SUPPORTING INFORMATION

Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications made by MS/MS and Database Search

Andrew Keller[†*], Alexey I. Nesvizhskii[†*], Eugene Kolker, and Ruedi Aebersold

Institute for Systems Biology
1441 North 34th Street
Seattle, WA 98103 USA

[†]contributed equally to this work

[*]corresponding authors: akeller@systemsbiology.org
nesvi@systemsbiology.org
Fax: 206-732-1299

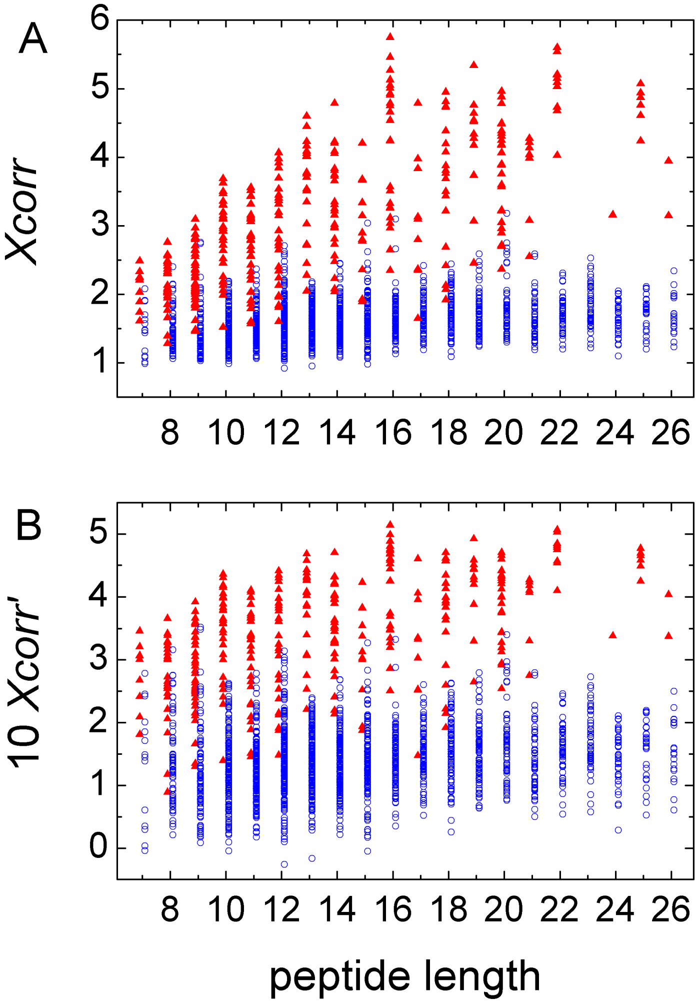

Figure S1. Transformation of *Xcorr* to *Xcorr′* reduces its dependence on peptide length. *Xcorr* (A) and *Xcorr′* (B) plotted versus length of assigned peptide for correct (red triangles) and incorrect (blue circles) training data search results for $[M+2H]^{2+}$ precursor ions. For clarity, only 20% of the data (randomly selected) is shown.

