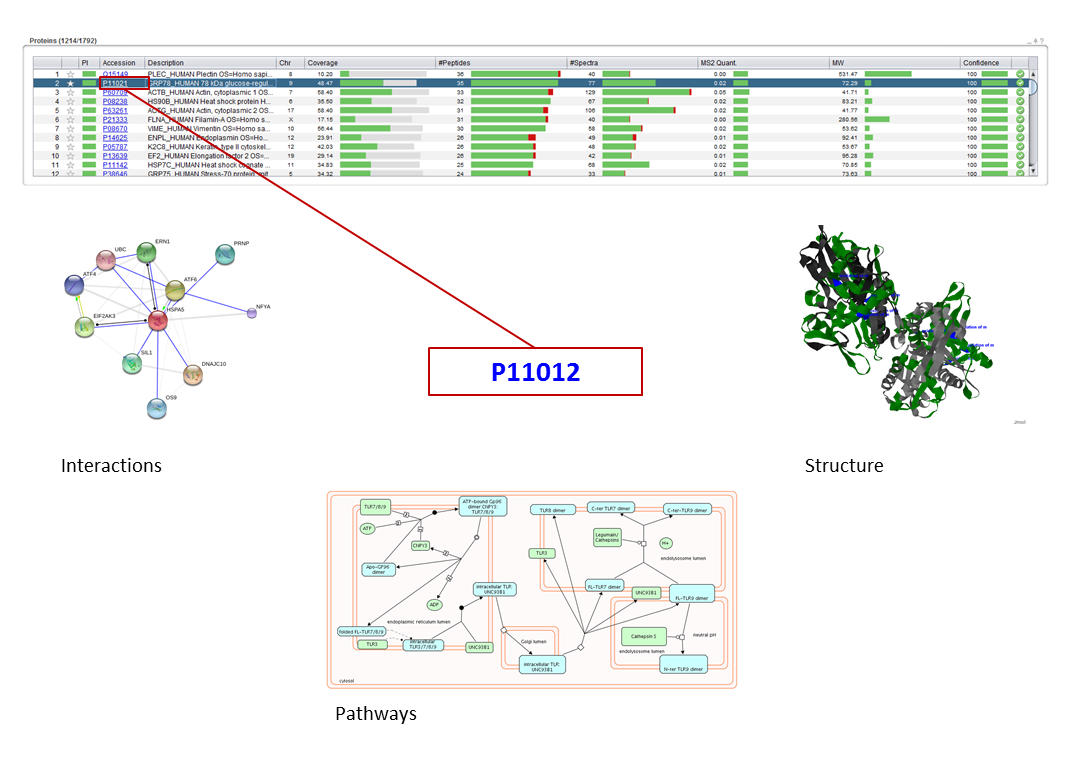
Functional Analysis

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:



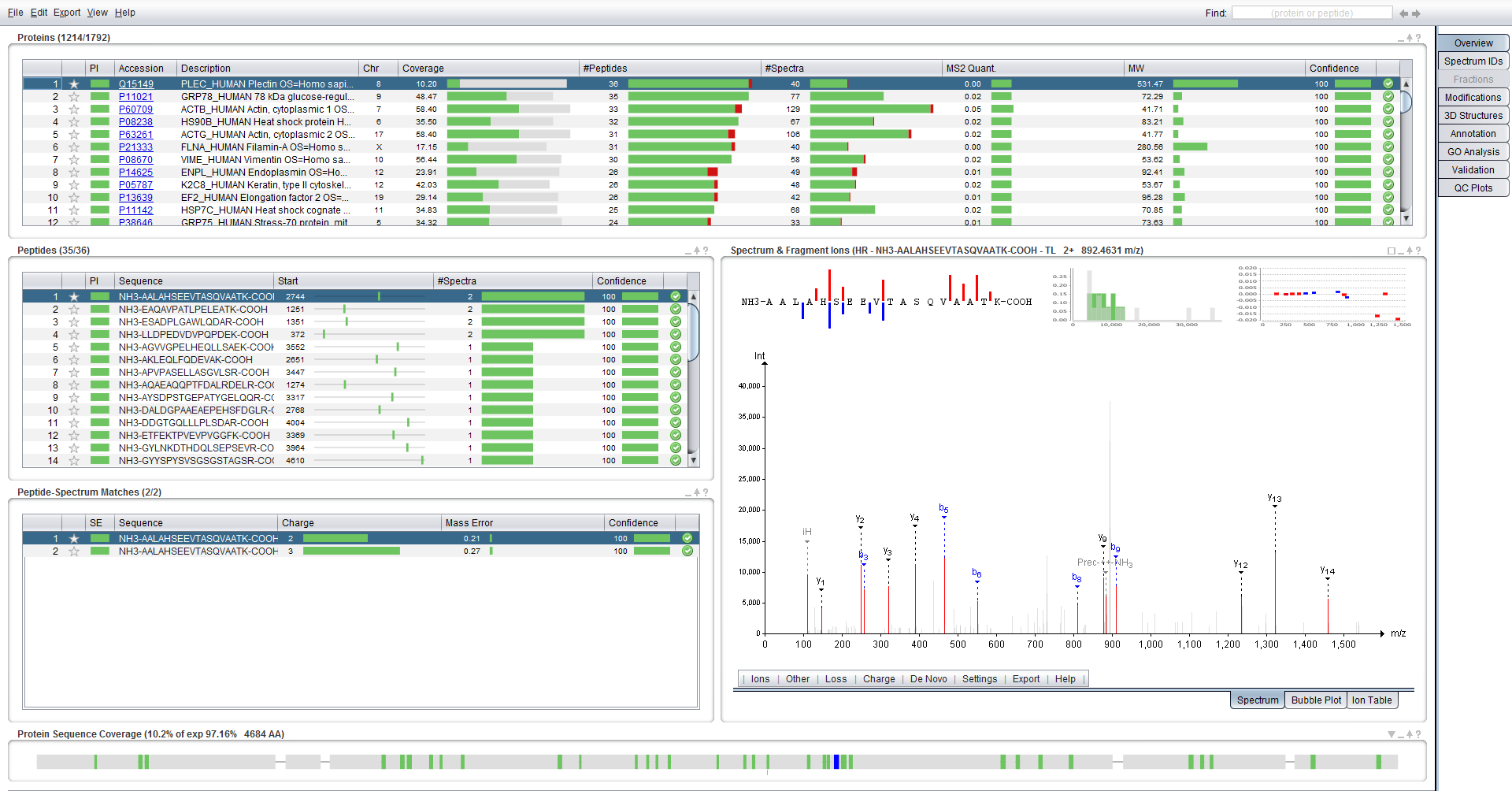
In this chapter we are going to look for further protein information, conduct a pathway analysis, view 3D structures of the proteins and finally inspect the Gene Ontology (GO) of our dataset.

