Web-accessible proteome databases for microbial research

Klaus-Peter Pleißner, Till Eifert, Sven Buettner, Frank Schmidt, Martina Boehme, Thomas F. Meyer, Stefan H. E. Kaufmann and Peter R. Jungblut

Max Planck Institute for Infection Biology, Berlin, Germany

The analysis of proteomes of biological organisms represents a major challenge of the post-genome era. Classical proteomics combines two-dimensional electrophoresis (2-DE) and mass spectrometry (MS) for the identification of proteins. Novel technologies such as isotope coded affinity tag (ICAT)-liquid chromatography/mass spectrometry (LC/MS) open new insights into protein alterations. The vast amount and diverse types of proteomic data require adequate web-accessible computational and database technologies for storage, integration, dissemination, analysis and visualization. A proteome database system (http://www.mpiib-berlin.mpg.de/2D-PAGE) for microbial research has been constructed which integrates 2-DE/MS, ICAT-LC/MS and functional classification data of proteins with genomic, metabolic and other biological knowledge sources. The two-dimensional polyacrylamide gel electrophoresis database delivers experimental data on microbial proteins including mass spectra for the validation of protein identification. The ICAT-LC/MS database comprises experimental data for protein alterations of mycobacterial strains BCG *vs.* H37Rv. By formulating complex queries within a functional protein classification database "FUNC_CLASS" for *Mycobacterium tuberculosis* and *Helicobacter pylori* the researcher can gather precise information on genes, proteins, protein classes and metabolic pathways. The use of the R language in the database architecture allows high-level data analysis and visualization to be performed "on-the-fly". The database system is centrally administrated, and investigators without specific bioinformatic competence in database construction can submit their data. The database system also serves as a template for a prototype of a European Proteome Database of Pathogenic Bacteria. Currently, the database system includes proteome information for six strains of microorganisms.

Received	10/10/03
Revised	22/12/03
Accepted	29/12/03

Keywords: Bioinformatics / Databases / Microorganisms / Proteomics / Two-dimensional gel electrophoresis

1 Introduction

The analysis of proteomes of diverse biological organisms represents one of the challenges in the post-genome era and is a rich source of biological information [1]. In contrast, to sequence data of more than 90 bacterial genomes [2, 3] accessible *via* public databases, proteome data are characterized by diverse data types and are stored in proprietary databases located worldwide. The diversity of data is due to the various methods applied for proteome research such as 2-DE/MS, isotope coded affinity tag (ICAT)-LC/MS, protein sequencing, and other methods. Proteome databases were established by different research groups and in different ways. Databases such as SWISS-2DPAGE, 2D-PAGE, HSC-2DPAGE referred to in the WORLD-2DPAGE index (http://www.expasy.org/ch2d/2d-index.html) or ProteomeWeb [4] and YPRC-PDB [5] can serve as examples. Additionally, efforts have been made to unite such heterogeneous databases by defining a set of rules for federation [6] or to create standards for modeling, capturing and disseminating proteome experimental data [7]. Further immunologic/bacterial

Correspondence: Dr. Peter R. Jungblut, Max Planck Institute for Infection Biology, Central Core Facility Protein Analysis, Schumannstrasse 21–22, D-10117 Berlin, Germany
E-mail: jungblut@mpiib-berlin.mpg.de
Fax: +49-30-28460507

Abbreviations: EBP, European Proteome Database of Pathogenic Bacteria; **ICAT**, isotope coded affinity tag; **LIMS**, laboratory information management system

proteome databases deal with human primary T helper cells [8], *Streptomyces coelicolor* (http://proteom.biomed.cas.cz/strepto/cc1_strep.php) and *Bacillus subtilis* (http://microbio2.biologie.uni-greifswald.de:8880/sub2d), for instance.

For immunologic research, we published our proprietary microbial proteome 2-DE database "2D-PAGE" [9, 10]. This database comprises data for *Mycobacterium tuberculosis, Helicobacter pylori, Chlamydophila pneumoniae, Borrelia garinii, Francisella tularensis* and *Mycoplasma pneumoniae*. Additionally, proteome data of Jurkat-T cells, mammary gland (mouse) and rat heart may be obtained. The 2D-PAGE database is mainly characterized by the consequent application of the relational data model for database construction and the application of open source software tools. Here we describe our web-accessible proteome database system for microbial research which reflects an effort to integrate 2-DE/MS, ICAT-LC/MS and functional classification data of proteins with genomic, metabolic and other knowledge sources in molecular biology, such as 3D-structure or protein-protein interaction databases. By formulating complex biological queries the researcher can gather information on genes, proteins, functional protein classes of *M. tuberculosis* and *H. pylori* and metabolic pathways at a glance. Thus, the storage, analysis and visualization of proteomic data contribute to a more holistic view on microorganisms.

