

Plant Proteomics Databases: Their Status in 2005

Setsuko Komatsu*

National Institute of Agrobiological Sciences, Tsukuba 305-8602, Japan

Abstract: Proteome analysis linked to genome sequence information is very useful for functional genomics. Since proteins are the major players in most processes of living cells, knowledge of the proteome has great relevance to the study of cells and organisms at the molecular level. The technique of proteome analysis using two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) has the power to monitor global changes that occur in the protein complement of tissues and subcellular compartments. As a complement to more focused studies, and to facilitate further advances in functional genomics, several databases based on 2D-PAGE are already available including those for plants. In this review, the rice proteome database and other plant proteome databases are discussed. Organizing and streamlining the access of information into plant proteome databases, especially the rice proteome database, will aid in cloning the genes and predicting the function of unknown proteins.

Keywords: Plant, rice, proteome, database, two-dimensional gel electrophoresis, mass spectrometry.

INTRODUCTION

Gaining an understanding of the biological functions of novel genes is a more ambitious goal than just obtaining their sequences; the wealth of information on nucleotide sequences being generated through genome projects far outweighs what is currently available on amino acid sequences of known proteins [1, 2]. Because the analysis of proteins is the most direct approach to defining the function of their associated genes, proteome analysis linked to genome sequence information is a very powerful tool for functional genomics; however, the genome and proteome of an organism do not correspond on a one-to-one basis. Alternative transcription initiation and splicing of mRNAs can produce multiple transcripts from a single gene. Alternative translation initiation sites may produce different proteins from each of these transcripts, and these protein variants can be targeted to different compartments in the cell and/or have different functions. Protein maturation does not stop with translation. Post-translational modifications, such as phosphorylation, acylation, ubiquitylation, or proteolytic processing, can alter protein activity, location, and stability. Proteins move in and out of protein complexes depending on post-translational modifications. Once a protein is produced, it can undergo a staggering array of highly regulated changes with enormous implications to biological processes. Since proteins are the major players in most processes of living cells, knowledge of the proteome has great relevance to the study of cells and organisms at the molecular level.

Also, the genome is static; the proteome is highly dynamic in its response to external and internal cellular events. These changes include the relative abundance of each protein and post-translational modifications. Many two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) reference maps have been built in order to catalog

the proteins expressed in an organ, a tissue, a cell, and even at the sub-cellular level [3]. Other studies focus on the variation of protein relative abundance [4]. 2D-PAGE comparative proteomics consist of the identification of proteins differentially expressed between constructed environmental conditions or in the course of developmental processes of a given organism or tissues [2]. Such a strategy may allow for the discovery of proteins of agronomical [5, 6] and medical interest [7]. However, a single protein profiling experiment can provide only a few indications regarding protein function. Integration of these data with results from previous experiments or from distinct proteomic approaches, along with genome and transcriptome databases, is critical to extract knowledge from proteomic experiments [8].

Several databases based on 2D-PAGE are already available on the internet. These are ECO-2DBASE [9], HSC-2DPAGE [10], YPD [11], SIENA-2DPAGE [12, 13], PHCI-2DPAGE [14, 15], and SWISS-2DPAGE [16]. For plant proteins, several databases based on 2D-PAGE are also available, such as WORLD-2DPAGE (<http://expasy.org/ch2d/2d-index.html>). In rice, catalogs of predicted membrane proteins, such as the Rice Membrane Protein Library (<http://www.cbs.edu/rice/>) are in the public domain, thus providing further support for rice proteomics efforts. In addition, the recently constructed Rice Proteome Database website (<http://gene64.dna.affrc.go.jp/RPD/>) provides extensive information on the progress of rice proteome research [17]. Proteomic analysis of select tissues and organelles has revealed diverse functional categories of proteins. In this review, the Rice Proteome Database and other plant proteome database are described.

