Research article

Assembling proteomics data as a prerequisite for the analysis of large scale experiments

Frank Schmidt*[1,4], Monika Schmid[1], Bernd Thiede[2], Klaus-Peter Pleißner[3], Martina Böhme[3] and Peter R Jungblut[1]

Address: [1]Max Planck Institute for Infection Biology, Core Facility Protein Analysis, Berlin, Germany, [2]The Biotechnology Centre of Oslo, University of Oslo, Oslo, Norway, [3]Max Planck Institute for Infection Biology, Core Facility Bioinformatics, Berlin, Germany and [4]Interfaculty Institute for Genetics and Functional Genomics, University of Greifswald, Greifswald, Germany

Email: Frank Schmidt* - Frank.Schmidt@uni-greifswald.de; Monika Schmid - Schmid@mpiib-berlin.mpg.de; Bernd Thiede - Bernd.Thiede@biotek.uio.no; Klaus-Peter Pleißner - pleissner@mpiib-berlin.mpg.de; Martina Böhme - Boehme@mpiib-berlin.mpg.de; Peter R Jungblut - jungblut@mpiib-berlin.mpg.de

* Corresponding author

Abstract

Background: Despite the complete determination of the genome sequence of a huge number of bacteria, their proteomes remain relatively poorly defined. Beside new methods to increase the number of identified proteins new database applications are necessary to store and present results of large- scale proteomics experiments.

Results: In the present study, a database concept has been developed to address these issues and to offer complete information via a web interface. In our concept, the Oracle based data repository system SQL-LIMS plays the central role in the proteomics workflow and was applied to the proteomes of *Mycobacterium tuberculosis*, *Helicobacter pylori*, *Salmonella typhimurium* and protein complexes such as 20S proteasome. Technical operations of our proteomics labs were used as the standard for SQL-LIMS template creation. By means of a Java based data parser, post-processed data of different approaches, such as LC/ESI-MS, MALDI-MS and 2-D gel electrophoresis (2-DE), were stored in SQL-LIMS. A minimum set of the proteomics data were transferred in our public 2D-PAGE database using a Java based interface (Data Transfer Tool) with the requirements of the PEDRo standardization. Furthermore, the stored proteomics data were extractable out of SQL-LIMS via XML.

Conclusion: The Oracle based data repository system SQL-LIMS played the central role in the proteomics workflow concept. Technical operations of our proteomics labs were used as standards for SQL-LIMS templates. Using a Java based parser, post-processed data of different approaches such as LC/ESI-MS, MALDI-MS and 1-DE and 2-DE were stored in SQL-LIMS. Thus, unique data formats of different instruments were unified and stored in SQL-LIMS tables. Moreover, a unique submission identifier allowed fast access to all experimental data. This was the main advantage compared to multi software solutions, especially if personnel fluctuations are high. Moreover, large scale and high-throughput experiments must be managed in a comprehensive repository system such as SQL-LIMS, to query results in a systematic manner. On the other hand, these database systems are expensive and require at least one full time administrator and specialized lab manager. Moreover, the high technical dynamics in proteomics may cause problems to adjust new data formats. To summarize, SQL-LIMS met the requirements of proteomics data handling especially in skilled processes such as gel-electrophoresis or mass spectrometry and fulfilled the PSI standardization criteria. The data transfer into a public domain via DTT facilitated validation of proteomics data. Additionally, evaluation of mass spectra by post-processing using MS-Screener improved the reliability of mass analysis and prevented storage of data junk.

Background

A major goal of proteomics is the large-scale study of proteins, particularly their structures and functions including the global qualitative and quantitative analysis of proteins in defined biological systems. The term proteomics was chosen to make an analogy with genomics, but proteomics is significantly more complex. As a result of alternative splicing, point-mutations, degradations and co- and post-translational modifications, the number of protein species [1] of a proteome exceeds by far the number of protein-coding genes of the corresponding genome. In the past, qualitative proteome profiling has overcome limitations in protein identification due to the amazing developments in mass spectrometry. Increased sensitivity and mass accuracy in conjunction with comprehensive database annotations allows the high-throughput identification of proteins. On the other hand, quantitative profiling, an essential part of proteomics, requires technologies that accurately, reproducibly, and comprehensively quantify proteins. During the past years, novel mass spectrometry-based methods such as ICAT [2], SILAC [3] and iTRAQ [4] were developed for relative quantification. The amount of identification and quantification data increased dramatically during the recent years and resulted in the accumulation of "metadata", which means data about data. The manufacturers of ESI-MS and MALDI-MS instruments and image analysis software have endeavored to close the gap between the increased amount of information and its interpretation. However, this mostly resulted in individual solutions for each company which hampered the exchange of experimental data. However, beside commercial solutions some open LIMS systems such as PROTEIOS [5] or the open source laboratory information management system for 2-D gel electrophoresis-based proteomics workflows [6] are available free of charge and some of them were compared in more detail by Piggee et al. [7]. The representation of protein data must be standardized to compare proteomics results worldwide. For this purpose, some solutions were proposed, such as the Proteome Standards Initiative (PSI) [8,9], and PEDRo [10]. The latter yielded to adapt XML or specialized mzXML [11] or mzML [12] which are open file formats for data exchange.

