APPLICATION NOTE

Genomics & Informatics
Vol. 9(4) 197-199, December 2011

http://dx.doi.org/10.5808/GI.2011.9.4.197
pISSN 1598-866X eISSN 2234-0742

2DSpotDB: A Database for the Annotated Two-dimensional Polyacrylamide Gel Electrophoresis of Pathogen Proteins

Dae-Won Kim, Won Gi Yoo, Myoung-Ro Lee, Yu-Jung Kim, Shin-Hyeong Cho, Won-Ja Lee and Jung-Won Ju*

Division of Malaria and Parasitic Diseases, Korea National Institute of Health, Osong 363-951, Korea

Abstract

The biological interpretation of two-dimensional (2D) gel electrophoresis experiments is a key step toward understanding the functions of biological systems. We here present a web-based integrated database, called 2DSpotDB, for the management of proteome data derived from several pathogens. The 2DSpotDB was established as a part of the management of a pathogen proteome project at the Korea National Institute of Health. The goals of the 2DSpotDB implementation are to store and define important pathogen genes, retrieve information obtained by 2D polyacrylamide gel electrophoresis and mass spectrometry, and create an integrated system to provide pathogen proteome information for biological scientists. This database currently contains 14 gels and information on 387 protein spots, among which 329 proteins were identified and annotated.

Keywords: database, 2D electrophoresis, LC-MS/MS, pathogen proteomics, functional annotation, data mining

Availability: The 2DSpotDB can be accessed at http://pathod.cdc.go.kr/2dspotdb.

Introduction

The post-genomic era has seen an increasing amount of effort on massive systematic surveys of various proteomes (Seger *et al*., 2009). Researchers in the biomedical sciences have increasingly been applying high-throughput proteome analysis techniques to study the direct products of gene expression, molecular interactions and the cellular environment since the introduction of two-dimensional (2D) gel electrophoresis in 1975 (Klose, 1975; O'Farrell, 1975; Gerner *et al*., 2000; Jiang *et al*., 2004; Wang *et al*., 2010). Proteome analysis by 2D gel image is one of the most efficient methods for analyzing the proteomes of cells, tissues or whole model organisms in proteomics studies. Simultaneously, many advances have been made in the field of liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) quantitative proteomics for large-scale complex samples using the observed strong correlation between the counts of protein-specific spectra or peptide quantitation obtained by mass spectrometry and the absolute and relative abundance of the proteins (Liu *et al*., 2004; Ishihama *et al*., 2005). With the rapid growth in proteomic raw data in pathogen laboratories, a current challenge is the meaningful management of the results of such large-scale studies, including spot identification and annotation information for numerous 2D gel images (Berth *et al*., 2007). Many databases, such as ProDB (Wilke *et al*., 2003), UAB (Hill and Kim, 2003), DynaProt 2D (Drews and Gorg, 2005) and 2DB (Allmer *et al*., 2008), have been developed to compete with SWISS-2DPAGE (Appel *et al*., 1996) and its protein identification in gel images in the past.

Here, we report the development of the 2DSpotDB, which provides a centralized database to serve the growing needs of the pathogen proteomic research community. The 2DSpotDB was developed as a part of the management of a proteome project for the important pathogens *Bacillus anthracis, Clonorchis sinensis, Neisseria gonorrhoeae* and *Streptococcus pneumoniae*. This database includes 2D gel images, spot identification information, annotation information and various statistical tables. The data contents have generally been deposited to share major sources of proteomic pathogen data from researchers in various laboratories. This information can enhance analyses related to the functions of pathogen-specific biological systems involving antigenic or pathogenic proteins with experimental evidence and shed light on the dynamic variations in the pathogen proteome during the complex physiological changes linked with different experimental stages and conditions.

