Analyze Protein-Protein Interactions with InSyBio Interact

InSyBio
Intelligent Systems Biology

User Manual

July 2024

Insybio Suite v3.3

www.insybio.com

Introduction

Interact is an intracellular interaction analysis tool for the

- analysis of proteins
- analysis of protein interactions
- prediction of proteins' functionality

Proteins are the functional components of many cellular processes and the identification of their physical Protein-Protein Interactions (PPIs) is an area of mature academic research.

InSyBio Interact enables users to combine a variety of sequential, structural and functional information using a high performance machine learning technique to predict a confidence score for each protein pair. This confidence score indicates the probability of these two proteins to interact and the strength of this interaction. Users are able to tune this confidence score and extract their own datasets as well as obtain only positives or negative interactions. Moreover, InSyBio Suite also provides visualization and meta-analysis of the Protein Interaction Networks. The protein interaction networks are clustered using a novel methodology to extract information about the protein complexes which the proteins form to perform specific tasks. These complexes and the overall PPI networks may be visualized using an interactive visualization module.

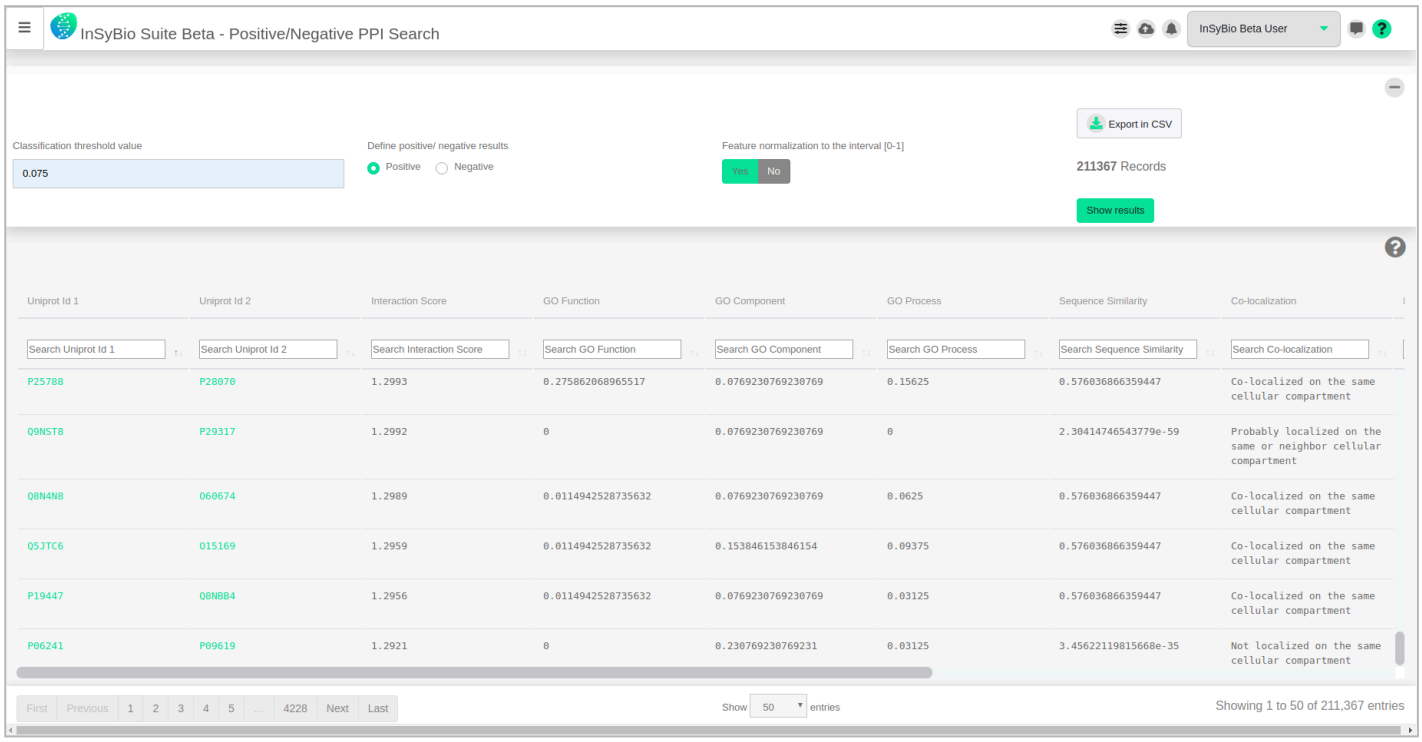
With Insybio Interact you can:

a) Search Positive/Negative PPIs
b) Retrieve information about a specific protein
c) View and download the list of clusters created by the EEMC algorithm
d) Search proteins and Complexes with a GO Term
e) Create and Save your own PPI dataset
f) Create a network from your set of biomarkers (genes, transcripts, proteins)
g) Perform functional enrichment analysis on your set of biomarkers
h) Annotate functionally a specific protein using a patent-pending method [Theofilatos, K., Dimitrakopoulos, C., Mavroudi, S., Korfiati, A. and Alexakos, C., Insybio Ltd, 2017.

Protein functional and sub-cellular annotation in a proteome. U.S. Patent Application 15/361,461.]

Positive/Negative PPIs

You can search for positive and negative Human Protein-Protein Interaction Data by specifying a threshold referring to the PPI confidence score.



You can search by choosing:

- The classification threshold value
- Positive or negative PPIs
- The application of normalization to the features to the interval [0-1]

The results are presented on your screen in a browse-able table or you can download them as a TAB delimited CSV file.

For each PPI the following are presented:

- The Uniprot IDs of the two proteins,

- 22 informative sequential, structural and functional features,
- Presence in the APID database,
- The protein-protein interaction score - a confidence score indicating the probability of a protein pair to be an interacting one and the strength of this interaction.

For every protein pair 22 informative features are calculated and provided including sequential, structural and functional features:

- **Gene Ontology (Co-function, Co-process, Co-localization)** (3 features). Gene Ontology database contains information of three types that include:
    - Molecular function of a gene product,
    - Biological process in which the gene product participates,
    - Cellular component where the gene product acts

      In this features the number of common GO terms existing in a protein pair are calculated.
- **Sequence Similarity** (1 feature). This feature was obtained using the NCBI-BLAST+ standalone executable BLASTP to perform the human to human protein alignment. All BLASTP hits obtained with the default parameters and the sequence alignment E-values of each protein-protein pair have been used to fill in the specific feature.
- **Homology Based PPI** (1 feature). For each human protein pair we have found its homology pairs in yeast by using the 0.001 cutoff on the E-value taken by similarity tests that were conducted using BLASTP. If at least one of the resulting homology pairs is referred as interacting in the Yeast organism's DIP dataset, a value equal to “1” is assigned to this feature, otherwise “0”.
- **Gene Expression Profile Similarity** (15 features). Fifteen gene expression datasets were used to estimate 15 gene expression profile similarity features. The datasets selected are extracted from NCBI Gene Expression Omnibus (GDS531, GDS534, GDS596, GDS651, GDS806, GDS807, GDS843, GDS987, GDS1085, GDS2855, GDS1402, GDS181, GDS1088, GDS841, GDS3257). For each dataset, the expression profile of each protein of every protein-pair was extracted and the Pearson correlation between the two profiles was estimated.

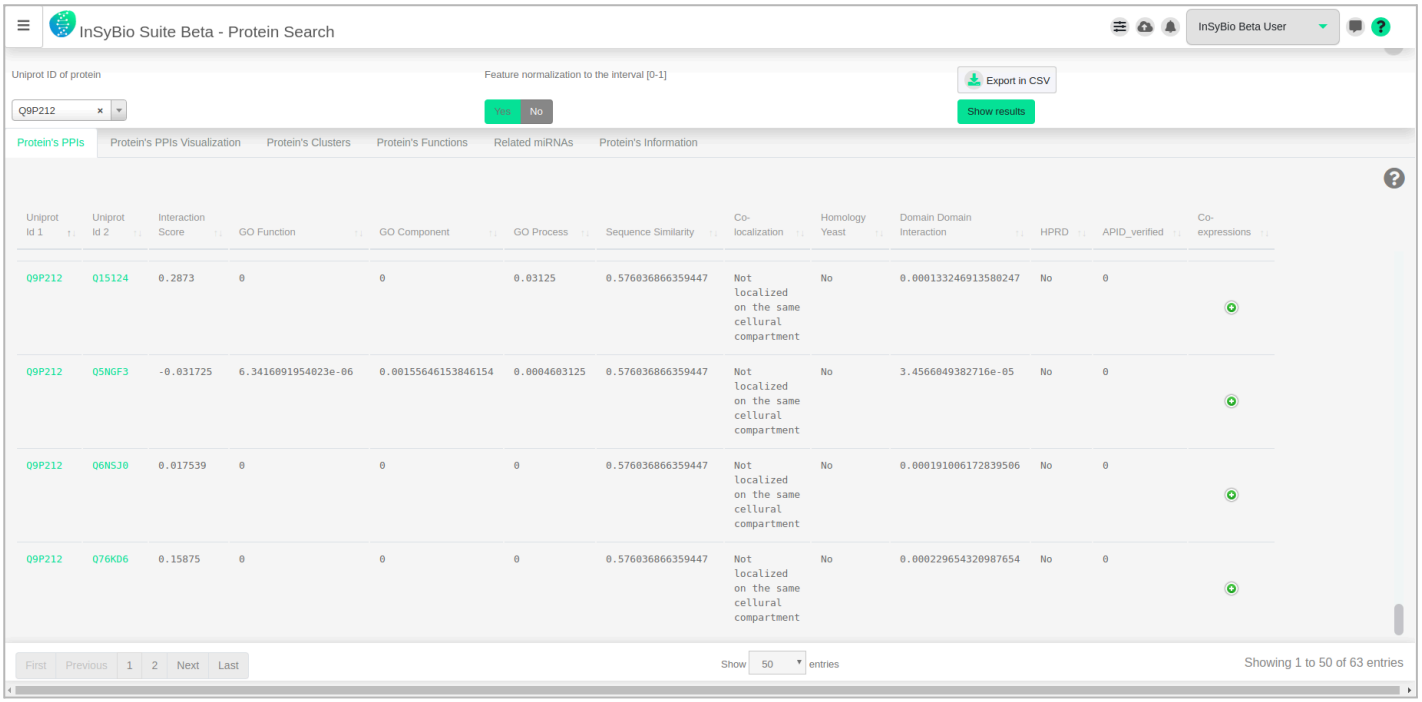
- **Co-localization** (1 feature). PLST tool (Scott et al, 2004) was used to predict the local cellular compartments in which each protein is likely to function. For every pair of proteins, a feature is calculated by taking the value of 1 if the two proteins share at least one local compartment and the value of 0 if they do not have any common local cellular compartment.
- **Domain-Domain Interactions** (1 feature). All possible domain pairs (downloaded from InterPro) were evaluated using the training set and the hypergeometric distribution to locate pairs that are interacting with high probability (p-values less than 0.05). Then for every protein pair, their domain combinations are computed and the number of the interacting ones is measured and used as the feature' s value.

