Proteomics Potential

Dr Herbert Thiele of Bruker and Martin Blüggel of Protogen AG argue that an advanced bioinformatics platform enhances pharmaceutical protein analytics

Proteomics activity in the biotech and pharmaceutical industries has shifted in recent years from technology-focused and centralised discovery programmes, to enabling facilities integrated into discovery and production departments, using highly specialised state-of-the-art proteomics techniques. These are utilised for biomarker and target discovery and the development of protein therapeutics. Applications for proteomics techniques include antigen discovery, protein-protein interaction studies, differential protein display between treated and non treated proteomes from preclinical studies (for example cell culture model), but also in upstream and downstream process optimisation. Also, state-of-the-art protein analytics developed for proteomics applications are applied for full characterisation of protein therapeutics in its structural characterisation, physiochemical properties and its process- and product-related impurities analysis following ICH Q6B.

State-of-the-art techniques for the analysis of complex proteomes and protein mixtures use multidimensional approaches: separation at a protein level (1D/2D PAGE); different separation principles at a peptide level (IEF, 1D LC and 2D LC workflows); and a combination of different mass spectrometry-based techniques (electrospray ionisation and MALDI-laser based ionisation, ion-trap or time-of-flight analysers to generate MS and MS-MS data). With the tremendous amount of heterogeneous data resulting from today's protein analysis due to these different experimental strategies, different MS-based techniques and different instrumental equipment, including sophisticated database driven warehousing and data mining strategies, are mandatory for the biopharmaceutical scientist.

New technology has enabled pharma researchers to manage proteomics and protein analytics data from the generation and data warehousing to a central data repository, with a focus on advanced precision on protein analytics by use of multiple search engines, decoy strategy, advanced algorithms for peptide to protein identification, various quantitation workflows and standardisation efforts.

ADVANCED PRECISION OF PROTEOMICS RESULTS

Multiple Search Engines

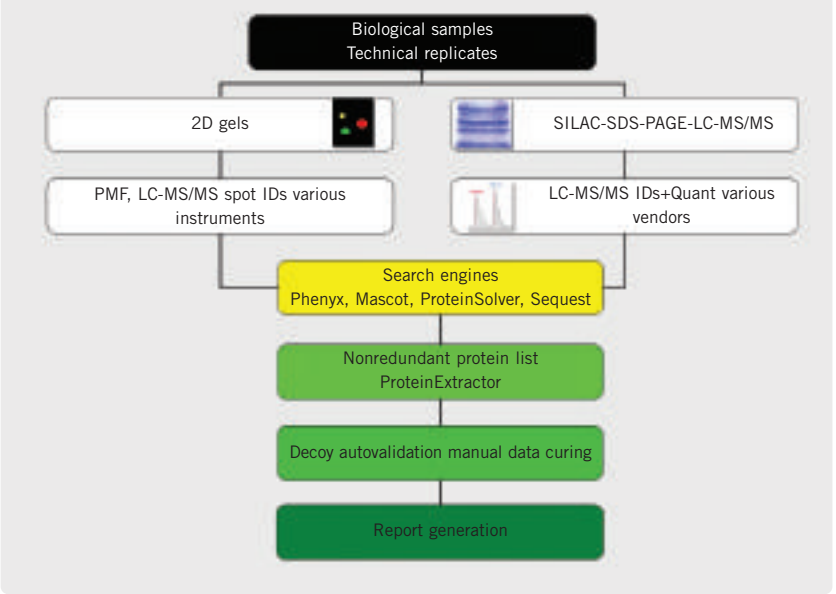
The key issue in MS-based protein identification is that peptide masses determined by MS are generally not unique and therefore each measured mass can randomly match a peptide from a sequence database. As a result, protein identification is probability-based and there remains a certain risk of obtaining a false positive. To measure the statistical significance of a match, the MS/MS search engines apply various different approaches to calculate scores.

