Answers

Below you will find answers to the questions asked as part of the individual sections of the tutorial. If you have any comments regarding the questions, or if the answers are not detailed enough, please let us know, and we will do our best to improve them.

**1.0 Identification**

**1.1 Database Generation**

**[1.1a]** It is also possible to identify spectra using so-called spectral libraries[1](#_ENREF_1), where experimental spectra are compared to already identified spectra. This approach is already widely used for the identification of small molecules and is now becoming a hot topic for peptides[2](#_ENREF_2).

Finally, de novo algorithms3-4 identify spectra by identifying mass signatures of single or series of amino-acids (so-called tags). These do not require the use of databases *a priori*.

**[1.1b]** UniProt provides a grand total of 134,116 protein entries for human. These sequences are inferred from the sequenced genome and curated algorithmically and manually. Interestingly, the entries labelled with a gold star (20,256) are manually reviewed, these proteins are historically called SwissProt entries. The silver star entries on the other hand are algorithmic prediction where no experimental validation is annotated in UniProt.

The identification efficiency is dependent on the size of the database. Notably, large databases (>100,000 sequences) are computationally demanding to search against and statistically result in low identification rates. Unless there is a really good reason to do so, it is hence advised to work with the reviewed sequences. Eventually, it is possible to add other sequences or research the data against the entire UniProt *a posteriori*.

Although the human proteome is one of the most extensively studied, it can be that a protein is missing or presents differences in the amino acid sequence. It is hence important to bear in mind that our reference does not necessarily perfectly reflect reality.

Due to the constant efforts at improving the quality of the database, the content of UniProt evolves with time. It is hence crucial to keep the same version of the database during the entire life of a project. It is also essential to note the date of creation of the database and report it in the publications.

**[1.1c]** UniProt can provide isoforms of protein sequences. These should be used with caution as they dramatically reduce the efficiency of the identification algorithms.

**1.2 Peak List Generation**

**[1.2a]** When programming the mass spectrometer, it will be indicated whether the spectra are recorded in profile mode (requiring peak picking) or in centroided mode (already peak picked). Depending on the instrument, using more advanced signal processing methods can improve the results.[5](#_ENREF_4) Generally, it is always recommended to inspect the data before processing – making it clear whether it needs to be peak-picked – and in case of doubt consult the manufacturer’s instructions.

**1.3 Peptide-Spectrum Matching**

**[1.3a]** (answer pending...)

**[1.3b]** Selecting the correct database is a crucial step in proteomics. First, it needs to be as comprehensive as possible: you cannot find a protein which is not in the database. Moreover, if a protein is missing, the search engines might attach spectra derived from this protein to another resembling protein – hence making a false identification. It is thus crucial that you leave enough room for the search engine to “distribute” mistakes. However, using a too large database will lower your probability to find your proteins.

Generally, it is recommended to use the reference database of your species of interest completed with the sequences of expected contaminants: keratin, proteases used for protein digestion, etc. See the “Database Generation” chapter for more details.

Finally, bear in mind that the content of sequence databases evolves with time. It is hence important to constantly use the same database for a given project and document its version in every communication.

**[1.3c]** There are two types of modifications: modifications induced by the experimental workflow and natural modifications of the sample. Among the modifications occurring when conducting the experiment, some are produced voluntarily like carbamidomethylation of cysteine here and some are experimental artefacts like oxidation of methionine here. These have hence to be selected in order to identify the proteins. The biological modifications on the other hand are selected in order to target biological functions. However these are typically low abundant: we have very little chance to identify a phosphorylated protein without enrichment[6](#_ENREF_4) – we actually here selected phosphorylation for illustrative purpose only.

Selecting variable modifications has a similar effect than using a large database: it increases the number of possible results, hence reducing our chances to identify our proteins. It is hence advised to reduce the number of variable modification. This can be done by selecting fixed modifications: for these, every targeted residue will be a priori considered as modified. Non-modified peptides will hence not be identified: it is to be used only when all peptides are expected to be modified. Here, carbamidomethylation is a high yield chemical process which will target all residues.

