Functional Analysis

In the previous steps of the tutorial, starting from the raw spectra, we established a list of proteins validated at a certain level of confidence. After all these efforts, all we have is just a list of protein accessions! How do we go about drawing any meaningful biological conclusions from this? As mentioned in the general introduction, many resources[1](#_ENREF_1) are available online:

**Proteomics Results**

In this chapter we are going to look for further protein information, conduct a pathway analysis, view 3D structures of the proteins and finally conduct a Gene Ontology (GO) enrichment analysis. All based on data processed in PeptideShaker. We are going to use a human sample as these are more biologically interesting than the earlier example data. Load the project human.cps located in the resources folder into PeptideShaker. *Note that parts of the tutorial can also be followed without PeptideShaker using the provided web links.*

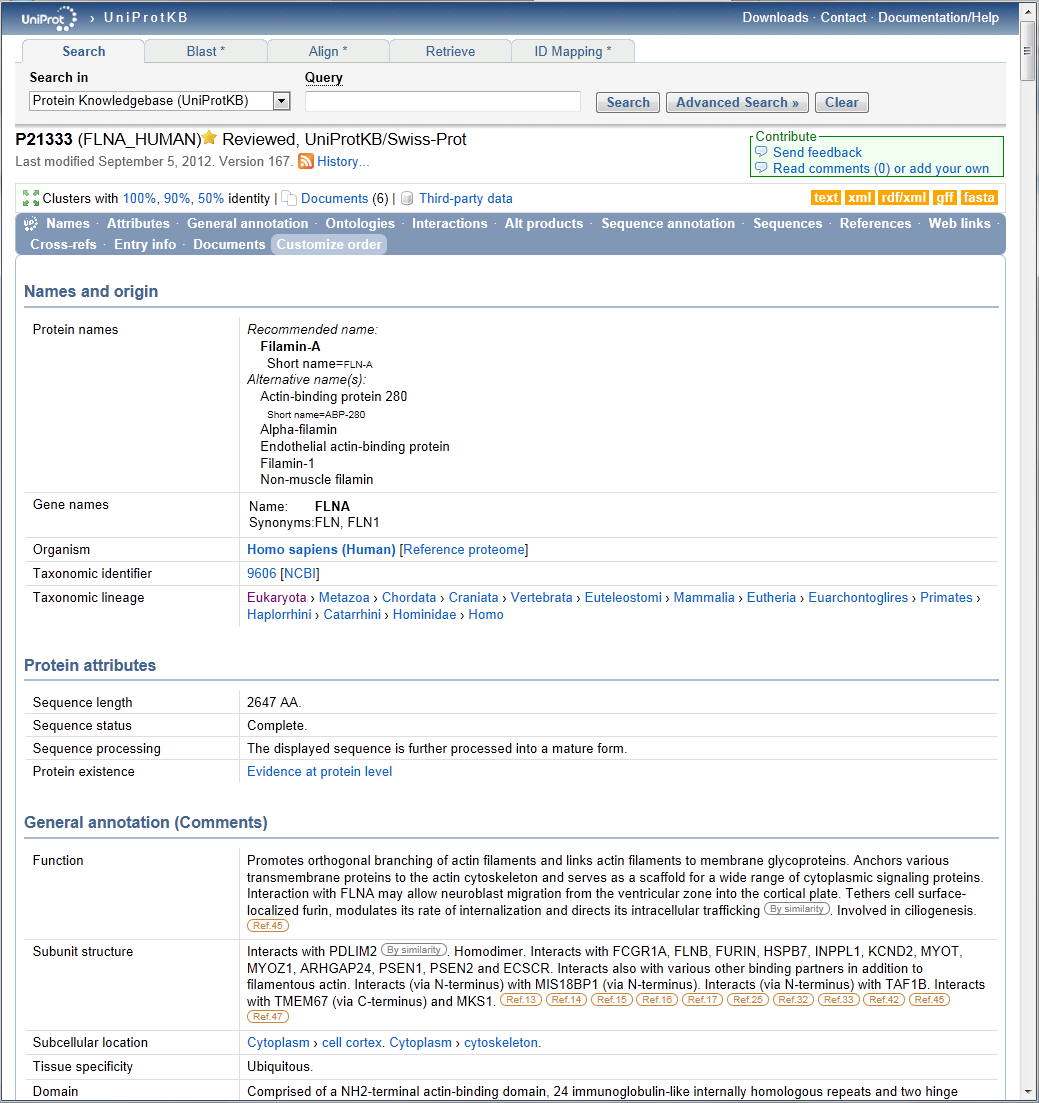
1. Protein Information

You should now have a project with 808 proteins validated at 1% FDR:



