

[The 2nd Annual Meeting of JHUPPO]

## Collaborative proteomics framework with XML databases and an integrated XML viewer for 2DPAGE

Hiraku Morisawa<sup>1</sup>, Hisashi Hisatomi<sup>1,2</sup>, Mikako Hirota<sup>1</sup> and Tosifusa Toda<sup>1</sup>

### SUMMARY

We developed a collaborative proteomics framework with XML-based 2DPAGE database and XML viewer optimized for the proteome database. Prior to the development of the framework, we optimized the XML format of our TMIG-XML-2DPAGE database for efficient data sharing in a collaborative research community. The data format is comprised with 2 layers of 2-D gel map and spot protein information. The function of clickable imagemap for representing spot protein information is available both in the proteome data management system with PHP and in the XML viewer with Object Pascal. Individual spots on an experimental 2-D gel image are easily matched to corresponding spots on a standard 2-D gel map in the TMIG-XML-2DPAGE database using the function of the XML viewer.

Key words: 2DPAGE, proteome database, XML viewer, informatics.

### INTRODUCTION

Proteomics<sup>1,2)</sup> is the most advanced methodology for comprehensive analysis of proteins in a complex protein mixture. Proteomic research generates large amounts of experimental data and related information from a small amount of sample in a short period of time. Therefore, it is getting much more important for us to construct cumulative proteome database and to develop an efficient information management system. 2DPAGE databases that consist of 2-D gel maps and spot protein information are expected to be useful knowledge resources for finding key proteins of cellular function and novel biomarkers of diseases. Actually many kinds of 2DPAGE databases<sup>3-6)</sup> have been already constructed on web servers and opened in the Internet to be accessed using a web browser. Those databases offer convenient Web interfaces such as clickable imagemap and key-word search. However, most of them are written in HTML or in a plain text format to be processed by Perl-based software. The TMIG (Tokyo Metropolitan Institute of Gerontology)-XML-2DPAGE database system<sup>5)</sup> has been improved to achieve efficient proteome data sharing in col-

laborative research community. In this paper, we report the concept and the architecture of the improved system of the collaborative proteomics framework.

### MATERIALS AND METHODS

The new TMIG-XML-2DPAGE database system was constructed on the RedHat Linux 9 operating system that was installed to an IBM-compatible computer with Pentium 4 Processor. The RedHat Linux 9 offers an excellent environment to develop open-source applications. The advanced database system is based on Apache, PHP and PostgreSQL. Apache is the most popular web server software. Upon receiving a request for a file from a remote PC, Apache returns its contents to the Web browser. PHP is a hypertext preprocessor which serves a powerful tool for adding scripts inside of HTML documents. PostgreSQL is a well featured relational database software. Functions of PHP and PostgreSQL can be tightly integrated in the open-source applications. The combination of them allows the Apache server to accomplish the user-friendly proteome database.

We optimized the XML format of our TMIG-XML-

<sup>1</sup> Proteomics Collaboration Research Group, Tokyo Metropolitan Institute of Gerontology.

<sup>2</sup> Analytical Center for Medical Science, SRL Inc.

Correspondence address: Hiraku Morisawa; Proteomics Collaboration Research Group, Tokyo Metropolitan Institute of Gerontology, 35-2 Sakaecho, Itabashi-ku, Tokyo 173-0015, Japan.

Abbreviations: TMIG, Tokyo Metropolitan Institute Gerontology; XML, Extensible markup language; PHP, Hypertext Preprocessor; RSS, Rich site summary.

JHUPPO: Japan Human Proteome Organization

(Received August 6, 2004, Accepted December 28, 2004, Published March 15, 2005)

2DPAGE database to achieve efficient data sharing in a collaborative proteome research community. The format comprised of 2-D gel map and spot annotating information as shown in Fig. 1. The PHP allows the PostgreSQL to output web pages prepared in the optimized XML format. It is actually able to transform experimental data described in a plane text format into the XML format.