In case of doubt, it is very easily to control the level of modifications by doing a pre-search with the modification of interest as variable.[7](#_ENREF_5) Here, searching with oxidation of methionine and carbamidomethylation of cysteine as variable modification returned >98% of cysteine residues modified. The modification can thus reasonably considered as fixed. Note that such quality control steps are for crucial importance when working with chemically labelled samples.[8](#_ENREF_6)

**[1.3d]** Missed cleavages are parts of the peptide sequence where one would expect the protease to cleave. Missed cleavages can occur due to incomplete digestion. Due to the impossibility for the protease to access some cleavage site or protease quality,[9](#_ENREF_7) some missed cleavages will always remain,[10](#_ENREF_8) in our experience up to two with trypsin.[7](#_ENREF_5)

**[1.3e]** With the first low resolution mass spectrometers, searches were conducted with a fixed tolerance in m/z – using the unit Dalton. With the advent of high resolution mass spectrometry, search engines adapted the tolerance to the m/z actually measured – one would allow a higher tolerance when measuring the mass of an elephant than the mass of a mouse – hence introducing ppm tolerance defined as:

The mass tolerances depend on the resolution of the mass spectrometer. Here, the data was recorded in the Orbitrap where a 10 ppm tolerance gives the best results on our setup. OMSSA and X!Tandem do not allow us to set the fragment ion tolerance in ppm so we use the value of 0.02 Da.

**[1.3f]** The data was acquired with higher-energy collision dissociation (HCD) fragmentation[11](#_ENREF_9) which principally generates b and y ions.

**[1.3g]** These modifications are all the OMSSA compatible modifications. Some of them will be better suited for your setup than others. Note that X!Tandem might not account for the difference between these OMSSA modifications. For more information on the handling of modifications by search engines, please contact the developers of these.

**[1.3h]** Before a peptide or a fragment ion is recorded, it can lose a moiety named neutral loss. Most encountered neutral losses are water (H2O) and ammonia (NH3) losses. Some modifications like phosphorylation can also generate neutral losses and these can be set in this dialog. Note that this information is not accounted for by OMSSA and X!Tandem.

Some modifications can also lose charged moieties, named reporter ions or diagnostic ions. This is for instance used for reporter ion based quantification.12-13

**[1.3i]** The search time usually scales with the number of spectra and their complexity. A similar effect goes for the database size. Notably, when using large databases, OMSSA will get stuck at ~98% progress during hours or days apparently doing nothing. Just be patient! There is a limitation in file size which can be processed by OMSSA. If this limit is reached, SearchGUI will propose to split the spectrum file. The splitting preferences can be modified in the additional settings. Also, bear in mind that the larger these files, the more challenging their post-processing. As a result, standard desktop computers are often simply unable to process large datasets.

**1.4 Identification Results**

**[1.4a]** After a standard search, X!Tandem performs a so-called second pass search where it automatically looks for extra peptides carrying these modifications. SearchGUI hence passed this information to PeptideShaker. This second pass search has the advantage to bring new identifications, however, note that it biases the way we estimate our error rates.[14](#_ENREF_2) This will be the subject of the next chapter.

**[1.4b]** Modern mass spectrometers have a high sequencing rate and it is normal to see multiple measurements of the same peptide. When optimizing the mass spectrometer settings, one tries to reduce this effect in order to improve sample coverage.

The notion of peptide is however not fixed with regards to charge and modification status. In PeptideShaker, a peptide is considered as able to carry different charges but the same sequence presenting different modification statuses will be considered as two different peptide entities. More details on peptide inference will be given in the “PTM Analysis” chapter.

**[1.4c]** Line 15: GYYSPYSVSGSGSGSTAGSR was found phosphorylated on serine 4. However, the localization of the phosphorylation is not confident: only the letter carries the color – more details on PTM localization will be given in the “PTM Analysis” chapter. Line 22: QLEMSAEAER was found oxidized on methionine 4. Line 36: ELYQQLQRGER was found phosphorylated on tyrosine 3. Peptides at lines 20, 23, 24, 25, 35 and 36 were carrying a pyro-cmc modification.

