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 to the identification results. Note however that such cross field workflows are relatively new and the connection between the various components is therefore not always as straightforward as one would expect.

There are several tools that partly or fully support such workflows. A detailed list mainly focused on free software is given in the appendix below. Among these, we recommend the use of the following tools:

1. To convert raw files we recommend **MSConvert** as part of the **Proteowizard1** package (<http://proteowizard.sourceforge.net>).
2. We recommend **UniProt2** ([www.uniprot.org](file:///C:\Users\hba041\Teaching\WT%20Course%20EBI%202012\www.uniprot.org)) databases and for their processing **dbtoolkit[3](#_ENREF_4" \o "Martens, 2005 #19)** (<http://dbtoolkit.googlecode.com>).
3. To match peptides to spectra, we will here use two distinct, freely available search engines: **OMSSA**[**4**](#_ENREF_5) and **X!Tandem[5](#_ENREF_6" \o "Craig, 2004 #46)**, both of which are made easily accessible *via* a free tool called **SearchGUI[6](#_ENREF_7" \o "Vaudel, 2011 #18)** ([http://searchgui.googlecode.com](http://dbtoolkit.googlecode.com)).
4. To analyze the search results, and to do the peptide and protein inference, we recommend the use of **PeptideShaker** (<http://peptide-shaker.googlecode.com>).
5. For the validation of the identifications we also recommend the use of **PeptideShaker** (<http://peptide-shaker.googlecode.com>).
6. Many external resources are available on the internet. Among these we will use:   
   **UniProt**[2](#_ENREF_3" \o "Apweiler, 2004 #45) (<http://www.uniprot.org>), **Reactome**[7](#_ENREF_9" \o "Haw, 2011 #139) (<http://www.reactome.org>),   
   **PICR**[8](#_ENREF_10) (<http://www.ebi.ac.uk/Tools/picr>) and **Dasty**[9](#_ENREF_11" \o "Jones, 2005 #81) (<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.
7. In order to make your data publicly available, you can upload them in public repositories. We recommend **ProteomeXchange** (<http://proteomexchange.org>) and **PRIDE**[**10**](#_ENREF_12) (<http://www.ebi.ac.uk/pride>).

This tutorial will guide you through these steps, separated into nine 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** *(in development)*
7. **Other Resources**
8. **Submission to PRIDE and ProteomeXchange**
9. **Protein Quantification - Reporter Ions**

You will find a folder named software containing all the software needed for this tutorial as well as nine folders corresponding to the nine chapters. Although it is recommended to follow the tutorial in its entirety, the chapters can be followed independently. For every chapter, the resources folder contained in the chapter folder provides all the files needed.

***All chapters are also available online:*** [***http://compomics.com/peptide\_and\_protein\_identification\_tutorial***](http://compomics.com/peptide_and_protein_identification_tutorial)

Appendix: Proteomics Software

The table below provides a (non-exhaustive) list of software dedicated to proteomics, with brief descriptions and corresponding references that will help you to get started.

