Proteomics data from ruminants easily investigated using ProteINSIDE

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

Farm animal proteomics 2014

Proceedings of the 5th Management Committee Meeting and
4th Meeting of Working Groups 1, 2 & 3 of COST Action FA 1002
Milano, Italy – 17-18 November 2014

edited by:
André de Almeida
Fabrizio Ceciliani
David Eckersall
Ingrid Miller
Jenny Renaut
Paola Roncada
Romana Turk

cost
EUROPEAN COOPERATION
IN SCIENCE AND TECHNOLOGY

Proteomics data from ruminants easily investigated using ProteINSIDE

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, 63122 Saint-Genès-Champanelle, France; nicolas.kaspric@clermont.inra.fr; muriel.bonnet@clermont.inra.fr

[2]Clermont Université, VetAgro Sup, UMR1213 Herbivores, BP 10448, 63000, Clermont-Ferrand, France

Objectives

A main challenge for scientists working on the efficiency of ruminant production 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 (Bonnet *et al.*, 2010). Their researches have produced huge amounts of genomic data requiring a lot of bioinformatics analyses to extract meaningful biological context for proteins or genes in ruminants. A strategy to increase the efficiency and the robustness of data mining from ruminant genomics data is to develop an online workflow that integrates several analysis steps in one package as it has been partly done in Human (BioMyn web service (Ramirez *et al.*, 2012)). Here, we present ProteINSIDE, an online workflow to analyse lists of proteins or genes from ruminant species by gathering biological information provided by an overview of data from myriad of databases, annotations according to Gene Ontology (GO), the prediction of tissue secretome and proteins interactions networks.

Methods

ProteINSIDE is an online tool, freely available at www.proteinside.org by using an internet browser. A flow chart (Figure 1) details the proposed basic or customizable analyses and the four main modules of the workflow. Whatever the type of analysis, the workflow uses data from the input file and runs default scripts (basic analysis) or scripts according to settings selected by the user (customs analysis). The four modules of analysis were developed as follow:

- 'Identifier mapping' module is an overview of available biological information thank to an assembly of reviewed biological data from UniProt and NCBI databases, and available on the 'ID resume' web page
- 'Gene Ontology' module queries QuickGO (Huntley *et al.*, 2009) database, with an option to unselect (basic analysis) or to select (custom analysis) electronic annotations. Then it ranks the over- represented GO terms by comparing the frequencies of a GO within a dataset and within the genome by a Fisher exact test. The most linked GO terms are viewed as a GOTree chart that is an ordered tree layout network provided by the custom analysis

- 'Secreted proteins' module searches for signal peptide on protein's fasta sequences thanks to SignalP (Petersen *et al.*, 2011) and TargetP (Emanuelsson *et al.*, 2000) algorithms. To support this prediction, ProteINSIDE selects GO terms related to secretion processes. The GO terms are also analysed to predict proteins that are secreted by processes that do not involve signal peptides (Nickel, 2003)
- 'Protein-protein interactions (PPi)' module searches for PPi recorded among the 26 PPi databases of Psicquic (Aranda *et al.*, 2011). ProteINSIDE only imports reviewed and curated PPi, and constructs PPi networks using an online cytoscape view (Lopes *et al.*, 2010) either between proteins of the dataset or between proteins of the dataset and outside of the dataset (in same species or using orthologous gene products)