In our concept, the Oracle-based data repository system SQL-LIMS™ (Applied Biosystems, Foster City, USA) plays the central role in the proteomics workflow and was applied to the proteomes of *Mycobacterium tuberculosis*, *Helicobacter pylori*, *Salmonella typhimurium* and protein complexes such as the 20S proteasome. Technical operations of our proteomics workflow were used as the standard for SQL-LIMS™ template creation. Post-processed data of different approaches, such as LC/ESI-MS, MALDI-MS and 2-DE gel electrophoresis were stored in SQL-LIMS™ by using a Jav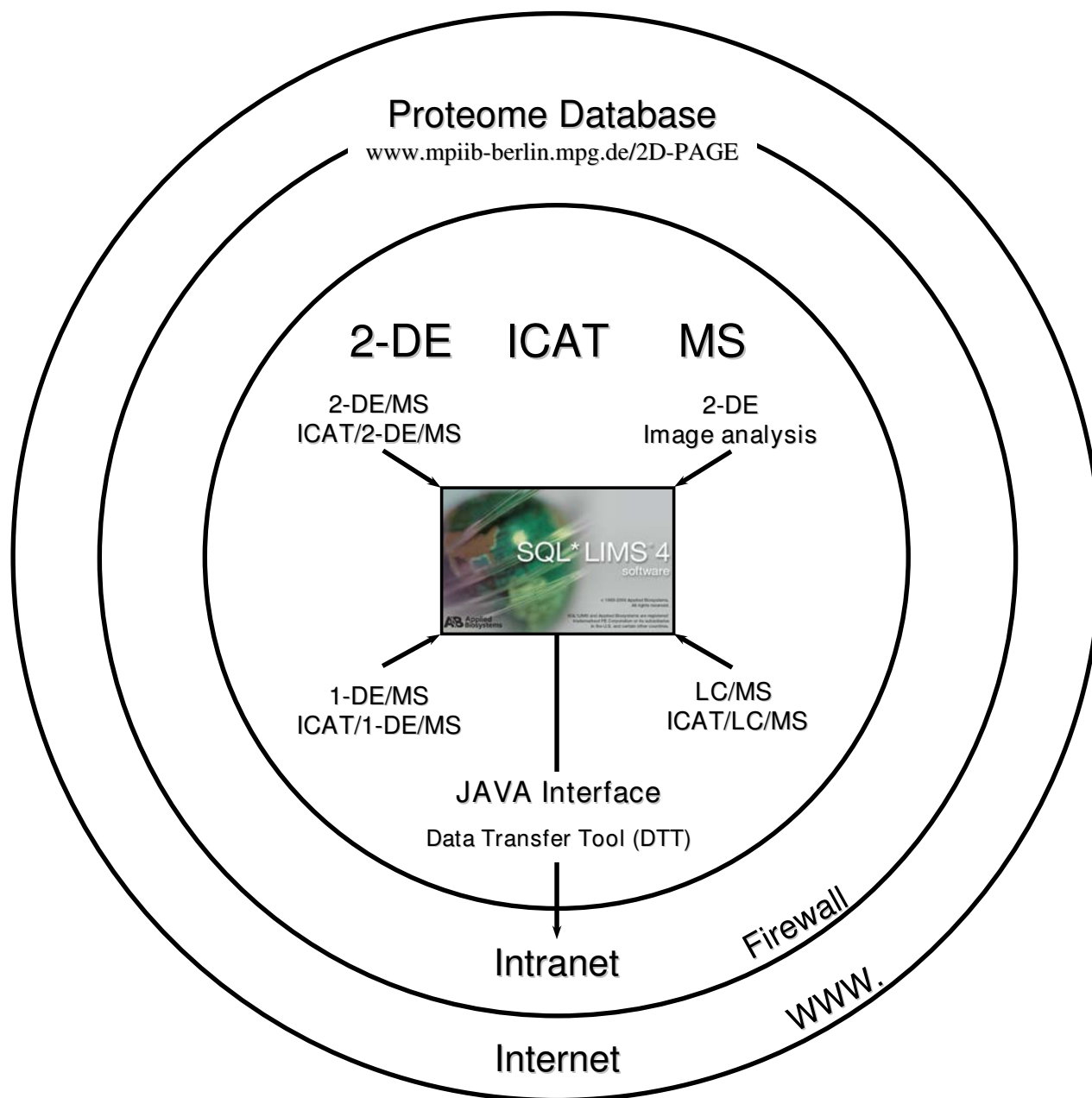a-based data parser. A minimum set of the proteomics data were transferred into the web-accessible Proteome Database System for Microbial Research http://www.mpiib-berlin.mpg.de/2D-PAGE/[13] using a Java-based interface (Data Transfer Tool) with the requirements of the PEDRo standardization. Furthermore, the stored proteomics data were extractable out of SQL-LIMS™ as XML documents.

Results and discussion

Concept for integration of proteomics data

We applied a variety of 2-DE and LC-based approaches for the comprehensive proteome analysis of microorganisms and other protein complexes. These technologies included 2-DE/MS coupled with image analysis, 1-DE/MS, ICAT/1-DE/MS, ICAT/2-DE/MS, LC/MS and ICAT/LC/MS (Figure 1). The enormous amount of information generated by these proteome analyses required the application of suitable programs for data integration and a repository in order to gain maximum benefit from the experimental results. In the past, the urgent need for such programs has often been emphasized, but the development of adequate programs was hampered by the large diversity of data formats. SQL*LIMS™ enabled the integration and storage of data emerging from all kinds of proteome analyses, e.g. sample preparation, 2-DE analyses, as well as raw and evaluated MS data. This allowed efficient data handling, in particular for the evaluation of large experimental datasets. Moreover, the storage of metadata produced during the laboratory work was established. Therefore, specific templates were created where the whole workflows as well as the protocols were implemented including all information about the biological samples and the applied preparation steps. In order to connect the metadata with the results from 2-DE and mass spectrometry measurements, binary files such as from image analysis calculations, mass spectrometry peak lists or identification results from Mascot or SEQUEST were parsed into an Oracle database. Raw data such as 2-DE gel images or MS spectra were stored as attachments. Furthermore, most of the experimental data were post-processed before storing into the SQL*LIMS™, e.g. by MS-Screener [14], to decrease the amount of data junk.

