Protein function easily investigated by genomics data mining using the proteINSIDE online tool

Nicolas Kaspric, Matthieu Matthieu.Reichstadt@inrae.Fr Reichstadt, Brigitte B. Picard, Jérémy Tournayre, Muriel Bonnet

GENOMICS AND COMPUTATIONAL BIOLOGY Vol. 1, No. 1 (2015): e16

Resource

Protein Function Easily Investigated by Genomics Data Mining Using the ProteINSIDE Online Tool

Nicolas KASPRIC[1, 2,*], Matthieu REICHSTADT[1, 2], Brigitte PICARD[1, 2], Jérémy TOURNAYRE[1, 2] and Muriel BONNET[1, 2,*]

[1]INRA, UMR1213 Herbivores, F-63122 Saint-Genès-Champanelle, France
[2]Clermont Université, VetAgro Sup, UMR1213 Herbivores, BP 10448, F-63000, Clermont-Ferrand, France

*To whom correspondence should be addressed: nicolas.kaspric@clermont.inra.fr; muriel.bonnet@clermont.inra.fr

Received 2015-02-13; Accepted 2015-06-09; Published 2015-09-18

ABSTRACT

Nowadays, genomic and proteomic studies produce vast amounts of data. To get the biological meaning of these data and to generate testable new hypothesis, scientists must use several tools often not designed for ruminant studies. Here we present ProteINSIDE: an online tool to analyse lists of protein or gene identifiers from well-annotated species (human, rat, and mouse) and ruminants (cow, sheep, and goat). The aims of ProteINSIDE modules are to gather biological information stores in well-updated public databases, to proceed to annotations according to the Gene Ontology consortium, to predict potentially secreted proteins, and to search for proteins interactions. ProteINSIDE provides results from several software and databases in a single query. From a list of identifiers, ProteINSIDE uses orthologs or homologs to extend analyses and biological information retrieval. As a tutorial, we presented how to launch, to recover, to view, and to interpret the results provided by the two types of analysis available with ProteINSIDE (basic and custom analyses). ProteINSIDE is freely available using an internet browser at www.proteinside.org. The results of this article are provided on the home page of ProteINSIDE website as the example of an analysis result.

AVAILABILITY AND REQUIREMENTS

- Project name: ProteINSIDE
- Project home page: http://www.proteinside.org
- Operating systems: Linux, MacOS, Windows
- License: No

KEYWORDS

Online tool, workflow, protein function, protein interaction, protein secretion, gene ontology, networks, ruminant, genomics.

INTRODUCTION

Given the increasing amount of genomic and proteomic data produced even in ruminants [1, 2, 3], there is a challenge for the bioinformatic data processing, which has not yet been completely solved [4]. Such bioinformatic data processing has to proceed to data gathering and database searching in order to produce a functional interpretation of large datasets. For this purpose, workflows integrating several bioinformatics analyses are now available [5-8] and were developed to mine dataset from specific species (BioMyn [9] for human, DroPNet [7] for Drosophila, TAIR [10] for *Arabidopsis thaliana*, EcoCyc [11] for *Escherichia coli* …) or to identify candidate genes related to diseases as ToppGene [12] or NetPath [13]. The few workflows currently used for the bioinformatics data processing of ruminant datasets are multispecies. Consequently, the data source of the results proposed is not available because of the privacy of databases (as the licensed software Pathway Studio [14] or Ingenuity Pathway Analysis (www.ingenuity.com; Redwood City, CA, USA). An alternative for scientists working with ruminant datasets is to use dedicated and complementary bioinformatics tools implemented as web services. These tools are dedicated to one type of analysis, as for example the annotation according to the Gene Ontology (GO) [15], the prediction of signal peptide to identify putative secreted proteins [16], or the molecular interactions identification [17] and visualization as networks [18, 19]. Whatever the analysis carried out, the first step is to connect a protein name to a unique identifier (ID). Conversely to gene names that have been standardized, protein names or IDs can differ between databases or tools, especially for ruminant data that remains to be largely curated in most of databases. Thus, the use of several bioinformatics tools to mine ruminant datasets leads to a substantial loss of information and time.

A strategy to perform a systematic and integrative analysis of biological protein information from ruminant datasets is to develop an online workflow that integrates several analysis steps in one package and from a unique ID. Thus, we propose ProteINSIDE [20], a web service dedicated to a systematic and integrative analysis of protein's biological information from ruminant datasets. Unlike human, mouse or rat, ruminant species are less annotated and protein sequences or information are not always curated. Often, scientists working with ruminant use orthologs or homologs with the aim to increase the meaningful

biological contexts for proteins thanks to knowledge available in well-annotated species. Thus, ProteINSIDE was designed to run using lists of protein or gene IDs from 6 species (cow, sheep, goat, human, rat, and mouse) to annotate biological and molecular functions and cellular location, predict secreted proteins, search for interactions between proteins within and/or outside a dataset. The objective of this article is to propose a tutorial to use ProteINSIDE and interpret generated results.

METHODS

This section lists necessary equipment, ProteINSIDE resources and describes the dataset used to assess the functionalities of the tool.

ProteINSIDE's features

ProteINSIDE is an online workflow with an interface devoted to accessible and fully customisable analyses from lists of protein or gene IDs. Registered users have access to an analyses manager to run and save analysis, and visualise the results. Unregistered users can use ProteINSIDE, but there is no analyses manager and analyses are deleted each month. ProteINSIDE is divided into three parts: the workflow, the database, and the web interface (Figure 1).

The web interface, designed to easily use ProteINSIDE, helps the user to create the analyses, to have access to the results thanks to a balance between technical functionalities and visual elements, and to inform about updates (Figure 1). ProteINSIDE proposed two types of analysis to be launched: the basic analysis (automatic settings) and the custom analysis (user's settings). There is also a pre-set analysis for registered users only who want to make a new analysis with settings of a previous analysis.

The basic analysis performs a:

- Functional annotation using GO terms by querying QuickGO database [21] without electronic annotation.
- Prediction of secreted proteins using SignalP [16] and TargetP [22] software. We improve the prediction by giving GO terms related to the cellular location of the protein and the processes of secretion.
- Search of proteins interactions curated and listed in IntAct [23], UniProt [24], and BioGrid [25] databases.

The custom analysis performs programs and their options that have to be selected by the user in order to:

- Perform a functional annotation using GO terms from QuickGO, with the options to select also electronic annotations (predicted and scripted annotations), and to generate a GOTree view of linked GO terms (pathways of functional annotation).
- Predict secreted proteins with the option to increase the software's sensitivity of prediction and by this way to increase the number of predicted proteins, however with a higher number of false positive results.
- Search for protein interaction within (core network) and outside (extended network) the uploaded dataset. Options propose to select interactions stored within 1 to 31 databases gathered by the PSICQUIC website [17]. User can select the databases depending on the type of interactions (PPi, Nucleic acid-Protein interaction (NPi), and Smallmolecule-Protein interaction (SPi)) and the data (curated, predicted, curated according to the IMEx project [26] or the MIMIx curation [27]). PSICQUIC service or some databases could be offline, that's why the status of each website is indicated in the table.

To submit an analysis, users either directly paste a list of IDs or upload a file of IDs. Inputs can be protein (e.g., ADIPO_HUMAN) or gene (e.g., ADIPO or gi|62022275) ID, or protein accession numbers (e.g., Q15848) from six species: cow, human, rat, mouse, sheep, and goat (Figure 2). A new analysis is run directly or is placed on a waiting list if the workflow is overloaded. Uploaded data and results remain confidential.

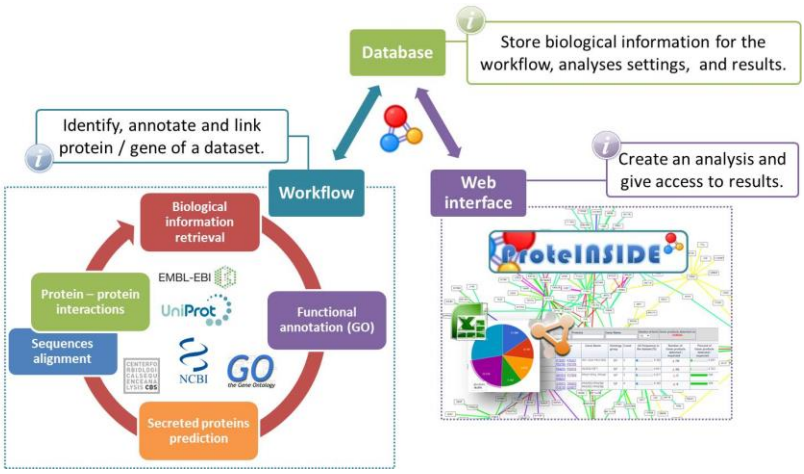

Figure 1. Flow chart of ProteINSIDE structure. The four workflow's modules are either all launched in the basic analysis or individually selected in the custom analysis. These modules aims to query the available biological information, annotate according to the gene ontology, predict signal peptide and visualized protein-protein interactions.

In addition to the web interface, ProteINSIDE is composed of a database and a workflow.