**Tip:**  
*Parts of the tutorial can also be followed without PeptideShaker using the web links.*

Load the example dataset into PeptideShaker. Note that we use human data, the amount of available information is extremely species dependent.

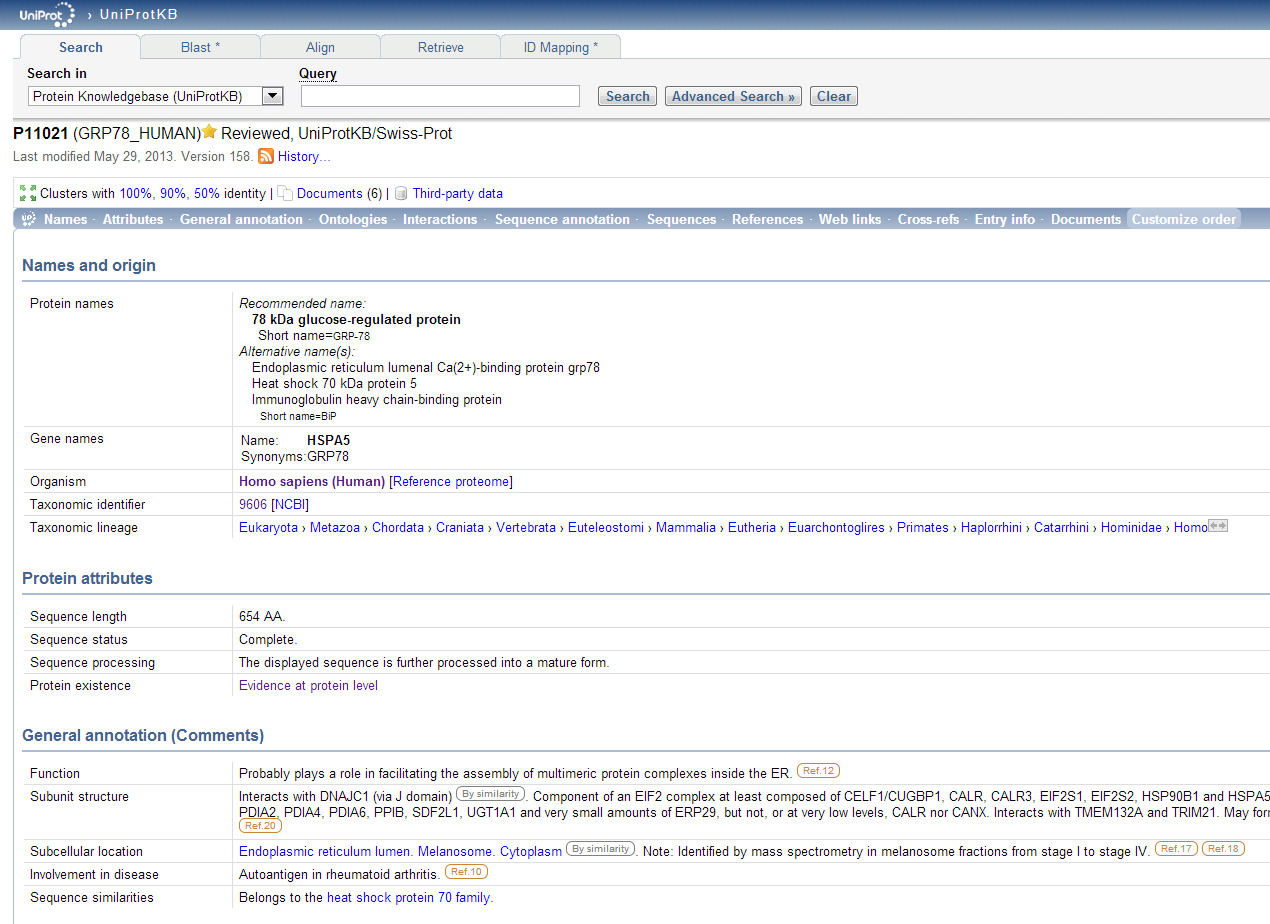
1. Protein Information

You should now have a project with 1,214 proteins validated at 1% FDR:



The first source of protein information is UniProtKB,[2](#_ENREF_2) 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.

