7

2D PAGE Databases for Proteins in Human Body Fluids

Christine Hoogland, Khaled Mostaguir, Jean-Charles Sanchez, Denis F. Hochstrasser, and Ron D. Appel

Summary

With the development of the Internet, a growing number of proteomics databases have become available. The web is a powerful tool for data integration because it links the components constituting these databases, in general gel images and protein information, while offering rapid means to navigate from one database to another. Unfortunately, with only 15 maps available as electronic resources, the human body fluids do not really benefit from this development. This chapter summarizes the state of the art of proteomics databases, with an emphasis on human body fluids. Insights into one of these databases, SWISS-2DPAGE, available for more than 10 yr now, are given to show current functionalities and usage examples. Some general thoughts are also given on how to improve sharing and publication of proteomics data through electronic media.

Key Words: Proteomics database; web services; human body fluids.

1. Introduction

In the postgenomic era, there has been increasing interest in human body fluids as outstanding sources of circulating biological markers. In particular, the identification and characterization of proteins in human body fluids and the interpretation of their relationships with diseases have promoted proteomics to the rank of the technology of choice for biomedical applications. Many laboratories have identified proteins on 2D electrophoresis (2-DE) maps of human body fluids, either to build a reference gel or as a source of potential biological disease markers. The present book is certainly a good proof (*see* the 16 chapters of Part II, each dedicated to a specific body fluid). Another example is given by a simple search in the PubMed resource ***(1)***. With appropriate key words, around 50 pertinent articles published during the last 2 yr can be retrieved in which

From: *Proteomics of Human Body Fluids: Principles, Methods, and Applications*
Edited by: V. Thongboonkerd © Humana Press Inc., Totowa, NJ

authors characterize proteomes of various human body fluids, not only to catalog proteins but also to compare their expression levels in different diseases. Unfortunately, very few of these data are available as electronic resources. In 2005, up to 50 2-DE databases were freely accessible on the Internet, for a total of nearly 300 annotated image maps (*see* http://www.expasy.org/ch2d/2d-index.html for an updated list ***[2]***). However, less than 15 annotated image maps concern human body fluid samples **(Table 1)**, with the most analyzed being plasma, urine and cerebrospinal fluid.

2. 2-DE Databases as Electronic Resources

The elements constituting a 2-DE database, i.e., gel images and protein information, can be conveniently integrated into a web database. Thanks to the active hypertext links, the user has a powerful tool for data integration, in addition to the opportunity to navigate from one database to another ***(3)***. Images are scanned representations of the gels provided with annotations. These represent general information about the identified protein (such as protein name, gene name, taxonomy, references) and experimental data, including the physical localization on the gel (i.e., isoelectric point and molecular weight) and the protein identification method. In addition, cross-references to other biological databases might be offered in some cases. These databases can be accessed in several ways, including both textual and graphical searches. SWISS-2DPAGE was the first federated 2-DE database available as an electronic resource ***(4)***.

2.1. A Detailed Example: The SWISS-2DPAGE Database

The SWISS-2DPAGE database started in 1993 and is still maintained as a collaborative work between the Swiss Institute of Bioinformatics and the Biomedical Proteomics Research Group of Geneva University. The first releases contained 10 human samples, i.e., plasma and cerebrospinal fluid (the others being human cells, tissues, or culture cell lines) ***(4)***. At that time, the plasma map totalized 20% of the total accesses to the SWISS-2DPAGE database through the ExPASy server ***(5)***. Ten years later, SWISS-2DPAGE contains 36 reference maps from various origins (seven species) ***(6)***. Nonetheless, the plasma map is still one of the most accessed, with approximately 30% of the total hits. The major feature of the SWISS-2DPAGE database, still with no equivalent today, is the high level of annotations accessible either in the given protein information (mapping methods, experimental data, and so on) or as supplementary materials (protocols, release notes with validation criteria and statistical resume, and so on).

