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.

There are different packages covering fully or partially this workflow. A detailed list mainly focused on free software is given in the appendix. Among these, 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](file:///C:\Users\hba041\Teaching\WT%20Course%20EBI%202012\www.openms.de)).
3. We recommend **UniProt[3](#_ENREF_3" \o "Apweiler, 2004 #45)** ([www.uniprot.org](file:///C:\Users\hba041\Teaching\WT%20Course%20EBI%202012\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 here use 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 **SearchGUI[7](#_ENREF_7" \o "Vaudel, 2011 #18)** ([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 these we will use:   
   **UniProt**[3](#_ENREF_3" \o "Apweiler, 2004 #45) (<http://www.uniprot.org>), **Reactome**[9](#_ENREF_9" \o "Haw, 2011 #139) (<http://www.reactome.org>),   
   **PICR**[10](#_ENREF_10) (<http://www.ebi.ac.uk/Tools/picr>) and **Dasty**[11](#_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.
8. In order to make your data publicly available, you can upload them in public repositories. We recommend **ProteomeXchange** (<http://proteomexchange.org>) and **PRIDE**[**12**](#_ENREF_12) (<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** *(in development)*
7. **Other Resources**
8. **Submission to PRIDE and ProteomeXchange**

You will find a folder named software containing all the software needed for this tutorial as well as eight folders corresponding to the eight 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[13](#_ENREF_13" \o "Cote, 2010 #150) | Mass spectrometry mzML file parser |
| General proteomics package | OpenMS[2](#_ENREF_2" \o "Bertsch, 2011 #15) | Package of tools for proteomics allowing the design  of workflows with a graphical interface |
| TPP[14](#_ENREF_14) | Package of tools for proteomics mainly command line driven |
| MaxQuant[15](#_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[16](#_ENREF_16" \o "Slotta, 2010 #128) | Identification processing software |
| De novo sequencing | PepNovo[17](#_ENREF_17" \o "Frank, 2005 #97) | De novo sequencing tool |
| PEAKS[18](#_ENREF_18) | De novo sequencing tool (commercial) |
| Tag sequencing | GutenTag[19](#_ENREF_19" \o "Tabb, 2003 #191) | Finds peptide patterns in spectra |
| DirecTag[20](#_ENREF_20" \o "Tabb, 2008 #190) | Finds peptide patterns in spectra |
| Database search engine | Sequest[21](#_ENREF_21" \o "Yates, 1995 #99) | Database search engine (commercial) |
| Mascot[22](#_ENREF_22) | Database search engine (commercial) |
| OMSSA[5](#_ENREF_5) | Database search engine |
| X!Tandem[6](#_ENREF_6" \o "Craig, 2004 #46) | Database search engine |
| Inspect[23](#_ENREF_23) | Database search engine |
| MyriMatch[24](#_ENREF_24) | Database search engine |
| MassWiz[25](#_ENREF_25" \o "Yadav, 2011 #101) | Database search engine |
| Andromeda[26](#_ENREF_26) | Database search engine (MaxQuant only) |
| User friendly interfaces | SearchGUI[7](#_ENREF_7) | Graphical interface for OMSSA and X!Tandem |
| PRIDE Inspector[27](#_ENREF_27) | Graphical interface for the inspection of PRIDE XML files |
| TOPPAS[28](#_ENREF_28) | Graphical interface for the design of OpenMS workflows |
| Spectral library searching | NIST MS search[29](#_ENREF_29) | Spectral libraries search engine |
| X!Hunter[30](#_ENREF_30" \o "Craig, 2006 #106) | Spectral libraries search engine |
| SpectraST[31](#_ENREF_31" \o "Lam, 2007 #108) | Spectral libraries search engine |
| Identification file parsers | MascotDatFile[32](#_ENREF_32" \o "Helsens, 2007 #59) | Java parser for Mascot .dat files |
| OMSSA parser[33](#_ENREF_33) | Java parser for OMSSA .omx files |
| X!Tandem parser[34](#_ENREF_34) | Java parser for X!Tandem XML files |
| Data structure | compomics-utilities[35](#_ENREF_35) | Java object structure for the handling and visualization  of identifications from different search engines |
| PSM rescoring | Percolator[36](#_ENREF_36) | Machine learning algorithm rescoring  PSMs and attaching them a p-value |
| PeptideProphet[37](#_ENREF_37" \o "Keller, 2002 #27) | Machine learning algorithm attaching  PSMs a PEP (integrated in TPP) |
| PepArML[38](#_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[4](#_ENREF_4" \o "Martens, 2005 #19) | Tool allowing the manipulation of databases  and creation of custom ones |
| Peptide inference | iProphet[39](#_ENREF_39" \o "Shteynberg, 2011 #126) | Tool for statistical post-processing of PSMs  (integrated in TPP) |
| Protein inference | ProteinProphet[40](#_ENREF_40" \o "Nesvizhskii, 2003 #10) | Tool for protein inference (integrated in TPP) |
| IDPicker[41](#_ENREF_41" \o "Ma, 2009 #127) | Tool for protein inference |
| MassSieve[16](#_ENREF_16" \o "Slotta, 2010 #128) | Identification processing software |
| Protein annotation | UniProtKB[3](#_ENREF_3" \o "Apweiler, 2004 #45) | Protein knowledge database |
| Dasty[11](#_ENREF_11" \o "Jones, 2005 #81) | Cross reference tool for protein databases |
| GO enrichment | GOTree[42](#_ENREF_42" \o "Zhang, 2004 #134) | GO enrichment tool |
| Onotologizer[43](#_ENREF_43" \o "Bauer, 2008 #135) | GO enrichment tool |
| DAVID[44](#_ENREF_44) | Interface for enrichment of identification results |
| 3D structures | jmol[45](#_ENREF_45" \o "Hanson, 2010 #138) | Tool for the display of 3D structures |
| Pathways | Reactome[9](#_ENREF_9" \o "Haw, 2011 #139) | Pathway investigation interface allowing the mapping  of one’s results and pathway coverage estimation |
| Interactions | STRING[46](#_ENREF_46) | Protein interaction investigation interface |
| Repository | PRIDE[12](#_ENREF_12) | Protein identification repository |
| PeptideAtlas[47](#_ENREF_47" \o "Deutsch, 2008 #142) | Peptide identification repository |
| GPMDB[48](#_ENREF_48) | Peptide and protein identification repository |
| Local data management | MASPECTRAS[49](#_ENREF_49) | LIMS system |
| Proteios[50](#_ENREF_50" \o "Hakkinen, 2009 #160) | LIMS system |
| ms\_lims[51](#_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. 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.

