.

that of the 58 ProteinProphet-only proteins, 7% contain only a single matched peptide; the corresponding percentage for Barista is 18%. The difference between these two sets of proteins is two-fold. First, Barista one-hit wonders are significantly shorter: 238 amino acids on average, compared with an average length of 602 amino acids for the ProteinProphet one-hit wonders. Second, the Barista one-hit wonders are supported by, on average, 3.2 additional peptides that receive PeptideProphet probabilities < 0.05. Indeed, among the set of 331 Barista-only proteins, only nine are true one-hit wonders, and these have an average length of 97 amino acids. For example, Figure 4(A) shows the matched spectrum for the protein URM1. This protein contains 100 amino acids but only two tryptic peptides longer than six amino acids. The spectrum includes nearly complete b-ion and y-ion series, and reassuringly, the two highest peaks are y-ions associated with glysine, which has previously been demonstrated to produce large peaks [Tabb et al., 2004, Klammer et al., 2008]. Similarly, Figure 4(B) shows the annotated spectrum for the single observed peptide in ATP15, a length-63 protein containing only four observable tryptic peptides. Again, nearly all of the b- and y-ion are observed. Thus, by taking into account weak matches and by normalizing with respect to the total number of peptides in the protein, Barista successfully discards long, absent proteins and retains short, present proteins. This approach agrees with recent evidence that requiring at least two peptides per protein unnecessarily eliminates many true identifications [Gupta and Pevzner, 2009].

We also used the abundance levels assigned to the proteins identified by antibody and GFP tagging experiments [Ghaemmaghami et al., 2003] to investigate the extent to which Barista scores correlate with protein abundance. Supplementary Figure 7 shows that when target proteins are ranked by Barista score, the top of the list is enriched with high-abundance proteins. This property, combined with Barista's ability to identify more target proteins at a fixed number of false positives, implies that Barista will successfully identify more low-abundance proteins than ProteinProphet. For example, on the trypsin-digested yeast dataset, for 10 false positives, ProteinProphet identifies 1087 proteins with average intensity of 31694. For the same data set, Barista identifies 1356 proteins at 10 false positives. When these proteins are ordered by the Barista-produced scores, the first 1087 of them (i.e., the same number as identified by ProteinProphet) have an average abundance of 34827, whereas the remaining 269 are low-abundance proteins with an average abundance of 5097.

As mentioned above, Barista's protein length normalization relies upon the total number of theoretical peptides that occur in the protein. This approach contrasts with ProteinProphet, which does not include a penalty for unmatched peptides. To test this protein score normalization, we investigated a variant of Barista, in which we modify the denominator of Barista's protein scoring to use the number of matched peptides, rather than the total number of database peptides. The results of this experiment, shown in Supplementary Figure 8, indicate that the modified form of Barista identifies fewer proteins than the original version. Furthermore, as shown in Supplementary Table 3, the average lengths of identified proteins is longer when Barista normalizes by the number of database peptides rather than the number of matched peptides. These results suggest that it may be possible to correct some of ProteinProphet's length bias by modifying its normalization.

Thus far, Barista focuses on optimizing a single value—the number of proteins identified from a shotgun proteomics experiment. This approach contrasts with previous applications of machine learning to this task [Anderson et al., 2003, Keller et al., 2002, Elias et al., 2004, Käll et al., 2007, Spivak et al., 2009], which optimize at the level of PSMs or peptides. In general, selecting one optimization target or the other will depend on the goal of the proteomics experiment. However, in some applications, it may be desirable to simultaneously achieve high levels of peptide and protein identification. In such applications, we can perform joint optimization on both the protein and peptide levels. We use multi-task learning [Caruana, 1997], training the protein and peptide ranking tasks in parallel using a shared neural network representation (see supplement for details). Supplementary Figure 9 compares the performance of three variants of Barista: optimizing protein identifications, optimizing peptide identifications and jointly optimizing both. The methods are evaluated both at the protein and peptide level. Comparing the multi-task and protein-only versions of Barista, we see that, as expected, the multi-tasking gives a slight improvement at the peptide level and slightly worse performance at the protein level. This is a natural consequence of jointly optimizing on two tasks at once. On the other hand, comparing the multi-task and peptide-only versions of Barista, we see that performing multi-tasking actually gives improved performance on both the protein and peptide levels. The unexpected improvement in peptide-level rankings occurs because the protein ranking task introduces higher-level information about the scores of all peptides belonging to the same protein. In general, choosing

whether to optimize for peptide identification, protein identification or both will depend upon the goal of the proteomics experiment.

Many algorithms designed for inferring a set of proteins from a collection of PSMs divide the problem into two stages: assessing the quality of the PSMs and then inferring the protein set [Nesvizhskii et al., 2003, Alves et al., 2007, Zhang et al., 2007, Li et al., 2008]. Subdividing the protein identification problem in this fashion results in a significant loss of information during the second stage of the analysis. For example, typically only a subset of spectra are assigned to a peptide during the peptide identification stage, so information about the unassigned spectra is not available to the protein identification algorithm. Also, if at most one peptide is assigned to each spectrum, and if for a particular spectrum that assignment happens to be incorrect, then information about the second-ranked, possibly correct peptide is not available during the protein identification stage. Finally, if the quality of the match between a peptide and a spectrum is summarized using a single score, such as the probability assigned by PeptideProphet, then detailed information about precisely how the peptide matches the spectrum is lost. In contrast, the machine learning approach described here directly optimizes the number of identified proteins, taking into account all available information to obtain the best possible result.