*Corresponding author: E-mail junomics@gmail.com
Tel +82-43-719-8524, Fax +82-43-719-8559
Received 12 November 2011, Revised 30 November 2011, Accepted 5 December 2011

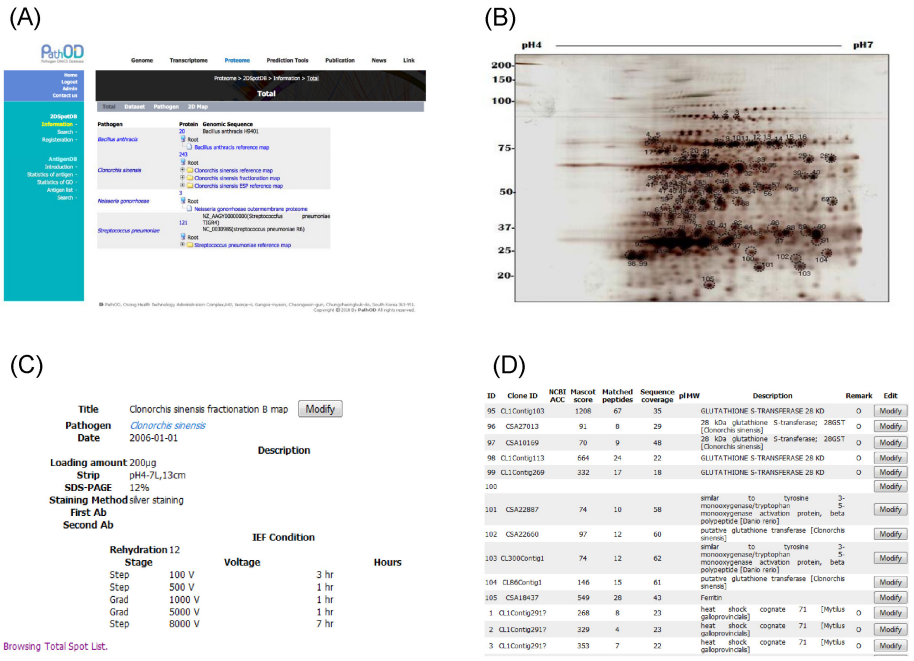

Fig. 1. The overview of 2-DSpotDB database. (A) User interface, (B) 2D spot image, (C) Experimental method, (D) Spot information.

Features and Results

Data acquisition and community involvement

Updated raw data and information for the 2DSpotDB is periodically provided from the pathogen proteomic research community at the Korea National Institute of Health (KNIH). Because the 2DSpotDB has no full-time automatic updater, expert users individually access and update this database to provide the pathogen community with integrated, comprehensive 2D proteomic spot information related to the experimental conditions.

Methods

The 2DSpotDB website is a relational database system (Oracle 11g) running on Dell PowerEdge 2580 servers using the Red Hat Linux operating system (Red Hat Enterprise Linux 5.5 platform). The 2DSpotDB is deployed in a '2-tier' architecture, divided neatly into the presentation tier (the user interface) and the data storage and access tier (the database). Users can submit experimental and analyzed data into the database or search queries through the web interface. The queries are then processed using the Apache web server, JSP (Java Server Pages), JavaServelet technology and cascading style sheet (CSS) properties. The results of each query are presented to the user in the web browser.

Implementations

Here, we report the archiving of proteomic 2D gel images and spot identification information for four major pathogenic organisms (*B. anthracis, C. sinensis, N. gonorrhoeae and S. pneumoniae*), which currently consists of information on 387 protein spots, of which 329 protein spots were identified and annotated from 14 gels.

The Web interface is divided into three major functional groups: (i) Proteome information (total statistics, 2D gel information and protein spot information), (ii) Integrated searching based on Clone ID, accession ID, or description, and (iii) Registration (web interfaces for registered users for manipulating and depositing their 2D experimental methods and spot information) (Fig. 1A). The 2DSpotDB was established to facilitate and encourage proteomics scientists to publish gel-based proteomics data. The 2D images and protein spot information are annotated, uploaded and curated by authorized users (Fig. 1B). The minimum information required spans the gel image(s), a spot identification list(s) with the corresponding spot identifiers, the X/Y coordinates on the gel and the protein identifiers and any relevant information on protocols and publications, if available.

For each specific 2D gel image and identified protein, the 2DSpotDB database contains the following information: (i) Detailed experimental descriptions (Loading amounts, Strip, SDS-PAGE, Staining Method and IEF Conditions) (Fig. 1C), (ii) General spot information (Clone ID, NCBI ACC, Mascot score, Matched peptides, Sequ-

ence coverage, pI, MW, Description and Remark) (Fig. 1D). All information can be manually curated.