TMIG-XML-2DPAGE was intended to be equipped with two ways of data retrieving functions. One is the function of clickable imagemap, which allows us to download the corresponding protein information by clicking a mouse button on a protein spot. Another function is the keyword search, which helps us to access to the relational information by sending a query of keyword in the categories of spot ID (ssp), protein name, Swiss-Prot accession number and EMBL accession number to our web server.

### RESULTS

We ultimately established the collaborative proteomics framework with XML-based TMIG-2DPAGE databases as shown in Fig. 2, and developed the integrated XML viewer. Members of a collaborative research community have been able to access to the database server and download all of the information in the server by using both the PostgreSQL application and the XML viewer. Public users are allowed to access to the database for downloading the limited contents using a web browser such as Internet Explorer or Netscape Navigator. The source program of the XML viewer was coded with Object Pascal, Delphi Studio Ver.7.0

enterprise edition in Windows XP. The research community member can use the XML viewer to open the JPEG-formatted 2-D gel maps and the XML-formatted data files. The function of clickable imagemap is also available in the XML viewer. Each protein data, which is linked to the corresponding spot on the 2-D gel map, is displayed in a sub-window by the function of clickable imagemap (Fig. 3). Individual protein spots on a user's 2-D gel image are easily matched to corresponding protein spots on the standard 2-D gel map of our TMIG-XML-2DPAGE proteome database

### Data structure : The TMIG-XML format

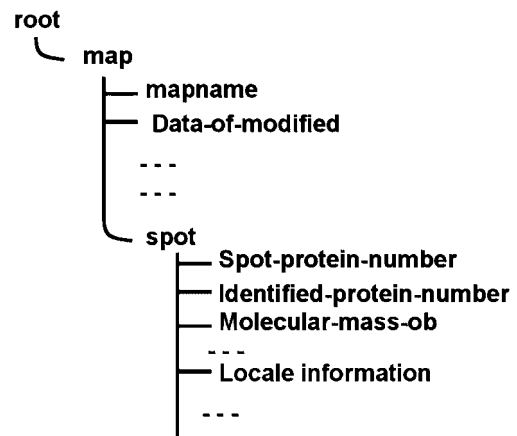


Fig. 1. Data structures of the TMIG-XML format for 2DPAGE. The format comprised of 2-D gel map and spot annotating information.

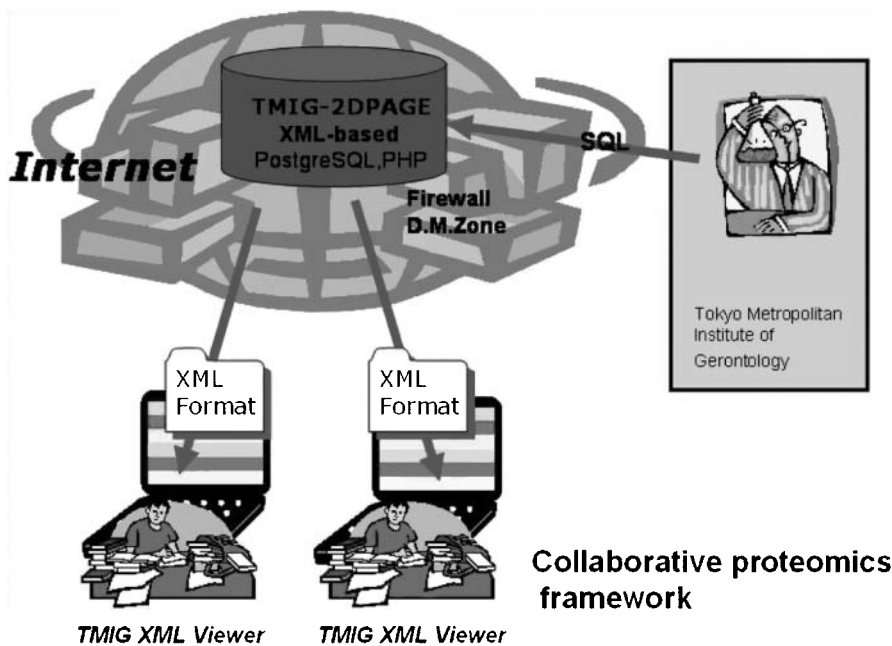


Fig. 2. Collaborative proteomics framework with an XML-based application.