**[1.4d]** Depending on the elution and ionization conditions, the exact same peptide can end up being recorded at two different charge states. Here, the spectra were recorded at time points separated by only three seconds.

**[1.4e]** At the top left of the screen, you can see which parts of the sequence are covered in the spectrum and at which intensity. Such a full coverage is very rare and leaves little doubt on the quality of the identification. In the middle, an histogram shows the distribution of the peak intensities – in green, identified peaks, in grey non identified. One clearly sees here that the most intense peaks are almost all annotated with a fragment ion which is again synonymous of quality for the identification. Finally, on the top right are displayed the fragment ion mass errors at their respective mass. One can see that all ions are very accurately identified, leaving little doubt on the peptide identification. Note that the error is increasing with the mass, as expected from the “Peptide-Spectrum Matching” chapter.

As a result, the spectrum is very nicely annotated with two series of b and y ions. These ions are the ones we used for the identification. PeptideShaker also annotates iF which is a commonly observed immonium ion for the amino-acid Phenylalanine.[15](#_ENREF_3) Some other ions presenting neutral losses are also annotated.

The ions detected are heavily dependent on the experimental workflow and the peptide species.

**[1.4f]** Peptides fragment at different places with different yields. As a result, some fragment ions are usually missing. The experimentalist optimizes the fragmentation conditions in order to get the best sequence coverage – but a full coverage is often impossible. In most cases however, a partial coverage is sufficient for confident peptide identification as only one candidate from the database would match the measured sets of fragment ions.

The ambiguous residues are amino-acids and sets of amino-acids presenting the same mass. The most famous case is the Isoleucine - Leucine couple. These can create systematic errors, hence biasing the error rate estimation[16](#_ENREF_4) and protein inference. The number of ambiguous cases obviously grows when taking into account more variable modifications.

**[1.4g]** For a trypsin digest, the C-terminus is more likely to carry a charge and hence more likely to be measured. As a result, y ions are typically more intense than b ions. The relative intensity levels are however heavily peptide, sample and experiment dependent.

**[1.4h]** With modern instruments, fragment ion intensities are extremely reproducible. These however strongly depend on the charge state of the precursor and modification status of the peptide.

**[1.4i]** The error clearly goes down for high masses as the PSM number increases. In fact, PSMs are sorted by increasing retention time: 1652 s, 1666 s, 1679 s and 1693. One observes here the fluctuation of the instrument calibration at high masses over time. This can be due to minor temperature fluctuations for instance. Note that the mass deviation stays between ±0.01 Da, safely below the ±0.02 boundaries set for the search.

**[1.4j]** It is peptide NGRVEIIANDQGNR at position 47.

**[1.4k]** 6 peptides were found oxidized (lines 3, 6, 7, 19, 20 and 34 in the peptide table) resulting in 5 oxidation sites on the protein sequence (M148, M153, M196, M263 and M541).

**[1.4l]** Search engines have complementary features, notably in terms of spectrum filtering and in-sillico fragmentations. Also, X!Tandem has an implemented second pass search bringing additional PSMs as illustrated on the Venn diagram. PeptideShaker takes advantage of these complementarities to increase the identification rate. Moreover, depending on the sample complexity, labelling or fragmentation methods, a search engine can underperform. Having different algorithms is a gage of stability. In such cases, the problem is easily spotted by the Venn diagram and a new project can be created excluding the underperforming search engine.

**[1.4m]** Here the hit proposed by X!Tandem is clearly better than the one found by OMSSA, as seen from the respective confidences and spectrum annotation. In fact, X!Tandem found this acetylated peptide during the second pass search – while OMSSA was not searching for acetylated peptides. OMSSA and X!Tandem are hence not looking at the spectrum with the same glasses, explaining the dramatic difference between the results.

