Proteins Page (Proteome Discoverer or PD)

Description of the worksheets:
Raw data – Unprocessed output from the PD software.
Result – Includes only the Master proteins and relevant data columns.

Definition of useful parameters:

Understanding Decoy Protein and FDR (for advanced reading):
During the database search, the Percolator node in PD estimates the number of false positive protein identifications by using a decoy database containing reversed protein sequences.

The Protein FDR Validator node in PD goes through the list of target proteins from top to bottom and calculates the false discovery rate (FDR) that would result if it used the target score of a particular protein as a threshold. It obtains this threshold by dividing the number of target proteins by the number of decoy proteins, or:

$$\frac{number\ of\ target\ proteins}{number\ of\ decoy\ proteins}$$

The application calculates this threshold for every target protein.

However, the FDR values do not monotonically increase when the score threshold decreases. When you lower the threshold further, the FDR usually decreases first, because you add additional target proteins that have a score above the threshold before you add the next decoy protein. For example, at a score threshold of 143, 1000 targets pass. With 10 decoys, an FDR of 1.0% results. Next the application lowers the score threshold by going to the next target of 142, then 141, then 137, and so forth. This methodology adds targets that pass the threshold. Because there are far fewer decoys, it is likely that it can take some time until you reach a score threshold where you add the next decoy score. Assume the next decoy has a score threshold of 90. From score 143 down, the FDR decreases at first as more targets, but not more decoys, pass. At score 91, 1050 targets but only 10 decoys yield an FDR of 0.95%. At score 90, 1100 targets but 11 decoys also yield an FDR of 1.0%.

If you want to filter by 1% FDR, which score threshold is best to use, 143 or 90? Both result in an FDR of 1%, but 90 would yield 1100 targets, and 143 would yield only 1000. To circumvent this problem, the application uses the q-value (see below), which is defined as the minimum FDR threshold at which a given target would be included in the results.

Protein FDR Confidence

Green/High: Displays high-confidence proteins, starting with either the first decoy protein or the protein FDR that reaches the high FDR threshold, whichever occurs later. **Default threshold is 1%.**

Yellow/Medium: Displays medium-confidence proteins, starting with the protein following the first decoy or highest protein threshold. **Default threshold is 5%.**

Red/Low: Displays low-confidence proteins, starting with the protein following the second decoy protein or the protein FDR that reaches the medium FDR threshold, whichever occurs later.

Protein FDR Validator node parameters

Parameter	Definition
Target FDR (Strict)	Specifies a target false discovery rate for protein matches of high confidence. High-confidence proteins are those with a q-value higher than the specified threshold. Range: 0.0–1.0 Default: 0.01
Target FDR (Relaxed)	Specifies a target false discovery rate for protein matches of medium confidence. Medium-confidence proteins are those with a q-value higher than the specified threshold. Range: 0.0–1.0 Default: 0.05

Understanding Protein Grouping:
During the database search, the Protein Grouping node in PD combines all proteins into one protein group that contains the same subset of peptides. As we set the Apply Strict Parsimony Principle parameter of the Protein Grouping node to True, the application removes all protein groups that have no unique peptides among the peptides that it considers for the protein grouping process.

Master protein
In a Protein Group, **the master protein is the protein with largest value in the # Protein Unique Peptides column and with the smallest value in the Coverage column (the longest protein).** If more than one protein in a group has the same score, an equal number of PSMs, and an equal number of peptides, the protein with the longest sequence is designated as the master protein.

Exp. q-value

It is defined as the minimum FDR threshold at which a given target would be included in the results. **For our workflows, q-values are less than 0.01 for high confidence hits and 0.05 for medium confidence hits.**

Sum PEP Score

The posterior error probability (PEP) is the probability that the observed PSM is incorrect. For example, if the PEP associated with (EAMRPK, s) is 5 percent, there is a 95 percent chance that the EAMRPK peptide was in the mass spectrometer when spectrum s was generated.

The Protein FDR Validator node calculates a new protein score by multiplying the PEP values of the peptides connected to the protein. To make it numerically more stable, because PEP values are very small numbers, it actually sums the logarithms of the PEP values as follows:

$$Sum\ PEP\ Score = \sum_{i} -\log_{10}(PEP_{best\ peptide,i})$$

For the calculation, the node first groups the PSMs of the protein by sequence, charge, and theoretical mass and then uses the best value—that is, the minimum PEP value within these groups—to calculate the sum PEP score:

$$PEP_{best\ peptide,\ i} = \min(PEP_{PSM,i,1} \ldots PEP_{PSM,i,n})$$

Decoy Proteins shows the number of decoy proteins above the given sum PEP score.

Coverage displays the percentage of the protein sequences covered by identified peptides.

Peptides displays the total number of distinct peptide sequences identified.

PSMs displays the number of identified peptide spectrum matches. **This is an indicator of the protein abundance.**

Protein Unique Peptides displays the total number of peptides that are truly unique to a particular protein.

Unique Peptides displays the total number of distinct peptide sequences unique to the protein group. **Since we only report the master proteins (Protein groups), this is the parameter to use.**

Protein Groups displays the total number of protein groups a particular protein belongs to.

Area displays the average area of the **three unique peptides** with the largest peak area.

emPAI
The exponentially modified protein abundance index (emPAI) is **a simple measurement of protein abundance that is based on the number of found peptides**. It correlates to the absolute amount of protein in a sample.
The PAI is calculated as follows:

$$\mathrm{PAI} = \frac{N_{obsd}}{N_{obsbl}}$$

where:
• Nobsd is the number of observed, or found, peptides.
• Nobsbl is the number of observable peptides.
The emPAI itself is an exponential transformation of the PAI:

$$\text{emPAI} = 10^{PAI} - 1$$

The application calculates the emPAI only for tryptic digests.

Score Sequest HT
For Sequest HT results, the score is the sum of all peptide XCorr values above the specified score threshold. The score threshold is calculated as follows:

0.8 + peptide_charge × *peptide_relevance_factor*

where peptide_relevance_factor is an advanced parameter of the Sequest HT node with a default value of 0.4. For each spectrum, only the highest-scoring match is used. For each spectrum and sequence, the application uses only the highest scored peptide.