The database (invisible to any user) collects and stores the information required for the proper functioning of ProteINSIDE. It stores analysis settings and results to reduce server load (Figure 1). The database stores also a gathering of biological information from the NCBI [28] (Gene, Protein, and HomoloGene for known orthologous proteins between the 6 species) and UniProt [24] databases (for the ID Mapping module), and QuickGO [21] and AmiGO [29] (for the GO annotation module). A script updates automatically and monthly the database by extracting IDs, homologs, biological function, FASTA sequence, and other information from the latest releases of these databases.

The workflow uses uploaded data. It is a combination of Perl and R scripts to query databases, recover protein data, perform calculations and run algorithms for signal peptide predictions and network visualisation (Figure 1). The workflow is invisible to any user. The workflow is composed of 4 parts: the “ID Mapping”, the search of annotations according to GO, the prediction of secreted proteins, and the search of protein-protein interactions (PPi). The workflow always starts by the ID Mapping program which searches the biological information available for each protein or gene of the input within the ProteINSIDE database. Gathered biological information is required to run the 3 other modules of the workflow: “Gene Ontology”, “Secreted Proteins”, and “Protein Interaction” (described in the “Results” section of this article). The GO program queries QuickGO and ProteINSIDE’s databases to perform the functional annotation. The GO program analyses over- and under-represented terms to highlight the most relevant GO terms related to the input. These statistical calculations are made with an R script performing a Fisher´s exact test (functional enrichment first proposed by FatiGO [30]) and the resulting *p*-value is corrected or not by the Benjamini & Hochberg (BH) test [31]. The prediction of secreted proteins is made using a local version of SignalP (version 4.1) that looks for a signal peptide on amino acid sequence of each protein [16] (cutoff of 0.45 and 0.34 for SignalP prediction, in the basic and custom analysis with the sensitive option selected, respectively; for more information see the tutorials of SignalP[a]. To ascertain that proteins are secreted, ProteINSIDE uses TargetP [22] (version 1.1) to predict the cellular location of each protein. ProteINSIDE uses a pre-set cutoff option to get a significant prediction (higher than 95%) according to TargetP instructions[b]. Protein interactions are searched using PSICQUIC service [17] and statistical calculations are made with an R script and the package “tnet” [32]. ProteINSIDE performs sequence alignment thanks to a local version of NCBI BlastP [33] against UniProt/Swissprot databases [24, 34]. Lastly and as an additional valuable tool, ProteINSIDE lists in one table all known IDs for an input of proteins or genes thanks to the ProteCONVERT module. This list is the result of a search and of a gathering of IDs thank to the ProteINSIDE biological database. Only registered users have access to the ProteCONVERT module.

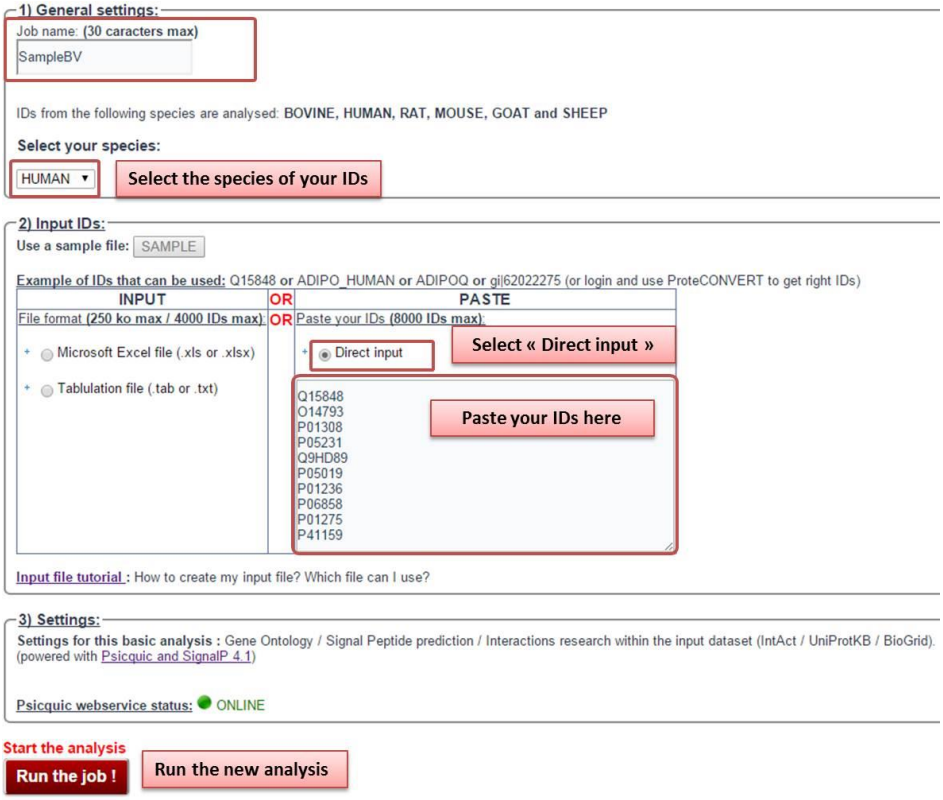

Figure 2. Setting up a basic analysis. First, enter a name for the analysis and select the species of study. There are two ways to submit a protein or gene list; you can use an input file or directly paste your IDs. The input file must be less than 250 kb and the file format must be specified. There is also a "Sample" button that loads parameters for an example analysis. Once everything is filled, click on the button "Run the job" to submit.

The results of an analysis with ProteINSIDE are available online or downloadable. Unregistered user gets results through a unique code and a link provided after the submission of a list. Registered users have access to the results through their analysis manager after their connection with their login and their password. Four separated pages provide results from the four analyses ("ID Mapping", "Gene Ontology", "Secreted Proteins", and "Protein Interaction"). The results are dynamic tables and charts that can be sorted and filtered by specific criterion such as biological function, protein or gene ID, or *p*-value. Tables and charts are made using Google Charts package (https://developers.google.com/chart/) and are downloadable, diagrams and histograms are printable. Networks are downloadable as image (.pdf or .png) or as network viewer input files (Cytoscape .cys or graphml and xgmml). The results interpretations are detailed in the results section.

Implementation

ProteINSIDE is freely available online at www.proteinside.org and doesn't require an installation on a computer. ProteINSIDE is completely adapted for any internet browser and tablet. We recommend multiprocessors computer with at least 2 GB of ram to get better performances for huge network visualization and filtering.

The web interface is programmed in PHP, HTML, and JavaScript. The workflow has been completely programmed in Perl (version 5.10.1; CPAN modules (Comprehensive Perl Archive Network) and BioPerl [35]) and R scripts (version 3.0.1). The database was made in MySQL (version 5.5) (Figure 1).

Sample dataset

We have created a dataset to assess ProteINSIDE performances. This dataset is composed of the UniProt accession numbers of 133 proteins (Table 1): 34 proteins related to the glycolysis cycle, 11 proteins from the respiratory chain, 5 proteins from the tricarboxylic acid cycle (TCA), 79 hormones or secreted proteins, and proteins with very specific functions unrelated to the others. We also included a duplicated ID among proteins of the glycolysis to verify its recognition by ProteINSIDE. ProteINSIDE is able to detect duplicate protein even if the IDs are different: a Gene Name, a UniProt accession number, and a Gene Identifier related to a same protein will be taken into account as a single protein.

We have created this dataset on bovine species, but the number of annotations and PPi weren't sufficient for a complete overview of the functionalities of ProteINSIDE. Then, we used the same proteins using human IDs to test ProteINSIDE with the basic and the custom analyses (Table 1).

RESULTS

Here we present how to run a basic or custom analysis and how to view the results. We explain how to interpret the results and we discuss the relevance of biological information extracted by ProteINSIDE for our sample dataset of 133 proteins.

Setting up a Basic Analysis: a standard overview of a dataset

ProteINSIDE performs a basic analysis (in which settings are locked and the workflow provides GO terms, list of putative secreted proteins, and PPi data from IntAct [36], UniProt [24], and BioGrid [37] databases). A basic analysis gives a complete overview of a dataset. To set up a basic analysis, user has to follow these steps (Figure 2):

- Click on "Basic Analysis" menu on the homepage of ProteINSIDE
- Fill in "the job name" box
- Select the species of study (the same species as the uploaded IDs)
- Upload an input file or directly paste IDs
- Click on the "Run the job" button to submit a new analysis

The analysis status is indicated by the colour of a button: red for "analysis on the waiting list", yellow for "the analysis is running" and green "analysis done". The blue globe is the link to access to the online results:

- Click on the blue globe button to view the results (or use the trash to delete them)
- Visualise the results summary produced by the four modules of analysis on the first default page (entitled "Results Summary", Figure 3)
- Navigate between the four module's results pages by clicking on the module's name on the toolbar menu.

Analyses and data	Glycolysis	Hormones	TCA	Analysis duration (min)
Dataset	33+1 duplicated	79	5	-
Basic analysis	27	78	3	2
Custom analysis	33	79	5	10

Table 1. Results summary of ProteINSIDE analysis performances. The numbers are the proteins that belong to main pathways in the sample dataset, that are properly annotated by GO terms relevant to glycolysis and tricarboxylic acid (TCA) pathways, and that have been predicted as secreted by SignalP (and confirmed by GO terms, TargetP, and subcellular location) for hormones.