On the other hand, when the search engines come up with different solutions with comparable confidence, the match can reasonably be considered as doubtful. This is notably the case when search engines infer conflicting PTM localizations – an effect which is translated in a score in PeptideShaker, the D-score.[17](#_ENREF_5)

**[1.4n]** Generally in proteomics, in order to avoid so-called one hit wonders, one requires two different peptides per protein. This is illustrated by the fact that our estimated number of validated false protein identification matches is solely found in the one peptide category. More details on the false and true positives will be given in the “Peptide and Protein Validation” chapter.

However, this does not imply that all single peptide hit proteins shall be discarded. They should be considered with care.

**[1.4o]** From the description, one can expect these proteins to be very similar, hence having high sequence similarity and being very difficult to distinguish by peptide centric mass spectrometry based proteomics.

**[1.4p]** Here is how: Case 1: A and B are identified, the group AB is deleted. Case2: A is identified and A or B is identified, the group AB remains. Case 3: A or B is identified, the group AB remains.

In all cases, the peptides of the group AB are also attached to A and B, hence visible in the table flagged with a different PI status than the unique peptides. The shared peptides are however not used for scoring purposes.

**[1.4q]** This sorting is a very imprecise sorting. It tends to be very conservative and flag more problematic cases as there actually are.

**[1.4r]** The unique peptide, LSVEGFAV, is flagged in green in the PI column of the Peptides table. Note that it presents a very low score and almost no annotated peaks in the spectrum supporting its identification. This group is thus clearly not reliable.

**[1.4s]** It is necessary to keep all groups for scoring reasons. This will be further detailed in the “Peptide and Protein Validation” chapter.

**[1.4t]** The protein inference problem is inherent to peptide-centric proteomics and can hence not be avoided. However, two factors dramatically reduce the prominence of that problem: (A) the improved identification of unique peptides which follows technical improvements and (B) the curation of databases: most of the secondary matches displayed in this tutorial are very unlikely to be identified when compared to the main match. Using a clean database hence dramatically simplifies the interpretation of the results.

When protein inference issues are actually impairing the scientific outcome of an experiment, it is possible to enrich for unique peptides like terminal peptides[18](#_ENREF_8) or to decipher the problem using targeted proteomics.[19](#_ENREF_9)

**1.5 Peptide and Protein Validation**

**[1.5a]** In order to maximize our proteome coverage, we will try to maximize the number of true positives while controlling our error rate: the share of false positives.

**[1.5b]** The decoy hits only indicate the propensity for the search engine to introduce random matches at a given score. In no way they indicate which target hit is the wrong one.

It is also possible to create decoy databases by randomizing amino acids. This is particularly easy with dbtoolkit.[20](#_ENREF_3) Both reverse and random decoy sequences were shown to perform equally well.16, 21 The random approaches present the advantage to allow the creation of different versions.

**[1.5c]** We expect a maximum of 12 false positives: 1% of 1214.

**[1.5d]** This value was the best below 1%. Including more proteins would have in all cases implied FDR > 1%. PeptideShaker hence stopped at 0.75% this is called a q-value.[22](#_ENREF_6)

**[1.5e]** As one can see on the right of the plot, the confidence can fluctuate at a given score. This shows that our estimation is not an exact estimation. In fact, PeptideShaker tells you that it estimates its resolution to 0.59 percentage points (pp). One can hence expect our confidence estimation to be percentage point accurate.

Including hundred hits at 95% confidence, we expect 95 true positives, hence 5 false positives. The complement of the confidence is named Posterior Error Probability (PEP): .[22](#_ENREF_6)

**[1.5f]** The new estimated FDR value is 11.48%, corresponding to an estimated FNR of 1.12%. We have hence included 150 false positives to rescue 38 true positives. The interest of this quantity-driven threshold is obviously disputable. However, there is no perfect threshold, it is up to the scientist to draw the line based on his experiment.

**[1.5g]** At 1% FDR, the lowest confidence retained is 63% estimated at an accuracy of approximately 1.5 percentage points. When thresholding at a minimal confidence of 95%, we obtain an estimated FDR of 0.06%.

**[1.5i]** For the proteins, the blue line clearly deviates from the black line. This is simply due to the fact that there are fewer proteins than spectra: the statistical estimation is hence less accurate. This deviation is directly linked to the deviation of the operating point of the ROC curve.

