An Open Source Protein Gel Documentation System for Proteome Analyses

Daniel Faller,*,† Thomas Reinheckel,‡ Daniel Wenzler,‡ Sascha Hagemann,‡ Ke Xiao,‡ Josef Honerkamp,† Christoph Peters,‡ Thomas Dandekar,‡,§ and Jens Timmer†

Freiburg Center for Data Analysis and Modeling, Eckerstrasse 1, 79104 Freiburg, Germany, Institut für Molekulare Medizin und Zellforschung, Hugstetterstrasse 55, 79106 Freiburg, Germany, and Lehrstuhl für Bioinformatik, Biozentrum, Am Hubland, 97074 Würzburg, Germany

Received August 11, 2003

Data organization and data mining represents one of the main challenges for modern high throughput technologies in pharmaceutical chemistry and medical chemistry. The presented open source documentation and analysis system provides an integrated solution (tutorial, setup protocol, sources, executables) aimed at substituting the traditionally used lab-book. The data management solution provided incorporates detailed information about the processing of the gels and the experimental conditions used and includes basic data analysis facilities which can be easily extended. The sample database and User-Interface are available free of charge under the GNU license from http://webber.physik.uni-freiburg.de/~fallerd/tutorial.htm.

The term proteomics denotes a systematic and integrated approach of high resolution protein separation and quantification in complex biological samples, which is usually achieved by two-dimensional polyacrylamide gel electrophoresis (2DE), spot detection and comparison by charge-coupled device (CCD) cameras or scanners and, finally, spot identification by mass spectrometry or Edman-based protein sequencing. The vast quantity of data generated by 2DE experiments requires integrated solutions that do not only allow for an easy access to measured protein abundances but also link these values with information about the biological samples and experimental conditions used to generate the data. The aim of the presented data management and analysis software is to facilitate the interpretation of 2DE experiments by providing all the information needed in a standardized and extensible way. Hence, it aims at substituting the traditionally used lab-book and additionally offers basic and extendable proteome data analysis facilities.

The presented data management solution evolved from the need of a typical academic laboratory group to document their 2DE experiments. The two main design goals were as follows: simple usability for laboratory staff and to be able to adapt the chemical information stored to user specified needs without having to change the structure of the database. Hence, the software provided can easily be adapted to fit the needs of different biochemical laboratories. To facilitate exchange and creation of such "electronic lab-books" the developed software is free of charge under the GNU license.

To allow easy portability we used a client server architecture. The client computer just needs a browser and Internet access. The communication with the server then uses an encrypted http connection and is handled by Java servlets. Currently the user interface supports two databases, Oracle 9i (for users preferring a high industrial standard) and the open source Postgres 7.3.2 which is freely available under the Postgres license. Extension to other databases should be straightforward. A flowchart of the program flow, technical details and an introductory tutorial as well as a concrete example are available in the Supporting Information.

The information stored in the database is organized in different experiments. Each experiment consists of two main building blocks, biological samples and 2DE gels, see Figure 1. Several biological samples with certain properties such as the genotype or the number of experimental replicates made are created within one experiment. Then a corresponding number of 2DE gels is added. The properties of the 2DE gels are organized in categories such as first and second dimension. These categories and their properties e.g. pH gradient or voltage can be added and changed interactively by the laboratory staff without having to change the database layout.

To create a new experiment, the name of the operator and a short description is entered. Then, the biological samples used in the experiment are described in detail, see Figure 1 for some example properties. In the next step, the 2D gels used to perform the experiments are entered, and some example properties of the 2D gel are given in Figure 1. In addition the files containing the scanned image of the 2D gel and a file containing the computed spot intensities are uploaded into the database. Afterward, the layout of the experiment, the assignment which 2DE gel corresponds to which biological sample, is defined. Once this information is entered, the analysis of the experiments can be performed, typically with examination, standardization and statistical analysis of the data. To this end, the output of standard image analysis software used to detect and match the spots on the different gels can be uploaded to the database. At the moment two types of analysis are possible, but further tools can be added easily. Matched spots can be selected according to their unique spot number and their detection frequency in different repeats. For example a query such as "All spots which were detected at least 3 times within 5 repeats" is

* Corresponding author e-mail: Daniel.Faller@physik.uni-freiburg.de.
† Freiburg Center for Data Analysis and Modeling.
‡ Institut für Molekulare Medizin und Zellforschung.
§ Lehrstuhl für Bioinformatik.

