Database Generation

In order to identify our peptides and proteins, we are going to compare the mass spectra to *in sillico* theoretic spectra deduced from a protein database. *Are there other database types that could be used to identify the spectra? Would it even be possible to identify the spectra without a database at all?*

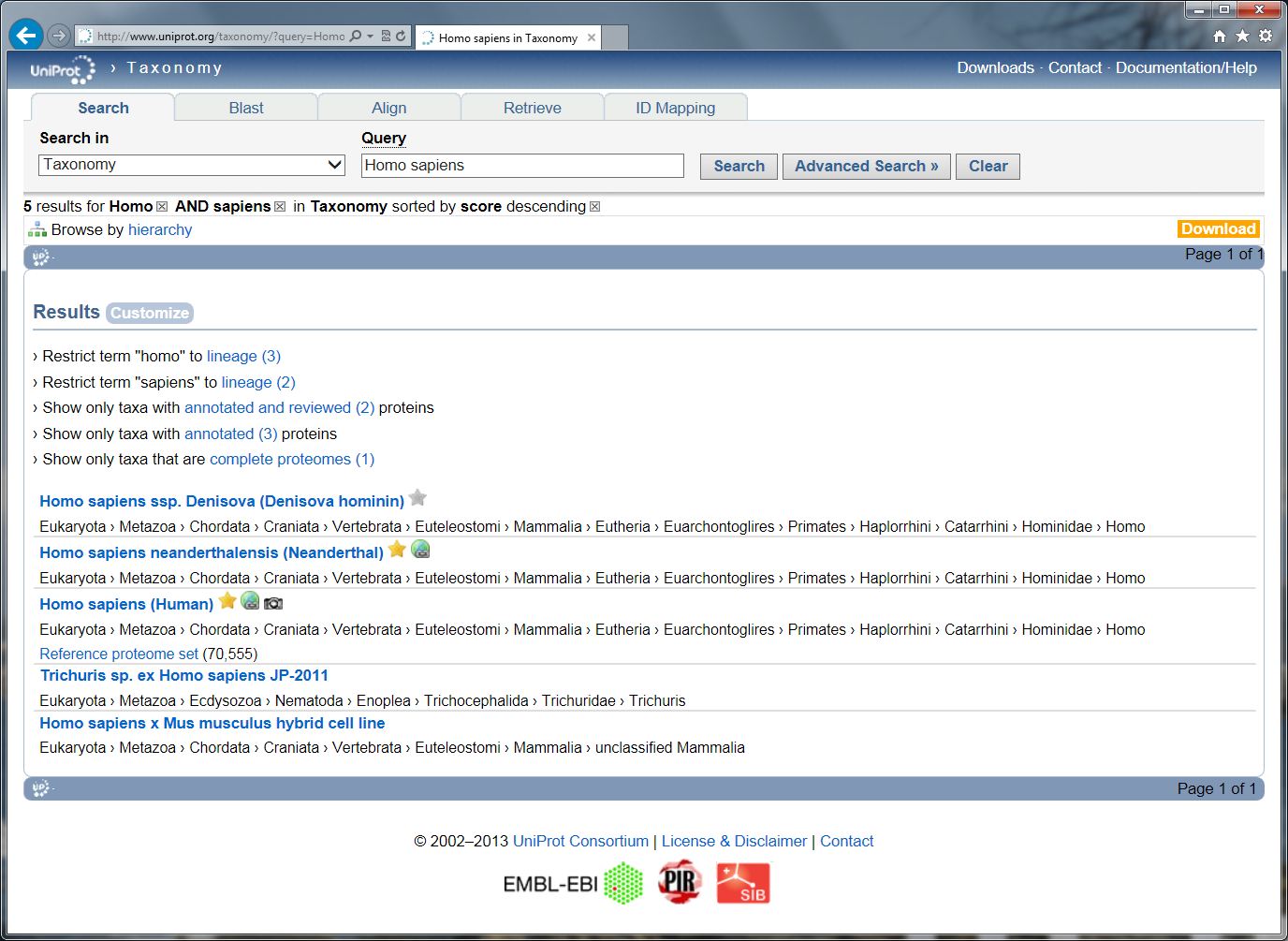
*It is also possible to identify spectra using so-called spectral libraries*[*1*](#_ENREF_1)*, then, experimental spectra are compared to already identified spectra. This approach is already widely used for the identification of small molecules and taking off for peptides*[*2*](#_ENREF_2)*.*