For our sample dataset, the "Results Summary" page reported that all 133 proteins were recognized by ProteINSIDE and the protein in duplicate was identified and excluded from the analysis (Figure 3). Thus, 132 proteins were submitted to the four modules of analysis.

The "ID Mapping's" module aimed to retrieve and gather basic biological knowledge, results are directly viewed on the "ID Resume" web page of ProteINSIDE. This module compares each submitted IDs to the database of ProteINSIDE to ascertain a match with genes or proteins from human, rat, mouse, cow, sheep or goat species. The local biological database of ProteINSIDE is a combination of NCBI Gene/Protein, NCBI HomoloGene [28], and UniProt [24] databases. These databases were chosen because data are easily extractable, curated and daily updated. For each uploaded ID, ProteINSIDE obtains and summarises as a downloadable table (Figure 4): gene or protein ID, gene or protein name, a summary of protein function, gene chromosomal location, and information on tissue expression and cellular location. The module also recovers the protein sequence of each input ID. Each protein and gene ID listed on this web page are linked to corresponding UniProt and NCBI web pages. FASTA amino acid sequences of each input are also downloadable.

The module dedicated to the functional annotation according to the GO consortium, produces results that are viewed on the "GO" web page of ProteINSIDE. ProteINSIDE imports GO terms by querying the QuickGO database [38]. QuickGO was chosen because of its daily update, accessibility, and performances. ProteINSIDE only imports GO terms that have been selected by evidence codes (GO Inferred from Electronic Annotation codes (IEA) are excluded by the basic analysis) and confirmed by curators. The GO script of ProteINSIDE analyses over- and under-represented terms to identify the most relevant and the most specific terms associated with the uploaded list. Within a GO, ProteINSIDE compares the number of genes or proteins from the dataset to the total number of gene products (for a species) declared in the AmiGO database [29] to provide a coverage frequency, and thus, to identify the most representative pathways associated to a dataset. The result is viewed on the "GO" web page of ProteINSIDE as tables and diagrams. Three tables (Figure 5) report the GO terms that annotated two or more proteins (Figure 5-B), the GO terms that annotate one protein (Figure 6-C), as well as all GO terms for a protein (Figure 5-D). Each annotation is informed with an evidence code (that reflects the type of experimental evidence or analysis to describe an annotation between a GO term and a gene product) and the database source. Tables are automatically sorted by the best enrichment *p*-value to help the user to view the most significant GO terms related to a dataset. Tables can also be sorted by ontology group, *p*-value range for enrichment, GO term description, gene name or any input IDs (Figure 5B). From the sample dataset of 132 proteins, ProteINSIDE annotated 128 proteins with 624 GO terms. The most significant enriched GO terms is "hormone activity" (that annotated 31 proteins over the 79 expected; not shown) and "glycolytic process" (that annotated 27 proteins over the 33 expected; Table 1). The low number of annotated proteins may be related to our choice to use only GO terms that have been confirmed by curator in the basic analysis. This means that the basic analysis doesn't use the annotation with IEA (Inferred by Electronic Annotation) evidence code. However, the option to use IEA is provided in the custom analysis to extend the annotations.

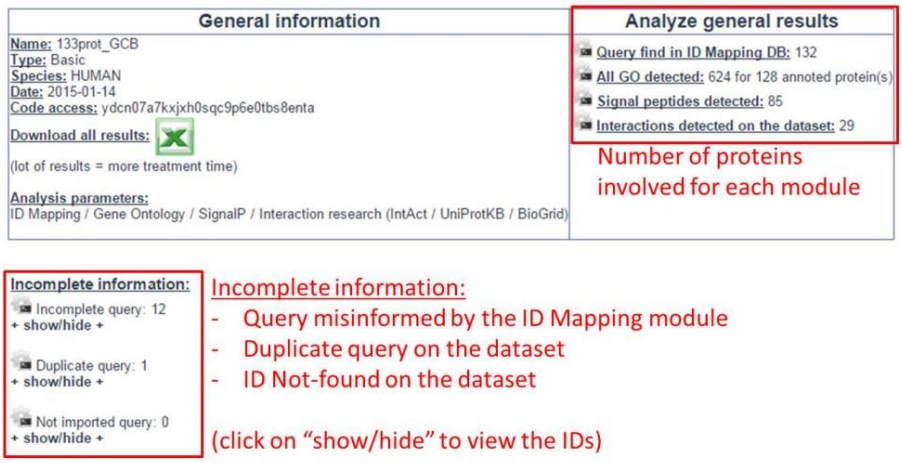

Figure 3. Main page of results produced by a basic analysis. This is the first page of the results. It shows the number of proteins or genes successfully analysed by each module.

The module that aims to predict potentially secreted proteins provides results on the "Secreted Proteins" web page of ProteINSIDE (Figure 6). To identify proteins that are putatively secreted, ProteINSIDE first predicts the presence of a signal peptide on a protein sequence (imported by the "ID Mapping" module) through a local version of the SignalP tool [16]. SignalP was chosen because of its high prediction score in comparison with other available tools [39, 40]. To ascertain the prediction, a local version of TargetP software [22] predicts the subcellular location of the proteins. ProteINSIDE also checks the subcellular location of the protein using UniProt source to confirm TargetP results. As a final verification step, ProteINSIDE selects the GO terms related to secretory pathways for each SignalP prediction. For this purpose, we have selected about 1,000 GO terms related to secretion (monthly updated) as for example: secretion, vesicle, or extracellular region. This four-step analysis improves the reliability of proteins proposed to be secreted thanks to a signal peptide and to our knowledge is unique to ProteINSIDE [40]. However, proteins are also secreted by pathways that do not involve signal peptide such as: endosomal recycling, plasma membrane transporter, membrane flip-flop, and membrane blebbing including the formation of vesicles or exosomes [41]. Thus, ProteINSIDE was designed to predict the proteins secreted by other pathways, by gathering the data of subcellular location provided by UniProt, GO terms, and TargetP results (Figure 6-B). From our sample dataset of 132 proteins, ProteINSIDE has predicted 85 proteins as potentially secreted outside the cell by a signal peptide, among them 78 over the 79 proteins that were expected (Table 1). This lack of perfect prediction can be explained by the false positive and false negative prediction rates of SignalP, as already evaluated by Petersen et al. (Supplementary materials and methods of [16]). Over the 85 predicted secreted proteins, 65 were also annotated by GO terms related to the secretion. The subcellular locations of 81 proteins were both confirmed by TargetP and UniProt source. Additionally, 30 proteins were predicted to be secreted without signal peptide.

The fourth module is dedicated to PPi analysis and results are viewed on "Protein Interaction" web page of ProteINSIDE. PPi identification and visualisation within a network conveyed how various genes or proteins contribute to cellular or metabolic processes. ProteINSIDE uses the PSICQUIC service [17] to identify PPi and imports PPi identified by their "interaction detection methods" with experimental proofs and confirmed by curator. The basic analysis identifies PPi within the uploaded dataset (core network) using the preselected databases IntAct, UniProt, and BioGrid. These PPi databases were chosen as a default option because there are daily updated and reviewed by curators as well as by the curation processes of the IMEx project (that ensures reliable interactions data using experts and curation rules shared between many interaction databases [26]) or MIMIx (a guideline of the minimum information required for reporting a molecular interaction experiment, thus advising the user on how to use the interaction data [27]). Moreover, BioGrid is the biggest PPi database that has its own curation workflow (more than 740000 curated PPi) and is not a partner of IMex curation program. IntAct is another big PPi database with more than 380000 PPi currated according to IMex and MIMIx curation rules and that are often listed in several databases. UniProt is a major database dedicated to the study of proteins. Thus, it possesses its own curated PPi but in lesser amounts compared to the two other specialized databases (less than 13 000 PPI; UniProt is a partner of IMEx project). By using 3 databases as a default option, the aim of ProteINSIDE is to favour the use of multiple PPi databases in order to improve the PPi data gathering [42].

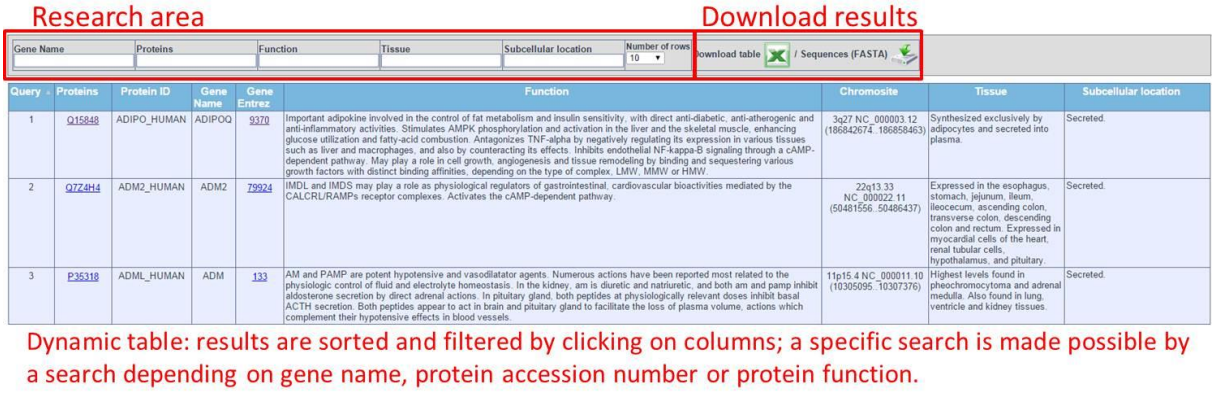

Figure 4. Biological knowledge retrieval. The ID Mapping module results are listed in a table. This table provides protein IDs, gene names, summaries the protein function, chromosomal locations, data on tissue expression, and subcellular location. User can sort the table by using the dynamic table research area.

