Protein function easily investigated by genomics data mining using the ProteINSIDE web service

Nicolas KASPRIC[1, 2], Brigitte PICARD[1, 2], Matthieu REICHSTADT[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

Corresponding authors: nicolas.kaspric@clermont.inra.fr & muriel.bonnet@clermont.inra.fr

Abstract. ProteINSIDE is a new workflow to analyse lists of protein or gene identifiers from ruminant species and gather biological information provided by functional annotations, putative secretion of proteins and proteins interactions networks. ProteINSIDE gets results from several software and databases with a single query. From a unique list, ProteINSIDE uses orthologs identifiers within well studied species (Human, Rat or Mouse) to extend analyses and biological information retrieval. ProteINSIDE is freely available at: http://www.proteinside.org.

Keywords: web service, workflow, protein-protein interaction, protein secretion, gene ontology, networks.

1 Introduction

The current challenge for scientists working on the efficiency of ruminant (cattle, sheep or goat) and the quality of their products (meat, milk...) is to understand which genes and proteins control nutrient metabolism and partitioning between tissues or which genes and proteins control tissues growth and physiology [1]. Such questioning leads to the genome annotation, the sequencing and the quantification of gene expression or protein abundance. The quantity of data produced by these genomic and proteomic studies increases continuously [2-4]. There is a necessity to analyse, understand and generate biological information and knowledge from these data [5]. This is possible by using a panel of tools requiring different identifiers (IDs) per protein or gene and time to read and analyse the results. Moreover, most databases (DB) like UniProtKB [6] or NCBI [7] possess a large quantity of information and most of existing bioinformatic tools implemented as web services are specific to one analysis: as the annotation according to the Gene Ontology (GO) [8] or the prediction of signal peptide [9] or the molecular interactions identification [10] and visualization as networks [11, 12]. Many workflows that integrate several analyses are available [13-16]

and are specific to a species (Drosophila, *Arabidopsis thaliana, Escherichia coli*…), and thus are not suitable for the analysis of genomic data from ruminant species. The few workflows working with ruminant data are multispecies, the results are not species-specific and the data source is not available because of the privacy of databases (as the license software Pathway Studio [17]). Other workflows are specialized on the identification of candidate genes related to diseases as ToppGene [18]. Thus, to date there is no workflow dedicated to the integrative analysis of genomic data from ruminant species.

Unlike Human or model species like mouse or rat, ruminant species are less annotated and protein sequences are not always verified. Often, scientists use orthologs with the aim to increase the meaningful biological contexts for proteins. For this purpose biologists query for annotations according to Gene Ontology, the putative secretion of proteins, protein-protein interactions (PPi) and network analysis first in ruminant and then in Human or in rodents. The integration within a workflow of gateways between proteins / genes from ruminants and their orthologs from Human and models species has never been done.

Here we propose ProteINSIDE, a web service dedicated to a systematic and integrative analysis of protein's biological information. ProteINSIDE works using lists of proteins or genes IDs from 6 species (Bovine, Ovine, Caprine, Human, Rat, and Murine) to annotate functions and cellular location, predict secreted proteins, search for interactions between proteins within and/or outside a dataset and allowing cross-species analysis using orthologs.

2 Materials and methods

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

2.1 Equipment

ProteINSIDE doesn't require an installation on a computer and the web service is available online at www.proteinside.org by using an internet browser. ProteINSIDE is completely adapted for any internet browser, but for better performances we recommend to use Firefox, Chrome, or Safari.

2.2 Implementation

ProteINSIDE is divided into three parts: the workflow, the database and the web interface. The workflow is a combination of Perl (version 5.10.1; CPAN modules (Comprehensive Perl Archive Network) used and BioPerl [19]) and R (version 3.0.1; with "tnet" package [20]) scripts to query databases, recover protein data, perform calculations and run algorithms for signal peptide predictions and network visualisation. The MySQL database aims to reduce server load and thus stores both available knowledge form major public databases and results (and settings) form queries. The

web interface is programmed in PHP, HTML, and JavaScript. It is devoted to the creation of a new analysis, the view of results and users information with updates.

