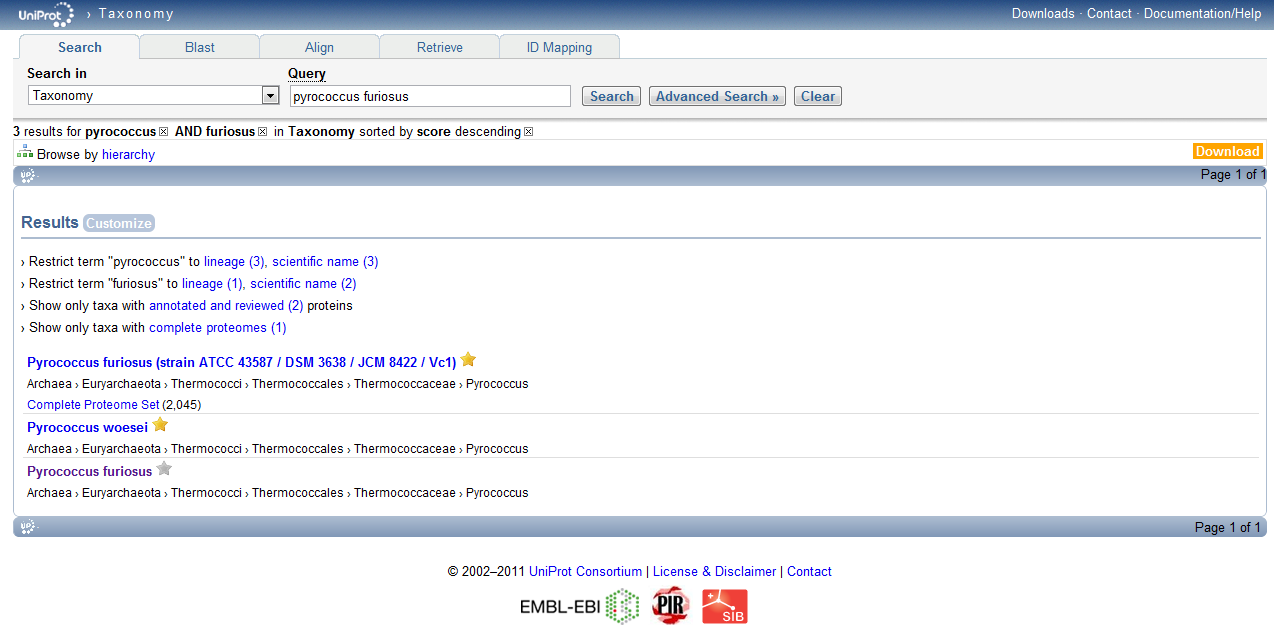
Database Generation

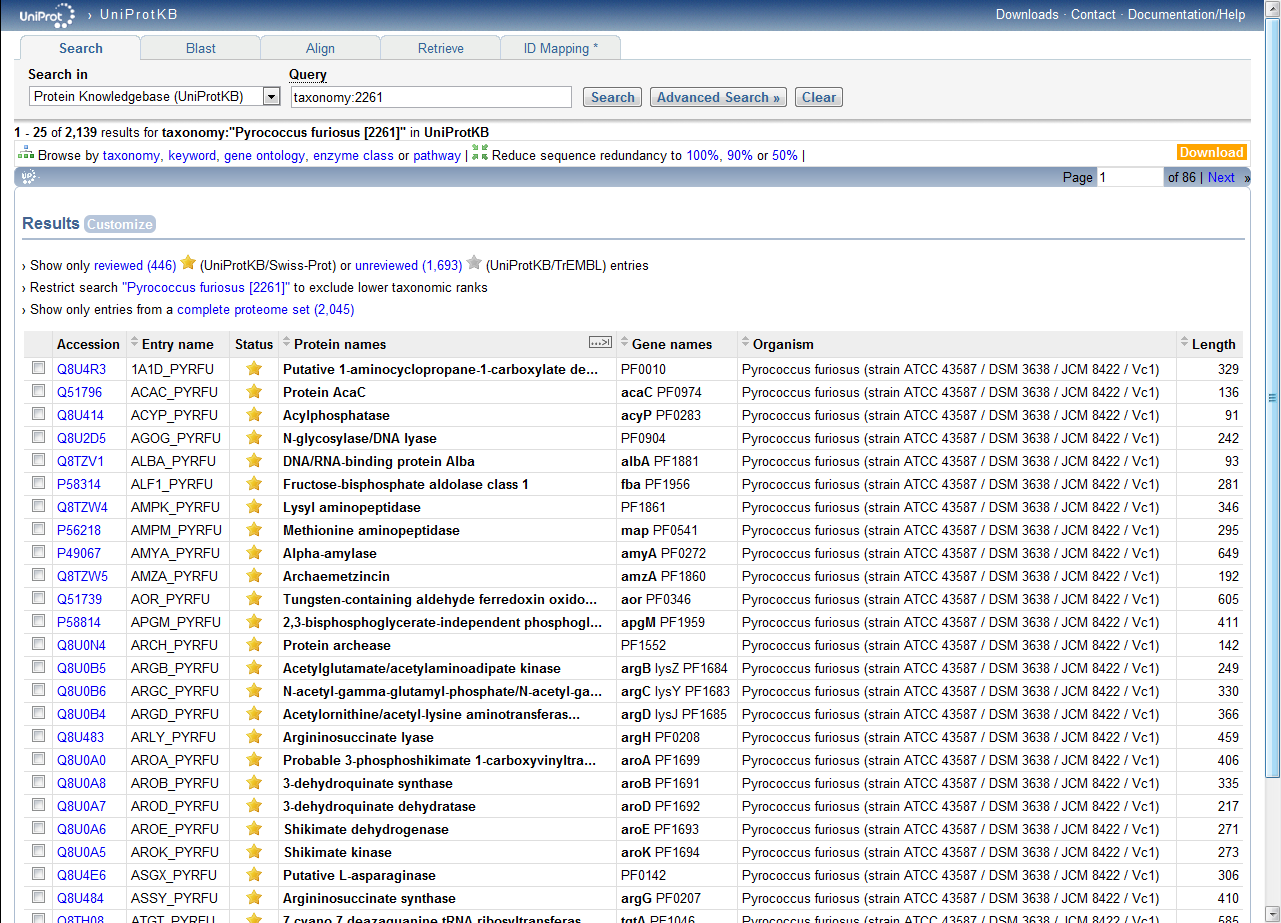
The choice of the database is critical for the identification procedure. Indeed, shotgun proteomics workflows will only retrieve proteins contained in the database: the database should contain all possible sequences. Yet, if the database is too large, the search engine will have more room for mistakes and will introduce false positive identifications. The UniProt[1](#_ENREF_1" \o "Apweiler, 2004 #45) database is a repository of choice for proteomics as it allies quality and quantity of protein sequences.

*Here we are using a protein sequence 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?*

In order to optimize the database size, we will select only the species needed. The spectra in our example were obtained on a *Pyrococcus furiosus* sample. Go on the UniProt webite (<www.uniprot.org>) and select *Taxonomy* under “Search in” and type *Pyrococcus furiosus* under “Query”. UniProt retrieves 3 hits:



Select the one named ‘*Pyrococcus furiosus*’, the website displays then the taxonomy tree. Click on ‘UniProtKB’. Uniprot now provides all the proteins expected for this organism:



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

Click on 'Download' in the upper right corner and select ‘FASTA Canonical sequence data in FASTA format’. We now have the desired FASTA file ready for searching. If you need to further process FASTA files, we recommend the use of dbtoolkit[2](#_ENREF_2" \o "Martens, 2005 #19) (<http://dbtoolkit.googlecode.com>).

Advanced

Repeat the same operation with *Homo sapiens*. *How many human sequences does UniProt contain?*

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

Using dbtoolkit (provided in the resources folder) compare the amount of sequences from both possibilities. *Which one would you choose?*

*For further information about databases, we refer to our Database Help page:* [*http://code.google.com/p/searchgui/wiki/DatabaseHelp*](http://code.google.com/p/searchgui/wiki/DatabaseHelp)

*There you will find information about how to setup your own custom databases   
for use in SearchGUI and PeptideShaker.*

References

(1) 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.

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