2 Materials and methods

2.1 Data generation and data storage

The proteome database contains diverse data types generated by 2-DE (gel images), MS (spectra), ICAT-LC/MS data (spectra), MS-database search results and textual information describing experimental protocols (for example, sample preparation) or results of protein identification. The sample preparation procedures and protocols of 2-DE are documented for each proteome [11, 12]. The 2-DE gels are scanned and analyzed either by our own developed gel image analysis software TopSpot [13] which can be downloaded from the "Download area" of the 2D-PAGE website free of charge or by PDQuest (Bio-Rad Laboratories, Hercules, CA, USA). The scanned images as well as the processed sets of gels (matchsets) containing the differentially regulated protein spots as markers for comparative 2-DE experiments are additionally stored in the laboratory information management system (SQL*LIMS; Applied Biosystems, Foster City, CA, USA) based on Oracle. Mass spectra used for protein identification are obtained using the mass spectrometer Voyager Elite (Perseptive Biosystems, Framingham, MS, USA) and also stored as proprietary binary files or attachments into the laboratory information management system (LIMS). MS-database search results from MASCOT (Matrix Science, London, UK; http://www.matrixscience.com) or MS-Fit (UCSF Mass Spectrometry Facility; http://prospector.ucsf.edu) and their descriptive information can be stored in the LIMS using a parser written in Python. Because only part of the complex data within the LIMS should be published worldwide, a Java-GUI enabling user-controlled data transfer from the LIMS into the corresponding tables of the 2D-PAGE database was developed. The usage of a LIMS assures a given quality of experimental data. Thus, the LIMS represents our central laboratory data repository. A detailed description of LIMS and its adaptation to our specific needs will be provided in the future.

2.2 Software tools

For the software development of our database system, only open source software tools were applied. Specifically, we are using the relational database management system MySQL (http://www.mysql.com), PERL, PHP, HTML, JavaScript, Java, the GD graphic library (http://www.boutell.com/gd/) and the language for data analysis and graphics R (http://www.r-project.org/) [14]. To integrate proteome research with genomic or metabolic data dynamically created HTML hyperlinks are essential. Therefore, the accession methods and URLs to public genomic and metabolic databases must also be known. Further, specific proteomic tools such as MASCOT for MS database searches and PDQuest or TopSpot for gel image analysis are applied.

3 Results and discussion

3.1 Data management, analysis and presentation

A schematic overview of our proteome informatics approach for microbial research is illustrated in Fig. 1. It mainly consists of three parts: data acquisition, data analysis, and web-accessible data presentation. For the acquisition of heterogeneous proteome data, a LIMS is applied. Results of gel image analysis performed by PDQuest can also be transmitted into the LIMS. A Java-GUI enables transfer of selected data from the LIMS into the database system consisting of 2D-PAGE, FUNC_CLASS (functional classification) and ICAT-LC/MS databases, whereby the 2D-PAGE plays the most important role. The database system is interconnected with public genomic, metabolic and other knowledge bases. To ac-

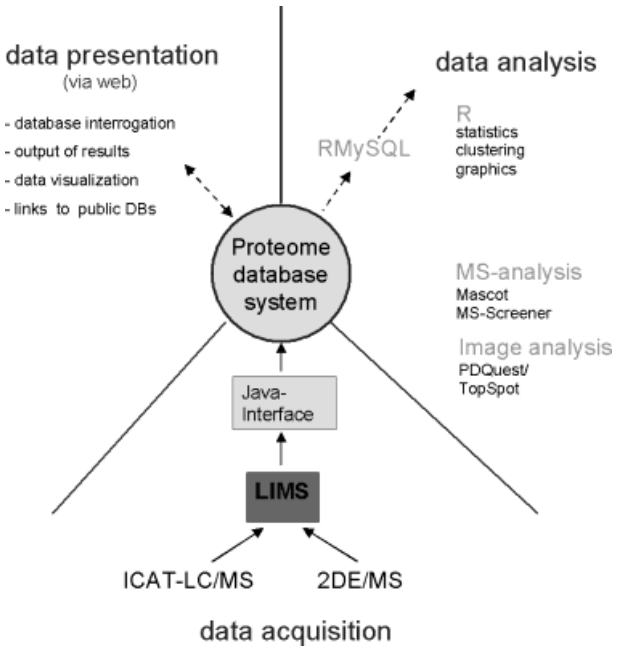

Figure 1. Schematic overview of our proteome informatics approach for microbial research.

complish data analyses, specific proteomic tools such as MASCOT, PDQuest or others are used. Furthermore, continuative explorative data analysis, for example calculation of statistical tests or determination of percent distribution of protein classes, is carried out using R, a language for data analysis and graphics. Web-accessible data presentation includes the possibilities of complex interrogations and the presentation of results (tables, graphics, download of tab-separated files) *via* a web browser enabling access to public databases by hyperlinks.

