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

Herbert Thiele, Jörg Glandorf, Peter Hufnagel

Bruker Daltonik GmbH, Fahrenheitstr. 4, 28359 Bremen, Germany

{ht,jg,ph}@bdal.de

Summary

With the large variety of Proteomics workflows, as well as the large variety of instruments and data-analysis software available, researchers today face major challenges validating and comparing their Proteomics data. Here we present a new generation of the ProteinScape™ bioinformatics platform, now enabling researchers to manage Proteomics data from the generation and data warehousing to a central data repository with a strong focus on the improved accuracy, reproducibility and comparability demanded by many researchers in the field. It addresses scientists` current needs in proteomics identification, quantification and validation. But producing large protein lists is not the end point in Proteomics, where one ultimately aims to answer specific questions about the biological condition or disease model of the analyzed sample. In this context, a new tool has been developed at the Spanish Centro Nacional de Biotecnologia Proteomics Facility termed PIKE (Protein information and Knowledge Extractor) that allows researchers to control, filter and access specific information from genomics and proteomic databases, to understand the role and relationships of the proteins identified in the experiments. Additionally, an EU funded project, ProDac, has coordinated systematic data collection in public standards-compliant repositories like PRIDE. This will cover all aspects from generating MS data in the laboratory, assembling the whole annotation information and storing it together with identifications in a standardised format.

1 Data Warehousing Concept: Supporting Complexity in Proteomics Workflows

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 (PAGE) and one- to multidimensional liquid chromatography (LC) techniques in combination with different mass spectrometry (MS) and tandem mass spectrometry (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. Nowadays, recent improvements in MS instrumentation and nano-LC reproducibility make a label-free MS based quantification approach feasible. The high throughput compatibility of a label-free approach allows processing large numbers of samples, which is required to obtain statistically valid quantifications from typical biological sample heterogeneity. Handling these workflows from data processing to statistical validation and quantification results is a big challenge. Any kind of software solution for data warehousing and analysis should address these different workflows in a flexible manner. The bioinformatics platform ProteinScape™ (Bruker Daltonics) supports these various discovery workflows in Proteomics through a flexible *analyte hierarchy concept* (Fig.1).

doi:10.2390/biecoll-jib-2010-141

Fig. 1: Quantitation Project in ProteinScape. All Views can be easily switched ON and OFF and changed in size to see all relevant information simultaneously:

1. **Project Navigator showing the gel and the 21 bands;**
2. **List of identified proteins, the peptides of one selected protein;**
3. **Sequence coverage map of the selected protein;**
4. **Distribution of the accepted peptides in the LC-MS/MS run;**
5. **Spectrum with fragment annotations for one selected peptide;**
6. **Box-Whisker plot showing the quantitative data for the selected protein. Does the Median equal the Average? Is the Quartile distribution symmetric?**
7. **Image of the 1D gel.**

2 Standardized Data Processing

To generalize the reprocessing of diverse data sets, a guideline (http://www.hbpp.org) has been set up defining the workflow of protein identification. Well defined data processing procedures and standardized operations (*processing pipeline*) significantly help to increase comparability and to improve the protein identification results. A data warehousing system including a data processing pipeline is mandatory for data comparison and validation. [1-3].

2.1 Protein List Compilation: From PeptideID to ProteinID

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. "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 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 has been developed for *ProteinExtractor*, a module of the ProteinScape bioinformatics platform. 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 An iterative approach has proved to be successful. ProteinExtractor uses only spectra, the assigned peptides and peptide scores as input. This strategy allows 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 allowing to combine the sensitivity and selectivity of each search engine.For a specific protein, some peptides are found e.g. only by Mascot, some other only by Phenyx. Thus, the number of identified proteins (at a given false discovery rate (FDR)) is higher when results of several search engines are combined [4-6].