(9) Haw, R.; Hermjakob, H.; D'Eustachio, P.; Stein, L. Reactome pathway analysis to enrich biological discovery in proteomics data sets. *Proteomics* **2011**, *11*, 3598.

(10) Cote, R. G.; Jones, P.; Martens, L.; Kerrien, S.; Reisinger, F.; Lin, Q.; Leinonen, R.; Apweiler, R.; Hermjakob, H. The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics* **2007**, *8*, 401.

(11) Jones, P.; Vinod, N.; Down, T.; Hackmann, A.; Kahari, A.; Kretschmann, E.; Quinn, A.; Wieser, D.; Hermjakob, H.; Apweiler, R. Dasty and UniProt DAS: a perfect pair for protein feature visualization. *Bioinformatics* **2005**, *21*, 3198.

(12) Martens, L.; Hermjakob, H.; Jones, P.; Adamski, M.; Taylor, C.; States, D.; Gevaert, K.; Vandekerckhove, J.; Apweiler, R. PRIDE: the proteomics identifications database. *Proteomics* **2005**, *5*, 3537.

(13) Cote, R. G.; Reisinger, F.; Martens, L. jmzML, an open-source Java API for mzML, the PSI standard for MS data. *Proteomics* **2010**, *10*, 1332.

(14) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; Tasman, N.; Sun, Z.; Nilsson, E.; Pratt, B.; Prazen, B.; Eng, J. K.; Martin, D. B.; Nesvizhskii, A. I.; Aebersold, R. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **2010**, *10*, 1150.

(15) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **2008**, *26*, 1367.

(16) Slotta, D. J.; McFarland, M. A.; Markey, S. P. MassSieve: panning MS/MS peptide data for proteins. *Proteomics* **2010**, *10*, 3035.

(17) Frank, A.; Pevzner, P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* **2005**, *77*, 964.

(18) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* **2003**, *17*, 2337.

(19) Tabb, D. L.; Saraf, A.; Yates, J. R., 3rd GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal Chem* **2003**, *75*, 6415.

(20) Tabb, D. L.; Ma, Z. Q.; Martin, D. B.; Ham, A. J.; Chambers, M. C. DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. *J Proteome Res* **2008**, *7*, 3838.

(21) Yates, J. R., 3rd; Eng, J. K.; McCormack, A. L.; Schieltz, D. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem* **1995**, *67*, 1426.

(22) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551.

(23) Tanner, S.; Shu, H.; Frank, A.; Wang, L. C.; Zandi, E.; Mumby, M.; Pevzner, P. A.; Bafna, V. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem* **2005**, *77*, 4626.

(24) Tabb, D. L.; Fernando, C. G.; Chambers, M. C. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res* **2007**, *6*, 654.

(25) Yadav, A. K.; Kumar, D.; Dash, D. MassWiz: a novel scoring algorithm with target-decoy based analysis pipeline for tandem mass spectrometry. *J Proteome Res* **2011**, *10*, 2154.

(26) Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **2011**, *10*, 1794.

