# PROTIC workshop: A bioinformatics environment for proteome data validation, analysis and integration

Renaud Flores, Johann J. Joets, Olivier Langella, Benoit B. Valot, Michel M. Zivy, Antoine de Daruvar, L. Gil, D. Jacob, Jean-Paul Bouchet, Mireille Faurobert, et al.

GENOPLANTE

AGENCE NATIONALE DE LA RECHERCHE ANR

# 2010 PLANT GENOMICS Seminar

Pierre & Vacances – Pont-Royal en Provence
Domaine et Golf de Pont-Royal
Mallemort – France
March 29-31, 2010

Scientific Director : Georges Pelletier, INRA
Organization : M. Barbou, D. Laborde, A. Sadi, I. Treton

| | | |
|---|---|---|
| GENOPLANTE ®<br>P025 | Thematic: New Tools<br>**ANR n°: ANR-07- GPLA- 009**<br>**Acronym: PROTICws**<br>Total Cost: 896 K€<br>Total Grant: 314 K€<br>Date of beginning: 01.01.2008 – End:31.12.2010 | ANR<br>**Edition 2007** |

# PROTIC workshop: A bioinformatics environment for proteome data validation, analysis and integration

**R. Flores[1], J. Joets[1] (coordinator), O. Langella[1], B. Valot[1], M. Zivy[1], A. de Daruvar[2], L. Gil[2], D. Jacob[2], J-P. Bouchet[3], M. Faurobert[3], D. Jeannin[3], K. Leyre[3], C. Lalanne[4], C. Plomion[4], D. Vincent[4].**

*[1] UMR de Génétique Végétale du Moulon, INRA, Univ ParisSud, CNRS, AgroParisTech, GifsurYvette (91)*
*[2] Centre de Bioinformatique de Bordeaux, Université Victor Segalen Bordeaux 2, Bordeaux (33)*
*[3] UR 1052, GAFL - Unité de Génétique et Amélioration des Fruits et Légumes, Montfavet (84)*
*[4] UMR BIOGECO, INRA, Cestas (33)*

## Aims

Because of the massive data generated by proteomic approaches, it was crucial to develop appropriate databases. Many bioinformatic projects endeavoured to fulfil such a need and, to date, a large body of data is being stored in labs databases. Yet, exploitation and valorisation of these data remain difficult.

Due to a prominent automated production mode, proteomic data must be validated and/or curated before being analyzed. This process, mostly manually performed, represents a bottleneck mainly resulting from a lack of suitable tools integrated into databases and aiming at facilitating expert inputs.

Numerous statistical methods exist and it is often necessary to combine several of them to extract as much information as possible from a given dataset. Statistical analyses may be cumbersome and time-consuming for scientists non familiar with them. Moreover, data downloading from databases can be complex and prone to errors. These issues are emphasized when one aims at simultaneously analyzing data generated by transcriptomic, proteomic and metabolomic experiments.

Finally, it is essential for the validated, cleaned and analyzed data to be accessible to as many scientists as possible. Consequently, proteomic databases should be integrated into international network databases, such as the « World-2D page » hosted by the Bioinformatics Swiss Institute.

To this end, we have proposed to develop PROTICws, a bioinformatics environment dedicated to validation, analysis and sharing of proteomic data. This novel tool will be based on PROTICdb2, already used by the partners of this project

## Results

### Development of PROTICstat, a statistical analysis environment

PROTICdb harbors many quantitative data regarding protein quantification as well as a detailed description of sample origin and treatment. We have developed the PROTICstat module that offer end-user to analyze data from PROTICdb through several statistical workflows. The first step is to extract dataset from the database thanks to the X2DBI module. Analysis results are available as tables or graphics. PROTICstat allow end user to develop its own R-based statistical workflows.