Protein Search

You can extract information about a specific Human protein by providing its UNIPROT identifier.

You can:

a) Request the PPIs that are related to the specific protein and download the results;



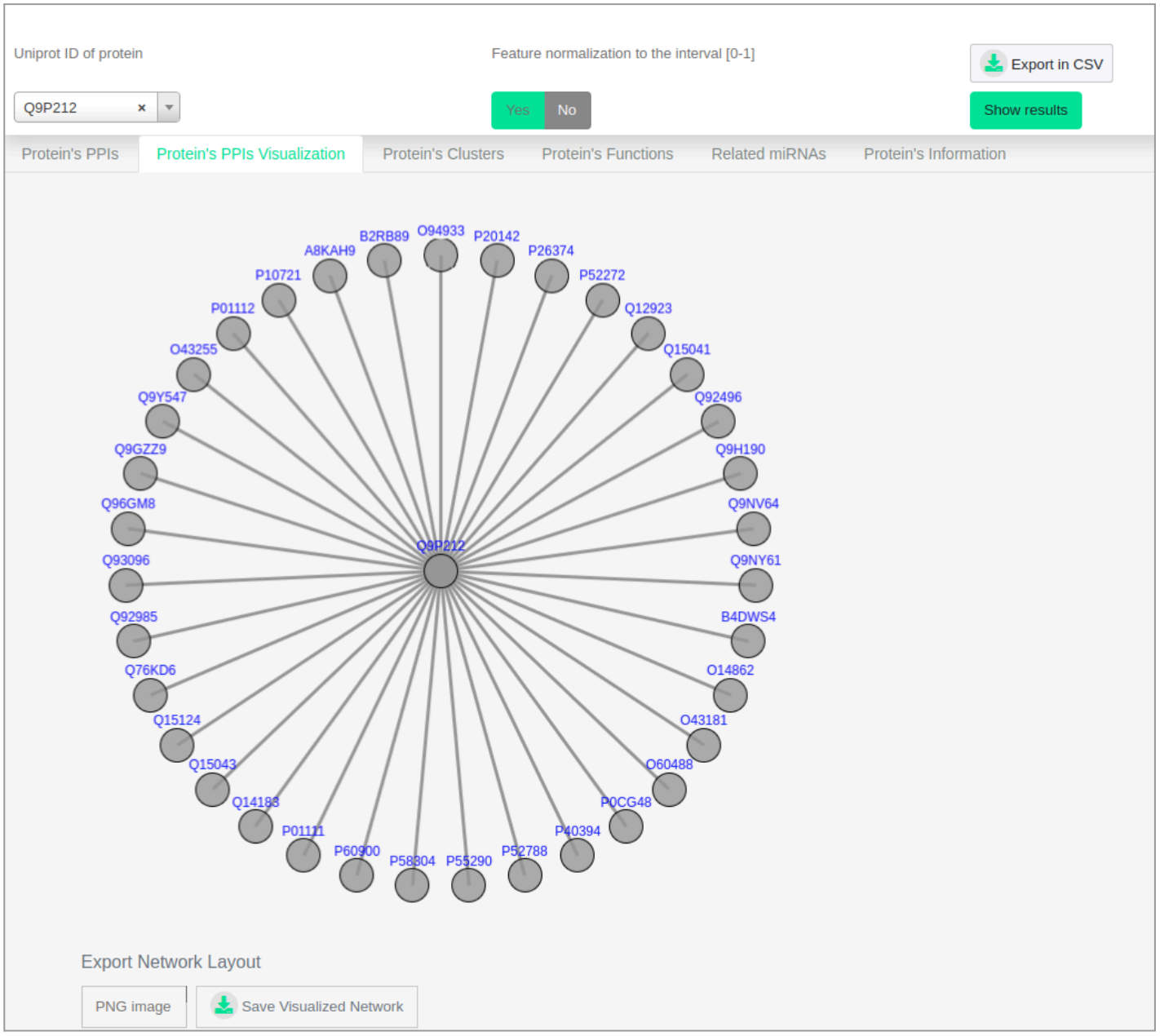
The results are presented on your screen in a browse-able table or you can download them as a TAB delimited CSV file.

For each PPI the following are presented:

- The Uniprot IDs of the two proteins,
- 22 informative sequential, structural and functional features,
- Presence in the APID database,

- The protein-protein interaction score - a confidence score indicating the probability of a protein pair to be an interacting one and the strength of this interaction.

b) Visualize the protein interaction network related to the specific protein and download it in various graph-representation file formats;



The result is an interactive graph displaying the first or the second neighborhood of the protein of interest (after user choice).

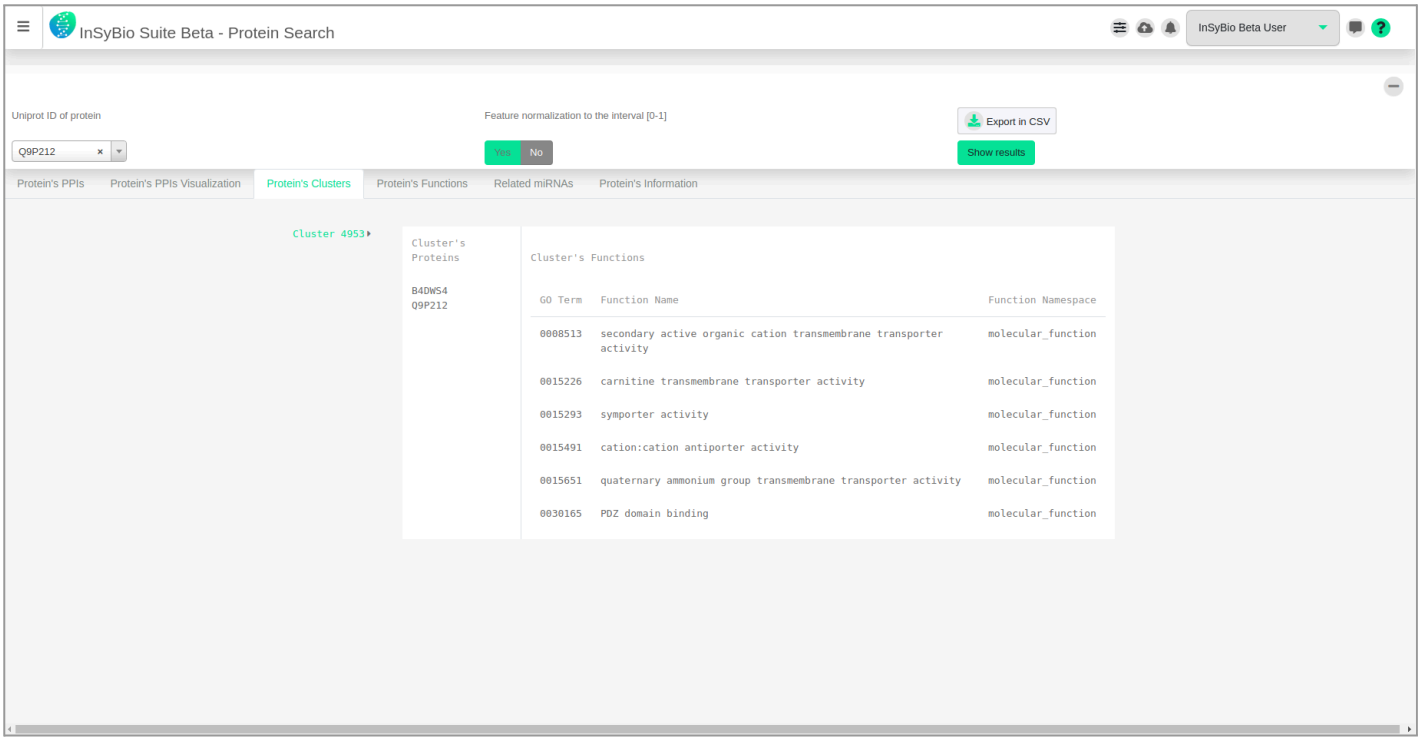
Clicking on a node-protein or on an edge-interaction, the respective information is presented.