References

M. Adamski, T. Blackwell, R. Menon, L. Martens, H. Hermjakob, C. Taylor, G. S. Omenn, and D. J. States. Data management and preliminary data analysis in the pilot phase of the HUPO Plasma Proteome Project. *Proteomics*, 5(13):3246–3261, 2005.

P. Alves, R. J. Arnold, M. V. Novotny, P. Radivojac, J. P. Reilly, and H. Tang. Advancement in protein inference from shotgun proteomics using peptide detectability. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 409–420, Singapore, 2007. World Scientific.

D. C. Anderson, W. Li, D. G. Payan, and W. S. Noble. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *Journal of Proteome Research*, 2(2):137–146, 2003.

R. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

H. Choi and A. I. Nesvizhskii. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *Journal of Proteome Research*, 7(1):47–50, 2007.

H. Choi and A. I. Nesvizhskii. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *Journal of Proteome Research*, 7(1):254–265, 2008.

H. Choi, D. Ghosh, and A. Nesvizhskii. Statistical validation of peptide identifications in large-scale proteomics using target-decoy database search strategy and flexible mixture modeling. *Journal of Proteome Research*, 7(1):286–292, 2008.

J. E. Elias and S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4(3):207–214, 2007.

J. E. Elias, F. D. Gibbons, O. D. King, F. P. Roth, and S. P. Gygi. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nature Biotechnology*, 22:214–219, 2004.

D Fenyo and R C Beavis. A method for assessing the statistical significance of mass spectrometry-based protein identification using general scoring schemes. *Analytical Chemistry*, 75:768–774, 2003.

M. Fitzgibbon, Q. Li, and M. McIntosh. Modes of inference for evaluating the confidence of peptide identifications. Submitted, 2007.

L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant. Open mass spectrometry search algorithm. *Journal of Proteome Research*, 3:958–964, 2004.

S. Ghaemmaghami, W. K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, and J. S. Weissman. Global analysis of protein exression in yeast. *Nature*, 425:737–741, 2003.

N. Gupta and P. Pevzner. False discovery rates of protein identifications: a strike against the two-peptide rule. *Journal of Proteome Research*, 2009. Epub ahead of print.

F. C. P. Holstege, E. G. Gennings, J. J. Wyrick, T. I. Lee, C. J. Hengartner, M. R. Green, T. R. Golub, E. S. Lander, and R. A. Young. Dissecting the regulatory circuitry of eukaryotic genome. *Cell*, 95:717–728, 1998.

E. L. Huttlin, A. D. Hegeman, A. C. Harms, and M. R. Sussman. Prediction of error associated with false-positive rate determination for peptide identification in large-scale proteomics experiments using a combined reverse and forward peptide sequence database strategy. *Journal of Proteome Research*, 2006.

L. Käll, J. Canterbury, J. Weston, W. S. Noble, and M. J. MacCoss. A semi-supervised machine learning technique for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4:923–25, 2007.

L. Käll, J. D. Storey, M. J. MacCoss, and W. S. Noble. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of Proteome Research*, 7(1):29–34, 2008.

A. Keller, A. I. Nesvizhskii, E. Kolker, and R. Aebersold. Empirical statistical model to estimate the accuracy of peptide identification made by MS/MS and database search. *Analytical Chemistry*, 74:5383–5392, 2002.

A. A. Klammer, S. R. Reynolds, M. Hoopmann, M. J. MacCoss, J. Bilmes, and W. S. Noble. Modeling peptide fragmentation with dynamic Bayesian networks yields improved tandem mass spectrum identification. *Bioinformatics*, 24(13):i348–i356, 2008.

Y. F. Li, R. J. Arnold, Y. Li, P. Radivojac, Q. Sheng, and H. Tang. A Bayesian approach to protein inference problem in shotgun proteomics. In M. Vingron and L. Wong, editors, *Proceedings of the Twelfth Annual International Conference on Computational Molecular Biology*, volume 12 of *Lecture Notes in Bioinformatics*, pages 167–180, Berlin, Germany, 2008. Springer.

Z.-Q. Ma, S. Dasari, M. C. Chambers, M. Litton, S. M. Sobecki, L. Zimmerman, P. J. Halvey, B. Schilling, P. M. Drake, B. W. Gibson, and D. L. Tabb. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *Journal of Proteome Research*, 2009.

R. E. Moore, M. K. Young, and T. D. Lee. Qscore: An algorithm for evaluating SEQUEST database search results. *Journal of the American Society for Mass Spectrometry*, 13(4):378–386, 2002.

A. I. Nesvizhskii, A. Keller, E. Kolker, and R. Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry*, 75:4646–4658, 2003.

C. Y. Park, A. A. Klammer, L. Käll, M. P. MacCoss, and W. S. Noble. Rapid and accurate peptide identification from tandem mass spectra. *Journal of Proteome Research*, 7(7):3022–3027, 2008.

J. Peng, J.E. Elias, C.C. Thoreen, L.J. Licklider, and S.P. Gygi. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS-MS) for large-scale protein analysis: the yeast proteome. *Journal of Proteome Research*, 2:43–50, 2003.

D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20:3551–3567, 1999.

M. Spivak, J. Weston, L. Bottou, L. Käll, and W. S. Noble. Improvements to the percolator algorithm for peptide identification from shotgun proteomics data sets. *Journal of Proteome Research*, 8(7):3737–3745, 2009.

D. L. Tabb, Y. Huang, V. H. Wysocki, and J. R. Yates, III. Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Analytical Chemistry*, 76:1243–48, 2004.

B. Zhang, M. C. Chambers, and D. L. Tabb. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *Journal of Proteome Research*, 6(9):3549–3557, 2007.