3.2 2D-PAGE database

The 2D-PAGE proteome database (http://www.mpiib-berlin.mpg.de/2D-PAGE/), first published *via* the internet in 1999 [9], is a multispecies database containing proteome information on diverse strains of microorganisms, tissues and cells. The number of identified spots, the methods of protein identification, the number of antigens and other information on each strain of microorganism are summarized on the statistics page of the database. A part of the database structure is shown in Fig. 2 where the main tables and their relations are schematically depicted. The self-explanatory headings of the tables show how the information on gels, spots, proteins, identification methods, MS, public sequence, literature databases, functional protein classes, and organisms are stored. The fact that one spot may contain several

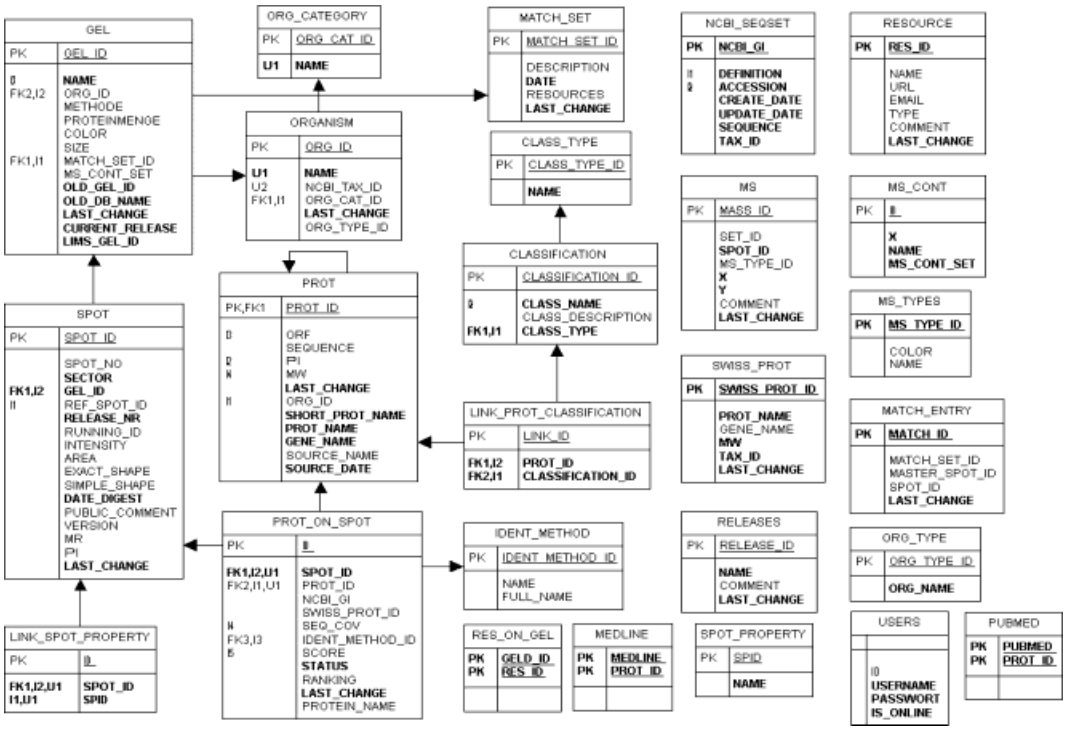

Figure 2. Part of the structure of the 2D-PAGE database.

proteins or one protein may be found in different spots is also taken into consideration in the development of the database schema.

Preparing 2-D gel images for database construction is usually done by the TopSpot gel image processing system. If a spot list was generated by other gel image analysis systems (for example, PDQuest or MELANIE) this data may also be accepted. The structure of a spot list is described in the "Technical Description of Data Submission" at our website. By using TopSpot, for example, the spot detection procedure yields both the spot positions and also describes the spot contours by polygons. These polygon approximations of protein spots serve as sensitive clickable areas within web-accessible gel images and provide the links to the annotated information which is stored in corresponding database tables.

Beside the specific interactions between descriptive information and gel image manipulations our database includes information on mass spectrometric data (peptide mass fingerprints) to comprehend the process of protein identification. Thus, synthetic MS spectra are generated from peak lists of peptide mass fingerprint data. 2D-PAGE is a centrally administrated database. The preparation of gel images and acquisition of descriptive data should be carried out according to data submission rules. The image and data files must be transferred to the database administrator by email or ftp. We are then able to establish a user-specific 2-DE database without further user interactions at the administration level. Thus, we offer a service to build up 2-DE databases for investigators without bioinformatic background. 2D-PAGE also serves as a template for a prototype of a European Proteome Database of Pathogenic Bacteria (EBP). Currently, the database system contains proteomic data on six strains of microorganisms and four eukaryotic proteomes.

