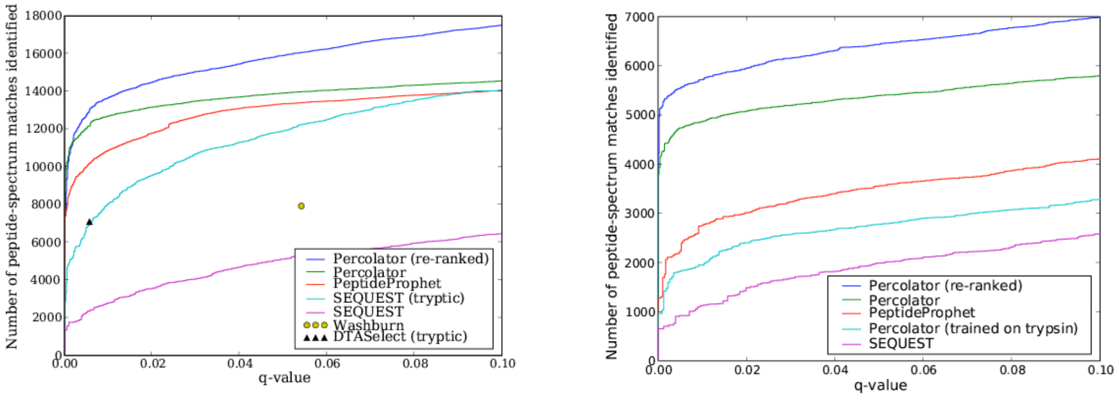
A semi-supervised machine learning technique for peptide identification from shotgun proteomics datasets

Lukas Käll[1], Jesse Canterbury[1], Jason Weston[2], Michael J. MacCoss[1] and William Stafford Noble[1,3]
[1]Department of Genome Sciences, University of Washington, Seattle, WA, USA [2]NEC Laboratories America Princeton, NJ, USA
[3]Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA

An essential component of proteomics is the efficient translation of mass spectrometry data into peptide identifications. We have developed a semi-supervised machine learning algorithm, Percolator, that dynamically learns to combine multiple scores from mapping a fragmentation spectrum to a peptide in a sequence database. Since it is difficult *a priori* to determine which spectra are well matched to a target sequence database, the method instead makes use of the fact that matches of the observed spectra against a shuffled sequence database can be trusted as examples of bad identifications. The SVM-based method dramatically increases the number of true spectrum identifications for a given false discovery rate, relative to the current state of the art. Percolator's power derives primarily from its ability to learn the unique features of each data set, a property that current methods are unable to do. Alternative methods such as PeptideProphet [*Anal. Chem.* 74:5383] are pre-trained using a collection of "true" and "false" peptide identifications.

We tested Percolator's and PeptideProphet's performance on a collected set of spectra from yeast digested with trypsin (Figure 1A). The results show that Percolator clearly outperforms PeptideProphet. However, Percolator's real strength is apparent when the data set differs substantially from the data set used to train PeptideProphet. In a second experiment, we ran Percolator and PeptideProphet on a data set with very different characteristics from the dataset that PeptideProphet was trained on: a yeast set digested with elastase instead of trypsin (Figure 1B). As expected, PeptideProphet's performance degrades on this data set, relative to Percolator, even though we ran PeptideProphet in its "elastase" mode: Percolator finds 75% more peptide-specrum matches than PeptideProphet at a false discovery rate of 1%. Finally, to show that the semi-supervised learning is the source of Percolator's good performance, we trained Percolator on the tryptic data set and tested on the elastase dataset. As expected, Percolator does not perform very well in this case (Figure 1B; light blue curve).

In summary, Figures 1A and 1B show that Percolator performs significantly better than PeptideProphet for two reasons: because of its semi-supervised nature, Percolator is free to use a larger set of features, and Percolator learns the relative weights of these features directly from the data being analyzed.



(A) (B)

Figure 1: Performance of Percolator and and PeptideProphet. Each panel plots the number of peptide-spectrum matches versus a metric closely related to the false discovery rate, the q-value [*PNAS* 100:9440]. Performance was measured on yeast sets digested with (A) trypsin (B) elastase. PeptideProphet (red curve) is not able to generalize to the new kind of data even though the method is run in elastase mode. The same behavior is observed when training Percolator on tryptic data but testing on the elastase data (B; light blue curve).