PLANT PROTEOME DATABASES

The genomes of rice and Arabidopsis have been sequenced. Powerful genomic tools for these plants include microarrays to examine changes in transcript levels and knockout lines for most of the genes. Because proteins are the major players in most processes of living cells, many plant proteomic investigations have been reported recently, and also several plant proteome databases have been

*Address correspondence to this author at the Department of Molecular Genetics, National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba 305-8602, Japan; Tel: 81-029-838-7446; Fax: 81-029-838-7408; E-mail: skomatsu@affrc.go.jp

constructed. The Arabidopsis proteome databases were more extensive than other plant proteome databases, because the Arabidopsis genome has been sequenced completely and deduced protein sequences could be retrieved from the EMBL proteome site (<http://www.ebi.ac.uk/proteome>). In addition to the EMBL resources, many other proteome data sets for Arabidopsis have also been accumulating [18]. The protein sequences from other plants were retrieved from the Swiss-Prot (release 40) and TrEMBL (release 20) data banks [19]. 2D-PAGE gel protein reference maps of sub-proteomes of different plant species are expected to become a central tool for organizing and understanding the plant proteome. Web sites with organized 2D-PAGE database are already available (<http://sphin.rug.ac.be:8080/ppmdb/index.html>, <http://www.biokemi.su.se/chloroplast/> and <http://www.expasy.ch/ch2D/> and so on). In the future, reference 2D-PAGE maps will be used to follow differential protein expression and post-translational modifications. In this review, several specific examples of plant proteome databases are compared as follows:

***Nicotiana tabacum*, cell suspension culture** - Although the tobacco genome is not sequenced yet, proteins from *Nicotiana tabacum* cv. Bright Yellow-2 (BY-2) cell suspension culture were analyzed using 2D-PAGE and electrospray mass spectrometry (ESI-MS/MS). These data were integrated in a database that can be accessed at <http://www.pdata.ua.ac.be/BY2/modules/main/>. It includes 73 proteins from cell suspension culture. At the on-line reference map, the identified protein spots are hyperlinked to individual protein entries in other on-line databases. Comprehensive search functions are integrated into the database design. Especially for an unsequenced but widely used model organism like tobacco BY-2, such a reference database is a convenient source for protein information that brings protein identification within reach without the need for extensive MS [20].

Soybean, seed filling - A high-throughput proteomic approach was employed to determine the expression profile and identity of 100 proteins during seed filling in soybean (*Glycine max* cv. Maverick). Soybean seed proteins were analyzed at 2, 3, 4, 5 and 6 weeks after flowering using 2D-PAGE and matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI TOF MS). A user-intuitive database (<http://oilseedproteomics.missouri.edu>) was developed to access these data for soybean and other oilseeds currently being investigated. This led to the establishment of high-resolution proteome reference map, expression profiles of 679 spots, and corresponding MALDI TOF MS spectra for each spot [21].

Arabidopsis mitochondrial protein database - The Arabidopsis Mitochondrial Protein Database is an internet-accessible relational database containing informed protein complement of mitochondria from the model plant Arabidopsis (<http://www.ampdb.bcs.uwa.edu.au/>). It includes 117 mitochondrial encoded proteins. The database was formed using the total non-redundant nuclear and organelle encoded sets of protein sequences and allows relational searching of published proteomic analyses of Arabidopsis mitochondrial samples, a set of predictions from six independent subcellular-targeting prediction

programs, and orthology prediction based on pairwise comparison of the Arabidopsis protein set with known yeast and human mitochondria proteins [22].

Arabidopsis nucleolar protein database - The Arabidopsis Nucleolar Protein Database (<http://bioinf.scri.sari.ac.uk/cgi-bin/atnopdb/home>) provides information on 217 proteins identified in a proteomic analysis of nucleoli isolated from Arabidopsis cell culture. The database is organized on the basis of the Arabidopsis gene identifier number. The information provided includes protein description, protein class, whether or not the plant protein has a homolog in the most recent human nucleolar proteome and the results of reciprocal BLAST analysis of the human proteome [23].

Until now, the freely available plant 2D-PAGE proteome databases are limited in the diversity and level of detail in stored data. For example, proteins analyzed using cell suspension culture of tobacco and rice were 73 and 245, respectively; similarly proteins during seed filling stage in soybean and rice were 100 and 446 in number, respectively. Concerning rice and Arabidopsis protein database, rice mitochondria and nucleolar consisted of 121 and 503 proteins in comparison to 117 and 217 proteins in Arabidopsis organelle.

RICE PROTEOME DATABASE

Rice is not only a very important agricultural resource; it is also a model plant for biological research because its genome is smaller than those of other cereals [24]. Publication of draft genome sequences for *Oryza sativa* L. ssp. indica [25] and for *Oryza sativa* L. ssp. Japonica [26], and a complete map-based sequence of chromosome 1 [27] and chromosome 4 [28] for *Oryza sativa* L. cv. Nipponbare provide a rich resource for understanding the biological processes of rice. Once the rice genome is completely sequenced [29], the challenge ahead for the plant research community will be to identify the function, regulation, and type of post-translational modification of each encoded protein. During the last couple of years, considerable research effort has been applied to analysis of the rice proteome [6]. Just recently, remarkable progress has been made in the systematic functional characterization of proteins in various tissues and organelles in rice [16].