The combination of the XML viewer and the optimized TMIG-XML-2DPAGE database system allows restricted users of the database to access accumulated information about identified protein spots in the database by easy matching of query spots on their own 2-D gel images to the database 2-D gel maps.

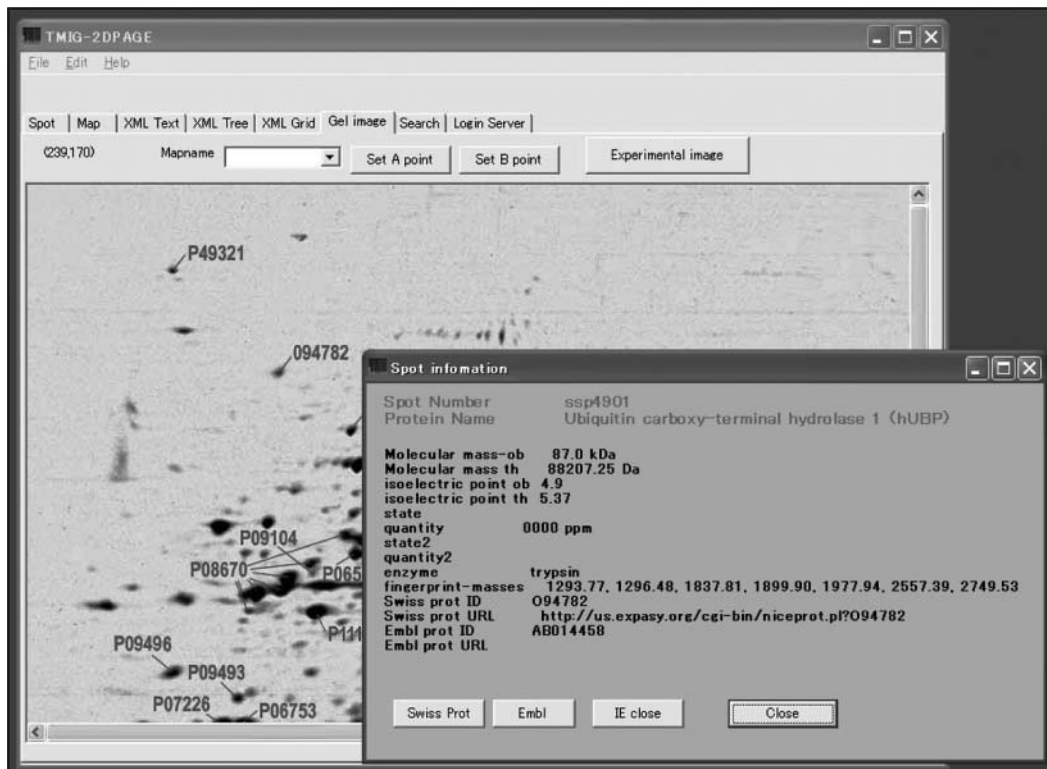


Fig. 3. Spot information of 2DPAGE gel in the viewer.

Each protein data is displayed in a sub-window by the function of clickable imagemap.

