© 2007 Nature Publishing Group http://www.nature.com/naturemethods

Semi-supervised learning for peptide identification from shotgun proteomics datasets

Lukas Käll[1], Jesse D Canterbury[1], Jason Weston[2], William Stafford Noble[1,3] & Michael J MacCoss[1]

Shotgun proteomics uses liquid chromatography–tandem mass spectrometry to identify proteins in complex biological samples. We describe an algorithm, called Percolator, for improving the rate of confident peptide identifications from a collection of tandem mass spectra. Percolator uses semi-supervised machine learning to discriminate between correct and decoy spectrum identifications, correctly assigning peptides to 17% more spectra from a tryptic *Saccharomyces cerevisiae* dataset, and up to 77% more spectra from non-tryptic digests, relative to a fully supervised approach.

Mass spectrometry has become the most widely used tool for the characterization of proteins within complex mixtures. Integral to the broad acceptance of mass spectrometry for protein characterization has been the development of database searching software such as SEQUEST[1] and MASCOT[2]. These algorithms assign a peptide sequence to each tandem mass spectrometry (MS/MS) spectrum by comparing experimentally acquired spectra against theoretically predicted spectra of peptides derived from a sequence database. Scores reflecting the similarity between the measured and predicted spectra are used to discriminate between correct and incorrect peptide sequence assignments.

Although database searching algorithms work well, there remains substantial room for improvement. In particular, current scoring methods produce significant overlap between the scores of correct and incorrect peptide identifications[3,4]. Thus, to ensure that a large fraction of the true positive identifications are retained, a score threshold must be selected such that a percentage of the peptide identifications are incorrect.

To estimate the number of false positive protein identifications in a more systematic fashion, an approach using a decoy database containing reversed protein sequences was developed[5]. Since this initial application, many other researchers have used decoy searches to estimate the number of incorrect peptide-spectrum matches (PSMs) that exceed a given threshold[6]. This approach allows the user to adjust the score threshold to obtain a target false discovery rate.

Because most database search algorithms return multiple scores (for example, XCorr, Sp, and ΔC_n for SEQUEST), most proteomics studies apply separate thresholds to each score. Using multiple orthogonal score criteria is useful for eliminating false discoveries that might exceed one threshold but not another. However, in most cases these orthogonal scores are considered independently, ignoring the benefits that can be obtained if the features are considered jointly.

An alternative approach is to use machine learning methods to re-rank the PSMs and then set a threshold automatically in the re-ranked list[4,7]. This approach uses a supervised classification algorithm to discriminate between correct and incorrect PSMs. Each PSM is characterized by a fixed-length vector of features, and the relative weights of the individual features are learned from a training set of manually curated PSMs. This approach provides substantially greater confidence in peptide identification than using SEQUEST alone; however, obtaining a high-quality training set is complicated. Furthermore, the characteristics of correct and incorrect PSMs vary from laboratory to laboratory and from one experiment to the next. Sample type (for example, soluble versus membrane proteins), enzyme specificity, modified versus unmodified peptides, mass spectrometer type, database size, instrument calibration and other parameters alter the optimal weighting of features. Unfortunately, it is impractical to generate a high-quality training set for every anticipated micro liquid chromatography–MS/MS analysis.

In this study, we describe a solution to these problems that uses a software post-processor that can be appended to any existing database search algorithm. The algorithm, called Percolator, uses a semi-supervised learning method that eliminates the need to construct a manually curated training set. The PSMs derived from searching a decoy database consisting of shuffled protein sequences are used as negative examples for the classifier, and a subset of the high-scoring PSMs derived from searching the target database are used as positive examples. Percolator trains a machine learning algorithm called a support vector machine (SVM)[8] to discriminate between positive and negative PSMs. One benefit of the semi-supervised learning paradigm is that the classifier is free to exploit a variety of specific features of the data, without overfitting to a particular type of spectrum. Percolator represents each PSM using a rich vector of 20 features. These features, as well as a detailed description of the algorithm, are provided in **Supplementary Table 1** and **Supplementary Methods** online.

Percolator is fully automated and significantly improves the sensitivity of existing database search algorithms at a constant false discovery rate. Furthermore, Percolator assigns a statistically

[1]Department of Genome Sciences, University of Washington, 1705 NE Pacific St., William H. Foege Building, Seattle, Washington 98195, USA. [2]NEC Laboratories America, Inc., 4 Independence Way, Suite 200, Princeton, New Jersey 08540, USA. [3]Department of Computer Science and Engineering, University of Washington, AC101 Paul G. Allen Center, 185 Stevens Way, Seattle, Washington 98195, USA. Correspondence should be addressed to M.J.M. (maccoss@u.washington.edu).