Right clicking on a node-protein, its first neighborhood is presented.

There are options for decreasing opacity on mouseover and changing layout (force directed, radial, or circle).

The graph can be saved in various formats (PNG image, SVG image, PDF file, XGMML document, GraphML document, or Simple Interaction Format (SIF) document).

c) Find the clusters related to the specific protein;



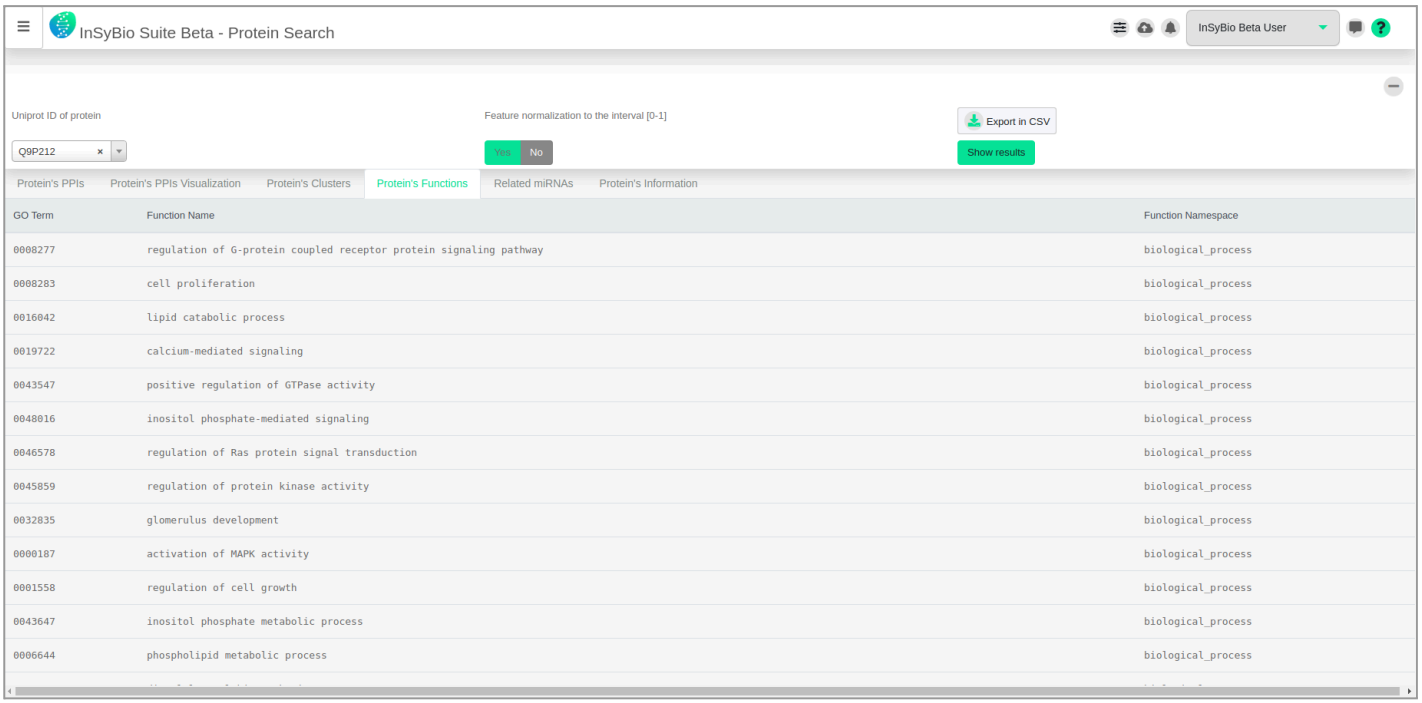
The result is a list of the clusters the protein of interest takes part in.

Clicking on a cluster, the following information is presented:

- The proteins that form the cluster,

- A table of the cluster's functions.

d) List all the functions related to the specific protein;

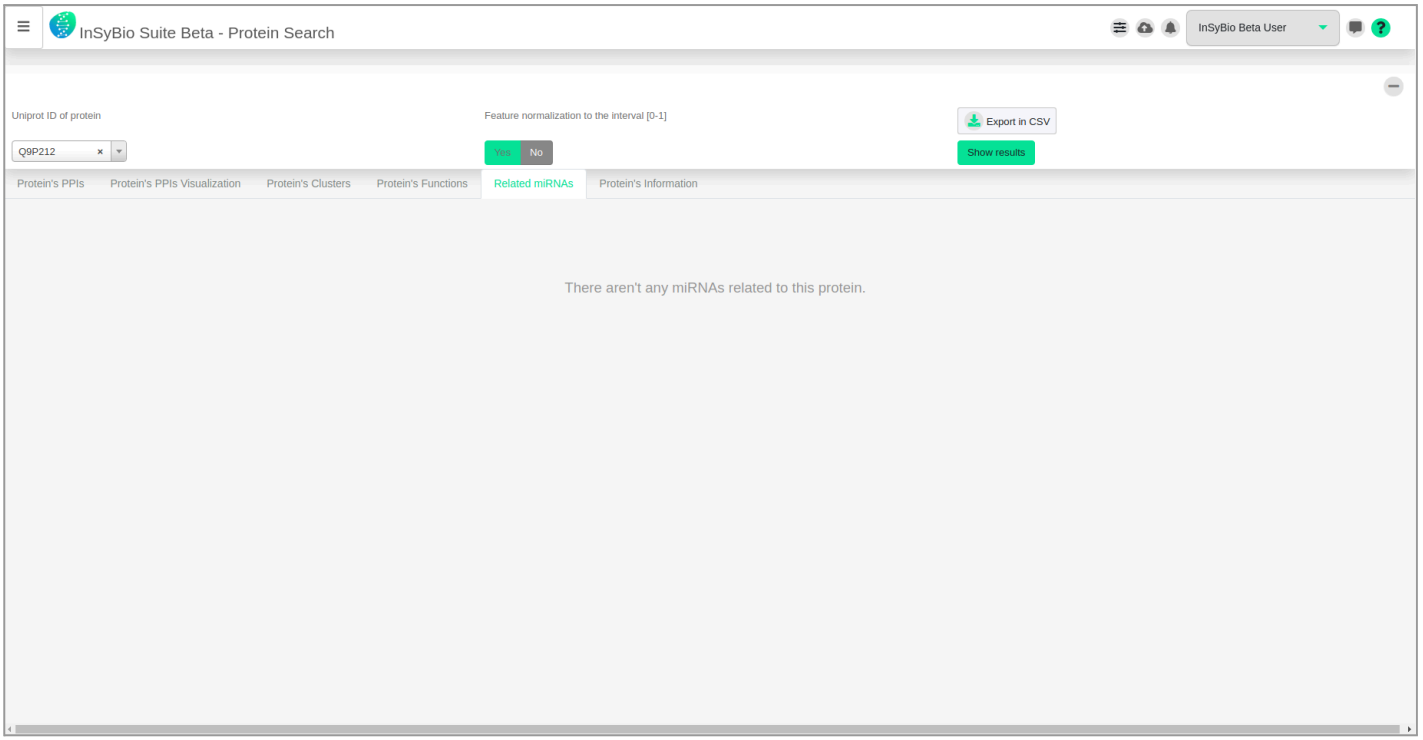


The result is a table with the functions that the protein of interest has.

For each function, the following information is presented:

- The function's GO term,
- The function's name, and
- The function's namespace.

e) List all the miRNAs related to the specific protein;

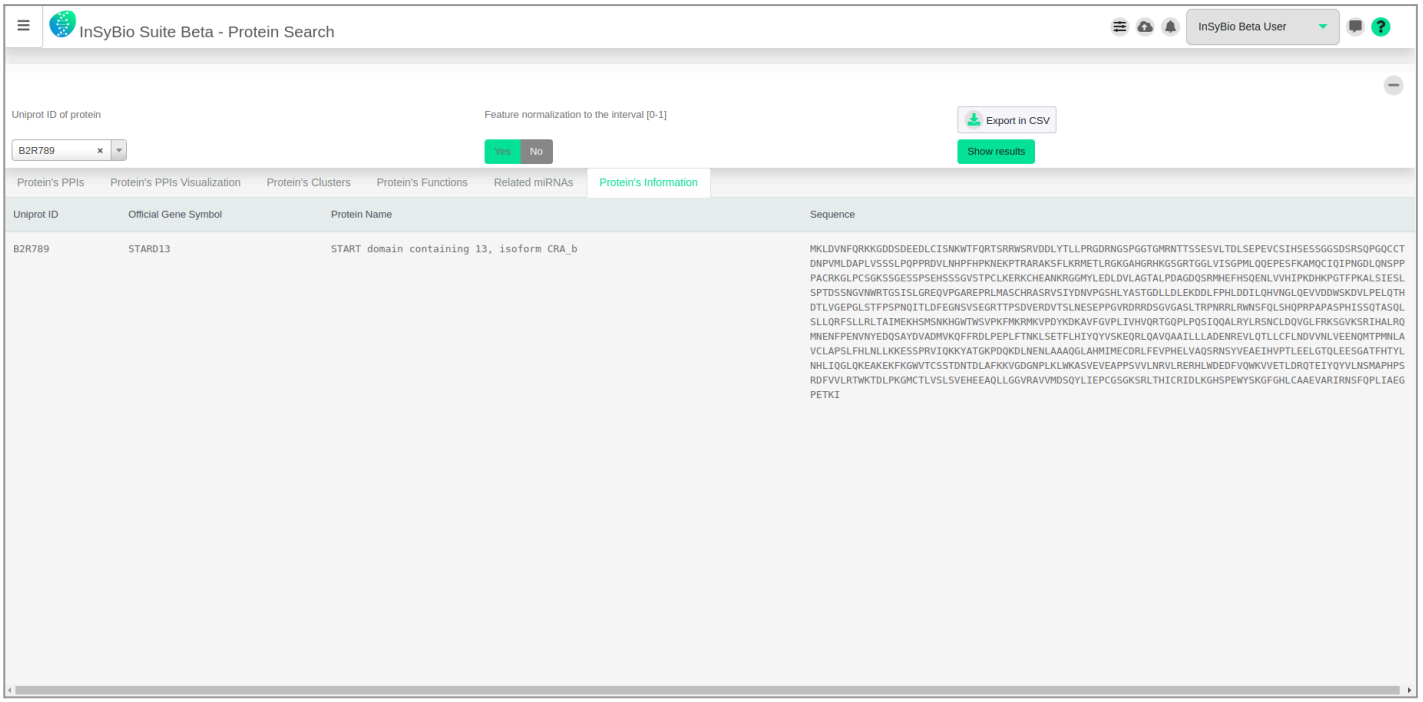


The result is a table with the miRNAs that regulate the expression of the protein of interest.