2.3 ProteINSIDE structure and interface

A flow chart (Fig 1) details the type of analysis (basic or customizable) and the four main queries proposed to the user. Whatever the type of analysis, the workflow uses data from the input file and runs default scripts (basic analysis) or scripts and options selected through the settings (customs analysis). At analysis completion, results are created and uploaded on ProteINSIDE database to decrease web interface treatment duration (results have to be deleted by the user; visitors results are automatically deleted monthly).

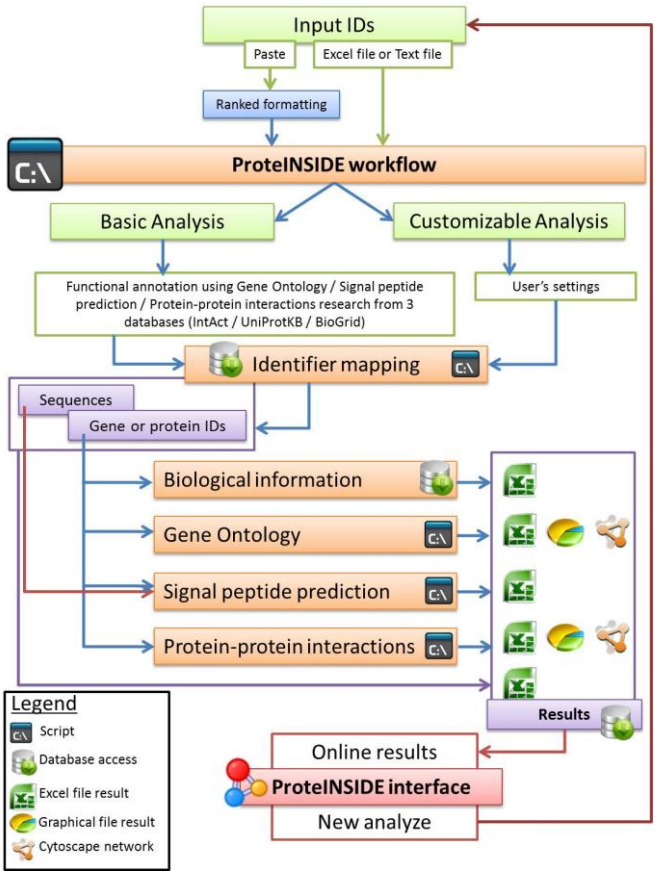


Fig. 1. Flow chart of ProteINSIDE structure. The four modules for querying the available biological information, annotations according to the gene ontology, signal peptide predictions and protein-protein interactions are either all present in the basic analysis or individually selected in the custom analysis.

ProteINSIDE is easily run by biologist through the interface. Registered user or visitor run a new analysis by using the web interface menus “Basic Analysis” (automatic settings) or “Custom Analysis” (user selects the settings):

1. Click on “Basic Analysis” menu on the homepage of ProteINSIDE
2. Fill in “the job name” box
3. Select the species for the analysis (related to the IDs that will be used on this query)
4. Upload your input file or directly paste your IDs
5. Click on the “Run the job” button to submit a new analysis

ProteINSIDE gives a link and an access code to view analysis status and get the results. 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 completion”. The blue globe is the link to access to the online results views:

1. Click on the blue globe button to view the results (use the trash to delete them)
2. Visualise the results summary produced by selected modules on the first default page
3. Navigate to module’s results pages by clicking on the module’s name on the toolbar menu.

2.4 The input and the output of ProteINSIDE

ProteINSIDE inputs are genes or proteins IDs (e.g. ADIPO or ADIPO_HUMAN) or UniProtKB protein accession numbers (e.g. Q15848). These IDs are uploaded as text tabulated files (extension .tab or .txt) or as Excel files (.xls or .xlsx). The input files have to be ranked as three columns (Fig. 2) because of the database format. Alternatively, the IDs are directly pasted.

	A	B	C
1	1	1	A4D1N9
2	2	2	ADIPO_HUMAN
3	3	3	C1T9A_HUMAN
4	4	4	F261_HUMAN
5	5	5	F262_HUMAN
6	6	6	PFKFB3
7	7	7	PFKFB4

Fig. 2. Example of an input files made using Excel 2010 and formatted for an upload.