These 3 PPi databases ensure the good recovery of known interactions for an overview of interactions within or/and outside of a new dataset. Then, ProteINSIDE lists pairs of proteins known to interact between each other in a downloadable table (Figure 7) and constructs a network (Figure 8) using the PPi identified within the uploaded list. The dynamic network is available by using the "Cytoscape" button on the "Protein Interaction" web page ("Dynamic Cytoscape view of PPi", Figure 7-A). Within the network, edges are experimental detection methods used to identify the PPi. Consequently, several edges may link two proteins. Network can be sorted by the number of interactions by node, the proximity of a node to other nodes (closeness centrality; CC) and the shortest paths between nodes (betweenness centrality; BC) (Figure 8-A). These centralities criteria were already proven to be efficient to select key nodes/proteins within a pathway [43]. From our sample dataset of 132 proteins, ProteINSIDE has identified 29 PPi that involved 28 different proteins (Figure 7-B). As expected from our small dataset, ProteINSIDE linked, within sub networks, proteins involved in glycolysis, TCA or respiratory chain as protein complexes (partially on Figure 8-B).

Setting up a Custom Analysis: an added-value provided by the extension of the analysis

We made a custom analysis using the same major settings as for the basic analysis with additional options (GO network, GO electronic annotations, and extension of PPi to proteins outside of the dataset in the same species, extended network). To set up the custom analysis, user has to follow these steps (also explained by Figure 2):

- Click on the "Custom Analysis" menu on the homepage of ProteINSIDE
- Fill in "the job name" box
- Select the species of study (the same species as the uploaded IDs)
- Upload an input file or directly paste IDs

Then, user has to select the settings of either all or only one module of analysis on the section "4" of the page, by following these steps (Figure 9):

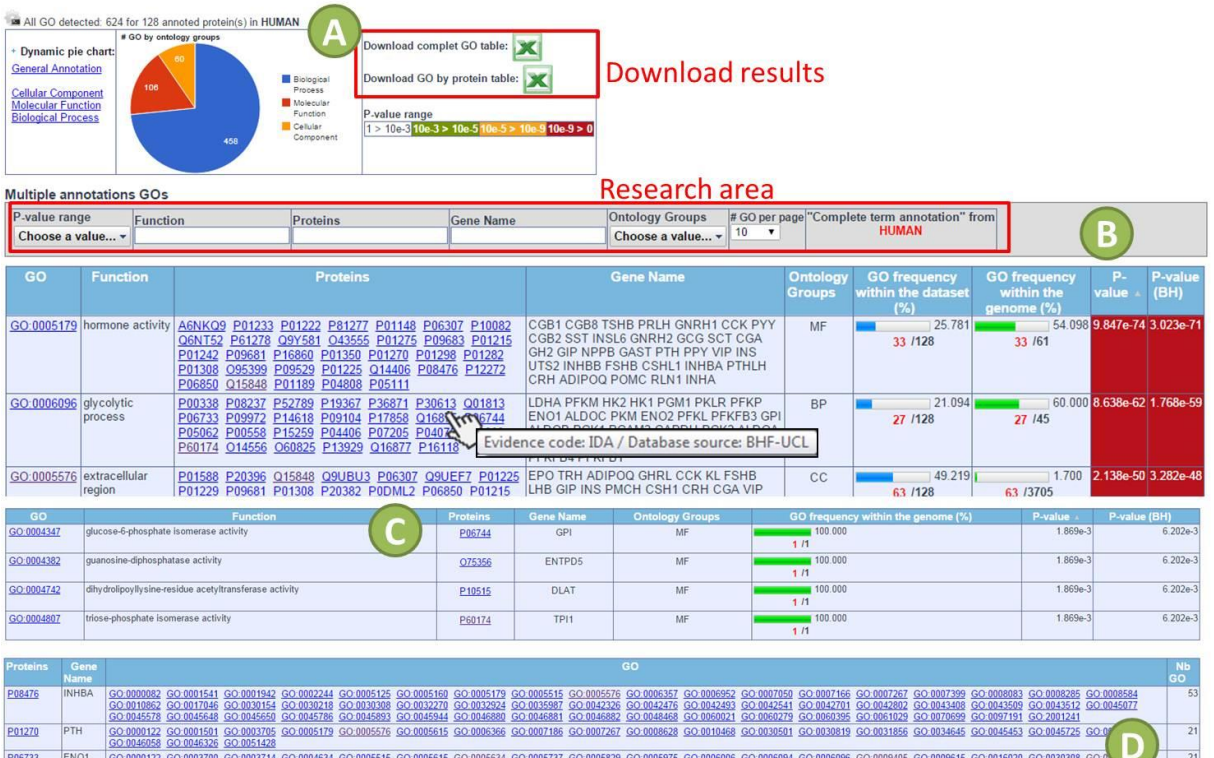

Figure 5. Functional annotation according to the Gene Ontology. GO results are first extracted and classified by the number of GO terms related to Molecular Functions, Biological Processes, and Cellular Components, then visualised as diagrams or downloadable tables. (A) Main menu of GO results page, to download the results as Excel files, to view the significance of *p*-value range colours, or a proportion of major annotation categories as diagram. (B and C) GO terms are also sorted as two dynamic tables (a table for GO terms that annotate more than one protein on the dataset - B, and a second table for the GO terms with a single protein annotated - C). Tables can be sorted by GO term, function, protein name or ID, gene name, number of annotations, annotation frequency or annotation enrichment. (D) A third table lists all GO codes for a given protein. Users can move the cursor over a protein to be informed about the evidence code and the database source of the GO annotation (B; where IDA means "Inferred from Direct Assay").

- Select Gene Ontology module
  - Select "GO electronic annotation (IEA)" if you aim to use GO annotation inferred from electronic annotation.
  - Select "Gene Ontology Tree network" to view linked GO terms
- Select Signal peptide module to use the basic cutoff value of prediction
- Select Protein - protein interaction module
  - Select "Protein - protein interaction custom analysis"
  - Select "Extend PPi research using protein outside of the dataset", if wanted
  - Select "Human species" to analyse PPi using data available in Human, for example
  - Select either the 3 most used databases (IntAct, UniProt, and BioGrid as used in the basic analysis) or from 1 to 31 databases (PPi are daily updated in each database)

Alternatively, user can load automatically the same settings as those already used in a previous custom analysis, by clicking on the “pre-set” button.

At the completion of the custom analysis, the “ID Resume” web page provides the same information than the basic analysis (Figure 3).

Within the GO module, the choice to use the electronic annotation option has increased both the number of annotated proteins (132 rather than 128 without IEA in the basic analysis) and the number of annotations by around 40% (1080 unique GO terms rather than 624 in the basic analysis). Thanks to IEA option, ProteINSIDE correctly retrieved the 33 expected proteins related the glycolytic process and the 79 proteins related to a hormone activity (Table 1). The GOTree network linked 570 GO terms. A link between 2 terms is represented by an "is_a" relation: "Diuretic hormone activity" linked to "Hormone activity" means that "Diuretic hormone activity" is a "Hormone activity" pathway. The network can be sorted by ontology group, by *p*-value range (to select and to link only the most enriched GO terms), by the number of directly linked terms or also by the number of GO terms linked together (to select group of GO terms involved in the same biological function). From our sample dataset, we have chosen to illustrate the GO tree of the “Molecular Function” group (Figure 10). In this visualisation, squares with dark red colour were GO terms which have annotated the highest number of proteins. Among them and as expected the GO:0005179 with the best *p*-value and the darkest red colour was “Hormone activity”, in agreement with the over representation of hormones in our sample dataset.

The “Secreted Proteins” module has predicted the same 85 proteins as the basic analysis as being secreted. By comparison with the basic analysis, the use of IEA option has allowed to confirm this prediction for 82 proteins that were also annotated with GO terms related to a “secretion” function.

By comparison with the basic analysis, the settings selected within the “Protein Interaction” module provided PPi within the dataset (between proteins of the dataset, as the basic analysis) and PPi between proteins from the dataset and outside of the dataset. For the extended network, ProteINSIDE retrieved 688 PPi made by 500 proteins. Among them, 61 proteins were from our uploaded sample dataset. By using PPi outside of the dataset in Human species, we got 95% more PPi that involved 60% more proteins from the sample dataset than the PPi recovered with the basic analysis. The extended network (Figure 11) highlighted major subnetworks related to the respiratory chain (Figure 11-A), hormone activity such as signalization pathways of adipokines (Figure 11-B), growth hormone (Figure 11-C), thyroid hormones (Figure 11-D), glycolysis (not highlighted), and carbohydrate metabolism (not highlighted). This is consistent with the over selection of proteins from glycolysis, TCA or hormones or adipokines. Betweenness and closeness centralities were used to sort the most central proteins of this extended network (Figure 11-E). By this way, we identified 22 highly central proteins, 13 of them coming from the uploaded sample dataset and involved in respiratory chain and glycolysis as protein complexes.