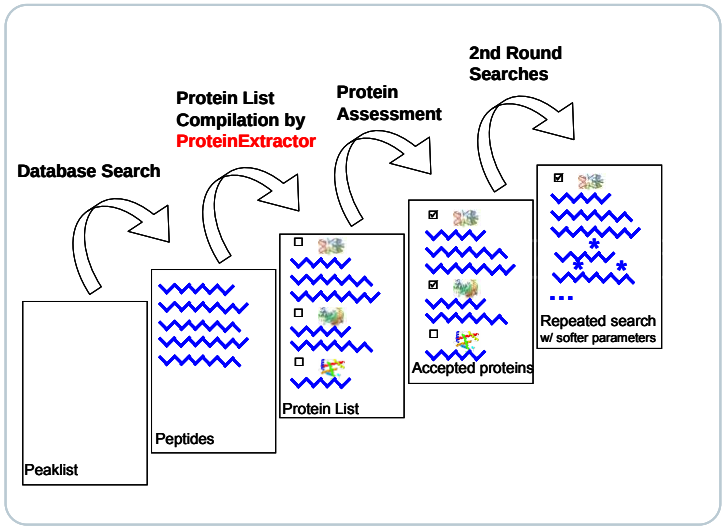

Fig. 2: Strategy for Protein Identification:

1. **Search engines report identified peptides.**
2. **The ProteinExtractor combines the peptides into a non-redundant protein list.**
3. **The proteins are assessed following a Decoy strategy for a minimized False Positive Rate.**
4. **Selected proteins can be submitted to a 2nd Round search for the detection of PTMs etc.**

2.2 Protein Identification Enhancement: Integrating Multiple Search Engines

The key problem 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 hit. To measure the statistical significance of a match, search engines (e.g. Sequest [7], Mascot [8] (http://www.matrixscience.com/), Phenyx [9] (http://www.phenyx-ms.com/), ProteinSolver [10]) for retrieval of peptide sequences from a sequence database apply various different approaches to calculate search scores. To get the most accurate and reliable information ProteinScape integrates several MS/MS search engines (including Mascot, Phenyx, Sequest, ProteinSolver) to allow cross-validation and consolidation of the identification results through the complementary use of these engines (Fig. 2).

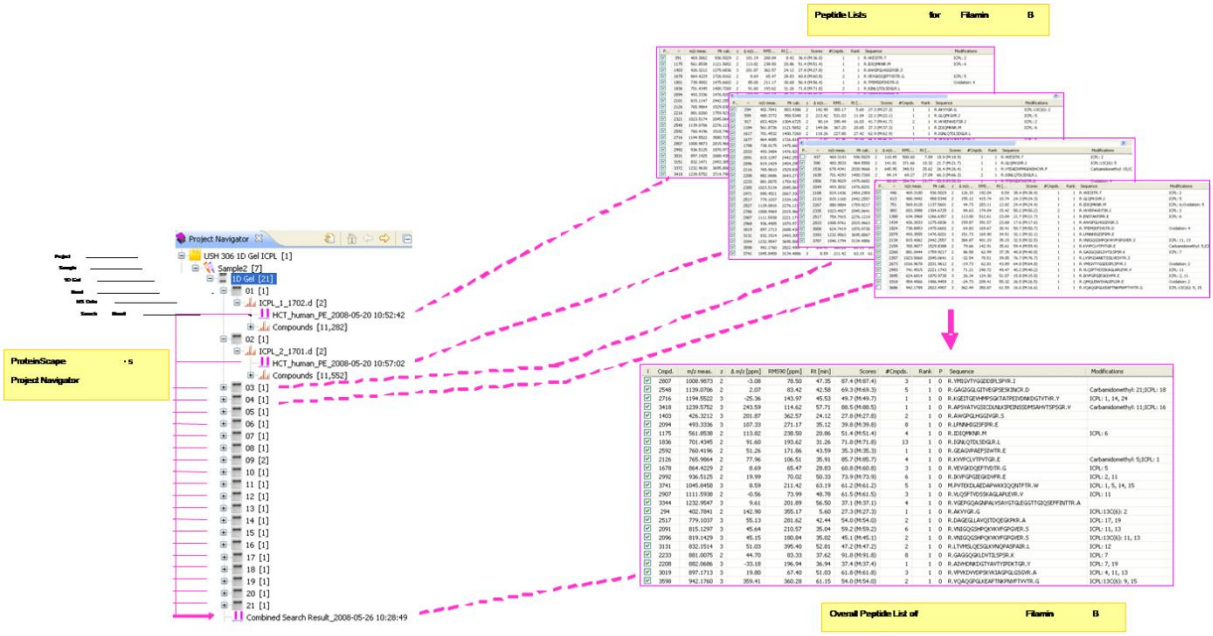

Fig. 3: Data handling in ProteinScape. The Project Navigator of ProteinScape (left) displays the 1D gel workflow incl. LC-MS/MS data and search results which have been compiled by the Protein Extractor(1) into one respective protein list for each band and (2) into one integrated protein list for the whole gel. On the right, the peptide lists for Filamin B for bands 1 to 4 and the overall peptide list are shown.