However, there is no doubt that administration of programs such as SQL*LIMS™ are time consuming due to difficulties in template and interface programming. Thus, SQL*LIMS™ needed to be maintained by at least one full time administrator and specialized lab-manager. To overcome extensive training in SQL*LIMS™ and to make proteomics data available, we have developed a data transfer tool (DTT) as shown in Figure 1. This interface means that experimental data stored in SQL*LIMS™ can automatically be transferred into the Proteome Database System, which makes the results easily accessible. In this domain, authorized persons have access to all evaluated data. In

Figure 1
Concept of proteomics data assembling. Data of the proteomics workflow such as 2-DE/MS coupled with image analysis, ICAT/2-DE/MS, ICAT/1-DE/MS, LC/MS, and ICAT/LC/MS was stored in SQL*LIMS™. Raw data were stored as an attachment or link to their repository. As a result, the data were transferred and unified into one system. Subsequently, a minimum set of information about the proteomics experiments and results were transferred into the web-accessible Proteome Database System for Microbial Research using the data transfer tool (DTT).

the Proteome Database, experimental data were linked with protein databases, such as Swiss-Prot/UniProt, NCBI or KEGG. Moreover, a higher-level investigation of the data can be performed using the large number of sophisticated functions and packages of the software environ- ment for statistical computing and graphics R http://www.r-project.org/. The advantage of this concept is that all information from different experiments is gathered in one system used for daily laboratory needs and which complements the web-accessible database system used for

data dissemination. The users have a unique and easy access to complex data sets. Moreover, already published experimental data can be transferred into the public internet domain.

Data storing in SQL*LIMS™

The requirements for data storing in SQL*LIMS™ depend on the experimental workflow. As a result, the data management system must contain specifically designed features (Figure 2). In order to structure different experiments, SQL*LIMS™ allowed to define studies. During the study initialization, only predefined attributes must be stored to save information about the frame and the goal of the experiment. For sample preparation, flexibility was also very important to track experimental workflows which often included a complex sequence of

operations. To meet these requirements, predefined basic sample types were combined in a hierarchical parent-child relationship tree and new attributes can be added to the predefined types. The following protein separation step required a full integration with 2-DE gel image analysis tools (e.g. Topspot, and PDQuest). Data of detected 2-DE spots or 1-DE bands were automatically acquired (uploaded) from image processing tool output files along with gel image pictures. In this case, structured data with fixed formats, such as spot coordinates, intensities or spot volumes, come together with unstructured raw data which have no common format, such as gel images and native report files. The information was spread into all these different records in the storage system, but it was available for the user as a single bulk through a gel viewer widget that allowed the drilling of spot information. In the MS
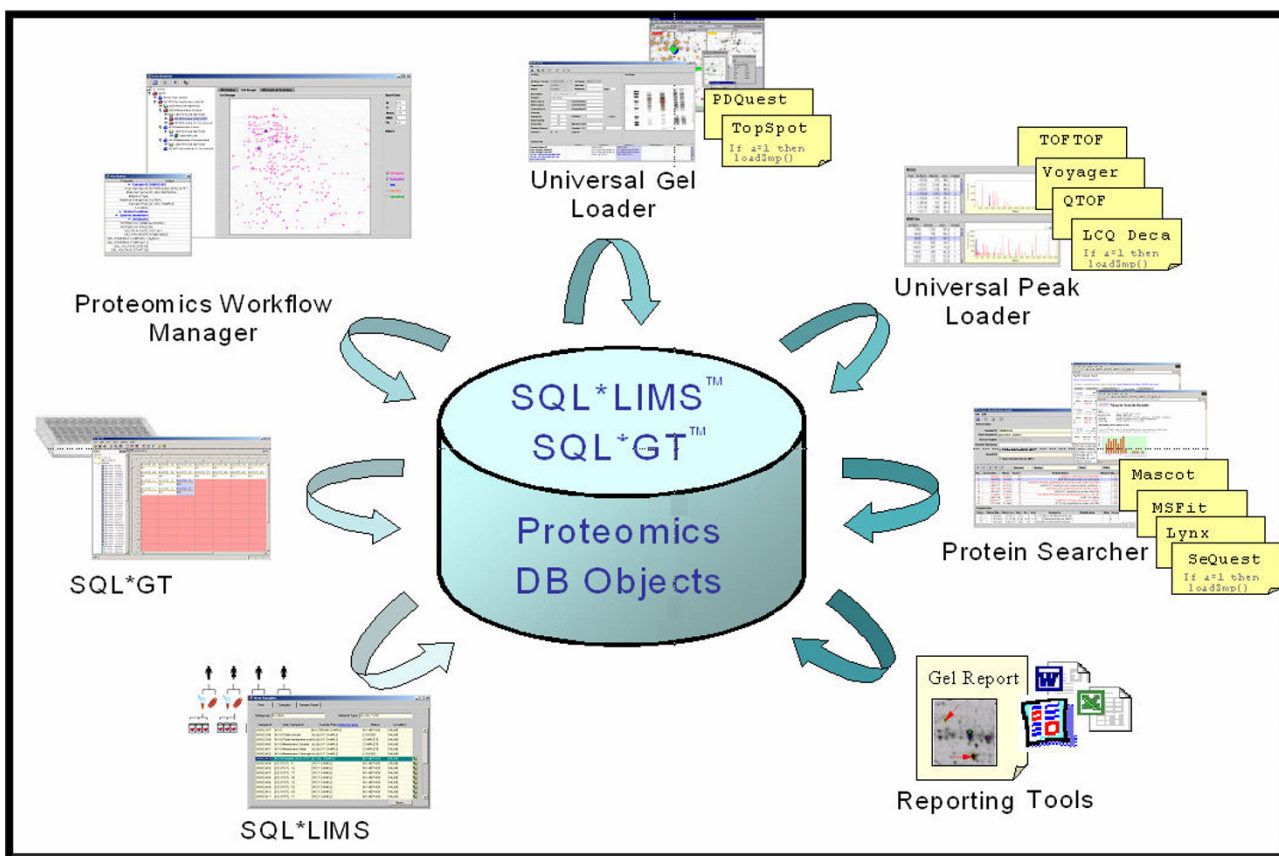
Figure 2
The core of the SQL*LIMS™ system and the integrated applications SQL*GT (microtiter plate solution) and Proteomics Solution. In a first step, proteomics studies can be defined by SQL *LIMS™ or SQL*GT for microtiter plates and structured by the Proteomics Workflow Manager. As a result, any experiments get a unique submission and sample identifier. Using the Proteomics DB Objects, gel images (Universal Gel Loader) and mass spectra (Universal Peak Loader) can be assigned to a specific study and evaluated by the Protein Searcher. Moreover, existing identification results from Mascot (.html, .dat), MS-Fit (.html), SEQUEST (.xls), or Lynx (.txt) can be parsed to SQL*LIMS™. Furthermore, experiment specific data can be queried and reported by the program Query Builder.

