Bioinformatics Strategies in Life Sciences: From Data Processing and Data Warehousing to Biological Knowledge Extraction

Herbert Thiele, Jörg Glandorf, Peter Hufnagel

Bruker Daltonik GmbH
Fahrenheitstrasse 4
28359 Bremen
ht@bdal.de, jg@bdal.de, ph@bdal.de

The extreme complexity of the Proteome calls for different multistep approaches for separation and analysis on protein and on peptide level. These are usually combinations of 1D or 2D gel electrophoresis and one- to multidimensional LC techniques in combination with different MS and MS/MS techniques. A database driven solution is the most effective way to manage these data, to compare experiments, and to extract and gain knowledge based on experiments already done in the past. The bioinformatics platform ProteinScape^{TM} (Bruker Daltonics) supports these various discovery workflows in Proteomics through a flexible *analyte hierarchy concept.*

To generalize the reprocessing of diverse data sets, a guideline (http://forum.hbpp.org) has been set up defining the workflow of protein identification. A data warehousing system including a data processing pipeline is mandatory for data comparison and validation. In a conceptional view, the general data flow in proteomics consists of three basic elements: (i) generating raw data on different types of MS and MS/MS instrumentation; (ii) the local database solution that handles the set of heterogeneous data supplying different vendors instruments, different types of MS based techniques and all possible workflows for protein identification and quantification with the support of sophisticated algorithms for standardized generation of validated results; and (iii) standard submission tools to submit the results to the global data repository PRIDE (PRoteomics IDEntifications database at the European Bioinformatics Institute (Hinxton/UK) (http://www.ebi.ac.uk/pride).

In most cases results from proteome-wide experiments result in a complex array of information represented as a set of identified, characterized or differentially expressed proteins. Whilst important, this work represents only a first step towards the goals in proteomics experiments, where one ultimately aims to obtain knowledge about the biological role of the proteins within the specific topic of the experiment.. This is currently not routinely performed in the proteomics community and a pressing need exists to develop sophisticated software that allows researchers to control, filter and access specific information from genomics and proteomics databases.