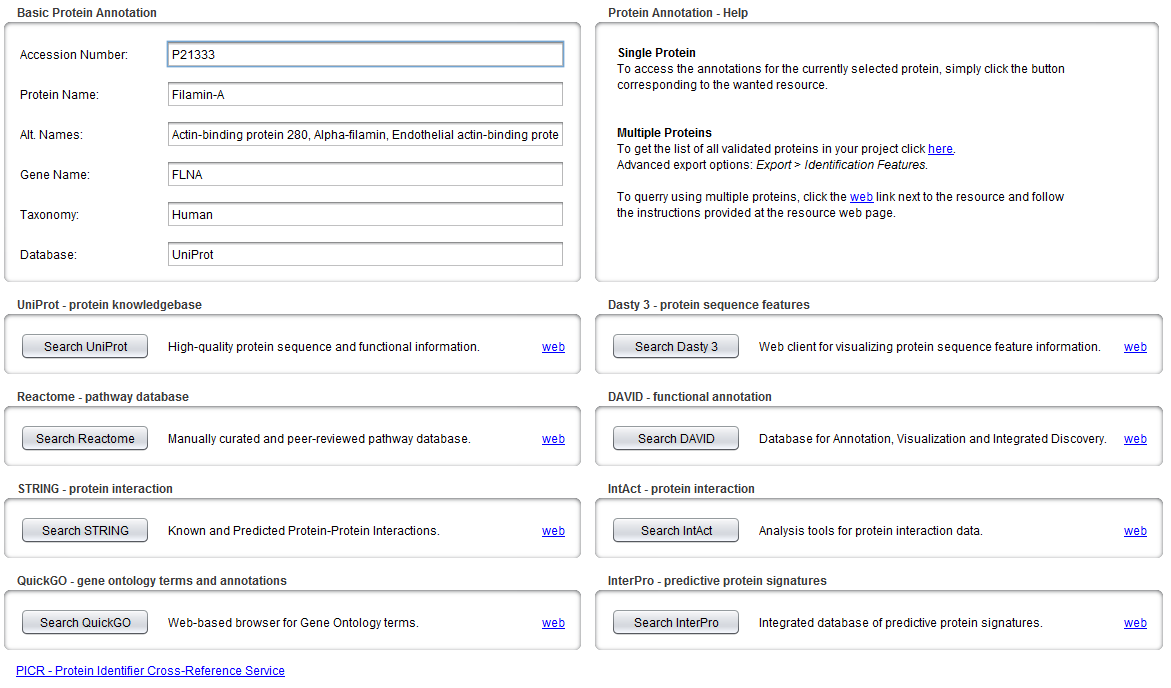
The first source of protein information is UniProtKB[2](#_ENREF_2" \o "Apweiler, 2004 #45), the Universal Protein Resource Knowledge Database which was also used to obtain the list of sequences in the first step of this tutorial. It contains a vast amount of details on every protein. It is accessible from PeptideShaker by clicking on a given protein accession number.

Click on the first accession number in the protein table: P21333. The UniProt entry for this protein is then opened in a web browser:



Scrolling down will show you all available information about this protein. Note especially the Sequence Annotations with references.

In some cases your protein accession number is not recognized in UniProt. Indeed, across resources and database versions, people use different references for the same protein. The Protein Identifier Cross-Reference (PICR) service from the EBI ([http://www.ebi.ac.uk/Tools/picr](http://www.ebi.ac.uk/Tools/picr/)) is a very helpful tool which will help you to find the missing accession number. The link to PICR is given in PeptideShaker along with other available resources in the ‘Annotation’ tab:



*Note that some basic details from UniProt are automatically loaded by PeptideShaker (if the sequence database used is a UniProt database), and displayed in the upper left corner.*