|  |  |  |
| --- | --- | --- |
| Type | Software | Description |
| Converter | ProteoWizard[1](#_ENREF_1" \o "Kessner, 2008 #14) | Converter accepting most mass spectrometer proprietary formats and converting them into open formats |
| mzML parser | jmzML[11](#_ENREF_13" \o "Cote, 2010 #150) | Mass spectrometry mzML file parser |
| General proteomics package | OpenMS[12](#_ENREF_2" \o "Bertsch, 2011 #15) | Package of tools for proteomics allowing the design  of workflows with a graphical interface |
| TPP[13](#_ENREF_14) | Package of tools for proteomics mainly command line driven |
| MaxQuant[14](#_ENREF_15" \o "Cox, 2008 #103) | Package for identification and quantification  of entire proteomes |
| PeptideShaker\* | Interpretation of proteomics identifications  from multiple search engines |
| Identification post-processor | MassSieve[15](#_ENREF_16" \o "Slotta, 2010 #128) | Identification processing software |
| De novo sequencing | PepNovo[16](#_ENREF_17" \o "Frank, 2005 #97) | De novo sequencing tool |
| PEAKS[17](#_ENREF_18) | De novo sequencing tool (commercial) |
| Tag sequencing | GutenTag[18](#_ENREF_19" \o "Tabb, 2003 #191) | Finds peptide patterns in spectra |
| DirecTag[19](#_ENREF_20" \o "Tabb, 2008 #190) | Finds peptide patterns in spectra |
| Database search engine | Sequest[20](#_ENREF_21" \o "Yates, 1995 #99) | Database search engine (commercial) |
| Mascot[21](#_ENREF_22) | Database search engine (commercial) |
| OMSSA[4](#_ENREF_5) | Database search engine |
| X!Tandem[5](#_ENREF_6" \o "Craig, 2004 #46) | Database search engine |
| Inspect[22](#_ENREF_23) | Database search engine |
| MyriMatch[23](#_ENREF_24) | Database search engine |
| MassWiz[24](#_ENREF_25" \o "Yadav, 2011 #101) | Database search engine |
| Andromeda[25](#_ENREF_26) | Database search engine (MaxQuant only) |
| User friendly interfaces | SearchGUI[6](#_ENREF_7) | Graphical interface for OMSSA and X!Tandem |
| PRIDE Inspector[26](#_ENREF_27) | Graphical interface for the inspection of PRIDE XML files |
| TOPPAS[27](#_ENREF_28) | Graphical interface for the design of OpenMS workflows |
| Spectral library searching | NIST MS search[28](#_ENREF_29) | Spectral libraries search engine |
| X!Hunter[29](#_ENREF_30" \o "Craig, 2006 #106) | Spectral libraries search engine |
| SpectraST[30](#_ENREF_31" \o "Lam, 2007 #108) | Spectral libraries search engine |
| Identification file parsers | MascotDatFile[31](#_ENREF_32" \o "Helsens, 2007 #59) | Java parser for Mascot .dat files |
| OMSSA parser[32](#_ENREF_33) | Java parser for OMSSA .omx files |
| X!Tandem parser[33](#_ENREF_34) | Java parser for X!Tandem XML files |
| Data structure | compomics-utilities[34](#_ENREF_35) | Java object structure for the handling and visualization  of identifications from different search engines |
| PSM rescoring | Percolator[35](#_ENREF_36) | Machine learning algorithm rescoring  PSMs and attaching them a p-value |
| PeptideProphet[36](#_ENREF_37" \o "Keller, 2002 #27) | Machine learning algorithm attaching  PSMs a PEP (integrated in TPP) |
| PepArML[37](#_ENREF_38" \o "Edwards, 2009 #179) | Machine learning algorithm merging results from  different search engines with web interface: https://edwardslab.bmcb.georgetown.edu/pymsio |
| Database manipulation | dbtoolkit[3](#_ENREF_4" \o "Martens, 2005 #19) | Tool allowing the manipulation of databases  and creation of custom ones |
| Peptide inference | iProphet[38](#_ENREF_39" \o "Shteynberg, 2011 #126) | Tool for statistical post-processing of PSMs  (integrated in TPP) |
| Protein inference | ProteinProphet[39](#_ENREF_40" \o "Nesvizhskii, 2003 #10) | Tool for protein inference (integrated in TPP) |
| IDPicker[40](#_ENREF_41" \o "Ma, 2009 #127) | Tool for protein inference |
| MassSieve[15](#_ENREF_16" \o "Slotta, 2010 #128) | Identification processing software |
| Protein annotation | UniProtKB[2](#_ENREF_3" \o "Apweiler, 2004 #45) | Protein knowledge database |
| Dasty[9](#_ENREF_11" \o "Jones, 2005 #81) | Cross reference tool for protein databases |
| GO enrichment | GOTree[41](#_ENREF_42" \o "Zhang, 2004 #134) | GO enrichment tool |
| Onotologizer[42](#_ENREF_43" \o "Bauer, 2008 #135) | GO enrichment tool |
| DAVID[43](#_ENREF_44) | Interface for enrichment of identification results |
| 3D structures | jmol[44](#_ENREF_45" \o "Hanson, 2010 #138) | Tool for the display of 3D structures |
| Pathways | Reactome[7](#_ENREF_9" \o "Haw, 2011 #139) | Pathway investigation interface allowing the mapping  of one’s results and pathway coverage estimation |
| Interactions | STRING[45](#_ENREF_46) | Protein interaction investigation interface |
| Repository | PRIDE[10](#_ENREF_12) | Protein identification repository |
| PeptideAtlas[46](#_ENREF_47" \o "Deutsch, 2008 #142) | Peptide identification repository |
| GPMDB[47](#_ENREF_48) | Peptide and protein identification repository |
| Local data management | MASPECTRAS[48](#_ENREF_49) | LIMS system |
| Proteios[49](#_ENREF_50" \o "Hakkinen, 2009 #160) | LIMS system |
| ms\_lims[50](#_ENREF_51" \o "Helsens, 2010 #161) | LIMS system |

*\* PeptideShaker is not yet published, available at* [*http://peptide-shaker.googlecode.com*](http://peptide-shaker.googlecode.com)*.*

References

1. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P., *Bioinformatics* **24**, 2534-2536 (2008).