analysis step, both structured and unstructured data must be managed as in the protein separation step. In addition, features for the direct exchange of real-time bi-directional data with MS instruments were provided for the work lists uploading and the peak lists downloading. A more complex strategy for unstructured data management was required due to the massive amounts of raw data with proprietary format. Instead of passively storing all the data in the SQL*LIMS™ database without any chance to extract the content for data searching, raw data files were saved into the storage server and/or the permanent storage supports. The locations can then be tracked into the SQL*LIMS™. For protein identification, MS peak lists were submitted to search database engines such as Mascot, MS-Fit, MS-Tag, and SEQUEST. Queries can be performed directly from SQL*LIMS™ or by using the search engine front-end interface. In both cases, protein identification results were stored in the SQL*LIMS™ database. Again, the need for managing unstructured electronic records was coped. The stored gel information and protein identification results of proteomics studies can now easily be reported by the Report Builder, printed out or sent as an e-mail.

Transfer of SQL*LIMS™ data into the Intranet/Internet database via DTT

In order to share the experimental results with other laboratories rather easily, the DTT was designed to facilitate the transfer out of the SQL*LIMS™ into the proteome database system (Figure 3). The DTT provides a GUI to enable the user to select datasets to be transferred. First, only necessary data records were selected among the vast amount of data stored in the SQL*LIMS™. The DTT displayed the gel data from SQL*LIMS™ corresponding to the identification numbers in the 2-DE database (Figure 3B). The user can choose a gel and the relevant data for the displayed spots. This included for example the sequence coverage, score, rank and molecular weight of identified proteins for each spot on the gel in question (Figure 3C). It is also possible to check newer entries for a spot in the 2-DE database which will be updated in the transfer process. A release number was assigned to selected protein identification data for the transfer. These release numbers can be used to control the degree of accessibility of the data in the 2-DE database. Thus, it was possible to restrict the view for certain release numbers. The usage of the DTT was password protected and all data transactions were logged. If the button "Transfer" was selected and "Save" was pressed, an existing release number was selected for this spot or assigned to a new one. The concept of the release numbers was established in order to control public access and to inform SQL*LIMS™ managers if a new protein could be identified due to new batch searches using the newest database releases. Currently, only data displayed in Figure 3C can be transferred but it is planned to

include MS-data extracted from the ms peak SQL*LIMS™ table (Table 1).

Pre and post-processing LC/ESI-MS/MS data

Tandem mass spectrometry has been particularly useful for determining the protein components of complex mixtures. The following strategy was applied to evaluate LC/ESI-MS/MS peak lists data: MS/MS spectra were automatically transformed into peak lists (.dta-files) by SEQUEST and subsequently imported into MS-Screener for generating data matrices. The binary matrices were subjected to hierarchical agglomerative cluster analyses performed by means of the hclust- function within R. To illustrate an example of cluster analyses, Figure 4 depicts a result of a dataset comprising 873 MS/MS peak lists. The spectra were derived from a comparative analysis of two rat liver proteasome subtypes [15] measured by ICAT/LC/ESI-MS/MS. The cluster analysis was applied to determine which spectra showed some degree of similarity, i.e. common peptide or experiment specific mass peaks. The analysis resulted in a dendrogram of mass spectra with high similarity formed branches. The subset of spectra framed in Figure 4 shared numerous polymer contaminant masses, caused by avidin chromatography which was used for the separation of ICAT labeled peptides. As a result, more than half of the 125 polymer mass spectra have been falsely fitted to a rat specific peptide and the Mascot MS/MS ion search result shown in Figure 4 represents a false positive match of a MS/MS spectrum. After removing the polymers containing mass spectra from the data set, the remaining 748 of 873 mass spectra were searched again. Subsequently, only the spectra of relevance and their SEQUEST or Mascot identification results were stored in SQL*LIMS™. The strategy outlined in this section demonstrated the capability to clean peak list data sets from contaminants in order to improve the reliability of identifications and to reduce the number of stored data.

