Biological Databases for Plant Research

Biology has undergone several rounds of transformation in terms of the research paradigms it has operated, ranging from theoretical to experimental, in the pursuit of discovering new molecular mechanisms that regulate biological form and function. In the decades to come, it will take on another transformation to understand the modes of action of biological processes at the organismal level, where computational models of systems-wide properties could serve as the basis for prediction of biological behavior, leading to new experimentation and discovery. For this transformation to occur, it is essential to facilitate and enhance the processing, integration, and interpretation of the massive amounts of biological data by the life science research community. Databases have been a standard way of managing and processing large amounts of information in diverse arenas, including academic disciplines, industry, and government sectors, for many years. The use of database technologies has drawn the attention of a subset of the biological community, but its use has been limited to a small sector of the community—mainly those involved in the organization and distribution of data resources. While these resources are perused by a great number of the research community, the majority of these users are relatively unaware of the initiatives undertaken to acquire, curate, and enhance the content of these databases in service to the wider research community. This can both limit the uses of these data to its maximal capacity as well as lead to misuse of the data. In addition, more and more experimental biologists are generating data on a large scale and are in need of developing and managing databases of their own.

The motivation of organizing this focus is severalfold: (1) to demystify how the major database resources relevant for plant research today acquire, process, and make available their data, what the current limitations and caveats are of these resources, and what the future directions of these resources are; (2) to bring forward the general issues of databases and data management today to the larger plant research community; (3) to engage the general readership of *Plant Physiology* in thinking about large datasets and how to apply them to their research problems; and (4) to encourage researchers in the use and development of databases to further their research goals. This focus issue will be published over three issues: the current May issue, July issue, and August issue. The articles that will be featured in the focus issue represent some the major database resources available today and serve to introduce this topic to the general readership of this journal. However, by no means do they complete the picture of the diversity of database resources that are currently available. To accommodate the ongoing research in the generation of large datasets and development of biological databases, *Plant Physiology*'s Bioinformatics section will have a subsection called Plant Databases starting September 2005.

There are three main types of biological databases that have been established and are being developed—large-scale public repositories, community-specific database resources, and project-specific databases—although the lines among these categories are becoming less clear. Large-scale public repositories are usually developed and maintained by government agencies or international consortia. Examples include GenBank (July Issue, 2005), which is an international nucleotide sequence repository developed and maintained as a collaboration between the National Center for Biotechnology Institute in the United States, EMBL in Europe, and DDBJ in Japan. Other examples include UniProt (Schneider et al., 2005) that stores protein data and ArrayExpress (July Issue, 2005) that stores microarray data. There are a number of community-specific database resources, a key example being model organism databases that cater to researchers focused on specific model species such as maize (*Zea mays*; Lawrence et al., 2005), Medicago (Cannon et al., 2005), rice (*Oryza sativa*; July Issue, 2005), and Arabidopsis (*Arabidopsis thaliana*; August Issue, 2005). The concept of community-specific databases is subject to change as researchers are widening their scope of research. For example, the explosion of available sequence data from many organisms has enabled researchers to more readily compare sequences of interest from many different species in combination with a number of model organism databases (July Issue, 2005). In addition to databases focused on a single species, databases that deal with taxonomically related species have emerged recently, which include databases for cereals (July Issue, 2005), grains (August Issue, 2005), and night shades (August Issue, 2005). Other examples of community-specific databases include those that are focused on specific classes of data, such as metabolism (Zhang et al., 2005), genome annotation (Yuan et al., 2005; July Issue, 2005), orthologous relationships (Horan et al., 2005), and germplasms (August Issue, 2005). The third category of databases includes smaller-scale and often short-lived databases that are developed for the management of project data during the funding period. Often these databases and Web resources are

www.plantphysiol.org/cgi/doi/10.1104/pp.104.900158.

not maintained beyond the funding period of the project, and currently there is no standard way of depositing or archiving these projects or the data stored within. The preservation and ongoing availability of such information has become a problem in recent years. For example, the average lifetime of viable Web sites in publications is about 18 months post publication. This is a problem that does not yet have a clear solution, and innovative and creative methods and technologies of overcoming this issue should be sought after.

These databases have been and are being developed fairly independently, and there is a general lack of good documentation on the rationale of the design and implementation and community-wide standards for operation in annotation and data exchange. Part of this problem comes from the lack of recognition of this work as a legitimate scientific endeavor. Most of the databases described above are public efforts carried out by biologists and software developers in academic settings, and more effort to share their development experiences via conferences and publications would help alleviate the problem. The majority of papers on databases describe mostly the content and user functionality available from the databases and their attendant query interfaces, and offer little information on the design and implementation of the software. Also, there is no standard in making database software and schema available. This is a particularly acute problem for emerging data types, such as those resulting from metabolite profiling experiments. Recently, standards in data description and exchange for plant metabolomics have been proposed (Bino et al., 2004), and a computer-readable data model has been developed (Jenkins et al., 2004). In this issue, specification and example implementation of metabolite profiling data acquisition applications using the proposed standard are described (Jenkins et al., 2005). Another example of emerging standards lies in the development of controlled vocabulary terms for annotation of molecular functions associated with genes and proteins, as developed by the Gene Ontology Consortium (July Issue, 2005).

