Peptide and Protein Identification

The process of searching mass spectral data for the purpose of peptide and protein identification can roughly be divided into six steps:

* **Step 1: Convert the raw, typically binary, output from the MS instrument into open formats.**
* **Step 2: Process the MS/MS spectra into peak lists.**
* **Step 3: Download the desired sequence database and adapt it to your identification strategy.**
* **Step 4: Search the peak lists against a sequence database using one or more search engines.**
* **Step 5: Identify the peptides and infer the proteins.**
* **Step 6: Validate the detected peptides and proteins.**

**(1) Convert   
Raw Files**

**(3) Generate Database**

**(2) Process MS/MS Spectra**

**(4) Match Peptides   
to Spectra**

**(5) Infer Peptides   
and Proteins**

**(6) Validate Peptides and Proteins**

In the past years tremendous efforts were made at connecting the various resources of the “omic” fields. Once the protein workflow is set up, it is thus possible to enrich your results with biological information. Then, your data begins to make a lot more sense!

**Proteomics Results**

We will introduce various external resources and show how to link them with the identification results. Note however that these cross field workflows are very young and the connection between the various components is sometimes not as straightforward as one would expect.

In order to successfully conduct these investigations, we recommend the use of the following tools:

1. To convert raw files we recommend **MSConvert** as part of the **Proteowizard[1](#_ENREF_1" \o "Kessner, 2008 #14)** package (<http://proteowizard.sourceforge.net>).
2. To process the MS/MS spectra, we recommend **OpenMS[2](#_ENREF_2" \o "Bertsch, 2011 #15)** (<www.openms.de>).
3. We recommend **UniProt[3](#_ENREF_3" \o "Apweiler, 2004 #45)** (<www.uniprot.org>) databases and for their processing **dbtoolkit[4](#_ENREF_4" \o "Martens, 2005 #19)** (<http://dbtoolkit.googlecode.com>).
4. To match peptides to spectra, we will use here two distinct, freely available search engines: **OMSSA**[**5**](#_ENREF_5) and **X!Tandem[6](#_ENREF_6" \o "Craig, 2004 #46)**, both of which are made easily accessible *via* a free tool called **SearchGUI7** ([http://searchgui.googlecode.com](http://dbtoolkit.googlecode.com)).
5. To visualize the search results, and to do the peptide and protein inference, we recommend the use of **PeptideShaker** (<http://peptide-shaker.googlecode.com>).
6. For the validation of the identifications we recommend the use of **PeptideShaker** (<http://peptide-shaker.googlecode.com>) and **Peptizer[8](#_ENREF_8" \o "Helsens, 2008 #47)** (<http://peptizer.googlecode.com>).
7. Many external resources are available on the Internet. Among them we will use:   
   UniProt (<http://www.uniprot.org>), Reactome (<http://www.reactome.org>),   
   PICR (<http://www.ebi.ac.uk/Tools/picr>) and Dasty (<http://www.ebi.ac.uk/dasty>).   
   Note that additional resources are listed in **PeptideShaker**, and will also be used to conduct the gene ontology analysis of the data.
8. In order to make your data publicly available, you can upload them in public repositories. We recommend PRIDE (<http://www.ebi.ac.uk/pride>).

This tutorial will guide you through these steps, separated into eight chapters:

1. **Database Generation**
2. **Peak List Generation**
3. **Peptide to Spectrum Matching**
4. **Browsing Identification Results**
5. **Peptide and Protein Validation**
6. **PTM Analysis**
7. **Other Resources**
8. **Submission to PRIDE**

Although it is recommended to follow the tutorial in its entirety, the chapters can be followed independently. The corresponding sections can be found along with the required example files in the respective folders.

References

(1) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **2008**, *24*, 2534.

(2) Bertsch, A.; Gropl, C.; Reinert, K.; Kohlbacher, O. OpenMS and TOPP: open source software for LC-MS data analysis. *Methods Mol Biol* **2011**, *696*, 353.

(3) Apweiler, R.; Bairoch, A.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L. S. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **2004**, *32*, D115.

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

(5) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm. *J Proteome Res* **2004**, *3*, 958.

(6) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466.

(7) Vaudel, M.; Barsnes, H.; Berven, F. S.; Sickmann, A.; Martens, L. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* **2011**, *11*, 996.

(8) Helsens, K.; Timmerman, E.; Vandekerckhove, J.; Gevaert, K.; Martens, L. Peptizer, a tool for assessing false positive peptide identifications and manually validating selected results. *Mol Cell Proteomics* **2008**, *7*, 2364.