BRUKER DIRECTOR DESCRIBES DESIGN METHODOLOGY OF ADVANCED PROTEINSCAPE 2 BIOINFORMATICS PLATFORM

Spectroscopy Focus

DATA WAREHOUSING CONCEPT SUPPORTS COMPLEXITY IN PROTEOMICS WORKFLOWS

Bruker Daltonics has just recently disclosed the design methodology of its innovative ProteinScape 2 bioinformatics platform.

The tremendous amount of data from today´s expression proteomics requires a database solution with sophisticated data-warehousing and data-mining capabilities. ProteinScape$^{\text{TM}}$ (co-developed with 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 ProteinScape.

Entire workflows of pre-fractionation, detailed LC/MS/MS (Liquid Chromatography/Mass Spectrometry) separation and post-processing with bioinformatics tools are merged and can be easily controlled and reviewed.

DIFFERENT PROTEOMICS WORKFLOWS FOR IDENTIFICATION AND QUANTIFICATION

"ProteinScape 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 labelling) 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," said Professor Dr. Herbert Thiele, Bruker Daltonics Director of Bioinformatics.

Professor Dr. Herbert Thiele, Bruker Daltonics Director of Bioinformatics.

VISUALISATION

ProteinScape 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.

ACCESS TO RAW DATA

In the course of a full scale proteomics experiment the handling of the data as well as the retrieval of the relevant information from the results is a major challenge due to the massive amount of generated data (gel images, chromatograms, and spectra). as well as associated result information (sequences, literature etc.). The variety of LC/MS mass spec techniques is producing vast volumes of data, posing two major issues in bioinformatics. First: Do we need all the raw data in the database? To cope with the huge amount of data the processing pipeline has to be able to condense the data starting with a real time MS peak picking done at acquisition time. This is followed by a detection of chromatographic compounds which have to be arranged to charge states and finally molecular features which can be used for data base searches and multivariate statistical analysis.

Following automatic data reduction, it is necessary to have software tools to validate the generated results. These validation tools should be able to go back to visualize the raw data and correlate the results on the basis of the raw data. Especially for applying quantification algorithms, the access to MS raw data is mandatory to make sure the information contained in the raw data is not disturbed by processing.

VALIDATION

BioTools integrates with ProteinScape for advanced sequence validation, PTM discovery, de novo sequencing and MS-BLAST searches for full structure elucidation functionalities.

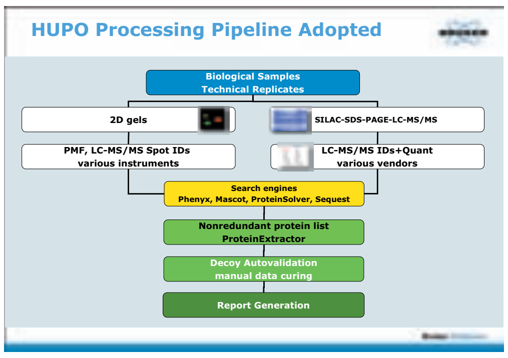
BioTools provides customisable views of sequence annotated raw spectra, interactive peak editing capabilities and error plots that permit interactive operator validation of MS data. MS/MS spectra that were not identified in automated procedures in ProteinScape can be further evaluated by de novo sequencing in conjunction with MS-BLAST or Sequence Queries of Mascot (Matrix Science). The relative scoring of, eg., different phosphorylation site isoforms permits the interactive validation of PTM attachment sites based on MS/MS spectra. BioTools permits the use of custom protein structures (sequence plus a particular set of modified amino acids) for quality control work independent of proteomics approaches, linking ProteinScape's database properties with dedicated work in protein structure analysis.

Thorough integrated access to the WARP-LC software package of quantification workflows that utilise labelling technologies combined with protein separation lead to greatly reduced analysis and validation time.

Author Details:
Professor Dr. Herbert Thiele
Director of Bioinformatics
Bruker Daltonics
Tel: +49 421 2205 187
Email: ht@bdal.de
Web: www.bdal.com

IDENTIFICATION & CHARACTERISATION: STANDARDISED DATA PROCESSING PIPELINE

The processing pipeline for protein identification implemented in ProteinScape has adopted the Human Proteome Organisation (HUPO) Brain Proteome Project (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 ProteinExtractor algorithm which combines peptide lists from different MS/MS search-engines to a combined protein result list. Protein list validation is based on the Decoy Database Concept.