As an example of the performance of the ProteinExtractor algorithm human lung carcinoma cell lysates (line A549; TGF β treated) were analyzed in a first separation step by a 1D SDS PAGE resulting in 21 bands followed by a tryptic digest of each band. The peptides were separated on a nano-LC system and supplied to an ion trap mass spectrometer (HCTultra, Bruker). ProteinScape performed database searches using a composite Decoy IPI human database on a Mascot search engine on the data of each gel band separately. The 21 peptide lists were combined by the ProteinExtractor algorithm, which generated one integrated protein list. 894 proteins and 5472 peptides resulted from the combination of the 21 gel band results. Details for one protein are shown in Fig. 3, right. As an example the protein Filamin B has been identified from four bands, each with 15 to 25 peptides. ProteinExtractor has created a common non-redundant peptide list for Filamin B with 38 peptides. 35 were accepted in the automatic assessment step. The fact that Filamin B peptides are found in various bands shows that the protein separation power of the gel was not sufficient.

2.3 Protein Assessment: Decoy Strategy for Validation of the Protein List

Combined MS/MS searches result in long protein lists, sorted by descending protein score. However, the question remains open as to which proteins are “really” identified. Where is the threshold that separates correct hits from random matches [11-12]?

Valid protein identification is reasonable only if some quality assessment of the resulting protein list is performed [13-14]. The quality of protein identification can be measured automatically by determining the false positive rate (FPR) or false discovery rate (FDR). 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. The following calculation is based on the assumptions: (i) every match to a “decoy” entry is a wrong match (false positive); (ii) 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. Thus, the

number of decoy matches gives a good estimation of the number of incorrect identifications [15].

The parameters of the Protein Extractor are typically set in a **stringent** way. This means, that the result list contains only proteins with at least two reliably identified peptides by just one search engine (Mascot score >35; Fig. 4, right). For demonstrating the Decoy strategy, we also applied a more **relaxed** setup with a softer score threshold (only one peptide with Mascot score >35; Fig. 4 left). This resulted in a longer protein list, however it contained more Decoy proteins. The complete protein list in the relaxed setup consists of 1623 proteins. Applying the FDR = 5% criteria, the list reduces down to 542 statistically validated proteins. In the case of the stringent setup the complete list consists of 894 proteins, which contains clearly less than 5% FDR identified proteins.

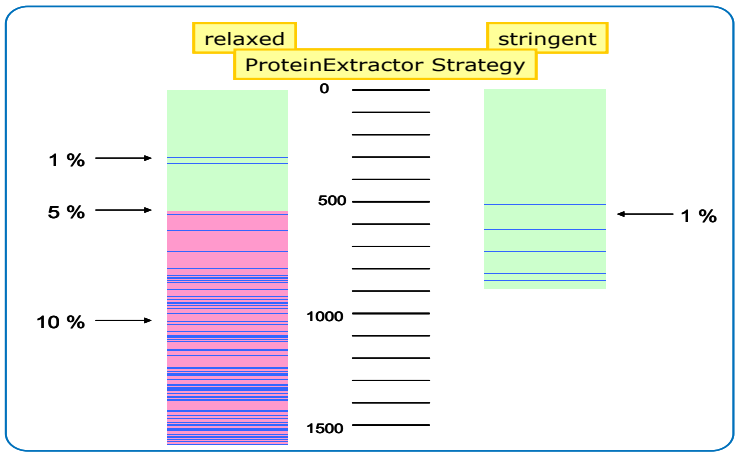

Fig. 4: ProteinExtractor setups. The more relaxed score criteria lead to a longer protein list with more False Positives. The blue lines mark the artificial protein entries in the composite database. The green and the red area mark the ranges with a False Discovery Rate below and above 5 %.

