Invited Review

Omics databases in plant science: key to systems biology

Keita Suwabe[1,a], Kentaro Yano[2,a,*]

[1] Graduate School of Life Sciences, Tohoku University, Sendai, Miyagi 980-8577, Japan; [2] Faculty of Agriculture, Meiji University, Kawasaki, Kanagawa 214-8571, Japan
* E-mail: kyano@isc.meiji.ac.jp Tel: +81-44-934-7046 Fax: +81-44-934-7046

Received June 20, 2008; accepted July 17, 2008 (Edited by D. Shibata)

Abstract Recently, high-throughput technologies for comprehensive genomic, transcriptomic and proteomic analyses have been developed. Large-scale 'omics data' obtained from such experimental methods have rapidly accumulated and been stored into web databases. In addition, the number of web databases for molecular biology has rapidly increased, since most new projects construct and maintain new web databases for their own large-scale omics data. This wealth of comprehensive online resources and web databases allow us to extract essential new biological information beyond a dataset obtained from an individual study. For effective and efficient handling of such large sets of data, various kinds of bioinformatics tools are being developed. Here, we review the current state of web databases and bioinformatics tools for plant biosciences and systems biology.

Key words: Bioinformatics, database, genome sequence, Internet, omics.

Recently, high-throughput technologies for comprehensive genomic, transcriptomic and proteomic analyses have been developed. Large-scale 'omics data' obtained from such experimental methods have rapidly accumulated and been stored into web databases. For nucleotide sequence data, the number of entries in the International Nucleotide Sequence Databases (INSD) (Brunak et al. 2002), maintained by DDBJ (Sugawara et al. 2008), EMBL (Cochrane et al. 2008) and GenBank (Benson et al. 2008), has steadily increased. There were 86,099,950,395 bases in 83,167,582 DNA sequence records in DDBJ Release 73.0 as of March 2008. This latest release contains almost twice the number of bases and entries of Release 61.0 (March 2005). The volume of omics data will continue to grow exponentially with further improvements in experimental methodology.

Throughout the last decade, numerous research projects have been launched, including complete genome sequences and functional and structural proteomic analyses. Since most projects construct and maintain web databases to manage their own large-scale data, the number of web databases for molecular biology has also proliferated. The *Nucleic Acids Research* journal's online Molecular Biology Database Collection listed more than 1,000 databases in 2008 (Galperin 2008), nearly quadrupling the 281 databases listed in the Collection in 2001 (Baxevanis 2001).

This wealth of comprehensive online resources and web databases allows us to extract new essential biological information beyond what is gleaned from the dataset obtained from a single study. The large-scale data collected from web databases can be used to detect patterns such as consensus sequences of expressed sequence tags (ESTs) (e.g., Lazo et al. 2004; Lee et al. 2005; Lopez et al. 2004), global patterns of gene expression (Obayashi et al. 2007) and orthologous DNA and protein sequences among different genomes (Tatusov et al. 2003). To handle such large amounts of data effectively and efficiently, many new bioinformatics tools are under development. Here, we review the current state of web databases and bioinformatics tools available for plant research. The websites mentioned in this article are summarized in Table 1.

Collections of databases, tools, and experimental materials

The Molecular Biology Database Collection contains links to web databases described in the annual database issues of *Nucleic Acids Research*. More than 1,000 databases have been listed in the collection to date. Users can browse the list and search databases by the

[a] These authors contributed equally to this review article.

Abbreviations: BAC, bacterial artificial chromosome; CDS, coding sequence; EST, expressed sequence tag; GO, gene ontology; INSD, International Nucleotide Sequence Database; NCBI, National Center for Biotechnology Information; RAP, Rice Annotation Project; SGN, The SOL Genomics Network; SNP, single nucleotide polymorphism; SOL, International Solanaceae Initiative; SSR, simple sequence repeat; TAIR, The Arabidopsis Information Resource; TIGR, The Institute for Genomic Research.

This article can be found at http://www.jspcmb.jp/

Table 1. Web resources for omics data and experimental materials.

