Shotgun proteomic analysis of S-thiolation sites of guinea pig lens nuclear crystallins following oxidative stress in vivo

Giblin, David, Wilmarth, Leverenz, and Simpanya

2012

Appendix 1, Supplemental Methods

Mass spectrometry data analysis

Databases: Version 62 of the Ensembl FASTA protein database (19,774 *Cavia porcellus* sequences) was downloaded on 6/1/2011. We used reversed databases to estimate error thresholds (Elias and Gygi, 2007). The database sequences and their reversed sequences were appended to 179 common contaminant sequences and their reversed forms for a final database of 39,906 sequences. The database processing was performed with Python scripts available at http://www.ProteomicAnalysisWorkbench.com.

SEQUEST searching: RAW data from the mass spectrometer were converted to DTA files representing individual MS2 spectra using DTA Extract in Bioworks (version 3.3; Thermo Scientific); charge state analysis performed using the ZSA option in BioWorks. The group scan minimum count was 1, a minimum of 25 ions were required, the mass tolerance for combining DTAs was set to 0.0001 Da to prevent combining DTA files, an absolute intensity of greater than 500 was required, and MH+ values had to be in the range of 550 to 4000 Da.

Lens proteome determination: SEQUEST (version 28, revision 12, Thermo Scientific) searches for all samples were performed with trypsin specificity; the maximum number of missed cleavages allowed was 2. Average parent ion mass tolerance was 2.5 Da. Monoisotopic fragment ion mass tolerance was 1.0 Da. The ion series used in scoring were b and y. A static modification of +57 Da was added to all cysteine residues.

We used a linear discriminant transformation to improve the identification sensitivity from the SEQUEST analysis (Keller et al., 2002; Wilmarth et al., 2009). SEQUEST DTA and OUT files were compressed, converted to SQT and MS2

files (McDonald et al., 2004), linear discriminant function scores computed from SEQUEST scores, and discriminant score histograms created separately for each peptide charge state (1+, 2+, and 3+) using in-house Python programs described previously (Wilmarth et al., 2009). Separate histograms were created for matches to forward sequences and for matches to reversed sequences for all peptides of 7 amino acids or longer. The score histograms for reversed matches were used to estimate peptide false discovery rates (FDR) and set score thresholds for each charge state that achieved the desired 1% peptide FDR. Much smaller SQT and MS2 files were written that contained only the spectra passing the score thresholds. The sets of confidently identified peptides for each lens sample were collectively mapped to the protein database. Any proteins identified by identical sets of peptides were grouped together as redundant proteins. Any proteins identified by a peptide set that was a formal subset of another protein's peptide set were removed (parsimony principle). Any proteins that were not identified by at least two distinct peptides having two tryptic termini per sample were removed from the final list of 520 identified lens proteins.

Modified peptide detection: A lens proteome database consisting of the 520 lens proteins and their reversed forms was used in SEQUEST searches configured for no enzyme cleavage specificity and with several variable modifications. The variable modifications were cysteine residues with an additional mass of 248 (the net mass of GSH adducts in excess of the static cysteine alkylation mass of 57 Da), cysteine residues with an additional mass of 62 (net mass increase of cysteinylation given a static C+57 alkylation mass), and methionine with an additional mass of 16 Da. Score histograms were created for each charge state (1+, 2+, or 3+), for each number of tryptic termini (2, 1, or 0), and for each homogeneously modified peptide form having at most two modifications per peptide. Score thresholds were set at a 1% peptide FDR independently across the 36 score histograms. Any peptide classes with score histograms that lacked fewer than 20 target peptide match scores in excess of the highest scoring decoy matches were excluded. Sample score histograms for 30-HBO-treatment 2+ peptides are shown in Figures 1-12 below.

Use of the very small 1040 protein database was necessary given the several fold increase in search times due to non specific enzymatic cleavage and several variable modifications; however, it increased the chance that the small numbers

of incorrectly identified peptides might match to one of the 520 target lens proteins. Three distinct peptides per protein were used during results reporting to reduce noise at the protein level. The concept of protein FDR is not applicable in searches using databases of identified proteins and was not computed. Table 1 lists dataset sizes and numbers of identified spectra at a 1.0% peptide FDR.