Because of the different approaches to scoring, different search engines frequently report different proteins and apply a different ranking. An important criterion for judging the performance of search engines is their ability to discriminate correctly identified proteins from randomly matched proteins. This approach enables the sensitivity and selectivity of the algorithms to be assessed. To get the most accurate and reliable information, new bioinformatics platforms integrate several MS/MS search engines in order to allow cross-validation and consolidation of the identification results through the complementary use of these engines (see Figure 1).

The result is the most correct protein hits found by all or at least several protein identification software programs, and the most random hits only by a single software program. This allows for a better and bigger merged protein list than any of the single-software protein lists. All peptides of the proteins identified by at least one of the protein identification software program are

Figure 1: Workflow for processing MS (PMF) and MS/MS spectra based on a processing guideline implemented within ProteinScape

The approach for analysing the MS/MS spectra is slightly different from the approach of the MS spectra. All spectra are searched against four search engines and the resulting peptides are used with the ProteinScape algorithm ProteinExtractor to end up with protein lists. These lists for each search engine are used again to merge the protein lists based on the peptides of the identified proteins. The four protein lists and the merged protein list are cut off by a False Positive Rate [FPR] (for example one per cent) on the protein level. The processing can be fully automated.



used, and protein scores are calculated as a (weighted) sum of the scores of all peptides matching to a protein from all algorithms.

Additionally, integrating interfaces to PTM discovery algorithms allows the further characterisation and validation of proteins, and the correlation of functions.

Decoy Strategy

The use of the 'decoy' approach allows for the measurement of the rate of wrong identifications (false positives) by means of artificial (obviously false) protein sequences mixed into the protein database. For every original protein sequence, a decoy entry is generated that contains the same amino acids in a random order – a protein with the same mass and amino acid content, but with an artificial sequence. The false positive rate (FPR) of protein searches can be estimated by searching decoy databases containing entries with 'right' (target) and 'false' (decoy) protein sequences. At any score threshold, the number of decoy proteins can be counted which indicates the rate of false positives.

A protein search against the 'decoy' database will result in a protein list that contains a certain number of decoy entries, clearly detectable by their accession number with the prefix. Following the assumption that every match to a 'decoy' entry is a wrong match (false positive) and that the number of random identifications in the 'original' part of the sequence database will be similar (or less) to the number of decoy entries found, the number of decoy matches allows a good estimation of the number of incorrect identifications and the calculation of the false positive rate.

From Peptide ID to Protein ID

In MS/MS experiments only peptides are identified, not proteins. A search engine identifies a list of different peptides for each single MS/MS spectrum. The mapping of peptides to proteins is not a one-to-one mapping, leading to ambiguities. Spectra match to several peptide candidates, and each peptide in turn matches to several proteins or protein isoforms. Generating a non-redundant list of proteins (containing only those proteins and protein isoforms which can be distinguished directly by MS/MS data) from peptides is therefore very difficult and has to be conducted through bioinformatics methods. Current approaches show very little transparency, disregard isoform distribution, utilise rough estimates or need sophisticated training.

Some bioinformatics platforms use an empiric method to derive protein identification lists from peptide search results. Prior to designing such a programme, a team of experts must define a set of rules in order to define a minimal protein list, which contains only those proteins (and protein variants), which can be distinguished directly by MS/MS data. Every protein reported should be identified by at least one (or more) spectrum with significant peptide score, which cannot be mapped to a higher-ranking protein already in the result list. The algorithm was then implemented to follow these rules.

Additionally, such platforms enable an automatic data processing and result validation by generating protein identification extracted from redundant information and multiple search engines, merging peptide lists from different workflows or studies into one compiled protein list and automatically cutting off identification results for which a pre-selected false positive rate has been reached. The combination of datasets from repeated experiments or complementary workflows – for example from different MS experiments (2D LC-ESI-MS/MS and LC-MALDI-TOF/TOF) – increases the sensitivity of the analysis as well as validating results. The use of decoy strategies as well as application of such platforms to overcome the protein inference problem minimises the need for manual validation, which is nevertheless easily possible by accessing the linked raw spectra information.

In the near feature it will be common practice for all protein identifications to come with a statistical significance value, or a specified false positive rate (FPR), so the validity and statistical relevance of the information is well described.