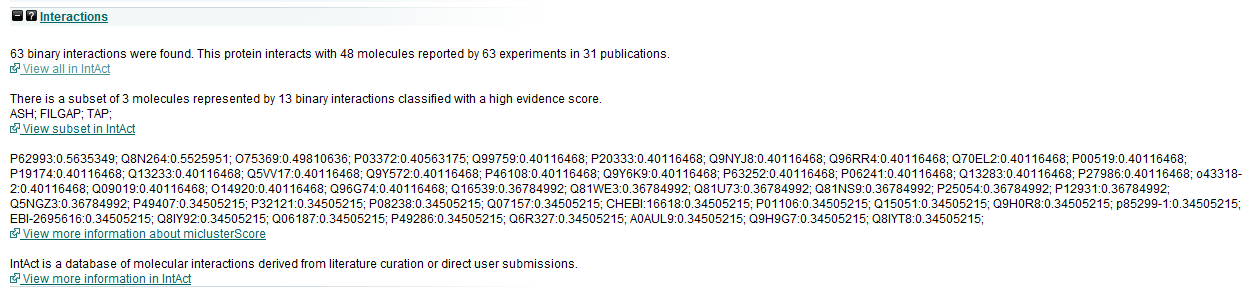
You can use the tool Dasty[3](#_ENREF_3" \o "Jones, 2005 #81) (<http://www.ebi.ac.uk/dasty>), another service from the EBI, to navigate existing UniProtKB information about this sequence. Dasty annotates the sequence of proteins with various features.

Click on ‘Search Dasty 3’, your web browser should then display the following:



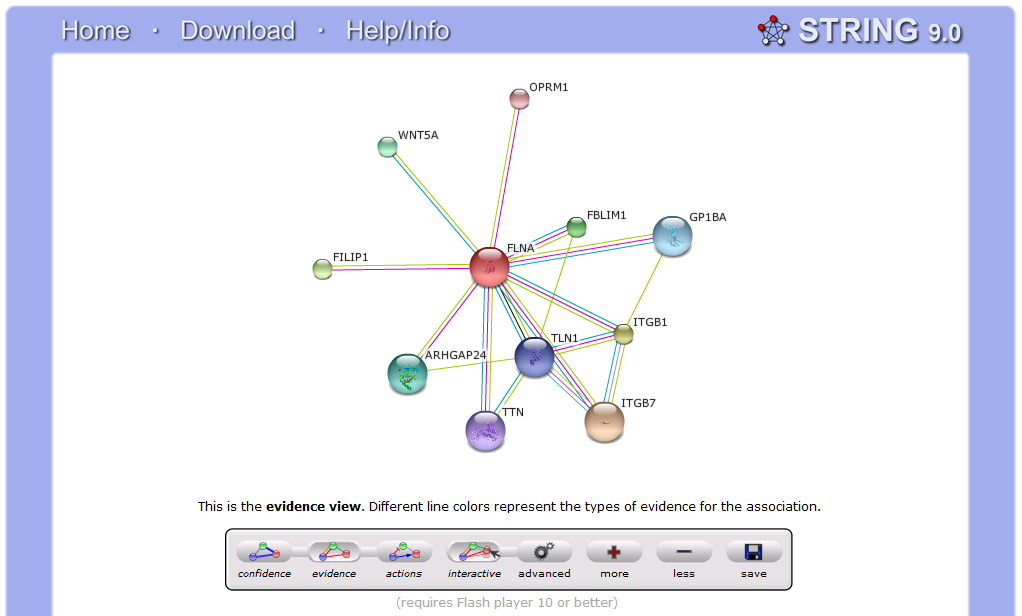
If you hold your mouse over a feature you can for instance see registered variations in the sequence already noticed by users. This function can explain unexpected identification results or modifications. If you scroll down, more information is available like interactions.

Maximize the ‘Interaction' section by clicking the plus icon. You will see the following information:



Interactions of proteins with other molecules are crucial for biological studies. You can obtain more details by following the link to 'Intact'. *Note that the link to Intact is also available directly from PeptideShaker.*

