Supplemental Material to "Improving Sensitivity by Probabilistically Combining Results from Multiple MS/MS Search Methodologies"

Brian C. Searle[1*], Mark Turner[1], Alexey I. Nesvizhskii[2,3]

[1]Proteome Software Inc., 1340 SW Bertha Blvd. Suite 201, Portland, OR, 97219-2039, USA
[2]Department of Pathology and [3]Center for Computational Medicine and Biology, University of Michigan, 1301 Catherine Road, Ann Arbor, Michigan 48109-0602, USA
[*]Primary contact. E-mail: Brian.Searle@ProteomeSoftware.com, Telephone: (503) 244-6027, Fax: (503) 245-4910

Experimental Data Availability

All experimental data presented in this manuscript has been made available in the form of Scaffold .SFD files. These files contain all the necessary publication standards information for proteomics data and can be examined by anyone using the free Scaffold viewer software (www.proteomesoftware.com). The original MS/MS peak lists can be exported through the viewer and all of the search engine identification results used in this manuscript can be analyzed. These files have been made available at:
http://www.proteomesoftware.com/manuscripts/scaffold_result_files.zip

Alterative Ways of Computing Search Engine Agreement

Although we present only one method for computing the agreement between database search engines (Equation 6), one could consider several different alternative agreement formulas. In the manuscript we used a discretized approach and define the Agreement Score for peptide assignment *j* by search engine *k* to spectrum *i* to be:

$$A_{i,\,j,\,k} = \sum_{k' \neq k} \begin{Bmatrix} p(+ \mid D_{i,\,j,\,k'}) < 0.05 & 0.0 \\ 0.05 \le p(+ \mid D_{i,\,j,\,k'}) < 0.5 & 0.5 \\ 0.5 \le p(+ \mid D_{i,\,j,\,k'}) & 1.0 \end{Bmatrix} \qquad \text{(Discrete 50)}$$

where $A_{i,j,k}$ is calculated for each spectrum assignment across all database-searching algorithms, except for *k*. Discrete 50 breaks agreement into "high agreement" (1.0), "low agreement" (0.5), and "no agreement" (0.0). An alternative discrete method was also tested:

$$A_{i,\,j,\,k} = \sum_{k' \neq k} \begin{Bmatrix} p(+ \mid Di,j,k') < 0.05 & 0.0 \\ 0.05 \le p(+ \mid Di,j,k') & 1.0 \end{Bmatrix} \qquad \text{(Discrete 100)}$$

Here, $A_{i,j,k}$ does not include an intermediate term for "low agreement".

Two similar techniques were tested (Discrete 50 w/ Neg and Discrete 100 w/ Neg) that penalize peptides for "no agreement". These formulas are:

$$A_{i,\,j,\,k} = \sum_{k' \neq k} \begin{Bmatrix} p(+ \mid D_{i,\,j,\,k'}) < 0.05 & -0.5 \\ 0.05 \le p(+ \mid D_{i,\,j,\,k'}) < 0.5 & 0.5 \\ 0.5 \le p(+ \mid D_{i,\,j,\,k'}) & 1.0 \end{Bmatrix} \qquad \text{(Discrete 50 w/ Neg)}$$

and

$$A_{i,\,j,\,k} = \sum_{k' \neq k} \begin{Bmatrix} p(+ \mid Di,j,k') < 0.05 & -0.5 \\ 0.05 \le p(+ \mid Di,j,k') & 1.0 \end{Bmatrix} \qquad \text{(Discrete 100 w/ Neg)}$$

A fifth method is to compute a non discrete sum of the peptide identification probabilities that other programs have made to the same peptide. This is analogous to the NSP calculation in ProteinProphet (1) where the Agreement Score is defined to be:

$$A_{i,j,k} = \sum_{k' \neq k} \{ p(+ \mid Di, j, k') \} \qquad \text{(Non-Discrete)}$$

Finally, using a penalty for "no agreement" was also tried:

$$A_{i,j,k} = \sum_{k' \neq k} \begin{Bmatrix} p(+|Di,j,k') \\ -p(+|Di,j,k') \end{Bmatrix} \begin{matrix} agreement \\ disagreement \end{matrix} \qquad \text{(Non-Discrete w/ Neg)}$$

Of these six methods, the three that performed best on the 18 protein control mixture were Discrete 50, Discrete 50 w/ Neg, and Non-Discrete (Figure S1).

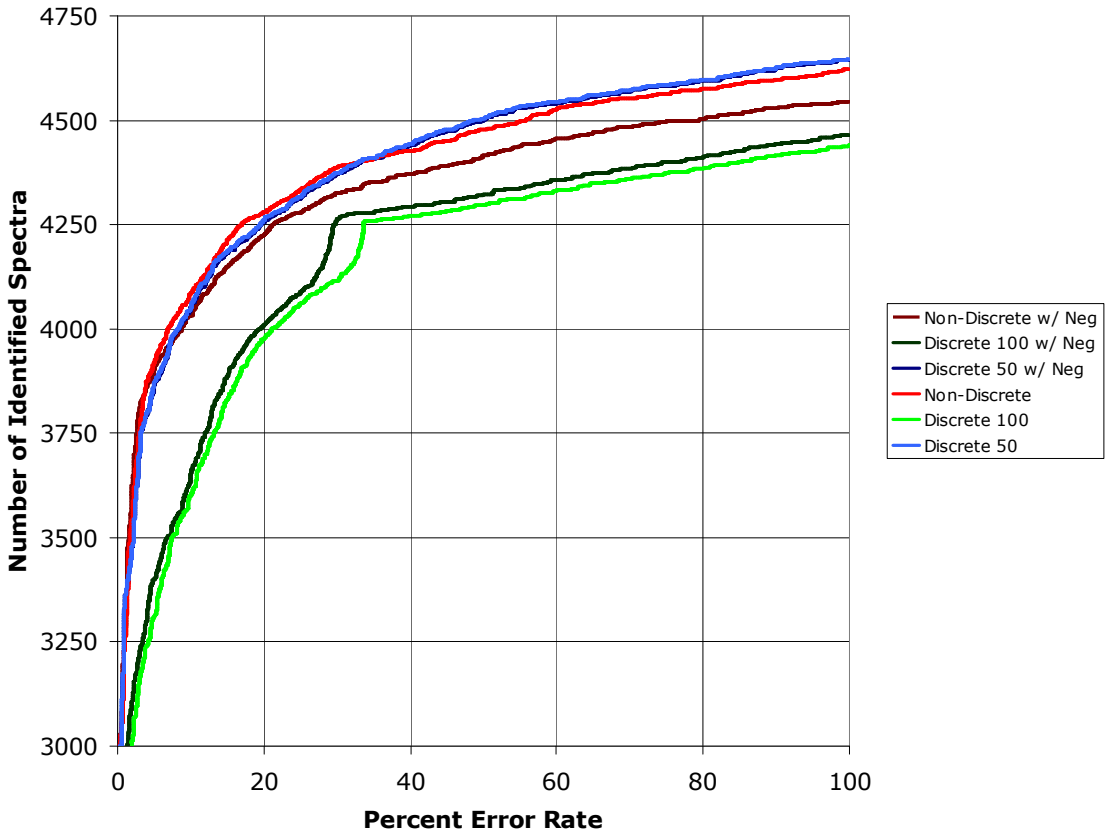

Figure S1 Comparison between the number of correctly identified spectra and the cumulative error rate in both the 18 protein known mixture. Higher curves indicate better sensitivity given a specified error rate.

(1) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R.; *Anal. Chem*. 2003, 75, 4646-4658.