Format and content of the "Rice Proteome Database" - As a complement to more focused studies, and to facilitate future advances in rice functional genomics, the Rice Proteome Database has been constructed. The Rice Proteome Database [30] compiles information about proteins identified on 2D-PAGE maps of protein extracts from a wide variety of rice tissues and subcellular compartments. Each entry in the Rice Proteome Database corresponds to one protein from the 2D-PAGE image file. The following three features are specific to the Rice Proteome Database:

- (i) The reference 2D-PAGE map shows the position of the identified entry. Spot numbers are displayed on this 2D-PAGE image. The spot list contains a table listing the number of proteins on each 2D-PAGE map in the Rice Proteome Database. Experimental protocols used for protein purification and 2D-PAGE, with either IEF or IPG in the first dimension,

are shown on this page. The 2D-PAGE image was synthesized as a composite of gels run using the two different first-dimension methods and the positions of individual proteins on the gels were evaluated using Image Master 2D Elite software.

- (ii) The spot information pages provide a range of information about each protein spot, including mapping procedure and spot coordinates; the calculated properties of the protein such as molecular weight, isoelectric point, and expression level; the experimentally determined properties, such as amino acid sequences and peptide masses obtained using protein sequencers and mass spectrometry, respectively, and the homologous proteins predicted by these two methods and other information. The accession number of each homologous protein links to the NCBI site (<http://www.ncbi.nlm.nih.gov/>). Other information shows the accession number and the percent identity of the homologous full-length cDNA in rice, and biological information such as the known function or functions obtained *via* experimentation.
- (iii) The Mascot Search results page displays the peptide masses derived from mass spectrometry. This page brings together the Mascot Search Results such as the accession numbers of homologous proteins, scores, sequence coverage, and predicted peptides. This page is also linked to the Mascot Web site (<http://www.matrix-science.com/>).

How to use the “Rice Proteome Database” - The Rice Proteome Database can be reached on the World Wide Web through the Rice Proteome Database Home Page at <http://gene64.dna.affrc.go.jp/RPD/>. The Rice Proteome Database Home Page and the contents of the Rice Proteome Database are maintained by the authors. The Rice Proteome Database Home Page provides introductory material on the Rice Proteome Database. A Rice Proteome Database entry may be obtained from the server in one of four ways:

- (i) By selecting a spot on one of the 2D-PAGE reference maps. The Rice Proteome Database contains information on proteins identified from several tissues and organelles on 2D-PAGE reference maps. These 2D-PAGE maps can be reached by clicking the individual tissues/organelles denoted by red boxes. Only spots with sequence data are highlighted and labeled “Annotation Data Available”.
- (ii) By “protein keyword” or “protein database accession identifiers” using the protein name or accession number. The Rice Proteome Database can be searched using protein names as keywords.
- (iii) By isoelectric point and molecular weight for any protein. The Rice Proteome Database can be searched for proteins with a range of isoelectric points and molecular weights.
- (iv) By similarity search with the user’s amino acid sequences. The query sequence can be searched using the homology search tools BLASTP and

BLASTX for the presence of amino acid sequences identical to or similar to previously reported amino acid sequences in the Rice Proteome Database.

Cataloguing of rice proteins in the “Rice Proteome Database” - The current release contains 23 reference maps from rice biological samples that are either tissue-specific, such as cultured suspension cells, endosperm, embryo, crown which is the basal part of leaf sheath in young seedling, seedling root, seedling leaf sheath, seedling leaf blade, stem, mature plant root, mature plant leaf sheath, mature plant leaf blade, anthers, panicle before heading, panicle after heading and 1 week after flowering, or specific to a subcellular location, such as cell wall, plasma membrane, vacuole membrane, Golgi membrane, mitochondrion, chloroplast, nucleus, and cytosol. These reference maps of proteins from various tissues and subcellular fractions have a total of 13,129 identified protein spots, corresponding to 5,236 separate protein entries in the database (Table 1). The information on amino acid sequences is updated frequently. The Rice Proteome Database has links to the NIAS Rice genome tools, which are the Rice Expression Database (RED), the Rice Full-length cDNA Database (KOME), the Rice Genome Integrated Map Database (INE), the Rice Mutant Panel Database (Tos17), the Rice Genome Annotation Database (RiceGAAS), and DNA Bank. The Rice Proteome Database also links to many useful proteomics tools and other proteomics databases.

