Answers [April 17th 2015]

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*. See for example DeNovoGUI5 (<http://denovogui.googlecode.com>) for a graphical user interface for *de novo* sequencing.

**[1.1b]** UniProt provides a grand total of 145 672 protein entries for human (as of April 2015). These sequences are inferred from the sequenced genome and curated algorithmically or manually. Interestingly, the entries labeled with a star (20 204) are manually reviewed, these proteins are historically called SwissProt entries. The entries without a star on the other hand (125 468) are algorithmic prediction where no experimental validation is annotated in UniProt, and this part of UniProt is called TrEMBL.

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 part of the data against bigger databases or 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 version or date of the database and report it in the publications.

**[1.1c]** Most proteins exist in various forms, with minor or major variations. UniProt therefore has to decide on a common protein sequence to represent all these variations, referred to as the canonical sequence. All the annotations are then described in relation to this common sequence.

UniProt also provides isoforms of the protein sequences. To include these in your database choose the "Canonical & isoform" option. However, this should be used with caution as including the isoforms dramatically reduces the efficiency of the identification algorithms.

**[1.1d]** As we are searching against human data the human keratin proteins are already in our protein sequence database, hence adding them would just duplicate the keratin entries.

**1.2 Peak List Generation**

**[1.2a]** Profile mode data is a continuous line of data points, while in centroid mode the data is converted into single data points, i.e., a peak list.

**[1.2b]** When programming the mass spectrometer, it will be indicated whether the spectra are recorded in profile mode (requiring peak picking) or in centroid mode (already peak picked). Depending on the instrument, using more advanced signal processing methods can improve the results.[6](#_ENREF_5) 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]** X!Tandem, MyriMatch, MS Amanda, MS-GF+, Comet, Tide and OMSSA, and indeed (almost) all of the tools used in this tutorial, are freely available tools, referred to as freeware. This means that anyone can download and use such tools without having to pay a license fee. In many cases such programs are also open source, meaning that also the source code is freely available. Freeware is not completely without a license though, and before using such tools as part of your own commercial software, be sure to check if the license allows for this. In the academic world this is rarely a problem though, and as long as you do not go around selling it as your own software you should be ok. However, it is always recommended to refer to the software you are using, for example when publishing results where freely available software was used to arrive at the results.

**[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 and some are experimental artefacts like oxidation of methionine. 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[7](#_ENREF_6) – we actually here selected phosphorylation for illustrative purposes only.

Selecting variable modifications has a similar effect as 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 possible to control the level of modifications by doing a pre-search with the modification of interest as variable.[8](#_ENREF_7) Here, searching with oxidation of methionine and carbamidomethylation of cysteine as variable modification returned >98% of cysteine residues modified. The modification can thus reasonably be considered as fixed. Note that such quality control steps are for crucial importance when working with chemically labelled samples.[9](#_ENREF_8)

**[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,[10](#_ENREF_9) some missed cleavages will always remain,[11](#_ENREF_10) in our experience up to two with trypsin.[8](#_ENREF_7)

**[1.3e]** With the first low resolution mass spectrometers, searches were conducted with a fixed tolerance in mass – using the unit Dalton. With the advent of high resolution mass spectrometry, search engines adapted the tolerance 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[12](#_ENREF_11) which principally generates b and y ions.

**[1.3g]** The phosphorylation options are all different, either targeting different amino acids or with different losses. The listed 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 the search engines.

**[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.13-14

**[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 can 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 Advanced 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 very large datasets.

**1.4 Identification Results**

**[1.4a]** Ensembl provides a centralized resource for information about genomes, both for human and other vertebrates and model organisms. For more details see: <www.ensembl.org>.

**[1.4b]** Additionally to the given search settings, X!Tandem performs by default a quick search for protein terminal acetylation and peptide pyrolidone. After a standard search, X!Tandem also performs a so-called second pass search where it automatically looks for extra peptides carrying terminal acetylation. SearchGUI passed this information to PeptideShaker, that is why the search parameters changed.

Note that the second pass search has the advantage to bring new identifications, however, it biases the way we estimate our error rates.[15](#_ENREF_14) This will be the subject of the next chapter.