3.3 ICAT-LC/MS database

The ICAT-LC/MS technology is a powerful tool to complement classic proteomic methods [15, 16]. For ICAT-LC/MS we have developed a relational database which contains information on mass error, coverage of predicted ions, gene name, protein name, sequence and the ratio of intensity for light/heavy isotope tags for mycobacterial strains BCG *vs.* H37Rv. The light/heavy isotope ratio represents the main information source in the ICAT-LC/MS technology. Because of the huge amount of generated spectra produced by ICAT-LC/MS it is advantageous to store and query data *via* a database that is more comfortable in comparison to an EXCEL approach. Furthermore, the formulation of complex queries can be carried out easily by a web-accessible interface. Additionally, the output of resulting records matching the search criteria can be limited to relevant data field descriptions. Currently, the ICAT-LC/MS database (http://www.mpiib-berlin.mpg.de/bioinfo/ICAT/) contains 1894 entries, of which 183 were manually evaluated.

3.4 FUNC_CLASS database

After sequencing the genomes of *M. tuberculosis* and *H. pylori,* protein coding genes were automatically predicted and functionally categorized or classified. The information on protein coding genes with their functional categories is available for *M. tuberculosis* at (http://www.sanger.ac.uk/Projects/Mtuberculosis/Genelist/) and for *H. pylori* at the comprehensive microbial resource of TIGR (http://www.tigr.org/tigr-scripts/CMR2/gene_attribute_form.dbi). We used these information sources to establish a functional classification database (FUNC_CLASS), consisting of CLASS_Mtb and CLASS_HP26695 for the functional classification of *M. tuberculosis* and *H. pylori,* respectively. The functional classification databases were also linked with our local databases 2D-PAGE and ICAT-LC/MS for mycobacterial strain BCG *vs.* H37RV and KEGG (http://www.genome.ad.jp/kegg/) for metabolic pathways and with the Protein Extraction, Description, and Analysis Tool (PEDANT) (http://pedant.gsf.de/) [17]. Using the capabilities of PEDANT, a comprehensive analysis of complete genomic sequences can be provided. For example, the sequence positions of protein coding genes or the 3-D presentation of protein structures can be shown. The CLASS_Mtb contains information on ORFs, short names of genes, protein names, class_ID, class name (category), M_r and p*I*. In total, 3924 genes are stored.

For *H. pylori* 1563 protein coding genes are available with information on locus, gene length, protein length, GC content, p*I*/M_r values, cellular roles, GenBank/SWISS-PROT ID, and putative identification (protein name).

To retrieve data from these functional classification databases we have written a web form which enables the formulation of multicriteria queries by extending search criteria to refine queries interactively. Additionally, a list of genes given, for instance, in an EXCEL table, can be copied and pasted in a text area. All records matched to this list are displayed. Finally, records can be exported as tab-separated files. The opportunity to ask complex questions and the retrieval of information on genes, proteins and metabolism is permitted by FUNC_CLASS.

3.5 Data analysis and visualization

A major challenge in bioinformatics for proteome research is intelligent data analysis and visualization. We use the R language which comprises a large number of data analysis and graphic tools that are available as packages or functions. Thus far, there is no task for data analysis needed for proteome research that is not solvable in R. Currently, this language has also gained importance for microarray analysis. Additionally, R is characterized by relative simplicity in software function calls for accomplishing complex tasks such as clustering, model fitting, statistical testing, plotting *etc.* Using the RMySQL package we can retrieve data from our MySQL databases and transfer these data into R data frames (matrix). RMySQL also enables the querying of databases using SQL statements. Thus, the retrieval of selected data can be provided. To generate a theoretical 2-DE pattern (virtual 2-DE plot) based on the theoretical p*I*/M_r values of proteins stored in our databases, an R-program is generated by means of a Perl-written CGI-script running on the web server. R dynamically creates the 2-DE plot as an image file (png format) that can be visualized directly *via* a browser. The virtual 2-DE patterns of *M. tuberculosis* or *H. pylori* are additionally overlaid with mouse-sensitive crosses showing those proteins that match the search criteria formulated by SQL statements. Thus, one can easily evaluate where specific protein spots are expected on a 2-DE gel and which proteins are outside of the analysis window of 2-DE.

The full capabilities of our interconnected proteome databases can only be demonstrated online. Here, we show some examples to gain information on microbial proteomes. One can search for information on proteins in 2-DE gel images of selected species by clicking on identified spots (Fig. 3). If peptide mass fingerprinting (PMF) data are available a synthetically generated MS spectrum

