

How to interpret results from shotgun MS analysis

A quick guide

MQ-PW, 27.01.2020, version 5.6

Results for “shotgun” protein identification experiments are obtained through database searches with the software **MASCOT**. For more information and a useful help section on protein identification by MS, please see the MASCOT site (<http://www.matrixscience.com/>).

The data in its simplest form is delivered as an Excel table. In the top part of the table you can find information on the data analysis process such as the database used, taxonomy, mass accuracy and the criteria for validation of peptide and protein identification. Then the list of matched proteins follows, with a description, number of similar matches, accession numbers and a column for every sample that was submitted. These columns contain usually the number of spectra (thus not peptides) that were matched to that particular protein in every sample. The full dataset is also delivered in the format for the Scaffold Viewer software tool (.sf3), which is described below.

Some important background concepts

1) This is not « true » protein sequencing. What we are doing is matching fragmentation spectra for trypsin fragments (peptides) of proteins to a database sequence. These spectra do contain sequence information which is however of variable quality and completeness. Since identity is only established by a match to the database, the results can only be good if the correspondence between the database and the organism being studied is good. In other words, if the database does not contain the sequence(s) of the protein(s) you are analyzing nor a close homologue, there will be no match and no results, even with the best of data. Now, if you are working with one of the common model organisms with sequenced genomes, databases are fairly complete and this is not a concern.

2) Strictly speaking, we are not identifying proteins by mass spectrometry, but peptides. These peptides are mostly between 7 and 30 amino acids long, the average is 11 AA. After having matched peptides, the software we use (MASCOT) proceeds to carry out protein inference. This means to derive which sequence(s) in the database contain(s) a given set of peptides. This can yield very univocal identifications if there are enough unique sequences matched. However several database sequences are often matched by the same set of peptides. This can happen because: i) highly homologous protein families exist, and these protein differ only by a few AA (ex. the tubulins) and ii) databases can be redundant and contain several nearly-identical sequences. It is also important to know that the software reports the **minimal set of protein sequences, which explain the maximum number of identified peptides** (principle of parsimony).

3) The data acquisition process is - to a certain degree – random. This means that in a very complex mixture of peptides not always the same ones are chosen for « sequencing », in particular with low abundant peptides. If a certain redundancy of sampling is present, the data can be very reliable, but on the other hand identifications with low number of peptides must be taken with a lot of caution (see below).

Important things to know for interpretation of a list of proteins

1) This table contains the essential results about your experiment. At the same time, you should receive from us a link to download a file containing the full data (but not the raw data). This file will be in a format for the software Scaffold viewer. You can download and use Scaffold for free from www.proteomesoftware.com. It is a fairly intuitive software which will also allow you to export your data very conveniently (see below). But the advice of an experienced mass spectrometrist can still be necessary to decide to validate or not borderline identifications.

2) The database we use in most cases is UniProt (<http://www.uniprot.org/>), which is a fusion of Swiss-Prot and TrEMBL databases.

* Swiss-Prot is a curated protein sequence database, which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases.

* TrEMBL is a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.

* Additionally, UniProt provides for an increasingly number of organisms a reference proteome database, which includes both manually reviewed (UniProtKB/Swiss-Prot) and a selection of unreviewed (UniProtKB/TrEMBL) entries.

3) Check the taxonomy of the species we have used for database search. Lets us know if it is not the correct one !

4) Identified proteins: in the « accession number » (2nd column) you can see if more than one database entries was matched by the same set of peptides. Column one («identified proteins ») gives you the annotation available for the accession number shown. Unfortunately the software does not always choose the best annotated database entry to report. So it is worth looking directly at the Scaffold file to see if a more informative sequence is reported.

Example:

Identified proteins	Accession number	Molecular Weight	Sample 1	Sample 2
cDNA FLJ78508	A8K7C2_HUMAN (+2)	42 kDa	23	31

In this example a TrEMBL sequence is reported while there are Swiss-Prot entries also matched (+2). These usually have a much better annotation.