Combining similar proteins. Peptides that were shared by more than one protein were split in PAW analysis based on relative unique peptide counts of the sharing proteins. This method can become less reliable if there are large numbers of shared peptides and low numbers of unique peptides. From supersets of all proteins containing any shared peptides with each other, we identified pseudo-redundant proteins, pseudo-subset proteins, and highly similar "sibling" proteins which were iteratively combined into single entries until the number of protein "families" were stable. Unique and shared peptide status was updated accordingly and split peptide counts recomputed for protein expression estimates.

Spectral count normalizations: An identical amount of digested protein was used for each mass spectrometry lens sample. Assuming that the number of confidently detected peptides is proportional to the total amount of protein present in each sample, each sample should ideally generate similar total numbers of identified peptides after contaminants such as trypsin are removed from the totals. Mass spectrometer and HPLC performance is somewhat variable so that each sample produced different total numbers of detected peptides (see Table 1 below). In this experiment, the total number of non-contaminant spectral counts differed by less than 2% between the two samples, and no normalization was necessary.

References:

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

Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem *74*, 5383-5392.

McDonald, W. H., Tabb, D. L., Sadygov, R. G., MacCoss, M. J., Venable, J., Graumann, J., Johnson, J. R., Cociorva, D., and Yates, J. R. r. (2004). MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. Rapid Commun Mass Spectrom *18*, 2162-2168.

Wilmarth, P. A., Riviere, M. A., and David, L. L. (2009). Techniques for accurate protein identification in shotgun proteomic studies of human, mouse, bovine, and chicken lenses. J Ocul Biol Dis Infor *2*, 223-234.

Table 1: Dataset sizes and identified spectra counts (1% peptide FDR).

Sample	MS2 scans	IDed scans	IDed lens protein scans
30T CTL	303,067	34,901	32,946
30T HBO	288,647	34,831	33,572

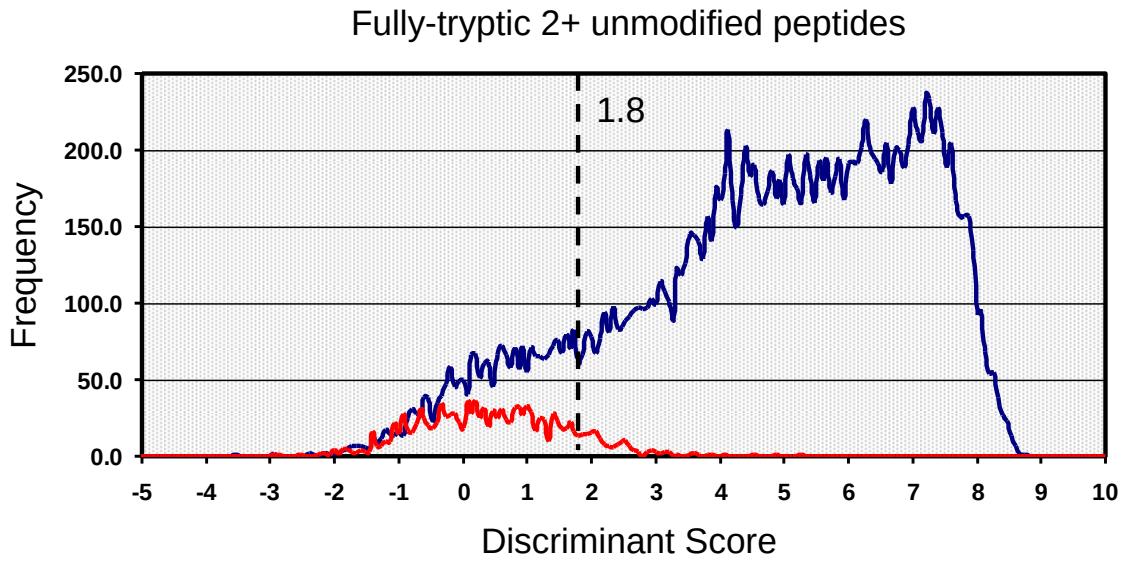

Figure 1: Unmodified fully-tryptic 2+ peptide score distributions. Target matches are in blue, decoy matches are in red. Score threshold was 1.8000.