DISCUSSION

Currently, most genomic and proteomic studies increasingly generate data which have to be gathered, filtered, and analysed using one or more softwares [44-46]. The major and widely used strategies to systematically study proteins [47] and genes [48] in a cell are based on functional annotation, proteins interactions and pathways analysis. The literature describes many tools for genomic and proteomic data analysis [4]. Scientists have to select appropriate tools among those for either the GO annotation [15, 30, 49, 21, 29, 50], the prediction of secreted proteins [51, 52, 39, 53], or the search of protein - protein interactions [54, 55, 36, 56, 37, 57, 58].

ProteINSIDE is not just an additional resource since it was designed to provide efficient and original strategies to run in a single query, biological knowledge gathering, GO terms annotation, secreted protein prediction, and protein interaction. The DAVID [59], ToppGene [12] or Babelomics [60] software resources are often mentioned for the biological knowledge gathering, functional annotation using GO terms or searches for proteins interactions. By comparison to these tools, added-values of ProteINSIDE have to be highlighted.

ProteINSIDE provides a functional annotation using a monthly updated GO terms database and enrichment calculation. Indeed, the list of GO terms is in constant evolution and GO terms could become redundant or obsolete the next month [15]. This could induce bad information in the results of an analysis if the database is not often updated. Each result of the annotation is easily readable thanks to dynamic tables and diagrams which can be sorted with many options and can be downloaded to work offline. The GO tree visualization of the most often associated GO terms with a list of IDs, is another added-value of ProteINSIDE. Tree networks of GO terms are also done by AmiGO or QuickGO to get an ancestor chart of a single term. However, ProteINSIDE is the only tool which highlights biological pathways of a dataset using linked GO terms and their representativeness rate (using *p*-values and number of annotations). This network visualization is also easy to use thanks to the friendly user interface that gives access to the sort

options. For the PPi research and visualization, ProteINSIDE uses only interactions that are based on experimental observations. The drawback is that the number of PPi identified by ProteINSIDE could be lower than those proposed by other resources that also list predicted interactions inferred from literature mining. Furthermore, ProteINSIDE is also capable to draw large interaction networks thanks to the use of the powerful graphical Cytoscape application. ProteINSIDE provides different options to filter large networks, making it as easy to use as the widely used resource STRING [57], and efficient to select keys proteins in a network. Moreover, to analyse locally the networks, files (e.g. .cys, xgmml, graphml) are ready to be open by a network viewer like Cytoscape (and its numerous plugins) and are downloadable from the PPi page result. To our knowledge, among the tools to mine genomics data from mammals, ProteINSIDE is the only resource that allows a very simple view and analysis of network, and prepares data for their further download and analysis by other network viewer software as Cytoscape. These features may be valuable for biologists without a strong bioinformatics background. For the less informed species, ProteINSIDE allows searching PPi in well-informed species thanks to homologous IDs. For this, ProteINSIDE automatically selects homologous IDs from its database for the wanted species. Nevertheless, user can choose to run a local Blastp to select the species with the highest sequence homology with the proteins of the input dataset, and then ProteINSIDE proceed to the selection of orthologous IDs for this species. A functional annotation of all proteins from an extended network (PPi between proteins within and outside of the dataset) is done by clicking a button on the network visualisation. Results of this annotation are available as a new analysis. In addition to biological knowledge gathering, GO annotation, and analysis of PPi, ProteINSIDE also proceeds to an *in silico* secretome analysis [40]. For this purpose, ProteINSIDE merges four strategies of analysis: signal peptide [16] and cellular location [22] predictions, as well as a review of GO term annotation and cellular location recorded in UniProt. This four-step analysis provides a reliable prediction of proteins secreted thanks to a signal peptide. To our knowledge, ProteINSIDE is the unique all-in-one tool that predicts secretome from a list of gene or protein IDs [40].

Scientists are dependent on the species of study when they choose among resources available for their genomic and proteomic data analysis. Indeed, many tools are dedicated to only one species such as BioMyn for the Human [9] or DroPNet for the Drosophila [7]). Moreover, many tools are dedicated to diseases studies such as NetPath [13] and ToppGene [12]. ProteINSIDE has been first tool designed for genomic and proteomic data analysis in ruminant species namely cattle, sheep, and goat. However, the lack of information on these species required us to add human, rat, and mouse species to do homologous analysis. Thus, IDs from these species are perfectly recognized and analysed by ProteINSIDE. To our knowledge, ProteINSIDE is the only resource that allows the user to recover biological knowledge from well-known species (human, rat or mouse) using IDs from ruminant species. This avoids losing information since many sequences or annotations remain to be stored in public databases for ruminant species and especially for goat. To our knowledge, only AgBase [61], a manually curated gene annotation database for farm species, including cattle and sheep, is available for functional annotation. However, AgBase does not perform analysis of PPi or prediction of secreted proteins.

In this article we have presented the performances of ProteINSIDE, a new powerful workflow which gathers tools and public databases to retrieve biological information of genes or proteins lists from 6 species (Bovine, Ovine, Caprine, Human, Rat, and Murine). We have reported a tutorial to describe how to get and interpret the results of a basic and a custom analysis with ProteINSIDE. Currently, there is no tool that performs in one query the analyses proposed by ProteINSIDE. ProteINSIDE offers a friendly-user interface where user can view, work, and download the results of an analysis. ProteINSIDE gives also a single file containing all results of an analysis. Thus, ProteINSIDE offers a great support to analyse efficiently a large quantity of data from genomic and proteomic studies to gather and interpret results necessary to construct a new research hypothesis or answer to a single question.

ACKNOWLEDGEMENTS

N. Kaspric's PhD grant is provided by the regional council of Auvergne in France, APIS-GENE and the regional information system Lifegrid with the help of Feder has provided the grant of J. Tournayre.

AUTHOR CONTRIBUTIONS

Conceive and design ProteINSIDE web service: NK, JT, MR, and MB. Perform experiments and analysed the data: NK and MB. Write the paper: NK, MB, and BP.

CONFLICT OF INTEREST DECLARATION

The authors declare no conflict of interest.

SUPPLEMENTARY DATA

High resolution files of the main figures on the paper are available for download at Genomics and Computational Biology online.

ABBREVIATIONS

BC: betweenness centrality

CC: closeness centrality

GO: gene ontology

ID: identifier

PPi: protein-protein interaction

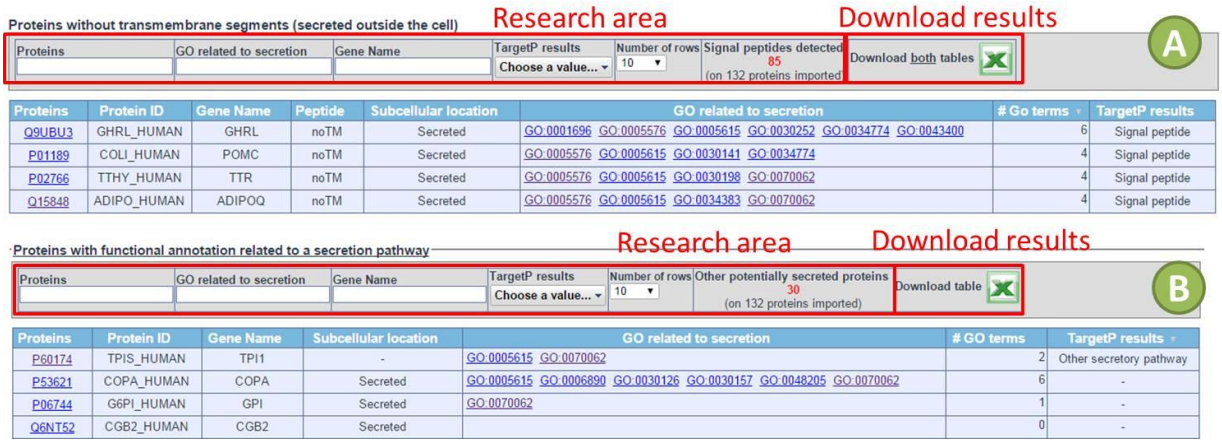

Figure 6. Prediction of secreted proteins. Proteins potentially secreted are listed as two or three downloadable dynamics tables. (A) The first table lists proteins predicted as secreted by SignalP. The column "Peptide" provides the results for a positive identification of a signal peptide on a protein sequence as provided by SignalP. Identified peptides can be "noTM" (not transmembrane) or "TM" (transmembrane), only "noTM" are listed in the first table. The column "Subcellular location" provides the location of the protein declared in the UniProt database. The column "TargetP" provides the prediction of the subcellular location of the protein by TargetP software, and GO related to secretion are also listed to improve the prediction. A second table lists proteins with the "TM" prediction of SignalP, not shown in the figure since there was no result with the sample dataset. (B) A third table lists proteins potentially secreted by secretory pathways that do not involve signal peptide. In this table, GO terms, TargetP prediction, and subcellular location are also selected to improve the prediction.