For each miRNA, the following information is presented:

- The miRNA's accession number, and
- The miRNA's id.

f) Find the protein's information;



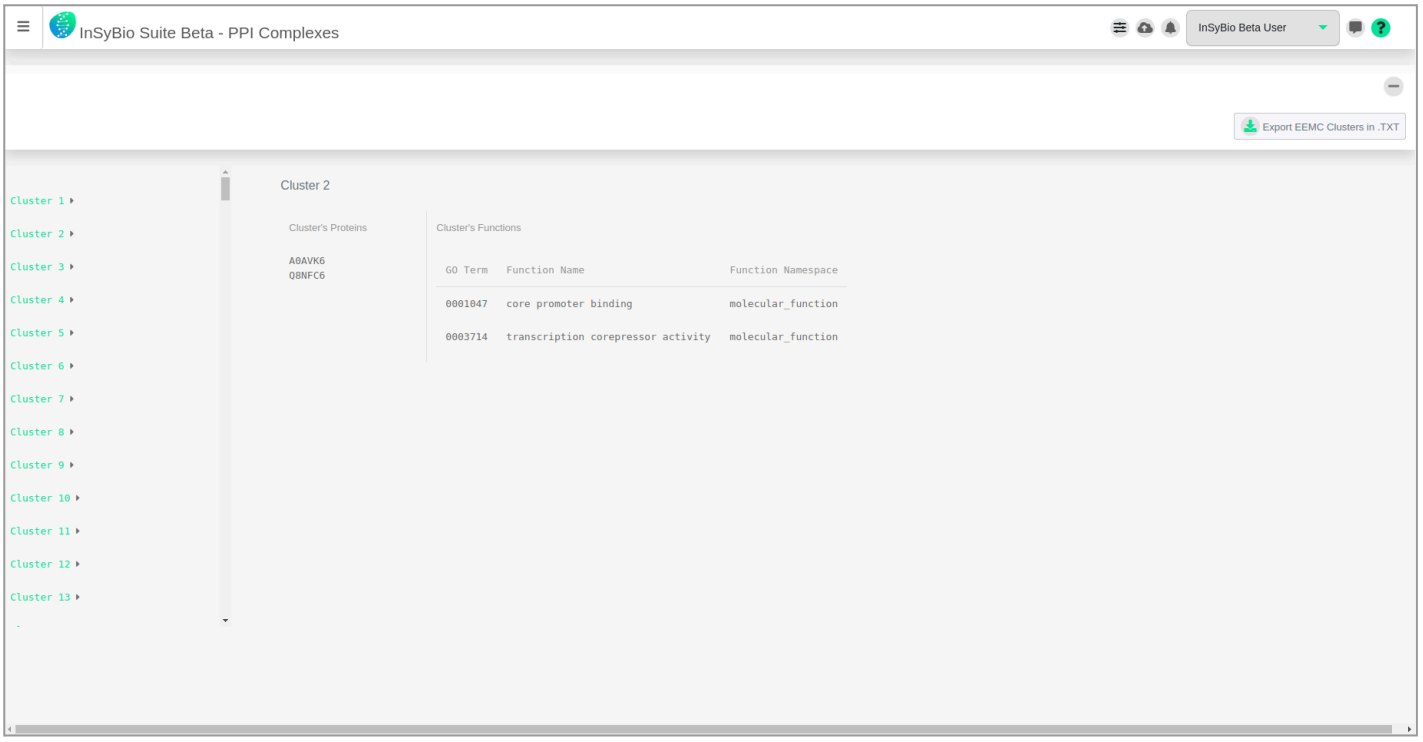
The result is a table containing information about the protein of interest.

For each protein, the following information is presented:

- The uniprot ID,
- The official gene symbol
- The protein name, and
- The protein's sequence.

PPI Complexes

You can view a list of all the clusters computed from the algorithm EEMC. EEMC (Evolutionary Enhanced Markov Clustering) is a novel fully unsupervised methodology that InSyBio team has proposed for the prediction of protein complexes from weighted PPI graphs. It is a hybrid combination of an adaptive evolutionary algorithm and a state-of-the-art clustering algorithm named Enhanced Markov Clustering. When applied to new human datasets EEMC' s performance was encouraging in the prediction of protein complexes which consist of proteins with high functional similarity. From EEMC' s analysis, we suggest new potentially true Human protein complexes which should be further validated using experimental techniques.



The result is a list of the clusters calculated using the EEMC algorithm. Clicking on a cluster, the user can be informed about the proteins that form the cluster and the cluster's functions.

The entire set of clusters can be downloaded in a txt file with the following format:

```
cluster_id,proteins,go_terms
1,A0AV96 Q9C0H2,0005254
2,Q8NFC6 A0AVK6,0001047 0003714
3,A0AVT1 Q7Z6D5,0019780 0016874
...
```

PPI Functions

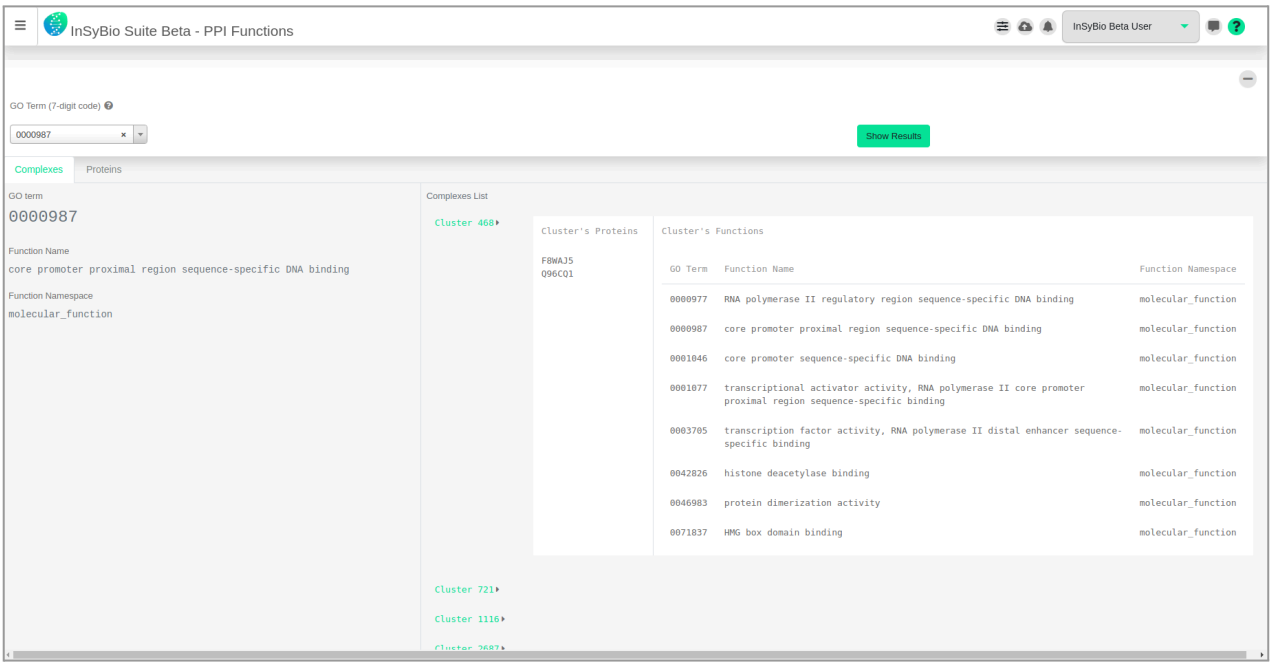
You can derive the details of your function of interest by providing its GO-term.

For each function, the following information is presented:

- The function's GO term,
- The function's name, and
- The function's namespace.