**[1.4c]** 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 upcoming “PTM Analysis” chapter.

**[1.4d]** 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 upcoming “PTM Analysis” chapter.) Line 22: QLEMSAEAER was found oxidized on methionine 4. Line 37: ELYQQLQRGER was found phosphorylated on tyrosine 3. Line 38: TSCIAGVFVDATK was found phosphorylated on tyrosine 1. (Again, the localization of the phosphorylation is not confident). Peptides at lines 20, 23-25, 27 and 37 were found carrying a pyro-cmc modification. (Terminal PTMs are not color coded.)

**[1.4e]** 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.

**[1.4f]** 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.[16](#_ENREF_15) Some other ions presenting neutral losses are also annotated. The common ions and their relation to the peptide sequence are shown in the figure below.

x3 y3 z3 x2 y2 z2 x1 y1 z1

a1 b1 c1 a2 b2 c2 a3 b3 c3

H

H

H

H

H

H

H

R1

R4

H2N – C – C – N – C – C – N – C – C – N – C – COOH

–

–

–

–

–

–

–

–

–

–

R2

R3

O

O

O

–

The ions detected are heavily dependent on the experimental workflow and the peptide species. For more details about the fragment ions see: www.matrixscience.com/help/fragmentation\_help.html.

**[1.4g]** 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[17](#_ENREF_16) and protein inference. The number of ambiguous cases obviously grows when taking into account more variable modifications.

An immonium ion is an internal fragment with just a single side chain formed by a combination of *a* type and *y* type cleavage. These ions are usually labelled with the one letter code for the corresponding amino acid, and sometimes with a lower case 'i' in front, e.g., iF. The presence of an immonium ion can be used as an indicator that the given amino acid should be in the identified peptide sequence. The opposite is however not true. In our example we find the immonium ion of Phenylalanine (iF), and we have a Phenylalanine (F) in our sequence, so there is a correspondence between the two.

**[1.4h]** 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 dependent on the peptide, sample and experiment.

**[1.4i]** The table displays exactly the same intensities as the ones used for *de novo*.

**[1.4j]** 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.

**[1.4k]** 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 s. 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.

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

**[1.4m]** 6 peptides were found oxidized (lines 3, 6, 7, 19, 20 and 35 in the peptide table) resulting in 5 oxidation sites on the protein sequence (M148, M153, M196, M263 and M541).

**[1.4n]** Search engines have complementary features, notably in terms of spectrum filtering and in-sillico fragmentations. Also, X!Tandem includes a so-called second pass search bringing additional PSMs as illustrated in the Venn diagram. Multiple algorithms workflows take advantage of these complementarities to increase the identification rate. Moreover, depending on the sample complexity, labeling 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.

**[1.4o]** 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.[18](#_ENREF_17)

**[1.4p]** Generally in proteomics, in order to avoid false positives and especially so-called one hit wonders, one requires at least two different peptides per protein. This is illustrated by the fact that our estimated number of doubtful protein identification matches is mainly found in the low peptide count categories. 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, simply that they should be considered with care. As you can observe, PeptideShaker already flagged these matches as “doubtful” for you.

**[1.4q]** The evidence column represents the level of evidence that exists for this protein, from evidence at the protein level (the strongest evidence), to simply uncertain. The evidence level can be used as a rough guide to pick the most likely protein in a group, e.g., if one protein has evidence at the protein level, while the other only at the transcript level, it would in most case be safe to assume that the first protein is the one found in the sample. For more details see: <http://www.uniprot.org/manual/protein_existence>.

The Enz column shows if enzymatic peptides, i.e., peptides consistent with the enzyme's cleavage rules, have been found. If a protein is only detected with non-enzymatic peptides we might trust it less compared to a protein identified with enzymatic (and possibly non-enzymatic) peptides.

**[1.4r]** 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.

**[1.4s]** Such a sorting tends to be on the conservative side and flag more problematic cases than there actually are. In many cases all proteins in a group are related isoforms, and it is then up to the properties of the experiment to decide if distinguishing between such proteins is important or not.

**[1.4t]** The unique peptide, NH2-VPNSNPPEYEFFWGLR-COOH, is flagged in green in the PI column in the Peptide table. Note that it presents a low score and few annotated peaks in the spectrum supporting its identification. This group is thus clearly not reliable.