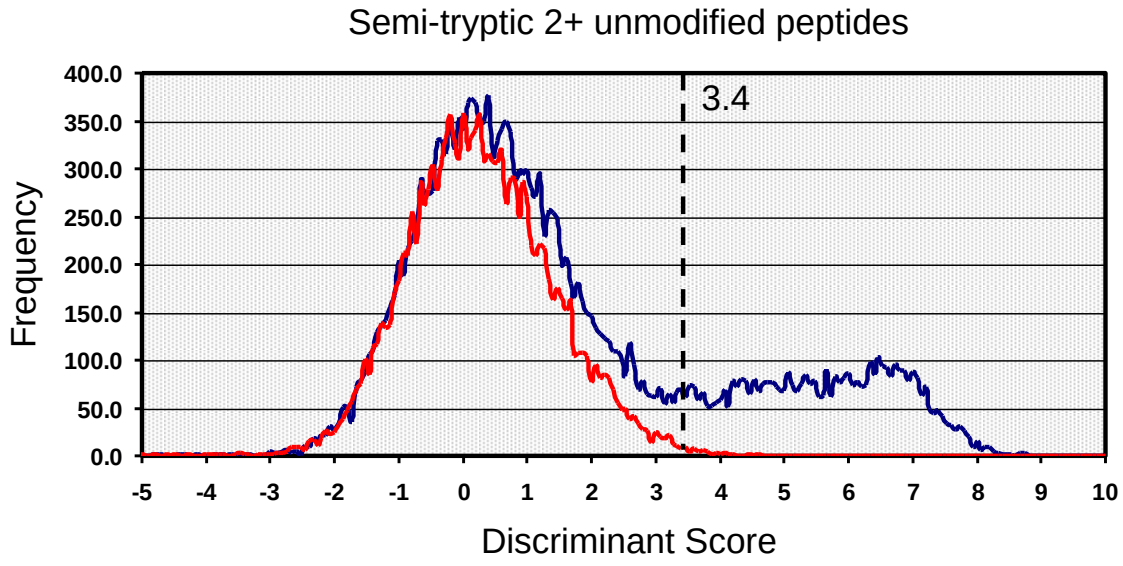

Figure 2: Unmodified semi-tryptic 2+ peptide score distributions. Target matches are in blue, decoy matches are in red. Score threshold was 3.4000.

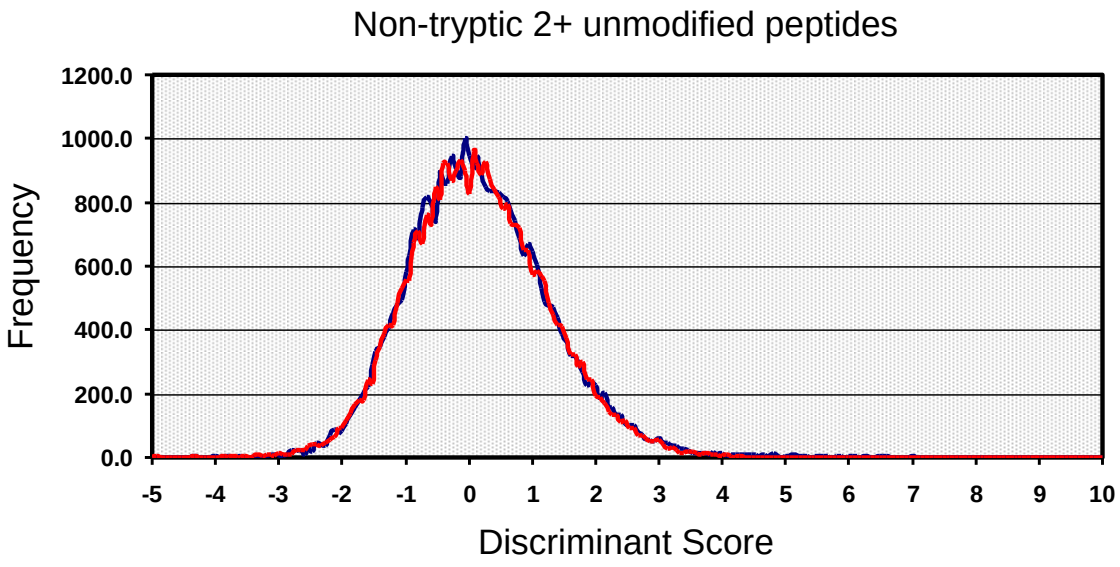

Figure 3: Unmodified non-tryptic 2+ peptide score distributions. Target matches are in blue, decoy matches are in red. Non-tryptic peptides were excluded.