*Finally,* de novo *algorithms*[*3*](#_ENREF_3)*,* [*4*](#_ENREF_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*.*

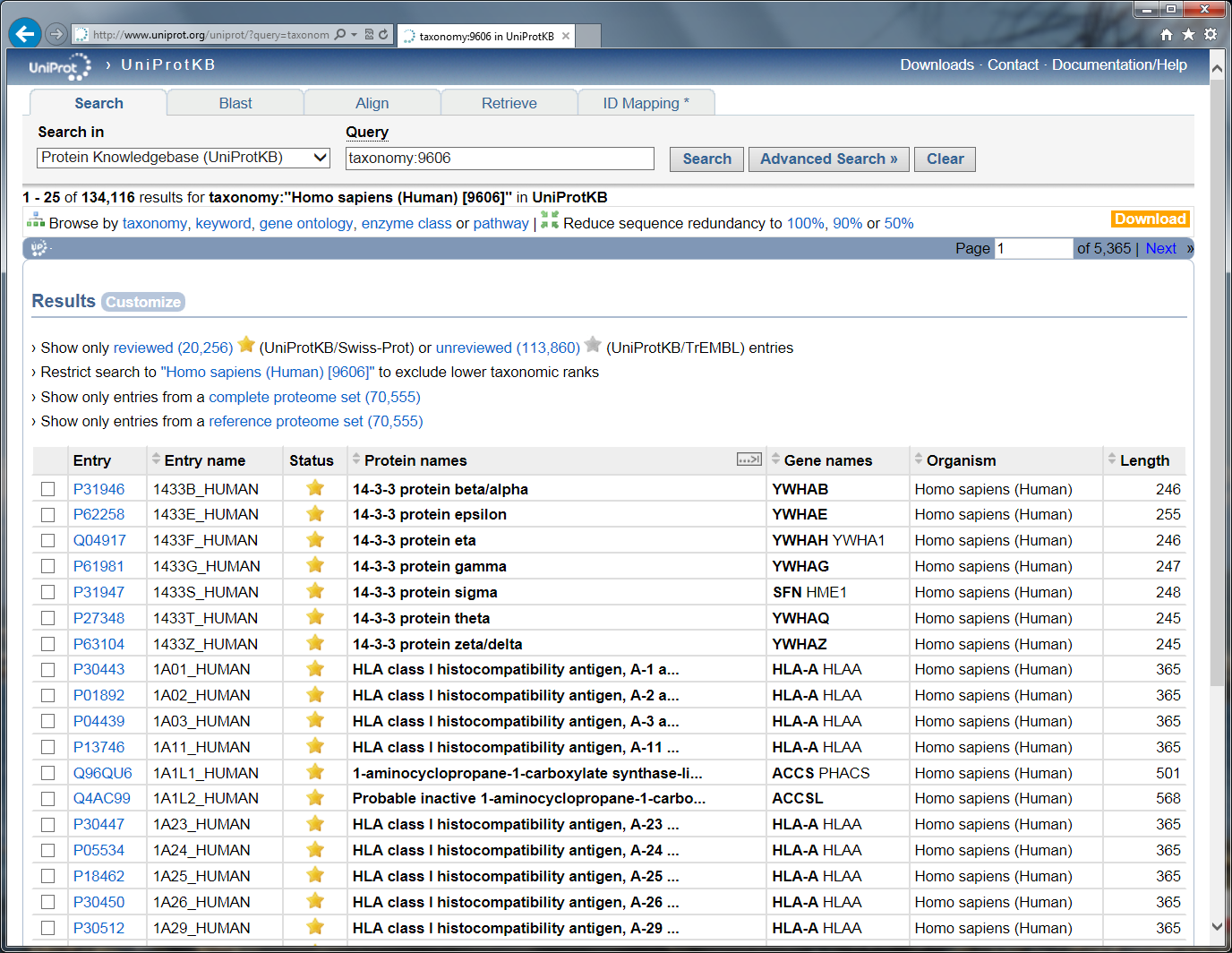
The choice of the database is crucial for the identification procedure. Indeed, shotgun proteomics workflows will only retrieve proteins contained in the database: the database should contain all possible sequences. Yet, if the database is too large, the search engine will have more room for mistakes and will introduce false positive identifications. The UniProt[5](#_ENREF_5" \o "Apweiler, 2004 #45) database is a repository of choice for proteomics as it allies quality and quantity of protein sequences.