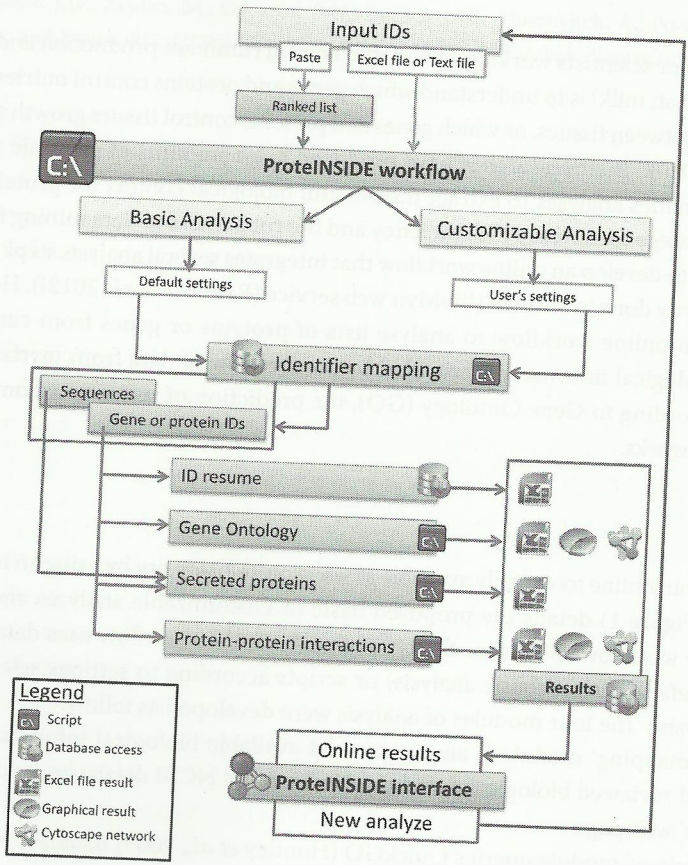

Figure 1. Flow chart of ProteINSIDE structure. The four modules for querying the available biological information, annotations according to the gene ontology, predictions of secreted proteins and protein-protein interactions are either all selected in the basic analysis or individually selected with specific settings in the custom analysis.

All results from ProteINSIDE's modules are viewed on the web page or can be downloaded. The output files are Excel file (.xls), Cytoscape file (.png or .pdf or .graphml or .xgmml), text or FASTA file (.txt or .fa) and pictures (.jpg or .png or .pdf).

ProteINSIDE inputs are proteins or genes identifiers (IDs) or names (e.g. ADIPO or ADIPO_HUMAN or e.g. gi|62022275) or UniProt protein accession number (e.g. Q15848) from 6 species (bovine, ovine, caprine, human, rat, and murine). To test ProteINSIDE's performances, we used a dataset composed of 133 proteins: 34 proteins related to the glycolysis cycle, 11 proteins from the respiratory chain, 5 proteins from the tricarboxylic 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 evaluation of the functionalities of ProteINSIDE. Then, we used the Bovine IDs for an analysis that use knowledge available in Human and ProteINSIDE automatically converted IDs from Bovine to Human using known orthologous gene products.

Results and discussion

Results from the 'Basic Analysis' shown that 133 proteins were recognized by ProteINSIDE, the protein in duplicate was identified and excluded from the analysis. Thus, 132 proteins were submitted to the analyses.

The ID Resume module of ProteINSIDE extracted and summarized biological information about each protein of the dataset. Results are available on the 'ID Resume' of the tool bar menu and are summarized in a table with: proteins and genes IDs, biological function, subcellular location, tissue specificity and chromosome location of the gene.

Among the 132 proteins submitted, ProteINSIDE annotates 123 proteins with 584 GO terms. GO terms are classed by the most significant p-value, by this way we retrieved the over-represented pathways expected for our dataset: GO terms relative to hormone activity and glycolytic process (that annotate respectively 33 and 27 proteins of the dataset). We have to note a lack of annotation for 12 proteins of the sample dataset, and a lack of annotation relative to glycolysis for 4 proteins (28 of the 33 expected proteins related to the glycolysis were annotated). 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' (no annotation with IEA (Inferred by Electronic Annotation) evidence code). All proteins were annotated when we selected the use of IEA for GO in custom analysis.

ProteINSIDE predicts 85 as potentially secreted. Among the 85 proteins, 81 are confirmed by TargetP and 65 of them are both confirmed by subcellular location and GO terms annotation related to a secretory pathway. By merging shared results from the 3 analysis, we get 78 of the 79 proteins expected on our dataset. ProteINSIDE also predicted 31 proteins as potentially secreted by signal peptide-independent pathways.

The interaction researches between proteins of our sample dataset on 3 databases (BioGrid, IntAct, UniProt) have identified 29 PPi that involved 23 different proteins. As expected, PPi within our dataset linked proteins known to contribute to the pyruvate dehydrogenase complex, the complexes IV and I of the respiratory chain, and also some proteins linked to the glycolysis and the carbohydrate oxidation.