2.4 Merging Peptide Lists from Different Search Engines

Most correct protein hits are found by all or at least several search engines, most random hits only by a single search engine. To determine the true protein content of a sample, independent protein lists based on different search engines have to be merged into a single protein list. ProteinExtractor merges first the peptide lists from all search engines, and then builds a new protein list. A new FDR calculation is done on this merged protein list, independent from the FDR of the individual engine protein lists.[11].

To study the effect of using more than one search engine on automatic result validation peptide search was performed for Mascot and Phenyx on each gel band separately resulting in 42 peptide lists. The conditions for protein list compilation was set to at least one peptide with score > 35 for Mascot and > 4 for Phenyx and for result assessment the false positive rate of 5% was set for automatic protein acceptance. ProteinExtractor compiled the Mascot and Phenyx search results for all 21 gel bands resulting in one comprehensive list with 1108 proteins and 6183 peptides. The use of two search engines leads to:

- an increased sequence coverage since the peptides identified by Mascot and Phenyx do not completely overlap, and
- more identified proteins, since the protein lists of the two search engines do not completely overlap either.

However, for proteins identified by only one search engine, a closer manual inspection is highly recommended.

3 Information Integration & Export to Public Data Repository

Producing large protein lists is not the end point in Proteomics research. To be of sustainable value, the results of an experiment should be stored in a utilizable manner. To enable result assessment and experiment comparison the experimental conditions have to be documented in a concise, reproducible and also machine readable way.

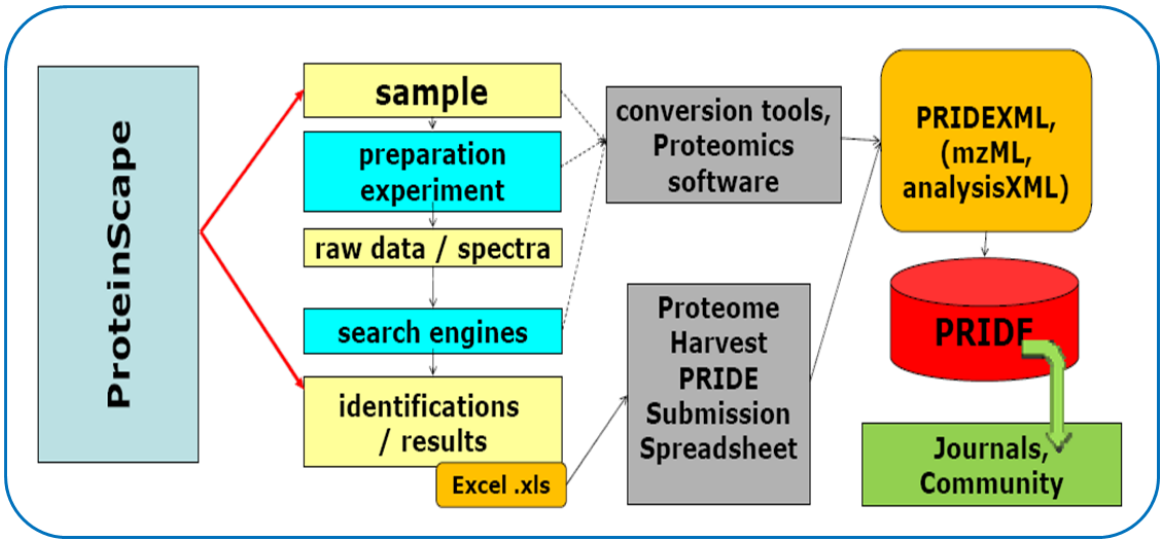

Fig. 5: Integration of Information - PRIDE as an interface between the local laboratory database and the scientific community.

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) [16].

3.1 Standardized Data Formats

Rapid developments in mass spectrometry instrumentation have enabled the acquisition of large data volumes in short times. Unfortunately, each vendor uses a different, proprietary format to present the data, leading to restrictions on both data (re-)use and data storage. To solve this problem, the Proteomics Standards Initiative (PSI, http://www.psidev.info) as a part of the Human Proteome Organisation (HUPO, http://www.hupo.og) was founded in 2002 [17]. The HUPO-PSI has since released the overall "Minimum Information About a Proteomics experiment (MIAPE)" guidelines along with several other community accepted standard formats (e.g. *mzML*, *AnalysisXML*) and reporting requirements [18]. *mzML* (http://psidev.info/) is a new data format for the storage and exchange of mass spectrometer output files. It follows on the successful mzXML and mzData formats but is not expected to completely replace vendor binary formats. mzML has been designed by merging the best aspects of both previous formats into a single unified format. A large number of different proteomics search engines are available that produce output in a variety of different formats. It is intended that *AnalysisXML* (http://psidev.info/) will provide a common format for the export of analysis results from any search engine (Fig. 5).