10.1021/ci034174m CCC: $27.50
Published on Web 11/26/2003

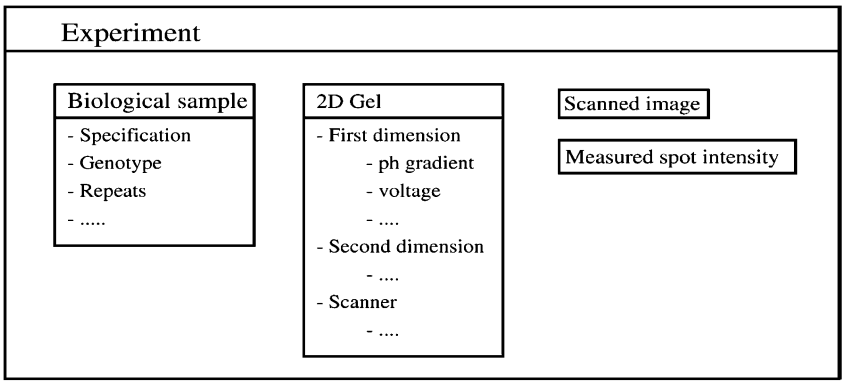

Figure 1. Layout of the database. Each experiments consists of the definitions for the biological sample and the 2DE gel.

possible. In addition, if a set of spot intensities with different experimental conditions has been selected a two sample *t*-test can be performed. This is done by a call to the external statistics package R,[1] and hence it is straightforward to implement other statistical tools into the user interface by just using the already available facilities of R.

There are several, generally commercial, database tools available for proteomic studies which offer a rich detail and variation of storage and database capabilities (e.g.[2,3] to cite two well-known tools). However, the tool presented here allows easy setup of an electronic lab-book: Like a book from a shelf, tutorial, setup protocol and executables allow the user the easy pickup and opening of the book. Furthermore, our software package is highly flexible including storage fields and formats for a number of other biochemical and experimental data, and it is freely available under the GNU license. Moreover, using the source code it can easily be customized to fit the specific needs of the user. In summary, here we present a versatile and academically free "electronic lab-book" that not only allows fast 2DE data retrieval and integration of proteomics data but also integrates (including specific query windows) the storage of any additional experimental data allowing context specific links, complex queries, average comparisons, and, importantly, correlation of proteome and gene-expression data in a standardized expandable database environment. We have to stress that the present version of the electronic lab-book was specifically geared toward 2DE gel data and accompanying experimental setups. Further extensions could include a wider variety of data using the same concept. However, already now the system can easily accommodate experimental data obtained by other gel based experiments such as PCR data (agarose gels), sequencing data (PAGE gels), and DNA fragment size analysis (agarose gels). For this the data fields used for 2D gel data storage are directly used (field for scanned gel image; optional: file for computed band intensities), while additional experimental data (e.g. PCR annealing conditions, DNA sample origin) go into their appropriate fields as provided by the system.

Supporting Information Available: A short tutorial describing the installation and the usage of the presented data management solutions. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) http://www.r-project.org.
(2) Appel, R. D.; Palagi, P. M.; Walther, D.; Vargas, J. R.; Sanchez, J. C.; Ravier, F.; Pasquali, C.; Hochstrasser, D. F. Melanie II: a third-generation software package for analysis of two-dimensional electrophoresis images: I. Features and user interface. *Electrophoresis* **1997**; *18(15)*, 2724−34.
(3) Raman, B.; Cheung, A.; Marten, M. R. Quantitative comparison and evaluation of two commercially available, two-dimensional electrophoresis image analysis software packages, Z3 and Melanie. *Electrophoresis* **2002**, *23(14)*, 2194−202.

CI034174M