Selecting the respective tab, you can also view the related

a) Complexes;

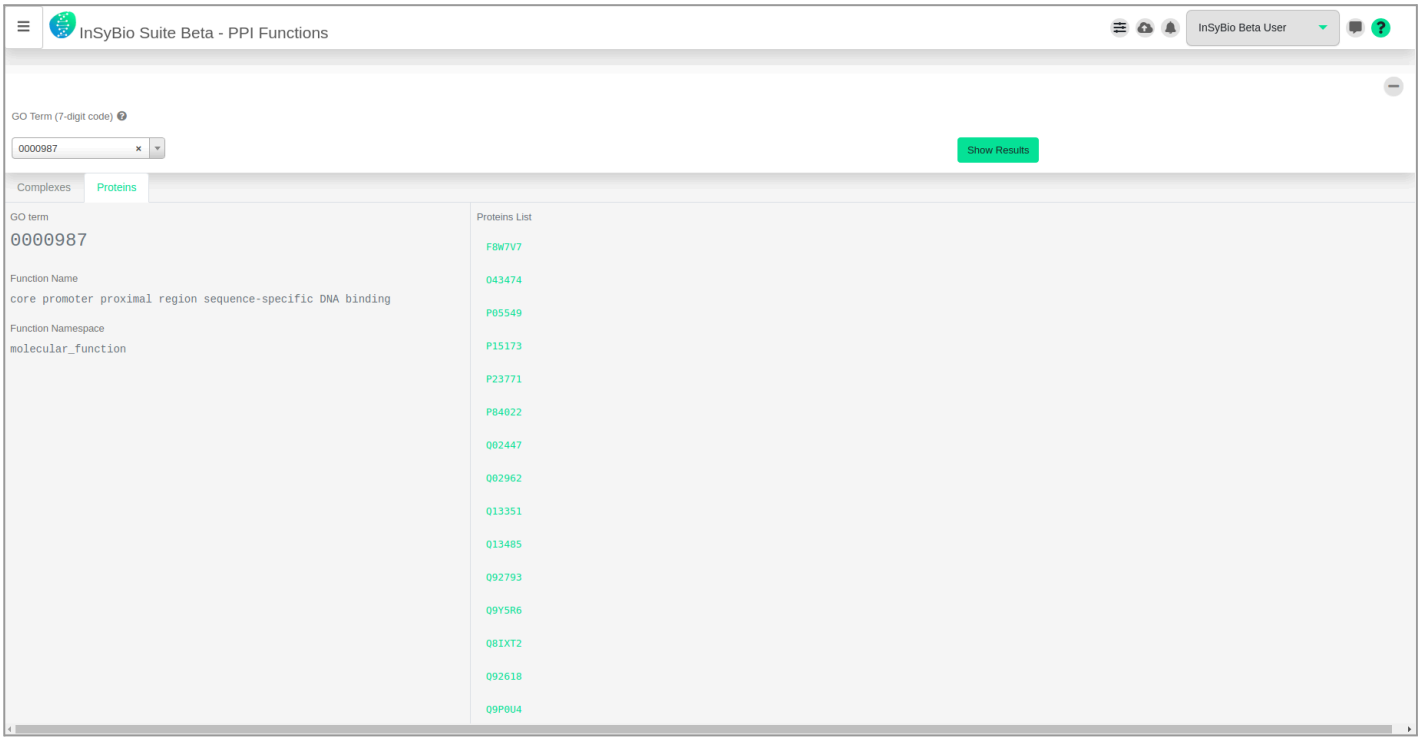


The result is a list of the complexes having the function of interest.

Clicking on a cluster, the following information is presented:

- The proteins that form the cluster, and
- A table of the cluster's functions.

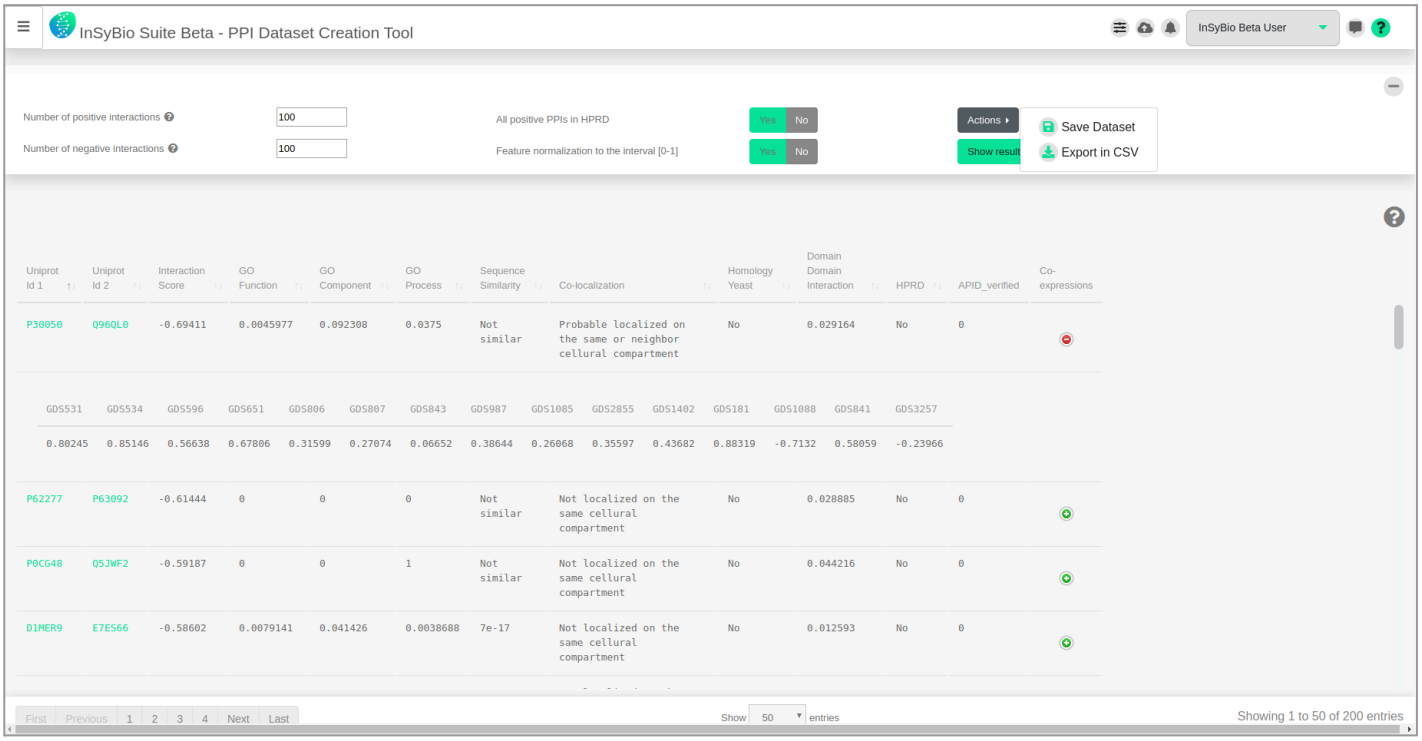
b) Proteins;



The result is a list of the proteins having the function of interest. Clicking on a protein the user will be guided to the protein's page.

PPI Datasets

You can create your own datasets by using a simple form to choose the number of positive and negative examples.



There are options for:

- All positive PPIs being in HPRD or not,
- The application of normalization to the features to the interval [0-1].

The results are presented on your screen in a browse-able table or you can download them as a TAB delimited CSV file through the Actions Button. The dataset created can also be saved in InSyBio Data Store for future use.

For each PPI the following are presented:

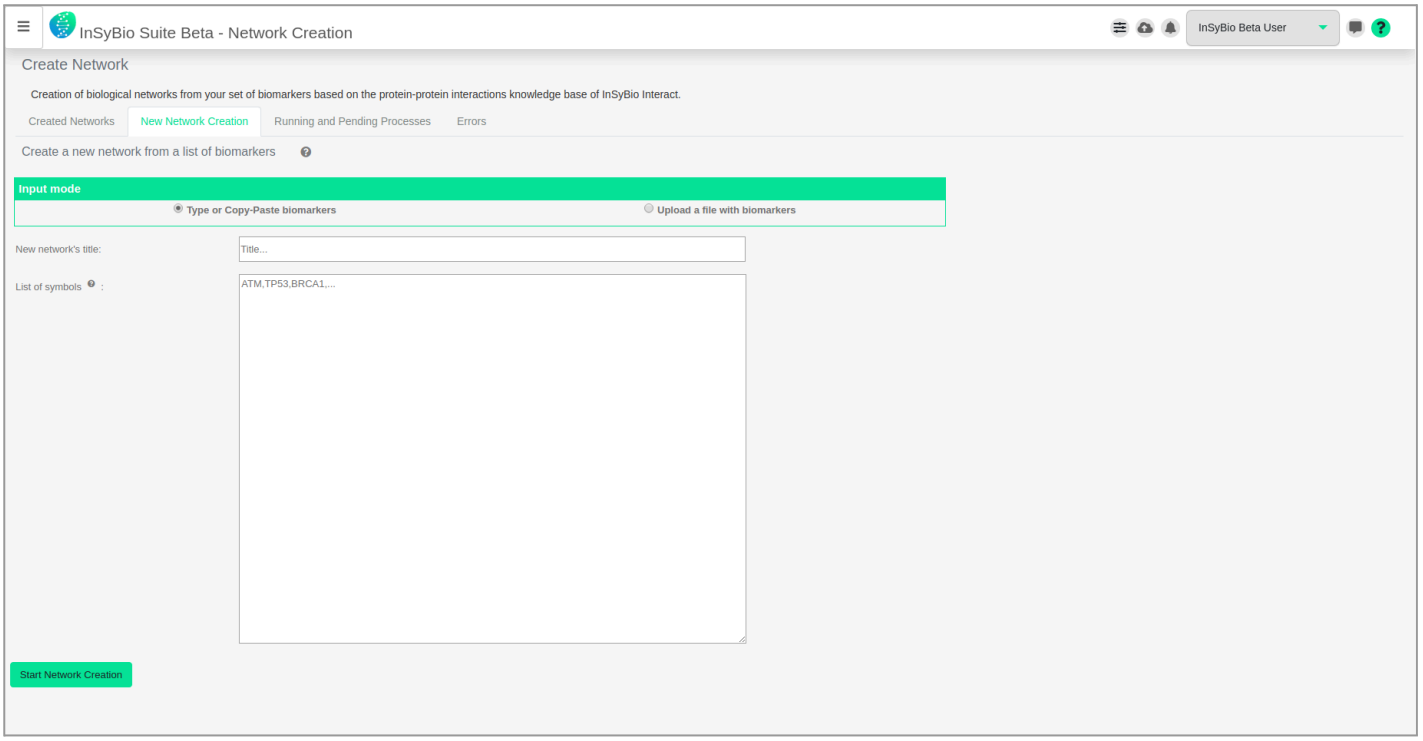
- The Uniprot IDs of the two proteins,

- 22 informative sequential, structural and functional features,
- Presence in the APID database,
- The protein-protein interaction score - a confidence score indicating the probability of a protein pair to be an interacting one and the strength of this interaction.