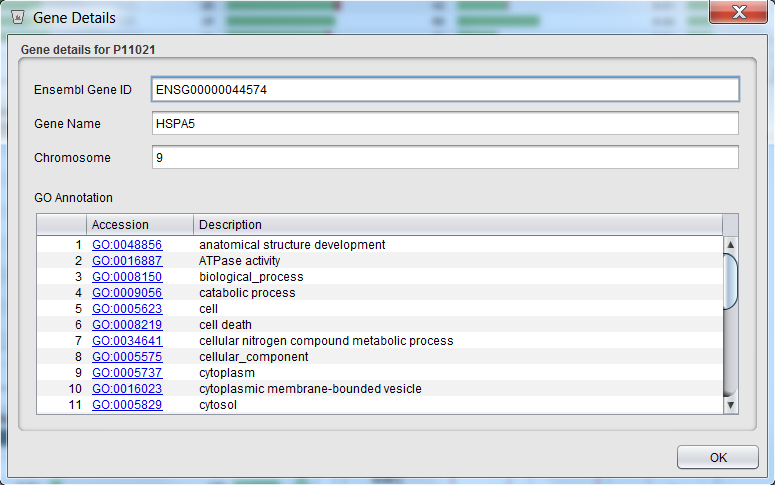
Click on the first accession number in the protein table: P11021. The UniProt entry for this protein is then opened in a web browser:



Scrolling down will show you all available information about this protein. *What is the function of this protein? In which cellular components are we likely to find this protein?*

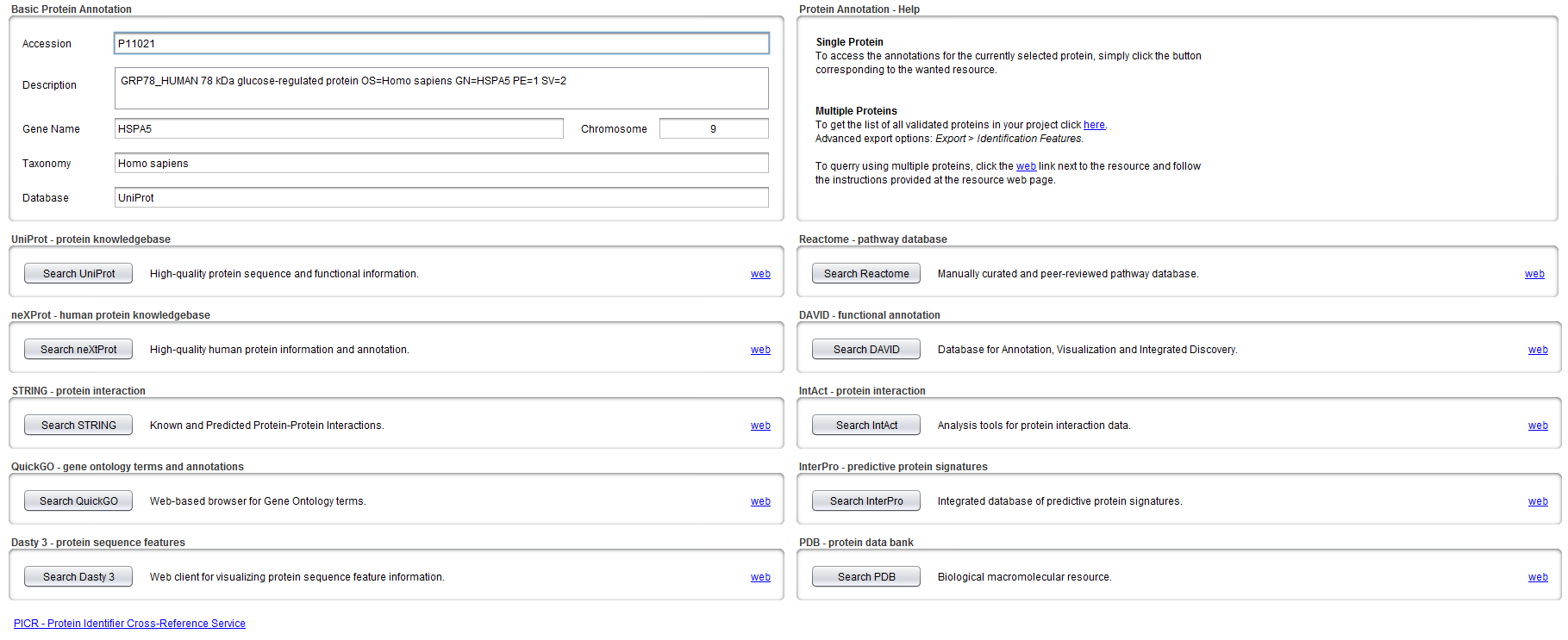
According to the protein attribtures, this protein “Probably plays a role in facilitating the assembly of multimeric protein complexes inside the ER” and was found in these subcellular locations: “Endoplasmic reticulum lumen. Melanosome. Cytoplasm.”. Note that more information is given in the “Ontologies” section of the protein report.