Table 1. Content of Rice Proteome Database: Its Status in March 2005

Map	Detected ^a	Identified ^b	Entries ^c
Cultured suspension cells	962	245	245
Endosperm	100	37	37
Embryo	639	409	409
Crown	700	480	480
Seedling root	508	48	48
Seedling leaf sheath	431	145	145
Seedling leaf blade	679	235	200
Stem	567	186	186
Mature plant root	265	100	100
Mature plant leaf sheath	509	115	115
Mature plant leaf blade	718	350	350
Anthers	1,080	365	365
Panicle before heading	704	441	441
Panicle after heading	559	361	361
One week after flowering	1,073	324	324
Cell wall	513	111	111
Plasma membrane	464	159	90
Vacuolar membrane	141	74	43
Golgi membrane	361	187	44
Mitochondria	672	369	121
Chloroplast	252	159	66
Nucleus	549	503	503
Cytosolic fraction	683	352	325
Total	13,129	5,755	5,236

The numbers in the table show the spot number.

^a) Detected protein spots on 2D-PAGE.

^b) Identified protein spots mean resolvable spots.

^c) Entered proteins in the Rice Proteome Database.

FUTURE PROSPECTS OF THE PLANT PROTEOME DATABASE

Information about post-translational modifications such as phosphorylation, glycosylation, and other modifications, obtained experimentally, is important to include in plant proteome databases. As new information from functional analyses of physiologically significant proteins becomes available, the number of identified proteins will increase. Similarly, it is expected that genomic databases will continue to be updated frequently; the frequency of updates to proteomic databases should follow closely behind.

Analysis by 2D-PAGE provides a convenient way to study the various proteins that are present in plants and to identify those that are regulated in response to different growth and/or stress conditions. Knowing where and when individual proteins are being synthesized in a plant, with respect to tissue, subcellular compartment, and developmental stage, can also provide new clues about their function. The partial amino acid sequences determined for these proteins will contribute greatly to the field of plant molecular biology, by allowing the identification of new plant proteins of interest through homology searches. The information thus obtained from the plant proteome databases will be helpful in predicting the function of plant proteins and will aid in their molecular cloning in future experiments. The present-day proteomics research promises a great deal for developing the high-yielding, sustainable agriculture of tomorrow.

ACKNOWLEDGMENTS

The author wishes to thank Dr. K. Higo at NIAS for valuable discussion, and Mr. K. Kojima, M. Yamamoto and K. Suzuki at Hitachi Soft. Co. for their technical assistance in database preparation. The author wishes to thank the members of the Rice Proteome Analysis Center for their assistance in protein identification. This work was supported by a grant from the rice genome project, Ministry of Agriculture, Forestry, and Fisheries, Japan.