Arabidopsis		
AraCyc	Metabolic pathways	http://www.arabidopsis.org/biocyc/index.jsp
ATTED-II	Co-expression	http://www.atted.bio.titech.ac.jp/
CSB.DB—A Comprehensive Systems-Biology Database	Co-expression	http://csbdb.mpimp-golm.mpg.de/index.html
KATANA	Annotations	http://www.kazusa.or.jp/katana/
RAFL	Full-length cDNAs	http://www.brc.riken.jp/lab/epd/catalog/cdnaclone.html
TAIR	Omics database	http://www.arabidopsis.org/
Rice		
The *Oryza* Map Alignment Project (OMAP)	Physical maps of the genus *Oryza*	http://www.omap.org/index.html
GRAMENE	An integrated database of grass species	http://www.gramene.org/
KOME	Full-length cDNAs	http://cdna01.dna.affrc.go.jp/cDNA/
OryzaBASE	A comprehensive database for rice	http://www.shigen.nig.ac.jp/rice/oryzabase/top/top.jsp
OryzaExpress	Annotations and co-expression	http://riceball.lab.nig.ac.jp/oryzaexpress/
RAP-DB	Genome annotations	http://rapdb.dna.affrc.go.jp/
RiceCyc	Metabolic pathways	http://www.gramene.org/pathway/ricecyc.html
TIGR	Genome annotations	http://www.tigr.org/
Tomato		
MiBASE	ESTs, unigenes, metabolic pathways	http://www.kazusa.or.jp/jsol/microtom/index.html
KaFTom	Full-length cDNAs	http://www.pgb.kazusa.or.jp/kaftom/
SGN	An integrated database of Solanaceae	http://www.sgn.cornell.edu/index.pl
LycoCyc	Metabolic pathways	http://www.gramene.org/pathway/lycocyc.html
Legumes and other plants		
Cassava full-length cDNA clone	Full-length cDNAs	http://www.brc.riken.jp/lab/epd/Eng/list/index.shtml
JGI	An integrated database of genomics	http://www.jgi.doe.gov/index.html
Legume base	Experimental materials	http://www.shigen.nig.ac.jp/bean/lotusjaponicus/top/top.jsp
Lotus japonicus	*Lotus japonicus* genome browser	http://www.kazusa.or.jp/lotus/index.html
Poplar full-length cDNA clone	Full-length cDNAs	http://www.brc.riken.jp/lab/epd/Eng/list/index.shtml
SABRE	An integrated database of plant resources	http://saber.epd.brc.riken.jp/sabre7/SABRE0101.cgi
Structure, function and expression of genes and proteins		
ArrayExpress	Gene expressions	http://www.ebi.ac.uk/microarray-as/aer/
BioCyc	Metabolic pathways	http://biocyc.org/
dbEST	ESTs	http://www.ncbi.nlm.nih.gov/dbEST/
Gene Ontology	GO	http://www.geneontology.org/
GeneIndex	ESTs and unigenes	http://compbio.dfci.harvard.edu/tgi/
GEO	Gene expression	http://www.ncbi.nlm.nih.gov/geo/
miRBase	microRNA	http://microrna.sanger.ac.uk/sequences/index.shtml
NASCarrays	Microarray data	http://affymetrix.arabidopsis.info
Plant Ontology	Plant Ontology	http://www.plantontology.org/
PRIME	Systems biology	http://prime.psc.riken.jp/
UniGene	ESTs and unigenes	http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene
DB links		
Molecular Biology Database Collection	Links of web databases	http://www.oxfordjournals.org/nar/database/c/
Bioinformatics Links Directory	Web-based software resources	http://bioinformatics.ca/links_directory/
SHIGEN	Experiment materials and databases	http://www.shigen.nig.ac.jp/
IUBMB	EC number	http://www.chem.qmul.ac.uk/iubmb/

categories shown in the website. The collection contains databases for model plants, including *Arabidopsis thaliana*, rice, *Brassica*, legumes, maize, *Medicago truncatula*, tomato and potato.

The annual web server issues of *Nucleic Acids Research* report web-based software resources for analysis of molecular biology data and provide links to them. The Bioinformatics Links Directory also has added links to the web servers described in the issues. The links are categorized by subjects (e.g., DNA, RNA, protein, expression) and are searchable. The compilation contained about 1,200 web servers as of 2007.

The SHared Information of GENetic resources (SHIGEN) Project has maintained a web server that provides information about experimental materials and databases. The experimental materials (live animal stocks, frozen embryos, plant seeds, cultured cells, DNA clones, etc.) described in the databases are available to researchers upon request.

Genome annotations and comparative genomics for model plants

TAIR

The *Arabidopsis* Information Resource (TAIR) maintains a database for the fully sequenced and extensively studied model plant *Arabidopsis* (Rhee et al. 2003). In addition to the completed and fully annotated *Arabidopsis* genome sequence, the database contains information about genes, gene expression, clones, nucleotide sequences, DNA markers, mutants, seed stocks, members of the *Arabidopsis* research community, and research papers. TAIR database contains various web pages for efficient and intuitive querying and browsing of datasets, graphical and interactive map viewers for genes and the genome, and downloading of data stored in the TAIR database. TAIR also includes the metabolic pathway database 'AraCyc' (Mueller et al. 2003).

The TAIR website is updated every two weeks with new information from research publications and community data submissions. Gene structures are also updated a couple of times per year using computational and manual methods as well as community submissions of new and updated genes. The latest version of the *Arabidopsis* genome annotation (TAIR8) was released at TAIR and NCBI.

The browser is simple and user-friendly. A user can retrieve information of interest after navigating only a few pages. By clicking a tab of interest such as 'search', 'tools' or 'stocks' in the top page, researchers can get basic information, with more detailed information accessible via hyperlinks to internal and external web pages.

RAP-DB

After sequencing of the rice genome was completed in 2004 (International Rice Genome Sequencing Project, 2005), the Rice Annotation Project (RAP) was launched to provide accurate and timely gene annotation of sequences of the rice genome. The RAP collaborated closely with the International Rice Genome Sequencing Project (IRGSP). Genome annotations stored in the database called the 'RAP-DB' include information about nucleotide and protein sequences, structures and functions of genes, gene families, RNA genes detected by massively parallel signature sequencing (MPSS), transposable elements, small RNAs detected from the microRNA database 'miRBase' and mutant line resources (Rice Annotation Project, 2008). In a continuing effort to update genome annotations, RAP Annotations (release 2), which contains information about 31,439 genes among the sequences identified in the IRGSP (version build 4 assembly), is now available to researchers. Search functions in the RAP-DB (e.g., BLAST, BLAT, keyword searches, the genome browser 'GBrowse') have also been improved. The RAP helps us to find gene families and genes whose biological functions have not yet been elucidated. It is noteworthy that one function of the RAP, called the 'identifier (ID) converter', allows us to retrieve IDs that are assigned by RAP and another rice annotation project, the 'TIGR Rice Genome Annotation Project' described below. For example, using the ID converter we can determine that the sequence Os01g0100100 in the RAP-DB and sequences LOC_Os01g01010.1 and LOC_Os01g01010.2 in the TIGR database are in fact assigned to the same gene.