3.2 Standardized Data Submission Pipelines

Previous experience has shown that tool providers sometimes claim compatibility to a certain standard, but actually only implement it in a very rudimentary way. 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 comparision of results and statistical relevance for relevant data, within the huge variety of proteomics data. The development of software to export, submit and import data sets using standard formats will significantly support, the development of custom data processing pipelines, and facilitate the submission of PSI conform data sets to public databases. However, the implementation of PSI standards in well-known proteomics tools is still a challenge.

As a result, the Proteomics Data Collection consortium (ProDaC, http://www.fp6-prodac.eu/) was initiated in 2006 with the aim to provide the additional infrastructure required for efficient data sharing. ProDac was a Coordination Action project within the EU 6th Framework Programme that linked academic researchers with community journals and industry. The European Commission-funded ProDaC consortium coordinated by Christian Stephan of Medizinisches Proteom-Center, Bochum, Germany aimed to achieve three main goals: (i) to support the PSI efforts, (ii) to provide user-friendly, standard-compliant data conversion tools and (iii) to promote and enable community-wide data collection in standards-compliant public proteomics databases. Ending in March 2009, ProDaC has delivered a comprehensive toolbox of standards and computer programs to achieve its goals [19]. This contains export from local LIMS systems like ProteinScape to standard file formats or direct upload into PRIDE (http://www.medizinisches-proteom-center.de/ProCon). It was already used to store the results and spectra of the Human Brain Proteome Project (http://www.hbpp.org/) into PRIDE (Fig. 5).

4 From Protein Lists to Biological Information & Knowledge

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. Access to popular databases can offer valuable information for proteome researchers. Some of these databases are: the Human Protein Reference Database (HPRD) [20], European Bioinformatics Institute`s suite of databases (EBI) IntAct [21], Search Tool for the Retrieval of Interacting Genes/Protein (STRING) or InterPro [22] for protein interactions; or Kyoto Encyclopedia of Genes and Genomes (KEGG) [23] for metabolic pathways.

There is a vast amount of information that exists in terms of content, annotation and relationships of entries in the various databases. This provides a wealth of information, but due to mixed use of nomenclature, data formats, and accession modes, it has also be proved to be a hurdle to garner functional and non-redundant information. 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 [24].

PIKE gets a commonly used protein list as input such as those obtained from ProteinScape or common search engines e.g. Phenyx, Mascot, or Sequest and then checks in real-time the information stored in a collection of databases mentioned above, and systematically extracts and reports functional and biological information. Fields of interest to be reported, according to the objectives of the current experiment, are: function, cellular location, disease and tissue specificity, Gene Ontology annotations according to biological process, molecular function and cellular component classification (http://www.geneontology.org/), OMIM references, KEGG references, protein and gene names, EBI IntAct molecular interactions.

PIKE also reports all available protein identifiers used in the commonly accessed databases for protein identification (NCBI nr, UniProt and IPI). This information is provided to the user in a wide range of standard output formats, that can be viewed, saved, exported, or downloaded. The system also provides methods to integrate PIKE data into ProteinScape to extend the level of information provided PIKE is freely available from the Spanish Centro Nacional de Biotecnologia (CNB-CSIC) Proteomics Facility website http://proteo.cnb.uam.es:8080/pike/. Other sources of protein meta-information and further biological and functional knowledge are the Protein Center (ProXeon), the NCBI (www.ncbi.nlm.nih.gov), IPI (http://www.ebi.ac.uk/IPI/) and UniProt (http://beta.uniprot.org/) pages that are accessible directly from individual proteins or whole result tables in ProteinScape.