STANDARDISED SAMPLE AND DATA PROCESSING

Different workflows create different kinds of information, so reproducibility and standardised ways to create confidence in the generated results are the great challenge for a relevant bioinformatics solution. An important aim is to define standardised analysis methods and validation techniques in SOPs. To standardise the processing of data sets, for example those acquired during protein production batches, a guideline like that suggested by the HUPO Brain Proteome project (1) has been used as a template. Well defined data processing procedures and standardised operations (processing pipeline) significantly help to increase comparability and to improve the protein identification results. It will allow comparison of results and statistical relevance for relevant data, within the huge variety of proteomics data.

MASS SPECTROMETRY-BASED QUANTIFICATION

Mass spectrometry-based quantification is becoming more and more powerful. Quantification based on label chemistry is divided into two classes:

- Non-isobaric labels (stable isotope label experiments – SILE – such as ICPL; stable isotope label analysis in cell culture – SILAC – such as Leu/Arg, 16O/18O labelling). Here, the proper signal pairs must be found in the spectra; intensity ratios are calculated on MS level.
- Isobaric labels (such as iTRAQ). Here, all labels of a pair or n-tuple have the same mass but generate different reporter fragments in the MS/MS spectra. Pair finding is much easier because the masses of the reporter fragments are known.

Labels that modify specific amino acid residues (such as ICPL that labels Lysines) (see Figure 2, page 72) are compatible with protein separation steps since they are introduced before the protein digest. Labels that specifically modify the N-termini of the peptides (as seen in the standard iTRAQ setup) must be introduced after the protein digest and thus rely on elaborate peptide separation techniques (2D LC, IEF+LC). All current label chemistries for protein quantification are complementary and have advantages in selected applications.

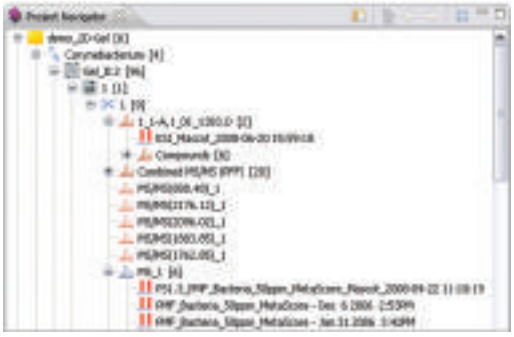
Recent improvements in MS instrumentation and nano-LC reproducibility make a label-free MS-based quantification approach

Figure 2: Display of the results of an LC-MALDI ICPL quantification of an artificial protein mixture in ProteinScape

The Protein Browser shows the basic numbers (for example ratio Median over all peptides) for each protein. For a selected protein, the sequence coverage map the respective Compounds in the LC-MS Survey and the quantification Box-Whisker Plot is shown. For a selected peptide, the MS and MS/MS spectrum with fragment ion annotation is available as well.

Figure 3: Example of a 2D gel project in the navigation tree

The project navigator of ProteinScape shows one project that contains one sample that has been submitted to 2D gel electrophoresis. 102 spots have been digested and analysed by MALDI TOF MS and LC ESI IT MS/MS. One spot is expanded, so the datasets and the underlying protein database search results can be seen.



feasible. This technology has the potential to become a significant complement to current quantification methods, such as label-based MS methods (ICAT, for example) or 2D-gel quantification methods. The high throughput compatibility of a label-free approach allows the processing of large numbers of samples, which is required in order to obtain statistically valid quantifications from typical biological sample heterogeneity. Handling these workflows from data processing (such as RT-alignment of different LC-MS data, compound detection and binning techniques) to statistical validation and quantification results is a major challenge of today. Using such platforms as those described above, with their advanced analysis tools for protein identification, quantification workflows that utilise labelling technologies combined with protein separation require greatly reduced analysis and validation time.

SUPPORTING COMPLEX WORKFLOWS

Data Warehousing Concept

A database-driven solution is the most effective way to manage heterogeneous data, to compare experiments, and to extract and gain knowledge based on experiments already done in the past. Bioinformatics platforms support various discovery workflows through a flexible analyte hierarchy concept.