In fact some of this information was readily displayed in the protein table. For instance, you can see the chromosome number attached to every protein according to Ensembl[3](#_ENREF_3" \o "Hubbard, 2002 #2) in the sixth column: chromosome number 9. Click on this number, you should see the following dialog appear:



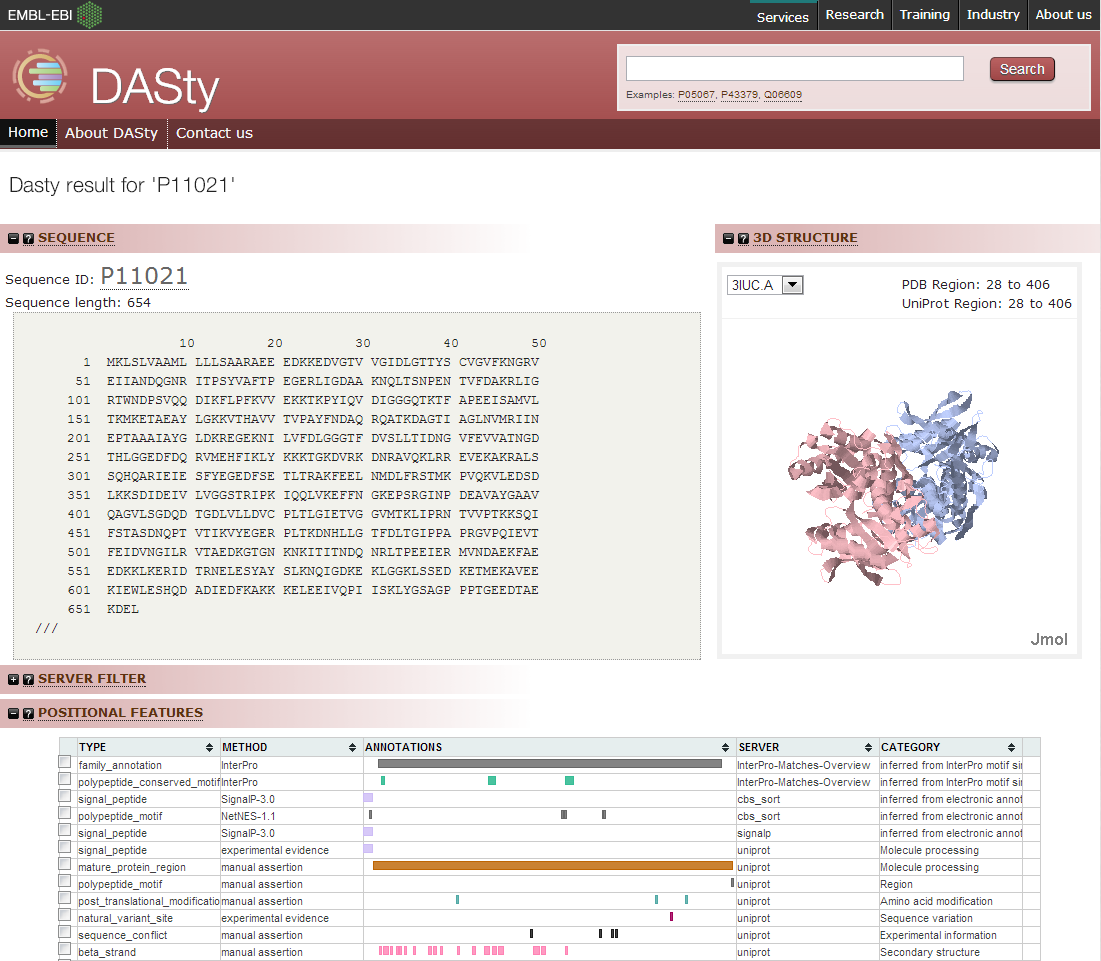
Here you can see the gene name, the chromosome it is attached to and a list of Gene Ontology (GO) terms attached to this entry. The GO terms will be further detailed later in this chapter. Note that the information displayed is strongly affected by the protein inference problem tackled in the “Identification” chapter.

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[4](#_ENREF_4) (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.*