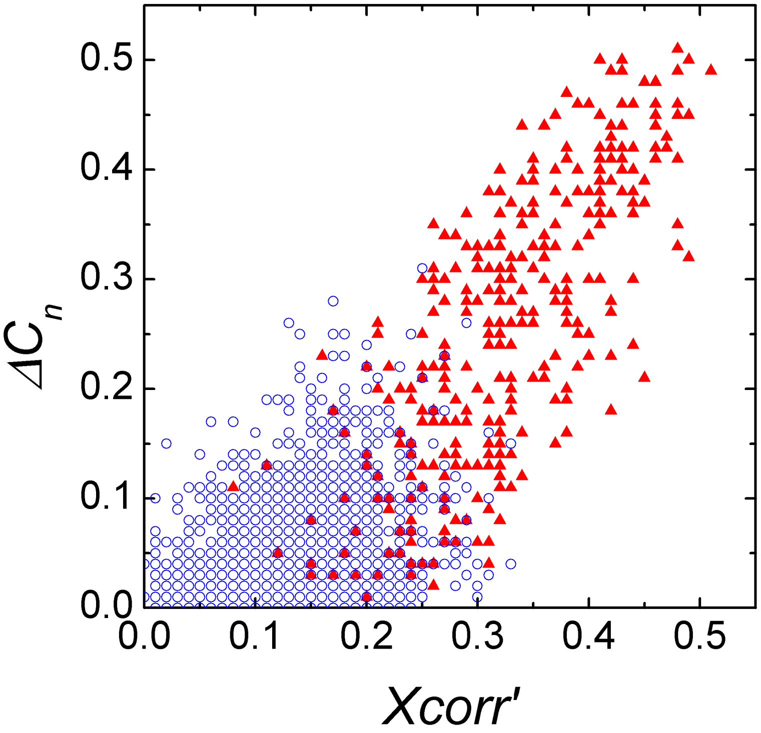

Figure S2. Separation between correct and incorrect results using *Xcorr'* and ΔC_n. ΔC_n plotted versus *Xcorr'* for correct (red triangles) and incorrect (blue circles) training data search results for $[M+2H]^{2+}$ precursor ions. For clarity, only 20% of the data (randomly selected) is shown.

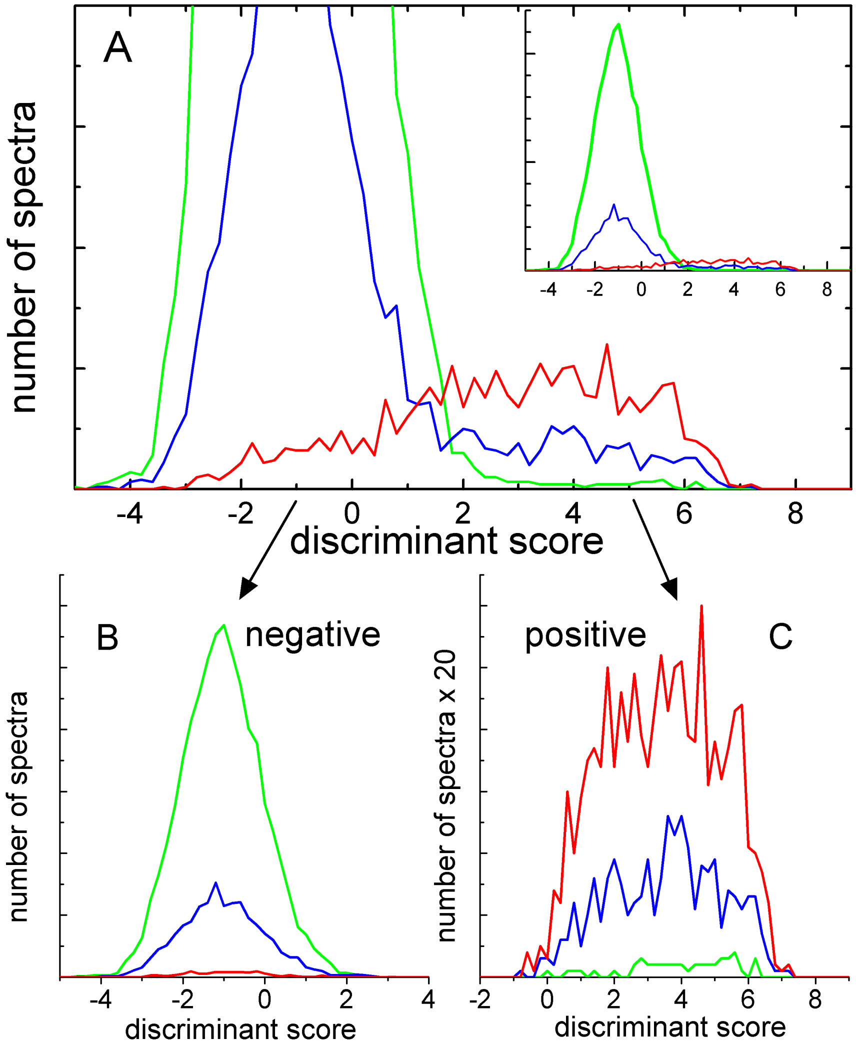

Figure S3. EM algorithm learns underlying negative (incorrect peptide assignments) and positive (correct peptide assignments) discriminant score distributions from the observed total distributions for training data spectra of $[M+2H]^{2+}$ ions. Observed total (A) and underlying negative (B) and positive (C) discriminant score distributions for peptide assignments with number of tryptic termini (*NTT*) equal to 2 (red), 1 (blue), and 0 (green). Inset to (A) shows full-scale observed distribution. The positive distribution was scaled by multiplying the number of spectra by 20.