**2.0 Quantification**

(answer pending...)

**3.0 Functional Analysis**

**[3.0a]** According to the protein attributes, this protein “Probably plays a role in facilitating the assembly of multimeric protein complexes inside the ER” and was found in these subcellular locations: “Endoplasmic reticulum lumen. Melanosome. Cytoplasm.”. Note that more information is given in the “Ontologies” section of the protein report.

**[3.0b]** A table lists all known possible partners inferred in this case from experiment, databases and text mining. Note that these interaction inference methods are not of the same trustfulness.

**[3.0c]** It is very rare to cover a pathway fully, and most often impossible. Indeed pathways also contain molecules like ADP which are not detected in proteomics experiments. Moreover, it can happen that an isoform of a given protein is expected whether we identify another. Here again, the protein inference problem is impairing our ability to map our data to external resources.

**[3.0d]** (answer pending...)

**[3.0e]** (answer pending...)

**[3.0f]** There are different structures inferred by different methods. Also, the mapping between the structure database and the sequence is not always perfect. Often, there is simply no structure available.

**[3.0g]** Here again, the sequence database and structure database do not fully overlap.

**4.0 Online Repositories**

**4.1 Submitting to Online Repositories**

**[4.1a]** If you happen to have a complex project combining different PeptideShaker projects, you will have more complex mappings. It is important to clearly document which files are related to each others.

**4.2 Browsing Online Repositories**

**[4.2a]** You can see detailed information about the project, notably, the publications it is attached to, contacts of the authors, type of sample, protocol used and statistics about the spectra and their identification. You see here how crucial it is to annotate your data in a meaningful way in order to make it comprehensible for others when viewing.

**[4.2b]** One of the main differences with PeptideShaker compared to PRIDE Inspector, is that PRIDE Inspector does not support the protein groups inferred during protein inference. Also, the interface does not display the result of the validation process. PeptideShaker still added all the available information as additional parameters for the matches which you can access at the end of every line. Note also that the m/z differences you see in the tables do not correspond to the ones used by the search engines so do not panic! (The issue is currently being looked into by the PRIDE team.)

**4.3 Reprocessing Public Experiments**

**[4.3a]** This dataset was part of a publication[23](#_ENREF_2) from 2005 as displayed in the References section in the lower right corner of the dialog. Reanalyzing it will hence give us an impression of what has changed in the field of proteomics since then.

One of the major differences comes from the instrumentation: 3565 MS/MS spectra were generated and searched with a tolerance of 0.3 Da. In comparison, the example dataset of the tutorial counts 11,332 MS/MS spectra (measured over a longer gradient however) searched with a tolerance of 10 ppm/0.01 Da. Since 2005, the resolution of the instrument was hence multiplied by more than 10 without decreasing the scan time.

Secondly, the sequence database used was the International Protein Index (IPI) which was discontinued and is now rather included in UniProt.[24](#_ENREF_3) You will also notice that the original data interpretation pipeline is complex and requires good computational skills. Especially, there was no user friendly interface allowing the intuitive browsing of proteins, peptides and spectra. Finally, note that there is no estimation of the error rate.

Finally, you will observe that this project has the same number of spectra as peptides. In fact, only the identified spectra were uploaded then. It is now required to provide all the raw data for publication – this will be further discussed in the tutorial.[25](#_ENREF_4)

In the end, the only thing that did not change is the search engine. It is actually quite an issue of the field, we are looking at high resolution data with tools designed on low resolution. Hence, new algorithms specifically designed for high resolution mass spectrometers are being developed and will be included in the present tutorial as soon as technically possible.

**[4.3b]** If you select the 'Spectrum IDs' tab, you will see that 1805 spectra out of 3565 spectra (50.6%) were identified at 1% FDR.

References

1. Yates, J.R., 3rd, Morgan, S.F., Gatlin, C.L., Griffin, P.R. & Eng, J.K. Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. *Analytical chemistry* **70**, 3557-3565 (1998).