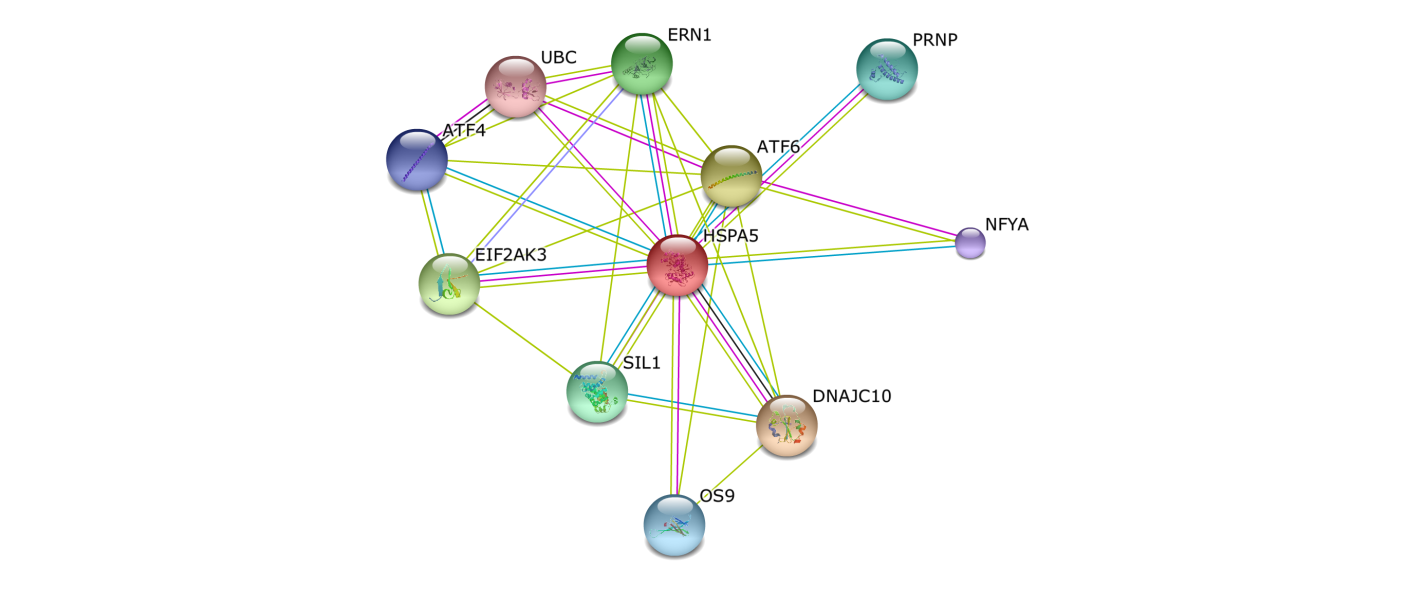
You can use the tool Dasty[5](#_ENREF_5" \o "Jones, 2005 #81) (<http://www.ebi.ac.uk/dasty>), another service from the EBI, to navigate existing UniProtKB information about this sequence. Dasty annotates the sequence of proteins with various features. Click on ‘Search Dasty 3’, your web browser should 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.

You can obtain more details by following the link to 'Intact'. *Note that the link to Intact is also available directly from PeptideShaker.*

STRING provides a graphical interface allowing browsing potein interactions in a more user friendly way. Click on the ‘Search STRING’ button in PeptideShaker and then on the ‘GO!’ button in the STRING interface:



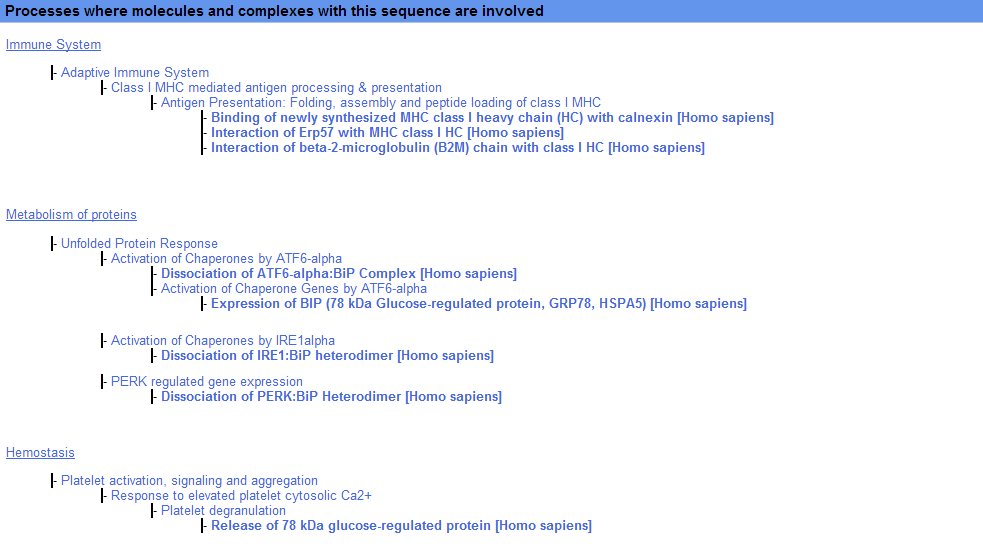
*What are the predicted functional partners? How are they inferred?*

A table lists all known possible partners inferred in this case from experiment, databases and text mining. Note that these interaction inference methods are not of the same trustfulness.