2.1.1. Annotation

SWISS-2DPAGE provides many annotations at various levels. The first level consists of general information about the protein, its name, function, taxonomy

Table 1
Electronic 2-DE Databases Containing Data from Human Body Fluids in 2005

Database name	Body fluid type[a]	Website URL	Ref. no.
BALF2D	Bronchoalveolar lavage fluid (−)	http://www.umh.ac.be/~biochim/BALF2D.html	***15***
HUPO Plasma Proteome Project	Plasma (+)	http://www.hupo.org/ http://www.bioinformatics.med.umich.edu/hupo/ppp/	***16***
Inner Ear Protein Database	Inner ear fluid (cochlear perilymph) (+)	http://oto.wustl.edu/thc/	***17***
ISPA gel gallery	Milk fat globule (+)	http://www.csaapz.to.cnr.it/proteoma/2DE/	***18***
LeeLab CSF	Cerebrospinal fluid (+/−)	http://www.leelab.org/csfmap/	***19***
LECB gel gallery	Serum (−), urine[b]	http://www.lecb.ncifcrf.gov/2DgelDataSets/	***20***
Proteomics Danish center	Urine (−)[b]	http://proteomics.cancer.dk/	***21***
SIENA-2DPAGE	Amniotic fluid (+)	http://www.bio-mol.unisi.it/2d/2d.html	***22***
SWISS-2DPAGE	Plasma (+), cerebrospinal fluid (+)	http://www.expasy.org/ch2d/	***6***
UAB database	Urine (−)[b]	http://www.uab.edu/proteinmenu	***23***
UCHSC gel gallery	Seminal fluid (+)[b]	http://proteomics.uchsc.edu/2Dexample/	***24***

[a](+), control; (−), disease.
[b]Not available at the time of writing.

origin, quaternary structure, posttranslational modifications, and so on. The second level relates to gel electrophoresis and spot identification. For each protein that has been experimentally identified in one or many gels, the information gathered includes its physical location on the gel (isoelectric point [p*I*]/molecular weight [Mw]), its unique spot serial number, and the identification method (i.e., gel matching, amino acid composition, immunoblotting, microsequencing, peptide mass fingerprinting, tandem mass spectrometry, or a combination of these) together with the experimental data itself (currently the amino acid composition in percent, the peptide masses that allowed the protein identification, and the peptide sequences identified from tandem mass spectrometry spectra). Annotations regarding protein expression levels, either normal, pathological, or after a treatment, as well as descriptions of variant polymorphisms, either physiological or related to disease, are also documented in some cases. A certain degree of data integration, the third level of annotations available in SWISS-2DPAGE, comes from database cross-references. Instead of gathering general or specific information existing in other resources, a choice has been made to provide pointers to this information directly in its original database, leading to a better reliability. Currently, the SWISS-2DPAGE links to about 10 different databases, either sequence related (i.e., the UniProt Knowledgebase ***[7]***) or other 2-DE collections (COMPLUYEAST-2DPAGE, HSC-2DPAGE, LENS-2DPAGE, OGP-WWW, PHCI-2DPAGE, PMMA-2DPAGE, Siena-2DPAGE, and others). Last but not least, technical information related to proteomics experiments (protocols, chemicals, apparatus, identification software, and validation criteria, and so on), to database structure schema, and to the current content (number of gels, number of identified spots and proteins, species distribution, and so on) are supported by separate documents. These can be viewed either in the contextual help accessible when browsing the data or as full material for a better comprehension of the database content. Finally, in all cases, relevant literature references are provided, with full author lists, titles, and citations, accompanied by the PubMed identifier for further access to the article abstract.

2.1.2. Availability

Since the beginning, SWISS-2DPAGE was hosted by the ExPASy website (http://www.expasy.org/ch2d/), and users could explore the data with various search engines, either by clicking on an image map or by textual queries. Typical browsing may start in different ways:

1. The simplest approach, well suited for gel exploration, is to navigate on the reference map. After choosing the gel of interest (from http://www.expasy.org/cgi-bin/map1), the user can see an image with identified spots marked by crosses, which facilitates their localizations. (Alternatively, a raw image is also available.)

Moving the mouse over the spots of interest displays a summary of the experimental information in a floating box. To access the full data content related to a specific spot, one could just click on it.

2. The second approach toward interaction with the SWISS-2DPAGE database is through textual queries. Many search engines are provided from the top page (http://www.expasy.org/ch2d/). If the user is aware of identifiers (either protein accession number or spot number) for a spot or protein of interest, he/she can go directly to the corresponding database entry. However, very often, the user has only very partial knowledge about the protein(s) for which he/she wants to have details. In this case, the web interface offers searches either by protein description, gene name, UniProtKB/Swiss-Prot keyword, author name, identification methods, p*I*/Mw range, or a combination of all these. In any case, the result consists of a protein list matching the querying criteria with access to the detailed information for each protein entry.
3. The third way to browse the SWISS-2DPAGE data is to search out the protein list for the reference gel of interest (http://www.expasy.org/cgi-bin/get-ch2d-table.pl) and then to click on the specific protein entry to display the full information.