5) The **Molecular Weight** in the third column of the table is a theoretical one calculated from the database sequence. Of course, this does not imply that the real mass of the protein actually corresponds to this value. Databases usually list the precursor (unprocessed) sequence. Also, if the protein detected in your sample is a fragment corresponding to only a portion of the sequence, this can only be deduced by looking in detail at the sequence coverage (i.e. where the matched peptide are located in the sequence). You need to check the full data (Scaffold file) for this.

6) The numbers of assigned spectra give an idea of the confidence of the identification. Although we can identify proteins with only one spectrum, confident identifications need usually a minimum of **two distinct peptides**. Remember that what you have listed in this table is the number of assigned spectra, not peptides. One and the same peptide can be matched several times. To know exactly how many distinct peptides have been matched you need the full data (Scaffold file). However as a rule of thumb for evaluating your Excel table, **you should take as good all identifications with 4 or more spectra. Consider with a lot of caution all those underneath this value.**

7) A certain linearity exists between the number of spectra assigned to a certain protein and its concentration. The numbers of matched spectra can thus be used to make **semi-quantitative estimates** of protein amount in the sample (this approach is called "**spectral counting**"). Of course this linearity is good when the numbers of spectra matched are sufficiently high (>10). With lower numbers the relationship is a lot less reliable. In other words, if you are comparing Tubulin in samples A and B where it is identified with 100 and 300 spectra, respectively, you can assume that there is a certain difference (2-3x) in the amount of tubulin present. However if the insulin receptor was identified with 3 spectra in sample A and 1 spectrum in sample B, you cannot really conclude that the difference in concentration is significant. The same is true if the values are 3 vs 0 spectra. You cannot really conclude that the protein is really absent in sample B in this case. In all cases, spectral counting is even much less (or not at all) reliable to compare amount of different proteins in the same sample.

8) For protein pull-down (IP etc) studies : the question of whether an "interactor" protein that you find is really specific is not a trivial one. Pull-down experiments are subjected to a lot of possible artifacts. The CRAPome database (<https://reprint-apms.org>) contains extensive data about the frequency of unspecific proteins identified in a large number of negative control experiments and can be a useful resource to eliminate unspecific binders.

9) Some links to databases useful to know more about a protein of interest:

General link through EXPASY: <http://www.expasy.org/proteomics>

Protein Databases: <http://www.uniprot.org>
<http://www.nextprot.org> (for human proteins)