The output files are Excel file (.xls), Cytoscape file (.cys or .xgmml), text or FASTA file (.txt or .fa) and pictures (.jpg or .png or .pdf). They are downloadable from the page results of each module of analysis.

2.5 The sample dataset

We created a dataset to assess ProteINSIDE performances. This dataset is composed of the UniProtKB accession numbers of 133 proteins (Table 1): 34 proteins related to the glycolysis cycle, 11 proteins from the respiratory chain, 5 proteins from the tricar-

boxylic acid cycle, 79 hormones or secreted proteins and proteins with very specific functions unrelated to the others. We also included a duplicated protein among proteins of the glycolysis to verify its recognition by ProteINSIDE.

We created this dataset on bovine species, but the numbers of annotations and PPi weren't sufficient for a clear representation of the functionalities of ProteINSIDE. Then, we used the same proteins in Human to test ProteINSIDE with the "Basic" and the "Custom Analysis" (Table 1).

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) functions, and that have been predicted as secreted by SignalP for hormones.

Analyses and data	Glycolysis	Hormones	TCA	Analysis time (min)
Dataset	33+1 (duplicate)	79	5	-
Basic Analysis	29	78	3	2
Custom Analysis	33	78	5	10

3 Results and discussion

Here we present the results produced by a "Basic Analysis" and a "Custom Analysis" from our sample dataset, and we discuss the relevance of biological information extracted by ProteINSIDE. All of the 133 proteins were recognized by ProteINSIDE, the protein in duplicate was identified and excluded from the analysis (Fig. 3). Thus, 132 proteins were submitted to the analyses. The numbers of proteins / genes submitted to the analyses, numbers of annotations, PPi and predicted secreted proteins are recorded on the default page following the access to the results (Fig. 3).

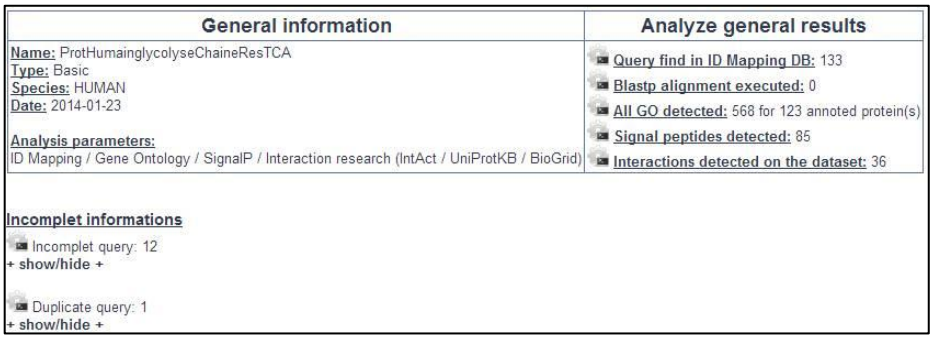


Fig.3. A table is provided by the default page after the access to the results both for a basic and a custom analysis. In addition to general information, the table provide counts of results retrieved by each module that has been run, incomplete query (IDs with missing biological information) and duplicate query.

3.1 Results of the Basic Analysis

The first module of analysis has extracted and summarized, as a downloadable table, other gene or protein IDs, gene or protein names, a summary for the protein function, the gene chromosomal location, information on tissue expression and cellular location, and the species in which orthologs have been identified. Thus on the "ID resume" page of the toolbar menu, a user has access at a glance to several information for a list of genes or their products, and also to direct links with the UniProtKB and the NCBI databases.