Similar results were obtained from bovine IDs submitted to analyses in bovine species with a custom analysis with settings that used electronic annotation and increased the numbers of PPi databases queried (BioGrid, IntAct, MINT, MatrixDB, STRING, Reactome, InnateDB, I2D and UniProt). This highlights that available data for ruminant species are poorly curated and mainly inferred from electronic annotations. Thus for ruminant IDs, we recommended to proceed to both a custom analysis in the ruminant species and to a basic analysis in a well annotated species (as Rat, Mouse or Human) and to compare the results.

Conclusion

In this work we present the performances of ProteINSIDE, a new powerful workflow which gather 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 known as hormones, and the putatively secreted proteins. ProteINSIDE has revealed the most common pathways related to our dataset by creating networks from PPi within the dataset. 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 analyses using ruminant IDs.

Acknowledgements

This work was supported by the regional council of Auvergne in France and APIS-GENE.

References

Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F.S., Ceol, A., Chautard, E., Dana, J.M., De Las Rivas, J., Dumousseau, M., Galeota, E., Gaulton, A., Goll, J., Hancock, R.E., Isserlin, R., Jimenez, R.C., Kerssemakers, J., Khadake, J., Lynn, D.J., Michaut, M., O'Kelly, G., Ono, K., Orchard, S., Prieto, C., Razick, S., Rigina, O., Salwinski, L., Simonovic, M., Velankar, S., Winter, A., Wu, G., Bader, G.D., Cesareni, G., Donaldson, I.M., Eisenberg, D., Kleywegt, G.J., Overington, J., Ricard-Blum, S., Tyers, M., Albrecht, M. and Hermjakob, H., 2011. PSICQUIC and PSISCORE: accessing and scoring molecular interactions. Nat Methods 8: 528-529.

Bonnet, M., Cassar-Malek, I., Chilliard, Y. and Picard, B., 2010. Ontogenesis of muscle and adipose tissues and their interactions in ruminants and other species. Animal 4: 1093-1109.

Emanuelsson, O., Nielsen, H., Brunak, S. and von Heijne, G., 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300: 1005-1016.

Huntley, R.P., Binns, D., Dimmer, E., Barrell, D., O'Donovan, C. and Apweiler, R., 2009. QuickGO: a user tutorial for the web-based Gene Ontology browser. Database (Oxford) 2009: bap010.

Lopes, C.T., Franz, M., Kazi, F., Donaldson, S.L., Morris, Q. and Bader, G.D., 2010. Cytoscape Web: an interactive web-based network browser. Bioinformatics 26: 2347-2348.

Nickel, W., 2003. The mystery of nonclassical protein secretion. A current view on cargo proteins and potential export routes. European Journal of Biochemistry 270: 2109-2119.

Petersen, T.N., Brunak, S., von Heijne, G. and Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8: 785-786.

Ramirez, F., Lawyer, G. and Albrecht, M., 2012. Novel search method for the discovery of functional relationships. Bioinformatics 28: 269-276.

Proteomics is the large-scale study of the proteome, i.e. a set of proteins being expressed in a certain fluid, tissue, organ or organism. The value of this advanced technology is being recognised in farm animal and veterinary sciences from 'farm to fork'. The potential of proteomics is unequivocal in holding a significant promise in applications such as vaccine and drug development, physiology, toxicology, animal product quality and food safety. Proteomics has been growing steadily during the last 3-4 years and, as time goes by, proteomics-based studies are more and more common, not just to scientists but to the general public as well, unravelling the full potential of this innovative technology.

This book reflects the will of a group of multi-disciplinary scientists that merge innovation with excellence of research and to whom the dissemination of knowledge and discovery through cooperation is a key point. It is of interest to scientists at the early stages of their careers as well as to researchers well established in the field and to whom proteomics may be the necessary next step towards more in-depth research activities. By providing a collection of diverse scientific interests, *Farm Animal Proteomics 2014* demonstrates the vitality of the area and the importance it holds to animal and food research, to science, industry, government agencies, the consumer and ultimately the society as a whole.

ISBN 978-90-8686-262-7

ESF provides the COST Office through an EC contract. COST is supported by the EU RTD Framework Programme

Wageningen Academic Publishers