In the lab, a proteomic workflow can be a manifold combination of various steps, including separation on protein level, protein labelling, protein digestion to generate peptides, separation at peptide level followed by analytical methods to identify peptides, proteins and their post translational modifications for example with mass spectrometry. In biologics platforms, a simple or complex workflow can be represented in the navigation tree (see Figure 3). However, two elements remain fixed: the topmost level is the project, which contains samples on the second level. Optionally, separation and digestion are created at the next level with mass spectrometric data below, or MS data can directly be located below a sample. An example of a standard workflow (2D gel workflow) is shown in Figure 3.

Visualisation

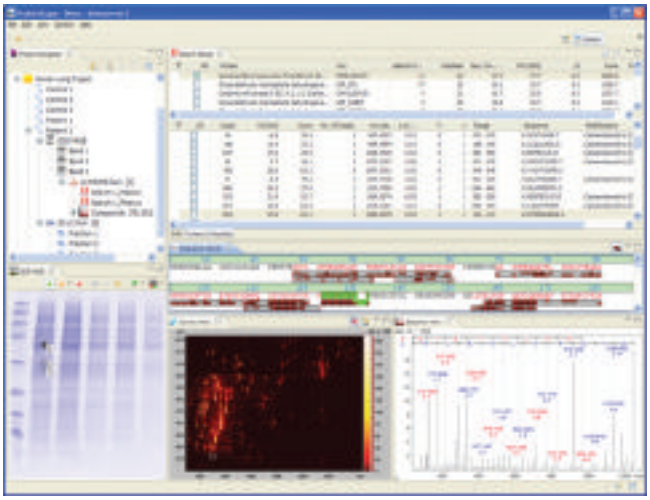
Visualisation of heterogeneous data, complex data processing and its results is the challenging task for the graphical user interface and is key for the acceptance of bioinformatics software. Whilst it provides an overview on all data, it has to be able to validate raw data at the same time. An intuitive user interaction and responsiveness is constantly under optimisation by systematic usability tests.

Most biologics platforms will have a number of dedicated data viewers (see Figures 2 and 4) that permit the evaluation and validation on each level of proteomics experiments, such as the navigation tree, a 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.

The new design concept allows visualisation of process proteomics data under different perspectives, which is a defined combination of view layouts. All the different views on the data are arranged within one general programme frame. Users running different analysis workflows or having preferences in the arrangement of views can easily define and store a suitable customised view layout as a perspective. This approach allows straightforward creation of any depth of information without switching between different software programmes.

Figure 4: Novel viewer concept

This example shows a study of three patient and three control samples that have been subjected to 1D SDS PAGE. The ProteinScape client program shows the Project Navigator, the gel image of one selected sample, and the list of identified proteins. For a selected protein it shows the list of identified peptides, the sequence coverage map (red bricks indicating matched fragment ions), and for a selected peptide its location in the LC-MS run and the respective MS/MS spectrum.



Different perspectives into the archived data include sample and project management, generating protein ID, protein quantitation, setting up data dependent queries including combining or grouping of samples, different data viewing tools (such as a spectrum viewer, protein viewer, LC- MS data viewer) including raw data access and various report options (such as protein ID, peptide ID, quantitation reports).