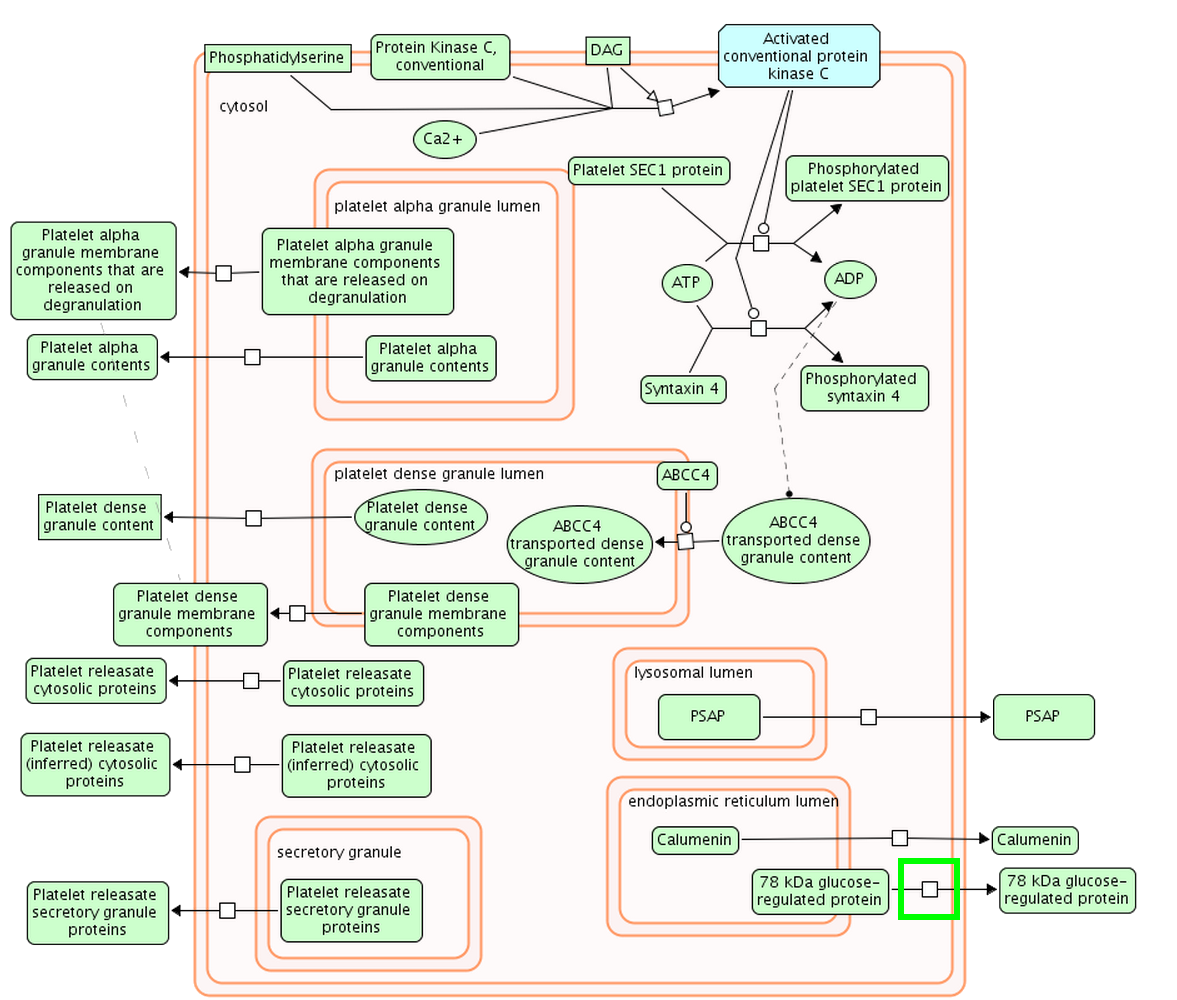
1. Pathways

In order to get a more functional insight of our experiment, we will browse the biological pathways it covers. Reactome[6](#_ENREF_6" \o "Haw, 2011 #3) is a manually curated and peer-reviewed pathway database. It allows us to see in which pathways our proteins are involved. Here again, 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:



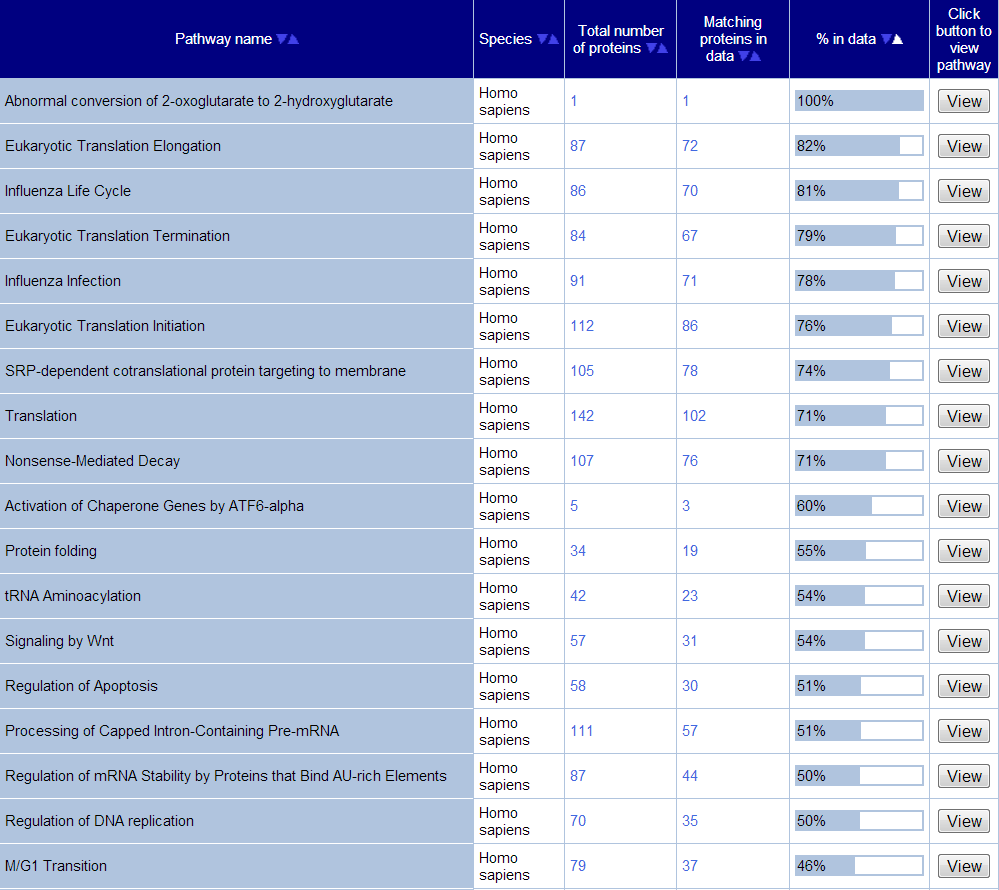
If you click on “Release of 78 kDa glucose-regulated protein” you will be able to browse the pathway and its main components:



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:

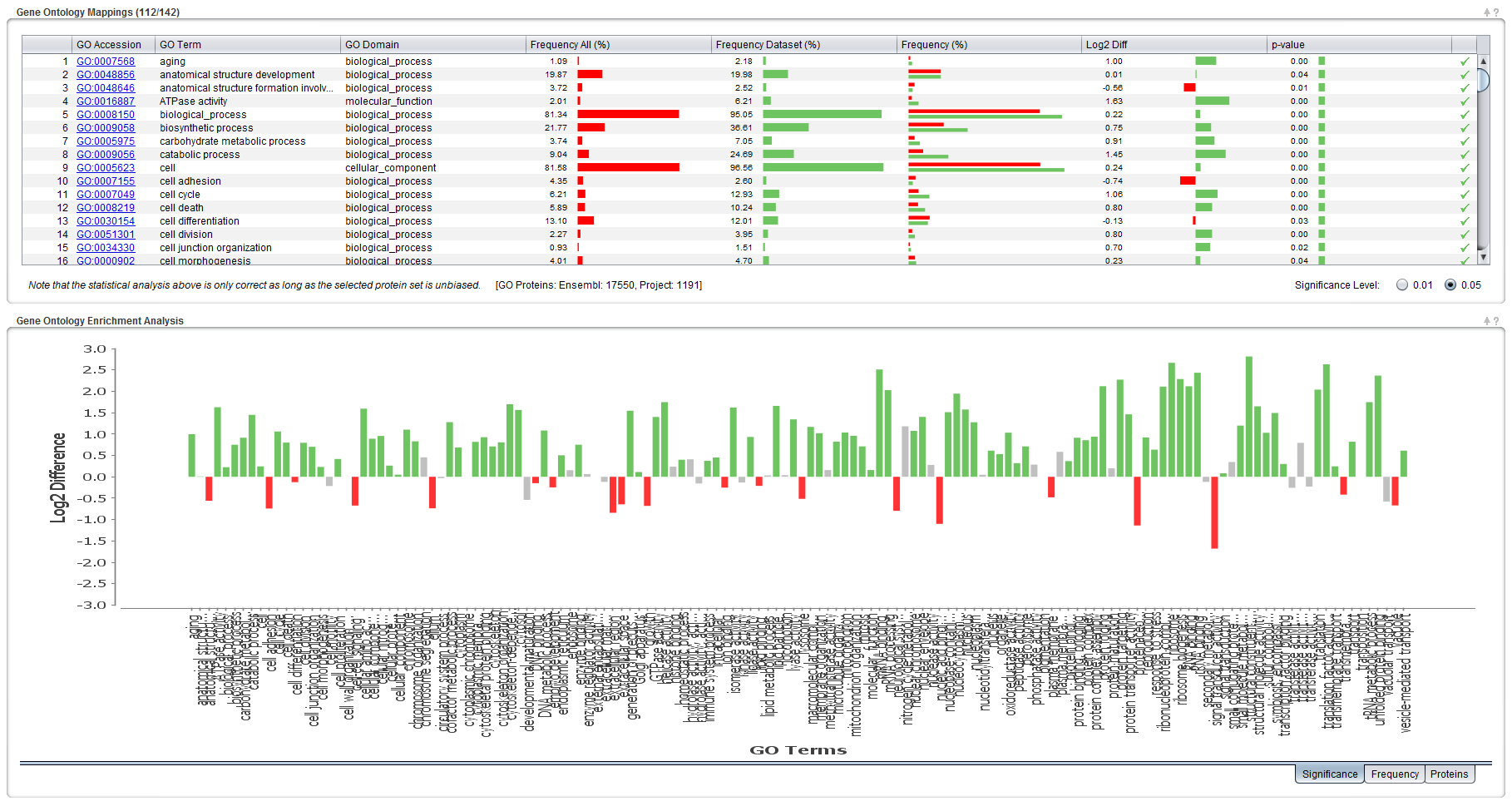


Note that the pathways covered are closely linked to platelets. *Why don’t we cover all pathways completely?*

It is very rare to cover a pathway fully, and most often impossible. Indeed pathways also contain molecules like ADP which are not detected in proteomics experiments. Moreover, it can happen that an isoform of a given protein is expected whether we identify another. Here again, the protein inference problem is impairing our ability to map our data to external resources.