**[1.4u]** It is necessary to keep all groups for the evaluation of our error rates. This will be further detailed in the “Peptide and Protein Validation” chapter.

**[1.4v]** There are three dotted lines, hence there are three non-enzymatic peptides detected for this protein.

**[1.4w]** There are three red peptides and no yellow, hence there are three non-validated peptides and no doubtful peptides.

**[1.4x]** Endoplasmin has 20 peptides that can only come from Endoplasmin (given the database used). Hence we are pretty confident that Endoplasmin has been detected in our sample. Three of the five additional proteins are only annotated as "Uncertain" in UniProt. Hence we can most likely ignore these. That leaves "P08238 - Heat shock protein HSP 90-beta (HS90B\_HUMAN)" and "Q58FF8 - Putative heat shock protein HSP 90-beta 2 (H90B2\_HUMAN)", which both have evidence at the protein level in UniProt. In order to conclude regarding where these shared peptides come from could rely on additional information, but in many cases it will be impossible to completely resolve such cases.

**[1.4y]** 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[19](#_ENREF_18) or to decipher the problem using targeted proteomics.[20](#_ENREF_19)

**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.[21](#_ENREF_20) Both reverse and random decoy sequences were shown to perform equally well.17, 22 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 1205.

**[1.5d]** The FDR limit used is the best below 1%. In some cases including more proteins will result in an implied FDR > 1%. PeptideShaker therefore sometimes stops below the actual limit, this is called a q-value.[23](#_ENREF_22)

**[1.5e]** As one can see in 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 minimal resolution to 6.67 percentage points (pp). One can hence expect our confidence estimation to be less than 6.67 percentage points accurate.

Including hundred hits at 95% confidence, we expect 95 true positives, 5 false positives. The complement of the confidence is named Posterior Error Probability (PEP): .[23](#_ENREF_22)

**[1.5f]** The new estimated FDR value at 1% FNR threshold is 0.84%, corresponding to an estimated FNR of 1.0%. We have hence removed 2 false positives but lost 13 true positives. Selecting a minimal confidence of 95% brings an estimated FDR of 1.065% and an FNR of 0.0%. Note that setting a threshold of 95% confidence with 5 percentage points estimation accuracy means that you are actually using a confidence between 90 and 100%.

Choosing between quality or quantity is always disputable, there is no perfect threshold; it is up to the scientist to draw the line based on his experiment. Keep in mind that statistical estimators are not perfect: it is always important to verify their accuracy and adapt the validation settings correspondingly. To that regard, FDR and FNR thresholds are usually more robust than confidence thresholds.

**[1.5g]** The main difference of PSM and peptide level statistics compared to protein level statistics is that there are more PSMs and peptides than proteins. The error rate estimation is thus more accurate, the minimal resolution is between 1 and 3 percent points for this dataset.

**[1.5h]** As we saw already, the statistical validation has its own resolution. Moreover, the target/decoy approach is only sensitive to random hits and does not account for minor sequence variations.[17](#_ENREF_16) Consequently, only high quality hits are hence kept for further investigation. For example, one typically requires two confident peptides per protein.

Note that whenever the database or the number of identified proteins does not allow confident validation, the hits will also be marked as doubtful.

**[1.5i]** For the proteins, the blue line deviates a lot more from the black line than for PSMs. 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 Functional Analysis**

**[2.0a]** According to information in UniProt, this protein “probably plays a role in facilitating the assembly of multimeric protein complexes inside the endoplasmic reticulum” and "is involved in the correct folding of proteins and degradation of misfolded proteins". It was found in these subcellular locations: “Endoplasmic reticulum lumen, Melanosome and Cytoplasm.”

**[2.0b]** A table lists all known possible partners inferred in this case from experiment, databases and text mining. Note that these interaction inference methods do not all have the same trustfulness.

**[2.0c]** 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 when we identify another. Here again, the protein inference problem is impairing our ability to map our data to external resources.

**[2.0d]** In the Significance plot all the green bars represent GO terms that are significantly more frequent in the example dataset compare to Ensembl, while red bars represent GO terms that are significantly less frequent in the example dataset compare to Ensembl.