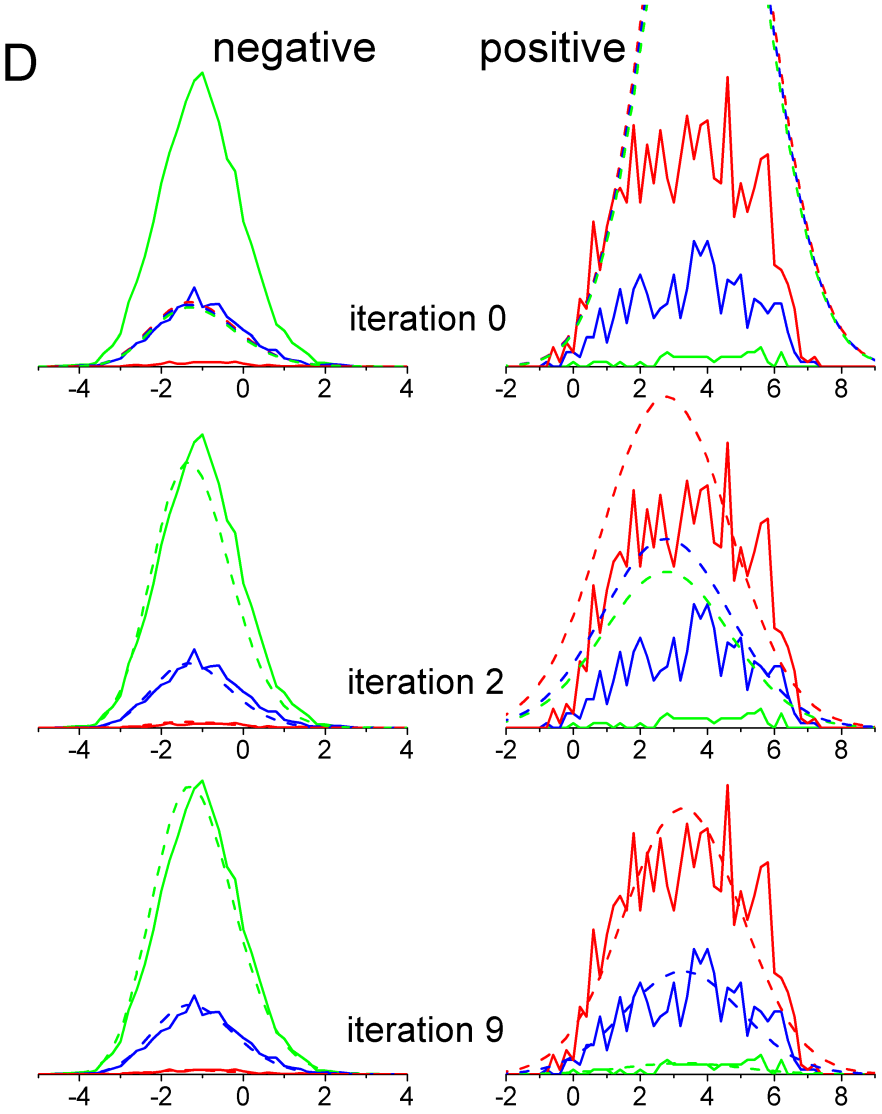

Figure S3 (continued). (D) Underlying negative and positive discriminant score distributions (solid lines), and those learned by the EM algorithm (dashed lines) after indicated number of iterations, for peptide assignments with number of tryptic termini (*NTT*) equal to 2 (red), 1 (blue), and 0 (green). Number of spectra are plotted along the vertical axis. The positive distributions were scaled by multiplying the number of spectra by 20. At iteration 0, the EM algorithm is initialized with distributions that are identical for all three values of *NTT*. At iteration 9, corresponding to termination of the algorithm, there is good agreement between the underlying and learned distributions.