RECEIVED 3 MAY; ACCEPTED 1 OCTOBER; PUBLISHED ONLINE 21 OCTOBER 2007; DOI:10.1038/NMETH1113

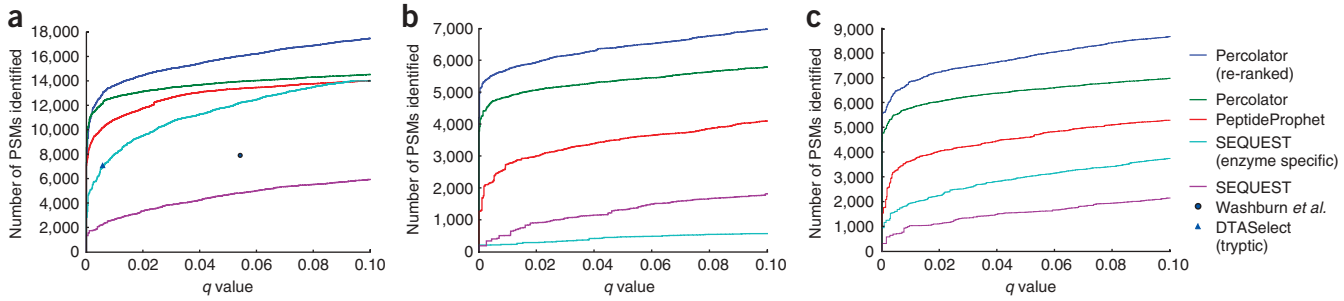

Figure 1 | Comparison of SEQUEST post-processing methods. The figure plots the number of identified PSMs as a function of the *q* value. (**a**–**c**) The data consist of 69,705 PSMs from yeast proteins digested with trypsin (**a**), 57,860 PSMs from yeast proteins digested with elastase (**b**) and 60,217 PSMs from yeast digested with chymotrypsin (**c**). In **a** the blue circle and blue triangle correspond to applying, respectively, the heuristic described previously[11] and the default DTASelect thresholds[10] to the SEQUEST output.

meaningful *q* value to each spectrum, which is defined as the minimal false discovery rate at which the identification is deemed correct[9]. These *q* values are estimated using the distribution of scores from the decoy database search.

We measured Percolator's ability to identify correct PSMs using a yeast dataset containing 35,236 spectra. These data were acquired from a tryptic digest of an unfractionated *S. cerevisiae* lysate and analyzed using a 4-h reverse-phase separation. We assigned peptides to spectra by using SEQUEST with no enzyme specificity, allowing multiple charge states for some spectra (see **Supplementary Methods**); this yielded 69,705 target PSMs. The SEQUEST analysis required ~3 d on an Athlon MP Opteron 842 CPU. The subsequent Percolator analysis required ~4 min on the same CPU. The results (**Fig. 1a**) show that Percolator markedly improved upon the initial SEQUEST scoring function. Unless otherwise specified, we consider a PSM to be correct if it achieves a *q* value <0.01. In this case, Percolator correctly identified 12,691 PSMs, corresponding to 8,197 unique peptides and 1,630 proteins. The number of proteins was computed using DTASelect[10], requiring one peptide per locus and removing 'subset' proteins. At the same *q* value, SEQUEST identified only 2,780 PSMs using XCorr alone.

We tried a number of methods for improving the performance of SEQUEST on this dataset. First, we filtered the SEQUEST identifications to allow only tryptic peptides. The results (cyan curve in **Fig. 1a**) were much better than the initial SEQUEST ranking (8,602 PSMs) but still not as good as those of Percolator. We also applied several published post-processing methods to the SEQUEST results. One heuristic[11] identified 7,926 PSMs at a *q* value of 0.054. At this same *q* value, Percolator identified 13,982 PSMs. Using default thresholds, DTASelect[10] identified 7,583 PSMs at a *q* value of 0.18, and 7,094 PSMs at a *q* value of 0.0057 when using only fully tryptic peptides. Finally, PeptideProphet[4], which uses linear discriminant analysis with a fixed set of coefficients, produced results that were better than SEQUEST's; however, even when restricted to tryptic peptides, PeptideProphet consistently identified fewer PSMs than Percolator. For example, at a *q* value of 0.01, Percolator identified 17% more PSMs (12,691 versus 10,863) and 15% more unique peptides (8,197 versus 7,120) than did PeptideProphet. Thus, none of the available algorithms that we tested identified as many peptides as Percolator.