2. Lam, H. Building and searching tandem mass spectral libraries for peptide identification. *Molecular & cellular proteomics : MCP* **10**, R111 008565 (2011).

3. Bartels, C. Fast algorithm for peptide sequencing by mass spectroscopy. *Biological Mass Spectrometry* **19**, 363-368 (1990).

4. Seidler, J., Zinn, N., Boehm, M.E. & Lehmann, W.D. De novo sequencing of peptides by MS/MS. *Proteomics* **10**, 634-649 (2010).

5. Lange, E., Gropl, C., Reinert, K., Kohlbacher, O. & Hildebrandt, A. High-accuracy peak picking of proteomics data using wavelet techniques. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 243-254 (2006).

6. Eyrich, B., Sickmann, A. & Zahedi, R.P. Catch me if you can: mass spectrometry-based phosphoproteomics and quantification strategies. *Proteomics* **11**, 554-570 (2011).

7. Vaudel, M., Burkhart, J.M., Sickmann, A., Martens, L. & Zahedi, R.P. Peptide identification quality control. *Proteomics* **11**, 2105-2114 (2011).

8. Burkhart, J.M., Vaudel, M., Zahedi, R.P., Martens, L. & Sickmann, A. iTRAQ protein quantification: a quality-controlled workflow. *Proteomics* **11**, 1125-1134 (2011).

9. Burkhart, J.M., Schumbrutzki, C., Wortelkamp, S., Sickmann, A. & Zahedi, R.P. Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics. *Journal of proteomics* **75**, 1454-1462 (2012).

10. Fannes, T. et al. Predicting tryptic cleavage from proteomics data using decision tree ensembles. *Journal of proteome research* **12**, 2253-2259 (2013).

11. Olsen, J.V. et al. Higher-energy C-trap dissociation for peptide modification analysis. *Nature methods* **4**, 709-712 (2007).

12. Ross, P.L. et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & cellular proteomics : MCP* **3**, 1154-1169 (2004).

13. Thompson, A. et al. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical chemistry* **75**, 1895-1904 (2003).

14. Everett, L.J., Bierl, C. & Master, S.R. Unbiased statistical analysis for multi-stage proteomic search strategies. *Journal of proteome research* **9**, 700-707 (2010).

15. Papayannopoulos, I.A. The interpretation of collision-induced dissociation tandem mass spectra of peptides. *Mass Spectrometry Reviews* **14**, 49-73 (1995).

16. Colaert, N., Degroeve, S., Helsens, K. & Martens, L. Analysis of the resolution limitations of peptide identification algorithms. *Journal of proteome research* **10**, 5555-5561 (2011).

17. Vaudel, M. et al. D-score: a search engine independent MD-score. *Proteomics* **13**, 1036-1041 (2013).

18. Gevaert, K. et al. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nature biotechnology* **21**, 566-569 (2003).

19. Picotti, P. & Aebersold, R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nature methods* **9**, 555-566 (2012).

20. Martens, L., Vandekerckhove, J. & Gevaert, K. DBToolkit: processing protein databases for peptide-centric proteomics. *Bioinformatics* **21**, 3584-3585 (2005).

21. Wang, G., Wu, W.W., Zhang, Z., Masilamani, S. & Shen, R.F. Decoy methods for assessing false positives and false discovery rates in shotgun proteomics. *Analytical chemistry* **81**, 146-159 (2009).

22. Kall, L., Storey, J.D. & Noble, W.S. QVALITY: non-parametric estimation of q-values and posterior error probabilities. *Bioinformatics* **25**, 964-966 (2009).

23. Martens, L. et al. The human platelet proteome mapped by peptide-centric proteomics: a functional protein profile. *Proteomics* **5**, 3193-3204 (2005).

24. Griss, J. et al. Consequences of the discontinuation of the International Protein Index (IPI) database and its substitution by the UniProtKB "complete proteome" sets. *Proteomics* **11**, 4434-4438 (2011).

25. Martens, L. et al. Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories. *Proteomics* **5**, 3501-3505 (2005).