Acknowledgements

The study of the effect of TGF-beta on human lung carcinoma cells was performed with Prof. H.E. Meyer (university Bochum)

References

[1] Stephan, C.; Hamacher, M.; Blüggel, M.; Körting, G.; Chamrad, D.; et.al.: Setting the Analysis Frame, Proteomics 5:3560-62, 2005

[2] Stephan, C.; Reidegeld, K.A.; Hamacher, M.; van Hall, A.; et.al.: Automated reprocessing pipeline for searching heterogeneous mass spectrometric data of the HUPO Brain Proteome Project pilot phase, Proteomics 6: 5015-29, 2006

[3] Hamacher,M., Stephan, C., Meyer, H.E., Eisenacher, M., : *Data handling and processing in proteomics*. Expert Rev Proteomics 2009,6,217-219

[4] Thiele, H.; Glandorf, J.; Hufnagel, P.; Körting, G.; Blüggel, M.: Managing Proteomics Data: From Generation and Data Warehousing to Central Data Repository, J Proteomics Bioinform 1: 485-507, 2008

[5] Peng, J.; Elias, J.E.; Thoreen, C.C.; Licklider, L.J.; Gygi, S.P.: Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large scale protein analysis: the yeast proteome, J Proteome Res 2:43-50, 2003

[6] Cargile, B. J., Bundy, J.L., Stephenson, J. L. Jr., Potential for false positive identifications from large databases through tandem mass spectrometry, J. Proteome Research, 3(5), 1082-85, 2005

[7] Eng, J.K.; McCormack, A.L 3rd; et.al.: 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, 1994

[8] Pappin, D.J.; Hojrup, P.; Bleasby, A.J.: Rapid identification of proteins by peptide-mass fingerprinting, CurrBiol., 3:327-32, 1993

[9] Colinge, J.; Masselot, A.; Cusin, I.; Mahe, E.; et.al.: High-performance peptide identification by tandem mass spectrometry allows reliable automatic data processing in proteomics, Proteomics 4:1977-84, 2004

[10] Chamrad, D,: PhD Thesis, Protagen AG, Germany, 2004

[11] Reidegeld, K.A.; Eisenacher, M.; Jung, K.; Chamrad, D.; Körting, G.; et.al.: An easy-to-use Decoy Database Builder software tool, implementing different decoy strategies for false discovery rate calculation in automated MS/MS protein identifications, Proteomics 8:1129-37, 2008

[12] Thiele, H.; Blüggel, M.: Proteomics Potential, European Biopharmaceutical Review, Dec. 68-73, 2008

[13] Benjamini, Y.; Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, Journal of the Royal Statistical Society 57: 289-300, 1995

[14] Levander, F., Krogh, M., Warell, K., Gärden, P., et. al.: Automated reporting from gel-based proteomics experiments using the open source Proteins database application, Proteomics 7:668-74, 2007

[15] Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search, Anal.. chem.. 74, 5383-5392, 2002

[16] Martens, L.; Hermjakob, H.; Jones, P.; Adamski, M.; Taylor, C.; et.al.: PRIDE: the proteomics identifications database, Proteomics 5:3537-45, 2005

[17] Orchard, S.; Kersey, P.; Hermjakob, H.; & Apweiler, R.: The HUPO Proteomics Standards Initiative Meeting: Towards Common Standards for Exchanging Proteomics Data, Comp. Functional Genomics 4: 16-19, 2003

[18] Taylor, C.F. et.al.: The minimum information about a proteomics experiment (MIAPE), Nat. Biotechnology 25: 887-93 , 2007

[19] Eisenacher, M.; Martens, L.; et.al.: Getting a grip on proteomics data – Proteomics Data Collection (ProDac), Proteomics 9,1-6, 2009

[20] Peri, S.;et.al.: Development of human protein reference database as an initial platform for approaching systems biology in humans, Genome Res 13(10):2363-71, 2003

[21] Hermjakob, H.; et.al.: IntAct: an open source molecular interaction database, Nucleic Acids Res 32 (Database issue): D452-5

[22] Mulder, N.J.; Appweiler, R.: New developments in the InterPro database, Nucleic Acids Res 35 (Database issue): D224-8

[23] Kanehisa, M.; Golo, S.: KEGG: Kyoto encyclopedia of genes and genomes, Nucleic Acids Res 28(1):27-30, 2000

[24] Medina-Aunon, J.A.; Paradela, A.; Thiele, H.; Corthals,G.; Albar, J.P.: Protein Information and Knowledge Extraction (PIKE), Molecular and Cellular Proteomics Journal, (submitted)