

Managing Proteomics Data from Data Generation and Data Warehousing to Central Data Repository and Journal Reviewing

● ABRF 2008, V24-T

H. Thiele¹, P. Hufnagel¹, J. Glandorf¹, G. Körting², M. Blüggel², M. Eisenacher³, H.E. Meyer³, C. Stephan³
¹Bruker Daltonik GmbH, Bremen, Germany
²Protagen AG, Dortmund, Germany
³Medical Proteom-Center, University of Bochum, Germany

Introduction

The tremendous amount of data from today's expression proteomics requires a database solution with sophisticated data-warehousing and data-mining capabilities. **ProteinScope™** (co-developed by Bruker and Protagen, Germany) provides a bioinformatics platform for in-house proteome studies as well as for large scale approaches, like the human brain proteome project (HUPO BPP).

The growing requirement for protein pre-fractionation to obtain more precise quantitative protein information is uniquely addressed in the new generation of **ProteinScope**. Entire workflows of pre-fractionation, detailed LC/MS/MS separation and post-processing with bioinformatics tools are merged and can be easily controlled and reviewed.

Methods

Different Proteomics Workflows for Identification and Quantification

ProteinScope 2 is the first bioinformatics platform addressing the current requirements for biomarker discovery, protein identification and quantification. It supports various discovery workflows through a flexible analyte hierarchy, various database search engines and quantification approaches. All current label chemistries for protein quantification are fully supported (ICPL, SILAC, iTRAQ, ICAT, and C-term 18O/16O-C-term labeling) and the software is prepared to include future label technologies. The support includes multiplexed quantification (e.g., ICPL triplex, iTRAQ or SILAC 4plex).

It enables the use of isobaric or non-isobaric label chemistries and it permits the targeted analysis of proteins in complex mixtures. Interactive validation of protein quantification based on raw LC/MS data is now simple and straight forward.

Visualization

ProteinScope has a number of dedicated viewers that permit the evaluation and validation on each level of proteomics experiments, such as the LC/MS survey viewer, the gel viewer and sequence annotated MS/MS spectra. All these views are linked and permit simple browsing through the proteomics data in the current projects and even allow retrieval of data generated years ago, allowing their joint reanalysis with novel capabilities and mining tools.

Validation

BioTools integrates with ProteinScope for advanced sequence validation, PTM discovery, de novo sequencing and MS-BLAST searches for full structure elucidation functionalities. Through integrated access to the **WARP-LC** Package quantification workflows that utilize labeling technologies combined with protein separation require greatly reduced analysis and validation time.

Identification & Characterization:

Standardized Data Processing Pipeline

The processing pipeline adopts the HUPO Brain (HUPO BPP) processing guidelines (forum.hbpp.org) and will facilitate the direct submission process of Proteomic project data adhering to HUPO/PSI publishing guidelines.

Part of this strategy is the new algorithm **ProteinExtractor** to combine peptide lists from different MS/MS search-engines to a combined protein result list. Protein list validation is based on the **Decoy Database Concept**.

ProteinExtractor

The mapping of peptides to proteins is not a one-to-one mapping, but often leads to ambiguities. A set of rules has been developed for ProteinExtractor in order to define a minimal protein list, which contains only those proteins (and protein variants) which can be distinguished by the MS/MS data. An iterative approach proved to be successful. ProteinExtractor uses only spectra, the assigned peptides and peptide scores as input. This gives the opportunity to create protein lists with the same algorithm and conditions regardless which search engine was used. ProteinExtractor has successfully been used and tested with Mascot, Phenyx, Sequest and ProteinSolver.

Data Integration:

Standardized Submission Pipelines

The European Commission-funded ProDaC consortium (Proteomics Data Collection, <http://www.fp6-prodac.eu/>) will finalize data storage and documentation standards, implement conversion tools and establish standardized submission pipelines into a central data repository. This contains export from local LIMS systems like ProteinScope to standard file formats or direct upload into PRIDE. With respect to this a tool is already implemented to upload ProteinScope data sets into PRIDE.

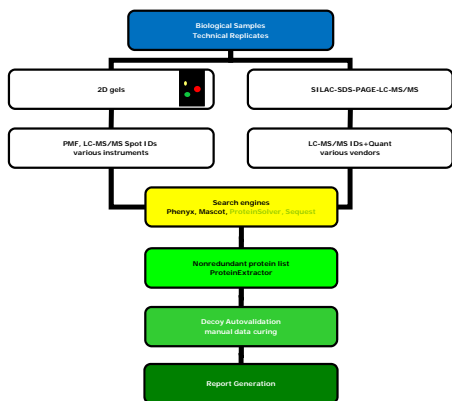


Fig. 1 The processing pipeline of ProteinScope adopts the HUPO HBPP standard guidelines for data processing.

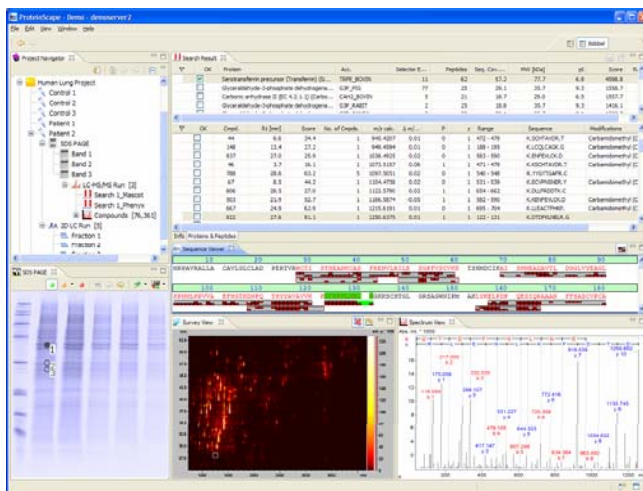


Fig. 2. The new look of ProteinScope 2. Various relevant viewers are integrated in a flexible yet easy to use client program.

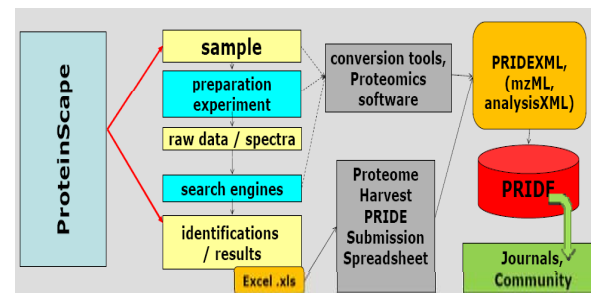


Fig. 3. Integration of Information - PRIDE as an interface between the local laboratory database and the scientific community.

Conclusions

- We present a database-driven **bioinformatics platform**, **ProteinScope**, that supports in particular research laboratories with a **heterogeneous instrumentation**.
- Flexible support of **different proteomics workflows**
- **Visualization tools** for gel-based and LC/MS workflows
- The **ProteinExtractor** integrates over various MS instruments and search engines.
- The **Decoy strategy** validates protein lists.
- **Complete Integration** into the ProDaC Data Submission Pipeline