TIGR and its rice genome database

The Institute for Genomic Research (TIGR) is one of the legacy organizations of the J. Craig Venter Institute founded in October 2006. The objective of TIGR is to collect various kinds of data, such as DNA and protein sequences, gene expression, cellular roles, protein families and taxonomic data for microbes, animals, humans, plants and other eukaryotes. Genomic databases for castor bean, wheat, *Arabidopsis*, rice, potato, maize, loblolly pine and *Medicago* have been established, and information based on microarray technologies is also now available for *Arabidopsis*, rice and maize. The goal of TIGR is to facilitate experimental validation of all gene predictions and to assign functional roles of those genes using microarray technology as a key tool. TIGR's outstanding compilation of large alignments among nucleotide and protein sequences makes it suitable for determining primary and secondary structures of genes and proteins.

TIGR maintains the TIGR Rice Genome Annotation Database, which provides information about nucleotide sequences and annotation data of the rice genome (Ouyang et al. 2007). Locus IDs (e.g. LOC_Os01g01010.1) assigned by the database differ from those in the RAP-DB as described above. Moreover, the methods of genome assembly and genome annotation in this database are distinct from those employed by RAP-DB. Consequently, the genome sequences and corresponding transcripts in RAP-DB and TIGR may not always have the same nucleotide sequences.

JGI

The US Department of Energy's Joint Genome Institute (JGI) was established in 1997 to integrate the expertise and resources of DNA sequencing, technology and informatics. The initial objective of the JGI was to generate complete sequences of human chromosomes 5, 16, and 19. Their mission has expanded to include other critical areas of genomics, including non-human sequences. The facility now has the ability to read over three billion nucleotides on a monthly basis. Their current targets include microbial genomes, communities of microbes and multicellular organisms. The community sequencing programs are chosen for other species based on proposals from researchers and peer reviews by outside scientists. A green alga (*Chlamydomonas reinhardtii*), a diatom (*Thalassiosira pseudonana*), the cottonwood tree (*Populus trichocarpa*) and various plant pathogens and agriculturally important plants (such as grape, soybean and poplar) have been included in the JGI's genome sequencing projects.

Solanaceae Genome Project Network

Tomato is a vegetable crop consumed worldwide. It is a model plant of the Solanaceae family, which includes other important crops such as potato, eggplant and pepper.

In 2004, the tomato genome sequencing program was launched by the internationally coordinated International Solanaceae Initiative (SOL) consortium (Mueller et al. 2005). The SOL Genomics Network (SGN) provides information relevant to the tomato genome sequencing project such as linkage maps containing DNA markers, bacterial artificial chromosome (BAC) clones anchored to these linkage maps (called 'seed BAC' clones) and BAC sequences with genomic and functional annotations.

Legumes

Lotus japonicus is a model legume that has a short life cycle (2–3 months) and self-fertilizes. Recently, structural features of *L. japonicus* were reported (Sato et al. 2008). The Kazusa DNA Research Institute has constructed a database that provides nucleotide sequences and annotations of its pseudomolecules as well as DNA markers and genetic linkage maps.

GRAMENE

GRAMENE is an integrated database for the genetics, genomics and comparative genomics of grasses, including rice, maize, rye, sorghum, wheat, and other close relatives (Liang et al. 2008). The data in GRAMENE (Release 27 in April 2008) include DNA/RNA sequences, functional annotations of genes and proteins, gene ontology (GO) (Gene Ontology Consortium, 2008), genetic and physical maps/markers, quantitative trait loci (QTLs), pseudomolecule assembly, genetic diversity among germplasms, and comparative genetics/genomics between rice and its wild relatives. To maximize its utility for comparative genomics, GRAMENE also provides a genome level comparison between rice and *Arabidopsis*. These data are shown in a chromosome map together.

In addition to the comprehensive characteristics mentioned above, it is noteworthy that information about QTLs and GO terms in rice has accumulated substantially in the GRAMENE database. The rice metabolic pathway database RiceCyc is also accessible via GRAMENE (Jaiswal et al. 2006). In RiceCyc, locus identifiers employed in the TIGR Rice Genome Annotation Database are assigned.

Since web pages within GRAMENE are easy to access by hyperlinks, detailed data can be obtained by general or in-depth searching. It also has the user-friendly feature of being available for download and local installation so that GRAMENE's data and tools can be customized to suit researcher's requirements.

Expressed sequence tags and unigenes

Information about expressed sequence tags (ESTs) and a non-redundant sequence set derived from these ESTs are provided by INSD and other public databases. ESTs generated from cDNA libraries give us information about transcript sequences and expression patterns in tissues and organs at various developmental stages (Ewing et al. 1999).

Databases dbEST (Boguski et al. 1993) and UniGene (Wheeler et al. 2003) in GenBank provide information about ESTs from a number of organisms, including some plants. The UniGene database provides a list of accession numbers including ESTs that appear to come from the same locus. In the database, information is also available about protein similarities, gene expression, cDNA clone reagents, and genomic locations.