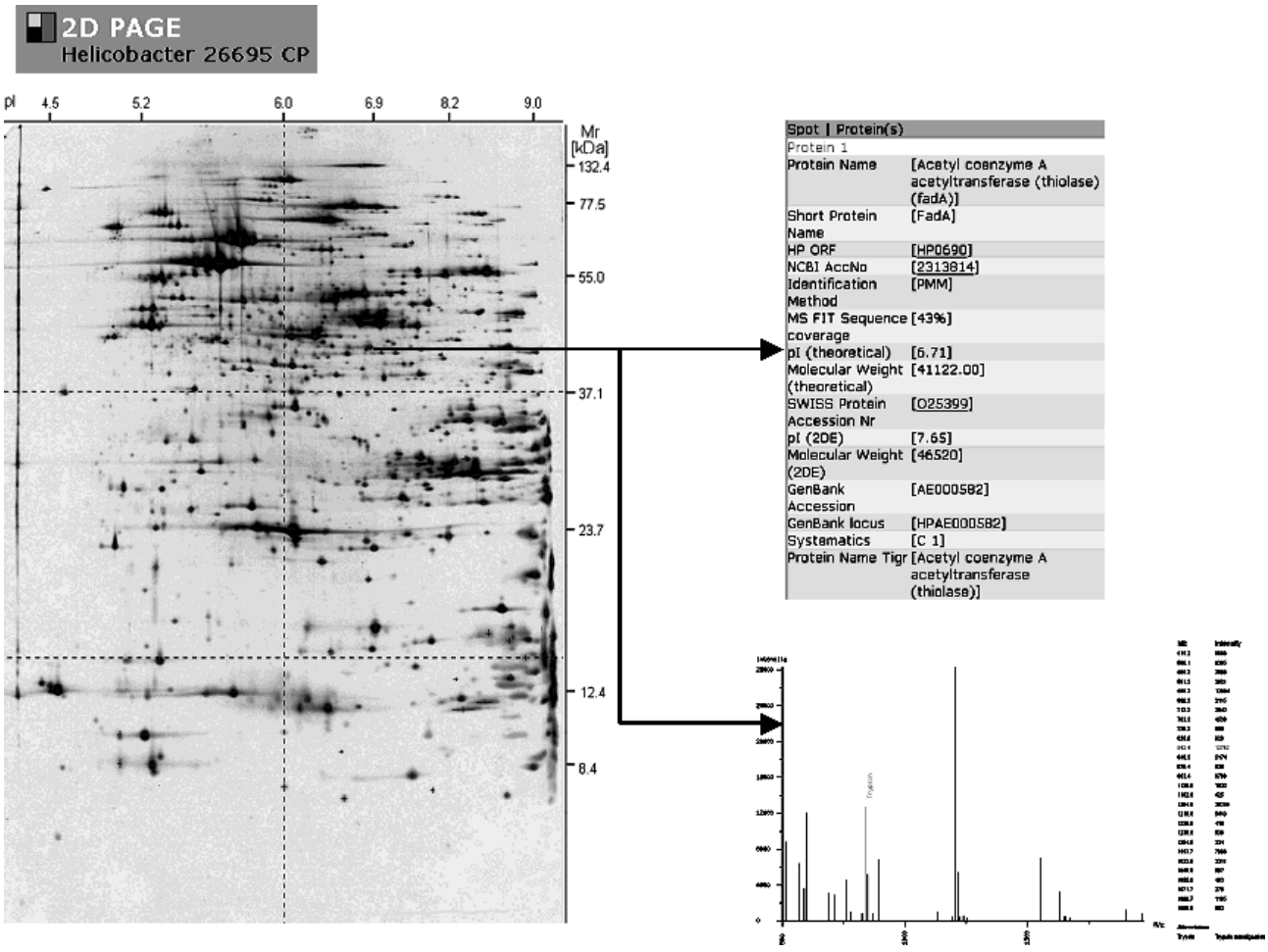

Figure 3. 2-D PAGE: *H. pylori* gel image. Protein spot information (annotation) on acetyltransferase and synthetically generated mass spectrum using peptide mass fingerprinting data.

is displayed for the stored PMF-list. Using functional classification database FUNC_CLASS, proteome relevant queries can be formulated such as: search for all acetyltransferase proteins with a molecular mass > 40 kDa for *M. tuberculosis* (Fig. 4, top). Thus, all entries in the database matching this query are displayed with their interconnections to the ICAT-LC/MS database (Fig. 4, middle) and to the 2D-PAGE database showing where the pro-