**[2.0e]** The Hypergeometric test used to compare the two groups rely on the fact that the selection of protein to compare against the distribution in Ensembl is randomly selected. If this is not the case, for example if only selecting a subset of the proteins with certain properties, the bases for the statistical test is no longer correct and the results of the test cannot be trusted. For more information about the test used see: <http://en.wikipedia.org/wiki/Hypergeometric_test>.

**[2.0f]** 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.

**[2.0g]** Here again, the sequence database and structure database do not fully overlap.

**3.0 Online Repositories**

**3.1 Submitting to Online Repositories**

**[3.1a]** Annotating your dataset will help other scientists to find and understand your data. For example, if you annotate a HeLa sample with the HeLa term, everyone searching “HeLa” will be able to find your dataset. A good annotation is also vital for dataset understanding and reprocessing: it will allow other scientists to know which modifications you have been looking for, which mass tolerances and which database. Note that PeptideShaker standardizes and annotates all the processing parameters for you!

**[3.1b]** 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.

**[3.1c]** This dataset is only an example and does not tackle a specific biological question. In real life experiments however, you might want to annotate the problematic of the project using disease keywords. If you proceeded to quantification you can select the method used. In case you are combining different PeptideShaker projects, you can annotate the different files in “Experimental Factor” using the same annotation as in PeptideShaker.

**[3.1d]** A public dataset is freely available for everyone and advertised on the ProteomeXchange webpage. By default, datasets are kept private. The PRIDE team will provide you with personal access codes which you should put in your manuscript during the review process. Once your manuscript is accepted, send its details to the PRIDE team who will attach the reference to the manuscript and make the data available. Finally, you can remove the review accession details from your manuscript during galley proof correction.

**3.2 Browsing Online Repositories**

**[3.2a]** You can see detailed information about the project, notably, the publications it is attached to (if any), 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.

**[3.2b]** One of the main differences with PeptideShaker compared to PRIDE Inspector, is that PRIDE Inspector does not display the result of the validation process.

**3.3 Reprocessing Public Experiments**

**[3.3a]** This dataset was part of a publication[24](#_ENREF_23) from 2005 as displayed in the Publication section as displayed by clicking the project accession link. 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.[25](#_ENREF_24) 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 data sharing chapters.[26](#_ENREF_25)

**[3.3b]** If you select the Spectrum IDs tab, you will see that 1882 spectra out of 3565 spectra (53%) were identified at 1% FDR. 432 protein groups were validated at 1% FDR, among them 118 were marked as confident.

**4.0 Quantification**

**4.1 Spectrum Counting**

**[4.1a]** The main advantage of this technique is obviously its simplicity: for a given protein, one simply needs to count spectra – no need for advanced experiments, signal processing or data mining. As a result, abundance indexes are straightforwardly available from any measurement in virtually no processing time.

This however comes at costs in quantification accuracy. Indeed, in terms of quantification performance, an abundance metric based on spectrum counts can simply not compare to intensity based metrics. For example, a protein where only one spectrum is validated will hence be seen two times down-regulated when compared to a similar one where two spectra are validated – although we saw in the previous chapters how subjective spectrum matches validation is.

Spectrum counting quantification accuracy is hence dramatically sensitive with regards to the identification process. Other factors can impair accuracy: different proteins do not have the same probability to generate observable tryptic peptides; these do not have the same probability to generate a validated identification. Also, by design, the protein inference problem tackled in the identification chapter dramatically biases the index. Finally, modern mass spectrometers are actually tuned toward the reduction of redundant peptide acquisition, hence biasing the base assumption of spectrum counting indexes.

**[4.1b]** The results of both indexes correlate well with an R2 value of 0.7665 which is not bad compared to the expected reliability of the method. Note that we cover four orders of magnitude, in order to fairly compare protein abundances on such a large range, we use the logarithm of the value.

**[4.1c]** This example consists of a single Q Exactive run (1 hour gradient) which does not allow full proteome coverage. In fact, only the most abundant proteins of the list were identified. We can clearly estimate our detection limit to 3E5 copies per cell. Note also that the probability of detecting a protein does not only depend on its abundance but also its chemical properties. Complementary experimental methods might help increasing proteome coverage

**[4.1d]** First, the different studies present different detection limit. The intensity based study identified very low abundant proteins (few thousands per cell) covering more than 10,000 proteins. This is mainly due to the experimental setup, much more complex than in our tutorial examples.