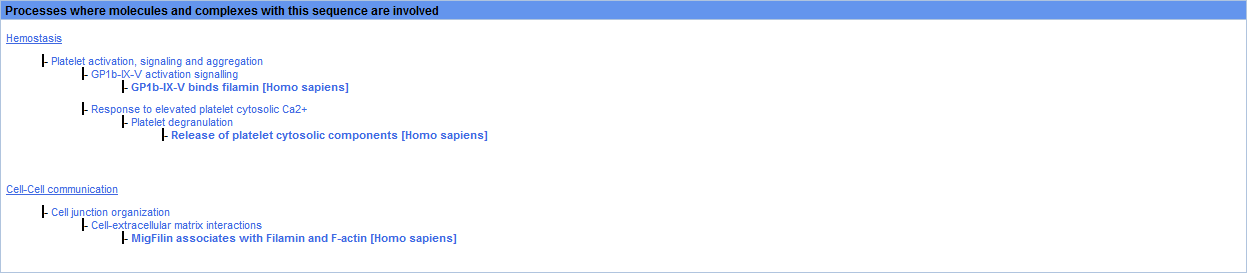
You can also visualize interactions in STRING, the link is provided in PeptideShaker (you just need to click the GO! button in the STRING interface):



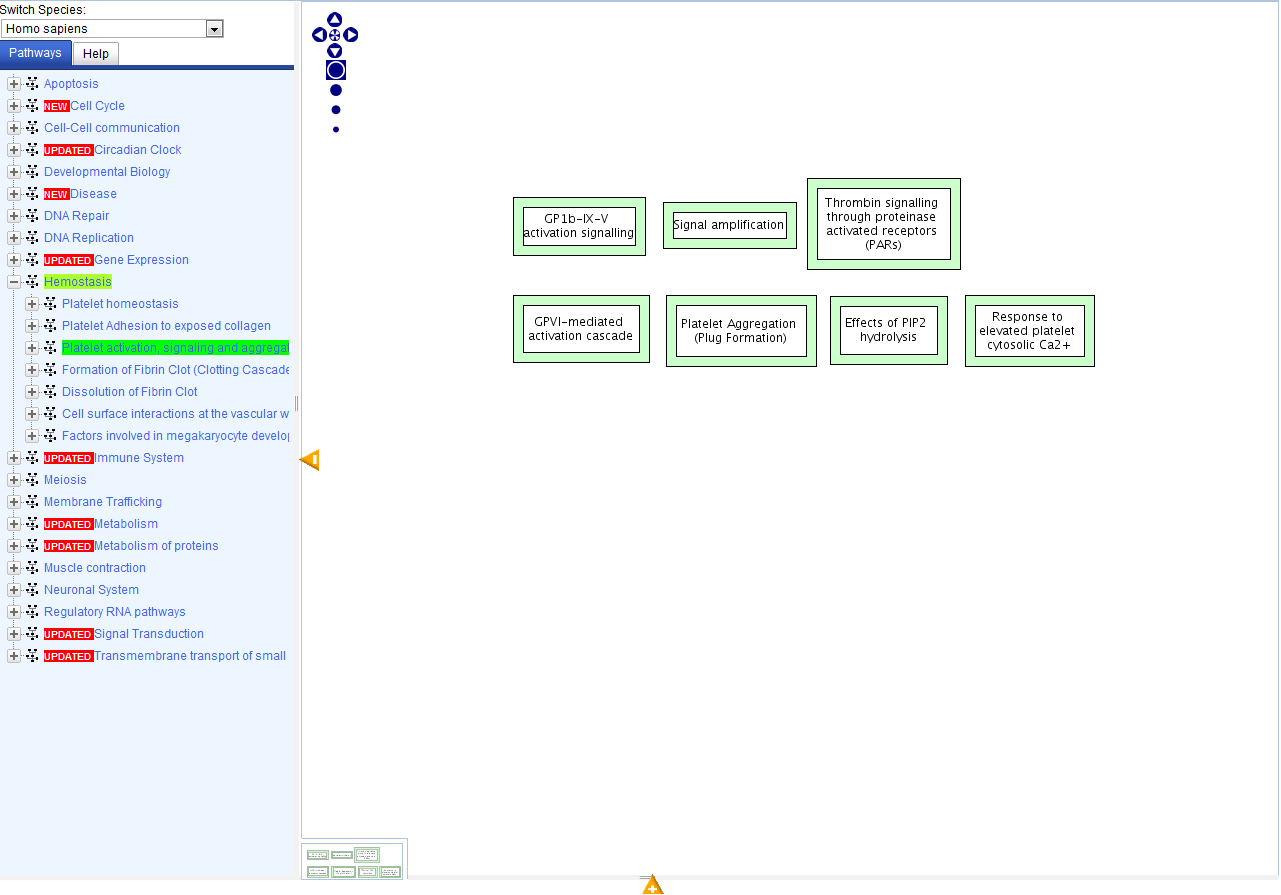
1. Pathways

Reactome is a manually curated and peer-reviewed pathway database. It allows us to see in which pathways our proteins are involved. However, since it relies heavily on previous experiments, not all species are available!