Pre and post-processing MALDI-MS data

Proteins separated by 2-DE were identified by peptide mass fingerprinting (PMF) after in-gel digestion. A Voyager Elite MALDI-TOF mass spectrometer and/or a 4700 Proteomics Analyzer MALDI-TOF/TOF instrument were used for this purpose. MS peak lists were generated by the program GRAMS or the peak-to-mascot script of the program 4700 Explorer™. In addition, the peak lists were evaluated by the program MS-Screener. Experimentally derived contaminant masses, e.g., masses matching to matrix, keratins, and autolysis products of trypsin or dye were detected and deleted from the spectra [14]. The simplified peak lists were analyzed by PMF using search algorithms, such as Mascot or MS-Fit. Subsequently, the modified peak list and the identification results were parsed and stored in SQL*LIMS™.
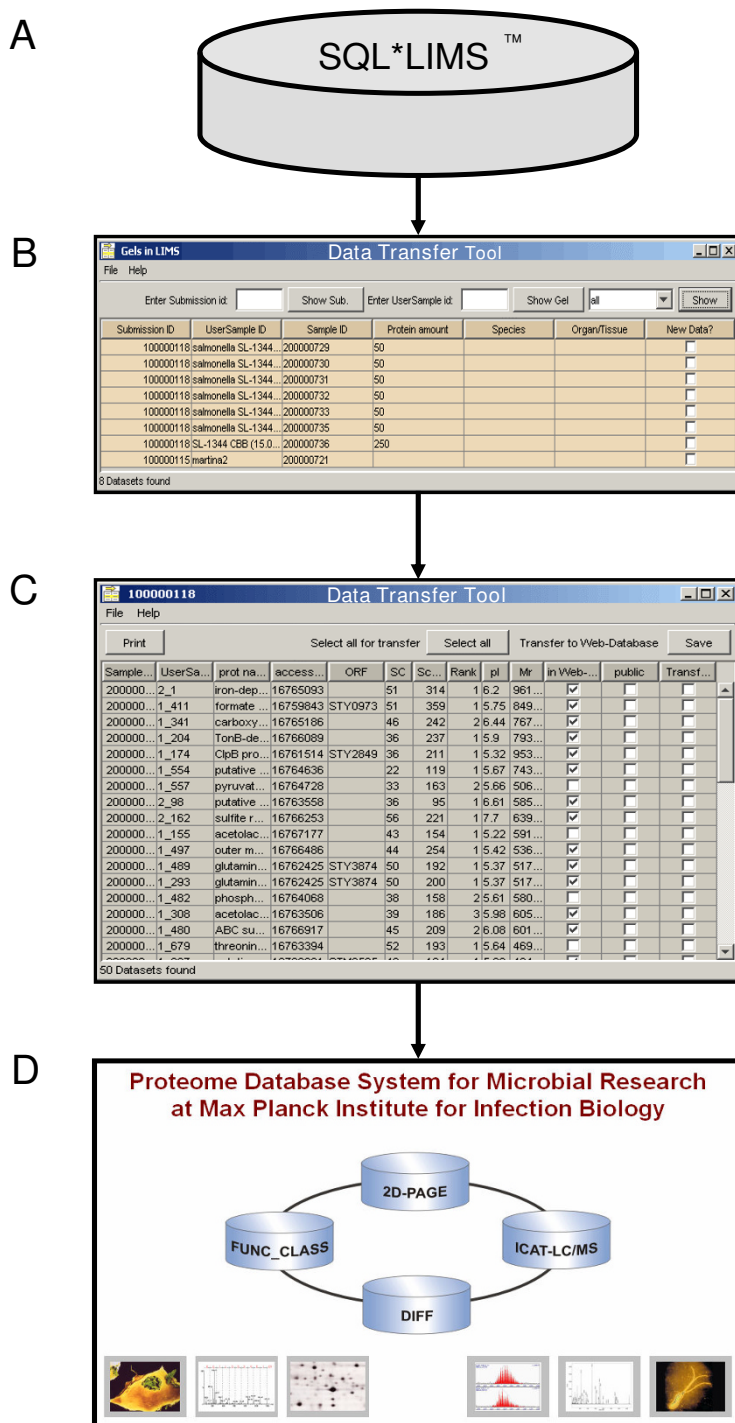
Figure 3
Proteome data transmission. A data transfer tool can be used to extract data from SQL*LIMS™ (A, B). In the first step, gels of a study were listed (B). In the second step, identified proteins and their main information e.g. name, accession number, ORF, sequence coverage, score and rank will be shown in a new window (C). Finally, protein identification data of interest can be selected for the transfer to the intra or public domain of the Proteome Database System (D).

Table 1: Overview of the different file formats stored in SQL *LIMS™

Workflow Step	Structured Data	Unstructured Data
Sample Preparation	- Combination of basic sample types - > 100 sample attributes	
Protein Separation	Spot data	- PDQuest® gel image (.tif files) - PDQuest® spot reports (text ASCII files) - Topspot gel image (.jpg files) - Topspot spot report (text ASCII files)
MS Analysis	MS peaklist data	- Thermo Finnigan LCQ™ DECA peaklist files (text ASCII files) - Thermo Finnigan LCQ™ raw data location reference - MALDI Voyager Elite raw peaklist files (text ASCII files)
Protein Identification	Protein data Peptide data	- Matrix Science Mascot® PMF v 1.8 search result (.html files) - Matrix Science Mascot® MSMS Query v 1.8 search result (.html files) - Matrix Science Mascot® Sequence v 1.8 search result (.html file) - ProteinProspector MS-Fit® v 3.2.1, 3.4.1 search result (.html files) - ProteinProspector MS-Tag® v 3.2.1, 3.4.1 search result (.html files) - SEQUEST® v 2.0 search result (.xls files)

Experimental

Two-dimensional electrophoresis (2-DE)

Protein samples from microorganisms were subjected to high-resolution 2-DE (gel size: 23 cm × 30 cm) [16]. Generated 2-DE gels were scanned and subsequently evaluated by image analysis programs Topspot (Algorithmus, Berlin, Germany) [17] and PDQuest (Bio-Rad Laboratories, Hercules, CA, USA).