Reactome (biological pathways): <http://www.reactome.org>

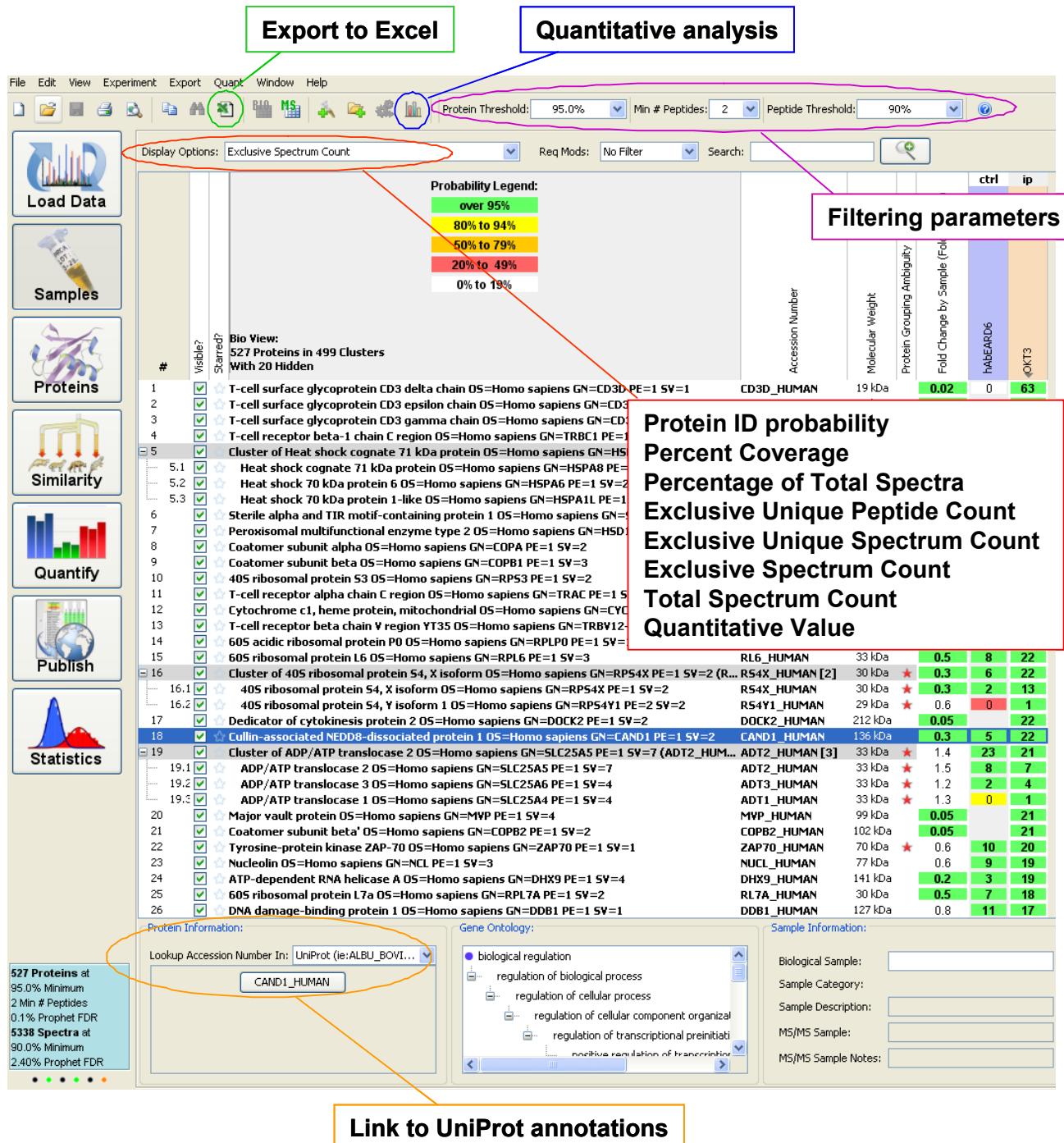
KEGG (biological pathways): <https://www.genome.jp/kegg/pathway.html>

Prediction of protein functional sites: <http://elm.eu.org>

Protein-protein interaction databases :
INTACT database : <https://www.ebi.ac.uk/intact>
HPRD database : <http://www.hprd.org>
DIP: <http://dip.doe-mbi.ucla.edu>
STRING: <http://string-db.org>
Human Proteinpedia: <http://www.humanproteinpedia.org>
BioPlex: <https://bioplex.hms.harvard.edu>

Scaffold software

1) Scaffold files capture the most important data produced by MS and the identification process and are an excellent way to distribute the results. Scaffold allows you to further look interactively at your data, and can be downloaded as a free viewer (<http://www.proteomesoftware.com>) for the Windows, MAC or Linux operating system. See below a snapshot of the Scaffold output main interface:



Please note that it is better to have a reasonably powerful computer with minimum of 8GB RAM to open large Scaffold files. We cannot provide here a full introduction in the capabilities of Scaffold, but the software is quite user-friendly and has a good Help menu. As Scaffold is regularly updated, be sure to have the latest version to be able to open your Scaffold files (.sf3).