1. GO Analysis

As we already saw in the protein table, it is possible to link 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, 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:

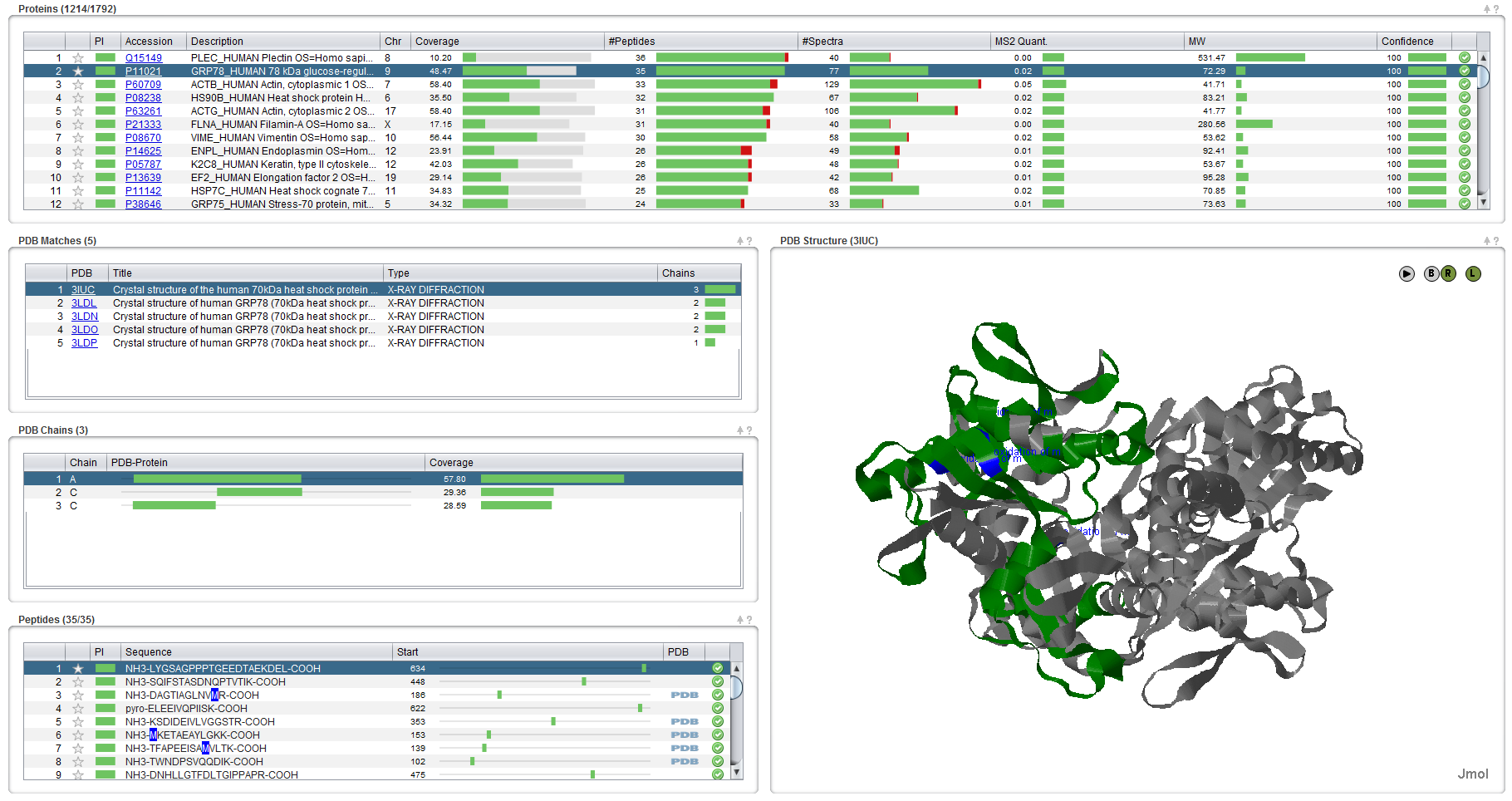
* Significantly higher in dataset (colored green).
* Significantly lower in dataset (colored red).
* Not significantly different (colored grey).

*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?*

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. Go to the ‘3D Structures’ tab in PeptideShaker, make sure that the protein P11021 is selected and chose the first available structure ‘3IUC’, you should see this:



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

There are different structures inferred by different methods. Also, the mapping between the structure database and the sequence is not always perfect. Often, there is simply no structure available.

Note that the identified peptides are displayed in green and modifications are mapped on the structure. If you select a peptide, it will be highlighted in blue. *Why are not all peptides available on the structure?*

Here again, the sequence database and structure database do not fully overlap.

References

1. Vizcaino, J.A., Mueller, M., Hermjakob, H. & Martens, L. Charting online OMICS resources: A navigational chart for clinical researchers. *Proteomics Clin Appl* **3**, 18-29 (2009).

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

3. Hubbard, T. et al. The Ensembl genome database project. *Nucleic acids research* **30**, 38-41 (2002).

4. Cote, R.G. et al. The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC bioinformatics* **8**, 401 (2007).

5. Jones, P. et al. Dasty and UniProt DAS: a perfect pair for protein feature visualization. *Bioinformatics* **21**, 3198-3199 (2005).

6. Haw, R., Hermjakob, H., D'Eustachio, P. & Stein, L. Reactome pathway analysis to enrich biological discovery in proteomics data sets. *Proteomics* **11**, 3598-3613 (2011).