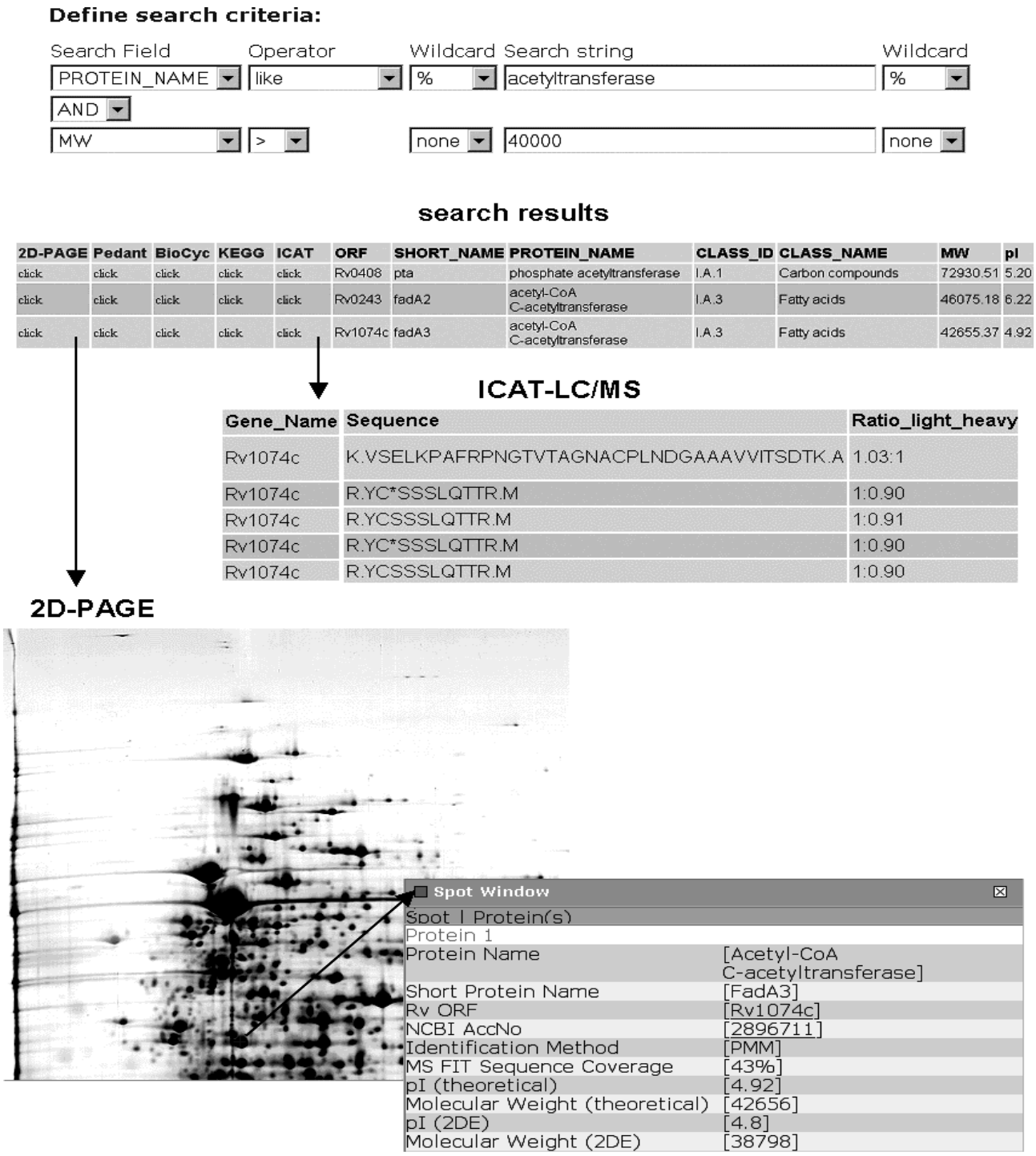

Figure 4. FUNC_CLASS database: search for all acetyltransferase proteins with a M_r > 40 kDa for *M. tuberculosis* and output search results matching the search criteria (top); ICAT-LC/MS databases results (middle); 2D-PAGE database showing the annotation and the location of the protein in the 2-DE gel of *M. tuberculosis* (bottom).