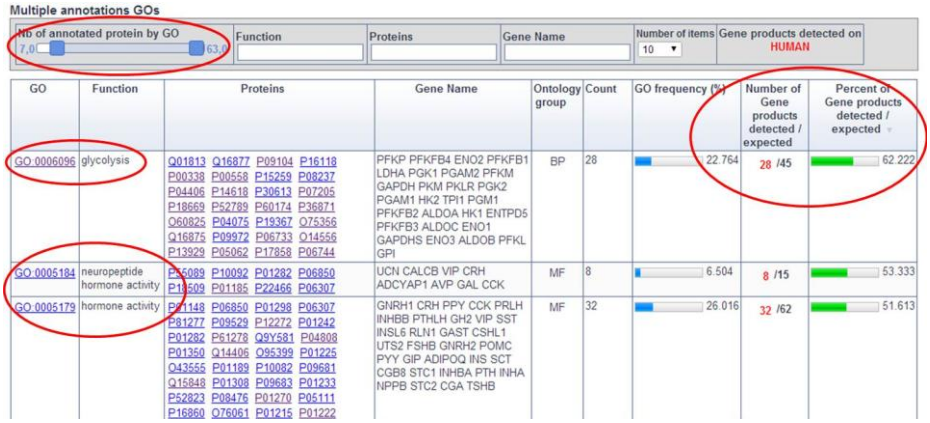
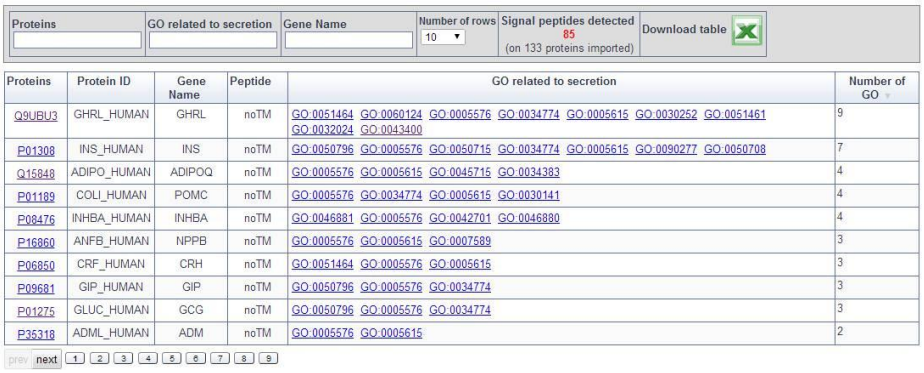


Fig. 4. Results of the functional annotations according to the Gene Ontology are available as dynamic tables. Results can be sorted by: the GO's identifier, the function, protein ID or gene name, the ontology group, the number of annotated proteins or the number of expected gene products.

On the "GO" page of the toolbar menu, we checked the relevance of the annotations extracted by ProteINSIDE by looking for the over-representation of annotations relative to glycolysis and hormones. First, among the 132 proteins submitted, ProteINSIDE annotates 123 proteins with 568 unique GOs (Fig 3). We classed these GO according to the number of proteins annotated by GO and the percentage of gene products detected/expected to identify the most common pathways associated to our sample dataset (Fig. 4). By this way, we retrieved as the most common pathways: glycolysis and hormone activity (about 62% and 52% of expected annotated gene products with these GO in Human, respectively). We have to note a lack of annotations for 12 proteins of the sample dataset, and a lack of annotations relative to glycolysis for 4 proteins (28 of the 33 expected proteins related to the glycolysis were annotated; Table 1). This lack of annotations is related to our choice to use only GO terms that have been agreed by review curator in the "Basic Analysis". This means that the "Basic Analysis" doesn't use GO annotations with IEA (Inferred by Electronic Annotation) evidence code, but the option to use IEA is provided in the custom analysis to extend the annotations.

On the "Secreted protein" page of the toolbar menu, the proteins potentially secreted are listed in a dynamic table (Fig. 3 and 5). From our sample dataset, 85 proteins were predicted as secreted by SignalP [9], among them 78 of the 79 proteins that were expected (Table 1). This lack of perfect prediction of the protein can be explained by the false positive and false negative prediction rates of SignalP, as already evaluated [21]. The prediction of secretion is then confirmed by a search for GOs related to the "secretion" function. Over the 85 predicted secreted proteins, 63 were annotated by GOs related to the "secretion" function. The double query of protein secretion both by the peptide signal prediction from protein sequence and the GO annotation is unique to ProteINSIDE.