Another key problem in this field is the limited ability to access and usefully integrate data from these myriad of databases in a seamless manner. The increasing number and types of databases and software applications make it more and more difficult for researchers to find out where to go for what information. In addition, the different ways in which data are presented and made accessible for many of these databases create an additional burden on researchers who seek to apply the available resources to their research. Emerging technologies to solve these problems have been proposed, such as the BioMOBY initiative (Wilkinson et al., 2005). In addition, the National Human Genome Research Institute has funded a collaborative project called the Generic Model Organism Database (www.gmod.org) to promote the development and sharing of software, schemas, and standard operation procedures. The project's major aim is to build a generic organism database toolkit to allow researchers to set up a genome database ''off the shelf.'' Currently, version 1 of the toolkit has been deployed, which includes a database schema (CHADO) for the organization of genome sequence, software to browse the genome sequence and annotation (GBROWSE), and a genome sequence annotation editor (APOLLO). This project also encourages curators of model organism databases to meet on a regular basis to discuss issues of curation. This, in conjunction with sponsorship from the Genetics Society of America, has recently culminated in the organization of a first international biological database curators meeting to be held at Asilomar in December 2005 (www.biocurator.org). It is hoped that these initial interactions and communications will continue and become a seed for establishing a sense of scientific community among all types of biological database projects.

It is clear that biological research is in an ongoing state of transition, where novel methods, technologies, and implementations will increasingly be deployed in the pursuit of an enhanced mechanistic understanding of biological systems. One of the most difficult hurdles to overcome in deploying new technologies and reaching new goals is the training and retraining of biologists to adapt to changing needs and environments. Social engineering and technology application will always occur more slowly than technological engineering. It is our hope that this focus issue and the ensuing Plant Databases section in this journal will contribute to the promotion of the social and technological engineering needed to help transition the plant research community to an enhanced awareness and application of database resources in support of its scientific endeavors.

LITERATURE CITED

Bino RJ, Hall RD, Fiehn O, Kopka J, Saito K, Draper J, Nikolau BJ, Mendes P, Roessner-Tunali U, Beale MH, et al (2004) Potential of metabolomics as a functional genomics tool. Trends Plant Sci **9:** 418–425

Cannon SB, Crow JA, Heuer ML, Wang X, Cannon EKS, Dwan C, Lamblin A-F, Vasdewani J, Mudge J, Cook AJ, et al (2005) Databases and information integration for the *Medicago truncatula* genome and transcriptome. Plant Physiol **138:** 38–46

Horan K, Lauricha J, Bailey-Serres J, Raikhel N, Girke T (2005) Genome Cluster Database: a sequence family analysis platform for Arabidopsis and *Oryza sativa*. Plant Physiol **138:** 47–54

Jenkins H, Hardy N, Beckmann M, Draper J, Smith AR, Taylor J, Fiehn O, Goodacre R, Bino RJ, Hall R, et al (2004) A proposed framework for the description of plant metabolomics experiments and their results. Nat Biotechnol **22:** 1601–1606

Jenkins H, Johnson H, Kular B, Wang TL, Hardy N (2005) Toward supportive data collection tools for plant metabolomics. Plant Physiol **138:** 67–77

Lawrence CJ, Seigfried TE, Brendel V (2005) The Maize Genetics and Genomics Database. The community resource for access to diverse maize data. Plant Physiol **138:** 55–58

Schneider M, Bairoch A, Wu CH, Apweiler R (2005) Plant protein annotation in the UniProt Knowledgebase. Plant Physiol **138:** 59–66

Wilkinson MD, Schoof H, Ernst R, Haase D (2005) BioMOBY successfully integrates distributed heterogeneous bioinformatics web services: the PlaNet exemplar case. Plant Physiol **138:** 5–17

Yuan Q, Ouyang S, Wang A, Zhu W, Maiti R, Lin H, Hamilton J, Haas B, Sultana R, Cheung F, Wortman J, Buell CR (2005) The Institute for Genomic Research Osa1 rice genome annotation database. Plant Physiol **138:** 18–26

Zhang P, Foerster H, Tissier CP, Mueller LA, Paley S, Karp PD, Rhee SY (2005) MetaCyc and AraCyc: metabolic pathway databases for plant research. Plant Physiol **138:** 27–37

Seung Yon Rhee
Monitoring Editor of *Plant Physiology*
rhee@acoma.stanford.edu

Bill Crosby
Monitoring Editor of *Plant Physiology*
bcrosby@uwindsor.ca