Reports

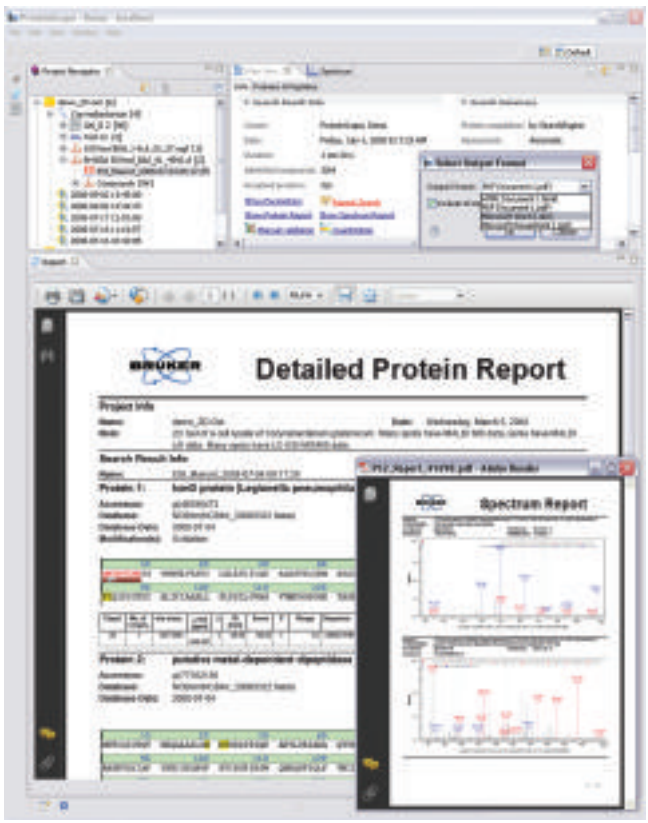
The results of analysis are documented in standardised reports enabling communication of results and their conclusion in a comprehensive way across departments and colleges with different backgrounds in their knowledge of proteomics technology.

The inbuild report generator allows maximum flexibility and standardisation of reports at the same time (see Figure 5). A set of typical reports serve as a template and are designed to fulfill the publication guidelines that were suggested by journals such as the *MCP* and *Proteomics Journal*.

Users wanting to submit their data to central repositories or publish their results are facing the problem of collecting all relevant information, methods, parameters, MS data, and search results. With the report template, everything is already in place because the methods, data and results are stored in a project-oriented manner. The search result, in particular, can be easily exported to MS Excel, or a well-structured PDF file can be generated. In addition, a dedicated spectrum report generates a more detailed view into the data.

Figure 5: An inbuilt report generator allows maximum flexibility and standardisation

Detailed protein reports can be generated in various formats (such as html, pdf, doc). A spectrum report can be generated to provide a more detailed view into the data.



About the authors

Dr Herbert Thiele was educated at the Institute of Organic Chemistry, TH Aachen, where he gained his diploma and wrote his doctoral thesis. He started out in 1977 as a teacher of Chemistry and Mathematics at the Fachschule des Heeres für Technik/ Aachen. He joined Bruker in 1978 and has held various management positions, as well as being made an Honorary Professor in Analytical Chemistry at the University of Bremen. In 2002 he was appointed Director of Bioinformatics.
Email: ht@bdal.de

Martin Blüggel earned a diploma in Chemistry at the University of Konstanz, specialising in proteins. After working as a researcher, he founded, along with Prof H E Meyer, Protagen GbR, a company devoted to peptide and protein analysis. In 1999, this venture became Protagen AG, and Martin was appointed Chief Operating Officer in 2002. He is responsible for management of customer projects, and coordinating Protagen´s resources across the various operating entities.
Email: martin.blueggel@protagen.de

CONCLUSION

With the large variety of workflows, as well as the multiplicity of instruments and data-analysis software available, researchers today face major challenges in validating and comparing their data. Using standardised data formats, but also with HUPO PSI, efforts in standardised processing and validation the generated data may be more accurate, reproducible and comparable.

Developed bioinformatics platforms for proteomics and protein analysis are addressing the current requirements of the biotech and pharmaceutical industries for identification, quantification, validation of biomarkers and detailed protein characterisation. Offering comprehensive solutions for qualitative and quantitative LC-MS/MS and gel-based protein analysis, this data warehousing and project management software supports all workflows through a flexible analyte hierarchy. A combination of different database search engines, scoring algorithms and quantification methods is combined with 'decoy' validation by a platform that produces non redundant protein result lists across entire projects. Additionally, capabilities in data visualisation, reporting, data integrity and data security match the current high standards of the pharmaceutical industry.

Note

ProteinScape is a trade mark of Bruker Daltonics GmbH

Reference

1. http://forum.hbpp.org