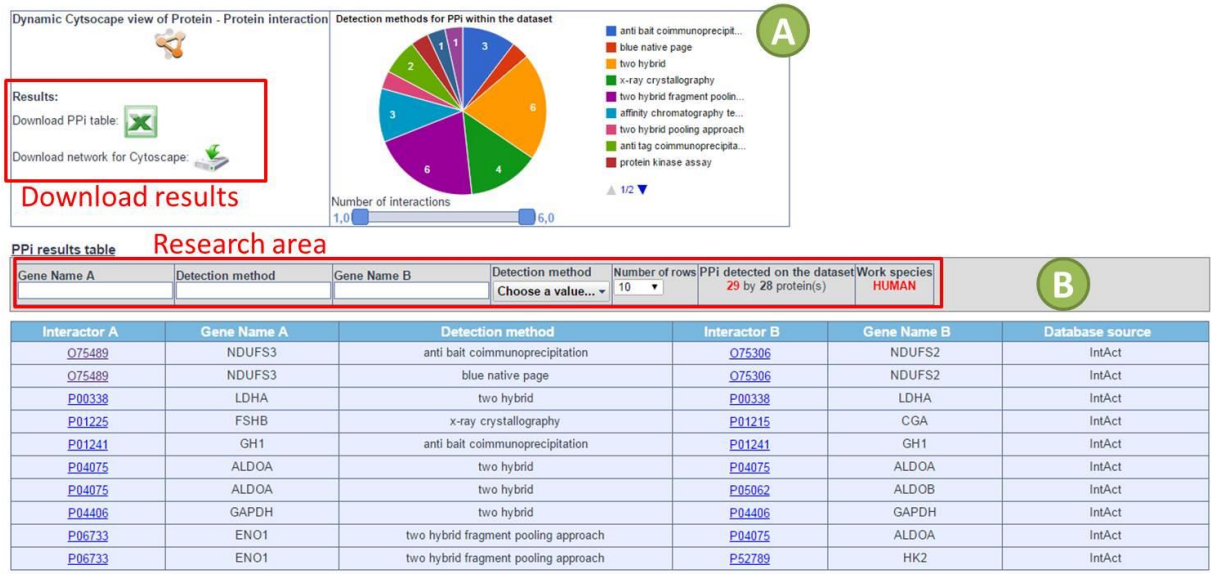

Figure 7. PPi results and visualisation. Results for PPi are summarised as a downloadable table and a diagram. (A) Main results are downloadable as table and network file that can be visualized using a network viewer (as Cytoscape). An online network view (made using the Cytoscape web application) is also proposed from this page result. A pie diagram indicates the number of PPi identified with the different detection methods. (B) A dynamic table lists linked proteins within the dataset, the detection method used to identify the interaction, and the database source of the interaction.

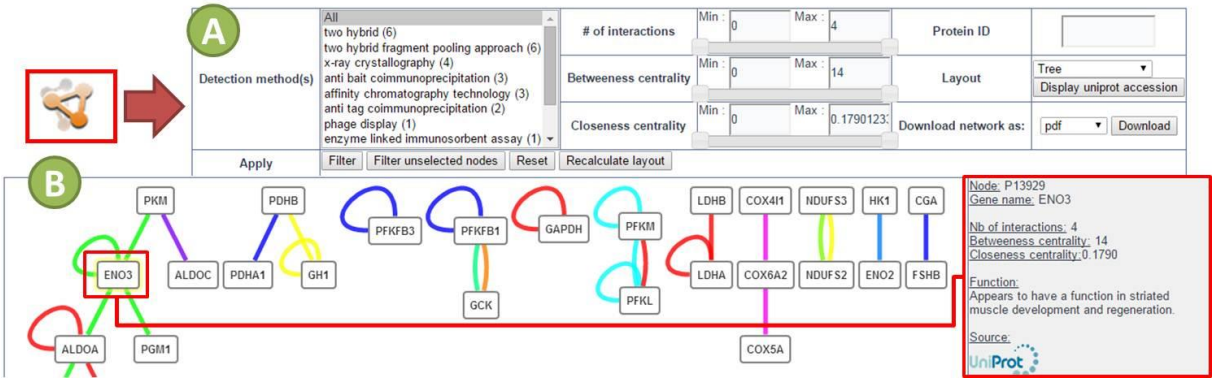

Figure 8. Network visualization of the PPi results. (A) This menu provides options to filter the network by: detection method, number of interactions for a protein, type of layout, protein ID, or the values of centralities. The centralities values are useful to sort large networks and to view only a central subnetwork. The betweenness centrality quantifies how frequently a node is on the shortest path between every pair of nodes for detecting bottlenecks in a network. The closeness centrality quantifies how distant minimal paths are from a given node to all others, a large closeness indicates that a node is close to the topological center of the network. (B) The network view is a dynamic image where user can access to a protein data by clicking on a node (name, function, statistic results, and database source and link of the protein).

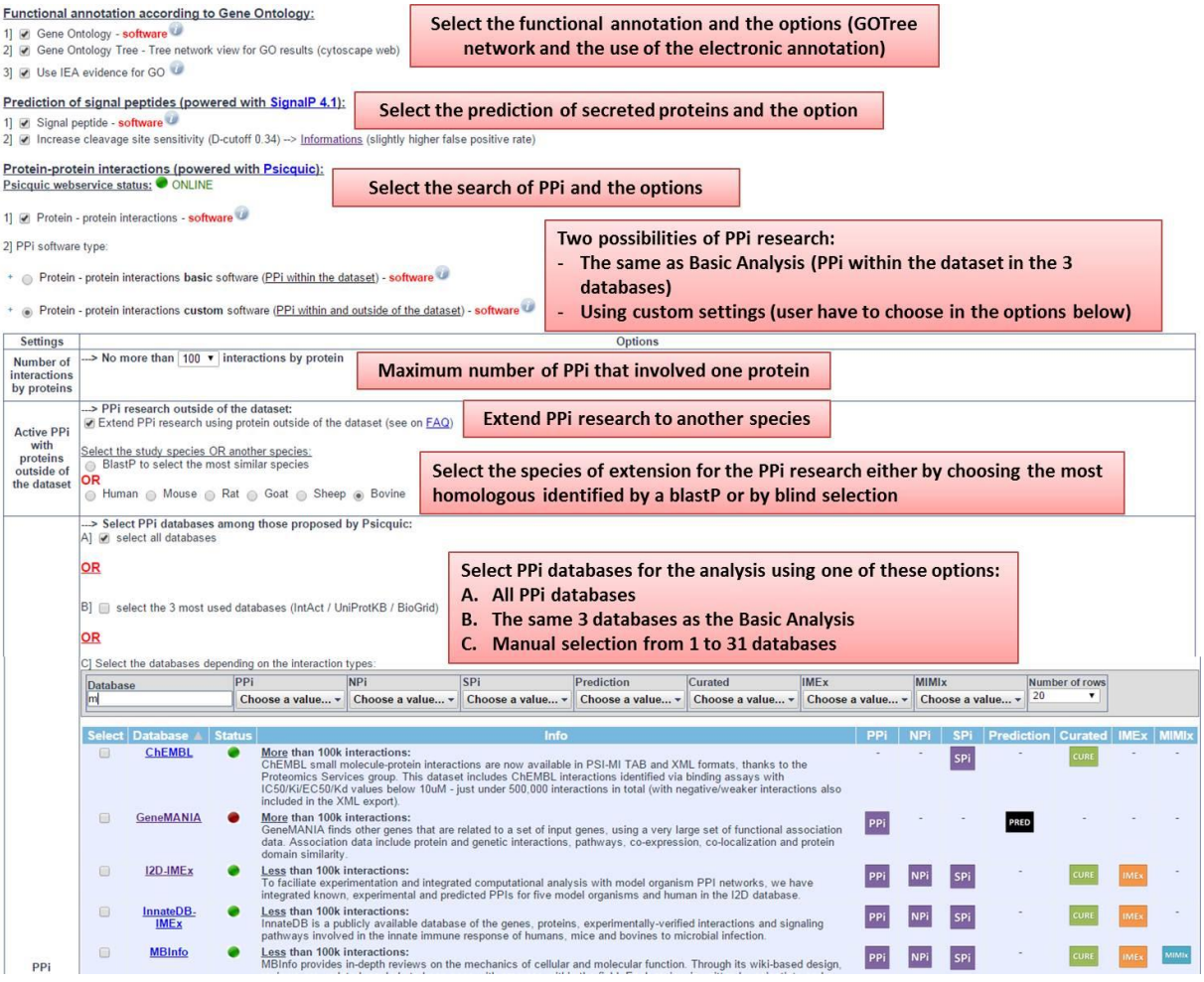

Figure 9. Setting up a custom analysis. Firstly, user has to enter a name for the analysis, select the species of study, and directly paste the input IDs (Figure 2). User has to select settings of the analysis: the setting followed by "software" mention activates the corresponding module in the workflow, and then user can select options for chosen module(s).