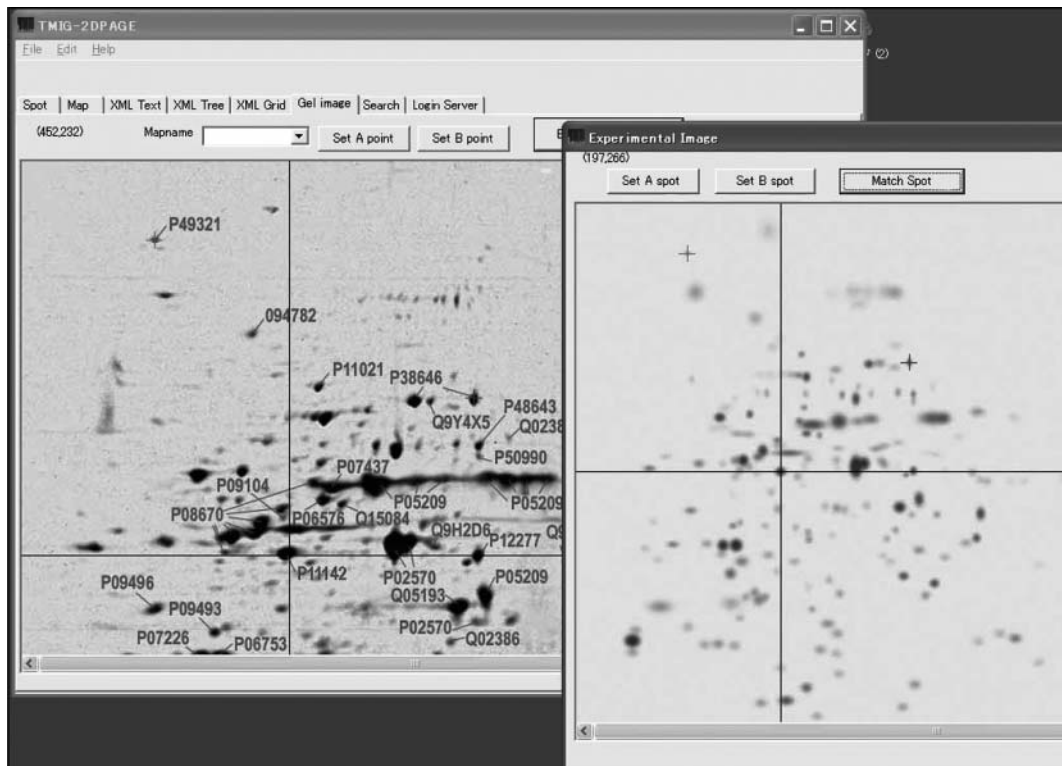


Fig. 4. Spot matting in the viewer.

Individual spots on a user's 2-D gel image are easily matched to corresponding spots on the standard 2-D gel map of our TMIG-XML-2DPAGE proteome database.

using the function of the XML viewer (Fig. 4). We can also look out the spot in the 2-D gel image by using the function of keyword search in the viewer. Furthermore, the downloaded files in the XML format can be used to supplement new data or correct the wrong data in the XML viewer. The XML viewer will be a convenient tool for database management in a closed research community. The combination of the XML viewer and the optimized TMIG-XML-2DPAGE database system allows restricted users of the database to access accumulated information about identified protein spots in the database by easy matching of query spots on their own 2-D gel images to the database 2-D gel maps. Thus the TMIG-XML-2DPAGE database system has been improved in performance, quality and expandability by integration of PostgreSQL and PHP.

### DISCUSSION

From the beginning of the Internet, many kinds of protocols have been introduced to achieve world-wide information sharing. In recent years, web-based database and Peer to Peer (PtoP) applications are getting general standards. However, PtoP applications still have some problems in administration of the data exchange. Most of the proteome research groups are wishing to have their own database server system to share their own proteome data. Actually in our institute, establishment of the collaborative proteomics framework has been required for sharing data in the members of collaborative research projects.