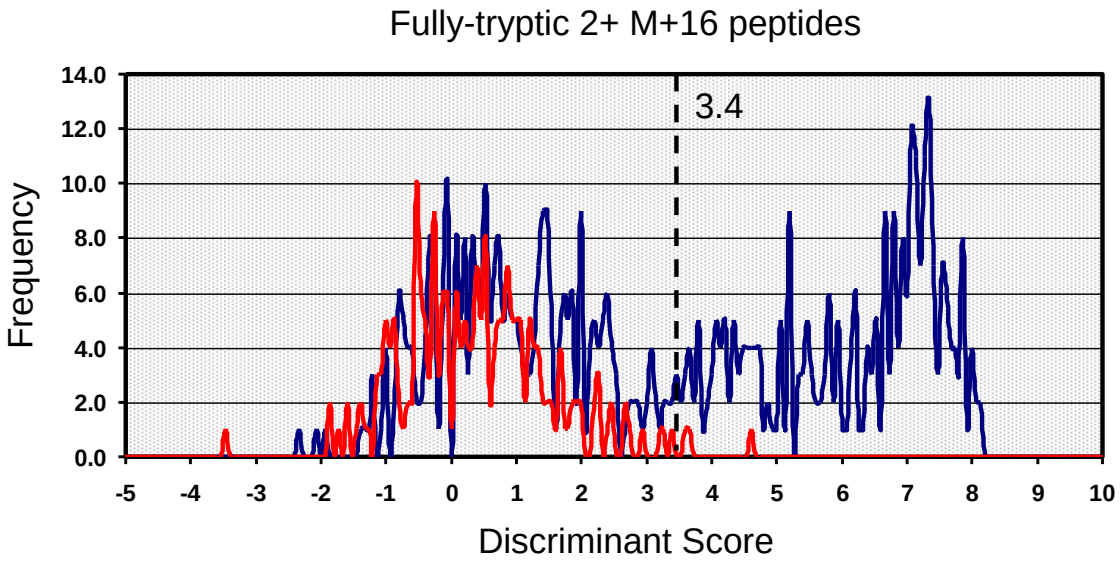

Figure 4: M+16 fully-tryptic 2+ peptide score distributions. Target matches are in blue, decoy matches are in red. Score threshold was 3.4000.