At the end of all possible ways of querying the database, the user will have the detailed information for a specific protein **(Fig. 1)**, in a friendly view, with all the annotations as described in the section above (*see* **Subheading 2.1.1.**). In addition, this view links to all the maps available in the database, on which the protein has been identified or not. When one clicks on the icons representing these maps, the entire gel images are displayed in which all the spots identified for that protein are highlighted. It displays a rectangle showing an estimated region where the protein is expected to migrate, according to its amino acid composition. Cross-references to external resources (PubMed, UniProt Knowledgebase, other 2-DE databases) are provided as hypertext links.

For people interested in having the SWISS-2DPAGE database locally, annotations and image files can be downloaded from the ExPASy FTP server (ftp://www.expasy.org/ftp/databases/swiss-2dpage/). Either through the website or as a download, the SWISS-2DPAGE database is free for nonprofit institutions. It has an annual subscription fee for commercial entities (for details, *see* http://www.genebio.com/).

2.1.3. Usage

With more than 1800 hits per day, the database is commonly used for preliminary investigations. Usually, users rely on the SWISS-2DPAGE reference maps annotations to start proteomics studies, ranging from biomarker discovery projects, to comparative studies, to theoretical analysis. A biomarker discovery project or a comparative study typically starts with a 2D gel electrophoresis, which is then compared with a reference map, established for a similar or related

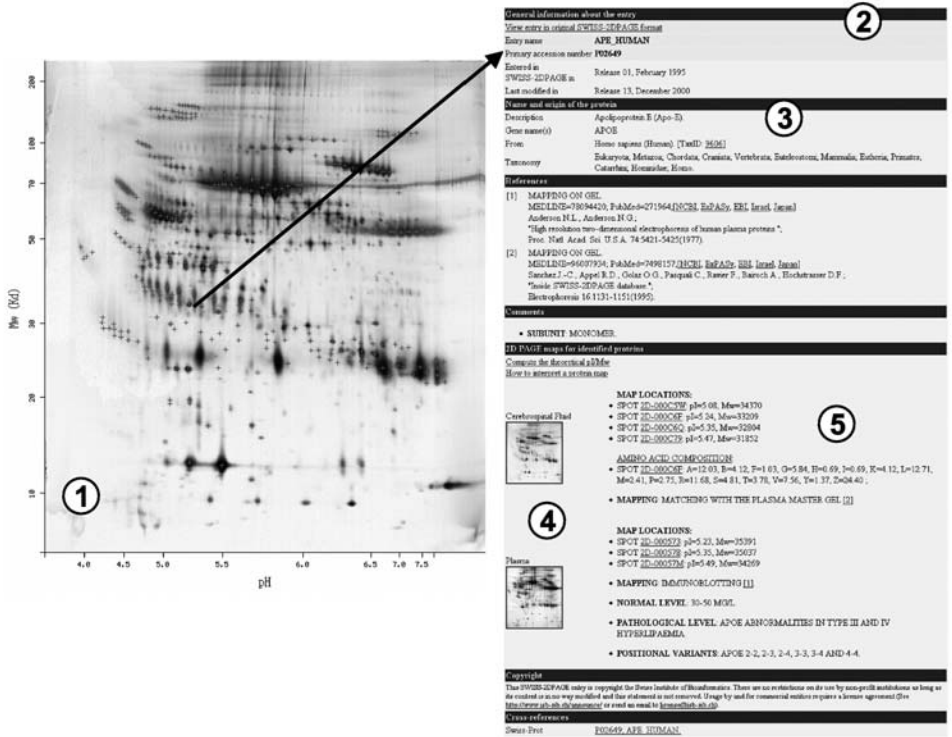

Fig. 1. Following a graphical search, the NiceView of a SWISS-2DPAGE entry shows its content in a user-friendly view. 1, Graphical search on the human cerebrospinal fluid reference map. Crosses show identified proteins, here nearly 300 spots. 2, NiceView of a SWISS-2DPAGE entry. 3, General information describing the protein entry (name, origin, references), provided with links to specific online databases such as PubMed or NCBI's taxonomic classification. 4, Icons of reference maps available for the protein shown. 5, Annotations including here experimental p*I*/Mw, amino acid composition, mapping methods, pathological and polymorphism information, and so on.

sample. When the reference map is well covered, with many identified spots, the comparative analysis with adequate software (e.g., ImageMaster™ 2D Platinum, developed by the Swiss Institute of Bioinformatics and commercialized by GeneBio and GE Healthcare) may reduce the number of spots that need further analysis. Such image analysis software is also able to highlight spots of interest that are over- or underexpressed, thanks to appropriate statistical differential analysis. With nearly 4000 identified spots on the 36 reference maps, SWISS-2DPAGE is widely used for such purposes.