Proteins | GO related to secretion | Gene Name | Number of rows 10 | Signal peptides detected 85 (on 133 proteins imported) | Download table

Proteins	Protein ID	Gene Name	Peptide	GO related to secretion	Number of GO
Q9UBU3	GHRL_HUMAN	GHRL	noTM	GO:0051464 GO:0060124 GO:0005576 GO:0034774 GO:0005615 GO:0030252 GO:0051461 GO:0032024 GO:0043400	9
P01308	INS_HUMAN	INS	noTM	GO:0050796 GO:0005576 GO:0050715 GO:0034774 GO:0005615 GO:0090277 GO:0050708	7
Q15848	ADIPO_HUMAN	ADIPOQ	noTM	GO:0005576 GO:0005615 GO:0045715 GO:0034383	4
P01189	COLI_HUMAN	POMC	noTM	GO:0005576 GO:0034774 GO:0005615 GO:0030141	4
P08476	INHBA_HUMAN	INHBA	noTM	GO:0046881 GO:0005576 GO:0042701 GO:0046880	4
P16860	ANFB_HUMAN	NPPB	noTM	GO:0005576 GO:0005615 GO:0007589	3
P06850	CRF_HUMAN	CRH	noTM	GO:0051464 GO:0005576 GO:0005615	3
P09681	GIP_HUMAN	GIP	noTM	GO:0050796 GO:0005576 GO:0034774	3
P01275	GLUC_HUMAN	GCG	noTM	GO:0050796 GO:0005576 GO:0034774	3
P35318	ADML_HUMAN	ADM	noTM	GO:0005576 GO:0005615	2

prev next 1 2 3 4 5 6 7 8 9

Fig. 5. Results as a dynamic table, of the potentially secreted proteins predict by SignalP.

On the "Protein interactions" page of the toolbar menu, proteins within the dataset are linked by the "Interaction detection methods" or reviewed by a curator (clicking on node gives information about the protein and a link to UniProtKB database). We selected to query BioGrid [22], UniProtKB [6] and IntAct [23] because these PPi databases are reviewed by curators and the query of PPi in 2 or 3 PPi database delivered best results [24]. PPi are listed by a dynamic table or viewed as a network (Fig. 6). The interactions research between proteins of our sample dataset has identified 36 PPi that involved 23 different proteins. As expected, PPi within the sample dataset linked proteins known to contribute to the pyruvate dehydrogenase complex (Fig 6A), the complexes IV (Fig 6B) and I (Fig 6C) of the respiratory chain, and also some proteins linked to the glycolysis and the carbohydrate oxidation (Fig 6D and 6E).

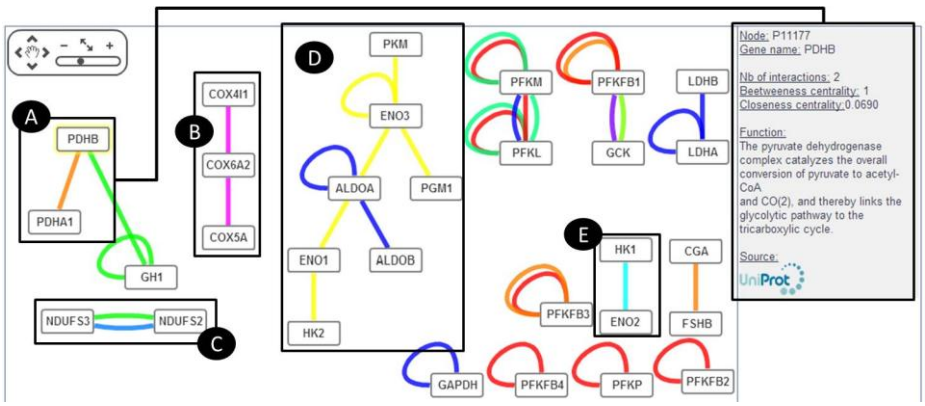


Fig. 6. Network of PPi retrieved by querying BioGrid, UniProtKB and IntAct databases. PPi are between proteins within the sample dataset. Information about a protein is obtained by clicking on a protein/gene or node. Edge colour depends on the detection method of the PPi.