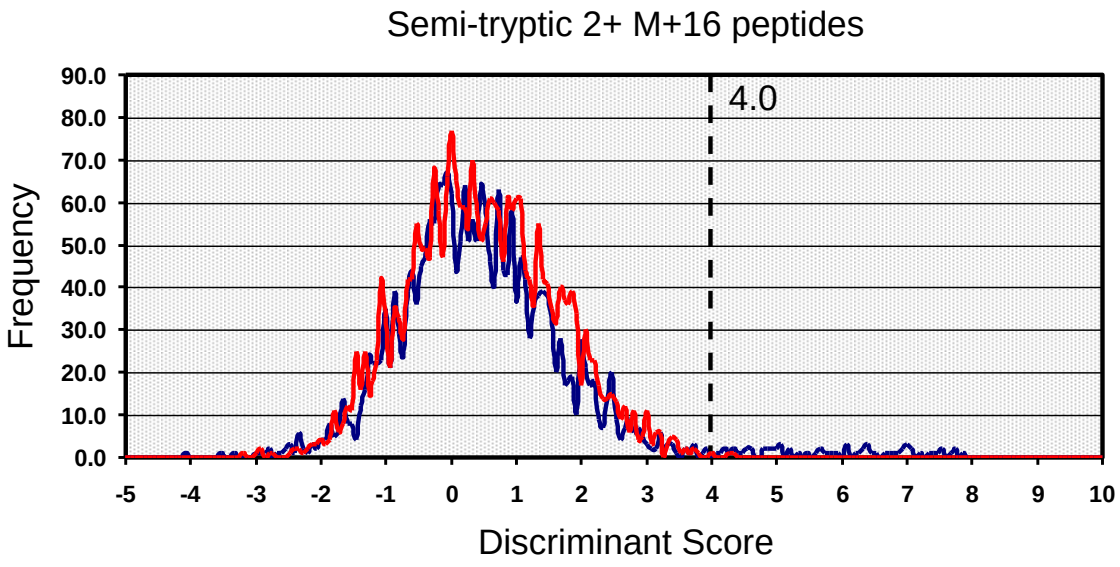

Figure 5: M+16 semi-tryptic 2+ peptide score distributions. Target matches are in blue, decoy matches are in red. Score threshold was 4.0000.

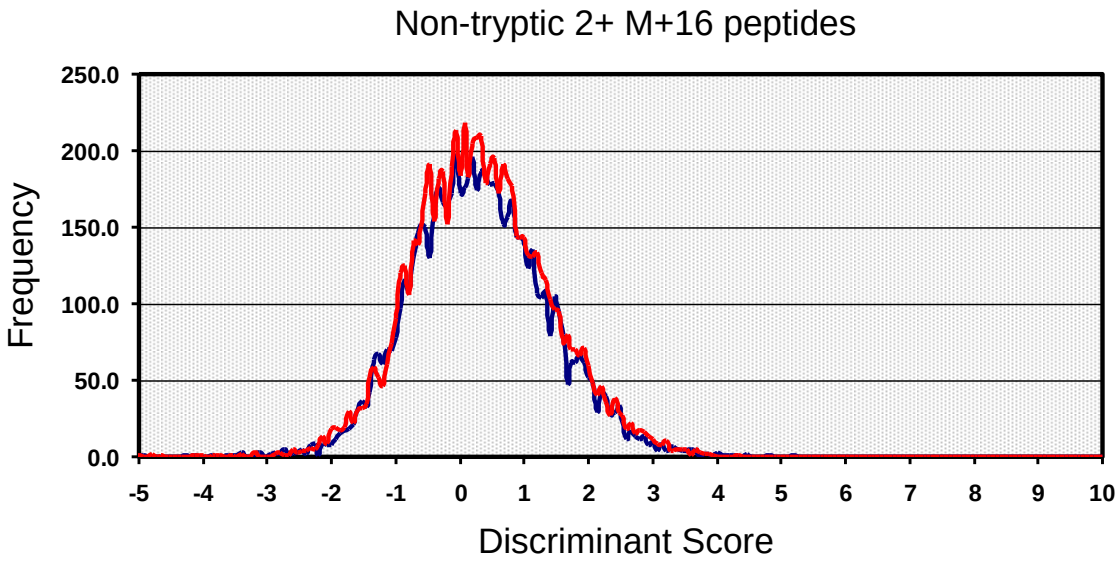

Figure 6: M+16 non-tryptic 2+ peptide score distributions. Target matches are in blue, decoy matches are in red. Peptides were excluded.