The availability of non-redundant consensus sequences obtained from ESTs allows us to use computational approaches such as homology searches and protein domain searches. The Dana-Farber Cancer Institute (DFCI) has released 'Gene Indices', databases that provide information about a non-redundant consensus sequence set called a 'tentative consensus' (TC) generated by assembling and clustering methods (Lee et al. 2005). In the databases, sequences of TCs are available together with their variants and functional and structural annotations. The detailed information contains homologous protein sequences, open reading frames (ORFs), GO terms, single nucleotide polymorphisms (SNPs), alternative splicing sequences, cDNA libraries, Enzyme Commission (EC) numbers by the International Union of Biochemistry and Molecular Biology (IUBMB), names of KEGG metabolic pathways (Okuda et al. 2008), unique 70-mer oligonucleotide sequences, and orthologs in other organisms.

One should keep in mind when using databases that methods for cataloguing and indexing sequences are still undergoing improvement. In addition, NCBI's UniGene does not provide consensus unigene sequences. Instead, UniGene is updated weekly or monthly, thus providing more recent entry information (accession numbers) for each unigene (cluster). Despite these issues, unigene sequences and a list of accession numbers in the same cluster serve as a very useful basis for analyzing ORFs, sequence similarities and functional domains in protein sequences.

Unigene databases for each species have been also constructed and maintained. For example, the tomato unigene database MiBASE provides information about unigenes obtained by assembling ESTs in tomato and a wild relative as well as information about SNPs, simple sequence repeats (SSRs), GO terms, metabolic pathway names, and gene expression data (Yamamoto et al. 2005, Yano et al. 2006a, Yano et al. 2006b).

Full-length cDNAs

Full-length cDNA clones are fundamental resources in molecular biology for experimental investigations of gene function as well as for detection of intron-exon structures in genes.

The RAFL database

Information about RIKEN *Arabidopsis* full-length (RAFL) cDNA clones is accessible from the RAFL database (Seki et al. 2002). Currently, 251,382 full-length cDNA clones are available in the database.

KOME

The Knowledge-based *Oryza* Molecular biological Encyclopedia (KOME) is a database that has collected information about full-length cDNAs of japonica rice (Rice Full-Length cDNA Consortium, 2003). In this database, 170,000 full-length cDNA clones have been grouped into 28,000 independent groups. All the representative clones have been completely sequenced.

KaFTom

A miniature tomato cultivar Micro-Tom has attracted attention as a model plant (Meissner et al. 1997) that can be cultivated even in ordinary laboratory spaces (Shibata 2005). Full-length cDNA libraries have been constructed from the fruit at different developmental stages and from pathogen-treated leaves of Micro-Tom. To date, information about 57,422 ESTs and 2,268 draft full-length sequences have been made available in the KaFTom database (Aoki et al. 2007; Tsugane et al. 2005; Yano et al. 2007).

Poplar and cassava

We can obtain information about full-length cDNA clones of poplar and cassava from the respective database (Table 1). The current versions of these databases provide information about 4,522 and 19,980 full-length cDNA clones in poplar and cassava, respectively.

Gene expression databases

GEO

The Gene Expression Omnibus (GEO) is a gene expression and molecular abundance repository at NCBI, supporting MIAME (Minimum Information About a Microarray Experiment) compliant data (Edgar and Barrett 2006). It covers a wide range of high-throughput experiments, including single- and dual-channel microarray for measuring mRNA, miRNA, DNA, Chip, SNP and protein abundance, as well as non-array based technologies such as serial analysis of gene expression (SAGE) and mass spectrometry peptide profile.

ArrayExpress

ArrayExpress is a public repository at the European Bioinformatics Institute (EBI) in the EMBL for transcriptomics, including gene expression data and data from comparative genomic hybridizations and chromatin immunoprecipitations (Parkinson et al. 2007). It encompasses most model species, including human, *Mus musculus*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Drosophila melanogaster* and *Arabidopsis*. The expression profiles for all species are stored in the Data Warehouse database and collaborates with the GEO for some experiments. ArrayExpress can be queried by keywords, for example gene, sample or experiment. Results of a query can be obtained as a graph of gene expression profiles. Recently a new component, the ArrayExpress Atlas of Gene Expression, has been developed for summary statistics based on meta-analyses. It can retrieve results for a condition-specific gene expression profile and biological interests of the researcher over the entire ArrayExpress Data Warehouse database.

ATTED-II

ATTED-II is a specialized database for the prediction of trans-factors and *cis*-elements in *Arabidopsis* (Obayashi et al. 2007). Based on the repositories of microarray data in TAIR and the Nottingham *Arabidopsis* Stock Centre Arrays (NASCArrays), ATTED-II predicts co-expressed gene networks that are estimated to be involved in the same and/or related biological pathways and any *cis*-elements that may exist up to 200 bp upstream of the transcriptional starting point. Although some databases for co-expression analysis, such as the Comprehensive Systems-Biology Database (CSB.DB), Botany Array Resource (BAR), *Arabidopsis* Co-expression Tool (ACT) and Genevestigator, have been developed, one helpful diagnostic feature of the ATTED-II is the visual representation of the results, a picture of the gene network and graphs of *cis*-element and gene expression. Because this kind of analysis is based on an enormous quantity of data, without such visual representation it is difficult for an experimental researcher to follow all the processes of an analysis and to surmise what the data indicate. A graphical display of the analysis makes it easier to see relationships and expression profiles of genes of interest.