Next, we tested a variant of the Percolator algorithm that re-ranks potential false negative identifications (**Supplementary Methods**). Rather than considering only the top-ranked SEQUEST PSM for each spectrum, Percolator scores the top five PSMs. The results (blue curve in **Fig. 1a**) were better than when Percolator was applied to only the top-ranked PSM. At a *q* value of 0.01, the re-ranking procedure produced 8% more PSMs relative to Percolator without re-ranking. We also investigated Percolator's behavior on a larger dataset, a 24-h MudPIT analysis of *Caenorhabditis elegans* proteins containing 207,804 spectra (**Supplementary Data**).

The primary advantage of re-training the classifier for each individual dataset, as we do with Percolator, is that we do not have to make the classifier general enough to handle all types of possible spectra. We thus expect Percolator's strength relative to PeptideProphet to be most apparent when the dataset being analyzed differs substantially from the dataset that was used to train PeptideProphet. We therefore ran Percolator on yeast sets digested with elastase (**Fig. 1b**) and chymotrypsin (**Fig. 1c**). As expected, Percolator's performance on these datasets improved relative to that of PeptideProphet, even though we ran PeptideProphet in its "elastase" and "chymotrypsin" modes, respectively. At a *q* value of 0.01, Percolator yielded 77% more PSMs and 48% more unique peptides in the elastase set and 58% more PSMs and 34% more unique peptides in the chymotrypsin set than did PeptideProphet. We performed a variety of control experiments to verify Percolator's performance and assess its robustness (**Supplementary Data**, **Supplementary Figs. 1**, **2** and **3** and **Supplementary Table 2** online).

SEQUEST is quite robust in assigning the correct peptide sequence from candidate sequences in a database. Nevertheless, there are occasions when the best peptide sequence is not ranked first using XCorr. In selected cases, Percolator can correct an incorrectly ranked peptide sequence obtained using XCorr alone. **Figure 2** illustrates an example in which Percolator correctly re-ranked the resulting PSMs. SEQUEST returned the sequence INSLNDSNLSIPE with an XCorr of 2.99 (**Fig. 2a**; the predicted *b*- and *y*-ion fragment ions are highlighted in red in the MS/MS spectrum). Several intense fragment ions were not accounted for by this peptide, indicating that this might not be a correct match. In contrast, Percolator assigned a *q* value of 0.95 for this peptide, but it assigned a better score (*q* value = 0) to SEQUEST's second-ranked peptide (LGANAILGVSMAAAR, XCorr = 2.87; **Fig. 2b**). In comparison to the sequence ranked first by SEQUEST, the predicted *b*- and *y*-ions from Percolator's top-ranked peptide accounted for a much larger fraction of fragment ions. We confirmed Percolator's top-ranked peptide assignment by acquiring