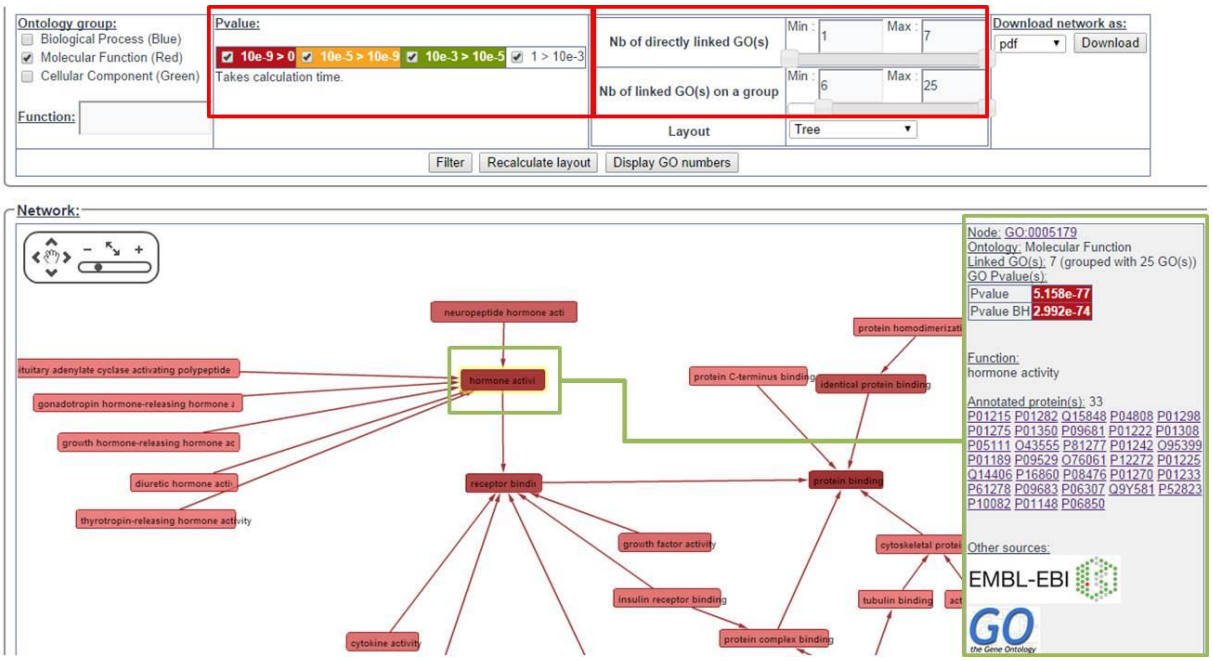

Figure 10. GOTree network visualization. Linked GO terms which annotate the dataset are linked using ancestor chart method. Each edge means that a term A is a subtype of a term B (is_a). Information about a GO is obtained by clicking on the GO or the node. Red colour is only for the GO terms relative to the Molecular Function. The degree of colour saturation is related to the number of proteins annotated by a GO (dark and clear for high and low numbers, respectively).

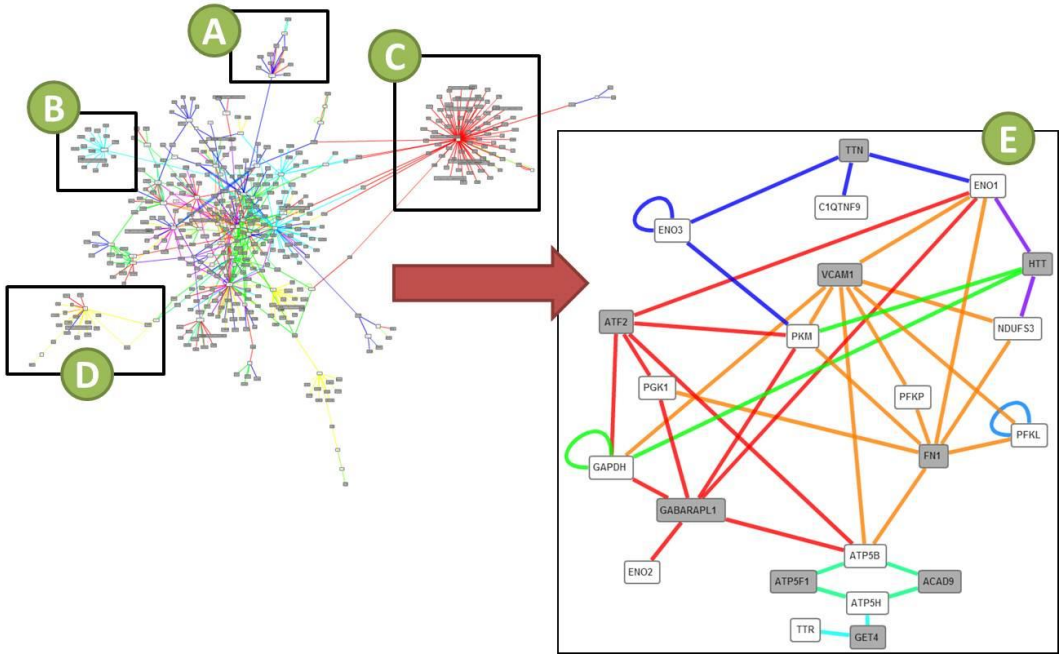

Figure 11. Extended network of PPi with proteins outside of the dataset. This network is made of PPi retrieved by querying the BioGrid, UniProt, and IntAct databases and using PPi with human proteins outside of the dataset. Grey squares are for proteins outside the dataset; white proteins are from the dataset. We have highlighted linked proteins that are involved in pathways such as: (A) glycolysis, (B) hormone activity, (C) the growth hormone signalling, and (D) thyroid hormones signalling. (E) We have used high values of betweenness and closeness centralities (BC: 3600; CC: 0.2) to get the most central proteins of this extended network.

REFERENCES

1. Chaze T, Meunier B, Chambon C, Jurie C, Picard B. **Proteome dynamics during contractile and metabolic differentiation of bovine foetal muscle**. Animal : an international journal of animal bioscience. 2009;3(7):980-1000. doi:10.1017/S1751731109004315.
2. Picard B, Cassar-Malek I, Guillemin N, Bonnet M. **Quest for Novel Muscle Pathway Biomarkers by Proteomics in Beef Production**. In: Moo-Young M, editor. Comprehensive Biotechnology (Second Edition). Burlington: Academic Press; 2011. p. 395-405.
3. Taga H, Chilliard Y, Meunier B, Chambon C, Picard B, Zingaretti MC et al. **Cellular and molecular large-scale features of fetal adipose tissue: is bovine perirenal adipose tissue brown?** Journal of cellular physiology. 2012;227(4):1688-700. doi:10.1002/jcp.22893.
4. Schmidt A, Forne I, Imhof A. **Bioinformatic analysis of proteomics data**. BMC Syst Biol. 2014;8 Suppl 2:S3. doi:10.1186/1752-0509-8-S2-S3.
5. Wong AK, Park CY, Greene CS, Bongo LA, Guan Y, Troyanskaya OG. **IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks**. Nucleic Acids Res. 2012;40(Web Server issue):W484-90. doi:10.1093/nar/gks458
6. Pache RA, Ceol A, Aloy P. **NetAligner--a network alignment server to compare complexes, pathways and whole interactomes**. Nucleic Acids Res. 2012;40(Web Server issue):W157-61. doi:10.1093/nar/gks446.
7. Renaud Y, Baillif A, Perez JB, Agier M, Mephu Nguifo E, Mirouse V. **DroPNet: a web portal for integrated analysis of Drosophila protein-protein interaction networks**. Nucleic Acids Res. 2012;40(Web Server issue):W134-9. doi:10.1093/nar/gks434.
8. Tuncbag N, McCallum S, Huang SS, Fraenkel E. **SteinerNet: a web server for integrating 'omic' data to discover hidden components of response pathways**. Nucleic Acids Res. 2012;40(Web Server issue):W505-9. doi:10.1093/nar/gks445
9. Ramirez F, Lawyer G, Albrecht M. **Novel search method for the discovery of functional relationships**. Bioinformatics. 2012;28(2):269-76. doi:10.1093/bioinformatics/btr631.
10. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R et al. **The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools**. Nucleic Acids Res. 2012;40(Database issue):D1202-10. doi:10.1093/nar/gkr1090.
11. Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martinez C et al. **EcoCyc: fusing model organism databases with systems biology**. Nucleic Acids Res. 2013;41(Database issue):D605-12. doi:10.1093/nar/gks1027.
12. Chen J, Bardes EE, Aronow BJ, Jegga AG. **ToppGene Suite for gene list enrichment analysis and candidate gene prioritization**. Nucleic Acids Res. 2009;37(Web Server issue):W305-11. doi:10.1093/nar/gkp427.
13. Kandasamy K, Mohan SS, Raju R, Keerthikumar S, Kumar GS, Venugopal AK et al. **NetPath: a public resource of curated signal transduction pathways**. Genome biology. 2010;11(1):R3. doi:10.1186/gb-2010-11-1-r3.
14. Nikitin A, Egorov S, Daraselia N, Mazo I. **Pathway studio--the analysis and navigation of molecular networks**. Bioinformatics. 2003;19(16):2155-7.
15. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM et al. **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. Nature genetics. 2000;25(1):25-9. doi:10.1038/75556.
16. Petersen TN, Brunak S, von Heijne G, Nielsen H. **SignalP 4.0: discriminating signal peptides from transmembrane regions**. Nature methods. 2011;8(10):785-6. doi:10.1038/nmeth.1701.
17. Aranda B, Blankenburg H, Kerrien S, Brinkman FS, Ceol A, Chautard E et al. **PSICQUIC and PSISCORE: accessing and scoring molecular interactions**. Nature methods. 2011;8(7):528-9. doi:10.1038/nmeth.1637.
18. Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD. **Cytoscape Web: an interactive web-based network browser**. Bioinformatics. 2010;26(18):2347-8. doi:10.1093/bioinformatics/btq430.
19. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. **Cytoscape 2.8: new features for data integration and network visualization**. Bioinformatics. 2011;27(3):431-2. doi:10.1093/bioinformatics/btq675.
20. Kaspric N. PB, Reichstadt M., Tournayre J., Bonnet M. , editor. **Protein function easily investigated by genomics data mining using the ProteINSIDE web service**. Proceeding of the 2nd International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO); 2014.
21. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. **QuickGO: a web-based tool for Gene Ontology searching**. Bioinformatics. 2009;25(22):3045-6. doi:10.1093/bioinformatics/btp536.
22. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. **Predicting subcellular localization of proteins based on their N-terminal amino acid sequence**. J Mol Biol. 2000;300(4):1005-16. doi:10.1006/jmbi.2000.3903.
23. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F et al. **The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases**. Nucleic Acids Res. 2014;42(Database issue):D358-63. doi:10.1093/nar/gkt1115.
24. UniProt C. **Activities at the Universal Protein Resource (UniProt)**. Nucleic Acids Res. 2014;42(11):7486. doi:10.1093/nar/gku469.
25. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D et al. **The BioGRID interaction database: 2015 update**. Nucleic Acids Res. 2015;43(Database issue):D470-8. doi:10.1093/nar/gku1204.
26. Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S et al. **Protein interaction data curation: the International Molecular Exchange (IMEx) consortium**. Nature methods. 2012;9(4):345-50. doi:10.1038/nmeth.1931.
27. Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, Oesterheld M, Stumpflen V et al. **The minimum information required for reporting a molecular interaction experiment (MIMIx)**. Nature biotechnology. 2007;25(8):894-8. doi:10.1038/nbt1324.
28. Coordinators NR. **Database resources of the National Center for Biotechnology Information**. Nucleic Acids Res. 2014;42(Database issue):D7-17. doi:10.1093/nar/gkt1146.
29. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S et al. **AmiGO: online access to ontology and annotation data**. Bioinformatics. 2009;25(2):288-9. doi:10.1093/bioinformatics/btn615.
30. Al-Shahrour F, Diaz-Uriarte R, Dopazo J. **FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes**. Bioinformatics. 2004;20(4):578-80. doi:10.1093/bioinformatics/btg455.
31. Benjamini Y, Hochberg Y. **Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing**. J Roy Stat Soc B Met. 1995;57(1):289-300.
32. Tore O. **Structure and Evolution of Weighted Networks**. 2009.
33. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. **Basic local alignment search tool**. J Mol Biol. 1990;215(3):403-10. doi:10.1016/S0022-2836(05)80360-2.
34. Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, de Castro E et al. **ExPASy: SIB bioinformatics resource portal**. Nucleic Acids Res. 2012;40(Web Server issue):W597-603. doi:10.1093/nar/gks400.

35. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C et al. **The Bioperl toolkit: Perl modules for the life sciences**. Genome research. 2002;12(10):1611-8. doi:10.1101/gr.361602.
36. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C et al. **The IntAct molecular interaction database in 2012**. Nucleic Acids Res. 2012;40(Database issue):D841-6. doi:10.1093/nar/gkr1088.
37. Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C et al. **The BioGRID interaction database: 2013 update**. Nucleic Acids Res. 2013;41(Database issue):D816-23. doi:10.1093/nar/gks1158.
38. Huntley RP, Binns D, Dimmer E, Barrell D, O'Donovan C, Apweiler R. **QuickGO: a user tutorial for the web-based Gene Ontology browser**. Database : the journal of biological databases and curation. 2009;2009:bap010. doi:10.1093/database/bap010.
39. Emanuelsson O, Brunak S, von Heijne G, Nielsen H. **Locating proteins in the cell using TargetP, SignalP and related tools**. Nat Protoc. 2007;2(4):953-71. doi:10.1038/nprot.2007.131.
40. Caccia D, Dugo M, Callari M, Bongarzone I. **Bioinformatics tools for secretome analysis**. Biochim Biophys Acta. 2013;1834(11):2442-53. doi:10.1016/j.bbapap.2013.01.039.
41. Nickel W. **The mystery of nonclassical protein secretion. A current view on cargo proteins and potential export routes**. Eur J Biochem. 2003;270(10):2109-19.
42. Martha VS, Liu Z, Guo L, Su Z, Ye Y, Fang H et al. **Constructing a robust protein-protein interaction network by integrating multiple public databases**. BMC Bioinformatics. 2011;12 Suppl 10:S7. doi:10.1186/1471-2105-12-S10-S7.
43. Hwang S, Son SW, Kim SC, Kim YJ, Jeong H, Lee D. **A protein interaction network associated with asthma**. Journal of theoretical biology. 2008;252(4):722-31. doi:10.1016/j.jtbi.2008.02.011.
44. Blake JA, Bult CJ. **Beyond the data deluge: data integration and bio-ontologies**. Journal of biomedical informatics. 2006;39(3):314-20. doi:10.1016/j.jbi.2006.01.003.
45. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W et al. **Big data: The future of biocuration**. Nature. 2008;455(7209):47-50. doi:10.1038/455047a.
46. Gobeill J, Pasche E, Vishnyakova D, Ruch P. **Managing the data deluge: data-driven GO category assignment improves while complexity of functional annotation increases**. Database-Oxford. 2013. doi:10.1093/database/bat041.
47. Carnielli CM, Winck FV, Paes Leme AF. **Functional annotation and biological interpretation of proteomics data**. Biochim Biophys Acta. 2015;1854(1):46-54. doi:10.1016/j.bbapap.2014.10.019.
48. Huang da W, Sherman BT, Lempicki RA. **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**. Nucleic Acids Res. 2009;37(1):1-13. doi:10.1093/nar/gkn923.
49. Maere S, Heymans K, Kuiper M. **BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks**. Bioinformatics. 2005;21(16):3448-9. doi:10.1093/bioinformatics/bti551.
50. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. **GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists**. BMC Bioinformatics. 2009;10:48. doi:10.1186/1471-2105-10-48.
51. Rice P, Longden I, Bleasby A. **EMBOSS: the European Molecular Biology Open Software Suite**. Trends in genetics : TIG. 2000;16(6):276-7.
52. Hiller K, Grote A, Scheer M, Munch R, Jahn D. **PrediSi: prediction of signal peptides and their cleavage positions**. Nucleic Acids Res. 2004;32(Web Server issue):W375-9. doi:10.1093/nar/gkh378.
53. Kall L, Krogh A, Sonnhammer EL. **Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server**. Nucleic Acids Res. 2007;35(Web Server issue):W429-32. doi:10.1093/nar/gkm256.
54. Hernandez-Toro J, Prieto C, De las Rivas J. **APID2NET: unified interactome graphic analyzer**. Bioinformatics. 2007;23(18):2495-7. doi:10.1093/bioinformatics/btm373.
55. Tarcea VG, Weymouth T, Ade A, Bookvich A, Gao J, Mahavisno V et al. **Michigan molecular interactions r2: from interacting proteins to pathways**. Nucleic Acids Res. 2009;37(Database issue):D642-6. doi:10.1093/nar/gkn722.
56. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E et al. **MINT, the molecular interaction database: 2012 update**. Nucleic Acids Res. 2012;40(Database issue):D857-61. doi:10.1093/nar/gkr930.
57. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A et al. **STRING v9.1: protein-protein interaction networks, with increased coverage and integration**. Nucleic Acids Res. 2013;41(Database issue):D808-15. doi:10.1093/nar/gks1094.
58. Zhang QC, Petrey D, Garzon JI, Deng L, Honig B. **PrePPI: a structure-informed database of protein-protein interactions**. Nucleic Acids Res. 2013;41(Database issue):D828-33. doi:10.1093/nar/gks1231.
59. Huang DW, Sherman BT, Lempicki RA. **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. Nat Protoc. 2009;4(1):44-57. doi:10.1038/nprot.2008.211.
60. Medina I, Carbonell J, Pulido L, Madeira SC, Goetz S, Conesa A et al. **Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling**. Nucleic Acids Res. 2010;38(Web Server issue):W210-3. doi:10.1093/nar/gkq388.
61. McCarthy FM, Gresham CR, Buza TJ, Chouvarine P, Pillai LR, Kumar R et al. **AgBase: supporting functional modeling in agricultural organisms**. Nucleic Acids Res. 2011;39(Database issue):D497-506. doi:10.1093/nar/gkq1115.

ENDNOTES

[a] http://www.cbs.dtu.dk/services/SignalP/performance.php
[b] http://www.cbs.dtu.dk/services/TargetP/instructions.php