OryzaExpress

OryzaExpress is a database providing functional annotations of genes, reaction names of metabolic pathways, locus IDs assigned from RAP and TIGR, and gene expression profiles in rice. Expression data from the GEO database are imported into OryzaExpress. The expression data were obtained from microarray platforms provided by the Affymetrix Rice Genome Array (http://www.affymetrix.com/products/arrays/specific/rice.affx) and Agilent Rice Oligo Microarray (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL892). Similar to ATTED-II, OryzaExpress provides a predicted co-expressed gene set detected from expression data by the Agilent Rice Oligo Microarray platform.

Searching by keywords or IDs in OryzaExpress, users can find genes, reaction names of metabolic pathways in KEGG and RiceCyc, locus IDs in RAP and TIGR and probe names of the Affymetrix Rice Genome Array and Agilent Rice Oligo Microarrays (22K). Although locus IDs and probe names for the same genes are different among the public databases and microarray platforms, OryzaExpress allows us to obtain simultaneously the information retrieved from distinct public databases.

Functional categories of genes and proteins

Gene Ontology and GO Slim

The GO project is a collaborative effort aiming to provide consistent descriptions of gene products in different databases (Gene Ontology Consortium, 2008). It was launched in 1998 as a collaboration of databases between *Drosophila*, *Saccharomyces* and Mouse. Since then the GO consortium has grown to include many other animal, microbe and plant databases. TAIR, TIGR and GRAMENE plant databases are members of the consortium at this time. The GO project describes gene products in terms of their biological role as a cellular component, a biological process or a molecular function. Each entry in GO has also a numerical ID and term name, such as “cell”, “signal transduction”, and “catalytic activity”. By matching corresponding terms and GO principles, GO facilitates uniform queries across collaborating databases so that genes can be queried at different levels.

GO is neither a database nor a catalog of gene sequences/products, so if researcher needs a list of genes or gene products that have been assigned with a particular GO term, the researcher should go to database links provided in the Current Annotations Table within GO.

GO Slim is a minimal version of GO that gives an overview of the ontology content. GO Slim is particularly useful for obtaining a summary of GO annotations.

KOG

The euKaryotic Orthologous Groups (KOG) is a eukaryote-specific version of the Clusters of Orthologous Groups (COG) (Tatusov et al. 2003), a phylogenetic classification database of proteins encoded in complete genomes. There are data sets for animals (including humans), plants and unicellular microbes. For plants, data for *Arabidopsis* have been updated. In the KOG browser, orthologous or paralogous proteins are assigned KOG IDs and are classified into four functional groups, “cellular processes and signaling”, “information storage and processing”, “metabolism”, or “poorly characterized”. Within each classification group, orthologous or paralogous proteins are listed. By clicking each KOG of interest, one can find detailed information for genes predicted by JGI (see above for details), transcripts, and descriptions.

Metabolic pathways

KEGG

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a bioinformatics database for predicting a higher-level complexity of cellular processes and organism behaviors from genomic and molecular information (Okuda et al. 2008). It consists of five categories: Atlas, PATHWAY, BRITE, GENES, and LIGAND. Each of these categories has subcategories for more detailed analysis. The KEGG Atlas is a graphical interface for the PATHWAY and BRITE databases and is suitable for use as a metabolic pathway map. KEGG PATHWAY is a collection of manually drawn pathway maps of molecular interactions and reaction networks. KEGG BRITE is a hierarchical classification database of various aspects of biological function. KEGG GENES is a database of gene catalogs for complete genomes, partial genomes and ESTs. KEGG LIGAND is for identification of chemical substances and reactions; it is composed of five other databases, COMPOUND for chemical compound structure, DRUG for drug structure, GLYCAN for glycan structure, REACTION for biochemical reaction, RPAIR for reactant pair alignment and ENZYME for enzyme nomenclature. For plant species, information about model plants can be referenced in the KEGG with a collaboration of other databases, such as NCBI and TIGR.

BioCyc

BioCyc is a collection of 371 pathway/genome databases for visualization of metabolic pathways and metabolic maps of an organism, as well as for an analysis of gene expression, proteomics and metabolomics. The BioCyc databases consist of three categories, each designated as a tier in the website according to the quality of data it contains. Tier 1 incorporates intensively supervised databases. At this time, Tier 1 includes two databases: EcoCyc for *Escherichia coli* K-12 and MetaCyc for metabolic pathways and enzymes from more than 900 organisms. Tier 2 contains computationally-derived databases that are subjected to moderate reviewing. Tier 2 currently consists of 20 databases. Tier 3 includes computationally-derived databases that are not reviewed. Tier 3 currently consists of 349 databases. BioCyc also includes other pathway/genome databases on the Internet: AraCyc for *Arabidopsis*, MedicCyc for *Medicago truncatula*, RiceCyc for rice and SolCyc for Solanaceae.

Methods for large-scale analyses of omics data

High-throughput sequencing

DNA sequencing is a technology for determining the nucleotide order within DNA fragments. Sequences are based largely on sequencing methods developed by Frederick Sanger in 1977; some methodologies have been developed specifically for Sanger sequencing. Recently, a novel technology, MPSS technology or Pyrosequencing, has been developed commercially. Several highly innovative high-throughput sequencing products now available are the Solexa, GS-FLX, and SOLiD analyzer platforms; others will be released in the near future. The systems exceed Sanger sequencing in their ability to produce nucleotide sequence data of 400,000 to 2 billion nucleotides per operation. This means that rapid genome sequencing will enable us to conduct many kinds of research from the micro- to the macro-level, and it will also be possible to sequence genomes among varieties, subspecies, and ecotypes. A good demonstration of this system's utility was the whole genome sequencing of scientist James Watson (Wheeler et al. 2008). In plant research, this methodology will facilitate further studies of genome sequences and DNA polymorphisms among different genomes.