3.2 The Custom Analysis: an added-value provided by the extension of the analysis

We made a "Custom Analysis" using the same major settings as the "Basic Analysis" but we used the proposed additional options:

1. The GO electronic annotation (IEA) evidence codes proposed to extend the annotation
2. GOTree network to view linked GOs
3. The search and the view of the PPi between proteins from our sample dataset and proteins outside the dataset (but still in the same species, here in Human) to extend the network and the biological information.

The use of electronic annotation has increased both the number of annotated proteins (132 rather than 123 without IEA in the basic analysis) and of annotations by around 50% since 1031 unique GOs were retrieved by ProteINSIDE. Over the 33 expected proteins related to the glycolysis, the Custom analysis of ProteINSIDE has annotated 32 proteins with the GO 0006096, glycolysis (Table 1). The GOTree network linked 236 GOs. We have chosen to visualize the GOs of the "Molecular Function" group (Fig. 7). In this visualisation, the dark red colour represents the most common GO associated to our sample dataset. As expected the GO:0005179, nominates "Hormone activity", which is consistent with the over-representation of hormones in our sample dataset. This network has also linked more specific GOs or child terms [25] of the "Hormone activity" GO, as for example "Neuropeptide hormone activity" (GO:0005184).

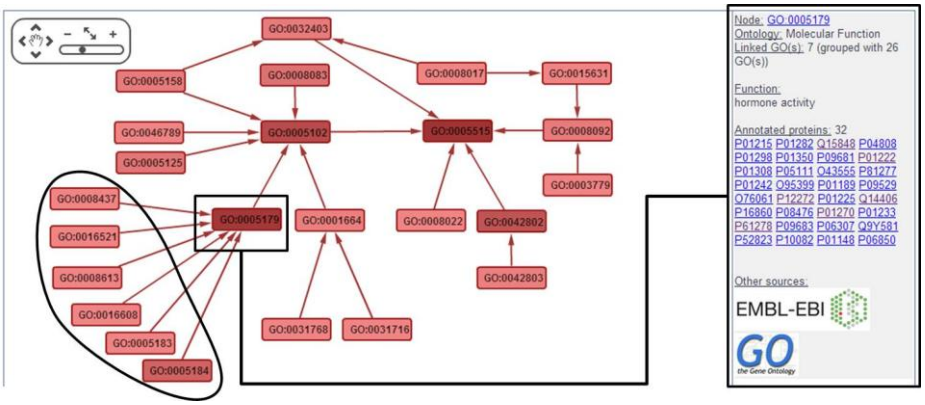


Fig. 7. A Network that links GOs used to annotate proteins of the sample dataset. Red colour is only for the GO terms relative to the Molecular Function. The degree of colour saturation represents the quantity of proteins annotated by a GO. Each edge means that a term A is a subtype of a term B (is_a). Information about a GO is obtained by clicking on the GO or the node.

Lastly, the same proteins as the "Basic Analysis" were predicted as secreted (Table 1). Thanks to the IEA electronic annotation, 82 proteins over the 85 proteins predicted to be secreted by SignalP, were also annotated by GOs related to the "secretion" function. It's 19 more than the 63 of the "Basic Analysis" because of the use of IEA evidence code for GO annotation.

By comparison with "Basic Analysis", "Custom Analysis" searches for PPi between the proteins within and outside the sample dataset by querying up to 28 DB. We chose to query the same 3 DB (BioGrid, UniProtKB and IntAct) to compare with the "Basic Analysis". ProteINSIDE retrieved 616 PPi made by 221 proteins among them 61 from the dataset. We visualized the network of the PPi (Fig. 8) and we retrieved some sub-networks relative to the respiratory chain (Fig. 8A), hormone activity such as signalization by adipokines (Fig. 8B), the growth hormone (Fig. 8C) and thyroid hormones (Fig. 8D), as well as sub-networks relative to glycolysis and carbohydrate metabolism (not highlighted).