As an example, Wattiez et al. ***(8)*** used the SWISS-2DPAGE human plasma reference map in a preliminary matching analysis to identify a number of proteins in their bronchoalveolar lavage fluid (BALF) 2-DE gel. They were able to identify 19 of 69 proteins of interest by matching the plasma and the BALF samples. The authors restricted the postseparation analysis (in this case microsequencing) to the only 50 remaining proteins. It was a first step to a better understanding of the normal BALF sample, which then allowed these researchers to analyze differential protein expression further between different lung pathological samples. In another example, Pietrogrande et al. ***(9)*** tested a mathematical method to extract information from 2-DE gels of complex protein mixture. The authors compared their simulated gels with three human reference maps from the SWISS-2DPAGE database (PLASMA_HUMAN, HEPG2_ HUMAN, DLD1_HUMAN) and thus confirmed their model hypothesis. As a third example (not from human body fluid samples but easily transferrable), Martens and co-workers ***(10)*** compared their human platelet protein list obtained with nongel technology with protein lists published earlier using gel technology, including the SWISS-2DPAGE platelet proteins. The low overlap (40% so far) observed between both technologies highlighted the advantage of nongel technology to identify proteins in complex mixtures even with low concentrations.

2.2. Other Useful Resources

Since 1993, a growing number of 2-DE images have been made publicly available, thanks to the web. Unfortunately, most of them are limited to one image and a protein list, thus not making the most of the Internet properties (data integration and navigation). With less than 20 image maps electronically available, temporarily or not, in the last 12 years, proteomics data from human body fluids are definitive examples of data regrettably lost or not fully exploited. Although bioinformatics tools have been developed to help scientists build their own database servers, it is clear that it is still not easy to do for institutes with limited access to the necessary infrastructure. These tools have similar functionalities. Starting from at least one image map with identified spots, they provide facilities to build a web server that allow easy browsing and

searching sets of 2-DE data. Make2D-DB (http://www.expasy.org/ch2d/make2ddb.html), developed by the Swiss Institute of Bioinformatics, is an integrated tool that allows anyone to get a SWISS-2DPAGE-like database ***(11)***. PROTICdb has been designed especially for plant samples (http://moulon.inra.fr/~bioinfo/PROTICdb). The French National Institute for Agricultural Research laboratories use it ***(12)***. Both tools are freely available for download and offer useful features to import new data and to browse the data by keyword or graphical searches. Public repositories, in which users can upload their own 2-DE data through a web form, are not yet currently used. So far, two such repositories are available through the Internet, namely, GELBANK (http://gelbank.anl.gov/) from the Argonne National Lab ***(13)*** and GelBank (http://www.gelscape.ualberta.ca:8080/htm/gdbIndex.html) from the University of Alberta ***(14)***. Both systems are proposed as public archives for 2-DE image maps and thus are open to host any data. However, the current data available are still limited to their own local or test samples.

Further measures need to be taken if scientists really want to build a large proteomics data network, which would be one of the objectives of proteomics. In this perspective, in late 2005 the Proteome Informatics group at the Swiss Institute of Bioinformatics started to provide the scientific community with some bioinformatics services related to proteomics. In particular, the group proposes various solutions to build proteomics electronic databases, ranging from assistance in the setup to data preparation for further remote installation, up to full data preparation and website hosting (http://www.expasy.org/ch2d/service/). We hope that with this contribution, more and more institutes will be able to give access to their proteomics data and thus promote this global virtual database that we all have dreamed of.

3. Conclusions

Even though bioinformatics have been developed to help scientists publish their own proteomics data on the web, giving access to a large proteomics data network, one must admit that we are still far from this reality. Assistance from bioinformatics groups should be reinforced to provide scientists with new tools. Some of the available tools already contain functionalities for better data reliability and consistency, and then integration of different "omics" will be the next step to a further understanding of systems biology.