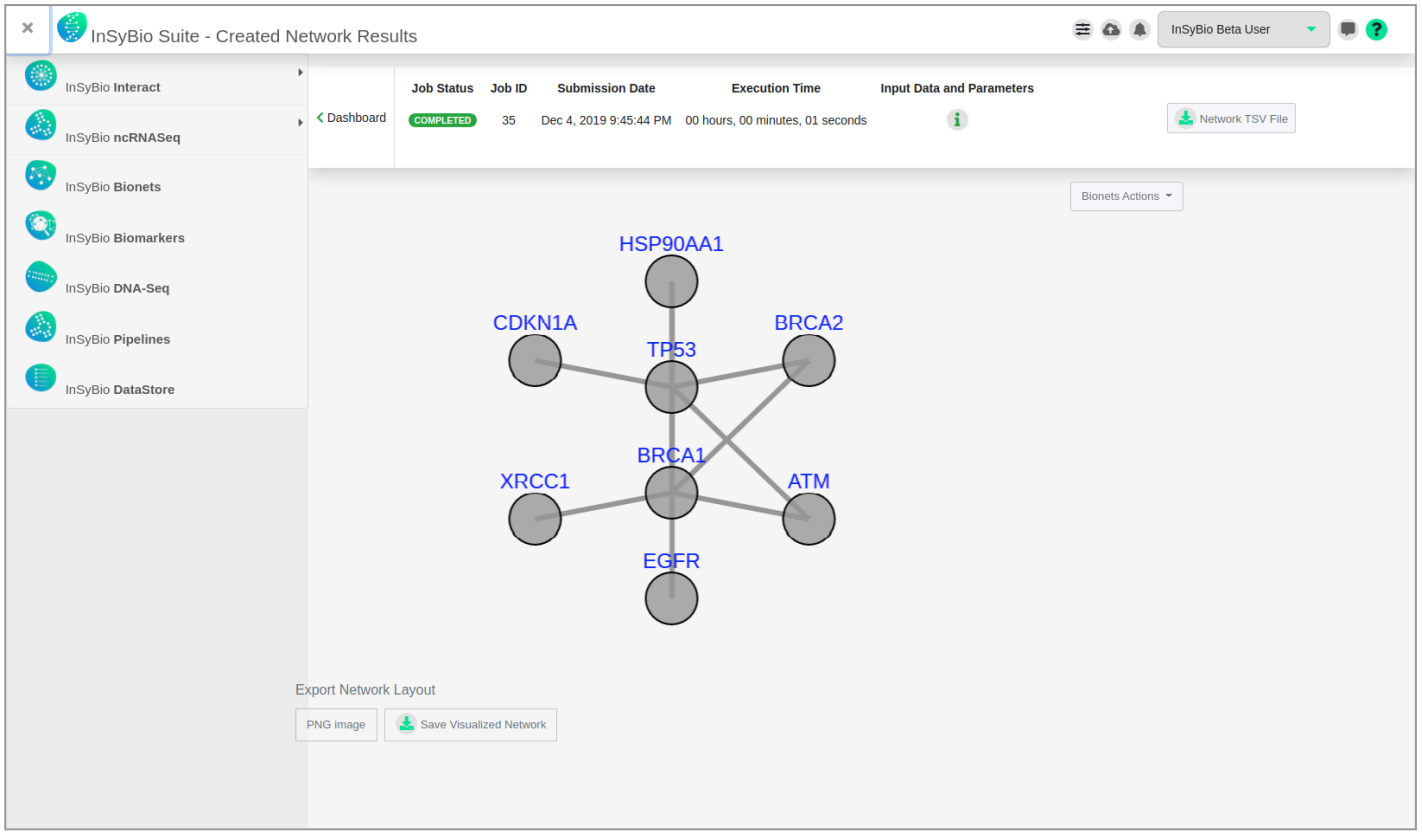
Network Creation

You can create and view biological networks from your set of biomarkers based on the protein-protein interactions knowledge base of InSyBio Interact.

In the tab New Network Creation you can either Type or Copy-Paste the biomarkers or Upload a file with biomarkers. You can use any known identifier for denoting your biomarkers: Uniprot IDs, gene symbols, RefSeq_id and so on. Mixed identifiers are not supported! Note that each symbol should be in one line or separated by comma.



After the network creation, the job will appear in the Created Networks tab. By clicking View Results a new page is opened.



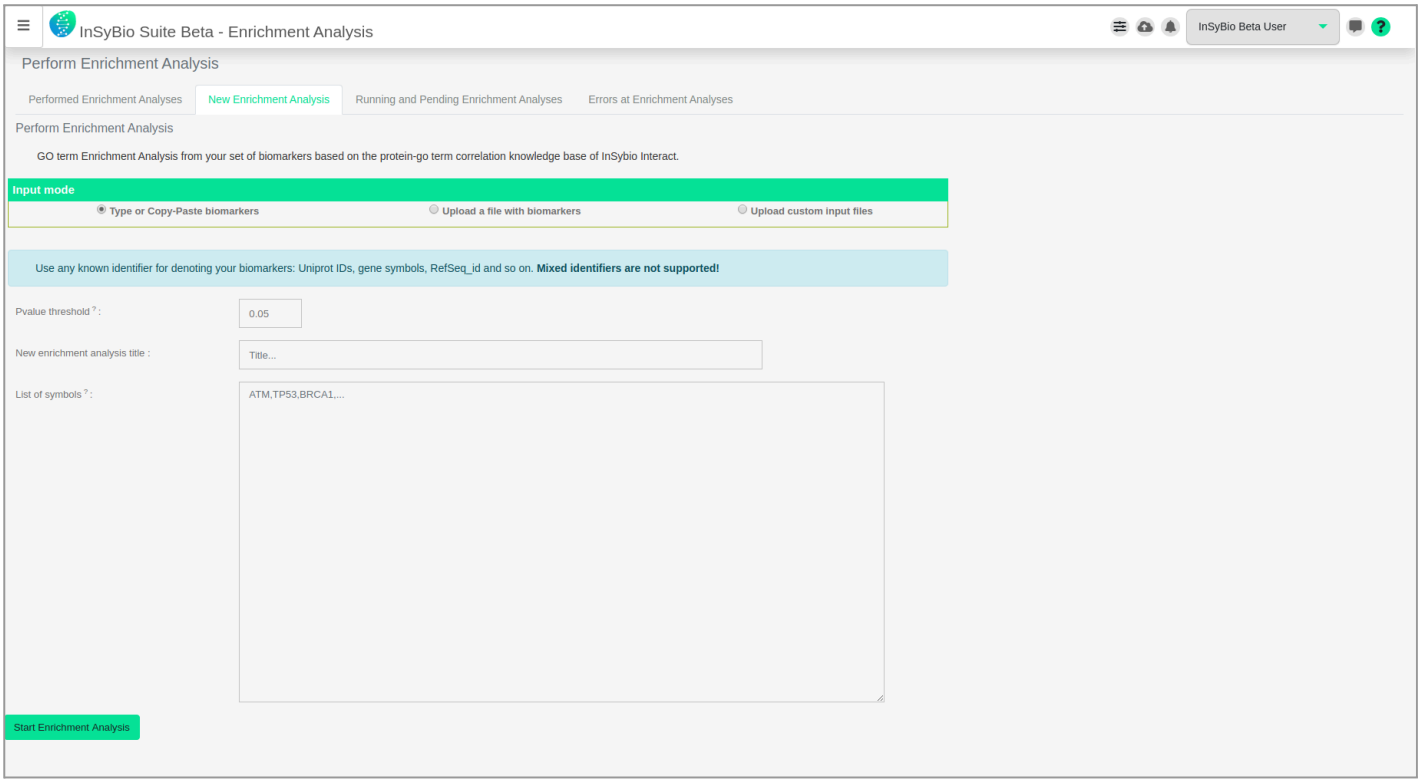
Networks' visualization provides an interactive graphical interface. Users can retrieve information about clicked nodes and edges, export the image in different formats (a PNG, SVG, PDF, XGMML, GraphML or SIF document), decrease opacity on mouseover and view the network using different visualization layouts (force-directed, circle or radial).

You can download the network in a TSV format and perform a list of supported actions in InSyBio BioNets tool.

