Proteios: an Open Source Proteomics Initiative

Per Gärdén*,[1], Rikard Alm†, and Jari Häkkinen*

*Complex Systems Division and Lund Swegene Bioinformatics Facility, Department of Theoretical Physics, Lund University, SE-223 62 Lund, Sweden

†Department of Biochemistry, Center for Chemistry and Chemical Engineering, Lund University, SE-221 00 Lund, Sweden

submitted to *Bioinformatics*

ABSTRACT

Summary: PROTEIOS is an initiative for the development of a comprehensive open source system for storage, organisation, 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 organise proteomics data in a standardised form is larger than ever (Prince *et.al.*, 2004). The advantage of standards is not only for software development, but standards also allow for seamless exchange of data between researchers and make data publication less strenuous. The Proteomics Standards Initiative (PSI)[2] (Orchard *et al.*, 2003) together with manufacturers of mass spectrometry equipment have recognised these benefits within mass spectrometry data and are consequently moving towards standardisation. Currently the PSI covers mass spectrometry (MS) experimental data in the mzData[2] 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.

[1]To whom correspondence should be addressed.
[2]http://psidev.sourceforge.net

Even after a standard has been developed, it will take time before it has been adopted by laboratories. Much effort has to be put on data exchange and developing solutions for organising data. In the experimental setting, the data of one experiment is aggregated from several inputs, where laboratory equipment typically covers 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 to track all information relevant to an experiment – sample, processing, mass spectrometry and protein identification. This scope sets it apart from other applications, most of which either focus on MS (e.g. Sashimi[3], OPD[4] (Prince *et.al.*, 2004), PROTEOME-3d (Lundgren, D.H *et al.*, 2003)) or do not enable automatised data capture (e.g. PEDRo[5] (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 avoids compromises in data models.

PROTEIOS manages biomaterials information, raw data, images, analysis results, and provides integrated "plug-in"-able protein identification, data viewing and analysis tools. The organisation 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 be extended or modified to fit local needs.

THE APPLICATION

PROTEIOS is a client-server application, with a many-to-many relationship between clients and

[3]http://sashimi.sourceforge.net
[4]http://bioinformatics.icmb.utexas.edu/OPD
[5]http://pedro.man.ac.uk

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. The mass spectrometry standard format mzData is also directly importable since PROTEIOS uses mzData to describe the MS part of 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[6] middle-ware. Hibernate adds a database abstraction layer that supports most SQL database providers, enabling a wide range of databases to be used as PROTEIOS backend servers.

Currently, two PROTEIOS client applications exist, a graphical user interface and a batch handling client.

Proteios graphical user interface. The graphical user interface (GUI) is the main PROTEIOS client and the common interface to view and analyse data. It presents data as graphical objects which make data viewing easy and intuitive. 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.

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. Data can be annotated and extended with, for instance, protein identifications from search engines like PIUMS[7] and Mascot[8] (Perkins *et al.*, 1999).

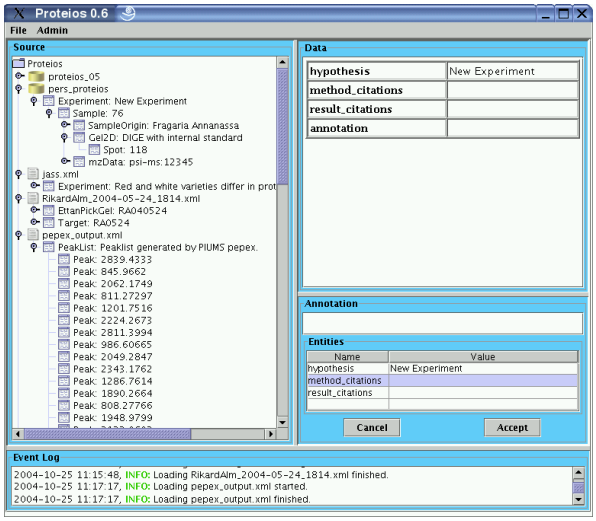

Figure 1: The PROTEIOS Graphical User Interface.

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

OUTLOOK

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

PROTEIOS is rapidly evolving - new features are constantly added. At the same time the aim of PROTEIOS is to remain compatible with upcoming PSI standards and turn them into useful functionality. Among other things, future features will include interactability with more protein identification search tools (e.g. Mascot[8] and Sequest[9] (Eng *et al.*, 1994)), better support for plug-ins as well as ontology handling.

PROTEIOS is freely available for download (in-

[6] http://www.hibernate.org

[7] http://idelnx81.hh.se/bioinf/mass_spectro.html

[8] http://www.matrixscience.com/

[9] http://fields.scripps.edu/sequest

cluding sample datasets) at the PROTEIOS web site http://www.proteios.org under GPL (Gnu Public License[10]). 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

Eng,J.K., McCormack,A.L. and Yates,J.R.,3rd (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, **5**, 976-989.

Lundgren,D.H., Eng,J, Wright,M.E. and Han,D.K. (2003) Proteome-3d: an interactive bioinformatics tool for large-scale data exploration and knowledge discovery. *Mol. Cell. Proteomics*, **2**. 1164-1176.

Orchard,S., Hermjakob,H. and Apweiler,R. (2003) The proteomics standards initiative. *Proteomics*,**3**, 1374-1376.

Perkins,D.N., Pappin, D.J., Creasy, D.M. and Cotrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551-3567.

Prince,J.T., Carlson,M.W., Wang,R., Lu,P. and Marcotte,E.M. (2004) The need for a public proteomics repository. *Nature Biotechnology*, **22**, 471-472.

Saal,L.H., Troein,C., Vallon-Christersson,J., Gruvberger,S., Borg,Å. and Peterson,C. (2002) Bioarray software environment: a platform for comprehensive management and analysis of microarray data. *Genome Biol.* **3**, software0003.1-0003.6.

Taylor,C.F., Paton,N.W., Garwood,K.L., Kirby,P.D., Stead,D.A., Yin,Z., Deutsch,E.W., Selway,L., Walker,J., Riba-Garcia,I., Mohammed,S., Deery,M.J., Howard,J.A., Dunkley,T., Aebersold,R., Kell,D.B., Lilley,K.S., Roepstorff,P., Yates 3rd.,J.R., Brass,A., Brown,A.J.P., Cash,P., Gaskell,S.J., Hubbard S.J. and Olover,S.G. (2003) A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nature Biotechnology*, **21**, 247-254.

[10]http://www.gnu.org/copyleft/gpl.html