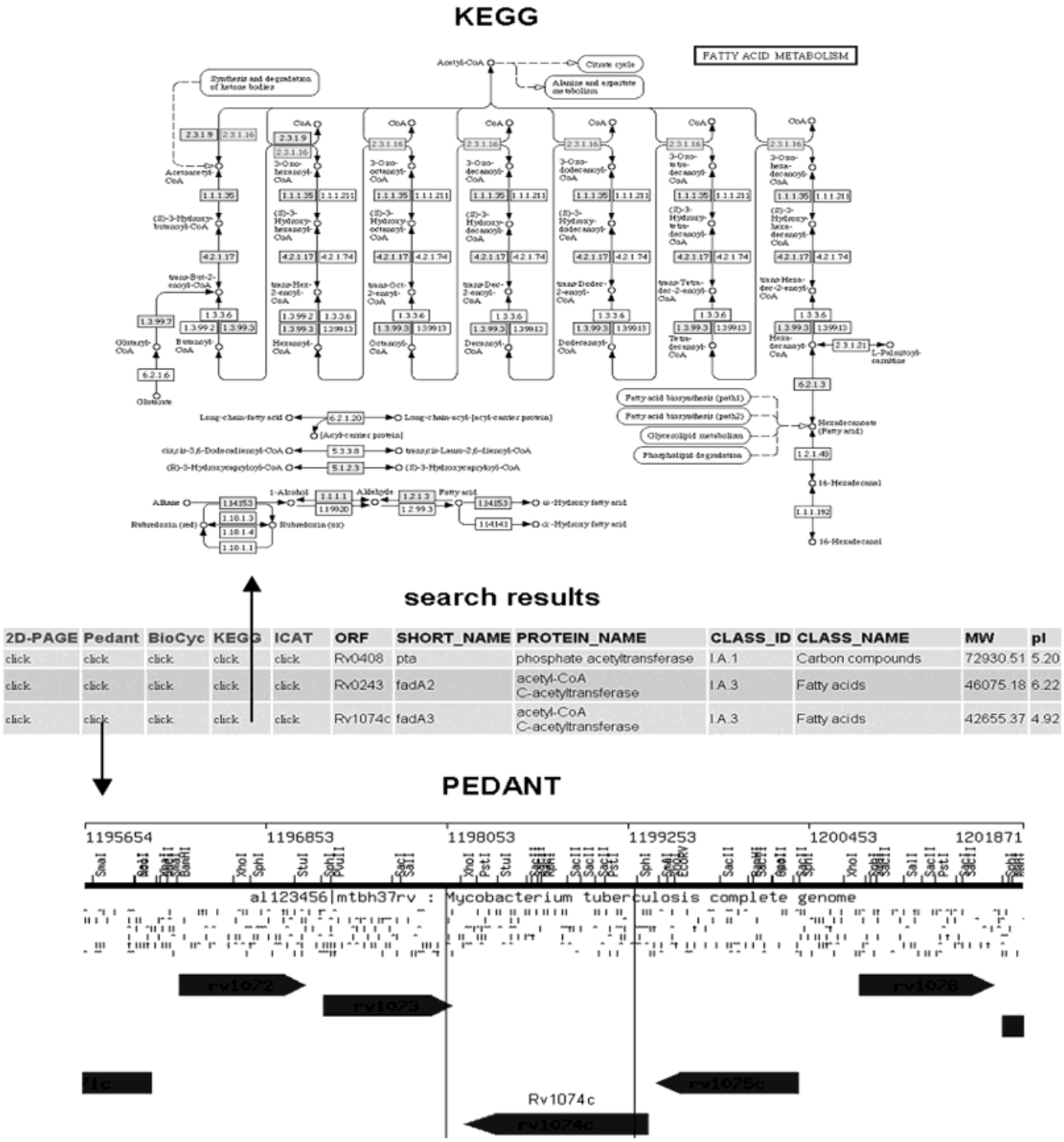

Figure 5. The role of acetyltransferase (ORF: Rv1074c) within the fatty acid oxidation pathway of *M. tuberculosis* generated by KEGG (top). The position of the Rv1074c gene within the genomic sequence provided by the DNA viewer of the PEDANT system (bottom).

teins are located in the 2-DE gel (Fig. 4, bottom). The role of the acetyltransferase within the fatty acid oxidation pathway of *M. tuberculosis*, automatically generated by the KEGG server, is depicted in Fig. 5, top. The position of the Rv1074c gene within the genomic sequence is shown by the PEDANT system (Fig. 5, bottom). The virtual 2-DE pattern (Fig. 6) created by all known proteins of *M. tuberculosis* is overlaid with markers (crosses) showing those protein entries found by the search criteria.