(27) Wang, R.; Fabregat, A.; Rios, D.; Ovelleiro, D.; Foster, J. M.; Cote, R. G.; Griss, J.; Csordas, A.; Perez-Riverol, Y.; Reisinger, F.; Hermjakob, H.; Martens, L.; Vizcaino, J. A. PRIDE Inspector: a tool to visualize and validate MS proteomics data. *Nat Biotechnol* **2012**, *30*, 135.

(28) Junker, J.; Bielow, C.; Bertsch, A.; Sturm, M.; Reinert, K.; Kohlbacher, O. TOPPAS: A Graphical Workflow Editor for the Analysis of High-Throughput Proteomics Data. *J Proteome Res* **2012**.

(29) Stein, S. E.; Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry* **1994**, *5*, 859.

(30) Craig, R.; Cortens, J. C.; Fenyo, D.; Beavis, R. C. Using annotated peptide mass spectrum libraries for protein identification. *J Proteome Res* **2006**, *5*, 1843.

(31) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **2007**, *7*, 655.

(32) Helsens, K.; Martens, L.; Vandekerckhove, J.; Gevaert, K. MascotDatfile: an open-source library to fully parse and analyse MASCOT MS/MS search results. *Proteomics* **2007**, *7*, 364.

(33) Barsnes, H.; Huber, S.; Sickmann, A.; Eidhammer, I.; Martens, L. OMSSA Parser: an open-source library to parse and extract data from OMSSA MS/MS search results. *Proteomics* **2009**, *9*, 3772.

(34) Muth, T.; Vaudel, M.; Barsnes, H.; Martens, L.; Sickmann, A. XTandem Parser: an open-source library to parse and analyse X!Tandem MS/MS search results. *Proteomics* **2010**, *10*, 1522.

(35) Barsnes, H.; Vaudel, M.; Colaert, N.; Helsens, K.; Sickmann, A.; Berven, F. S.; Martens, L. compomics-utilities: an open-source Java library for computational proteomics. *BMC Bioinformatics* **2011**, *12*, 70.

(36) Kall, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* **2007**, *4*, 923.

(37) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* **2002**, *74*, 5383.

(38) Edwards, N.; Wu, X.; Tseng, C.-W. An Unsupervised, Model-Free, Machine-Learning Combiner for Peptide Identifications from Tandem Mass Spectra. *Clinical Proteomics* **2009**, *5*, 23.

(39) Shteynberg, D.; Deutsch, E. W.; Lam, H.; Eng, J. K.; Sun, Z.; Tasman, N.; Mendoza, L.; Moritz, R. L.; Aebersold, R.; Nesvizhskii, A. I. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol Cell Proteomics* **2011**, *10*, M111 007690.

(40) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **2003**, *75*, 4646.

(41) Ma, Z. Q.; Dasari, S.; Chambers, M. C.; Litton, M. D.; Sobecki, S. M.; Zimmerman, L. J.; Halvey, P. J.; Schilling, B.; Drake, P. M.; Gibson, B. W.; Tabb, D. L. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J Proteome Res* **2009**, *8*, 3872.

(42) Zhang, B.; Schmoyer, D.; Kirov, S.; Snoddy, J. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics* **2004**, *5*, 16.

(43) Bauer, S.; Grossmann, S.; Vingron, M.; Robinson, P. N. Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics* **2008**, *24*, 1650.

(44) Huang da, W.; Sherman, B. T.; Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **2009**, *4*, 44.

(45) Hanson, R. Jmol - a paradigm shift in crystallographic visualization. *Journal of Applied Crystallography* **2010**, *43*, 1250.

(46) Szklarczyk, D.; Franceschini, A.; Kuhn, M.; Simonovic, M.; Roth, A.; Minguez, P.; Doerks, T.; Stark, M.; Muller, J.; Bork, P.; Jensen, L. J.; von Mering, C. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* **2011**, *39*, D561.

(47) Deutsch, E. W.; Lam, H.; Aebersold, R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep* **2008**, *9*, 429.

(48) Craig, R.; Cortens, J. P.; Beavis, R. C. Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res* **2004**, *3*, 1234.

(49) Hartler, J.; Thallinger, G. G.; Stocker, G.; Sturn, A.; Burkard, T. R.; Korner, E.; Rader, R.; Schmidt, A.; Mechtler, K.; Trajanoski, Z. MASPECTRAS: a platform for management and analysis of proteomics LC-MS/MS data. *BMC Bioinformatics* **2007**, *8*, 197.

(50) Hakkinen, J.; Vincic, G.; Mansson, O.; Warell, K.; Levander, F. The proteios software environment: an extensible multiuser platform for management and analysis of proteomics data. *J Proteome Res* **2009**, *8*, 3037.

(51) Helsens, K.; Colaert, N.; Barsnes, H.; Muth, T.; Flikka, K.; Staes, A.; Timmerman, E.; Wortelkamp, S.; Sickmann, A.; Vandekerckhove, J.; Gevaert, K.; Martens, L. ms\_lims, a simple yet powerful open source laboratory information management system for MS-driven proteomics. *Proteomics* **2010**, *10*, 1261.