In order to optimize the database size, we will select only the species needed. The spectra in our example were obtained from a human sample. Go to the UniProt website (<www.uniprot.org>) and select *Taxonomy* under “Search in” and type *Homo Sapiens* under “Query”.

UniProt retrieves 5 hits:



Select the one named ‘*Homo sapiens (Human)*’, the website then displays the taxonomy tree. Click on the ‘UniProtKB’ link in the upper left corner. UniProt now provides all the proteins expected for the selected organism:



*How many proteins can we find for this proteome? How is the protein sequence list established? Is it exhaustive? What is the difference between a gold star entry and a silver star entry?*

*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 Swiss-Prot entries. The silver star entries on the other hand are algorithmic prediction where no experimental validation is annotated in Uniprot.*

*As we will see in a following chapter, 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 preferably. Eventually, it is possible to add other sequences or research the data against 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 every publication.*

Select the "Show only reviewed" (UniProtKB/Swiss-Prot) option and then click on download. You can here choose between “Canonical sequences” and “Canonical and isoform” sequences. *What is the difference?*

*Uniprot can provide you isoforms of protein sequences. As for above, these should be used with caution as they dramatically reduce the efficiency of the identification algorithms.*

Download the file called ‘FASTA - Canonical sequence data in FASTA format’. You now have the desired FASTA file needed to search the mass spectrometry example dataset.

**Tip:**  
*Always document your database type and version.*

*Organize your databases in a rational way for you and your colleagues!*

Advanced – Database manipulation

If you need advanced re-processing of FASTA files, we recommend the use of dbtoolkit[6](#_ENREF_6" \o "Martens, 2005 #19) (<http://dbtoolkit.googlecode.com>).

Advanced – homemade databases

For some specific studies, one needs to create its own database. This is facilitated by the relatively simple syntax of the fasta format which can be edited in a normal text editor:

>header

SEQUENCE

As illustrated here with the sequence of a human protein:

>sp|A6NCN2|K121P\_HUMAN Keratin-81-like protein KRT121P OS=Homo sapiens GN=KRT121P PE=5 SV=4

MEANSGRLASELNHVQEVLEGYKKKYEEEVALRATAENEFVALKKDVDCAYLRKSDLEAN

VEALTQEIDFLRRLYEEEIRVLQSHISDTSVVVKMDNSRDLNMHCVITEIKAQYDDIATR

SRAEAESWYRSKCEEMKATVIRHGETLRRTKEEINELNRMIQRLTAEVENAKCQNSKLEA

AVAQSEQQGEAALSDARCKLAELEGALQKAKQDMACLIREYQEVMNSKLAWTLRSPPTGA

CWRARSRGCVRALVL

It is however vital that the syntax used for the header is compatible with the search engines. For homemade databases, we recommend a generic format as detailed in our Database Help page (<http://code.google.com/p/searchgui/wiki/DatabaseHelp>). There you will find information about how to set up your own custom databases.

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. Apweiler, R. et al. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32**, D115-119 (2004).

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