Automated 2-DE spot processing

High-throughput MALDI-MS PMF was performed as follows: Spots of interest were excised from 2-DE gels, transferred into 96-well microtiter plates, and digested with trypsin using a spot-cutter (Proteome Works, Bio-Rad, Hercules, CA, USA). Subsequently, equal volume of resulting peptides and α-cyano-4-hydroxycinnamic acid (CHCA) were mixed and spotted onto MALDI templates by the Ettan spot-handling workstation (Amersham Biosciences, Uppsala, Schweden). Subsequently, MALDI spectra were internally calibrated and the resulting peak lists exported using the "Peak-to-Mascot" script of the 4700 Explorer software (Version 2.0) (Applied Biosystems, Foster City, USA). The parameters applied for this process were optimized (signal-to-noise ratio, mass range, peak density, etc.). Afterwards, the MS-Screener program was used to determine and to remove common contaminant masses.

Data analysis by MS-Screener

The program MS-Screener (Version 1.0.1) was applied to evaluate large datasets of peak lists. This program comprised 162 Java classes and has been developed for Java 2 Runtime Environment (Version J2RE 1.4.1; http://java.sun.com/). MS-Screener offers a multi-platform support for Linux, Solaris and Microsoft Windows including a helpful graphical user interface (GUI). Graphical representations of peak lists as plot-views have been integrated using the JFreeChart class library (Version 0.9.13) http://www.jfree.org/jfreechart/index.html published under the GNU Lesser General Public License. MS-Screener facilitates the import and export of ASCII files (.pkm (GRAMS, Applied Biosystems, Framingham, USA), .pkt (Data Explorer, Applied Biosystems, Framingham, USA), .txt (Peak-to-Mascot, 4700 Explorer, Applied Biosystems, Framingham, USA) and .dta (SEQUEST, Thermo Finnigan, San Jose, USA)) and data exchange via other interfaces. MS-Screener was used for many tasks, e.g. the detection of common mass peaks, the elimination of contaminant masses, and the calculation of the half decimal places rule [14]. Furthermore, it was used to generate peak lists matrices as a prerequisite for cluster analyses using R. Moreover, the recalibration of binary peak lists and a peak pair comparison tool to determine ICAT ratios were applied. The MS-Screener results were transformed in tab-separated files (.txt) to transfer the data into SQL*LIMS™.

Mass spectrometry and protein identification/quantification

For protein identifications, 2-DE spots were analyzed by MALDI-MS or MS/MS or ESI-MS/MS [16,18-20]. In most cases, spots to be identified were digested by trypsin prior to MS analysis [21]. MALDI-MS was carried out using a Voyager Elite MALDI-TOF mass spectrometer or a 4700 Proteomics Analyzer MALDI-TOF/TOF (both from Applied Biosystems, Framingham, USA). Protein identifications were achieved by database comparisons using search algorithms such as Mascot [22] or MS-FIT http://prospector.ucsf.edu, whereby Mascot was available as in-house version. Searches were accomplished either individually or in batch mode (analysis of large datasets). In the latter case, Mascot-Daemon http://www.matrix_science.com was used as batch interface. Individual searches
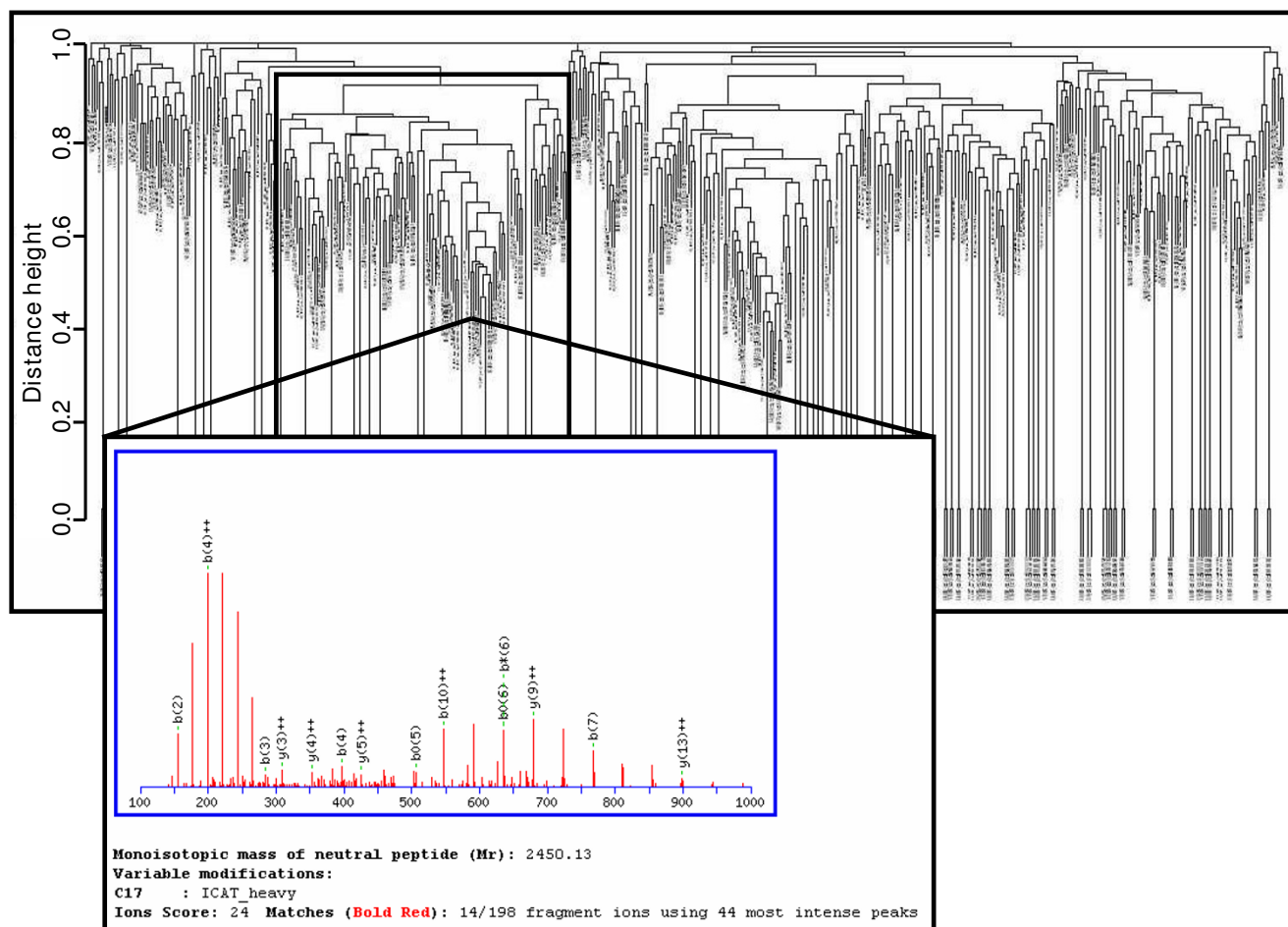
Figure 4
Post-processing: Clustering of polymer contaminants. A cluster dendrogram comprising 873 MS/MS mass spectra is shown. The data were recorded using a comparative ICAT/LC/ESI-MS study of proteasome subtypes from rat liver. The framed part of the dendrogram shows a cluster of 125 similar mass spectra originated by polymer contaminations due to avidin column purification steps. Half of these mass spectra falsely fitted to proteasome peptides using Mascot search algorithm. These junk mass spectra were eliminated before data were stored into SQL*LIMS™.

