PROTEIOS: an open source proteomics initiative
Per Gårdén1,∗, Rikard Alm2 and Jari Häkkinen1

1Complex Systems Division and Lund Swegene Bioinformatics Facility, Department of Theoretical Physics, Lund University, SE-223 62 Lund, Sweden and 2Department of Biochemistry, Center for Chemistry and Chemical Engineering, Lund University, SE-221 00 Lund, Sweden

Received on November 25, 2004; revised on January 11, 2005; accepted on January 25, 2005

Advance Access publication February 3, 2005

ABSTRACT

Summary: PROTEIOS is an initiative for the development of a comprehensive open source system for storage, organization, analysis and annotation of proteomics experiments. The PROTEIOS platform is based on commonly acknowledged principles for proteomics data publishing.

Availability: http://www.proteios.org

Contact: per@thep.lu.se

INTRODUCTION

The need to organize proteomics data in a standardized form is greater than ever (Prince et al., 2004). The advantage of standards is not limited to software development, but extends to allow the seamless exchange of data between researchers, thus making data publication less strenuous. The Proteomics Standards Initiative (PSI) (http://psidev.sourceforge.net) (Orchard et al., 2003) together with manufacturers of mass spectrometry (MS) equipment have recognized these benefits within MS data and are consequently moving towards standardization. Currently, the PSI covers MS experimental data in the mzData (http://psidev.sourceforge.net) standard and in a related initiative the PSI also covers protein–protein interactions. Future development of PSI standards will cover the larger experimental context, including parts dealing with samples and protein identification.

After a standard has been developed, it will still take time for its adoption by the laboratories. Much effort is needed to effect data exchange and develop solutions for organizing data. In the experimental setting, the data of one experiment is aggregated from several inputs, where laboratory equipment typically cover only certain parts of an experiment. Moreover, the data output is difficult to handle as it contains superfluous data and does not comply with open standards.

PROTEIOS supports the user in managing and connecting data from heterogeneous sources with the aim of tracking all information relevant to an experiment—sample, processing, MS and protein identification. This scope sets it apart from other applications, most of which either focus on MS [e.g. Sashimi (http://sashimi.sourceforge.net), OPD (http://bioinformatics.icmb.utexas.edu/OPD) (Prince et al., 2004), PROTEOME-3d (Lundgren et al., 2003)] or do not enable automatized data capture [e.g. PEDRo (http://pedro.man.ac.uk) (Taylor et al., 2003)]. In this respect, PROTEIOS aims to become for proteomics what BASE (Saal et al., 2002), also maintained by our group, is for microarray research. As data formats for microarray experiments differ from that of proteomic experiments, existing microarray database platforms should probably not be used. Rather than tweaking proteomics data into a tool like BASE, one is better off creating separate applications and thus avoid compromises in data models.

PROTEIOS manages biomaterials information, raw data, images and analysis results, and provides integrated ‘plug-in’-able protein identification, data viewing and analysis tools. The organization and interface of PROTEIOS is designed to closely follow the natural work-flow of the proteomics researcher, and is today compatible with both LC–MSMS and 2D-gel experiments. Being an open source software, PROTEIOS can be used independently of equipment manufacturers and extended or modified to fit local needs.

THE APPLICATION

PROTEIOS is a client–server application, with a many-to-many relationship between clients and servers. This architecture, where PROTEIOS handles import and export of data to and from databases, makes it possible for researchers to share data with colleagues worldwide. PROTEIOS maintains data ownership and accessibility on each database as one unit using standard SQL privilege mechanisms. The PROTEIOS data model is implemented as an XML-schema and as database tables. XML-schemas prescribe the format for files which are directly importable to PROTEIOS. Although XML-files can be very large, XML has a great advantage in that it allows data to be validated. This is important since validation prevents corrupt data from being added into the database. Although XML-files can be very large, XML has a great advantage in that it allows data to be validated. This is important since validation prevents corrupt data from being added into the database. The mass spectrometry standard format mzData is also directly importable. Since PROTEIOS uses mzData to describe the MS part of the data. Furthermore, the sample generation and sample processing parts of PEDRo (Taylor et al., 2003) can be readily imported. Other imports exist (e.g. mzXML as raw files) and further ones are easily added.

PROTEIOS is implemented in Java and SQL, and is thus platform-independent. Specifically, the PROTEIOS client runs as a Java application on virtually any workstation and connects to server database(s) through the Hibernate (http://www.hibernate.org) middleware. Hibernate adds a database abstraction layer that supports most SQL database providers, enabling a wide range of databases to be used as PROTEIOS back-end servers. Currently, two PROTEIOS
client applications exist, a graphical user interface (GUI) and a batch handling client.

PROTEIOS graphical user interface

The GUI is the main PROTEIOS client and the common interface to view and analyse the data. It presents the data as graphical objects which makes data viewing easy and intuitive. By interacting with these objects, a user can also import and export data. The GUI enables the user to tie together data from different experimental sources in a project (Fig. 1).

The data presentation is a tree structure, which can be rearranged to highlight items of interest. This functionality is part of the very flexible data import and export feature. The data can be annotated and extended with, for instance, protein identifications from search engines like PIUMS (http://idelnx81.hh.se/bioinf/mass_spectro.html) and Mascot (http://www.matrixscience.com/) (Perkins et al., 1999).

PROTEIOS batch handling

The same functionality as in the GUI is provided for batch processing.

OUTLOOK

So far most of the effort has been on developing the data repository infrastructure, including validation capabilities, import and export of data, and enabling asynchronous entry of experimental data. The focus is now on extending analysis features and incorporating third party tools. Future development will include a stand-alone PROTEIOS server, which will enable web interaction tools to be connected.

PROTEIOS is evolving rapidly—new features are constantly added. At the same time the aim of PROTEIOS is to remain compatible with the upcoming PSI standards and turn them into useful functionalities. Among other things, future features will include interactability with more protein identification search tools [e.g. Mascot and Sequest (http://fields.scripps.edu/sequest) (Eng et al., 1994)], better support for plug-ins and ontology handling. PROTEIOS is freely available for download (including sample datasets) at the PROTEIOS website http://www.proteios.org under GPL (Gnu Public License (http://www.gnu.org/copyleft/gpl.html)). GPL allows anyone to use the PROTEIOS software free of charge. Restrictions may apply on redistribution and modification of the application.

ACKNOWLEDGEMENTS

This work is in part supported by the Knut and Alice Wallenberg Foundation through the Swegene consortium and by grants from FORMAS (22.6/2002-0042).
REFERENCES