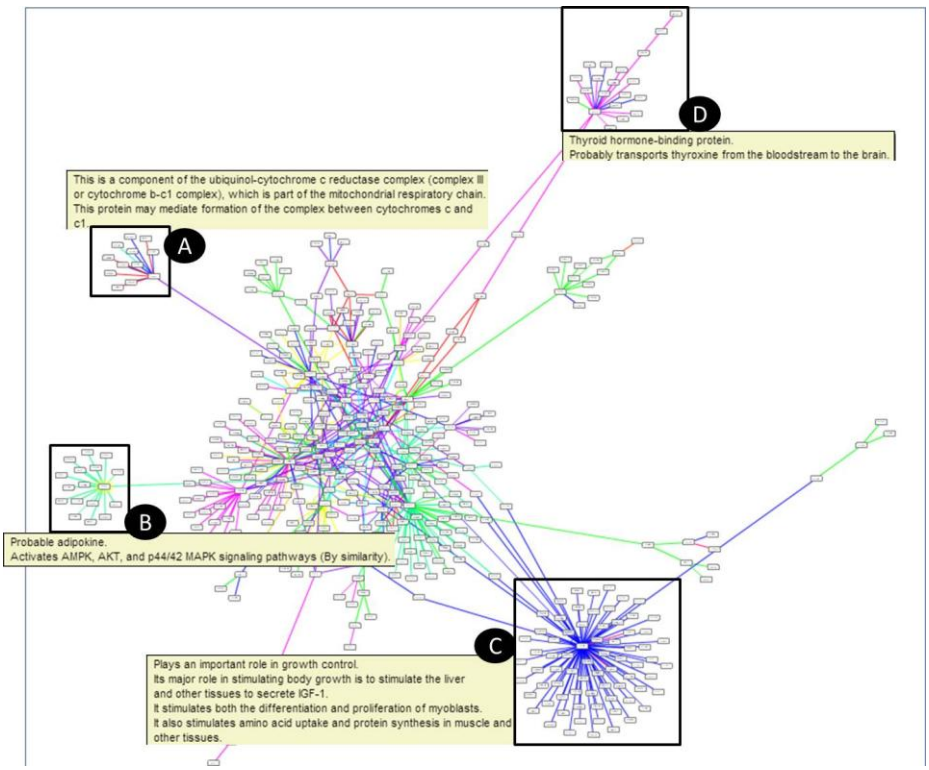


Fig. 8. Network of PPi retrieved by querying the BioGrid, UniProtKB and IntAct DB and using PPi with human proteins outside of the dataset.

4 Conclusion

In this work we present 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). The presented web service has correctly identified a dataset of 133 proteins, has excluded a duplicate query and has retrieved biological information for each protein. According to our dataset, ProteINSIDE properly annotates the proteins related to the glycolysis, the proteins affiliated as hormones, and the putatively secreted proteins. ProteINSIDE has revealed the most common pathways related to our dataset by creating networks from PPI interactions within and outside the dataset and from links between GOs. Each result is easily accessible and downloadable.

ProteINSIDE offers a great support to analyse a large quantity of data from genomic and proteomic studies. ProteINSIDE is also the unique web service that makes all of these analyse using ruminant IDs.

Acknowledgment

This work was supported by the Region Auvergne (FRANCE) and Apis-Gene (FRANCE).