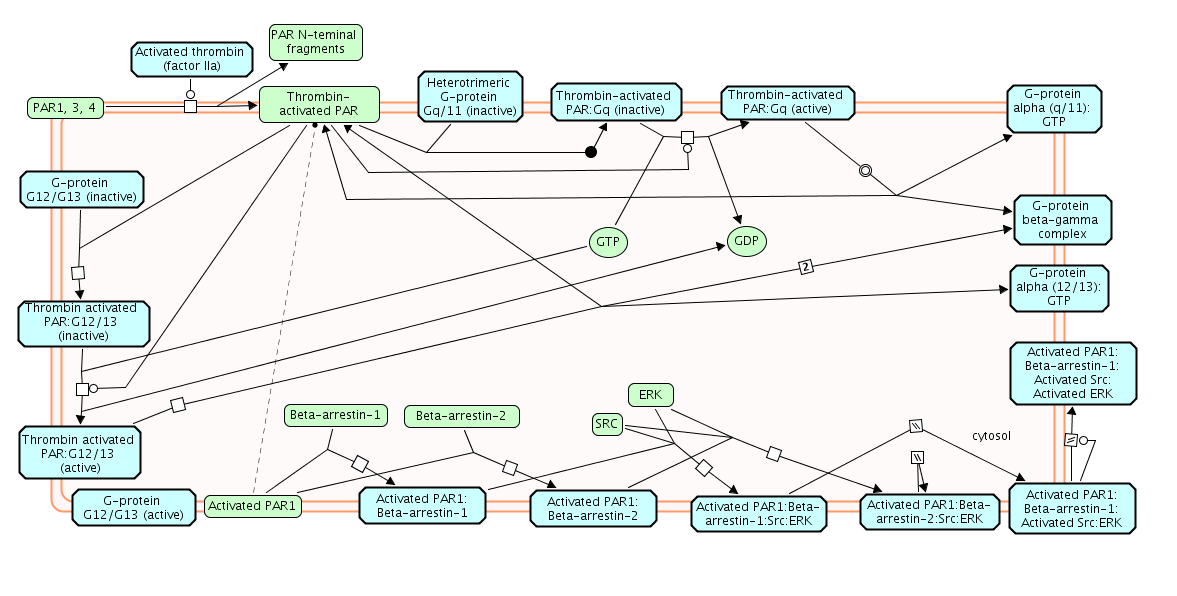
Click on the ‘Search Reactome’ button in PeptideShaker. You will see a summary of information from UniProt as well as a list of processes in which our protein is involved:



If you click on “platelet activation, signalling and aggregation” you will be able to browse the pathway and its main components:



For instance, expanding ‘Thrombin signalling through proteinase activated receptors (PARs)’ will give:



We indeed have here a platelet sample and it is not a surprise to see this pathway covered. Reactome can even quantify our coverage of pathways with our experiment. Go back to PeptideShaker and export all identified accessions, as available in the top right Help of the Annotation panel (the accessions are also available in the resources folder in accessions.txt):