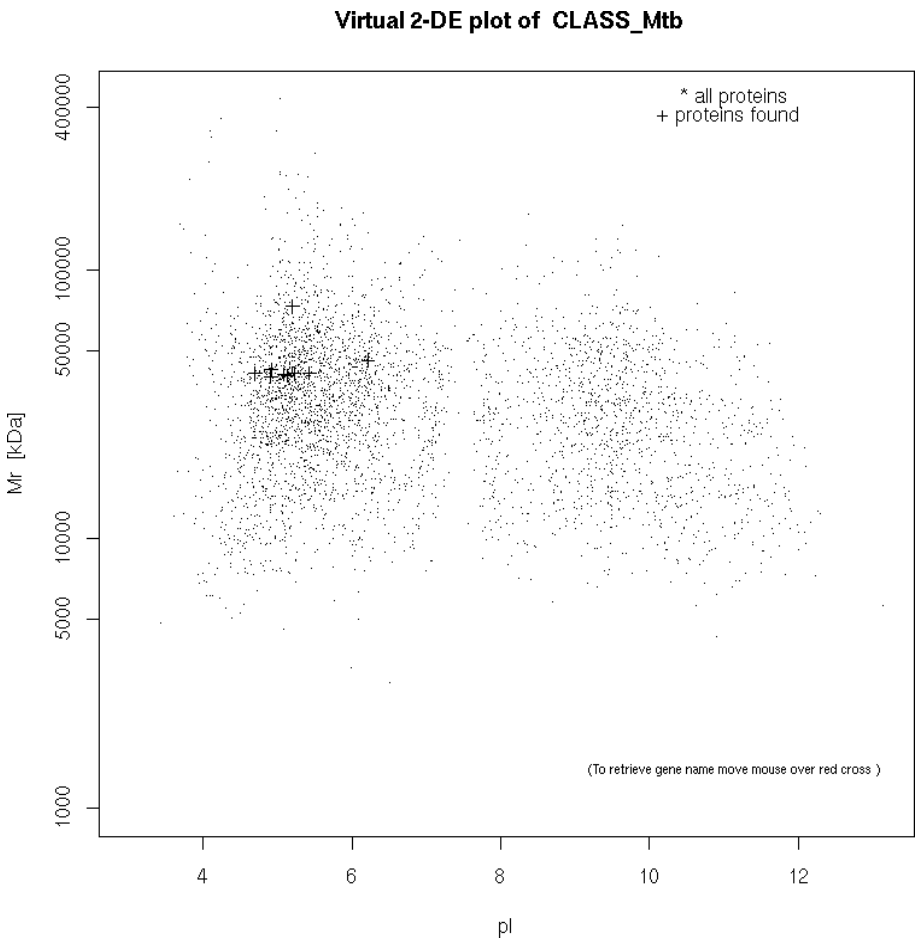

Figure 6. Virtual 2-DE pattern of *M. tuberculosis*: all proteins with theoretical p*I*/M_r values are overlaid with mouse-sensitive markers (crosses) showing those protein entries found by the search criteria. Using the language R all graphics are dynamically created.

4 Concluding remarks

A web-accessible system of proteome databases has been developed that comprises 2-DE/MS, ICAT-LC/MS and a functional classification database for microbial proteins. These local databases are also linked to a variety of external public genomic and metabolic databases. Hence, researchers can gather information on genes, proteins and metabolic pathways. Different access methods, such as clicking on a spot in a 2-D gel or asking complex questions were developed to obtain customized information. A LIMS is used as a central repository for experimental and processed data. A part of these data are transferred into the web-accessible 2D-PAGE database. Applying the language R, data analysis and data visualization tasks may be performed dynamically. The power of R fulfills the requirements of a high-level dynamic data analysis and data visualization tool for proteome data. Furthermore, the proteome database system, especially the 2D-PAGE database, serves as a template for a prototype of a European Proteome Database of Pathogenic Bacterial.

The authors would like to express their thanks to Dr. Ursula Zimny-Arndt and Monika Schmid for 2-DE and MS analysis. This work was partially funded by the German Federal Ministry for Eduction and Research (contract number: 031U107C/031U207C) and by the European Union (EBP-network – contract number: QLRT-1999-31536).

5 References

[1] Patterson, S. D., Aebersold, R. H., *Nat. Genet.* 2003, *33*, 311–323.

[2] Hiscock, D., Upton, C., *Bioinformatics* 2002, *16*, 484–485.

[3] Oh, J. M. C., Hanash, S. M., Teichroew, D., *Electrophoresis* 1999, *20*, 766–774.

[4] Babnigg, G., Giometti, C. S., *Proteomics* 2003, *3*, 584–600.

[5] Cho, S. Y., Park, K.-S., Shim, J. E., Kwon, M.-S. *et al.*, *Proteomics* 2002, *2*, 1104–1113.

[6] Appel, R. D., Bairoch, A., Sanchez, J. C., Vargas, J. R. *et al.*, *Electrophoresis* 1999, *17*, 540–546.

[7] Taylor, C. F., Paton, N. W., Garwood, K. L., Kirby, P. D. *et al.*, *Nat. Biotechnol.* 2003, *21*, 247–254.

[8] Nyman, T., Rosengren, A., Syyrakki, S., Pellinen, T. *et al.*, *Electrophoresis* 2001, *22*, 4375–4382.

[9] Mollenkopf, H. J., Jungblut, P. R., Raupach, B., Mattow, J. *et al.*, *Electrophoresis* 1999, *20*, 2172–2180.

[10] Pleißner, K.-P., Eifert, T., Jungblut, P. R., *Comp. Funct. Genom.* 2002, *3*, 97–100.

[11] Mollenkopf, H. J., Mattow, J., Schaible, U. E., Grode, L. *et al.*, *Methods Enzymol.* 2002, *358*, 242–256.

[12] Jungblut, P. R., Bumann, D., *Methods Enzymol.* 2002, *358*, 307–316.

[13] Prehm, J., Jungblut, P. R., Klose, J., *Electrophoresis* 1987, *8*, 562–572.

[14] Ihaka, R., Gentleman, R., *J. Computational Graphical Statistics* 1996, *5*, 299–314.

[15] Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F. *et al.*, *Nat. Biotechnol.* 1999, *10*, 994–999.

[16] Schmidt, F., Schmid, M., Jungblut, P. R., Mattow, J. *et al.*, *J. Am. Soc. Mass Spectrom.* 2003, *14*, 943–956.

[17] Frishman, D., Albermann, K., Hani, J., Heumann, K. *et al.*, *Bioinformatics* 2001, *17*, 44–57.