were performed by the Mascot web-front end or the SQL-LIMS™ clients, respectively, and both were connected with in-house Mascot server. The search parameters applied have previously been described [21]. Moreover, proteins were separated and identified by large-scale on-line LC/ESI-MS/MS. The protein samples were prepared as described [23] and measured by LCQ ion trap mass spectrometer (Thermo Finnigan, San Jose, USA). For peptide identifications, the generated MS/MS spectra were evaluated using the SEQUEST analysis program and/or Mascot. In order to quantify differences between 20S proteasome subtypes [15,24] and proteomes of *M. tuberculosis* and *bovis BCG* [23], proteins were labelled with the ICAT reagent and analyzed by LC/ESI-MS/MS. To calculate the relative ratios, MS-spectra were evaluated by the program Xpress. Furthermore, a complementary approach was

used to detect differences in protein abundance, which combines ICAT and 2-DE and were quantified by the program MS-Screener [24]. An iterative search procedure was applied for in-depth analysis of large 2-DE/MALDI-MS datasets [14].

SQL-LIMS™ Proteomics Solution
The workflow described above requires a suitable system for the integration and management of raw and processed experimental data. These issues were addressed by the Laboratory Information Management System (LIMS) in combination with an implemented SQL*LIMS™ Proteomics Solution, customized for our proteomics research laboratory. The implemented solution was based on the Applied Biosystems™ product suite for life science, including a core application (SQL*LIMS™). The latter was

designed for analytical laboratories, Pharma R&D and manufacturing environments. Furthermore, components specifically designed for microtiter plates (SQL*GT™) and proteomics (Proteomics Solution) data management were implemented. Operating flexibility and extensibility of this solution has minimized the requirement for code customization. The SQL*LIMS™ users are allowed to enter new or to amend existing workflows and to open interfaces providing an add-on and built-in mechanism for the integration of MS instruments and third-party tools. A highly integrated environment has been addressed from the very beginning as a key factor to enhance productivity by streamlining time consuming operations such as MS data exchange (work list uploading and peak list downloading) or protein search engines querying.

Data transfer tool Java interface (DTT)
The data transfer tool was designed to facilitate the data transfer from the SQL*LIMS™ into the public 2-DE database, which is the essential part of our Proteome Database System http://www.mpiib-berlin.mpg.de/2D-PAGE/. The DTT has been developed in Java using J2SE 1.4 http://java.sun.com/j2se/1.4 and Eclipse http://www.eclipse.org. The program comprised a graphical user interface (GUI) to enable the selection of datasets which were to be transferred. For safety reasons, the data transfers out of SQL-LIMS™ were protected by password accession.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
FS and MS carried out the proteomics studies, participated in the database structure and the template creation, and prepared the manuscript. KPP contributed to the concept and the realization of the 2D-PAGE and the SQL-LIMS database. MB participated in the DTT tool development. BT participated in the realization of the manuscript. PRJ coordinated and conceived of the study, and participated in its design. All authors read and approved the final manuscript.