- 2) The table can be sorted based on one particular column by clicking on the header.
- 3) Scaffold allows the validation of peptide & protein IDs, the alignment of samples and the quantitative comparison of samples based on spectral counting. Usually 95% protein probability, 90% peptide probability and min. 2 peptides are used as **filtering parameters**, but these settings can be relaxed if you are looking for a specific low abundant protein or peptide. Caveat : these recovered identifications should be used with caution: ask for the advice of an experienced mass spectrometrist!
- 4) Scaffold summarizes the protein list using two levels of hierarchy:
 - **Protein Group (PEG)** is a set of protein sequences that are associated with an identical set of peptides. Protein groups are by default represented by the sequence(s) with the highest probability and the largest associated number of spectra.
 - **Protein Cluster** – is a group of **PEGs** created using a hierarchical clustering algorithm. Proteins member of the cluster share some peptides but not all of them. Protein Clusters are by default represented by the protein that shows the highest associated probability. Clusters can be collapsed or expanded directly in the protein list (-/+ buttons on the left).
- 5) Various **display options** are available to look at the list of identified proteins (see below). "**Exclusive Spectrum Count**" or "**Total Spectrum Count**" is often preferred as it is also the classical basis of (semi-)quantitative comparison of proteins in the different samples (see the option **quantitative analysis**). Others are :
 - **Protein Identification Probability** - Scaffold's calculated probability that the protein identification for any of the MS Samples is correct. Results are color-coded to indicate significant differences in protein identification confidence.
 - **Percentage Coverage** - The percentage of all the amino acids in the protein sequence that were covered by identified peptides detected in the sample.
 - **Percentage of Total Spectra** - The number of spectra matched to a protein, summed over all MS Samples, as a percentage of the total number of spectra in the sample.
 - **Exclusive Unique Peptide Count** - The number of different amino acid sequences (unique peptides), regardless of any modification that are associated with a single protein/protein group.
 - **Total Unique Peptide Count** - *only available with clustering algorithm selected* – The number of different amino acid sequences (unique peptides) that are attributed to a single protein including those shared with other proteins.

- **Exclusive Unique Spectrum Count** - Number of unique spectra associated only with a single protein/protein group. Spectra are considered unique when:

- they identify different sequences of amino acids or peptides;
- they identify different charge states or a modified form of the peptide within the same identified sequences of amino acids.

- **Total Unique Spectrum Count** - *only available with clustering algorithm selected* – The number of unique spectra that are attributed to a single protein/protein group including those shared with other proteins.

- **Exclusive Spectrum Count** - The number of spectra, associated only with a single protein/protein group.

- **Total Spectrum Count** - The total number of spectra associated to a single protein/protein group, including those shared with other proteins.

- **Quantitative Value (Selected quantitative method)** - Scaffold will display the results of the Quantitative Method selected from the Quantitative Analysis Dialog Box.

6) The box “search”, for example by name or accession number, is very useful when you are looking at a specific protein in a list of a few thousands hits! It is also easy to sort the list by protein name, molecular weight, number of spectra, etc., by clicking on the column header.

7) For looking in more details at the sequence coverage, assigned peptides, or spectra of a protein, you can select it (click on it) and use the “Proteins” window of the left menu bar. Note that the **GO annotations** of all identified proteins can be downloaded into the Scaffold file (Menu bar → Experiment -> Add or Edit Annotations). You can also look at alternative accession numbers matched as well as at the detailed UniProt annotations of a selected protein, using the **button/link** in the Protein Information frame.

8) Various export options are available, the most useful being the **Samples Report (Excel table)**.

9) Samples can be grouped in categories (e.g. in the case of replicate samples) and more sophisticated semi-quantitative analyses and statistical tests can be performed in the **quantitative analysis** sub-menu.

10) For more information :

- Searle, B. C. (2010). Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics*, 10 (6), 1265–9.
- [Scaffold User’s Guide \(available from the Help menu in Scaffold\)](#)
- <http://www.proteomesoftware.com>

Then go to their [Technical resources](#) section.