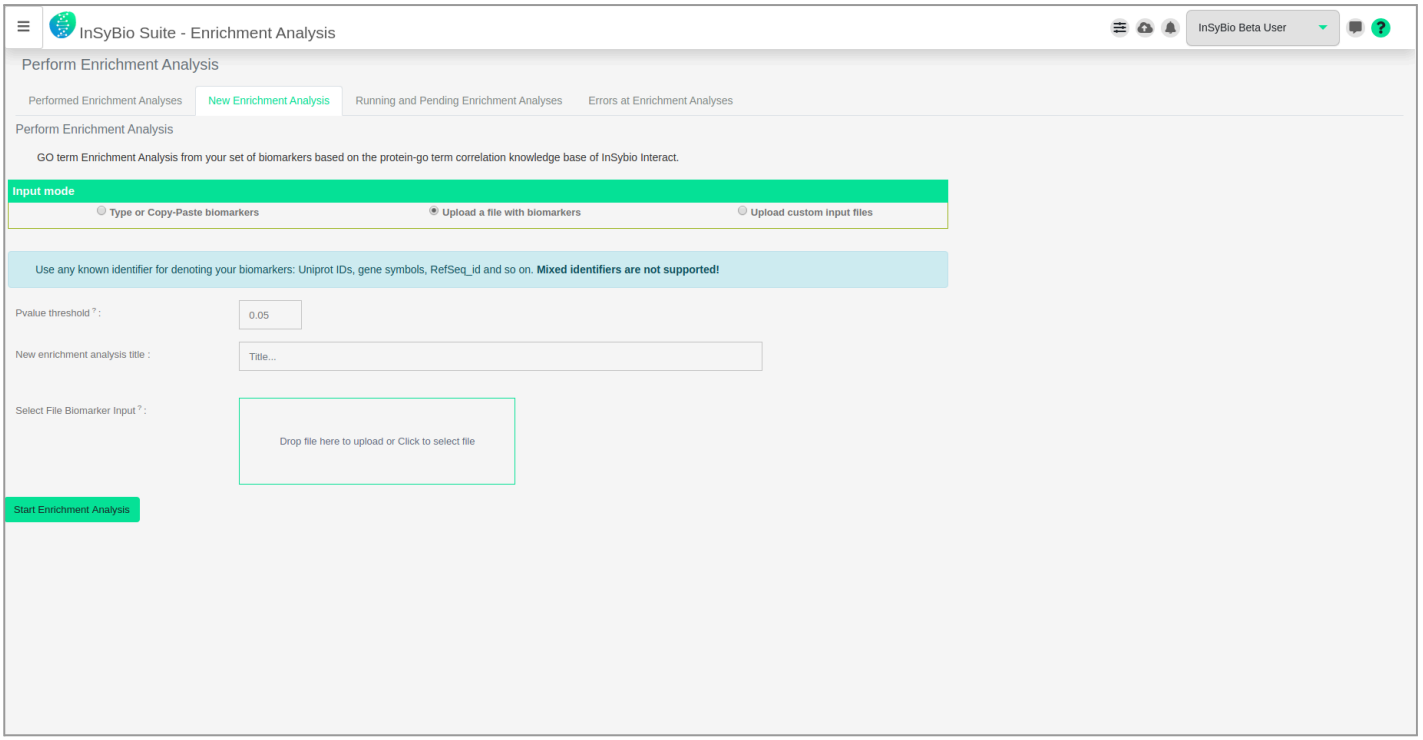
Enrichment Analysis Functionality

You can perform enrichment analysis with hypergeometric distribution on a given list of proteins, genes or transcripts and produce a list of GO terms associated with the list, with their term specificity and score in the distribution. You can also provide your custom annotation, term, term type and functional annotation of molecules files, that will be appended to the default files to perform the enrichment.

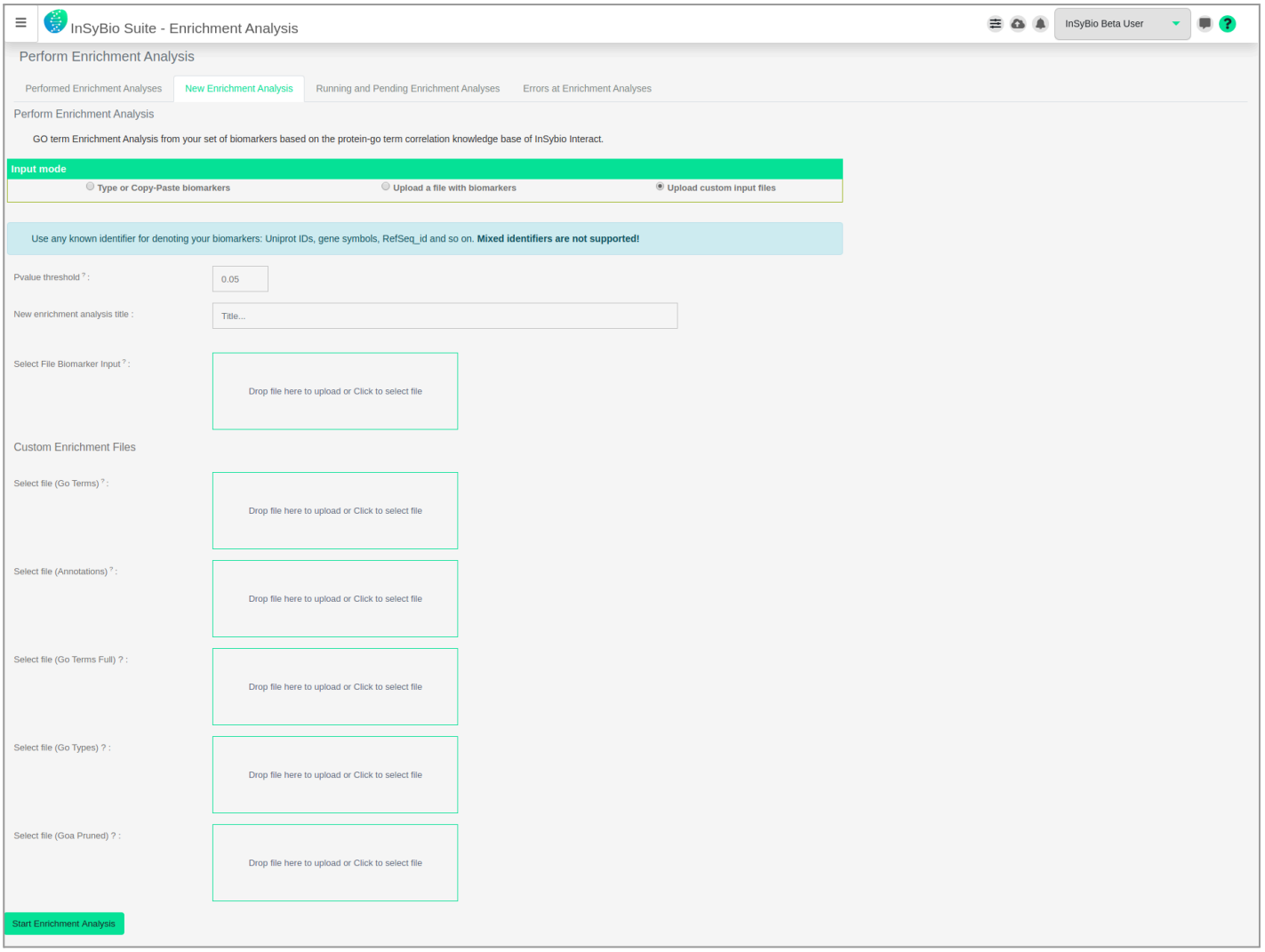
In the tab New Enrichment Analysis you can:



- Either Type or Copy-Paste biomarkers or



- Upload a file with biomarkers or

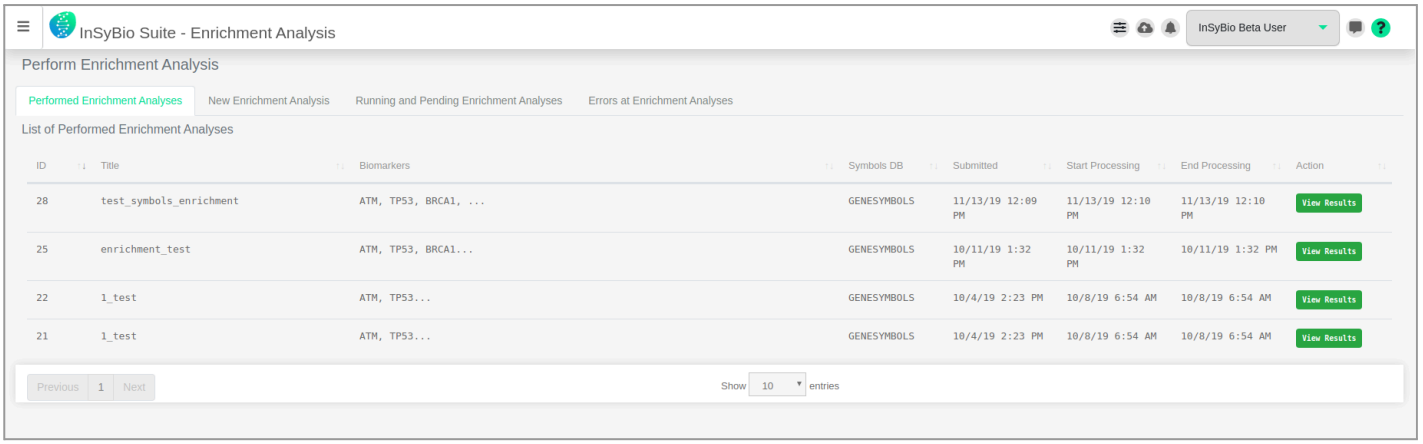


- Upload a file with biomarkers and other custom annotation, term, term type and functional annotation of molecules files

