MS & Proteomics Resource

Yale School of Medicine
Keck Biotechnology Resource Laboratory

Application Note 4: Understanding how the Mascot Search Algorithm makes protein identifications and using YPED to view the results

Mascot is a powerful search engine used to identify proteins from LC-MS/MS data. See Matrix Science - Home (http://www.matrixscience.com/) for more details on this analysis. The below are some bullet points to assist you in interpreting and understanding the search results.

- **Peptide Scoring** - In Mascot, the score for an MS/MS match is based on the absolute probability (P) that the observed match between the experimental data and the database sequence is a random event. The reported score is -10Log(P). So, during a search, if 1.5 x 10^5 peptides fell within the mass tolerance window about the precursor mass, and the significance threshold was chosen to be 0.05, (a 1 in 20 chance of a false positive), this would translate into a score threshold of 65.
- The use of **red** and **bold** typefaces is intended to highlight the most logical assignment of peptides to proteins. The first time a peptide match to a query appears in the report, it is shown in bold face. Whenever the top ranking peptide match appears, it is shown in red. This means that protein hits with peptide matches that are both bold and red are the most likely assignments. These hits represent the highest scoring protein that contains one or more top ranking peptide matches
- **Expectation value** for the peptide match: this is the number of times we would expect to obtain an equal or higher score, purely by chance. The lower this value, the more significant the result.
- **Sequence of the peptide** in 1-letter code. The residues that bracket the peptide sequence in the protein are also shown, delimited by periods. If the peptide forms the protein terminus, then a dash is shown instead.
- If any **variable modifications** are found in a peptide, these are listed after the sequence string.
- The **protein score threshold** varies for each database searched. The protein score is the sum of the ions scores of all the non-duplicate peptides. Where there are duplicate peptides, the highest scoring peptide is used.
- A mascot peptide score of 50 DOES NOT mean a probability of 0.05. See the first bullet above. But all searches are done using a significance threshold p<0.05
- The **Decoy score** tells you the false positive levels. This is against a database in which the sequences have been reversed or shuffled.
- The **identity threshold** is calculated from the # of trials (or # of candidate peptides with the same precursor MW in the database)
- The **homology threshold** is an empirical measure of whether the match is an outlier. Basically, the identity threshold is a conservative match while the homology threshold can provide a useful # of additional true positive matches without exceeding the specified false positive rate

YPED (the Yale Protein Expression Database)

YPED was designed to address the storage, retrieval, and integrated analysis of high throughput proteomic and small molecule analyses. For proteomics data, YPED handles data from the following analyses: LC-MS/MS protein identifications and protein posttranslational modification (i.e. phosphorylation, ubiquitinylation, acetylation, methylation, etc.); identification/quantitation results from label-based proteomics experiments (such as DIGE, iTRAQ, ICAT, and SILAC experiments); LCMS-based label-free quantitative proteomics; and targeted proteomics (MRM).

After logging into YPED (http://yped.med.yale.edu/), you need to click on the sample of interest. This will pull up the Mascot search results. **The Score and Expectation values are described above**. The highest score (in the first column) and with a low expectation value (in the second column) will be the best match(es). YPED is sortable so you can sort the column by score, protein ID etc. When the file is first opened, it is sorted by score. Clicking on the Protein ID will link out to the database used for searching and this accession number. The % coverage shows where the peptides that were identified match to the indicated sequence. Clicking on the view peptides, will show the peptides identified with the *peptide* score (not the protein score) and the *peptide* expectation values).

In the **peptide view table**, the m/z column is the mass to charge ratio. Using electrospray mass spectrometry, the masses are typically +2 and +3 charge states. The charge state is found in the last column in this view. Hence, a m/z value of 679.35 which is doubly charged, has a peptide mass (found in the Ion Mass column) of M=1356.7 and an M+H= 1357.7. The Ion Mass (calc) column shows the theoretical mass. The Delta column is the mass difference between the Ion Mass (which is from the experiment) and the Ion Mass (calc) and is in Da. The ppm column shows you the mass difference as calculated in parts per million.

The peptides in the **peptide view table** are grouped based on the identity and homology scores. Mascot calculates these 2 scores for each peptide. The scores are not fixed and will change. The homology threshold helps determine if the peptide match is an outlier. Hence, the best peptide matches will be above the homology score. Those below this score are shown in the bottom of the table. Peptides with scores higher than the homology score but less than the calculated identity score are in the middle section of the table. Again, the identify score is based on the probability of getting a false positive match. (see above). Hence, the best peptide matches are seen in the top part of this table with scores above homology and above the identity score.

Excel tables can be created by clicking on the Export options at the bottom the tables (for both the protein view and the peptide view).