Tiling array in Arabidopsis

The Tiling Array is a type of microarray chip in which short DNA fragments are designed to cover the whole genome, allowing the Tiling Array to investigate gene expression, genome structure and protein binding on a genome-wide scale. The procedure is similar to traditional microarray technology, but it differs in a variety of objectives: unbiased gene expression profile, transcriptome mapping, chromatin immunoprecipitation (Chip)-chip, Methyl-DNA immunoprecipitation (MeDIP)-chip and DNase Chip (e.g., Gregory et al. 2008, Stolc et al. 2005). Although it still has problems of operating cost and signal sensitivity, it will likely become one of the preferred tools for genome-wide investigation in the future.

CA method for large-scale microarray data

Since high-throughput technologies have rapidly developed, the sizes of omics datasets have continued to increase. Computer analyses of the larger datasets are still too time-consuming and impractical. For example, in microarray data analyses, hierarchical clustering methods have been widely used to cluster genes (probes) according to their expression profiles (Eisen et al. 1998) and construct dendrograms and expression maps for identification of significantly different expression patterns against other genes. However, a dataset obtained from Agilent Oligo Microarrays (4×44K) that includes results of several experimental treatments and replications would be too large to be handled by a general computer memory. Yano et al. (2006c) developed a high-throughput gene discovery method based on correspondence analysis (CA) with a new index for expression ratios [arctan (1/ratio)] and three artificial marker genes. This method allows us to quickly analyze a large-scale microarray dataset to identify up- or down-regulated genes related to a trait of interest. They also have developed and distributed a software tool for the calculation. Further development of various bioinformatics tools for large-scale omics data will accelerate the identification of novel, important genes, gene products and metabolic pathways.

Goal to global understanding of biological events

Systems biology

Program for Research on Immune Modeling and Experimentation (PRIME) is a free and user-friendly repository for biological pathways. An image of gene network pathways can be produced by CellDesigner version 4.0 software (http://www.celldesigner.org/) and converted into a user-determined publication style by a simple web-accessible application called BioPP (Biological Pathway Publisher) available in the PRIME website.

Oryzabase

Oryzabase is a comprehensive database for rice research established in 2000 by a committee of rice researchers in Japan (Kurata and Yamazaki 2006). It includes a variety of resources: genetic lines and wild germplasms collected from all over the world, mutant lines classified based on tissue in which the mutated genes disrupt the phenotype, genes annotated by PAP-DB and TIGR, literature references, linkage maps that are integrated into reference maps among ssp. *japonica* and between *japonica* and *indica*, physical maps of SHIGEN and TIGR databases, comparative maps between rice and other grass species, a DNA and organelle database, and analytical tools and protocols. It also has links with other databases such as RAP-DB, KOME, TIGR, and GRAMENE. Oryzabase would be the best choice when a researcher needs a fundamental database for analysis in rice research.

The Oryza Map Alignment Project (OMAP)

With the completion of the full sequencing of the rice genome, the DNA fingerprint/BAC-end sequencing of eleven wild relatives of rice was accelerated by a collaboration of the Arizona Genomics Institute, Arizona Genomics Computational Lab, Cold Spring Harbor Lab and Purdue University (Wing et al. 2005). The main objective of OMAP is to develop a database for physical maps of genomes of 11 wild species in the *Oryza* genus, namely *O. rufipogon*, *O. glaberrima*, *O. punctata*, *O. officinalis*, *O. minuta*, *O. australiensis*, *O. latifolia*, *O. schlechteri*, *O. ridleyi*, *O. brachyantha* and *O. granulata*. The database includes: 1) an alignment of maps between *japonica* and *indica* subspecies of *Oryza sativa*, 2) construction of high resolution physical maps of chromosomes 1, 3 and 10 across the 11 wild rice species, and 3) development of convenient bioinformatics and educational tools for understanding an *Oryza* genome. These achievements provide an experimental system for understanding the evolutionary history of the *Oryza* genome, including synteny, rearrangement and species-specific insertion and/or deletion. It will contribute to our understanding of how the *Oryza* species has adapted to diverse ecological habitats from tropical regions to upland over the course of evolution.

Perspectives

With the completion of full-genome nucleotide sequencing of the model plants, there is now a great need for systems that analyze these nucleotide sequence data in a public database. In addition, with improvements in technology for molecular biology, various kinds of data such as cDNA, EST, DNA markers and microarrays have been produced all over the world. These developments mean that an integrated assembly system for web databases is required.

What is the reality? Such data are still decentralized into individual databases developed by each research group. Thus, the omics research arena is still in flux, and this complicated situation sometimes makes it difficult for researchers to maximize productivity using the vast omics data available. Still, one's luck changes with each bit of data added to the databases. Since any finding contributes to progress and may do so at a tipping point, any such finding can shift the equilibrium and accelerate the rate at which valuable data accumulate for a particular subdiscipline.

Although the research environment for plant biology is improving, there still seems to be a large gap between a single experimental study and bioinformatics. That is a challenge of omics in plant biology. Bioinformatics researchers need to have knowledge of traditional plant biology, and experimental plant biologists need to understand bioinformatics. In reality, this is difficult to achieve. To bridge such gaps and to fully exploit the available data, all researchers in every specialized field need to contribute to such collaborative projects in order for plant biology to advance.