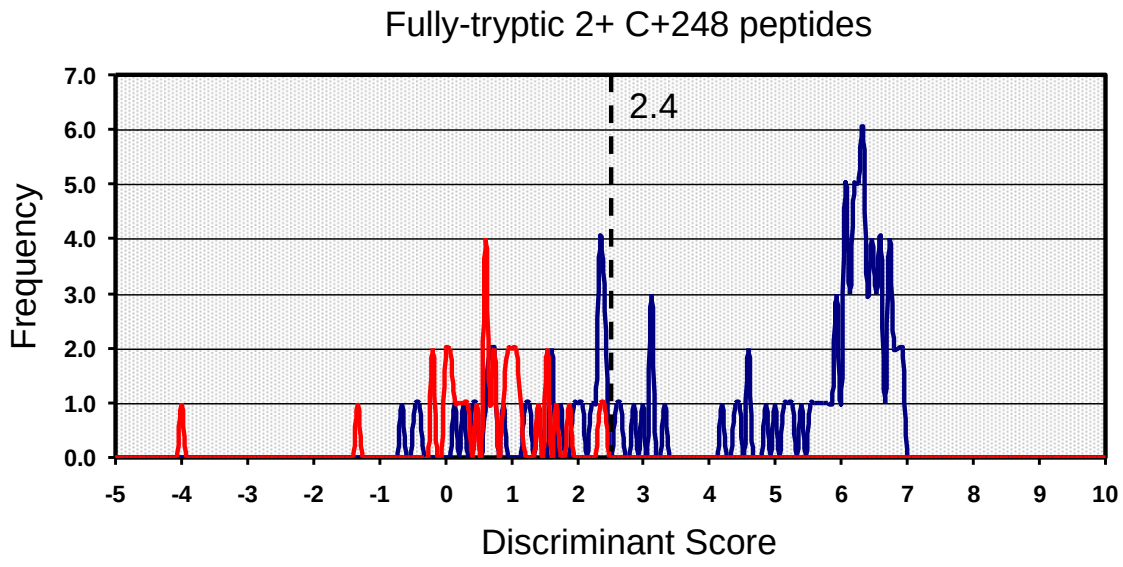

Figure 7: C+248 fully-tryptic 2+ peptide score distributions. Target matches are in blue, decoy matches are in red. Score threshold was set at 2.4000.

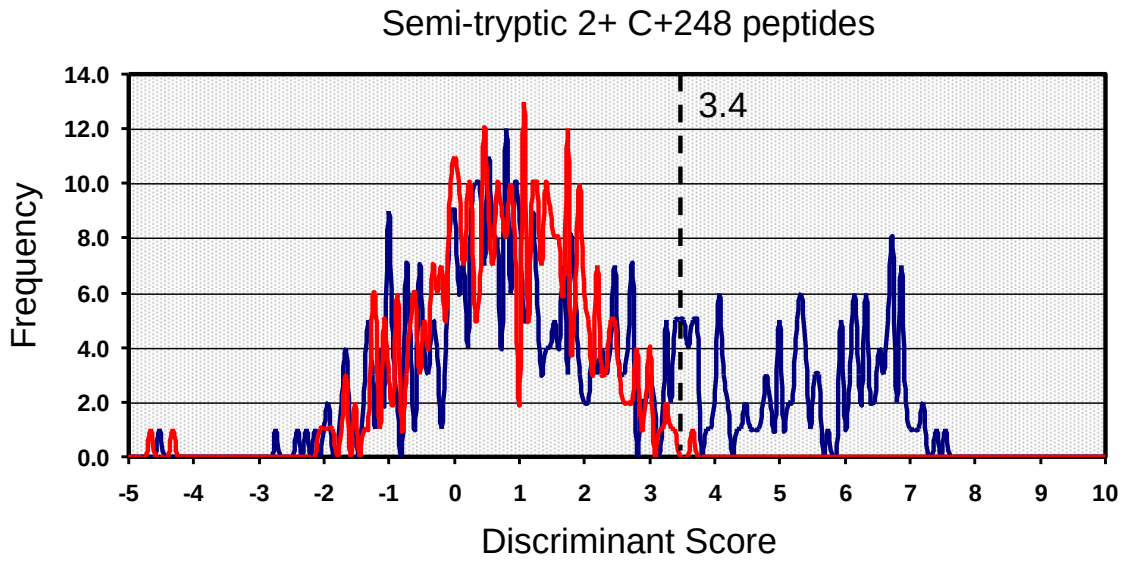

Figure 8: C+248 semi-tryptic 2+ peptide score distributions. Target matches are in blue, decoy matches are in red. Score threshold was set at 3.4000.