Acknowledgments

Our thanks go to Patricia M. Palagi, who read the manuscript and provided valuable comments.

References

1. Wheeler DL, Barrett, T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2005;33:D39–D45.
2. Hoogland C, Sanchez JC, Walther D, et al. Two-dimensional electrophoresis resources available from ExPASy. *Electrophoresis* 1999;20:3568–3571.
3. Appel RD, Bairoch A, Sanchez JC, et al. Federated two-dimensional electrophoresis database: a simple means of publishing two-dimensional electrophoresis data. *Electrophoresis* 1996;17:540–546.
4. Appel RD, Sanchez JC, Bairoch A, et al. SWISS-2DPAGE: a database of two-dimensional gel electrophoresis images. *Electrophoresis* 1993;14:1232–1238.
5. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* 2003;31:3784–3788.
6. Hoogland C, Mostaguir K, Sanchez JC, Hochstrasser DF, Appel RD. SWISS-2DPAGE, ten years later. *Proteomics* 2004;4:2352–2356.
7. Bairoch A, Apweiler R, Wu CH, et al. The Universal Protein Resource (UniProt). *Nucleic Acids Res* 2005;33:D154–D159.
8. Wattiez R, Hermans C, Cruyt C, Bernard A, Falmagne P. Human bronchoalveolar lavage fluid protein two-dimensional database: study of interstitial lung diseases. *Electrophoresis* 2000;21:2703–2712.
9. Pietrogrande MC, Marchetti N, Tosi A, Dondi F, Righetti PG. Decoding two-dimensional polyacrylamide gel electrophoresis complex maps by autocovariance function: a simplified approach useful for proteomics. *Electrophoresis* 2005; 26:2739–2748.
10. Martens L, Van Damme P, Van Damme J, et al. The human platelet proteome mapped by peptide-centric proteomics: a functional protein profile. *Proteomics* 2005;5:3193–3204.
11. 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.
12. Ferry-Dumazet H, Houel G, Montalent P, et al. PROTICdb: a web-based application to store, track, query, and compare plant proteome data. *Proteomics* 2005;5: 2069–2081.
13. Babnigg G, Giometti CS. GELBANK: a database of annotated two-dimensional gel electrophoresis patterns of biological systems with completed genomes. *Nucleic Acids Res* 2004;32:D582–D585.
14. Young N, Chang Z, Wishart DS. GelScape: a web-based server for interactively annotating, manipulating, comparing and archiving 1D and 2D gel images. *Bioinformatics* 2004;20:976–978.
15. Wattiez R, Hermans C, Bernard A, Lesur O, Falmagne P. Human bronchoalveoalr lavage fluid: two-dimensional gel electrophoresis, amino-acid microsequencing and identification of major proteins. *Electrophoresis* 1999;20:1634–1645.

16. Omenn GS. Advancement of biomarker discovery and validation through the HUPO plasma proteome project. *Dis Markers* 2004;20:131–134.
17. Thalmann I. Proteomics and the inner ear. *Dis Markers* 2001;17:259–270.
18. Fortunato D, Giuffrida MG, Cavaletto M, et al. Structural proteome of human colostral fat globule membrane proteins. *Proteomics* 2003;3:897–905.
19. Finehout EJ, Franck Z, Lee KH. Towards two-dimensional electrophoresis mapping of the cerebrospinal fluid proteome from a single individual. *Electrophoresis* 2004;25:2564–2575.
20. Robinson MK, Myrick JE, Henderson LO, et al. Two-dimensional protein electrophoresis and multiple hypothesis testing to detect potential serum protein biomarkers in children with fetal alcohol syndrome. *Electrophoresis* 1995;16: 1176–1183.
21. Rasmussen HH, Orntoft TF, Wolf H, Celis JE. Towards a comprehensive database of proteins from the urine of patients with bladder cancer. *J Urol* 1996;155: 2113–2119.
22. Liberatori S, Bini L, De Felice C, et al. A two-dimensional protein map of human amniotic fluid at 17 weeks' gestation. *Electrophoresis* 1997;18:2816–2822.
23. Hill A, Kim H. The UAB Proteomics Database. *Bioinformatics* 2003;19: 2149–2151.
24. Fung KY, Glode LM, Green S, Duncan MW. A comprehensive characterization of the peptide and protein constituents of human seminal fluid. *Prostate* 2004;61: 171–181.