In terms of quantification however, the variability of the intensity based quantification R2=0.6155 is very similar to the one obtained with spectrum counting, respectively 0.6091 and 0.5602 for a one and four hours gradient.

**[4.1e]** Spectrum counting quantification presents the advantage of being extremely simple and fast to compute. Its main shortcoming is the limited accuracy and lack of robustness as pointed out by the dependency on the identification procedure. Hence, it typically serves as a fast and rough abundance estimator. For more accurate results, it is preferable to set-up SRM experiments.

**4.2 Reporter Ions**

**[4.2a]** If the three spiked in proteins are not added to the default human database they will not be in the list of possible proteins the spectra can be matched against, and hence cannot be identified.

**[4.2b]** The four iTRAQ labels are all isobaric, meaning that they have the mass and thus appear as identical in the MS1 spectrum. When trying possible amino acid modifications for the peptide to spectrum matches it is therefore enough to only include one of the iTRAQ modifications, as a match against one of them will also match all the others. And at the MS1 level, i.e., when finding the mass of the precursor, this is all we need. (Note that the iTRAQ labels, unlike TMT, are not truly isobaric though, as the chemical modifications used to generate the different labels differ slightly. This is not usually an issue, but with the ever increasing accuracy of the instruments, there will come a time when iTRAQ labels can no longer be considered as isobaric.)

The iTRAQ labelling is considered as variable on the Y, because experiments have shown that it modifies the Y's in roughly 50% of the cases. While the iTRAQ modification on the K and n-term occurs in close to 100% of the cases and is thus considered as fixed.

**[4.2c]** All the spectra matching to a given peptide should have similar iTRAQ peak intensities. They do after all come from the same peptide, and assuming that this peptide is unique to a given protein, the intensities should reflect the protein amounts in the four labelled samples. It follows from this that different peptides from the same protein should all have similar iTRAQ peak intensities. However, there will be slight differences between peptides and because of this it is therefore important to have data from more than a single peptide when using iTRAQ for quantification.

The picture is very much complicated by the addition of shared peptides. A peptide that cannot be uniquely linked to a single protein, but rather maps to two or more proteins, will often end up having a deviating peak intensity relative to the unique peptides. The reason for this is that the amounts of the proteins the peptide maps to can differ. Let's say that we have the two proteins A and B, where A as a low abundance and B a high abundance. All peptide unique to protein A will thus have a low abundance and all proteins unique to protein B will have a high abundance. However, a peptide shared between the two proteins will in some cases have a low abundance, i.e., when it comes from protein A, and in some cases a high abundance, i.e., when it comes from protein B. The average abundance of the peptide will therefore be somewhere in between the low and high abundance, and including such peptides in the quantification must therefore be done with much care.

**[4.2d]** The isotope correction will compensate for impurities of the reporter ion reagents. The manufacturer provides you a quality control report along with the kits containing correction factors which can be implemented here in Reporter. Reporter will make this correction for you.

**[4.2e]** Similarly as for the identification, it is dangerous to rely on a single quantification event – a so-called one hit wonder. Quantification artefacts can easily appear on the peptide or PSM level. You should hence be very careful with proteins presenting low number of quantified peptides and PSMs.

**[4.2f]** The three spiked in proteins are: Hexokinase-1 (HXKA\_YEAST), Potassium-activated aldehyde dehydrogenase, mitochondrial (ALDH4\_YEAST) and Beta-galactosidase (H5Q9R5\_ECOLX).

**[4.2g]** The three spiked in proteins are the same as for [2.2f]. However, they are more difficult to detected, and the reason is the variation in the background. Each sample has a different background due to individual differences between the patients the samples come from. The data is thus a lot more noisy and it is much harder to separate the spiked in proteins from the rest of the data.

**4.3 Label Free**

(answers pending...)

**4.4 MS1 Labeling**

(answers pending...)

**4.5 Targeted Quantification**

(answers pending...)

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. Muth, T. et al. DeNovoGUI: an open source graphical user interface for de novo sequencing of tandem mass spectra. *J Proteome Res* **13**, 1143-1146 (2014).

6. 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).

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

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

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

10. 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).

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

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

13. 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).

14. 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).

15. 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).

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

17. 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).

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

19. 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).

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

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

22. 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).

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

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

25. 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).

26. 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).