Acknowledgements

This work was supported in part by Grants-in-Aid for Special Research on Priority Areas (19043015) to KY from the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT). This research was also supported in part by Grants-in-Aid for Exploratory Research (19651084) to KY, Grant-in-Aid for Scientific Research (B) (20380022) to KY and for Young Scientists (Start-up) (19880004) to KS from the Japan Society for Promotion of Science (JSPS). KS is the recipient of the JSPS Research Fellowship for Young Scientists.

References

Aoki K, Yano K, Sakurai N, Tsugane T, Watanabe M, Yin YG, Matsukura C, Shibata D (2007) Expression-based approach toward functional characterization of tomato genes that have no or weak similarity to *Arabidopsis* genes. *Acta Hortic* 745: 457–464

Baxevanis AD (2001) The Molecular Biology Database Collection: an updated compilation of biological database resources. *Nucl Acids Res* 29: 1–10

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2008) GenBank. *Nucl Acids Res* 36: D25–30

Boguski MS, Lowe TM, Tolstoshev CM (1993) dbEST—database for "expressed sequence tags". *Nat Genet* 4: 332–333

Brunak S, Danchin A, Hattori M, Nakamura H, Shinozaki K, Matise T, Preuss D (2002) Nucleotide Sequence Database Policies. *Science* 298: 1333

Cochrane G, Akhtar R, Aldebert P, Althorpe N, Baldwin A, Bates K, Bhattacharyya S, Bonfield J, Bower L, Browne P, Castro M, Cox T, Demiralp F, Eberhardt R, Faruque N, Hoad G, Jang M, Kulikova T, Labarga A, Leinonen R, Leonard S, Lin Q, Lopez R, Lorenc D, McWilliam H, Mukherjee G, Nardone F, Plaister S, Robinson S, Sobhany S, Vaughan R, Wu D, Zhu W, Apweiler R, Hubbard T, Birney E (2008) Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucl Acids Res* 36: D5–12

Edgar R, Barrett T (2006) NCBI GEO standards and services for microarray data. *Nat Biotechnol* 24: 1471–1472

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95: 14863–14868

Ewing RM, Ben Kahla A, Poirot O, Lopez F, Audic S, Claverie JM (1999) Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res* 9: 950–959

Galperin MY (2008) The Molecular Biology Database Collection: 2008 update. *Nucl Acids Res* 36: D2–4

Gene Ontology Consortium (2008) The Gene Ontology project in 2008. *Nucl Acids Res* 36: D440–444

Gregory BD, Yazaki J, Ecker JR (2008) Utilizing tiling microarrays for whole-genome analysis in plants. *Plant J* 53: 636–644

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* 436: 793–800

Jaiswal P, Ni J, Yap I, Ware D, Spooner W, Youens-Clark K, Ren L, Liang C, Zhao W, Ratnapu K, Faga B, Canaran P, Fogleman M, Hebbard C, Avraham S, Schmidt S, Casstevens TM, Buckler ES, Stein L, McCouch S (2006) Gramene: a bird's eye view of cereal genomes. *Nucl Acids Res* 34: D717–723

Kurata N, Yamazaki Y (2006) Oryzabase. An integrated biological and genome information database for rice. *Plant Physiol* 140: 12–17

Lazo GR, Chao S, Hummel DD, Edwards H, Crossman CC, Lui N, Matthews DE, Carollo VL, Hane DL, You FM, Butler GE, Miller RE, Close TJ, Peng JH, Lapitan NL, Gustafson JP, Qi LL, Echalier B, Gill BS, Dilbirligi M, Randhawa HS, Gill KS, Greene RA, Sorrells ME, Akhunov ED, Dvorák J, Linkiewicz AM, Dubcovsky J, Hossain KG, Kalavacharla V, Kianian SF, Mahmoud AA, Miftahudin, Ma XF, Conley EJ, Anderson JA, Pathan MS, Nguyen HT, McGuire PE, Qualset CO, Anderson OD (2004) Development of an expressed sequence tag (EST) resource for wheat (Triticum aestivum L.): EST generation, unigene analysis, probe selection and bioinformatics for a 16,000-locus bin-delineated map. *Genetics* 168: 585–593

Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J (2005) The TIGR gene indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucl Acids Res* 33: D71–74

Liang C, Jaiswal P, Hebbard C, Avraham S, Buckler ES, Casstevens T, Hurwitz B, McCouch S, Ni J, Pujar A, Ravenscroft D, Ren L, Spooner W, Tecle I, Thomason J, Tung CW, Wei X, Yap I, Youens-Clark K, Ware D, Stein L (2008) Gramene: a growing plant comparative genomics resource. *Nucl Acids Res* 36: D947–953

Lopez C, Jorge V, Piégu B, Mba C, Cortes D, Restrepo S, Soto M, Laudié M, Berger C, Cooke R, Delseny M, Tohme J, Verdier V. (2004) A unigene catalogue of 5700 expressed genes in cassava. *Plant Mol Biol* 56: 541–554

Meissner R, Jacobson Y, Melamed S, Levyatuv S, Shalev G, Ashri A, Elkind Y, Levy A (1997) A new model system for tomato genetics. *Plant J* 12: 1465–1472

Mueller LA, Zhang P, Rhee SY (2003) AraCyc: a biochemical pathway database for *Arabidopsis*. *Plant Physiol* 132: 453–460