HUPO Processing Pipeline Adopted

Biological Samples
Technical Replicates

2D gels

SILAC-SDS-PAGE-LC-MS/MS

PMF, LC-MS/MS Spot IDs
various instruments

LC-MS/MS IDs+Quant
various vendors

Search engines
Phenyx, Mascot, ProteinSolver, Sequest

Nonredundant protein list
ProteinExtractor

Decoy Autovalidation
manual data curing

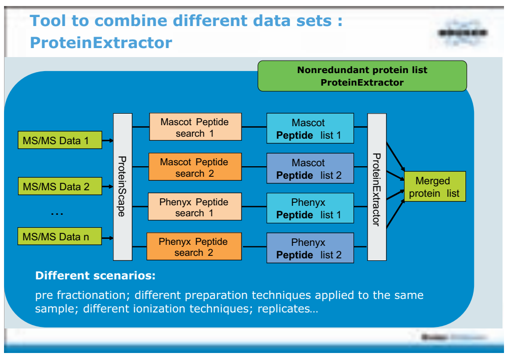
Report Generation

Figure 1. HUPO Processing Pipeline Adopted

PROTEINEXTRACTOR

"How sure can we be to have identified the right proteins with our mass spectrometric instrumentation? Can we expect valid data from the employed search algorithms?" research scientists ask.

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 has proved to be successful. ProteinExtractor uses only spectra, the assigned peptides and peptide scores as input. This gives scientists the opportunity to create protein lists with the same algorithm and conditions regardless of which search engine was used. ProteinExtractor can be used to combine the peptide search results of several search engines. With this the sensitivity and selectivity of each search engine can be combined. For a specific protein, some peptides are found e.g. only by Mascot, some others only by Phenyx. Thus, the number of identified proteins (at a given FPR) is higher when results of several search engines are combined. ProteinExtractor has successfully been used and tested with Mascot, Phenyx, Sequest and ProteinSolver.


Tool to combine different data sets :
ProteinExtractor

Nonredundant protein list
ProteinExtractor

MS/MS Data 1
MS/MS Data 2
...
MS/MS Data n

ProteinScape

Mascot Peptide search 1
Mascot Peptide search 2
Phenyx Peptide search 1
Phenyx Peptide search 2

Mascot **Peptide** list 1
Mascot **Peptide** list 2
Phenyx **Peptide** list 1
Phenyx **Peptide** list 2

ProteinExtractor

Merged protein list

Different scenarios:

pre fractionation; different preparation techniques applied to the same sample; different ionization techniques; replicates...

Figure 2. Tool to combine different data sets: ProteinExtractor

AUTOMATED DECOY ANALYSIS IN PROTEOMICS PROJECTS

The analysis of complex proteomes notoriously leads to large protein lists. However, the current scoring algorithms of the search engines do not allow the reliable discrimination between correct and false hits. Very often the scientist has to assess large numbers of potentially identified proteins – manually.

The Decoy Database Concept is a strategy for an automatic assessment of protein lists, allowing the measurement of false positive rates (FPR) in a very straightforward fashion.

The principle of artificial Decoy protein sequences is a widely accepted answer to the task of automated result validation. Nevertheless, the actual realisation of the Decoy approach varies among research groups and search engines. Here, the ideal solution applies a composite database of real and Decoy protein entries.

To assure a FPR for a particular experiment, e.g. 5%, the implemented algorithms allow scientists to order the identified proteins according to their identification score value and take only those proteins as correctly identified where the accumulated FPR does not exceed the pre-specified threshold (5% mark).

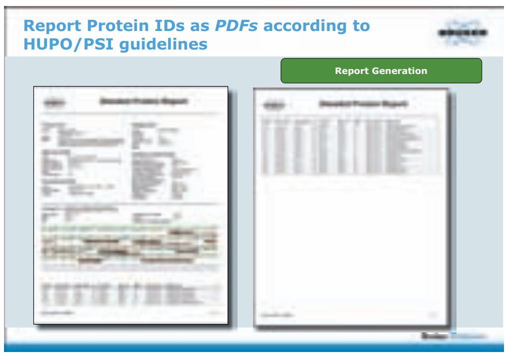
AUTOMATED RESULT VALIDATION

Result integration and validation are key issues for the identification and quantification of proteins in great numbers. For a maximized number of identified proteins, one strategy comprises biological and technical replicates, another involves separation steps on protein and/or peptide level. The resulting redundant search results need to be integrated on a peptide level.