The comprehensive 2DPAGE database was first reported by Celis JE et al.<sup>7)</sup>, however, they did not open the database in their web server at that time. The SWISS-2DPAGE<sup>8, 9)</sup> was the first database which was opened in their ExPASy Server for free access of public users in the Internet. Recently, Mostaguir K et al.<sup>10)</sup> introduced a new concept in their 2DPAGE database which was created and maintained using PostgreSQL. Their 2DPAGE database is superior to the previous databases in terms of information management system. However, it has still been insufficient for the requirements in our collaborative proteomics framework. Thus we applied further improvements to the PostgreSQL-based 2DPAGE system in our own database. Our collaborative proteomics framework provides a new environment for each client PC of research project members to share the 2DPAGE data. The potential of the XML-based application is high<sup>11)</sup>, however, the XML still has an insufficiency in the ability for searching the contents. RSS format, that is one of XML formats, provides a further progressive indexing function for collaborative network applications. We examined the same function as the RSS's to create various XML-based collaborative applications for our proteomics research projects. Our XML database not only replies for our request on the web but also provides us the other usage of proteome data in combination with our XML viewer. The development of the XML-based proteomics framework is

sure to promote our research projects in proteomics. The demonstrated version of the improved TMIG-XML-2DPAGE database system will be opened at [http://proteomeback.tmig.or.jp/2D/index\\_alt.html](http://proteomeback.tmig.or.jp/2D/index_alt.html) for everyone who is interested in our concepts of protein information management.

### REFERENCES

- 1) Wasinger VC, Cordwell SJ, Cerpa-Poljak A, Yan JX, Gooley AA, Wilkins MR, Duncan MW, Harris R, Williams KL, Humphery-Smith I. Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium*. *Electrophoresis* 1995;16:1090–1094.
- 2) Wilkins MR, Sanchez JC, Gooley AA, Appel RD, Humphery-Smith I, Hochstrasser DF, Williams KL. Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnol Genet Eng Rev* 1996;13:19–50.
- 3) Appel RD, Sanchez JC, Bairoch A, Golaz O, Ravier F, Pasquali C, Hughes GJ, Hochstrasser DF. The SWISS-2DPAGE database of two-dimensional polyacrylamide gel electrophoresis, its status in 1995. *Nucleic Acids Res* 1996;24:180–181.
- 4) Pleissner KP, Sander S, Oswald H, Regitz-Zagrosek V, Fleck E. The construction of the World Wide Web-accessible myocardial two-dimensional gel electrophoresis protein database "HEART-2DPAGE": a practical approach. *Electrophoresis* 1996;17:1386–1392.
- 5) Toda T, Kaji K, Kimura N. TMIG-2DPAGE: a new concept of two-dimensional gel protein database for research on aging. *Electrophoresis* 1998;19:344–348.
- 6) Hill A, Kim H. The UAB Proteomics Database. *Bioinformatics* 2003;19:2149–2151.
- 7) Celis JE, Ratz GP, Madsen P, Gesser B, Lauridsen JB, Hansen KP, Kwee S, Rasmussen HH, Nielsen HV, Cruger D, et al. Computerized, comprehensive databases of cellular and secreted proteins from normal human embryonic lung MRC-5 fibroblasts: identification of transformation and/or proliferation sensitive proteins. *Electrophoresis* 1989;102:76–115.
- 8) Appel RD, Sanchez JC, Bairoch A, Golaz O, Miu M, Vargas JR, Hochstrasser DF. SWISS-2DPAGE: a database of two-dimensional gel electrophoresis images. *Electrophoresis* 1993;14:1232–1238.
- 9) O'Donovan C, Martin MJ, Gattiker A, Gasteiger E, Bairoch A, Apweiler R. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief Bioinform* 2002;3:275–284.
- 10) Mostaguir K, Hoogland C, Binz PA, Appel RD. The Make 2D-DB II package: conversion of federated two-dimensional gel electrophoresis databases into a relational format and interconnection of distributed databases. *Proteomics* 2003;3:1441–1444.
- 11) Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, et al. The HUPO PSI's Molecular Interaction format—a community standard for the representation of protein interaction data. *Nat Biotechnol* 2004;22:177–183.