The Multinational Arabidopsis Steering Subcommittee for Proteomics Assembles the Largest Proteome Database Resource for Plant Systems Biology

Wolfram Weckwerth, University of Vienna; Sacha Baginsky; Eth Zurich; Klaas Van Wijk. Cornell University; Joshua Heazlewood, Joint BioEnergy Institute; Harvey Millar, University of Western Australia

In the past 10 years, we have witnessed remarkable advances in the field of plant molecular biology. The rapid development of proteomic technologies and the speed with which these techniques have been applied to the field have altered our perception of how we can analyze proteins in complex systems. At nearly the same time, the availability of the complete genome for the model plant Arabidopsis thaliana was released; this effort provides an unsurpassed resource for the identification of proteins when researchers use MS to analyze plant samples. Recognizing the growth in this area, the Multinational Arabidopsis Steering Committee (MASC) established a subcommittee for A. thaliana proteomics in 2006 with the objective of consolidating databases, technique standards, and experimentally validated candidate genes and functions. Since the establishment of the Multinational Arabidopsis Steering Subcommittee for Proteomics (MASCP), many new approaches and resources have become available. Recently, the subcommittee established a webpage to consolidate this information (www.masc-proteomics.org). It includes links to plant proteomic databases, general information about proteomic techniques, meeting information, a summary of proteomic standards, and other relevant resources. Altogether, this website provides a useful resource for the Arabidopsis proteomics community. In the future, the website will host discussions and investigate the cross-linking of databases. The subcommittee members have extensive experience in arabidopsis proteomics and collectively have produced some of the most extensive proteomics data sets for this model plant (Table S1 in the Supporting Information has a list of resources). The largest collection of proteomics data from a single study in A. thaliana was assembled into an accessible database (AtProteome; http://fgcz-atproteome.unizh.ch/index.php) and was recently published by the Baginsky lab.[1] The database provides links to major Arabidopsis online resources, and raw data have been deposited in PRIDE and PRIDE BioMart. Included in this database is an Arabidopsis proteome map that provides evidence for the expression of ~50% of all predicted gene models, including several alternative gene models that are not represented in The Arabidopsis Information Resource (TAIR) protein database. A set of organ-specific biomarkers is provided, as well as organ-specific proteotypic peptides for 4105 proteins that can be used to facilitate targeted quantitative proteomic surveys. In the future, the AtProteome database will be linked to additional existing resources developed by MASCP members, such as PPDB, ProMEX, and SUBA.

The most comprehensive study on the Arabidopsis chloroplast proteome, which includes information on chloroplast sorting signals, posttranslational modifications (PTMs), andprotein abundances (analyzed by high-accuracy MS [Orbitrap]), was recently published by the van Wijk lab.[2] These and previous data are available via the plant proteome database (PPDB; http://ppdb.tc.cornell.edu) for A. thaliana and maize. PPDB provides genomewide experimental and functional character-ization of the A. thaliana and maize proteomes, including PTMs and subcellular localization information, with an emphasis on leaf and plastid proteins. Maize and Arabidopsis proteome entries are directly linked via internal BLAST alignments within PPDB. Direct links for each protein to TAIR, SUBA, ProMEX, and other resources are also provided.

A comprehensive database on the subcellular localization of Arabidopsis proteins was extensively updated in 2007 (SUBA; www.suba.bcs.uwa.edu.au) by the Millar lab.[3] These data are linked to several other databases, and selected data are provided as web services via the BioMoby Dashboard. The database houses relational data with localization information from subcellular proteome studies, fluorescent protein targeting studies, AmiGO and UniProt information, as well as several bioinformatics prediction programs.

The recent use of high-throughput techniques for phosphoproteomic analyses led the Heazlewood, Weckwerth, and Schulze labs to establish a database of phosphorylation sites in A. thaliana (PhosPhAt; http://phosphat.mpimp-golm.mpg.de) earlier this year.[4] The database also incorporates a phospho-serine prediction algorithm that has been used to precompute phosphoserine sites across all TAIR gene models.

A central, searchable database of MS/MS reference spectra (mostly Orbitrap accurate precursor ion mass data) derived from A. thaliana, Chlamydomonas reinhardtii, Medicago trun-catula, potato, tomato, and other plants (ProMEX; http:// promex.mpimpgolm.mpg.de/home.shtml) was established last year by the Weckwerth lab.[5] The database provides an interface to match newly generated MS/MS spectra against

previously derived experimental MS/MS spectra; this process results in high-fidelity matching between real spectra and the identification of poor and unmatched spectra. In addition, the spectral count for each identified protein is exported into the search result table for semiquantitative analysis. The ProMEX database serves as a design tool for proteotypic peptides that can be used for targeted, accurate protein quantification in complex samples. Unidentified spectra of good S/N quality can be included in the ProMEX database; such spectra can facilitate the assign-ment of unknown spectra. The database cross-references the UniProt plant genome annotation initiative[6] and other resources.

A genome browser (AnnoJ Web 2.0; http://neomorph.salk.edu/ epigenome.html) developed by the Millar lab will be a tool to watch for in the future. It is currently being used for displaying deep-sequencing DNA and RNA data,[7] but the same interface will be used to establish genome browsing functionality for the proteogenomic mapping of MS/MS spectral data. This new function is expected to be completed by 2009.

Finally, several predicted protein-protein interaction sets have been developed for Arabidopsis proteins based on their homology with proteins from other organisms for which researchers have reported substantial experimental protein-protein interaction sets.[8,9] These are accessed via the Arabi-dopsis Interaction Viewer (http://bbc.botany.utoronto.ca/ interactions) and the A. thaliana protein interactome database (AtPID; http://atpid.biosino.org). The establishment of large protein interaction databases using real experimental data from Arabidopsis and their cross-links to proteomic resources is a major focus for MASCP in the next few years.

MASCP was established to facilitate and support the use of proteomics in the model plant A. thaliana. Since its inception, the subcommittee has been actively communicating with the Arabidopsis community through the MASC annual reports, the MASC proteomics website, and through proteomic workshops at the annual International Conference on Arabidopsis Research (www.plantconferences.org/Arabidopsis2008). As proteomics techniques are more widely adopted by plant researchers, it is hoped that the subcommittee will continue to provide guidance, expertise, and resources to aid the adoption of proteomics in plant research.

REFERENCES

(1) Baerenfaller, K.; et al. Science **2008**, 320, 938–941.

(2) Zybailov, B.; et al. PLoS One **2008**, 3, e1994.

(3) Heazlewood, J. L.; et al. Nucleic Acids Res. **2007**, 35, D213-D218.

(4) Heazlewood, J. L.; et al. Nucleic Acids Res. **2008**, 36, D1015-D1021.

(5) Hummel, J.; et al. BMC Bioinf. **2007**, 8, 216.

(6) Schneider, M.; et al. Plant Physiol. **2005**, 138, 59–66.

(7) Lister, R.; et al. Cell **2008**, 133, 523–536.

(8) Geisler-Lee, J.; et al. Plant Physiol. **2007**, 145, 317–329.

(9) Cui, J.; et al. Nucleic Acids Res. **2008**, 36, D999–D1008.

This work was supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

This work was supported by the Assistant Secretary for Energy Efficiency and Renewable Energy, Office of Building Technology, State, and Community Programs, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.