### Development of PROTICannotate, a curation and validation tool for mass spectrometry data

Whether proteins have been separated or not by methods such 2DE, scientists need to link each protein sample with a set of protein identifications. These protein identifications always result from the interpretation of mass spectrometry data obtained on peptides, compared to proteins reference databases. It often is important to be able to reanalyze mass spectrometry data with new proteins reference databases as more and more data are available. PROTICdb stores all these data and scientists need to annotate them and decide which identifications they consider as valid. With PROTICannotate they can easily explore the database and annotate spectra, peptides, protein identifications and mass spectrometry samples. Each annotation is the combination of a value selected from lists of controlled vocabulary terms, comments, author and date. Previous annotations are kept and may be consulted.

### Development of framework for automated functional annotation of biological sequences

Where possible, identification of protein corresponding to mass spectra are based on the most reliable sequence database available; UniProt SwissProt. This guarantees that the identification will come with the state of the art functional annotation. However, for most non-model organism SwissProt is of no help as very few sequences are available. As a consequence, identifications may be based on poorly annotated or not annotated at all sequences (EST. Predicted genes, ...). In such case it could be of interest to automatically annotate the sequences retained for identifications. Rather than devised a unique annotation pipeline, we have made the choice to integrate a workflow management tool within PROTICws. Users will be able to customize easily the workflows according their requirements. We are in the process of selecting a workflow framework. We are testing Taverna and Remora tools with a simple workflow including a search for domain with InterproScan.

Development of PROTICport, a set of web services for standardized programmatic access to PROTICdb

A set of web services have been developed to offer PROTICannotate and PROTICstat access to the PROTICdb database. These services are based on a library (API) that has been designed in order to allow for easy and quick development of any services required by external applications to extract or insert data in PROTICdb. To date two dozen of services are available.

In addition to integration of PROTICdb with PROTICstat and PROTICannotate modules, PROTICport also offer services that allow to integrate each instance of PROTICdb within the network of federative proteomics database organized by the SIB and queriable through the World-2Dpage web portal. We have developed several prototypes that are under evaluation at the SIB.

## Perspectives

PROTICannotate

Next versions of PROTICannotate will offer functions (i) to manage lists of terms used for expert annotation and (ii) to select subsets of data (MS samples, MS identification runs, proteins, peptides) and to apply them processes to make expert annotation easier.

PROTICstat

Coming months will be devoted to increase the level of integration between PROTICstat and the whole PROTICws interfaces.

PROTICport

New services will developed according to the needs of PROTICannotate and PROTICstat development teams.

Integration of PROTICws within the World-2Dpage will be achieved and tested by user from this project.

Automated sequence functional annotation

Once the workflow management technology will be implemented, several annotation workflow will be devised in coordination with the proteomic group of Gif, Bordeaux and Avignon.

A first release of PROTICws should be publicly available at the end of 2010.

## Publications / Congress

Congress :

Jean-Paul Bouchet, Antoine de Daruvar, Mireille Faurobert, Raphaël Flores' Daniel Jacob, Dominique Jeannin, Johann Joets, Céline Lalanne, Karine Leyre, Olivier Langella, Christophe Plomion, Patrick Rousselle, Benoît Valot, Delphine Vincent, Michel Zivy (2009) PROTIC workshop: A bioinformatics environment for proteome data validation, analysis and integration. Journées Ouvertes en Biologie, Informatique et Mathématiques Nantes France

Publication :

PROTICdb was used as data repository in this paper from a project of partner 4

Vincent D, Balesdent MH, Gibon J, Claverol S, Lapaillerie D, Lomenech AM, Blaise F, Rouxel T, Martin F, Bonneu M, Amselem J, Dominguez V, Howlett BJ, Wincker P, Joets J, Lebrun MH, Plomion C. Hunting down fungal secretomes using liquid-phase IEF prior to high resolution 2-DE. Electrophoresis. 2009 Dec;30(23):4118-36

## Total permanent scientist

Gif-sur-Yvette 4.5 ETP, Avignon 1.8 ETP, Bordeaux 1.8 ETP, Cestas 0.25 ETP

## Temporary contracts

| | | |
|---|---|---|
| Raphael Flores | Bioinformatics engineer | January 2009 – December 2010 |
| Dominique Jeannin | Bioinformatics engineer | July 2008 – June 2010 |
| Laurent Gil | Bioinformatics engineer | |