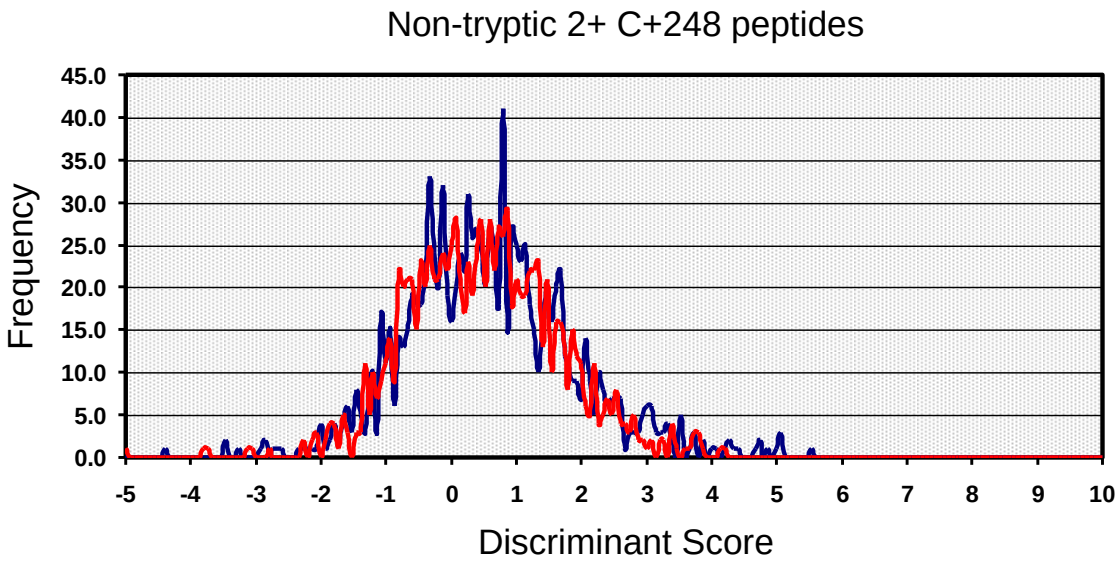

Figure 9: C+248 non-tryptic 2+ peptide score distributions. Target matches are in blue, decoy matches are in red. Peptides were excluded.

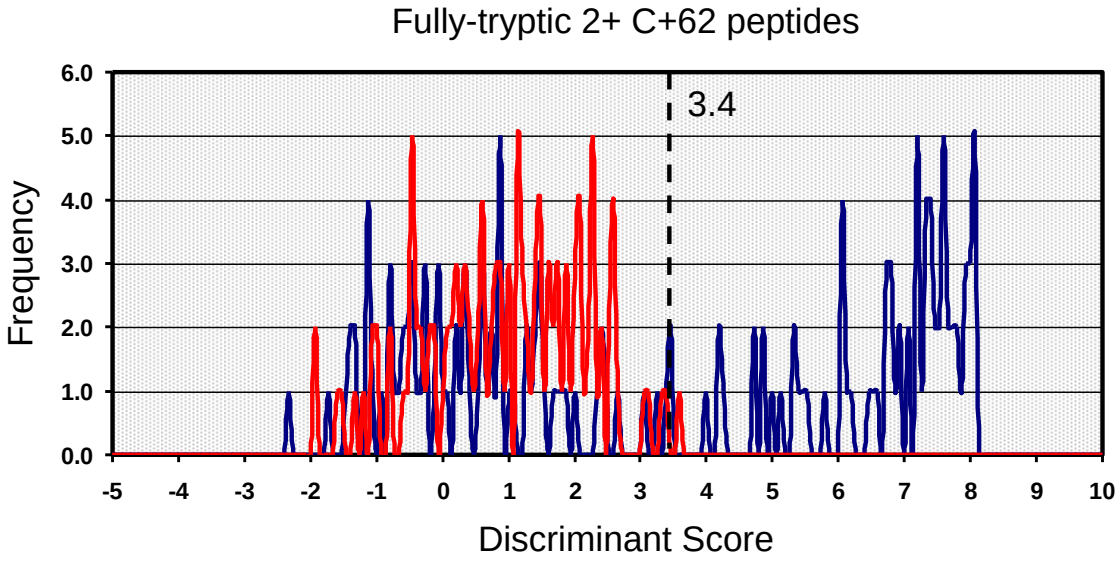

Figure 10: C+62 fully-tryptic 2+ peptide score distributions. Target matches are in blue, decoy matches are in red. Score threshold was set at 3.4000.