Mueller LA, Tanksley SD, Giovannoni JJ, Van Eck J, Stack S, Choi D, Kim BD, Chen M, Cheng Z, Li C, Ling H, Xue Y, Seymour G, Bishop G, Bryan G, Sharma R, Khurana J, Tyagi A, Chattopadhyay D, Singh NK, Stiekema W, Lindhout P, Jesse T, Lankhorst RK, Bouzayen M, Shibata D, Tabata S, Granell A, Botella MA, Giuliano G, Frusciante L, Causse M, Zamir D (2005) The Tomato Sequencing Project, the first cornerstone of the International Solanaceae Project (SOL). *Comparative and Functional Genomics* 6: 153–158

Obayashi T, Kinoshita K, Nakai K, Shibaoka M, Hayashi S, Saeki M, Shibata D, Saito K, Ohta H (2007) ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. *Nucl Acids Res* 35: D863–869

Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, Bork P, Goto S, Kanehisa M. (2008) KEGG Atlas mapping for global analysis of metabolic pathways. *Nucl Acids Res* 36: 423–426

Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, Orvis J, Haas B, Wortman J, Buell CR (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucl Acids Res* 35: D883–887

Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, Mani R, Rayner T, Sharma A, William E, Sarkans U, Brazma A (2007) ArrayExpress—a public database of microarray experiments and gene expression profiles. *Nucl Acids Res* 35: D747–750

Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, Miller N, Mueller LA, Mundodi S, Reiser L, Tacklind J, Weems DC, Wu Y, Xu I, Yoo D, Yoon J, Zhang P (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucl Acids Res* 31: 224–228

Rice Annotation Project (2008) The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucl Acids Res* 36: D1028–1033

Rice Full-Length cDNA Consortium (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science* 301: 376–379

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 74: 5463–5467

Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K, Fujishiro T, Katoh M, Kohara M, Kishida Y, Minami C, Nakayama S, Nakazaki N, Shimizu Y, Shinpo S, Takahashi C, Wada T, Yamada M, Ohmido N, Hayashi M, Fukui K, Baba T, Nakamichi T, Mori H, Tabata S (2008) Genome Structure of the Legume, *Lotus japonicus*. *DNA Res* 2008 May 28. [Epub ahead of print].

Seki M, Narusaka M, Kamiya A, Ishida J, Satou M, Sakurai T, Nakajima M, Enju A, Akiyama K, Oono Y, Muramatsu M, Hayashizaki Y, Kawai J, Carninci P, Itoh M, Ishii Y, Arakawa T, Shibata K, Shinagawa A, Shinozaki K (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science* 296: 141–145

Shibata D (2005) Genome sequencing and functional genomics approaches in tomato. *J Gen Plant Pathol* 71: 1–7

Stolc V, Samanta MP, Tongprasit W, Sethi H, Liang S, Nelson DC, Hegeman A, Nelson C, Rancour D, Bednarek S, Ulrich EL, Zhao Q, Wrobel RL, Newman CS, Fox BG, Phillips GN Jr, Markley JL, Sussman MR (2005) Identification of transcribed sequences in *Arabidopsis thaliana* by using high-resolution genome tiling arrays. *Proc Natl Acad Sci USA*102: 4453–4458

Sugawara H, Ogasawara O, Okubo K, Gojobori T, Tateno Y (2008) DDBJ with new system and face. *Nucl Acids Res* 36: D22–24

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4: 41

Tsugane T, Watanabe M, Yano K, Sakurai N, Suzuki H, Shibata D (2005) Expressed sequence tags of full-length cDNA clones from the miniature tomato (*Lycopersicon esculentum*) cultivar Micro-Tom. *Plant Biotechnol* 22: 161–165

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452: 872–876

Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L (2003) Database Resources of the National Center for Biotechnology. *Nucl Acids Res* 31: 28–33

Wing RA, Ammiraju JS, Luo M, Kim H, Yu Y, Kudrna D, Goicoechea JL, Wang W, Nelson W, Rao K, Brar D, Mackill DJ, Han B, Soderlund C, Stein L, SanMiguel P, Jackson S (2005) The oryza map alignment project: the golden path to unlocking the genetic potential of wild rice species. *Plant Mol Biol* 59: 53–62

Yamamoto N, Tsugane T, Watanabe M, Yano K, Maeda F, Kuwata C, Torki M, Ban Y, Nishimura S, Shibata D (2005) Expressed sequence tags from the laboratory-grown miniature tomato (*Lycopersicon esculentum*) cultivar Micro-Tom and mining for single nucleotide polymorphisms and insertions/deletions in tomato cultivars. *Gene* 356: 127–134

Yano K, Watanabe M, Yamamoto N, Tsugane T, Aoki K, Sakurai M, Shibata D (2006a) MiBASE: A database of a miniature tomato cultivar Micro-Tom. *Plant Biotechnol* 23: 195–198

Yano K, Tsugane T, Watanabe M, Maeda F, Aoki K, Shibata D (2006b) Non-biased distribution of tomato genes with no counterparts in *Arabidopsis thaliana* in expression patterns during fruit maturation. *Plant Biotechnol* 23: 199–202

Yano K, Imai K, Shimizu A, Hanashita T (2006c) A new method for gene discovery in large-scale microarray data. *Nucl Acids Res* 34: 1532–1539

Yano K, Aoki K, Shibata D (2007) Genomic Databases for Tomato. *Plant Biotechnol* 24: 17–25