Currently, proteomic data is a potentially rich resource, but proteome peptide identification has been considerably underexploited. For the identification steps, using bioinformatic software in many cases, individual protein spots containing protein identifiers and exact descriptions may be obtained from any reference public database, but it has frequently been observed that these protein spots have no protein ID due to unidentified genome and expressed sequence tag data. Thus, we permit the protein accession numbers of the identified proteins to be curated and updated by individual researchers upon further study.

Discussion

In our study, we aim to find out antigens which can be used to produce new candidate vaccines against diseases competing with SWISS-2DPAGE (Appel *et al*., 1996). Thus, we will present a database application which can be employed in many antigen-experimental contexts such as antigenicity and epitopes information. The 2DSpotDB database is expected to expand further over the coming years, incorporating new species including *Plasmodium vivax* and many viruses. Advanced features and data mining tools will continue to be introduced and improved, e.g., the ability to make biological classifications using gene ontology, a comparison system for diverse 2D images according to different experimental methods and the weighting of annotation results. We also anticipate that pathogenic proteomic data will continue to be deposited at an ever-increasing rate.

Acknowledgements

The authors wish to acknowledge the contributions of numerous members of the pathogen research community in the management of the pathogen proteome project. This study was supported by grant 2006-N54002-00 from the Korea National Institute of Health.

References

Allmer, J., Kuhlgert, S., and Hippler, M. (2008). 2DB: a Proteomics database for storage, analysis, presentation, and retrieval of information from mass spectrometric experiments. *BMC Bioinformatics* 9, 302.

Appel, R.D., Bairoch, A., Sanchez, J.C., Vargas, J.R., Golaz, O., Pasquali, C., and Hochstrasser, D.F. (1996). Federated two-dimensional electrophoresis database: a simple means of publishing two-dimensional electrophoresis data. *Electrophoresis* 17, 540-546.

Berth, M., Moser, F.M., Kolbe, M., and Bernhardt, J. (2007). The state of the art in the analysis of two-dimensional gel electrophoresis images. *Appl. Microbiol. Biotechnol.* 76, 1223-1243.

Drews, O. and Gorg, A. (2005). DynaProt 2D: an advanced proteomic database for dynamic online access to proteomes and two-dimensional electrophoresis gels. *Nucleic Acids Res.* 33, D583-587.

Gerner, C., Frohwein, U., Gotzmann, J., Bayer, E., Gelbmann, D., Bursch, W., and Schulte-Hermann, R. (2000). The Fas-induced apoptosis analyzed by high throughput proteome analysis. *J. Biol. Chem.* 275, 39018-39026.

Hill, A. and Kim, H. (2003). The UAB Proteomics Database. *Bioinformatics* 19, 2149-2151.

Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., and Mann, M. (2005). Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell Proteomics* 4, 1265-1272.

Jiang, X.S., Zhou, H., Zhang, L., Sheng, Q.H., Li, S.J., Li, L., Hao, P., Li, Y.X., Xia, Q.C., Wu, J.R., and Zeng, R. (2004). A high-throughput approach for subcellular proteome: identification of rat liver proteins using subcellular fractionation coupled with two-dimensional liquid chromatography tandem mass spectrometry and bioinformatic analysis. *Mol. Cell Proteomics* 3, 441-455.

Klose, J. (1975). Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik* 26, 231-243.

Liu, H., Sadygov, R.G., and Yates, J.R., III (2004). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* 76, 4193-4201.

O'Farrell, P.H. (1975). High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* 250, 4007-4021.

Seger, C., Tentschert, K., Stoggl, W., Griesmacher, A., and Ramsay, S.L. (2009). A rapid HPLC-MS/MS method for the simultaneous quantification of cyclosporine A, tacrolimus, sirolimus and everolimus in human blood samples. *Nat. Protoc.* 4, 526-534.

Wang, F., Chen, R., Zhu, J., Sun, D., Song, C., Wu, Y., Ye, M., Wang, L., and Zou, H. (2010). A fully automated system with online sample loading, isotope dimethyl labeling and multidimensional separation for high-throughput quantitative proteome analysis. *Anal. Chem.* 82, 3007-3015.

Wilke, A., Ruckert, C., Bartels, D., Dondrup, M., Goesmann, A., Huser, A.T., Kespohl, S., Linke, B., Mahne, M., McHardy, A., Puhler, A., and Meyer, F. (2003). Bioinformatics support for high-throughput proteomics. *J. Biotechnol.* 106, 147-156.