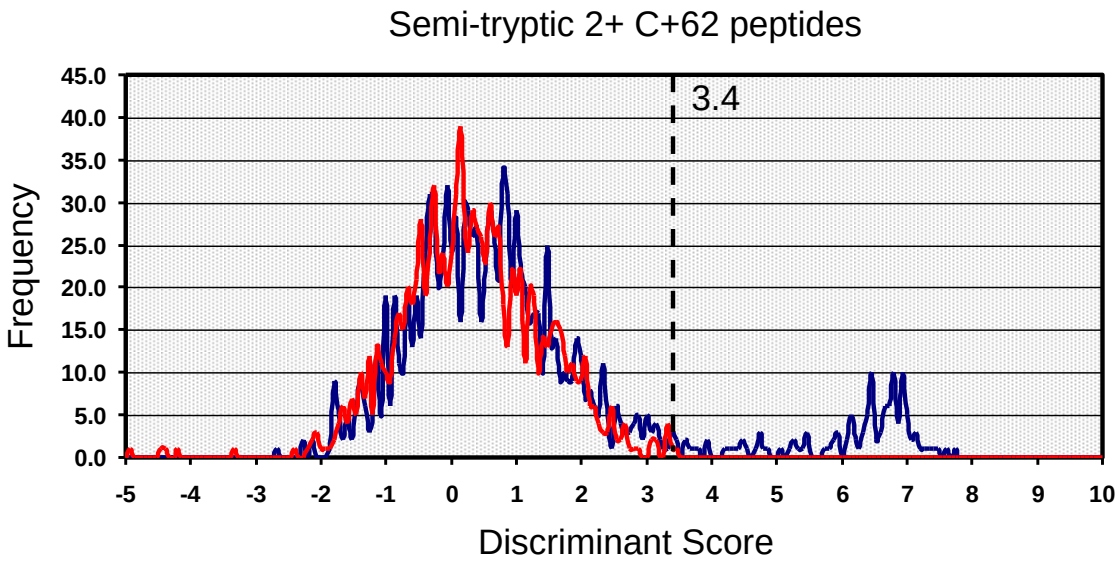

Figure 11: C+62 semi-tryptic 2+ peptide score distributions. Target matches are in blue, decoy matches are in red. Score threshold was set at 3.4000.

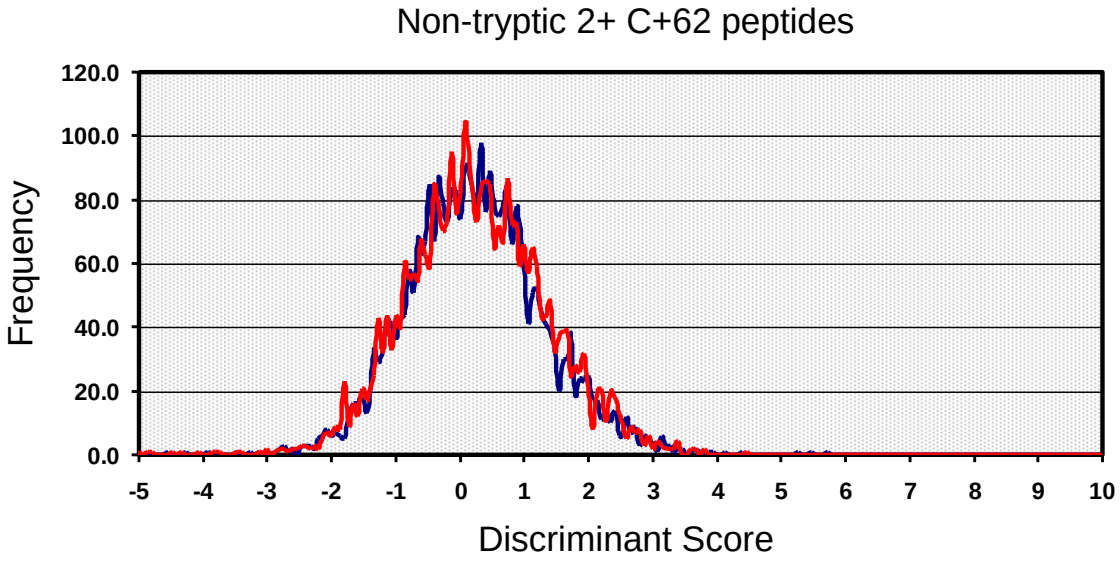

Figure 12: C+62 non-tryptic 2+ peptide score distributions. Target matches are in blue, decoy matches are in red. Peptides were excluded.

[prepared 5/3/2012 by Phil Wilmarth]