The ProteinExtractor compiles a non-redundant protein list from peptide lists of different origin. This allows the combination of data from different search engines as well as from different MS experiments (2D LC-ESI-MS/MS and LC-MALDI-TOF/TOF).

The use of decoy strategies to validate the number of identified proteins according to a desired False Positive rate. as well as application of the ProteinExtractor to overcome the protein inference problem minimises the need for manual data validation.

Days of manual processing time are condensed into hours of computing time. In parallel the use of a single data repository allows for easy access to the combined information from different workflows and links to external tools complement the system for project-spanning comparisons of data sets


Report Protein IDs as *PDFs* according to HUPO/PSI guidelines

Report Generation

Figure 3. Report Protein IDs as PDFs according to HUPO/PSI guidelines

QUERIES

Comprehensive proteomic analyses deal with large data sets comprising a range from very few up to thousands of MS and MS/MS spectra. Considering multiple measurements deriving from method optimisation, quality control and long-term studies, this number even multiplies and necessitates a structured overview and summary about the different results. The optimisation of a complex separation workflow is often a multidimensional task. Many parameters might have to be varied. In the end, a huge number of datasets is generated and must be compared in various ways. Even a rather simple experiment shows that a database-driven software platform makes life much easier. It keeps track of all data and allows the setup of simple, relevant queries. In ProteinScape the concept of comparative queries featuring proteomics-specific queries for mass spectrometric data is implemented. The queries allow investigating specific aspects with a focus on different sample preparations, certain peptides, or protein specific attributes including biological properties. The concept of comparative queries allows for quick and simple extraction of tailored and concise information, it gives an excellent overview about large data amounts and is an ideal tool for method optimisation.

PROVIDING A TRANSPARENT ACCESS TO PROTEOMICS DATABASES FOR RETRIEVING BIOLOGICAL INFORMATION

One of the main goals in advanced bioinformatics is to extract and collect all biological information available in public databases from a set of identified molecules (genes, proteins, etc.). Due to the complexity of this task and the huge amount of data available, it is not feasible to gather this information by hand, making it necessary to have automatic methods. PIKE (Protein Information and Knowledge Extractor) solves this problem by automatically retrieving via Internet all functional information on public information systems and databases, and then clustering this information according to the pre-selected criteria.

PIKE offers an easy and user friendly way to obtain protein functional information extracted from several internet sources. The user can improve the way to obtain knowledge about the biological role of the proteins within the specific topic of the experiment. The system also provides methods to integrate PIKE data into ProteinScape to extend the level of information provided. (http:proteo.cnb.uam.es) Other sources of protein meta-information and further knowledge are the Protein Center (ProXeon), the NCBI, IPI and SwissProt pages that are accessible directly from individual proteins or whole result tables in ProteinScape.

> ***How sure can we be to have identified the right proteins with our mass spectrometric instrumentation?***

New High Temperature Gpcir, Dedicated To Polyolefin Analysis

Polymer ChAR has developed a new high temperature GPC dedicated to polyolefin analysis, the GPCIR. It can be configured as a triple detector system, including viscometer, Light Scattering and IR detector, with both concentration and composition sensors for SCB measurement. IR detector is the new IR4+ and optionally the IR5-MCT, for highly demanding applications. GPCIR can also be configured as a GPC+TREF system.

GPCIR is a robust, compact and easy to use instrument; sample preparation is fully automated using an autosampler with two temperature zones to prevent sample degradation. Unattended sample dissolution and injection of 70 samples is possible in 10ml vials for sample representativeness, without requiring vials transfer. In addition, the pump system is highly stable and includes solvent preheating. Columns are located in a separated oven dedicated for them only, and the valves, detectors and the sample filtration system, which incorporates backflushing are located in another oven independent from the columns one. GPCIR Virtual Instrumentation Software provides instrument control, monitoring of the whole process and calculations, which integrates all detector signals in a single package.

Circle no. 37