© 2007 Nature Publishing Group http://www.nature.com/naturemethods

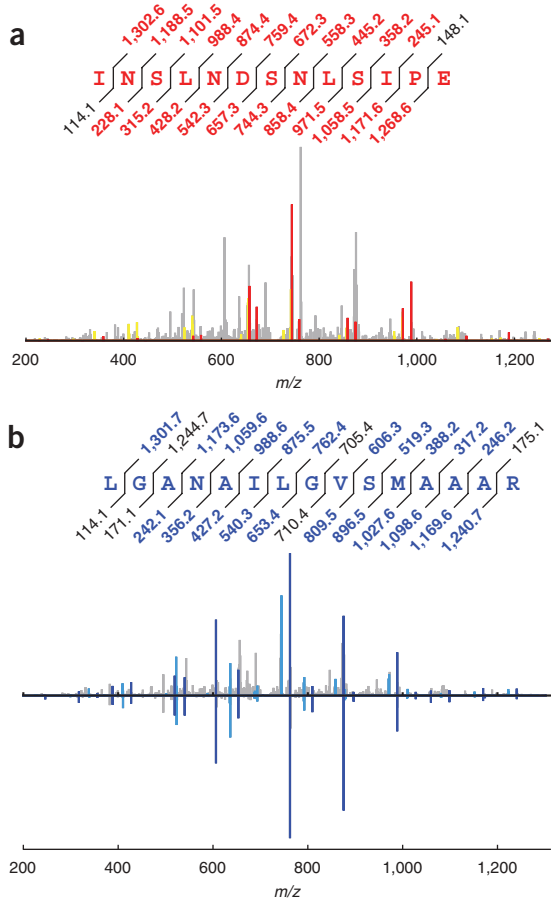

Figure 2 | A peptide that was re-ranked by Percolator. (**a**) The peptide that was returned with the highest XCorr by SEQUEST. The peptide is a non-tryptic peptide for which only a fraction of the total ion current is accounted for by the predicted *b*- and *y*-ions (labeled in red) and associated neutral losses (labeled in yellow). (**b**) The peptide was chosen by Percolator as the best-matching sequence for this spectrum. The predicted *b*- and *y*-ions from the peptide (labeled in blue) account for a significantly greater amount of the total signal in the spectrum. The peaks in the mirror image (**b**, below) are from a spectrum acquired on the purified peptide produced synthetically. Although there are a few chemical noise peaks unaccounted for in the spectrum acquired from the endogenous peptide, this is not unusual for spectra acquired in the context of an unfractionated mixture. The relative abundances of the peaks obtained for the endogenous and synthetic peptides are very similar and are indicative of a high-quality match.

an MS/MS spectrum of a synthetic analog of the peptide. Percolator can also be used for the interpretation of spectra containing multiple peptides (see **Supplementary Fig. 4** and **Supplementary Methods** online for further discussion).

We have described a statistically rigorous, computationally efficient machine learning method for increasing the number of confident peptide identifications from tandem mass spectra. The Percolator algorithm can be applied as a post-processor to a collection of target and decoy PSMs produced by any database search algorithm. We show that this approach improves the rate of peptide identification relative both to a static search procedure and to a fully supervised post-processing method. The improvement is greatest for non-tryptic digests, for which existing methods are not optimized. The test data used in this report are available in **Supplementary Data 2** online. The test data used in this report are available in **Supplementary Data 2** online. The software, including source code, is freely available for nonprofit use. The data used in the manuscript and information about obtaining the software are available online (http://noble.gs.washington.edu/proj/percolator).

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

This work was funded by US National Institutes of Health grants P41 RR011823 and R01 EB007057.

AUTHOR CONTRIBUTIONS

M.J.M. came up with the initial idea to use decoy PSMs as negative examples. L.K. and W.S.N. came up with the idea to use a support vector machine using semi-supervised learning. L.K. implemented Percolator and performed computational experiments. J.W. provided machine learning expertise. J.D.C. performed initial proof-of-concept experiment and provided mass spectrometry expertise. W.S.N., L.K. and M.J.M. wrote the article.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Eng, J.K., McCormack, A.L. & Yates, J.R. III. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
2. Perkins, D.N., Pappin, D.J.C., Creasy, D.M. & Cottrell, J.S. *Electrophoresis* **20**, 3551–3567 (1999).
3. MacCoss, M.J., Wu, C.C. & Yates, J.R. III. *Anal. Chem.* **74**, 5593–5599 (2002).
4. Keller, A., Nezvizhskii, A.I., Kolker, E. & Aebersold, R. *Anal. Chem.* **74**, 5383–5392 (2002).
5. Moore, R.E., Young, M.K. & Lee, T.D. *J. Am. Soc. Mass Spectrom.* **13**, 378–386 (2002).
6. Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J. & Gygi, S.P. *J. Proteome Res.* **2**, 43–50 (2003).
7. Anderson, D.C., Li, W., Payan, D.G. & Noble, W.S. *J. Proteome Res.* **2**, 137–146 (2003).
8. Boser, B.E., Guyon, I.M. & Vapnik, V.N. A training algorithm for optimal margin classifiers. in *5th Annual ACM Workshop on COLT* (ed. Haussler, D.) 144–152 (ACM Press, Pittsburgh, Pennsylvania, USA, 1992).
9. Storey, J.D. & Tibshirani, R. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
10. Tabb, D.L., McDonald, W.H. & Yates, J.R. III. *J. Proteome Res.* **1**, 21–26 (2002).
11. Washburn, M.P., Wolters, D. & Yates, J.R. III. *Nat. Biotechnol.* **19**, 242–247 (2001).

© 2007 Nature Publishing Group http://www.nature.com/naturemethods