REFERENCES

- [1] Lockhart JD, Winzler AE. Genomics, gene expression and DNA arrays. *Nature* **2000**; 405: 827-835.
- [2] Pandey A, Mann M. Proteomics to study genes and genomics. *Nature* **2000**; 405: 837-845.
- [3] Canovas FM, Dumas-Gaudot E, Recorbet G, *et al.* Plant proteome analysis *Proteomics* **2004**; 4: 285-298.
- [4] Zue H, Bilgin M, Snyder M. Proteomics *Annu Rev Biochem* **2003**; 72: 783-812.
- [5] Thiellement H, Bahrman N, Damerval C, *et al.* Proteomics for genetic and physiological studies in plants. *Electrophoresis* **1999**; 20: 2013-2026.
- [6] Komatsu S, Konishi H, Shen S, Yang G.. Rice proteomics: a step functional analysis of the rice genome. *Mol Cell Proteomics* **2003**; 2: 2-10.
- [7] Wulfskuhle JD, Paweletz CP, Steeg PS, *et al.* Proteomic approaches to the diagnosis, treatment, and monitoring of cancer *Adv Exp Med Biol* **2003**; 532: 59-68.
- [8] Ge H, Walhout AJ, Vidal M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet* **2003**; 19: 551-560.
- [9] VanBogelen RA, Abshire KZ, Moldover B, Olson ER, Neidhardt FC. Escherichia coli proteome analysis using the gene-protein database. *Electrophoresis* **1997**; 18: 1243-1251.
- [10] Dunn MJ, Corbett JM, Wheeler CH. HSC-2DPAGE and the two-dimensional gel electrophoresis database of dog heart proteins. *Electrophoresis* **1997**; 18: 2795-2802.
- [11] Payne WE, Garrels JI. Yeast protein database (YPD): a database for the complete proteome of *Saccharomyces cerevisiae*. *Nucl Acids Res* **1997**; 25: 57-62.
- [12] Bini L, Heid H, Liberatori S, Geier G, Pallini V, Zwilling R. Two-dimensional gel electrophoresis of *Caenorhabditis elegans* homogenates and identification of protein spots by microsequencing. *Electrophoresis* **1997a**; 18: 557-562.
- [13] Bini L, Magi B, Marzocchi B, *et al.* Protein expression profiles in human breast ductal carcinoma and histologically normal tissue. *Electrophoresis* **1997b**; 18: 2832-2841.
- [14] Shaw AC, Larsen MR, Roepstorff P, Holm A, Christiansen G, Birkelund S. Mapping and identification of HeLa cell proteins separated by immobilized pH-gradient two-dimensional gel electrophoresis and construction of a two-dimensional polyacrylamide gel electrophoresis database. *Electrophoresis* **1999a**; 20: 977-983.
- [15] Shaw AC, Larsen MR, Roepstorff P, Justesen J, Christiansen G, Birkelund S. Mapping and identification of interferon gamma regulated HeLa cell proteins separated by immobilized pH gradient two-dimensional gel electrophoresis. *Electrophoresis* **1999b**; 20: 984-993.
- [16] Hoogland C, Sanchez J-C, Tonella L, *et al.* The 1999 SWISS-2DPAGE database update. *Nucl Acids Res* **2000**; 28: 286-288.
- [17] Komatsu S, Tanaka N. Rice proteome analysis: A step toward functional analysis of the rice genome. *Proteomics* **2004**; 4: 938-949.
- [18] Peck SC. Update on proteomics in Arabidopsis. Where do we go from here? *Plant Physiol* **2005**; 138: 591-599.
- [19] Boeckmann B, Bairoch A, Apweiler R, *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl Acids Res* **2003**; 31: 365-370.
- [20] Laukens K, Deckers P, Esmans E, Onckelen HV, Witters E. Construction of a two-dimensional gel electrophoresis protein database for the *Nicotiana tabacum* cv. Bright Yellow-2 cell suspension culture. *Proteomics* **2004**; 4: 720-727.
- [21] Hajdich M, Ganapathy A, Stein JW, Thelen JJ. A systematic proteomic study of seed filling in soybean. Establishment of high-resolution two-dimensional reference maps, expression profiles, and an interactive proteome database. *Plant Physiol* **2005**; 137: 1397-1419.
- [22] Heazlewood JL, Millar AH. AMPDB: the Arabidopsis mitochondrial protein database. *Nucl Acid Res* **2005**; 33: 605-610.
- [23] Brown JWS, Shaw PJ, Shaw P, Marshall DF. Arabidopsis nucleolar protein database (AtNoPDB). *Nucl Acids Res* **2005**; 33: 633-636.
- [24] Devos MK, Gale DM. Genome relationships: The grass model in current research. *Plant Cell* **2000**; 12: 637-646.
- [25] Yu J, Hu S, Wang J, *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science* **2002**; 296: 79-92.
- [26] Goff SA, Ricke D, Lan T-H, *et al.* A draft sequence of rice genome (*Oryza sativa* L. ssp. japonica). *Science* **2002**; 296: 92-100.
- [27] Sasaki T, Matsumoto T, Yamamoto K, *et al.* The genome sequence and structure of rice chromosome 1. *Nature* **2002**; 420: 312-316.
- [28] Feng Q, Zhang Y, Hao P, *et al.* Sequence and analysis of rice chromosome 4. *Nature* **2002**; 420: 316-320.
- [29] International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **2005**; 436: 793-800.
- [30] Komatsu S, Kojima K, Suzuki K, Ozaki K, Higo K. Rice Proteome Database based on two-dimensional polyacrylamide gel electrophoresis: its status in 2003. *Nucl Acids Res* **2004**; 32: 388-392.