Acknowledgements

References
1. Jungblut PR, Holzhütter HG, Apweiler R, Schlüter H: **The speciation of the proteome.** *Chemistry Central Journal* 2008, **2**:1-10.
2. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R: **Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.** *Nat Biotechnol* 1999, **17**:994-999.
3. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M: **Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics.** *Mol Cell Proteomics* 2002, **1**:376-386.
4. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlet-Jones M, He F, Jacobson A, Pappin DJ: **Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-Reactive Isobaric Tagging Reagents.** *Mol Cell Proteomics* 2004, **3**:1154-1169.
5. Gärdén P, Alm R, Häkkinen J: **PROTEIOS: an open source proteomics initiative.** *BMC Bioinformatics* 2005, **21**:2085-2087.
6. Morisawa H, Hirota M, Toda T: **Development of an open source laboratory information management system for 2-D gel electrophoresis-based proteomics workflow.** *BMC Bioinformatics* 2006, **7**:430.
7. Piggee C: **LIMS and the art of MS proteomics.** *Anal Chemi* 2008, **1**:4801-4806.
8. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, Roechert B, Poux S, Jung E, Mersch H, Kersey P, Lappe M, Li Y, Zeng R, Rana D, Nikolski M, Husi H, Brun C, Shanker K, Grant SG, Sander C, Bork P, Zhu W, Pandey A, Brazma A, Jacq B, Vidal M, Sherman D, Legrain P, Cesareni G, Xenarios I, Eisenberg D, Steipe B, Hogue C, Apweiler R: **The HUPO PSI's Molecular Interaction format-a community standard for the representation of protein interaction data.** *Nat Biotechnol* 2004, **22**:77-183.
9. Orchard S, Martens L, Tasman J, Binz BA, Albar JP, Hermjakob H: **6th HUPO Annual World Congress – Proteomics Standards Initiative Workshop 6–10 October 2007, Seoul, Korea.** *Proteomics* 2008, **7**:1331-1333.
10. Taylor CF, Paton NW, Garwood KL, Kirby PD, Stead DA, Yin Z, Deutsch EW, Selway L, Walker J, Riba-Garcia I, Mohammed S, Deery MJ, Howard JA, Dunkley T, Aebersold R, Kell DB, Lilley KS, Roepstorff P, Yates JR 3rd, Brass A, Brown AJ, Cash P, Gaskell SJ, Hubbard SJ, Oliver SG: **A systematic approach to modeling, capturing, and disseminating proteomics experimental data.** *Nat Biotechnol* 2003, **21**:247-254.
11. Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R, Cheung K, Costello CE, Hermjakob H, Huang S, Julian RK, Kapp E, McComb ME, Oliver SG, Omenn G, Paton NW, Simpson R, Smith R, Taylor CF, Zhu W, Aebersold R: **A common open representation of mass spectrometry data and its application to proteomics research.** *Nat Biotechnol* 2004, **22**:1459-1466.
12. Deutsch E: **mzML: a single, unifying data format for mass spectrometer output.** *Proteomics* 2008, **14**:2776-2777.
13. Pleissner KP, Eifert T, Buettner S, Schmidt F, Boehme M, Meyer TF, Kaufmann SH, Jungblut PR: **Web-accessible proteome databases for microbial research.** *Proteomics* 2004, **5**:1305-1313.
14. Schmidt F, Schmid M, Mattow J, Facius A, Pleissner KP, Jungblut PR: **Iterative data analysis is the key for exhaustive analysis of peptide mass fingerprints from proteins separated by two-dimensional electrophoresis.** *J Am Soc Mass Spectrom* 2003, **14**:943-956.
15. Dahlmann B, Ruppert T, Kuehn L, Merforth S, Kloetzel PM: **Different proteasome subtypes in a single tissue exhibit different enzymatic properties.** *J Mol Biol* 2000, **10**:643-653.
16. Klose J: **Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues A novel approach to testing for induced point mutations in mammals.** *Humangenetik* 1975, **26**:231-243.
17. Prehm J, Jungblut PR, Klose J: **Analysis of two dimensional protein patterns using a video camera and a computer.** *Electrophoresis* 1987, **8**:562-572.
18. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM: **Electrospray Ionization-Principles and Practice.** *Science* 1989, **246**:64-71.
19. Tanaka K, Waki H, Ido Y, Akita S, Yoshida Y, Yoshida T: **Protein and polymer analyses up to m/z 100,000 by laser ionization time-of-flight mass spectrometry.** *Rapid Commun Mass Spectrom* 1988, **2**:151-153.
20. Karas M, Hillenkamp F: **Laser Desorption/Ionization of Proteins with Molecular Masses Exceeding 100,000 Daltons.** *Anal Chem* 1988, **60**:2299-2301.
21. Thiede B, Hohenwarter W, Krah A, Mattow J, Schmid M, Schmidt F, Jungblut PR: **Peptide mass fingerprinting.** *Methods* 2005, **35**:237-247.
22. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS: **Probability-based protein identification by searching sequence databases using mass spectrometry data.** *Electrophoresis* 1999, **20**:3551-3567.

23. Schmidt F, Donahoe S, Hagens K, Mattow J, Schaible UE, Kaufmann SH, Aebersold R, Jungblut PR: **Complementary analysis of the Mycobacterium tuberculosis proteome by two-dimensional electrophoresis and isotope-coded affinity tag technology.** *Mol Cell Proteomics* 2004, **3**:24-42.

24. Schmidt F, Dahlmann B, Janek K, Kloss A, Wacker M, Ackermann R, Thiede B, Jungblut PR: **Comprehensive quantitative proteome analysis of 20S proteasome subtypes from rat liver by isotope coded affinity tag and 2-D gel-based approaches.** *Proteomics* 2006, **6**:4622-4632.

25. Smolka MB, Zhou H, Purkayastha S, Aebersold R: **Quantitative Protein Profiling Using Two-dimensional Gel Electrophoresis, Isotope-coded Affinity Tag Labeling, and Mass Spectrometry.** *Mol Cell Proteomics* 2002, **1**:19-29.