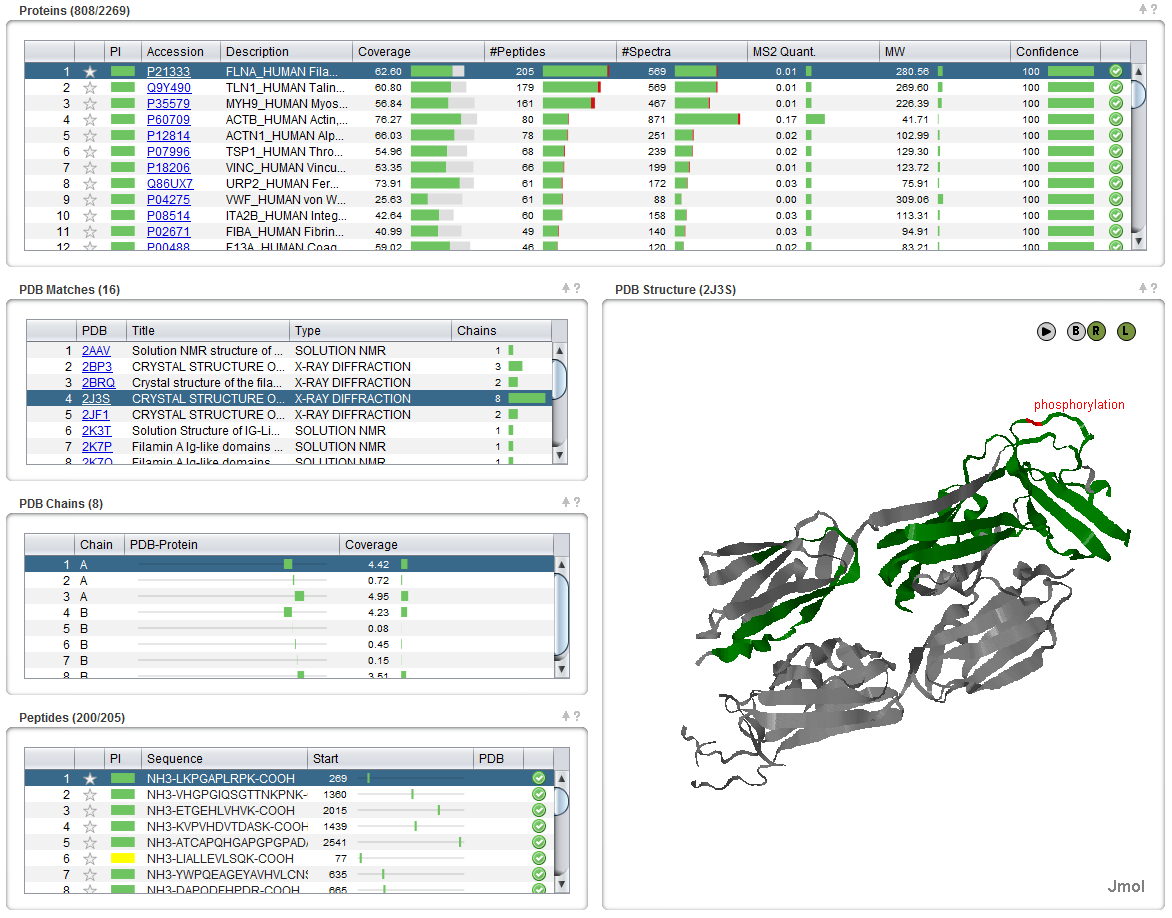
Now go to the Reactome home page (http://www.reactome.org) and go to ‘Analyse Expression Data’. Paste our list of accessions in the text field and click Analyse. When the calculation is finished, after sorting by “% in data”, you should see the following table:



Note that the pathways covered are closely linked to platelets. *Why don’t we cover all pathways completely?*

1. 3D Structures

It is possible to visualize 3D structures of proteins in PeptideShaker. More, the peptides you identified with their Post Translational Modifications are mapped to the structure. Select The ‘3D Structures’ tab in PeptideShaker, select the fourth available structure ‘2J3S’, you should see this:

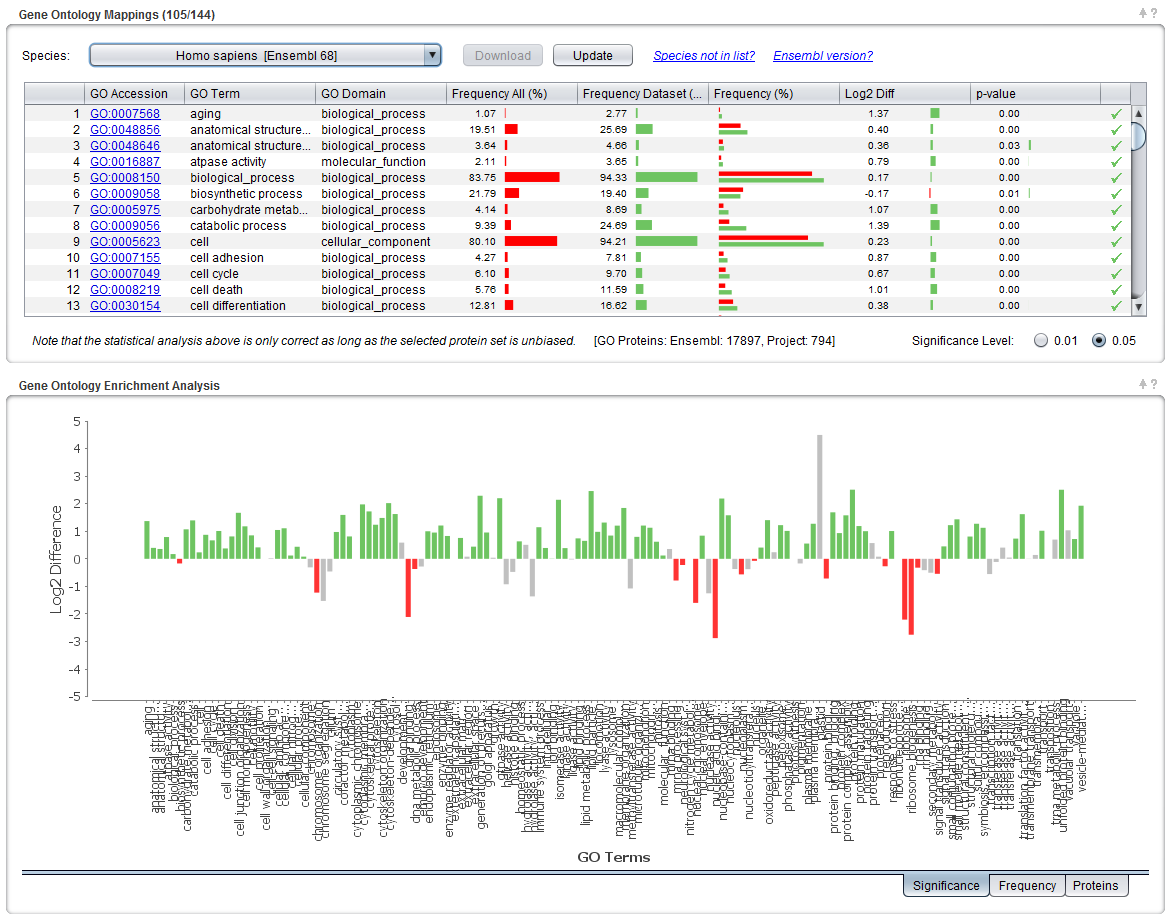


*Why are there many structures available for this protein? Is there a structure available for all proteins?*

Note that a phosphorylation was observed and is displayed on the structure. If you select the desired peptide (RAPSVANVGSHCDLSLK) it will be highlighted in blue. *Is it a surprise that the phosphorylation is located on the outside of the protein? Why are not all peptides available on the structure?*

1. GO Analysis

One of the main challenges in results interpretation is the link to the original gene and the functional interpretation of the results. PeptideShaker offers the possibility to conduct a Gene Ontology Enrichment Analysis of your validated proteins. Go to the ‘GO Analysis’ tab and select *Homo sapiens* as the species at the top. A list of GO terms will appear with their prevalence in our dataset when compared to all human genes:



The Gene Ontology Enrichment Analysis (GOEA) analyzes the frequencies of gene ontology terms in your dataset and compares these to the frequencies of the same GO terms in the species specific version of Ensembl (<http://www.ensembl.org>). In order to not get too many terms the GOSlim UniProtKB-GOA (<http://www.ebi.ac.uk/GOA>) is used.

GOEA shows if certain GO terms are found more or less often in your dataset compared to the distribution of the same terms in Ensembl. To calculate the significance PeptideShaker employs a Hypergeometric test (<http://en.wikipedia.org/wiki/Hypergeometric_test>). This divides the GO terms into three groups:

* Significantly higher in dataset (colored green).
* Significantly lower in dataset (colored red).
* Not significantly different (colored grey).

*Which GO terms are significantly more frequent in this dataset compared to the frequency in Ensembl? And which are significantly less frequent?*

*Does it always make sense to compare against the whole of Ensembl? What happens if we have a biased selection of proteins as input to the analysis? Can the results then be trusted?*

References

(1) Vizcaino, J. A.; Mueller, M.; Hermjakob, H.; Martens, L. Charting online OMICS resources: A navigational chart for clinical researchers. *Proteomics Clin Appl* **2009**, *3*, 18.

(2) 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.

(3) 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.