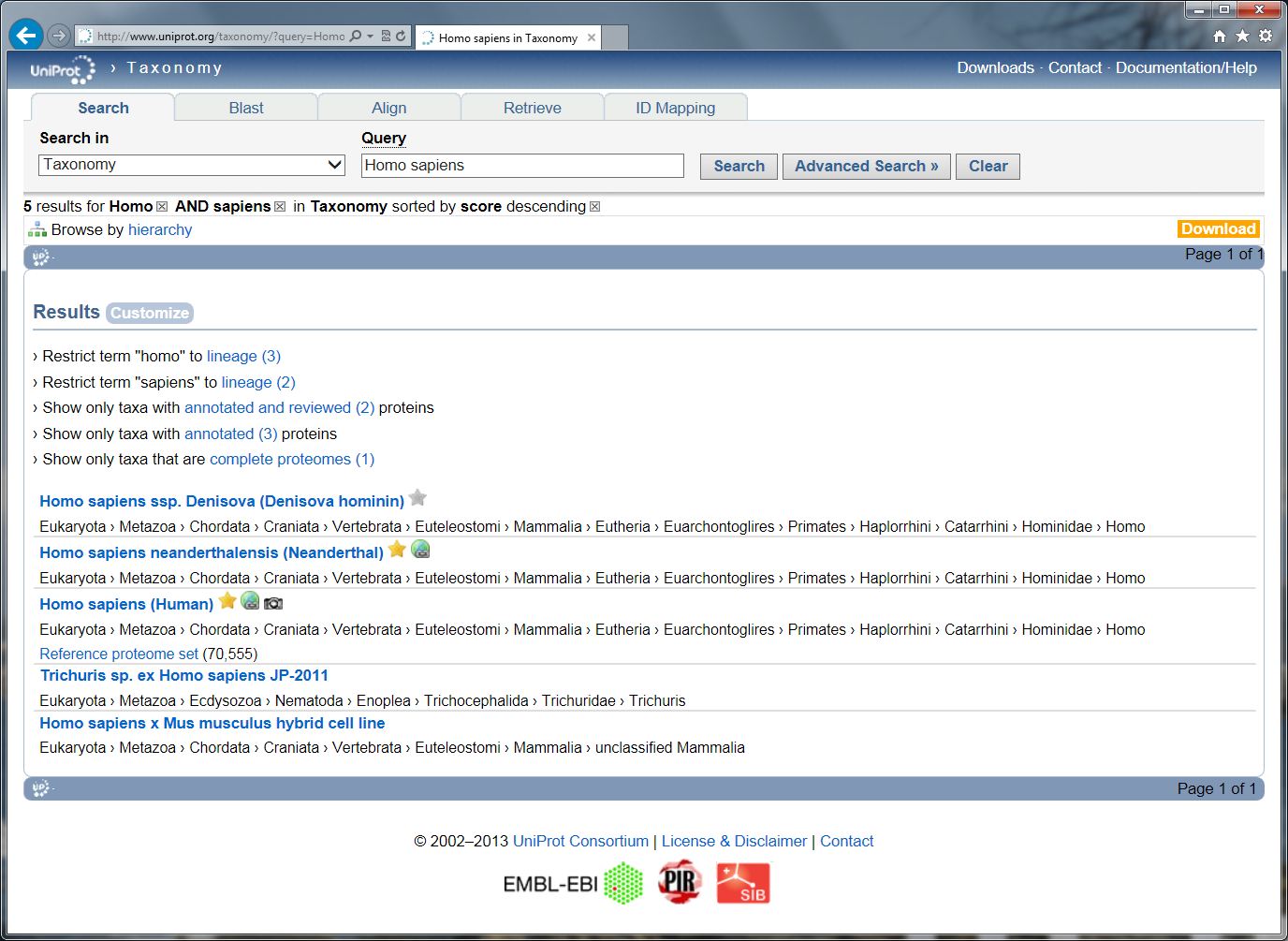
Database Generation

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? [1.1a]*

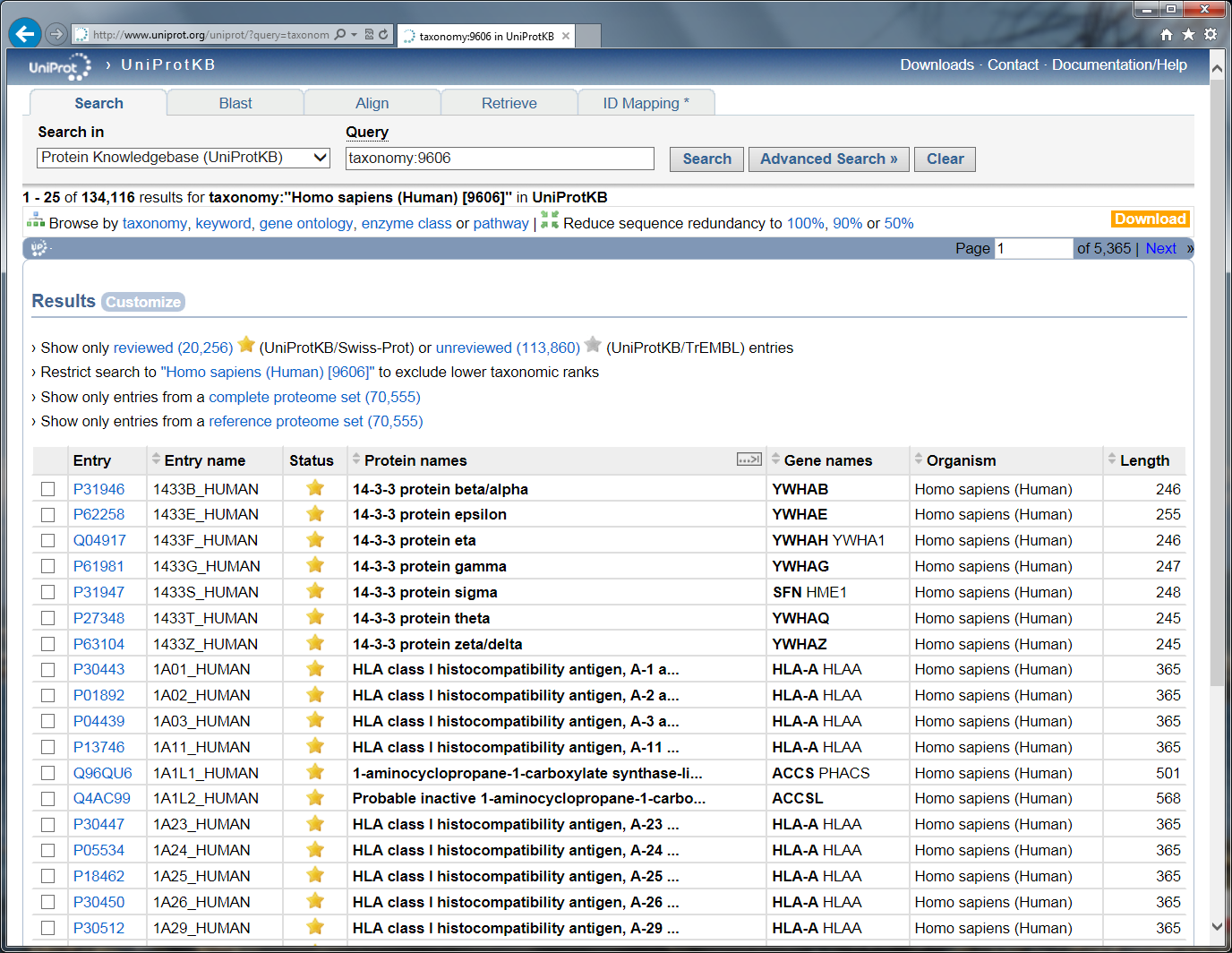
The choice of the database is crucial for the identification procedure. Indeed, shotgun proteomics workflows will only retrieve proteins contained in the database: the database should contain all possible sequences. Yet, if the database is too large, the search engine will have more room for mistakes and will introduce false positive identifications. The UniProt[1](#_ENREF_5) 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? [1.1b]*

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? [1.1c]*

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 – Non Standard Databases

For some studies, one has to create a non standard 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 and the tools used to process the search results. For homemade databases, we recommend a generic format as detailed on 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.

Advanced – Database Manipulation

If you need advanced re-processing of FASTA files, we recommend the use of dbtoolkit[2](#_ENREF_6) (<http://dbtoolkit.googlecode.com>).

References

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

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