2. Apweiler, R. et al., *Nucleic Acids Res* **32**, D115-119 (2004).

3. Martens, L., Vandekerckhove, J. & Gevaert, K., *Bioinformatics* **21**, 3584-3585 (2005).

4. Geer, L.Y. et al., *J Proteome Res* **3**, 958-964 (2004).

5. Craig, R. & Beavis, R.C., *Bioinformatics* **20**, 1466-1467 (2004).

6. Vaudel, M., Barsnes, H., Berven, F.S., Sickmann, A. & Martens, L., *Proteomics* **11**, 996-999 (2011).

7. Haw, R., Hermjakob, H., D'Eustachio, P. & Stein, L., *Proteomics* **11**, 3598-3613 (2011).

8. Cote, R.G. et al., *BMC Bioinformatics* **8**, 401 (2007).

9. Jones, P. et al., *Bioinformatics* **21**, 3198-3199 (2005).

10. Martens, L. et al., *Proteomics* **5**, 3537-3545 (2005).

11. Cote, R.G., Reisinger, F. & Martens, L., *Proteomics* **10**, 1332-1335 (2010).

12. Bertsch, A., Gropl, C., Reinert, K. & Kohlbacher, O., *Methods Mol Biol* **696**, 353-367 (2011).

13. Deutsch, E.W. et al., *Proteomics* **10**, 1150-1159 (2010).

14. Cox, J. & Mann, M., *Nat Biotechnol* **26**, 1367-1372 (2008).

15. Slotta, D.J., McFarland, M.A. & Markey, S.P., *Proteomics* **10**, 3035-3039 (2010).

16. Frank, A. & Pevzner, P., *Anal Chem* **77**, 964-973 (2005).

17. Ma, B. et al., *Rapid Commun Mass Spectrom* **17**, 2337-2342 (2003).

18. Tabb, D.L., Saraf, A. & Yates, J.R., 3rd, *Anal Chem* **75**, 6415-6421 (2003).

19. Tabb, D.L., Ma, Z.Q., Martin, D.B., Ham, A.J. & Chambers, M.C., *J Proteome Res* **7**, 3838-3846 (2008).

20. Yates, J.R., 3rd, Eng, J.K., McCormack, A.L. & Schieltz, D., *Anal Chem* **67**, 1426-1436 (1995).

21. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S., *Electrophoresis* **20**, 3551-3567 (1999).

22. Tanner, S. et al., *Anal Chem* **77**, 4626-4639 (2005).

23. Tabb, D.L., Fernando, C.G. & Chambers, M.C., *J Proteome Res* **6**, 654-661 (2007).

24. Yadav, A.K., Kumar, D. & Dash, D., *J Proteome Res* **10**, 2154-2160 (2011).

25. Cox, J. et al., *J Proteome Res* **10**, 1794-1805 (2011).

26. Wang, R. et al., *Nat Biotechnol* **30**, 135-137 (2012).

27. Junker, J. et al., *J Proteome Res* (2012).

28. Stein, S.E. & Scott, D.R., *Journal of the American Society for Mass Spectrometry* **5**, 859-866 (1994).

29. Craig, R., Cortens, J.C., Fenyo, D. & Beavis, R.C., *J Proteome Res* **5**, 1843-1849 (2006).

30. Lam, H. et al., *Proteomics* **7**, 655-667 (2007).

31. Helsens, K., Martens, L., Vandekerckhove, J. & Gevaert, K., *Proteomics* **7**, 364-366 (2007).

32. Barsnes, H., Huber, S., Sickmann, A., Eidhammer, I. & Martens, L., *Proteomics* **9**, 3772-3774 (2009).

33. Muth, T., Vaudel, M., Barsnes, H., Martens, L. & Sickmann, A., *Proteomics* **10**, 1522-1524 (2010).

34. Barsnes, H. et al., *BMC Bioinformatics* **12**, 70 (2011).

35. Kall, L., Canterbury, J.D., Weston, J., Noble, W.S. & MacCoss, M.J., *Nat Methods* **4**, 923-925 (2007).

36. Keller, A., Nesvizhskii, A.I., Kolker, E. & Aebersold, R., *Anal Chem* **74**, 5383-5392 (2002).

37. Edwards, N., Wu, X. & Tseng, C.-W., *Clinical Proteomics* **5**, 23-36 (2009).

38. Shteynberg, D. et al., *Mol Cell Proteomics* **10**, M111 007690 (2011).

39. Nesvizhskii, A.I., Keller, A., Kolker, E. & Aebersold, R., *Anal Chem* **75**, 4646-4658 (2003).

40. Ma, Z.Q. et al., *J Proteome Res* **8**, 3872-3881 (2009).

41. Zhang, B., Schmoyer, D., Kirov, S. & Snoddy, J., *BMC Bioinformatics* **5**, 16 (2004).

42. Bauer, S., Grossmann, S., Vingron, M. & Robinson, P.N., *Bioinformatics* **24**, 1650-1651 (2008).

43. Huang da, W., Sherman, B.T. & Lempicki, R.A., *Nat Protoc* **4**, 44-57 (2009).

44. Hanson, R., *Journal of Applied Crystallography* **43**, 1250-1260 (2010).

45. Szklarczyk, D. et al., *Nucleic Acids Res* **39**, D561-568 (2011).

46. Deutsch, E.W., Lam, H. & Aebersold, R., *EMBO Rep* **9**, 429-434 (2008).

47. Craig, R., Cortens, J.P. & Beavis, R.C., *J Proteome Res* **3**, 1234-1242 (2004).

48. Hartler, J. et al., *BMC Bioinformatics* **8**, 197 (2007).

49. Hakkinen, J., Vincic, G., Mansson, O., Warell, K. & Levander, F., *J Proteome Res* **8**, 3037-3043 (2009).

50. Helsens, K. et al., *Proteomics* **10**, 1261-1264 (2010).