References

1. Bonnet M, Cassar-Malek I, Chilliard Y, Picard B. "Ontogenesis of muscle and adipose tissues and their interactions in ruminants and other species". Animal : an international journal of animal bioscience, 4(7), pp. 1093-109. PubMed PMID: 22444612. (2010)
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. 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, 3(7), pp. 980-1000. PubMed PMID: 22444818. (2009)
4. 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, 227(4), pp. 1688-700. PubMed PMID: 21678425. (2012)
5. Woelders H, Te Pas MF, Bannink A, Veerkamp RF, Smits MA. "Systems biology in animal sciences". Animal : an international journal of animal bioscience, 5(7), pp. 1036-47. PubMed PMID: 22440099. (2011)
6. Magrane M, Consortium U. "UniProt Knowledgebase: a hub of integrated protein data". Database : the journal of biological databases and curation, 2011, pp. bar009. PubMed PMID: 21447597. Pubmed Central PMCID: 3070428. (2011)
7. Coordinators NR. "Database resources of the National Center for Biotechnology Information". Nucleic Acids Res, 41(Database issue), pp. D8-D20. PubMed PMID: 23193264. Pubmed Central PMCID: 3531099. (2013)
8. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. "QuickGO: a web-based tool for Gene Ontology searching". Bioinformatics, 25(22), pp. 3045-6. PubMed PMID: 19744993. Pubmed Central PMCID: 2773257. (2009)
9. Petersen TN, Brunak S, von Heijne G, Nielsen H. "SignalP 4.0: discriminating signal peptides from transmembrane regions". Nature methods, 8(10), pp. 785-6. PubMed PMID: 21959131. (2011)
10. Aranda B, Blankenburg H, Kerrien S, Brinkman FS, Ceol A, Chautard E, et al. "PSICQUIC and PSISCORE: accessing and scoring molecular interactions". Nature methods, 8(7), pp. 528-9. PubMed PMID: 21716279. Pubmed Central PMCID: 3246345. (2011)

11. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. "Cytoscape 2.8: new features for data integration and network visualization". Bioinformatics, 27(3), pp. 431-2. PubMed PMID: 21149340. Pubmed Central PMCID: 3031041. (2011)
12. Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD. "Cytoscape Web: an interactive web-based network browser". Bioinformatics, 26(18), pp. 2347-8. PubMed PMID: 20656902. Pubmed Central PMCID: 2935447. (2010)
13. 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, 40(Web Server issue), pp. W484-90. PubMed PMID: 22684505. Pubmed Central PMCID: 3394282. Epub 2012/06/12. eng. (2012)
14. Pache RA, Ceol A, Aloy P. "NetAligner--a network alignment server to compare complexes, pathways and whole interactomes". Nucleic Acids Res, 40(Web Server issue), pp. W157-61. PubMed PMID: 22618871. Pubmed Central PMCID: 3394252. Epub 2012/05/24. eng. (2012)
15. 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, 40(Web Server issue), pp. W134-9. PubMed PMID: 22641854. Pubmed Central PMCID: 3394298. Epub 2012/05/30. eng. (2012)
16. 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, 40(Web Server issue), pp. W505-9. PubMed PMID: 22638579. Pubmed Central PMCID: 3394335. Epub 2012/05/29. eng. (2012)
17. Nikitin A, Egorov S, Daraselia N, Mazo I. "Pathway studio--the analysis and navigation of molecular networks". Bioinformatics, 19(16), pp. 2155-7. PubMed PMID: 14594725. (2003)
18. Chen J, Bardes EE, Aronow BJ, Jegga AG. "ToppGene Suite for gene list enrichment analysis and candidate gene prioritization". Nucleic Acids Res, 37(Web Server issue), pp. W305-11. PubMed PMID: 19465376. Pubmed Central PMCID: 2703978. (2009)
19. 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, 12(10), pp. 1611-8. PubMed PMID: 12368254. Pubmed Central PMCID: 187536. (2002)
20. T. O. " Structure and Evolution of Weighted Networks". University of London (Queen Mary College), London, UK, pp. 104-22. (2009)
21. Emanuelsson O, Brunak S, von Heijne G, Nielsen H. "Locating proteins in the cell using TargetP, SignalP and related tools". Nat Protoc, 2(4), pp. 953-71. PubMed PMID: 17446895. (2007)
22. Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, et al. "The BioGRID interaction database: 2013 update". Nucleic Acids Res, 41(Database issue), pp. D816-23. PubMed PMID: 23203989. Pubmed Central PMCID: 3531226. (2013)
23. 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, 40(Database

issue), pp. D841-6. PubMed PMID: 22121220. Pubmed Central PMCID: 3245075. (2012)
24. 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, 12 Suppl 10, pp. S7. PubMed PMID: 22165958. Pubmed Central PMCID: 3236850. (2011)
25. 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, 25(1), pp. 25-9. PubMed PMID: 10802651. Pubmed Central PMCID: 3037419. (2000)