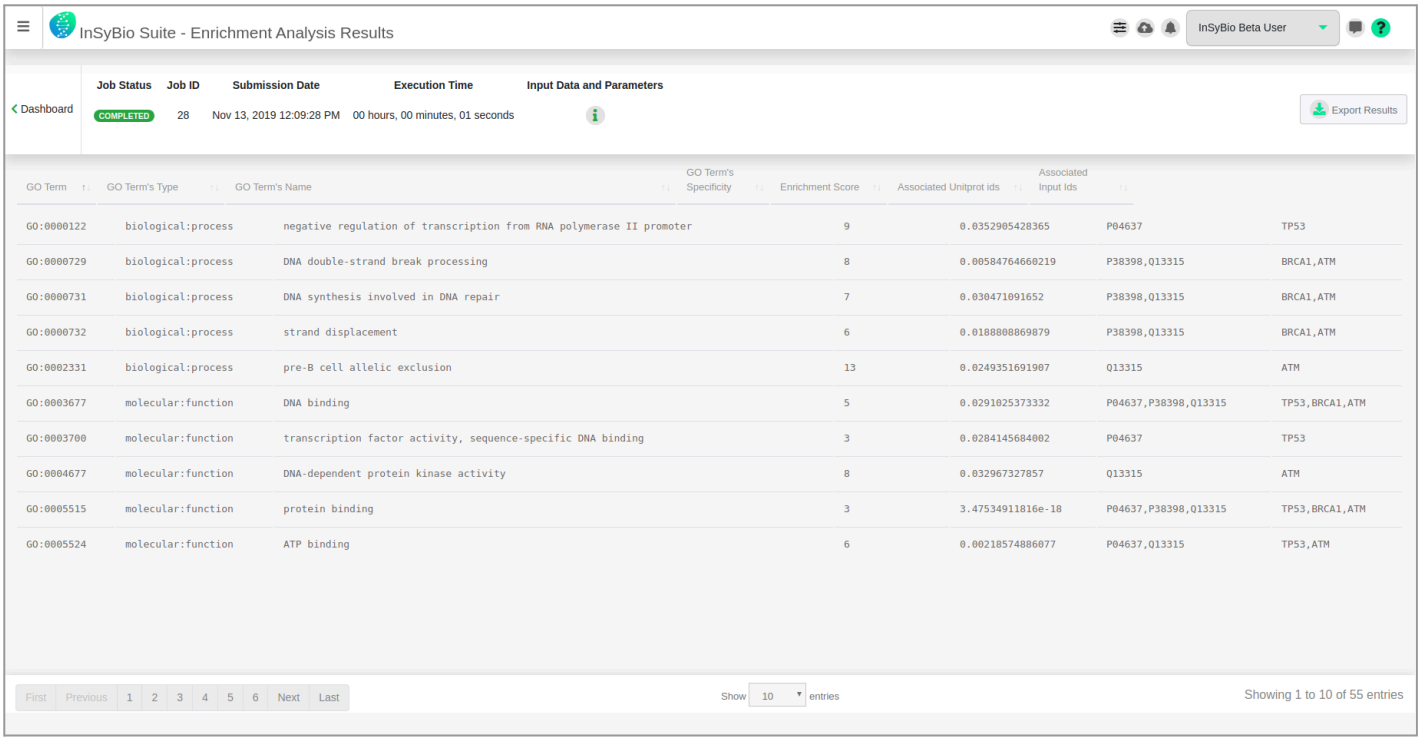
You can use any known identifier for denoting your biomarkers: Uniprot IDs, gene symbols, RefSeq_id and so on. Mixed identifiers are not supported! Note that each symbol should be in one line or separated by comma.

You can also define a pvalue threshold for the biomarker to GO terms association output.

After the enrichment analysis, the job will appear in the Performed Enrichment Analyses tab.

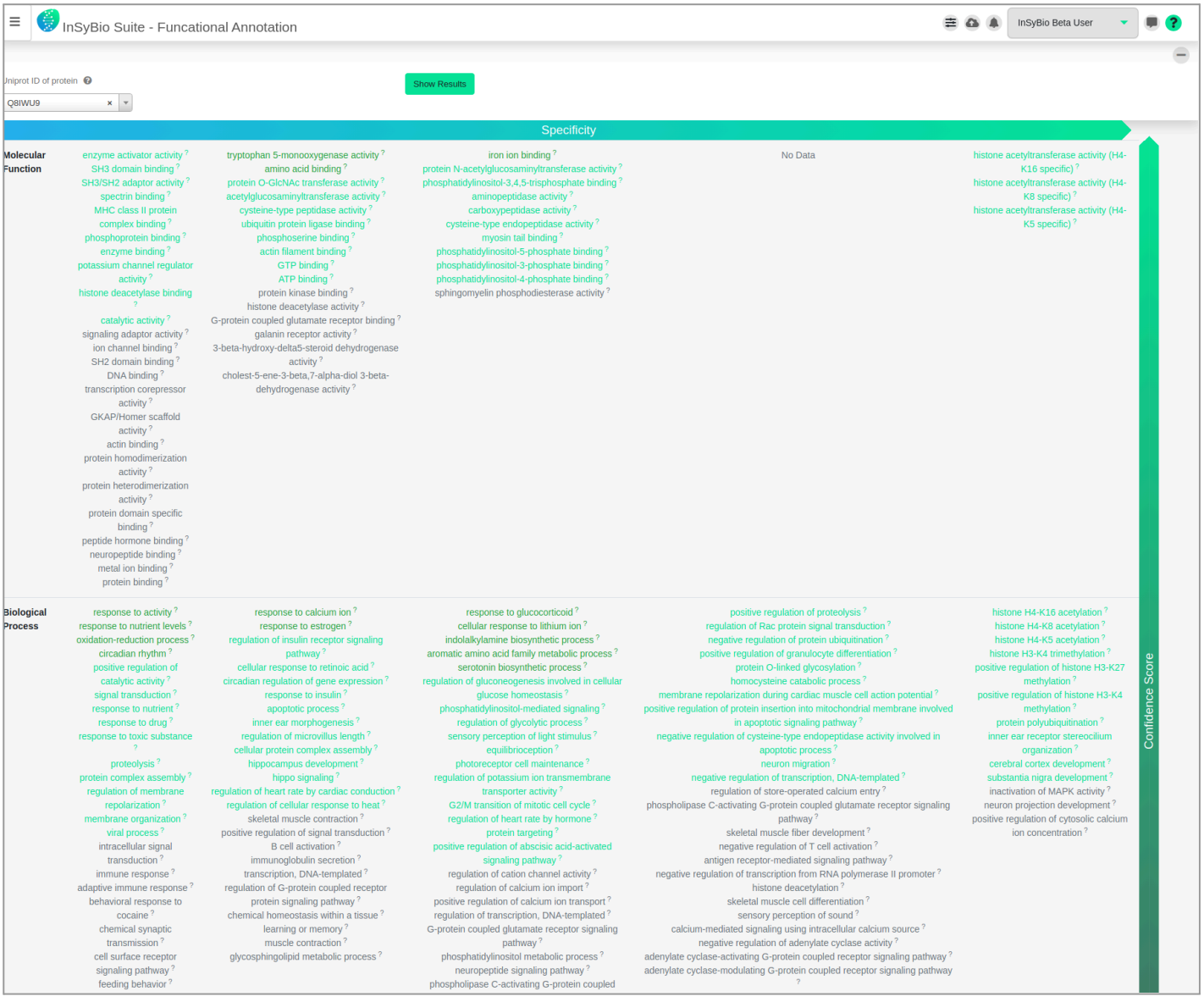


By clicking View Results you will be redirected to the Enrichment Analysis Results window, where you can view the results and download them in a tab delimited file. The results are a list of GO terms, terms type and name, specificity, enrichment score, associated Uniprot ids and input ids.



Functional Annotation Patented Method

You can view the confirmed and predicted functional annotation of a given protein, from our database.



You can provide the Uniprot ID of your protein of interest, and the results of our Functional Annotation Patented method will appear. The associated GO terms with

the protein are separated into 3 rows, one for each major term group (molecular function, biological process and cellular component terms). Going from left to right, the specificity of the term increases. Each term' s position in a cell represents its predicted confidence score in descending order, dark green are confirmed associations (score 1), light green high confidence and gray medium confidence predicted associations (with score near 0.5).

How to get InSyBio Interact

To request a free one month license of InSyBio Suite please email us at info@insybio.com .

To purchase InSyBio Interact commercial version 3.3 please contact us at sales@insybio.com.

About Us

InSyBio Ltd is a bioinformatics pioneer company (www.insybio.com) in personalized healthcare, that focuses on developing computational frameworks and tools for the analysis of complex life-science and biological data in order to develop predictive integrated biomarkers (biomarkers of various categories) with increased prognostic and diagnostic aspects for the personalized Healthcare Industry.

InSyBio Suite consists of tools for providing integrated biological information from various sources, while at the same time it is empowered with robust, user-friendly and installation-free bioinformatics tools based on intelligent algorithms and methods.

COPYRIGHT NOTICE

External Publication of InSyBio Ltd - Any InSyBio information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the InSyBio Ltd. A draft of the proposed document should accompany any such request. InSyBio Ltd reserves the right to deny approval of external usage for any reason.

Copyright 2024 InSyBio Ltd. Reproduction without written permission is completely forbidden.