Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

Bioinformatics and Its Applications in Plant Biology

Seung Yon Rhee,[1] Julie Dickerson,[2]
and Dong Xu[3]

[1]Department of Plant Biology, Carnegie Institution, Stanford, California 94305;
email: rhee@acoma.stanford.edu

[2]Baker Center for Computational Biology, Electrical and Computer Engineering,
Iowa State University, Ames, Iowa 50011-3060; email: julied@iastate.edu

[3]Digital Biology Laboratory, Computer Science Department and Life Sciences
Center, University of Missouri-Columbia, Columbia, Missouri 65211-2060;
email: xudong@missouri.edu

Annu. Rev. Plant Biol.
2006. 57:335–60

The *Annual Review of Plant Biology* is online at
plant.annualreviews.org

doi: 10.1146/
annurev.arplant.56.032604.144103

Copyright © 2006 by
Annual Reviews. All rights
reserved

First published online as a
Review in Advance on
February 28, 2006

1543-5008/06/0602-
0335$20.00

Key Words

sequence analysis, computational proteomics, microarray data analysis, bio-ontology, biological database

Abstract

Bioinformatics plays an essential role in today's plant science. As the amount of data grows exponentially, there is a parallel growth in the demand for tools and methods in data management, visualization, integration, analysis, modeling, and prediction. At the same time, many researchers in biology are unfamiliar with available bioinformatics methods, tools, and databases, which could lead to missed opportunities or misinterpretation of the information. In this review, we describe some of the key concepts, methods, software packages, and databases used in bioinformatics, with an emphasis on those relevant to plant science. We also cover some fundamental issues related to biological sequence analyses, transcriptome analyses, computational proteomics, computational metabolomics, bio-ontologies, and biological databases. Finally, we explore a few emerging research topics in bioinformatics.

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

Contents

INTRODUCTION

Recent developments in technologies and instrumentation, which allow large-scale as well as nano-scale probing of biological samples, are generating an unprecedented amount of digital data. This sea of data is too much for the human brain to process and thus there is an increasing need to use computational methods to process and contextualize these data.

Bioinformatics refers to the study of biological information using concepts and methods in computer science, statistics, and engineering. It can be divided into two categories: biological information management and computational biology. The National Institutes of Health (NIH) (**http://www.bisti.nih.gov/**) defines the former category as "research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, represent, describe, store, analyze, or visualize such data." The latter category is defined as "the development and application of data-analytical and theoretical methods, mathematical modeling, and computational simulation techniques to the study of biological, behavioral, and social systems." The boundaries of these categories are becoming more diffuse and other categories will no doubt surface in the future as this field matures.

The intention of this article is not to provide an exhaustive summary of all the advances made in bioinformatics. Rather, we describe some of the key concepts, methods, and tools used in this field, particularly those relevant to plant science, and their current limitations and opportunities for new development and improvement. The first section introduces sequence-based analyses, including gene finding, gene family and phylogenetic analyses, and comparative genomics approaches. The second section presents computational transcriptome analysis, ranging from analyses of various array technologies to regulatory sequence prediction. In section three, we focus on computational proteomics, including gel analysis and protein identification from mass-spectrometry data. Section four describes computational metabolomics. Section five introduces biological ontologies and their applications. Section six addresses various issues related to biological databases

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

ranging from database development to curation. In section seven, we discuss a few emerging research topics in bioinformatics.

SEQUENCE ANALYSIS

Biological sequence such as DNA, RNA, and protein sequence is the most fundamental object for a biological system at the molecular level. Several genomes have been sequenced to a high quality in plants, including *Arabidopsis thaliana* (130) and rice (52, 147, 148). Draft genome sequences are available for poplar (**http://genome.jgi-psf.org/Poptr1/**) and lotus (**http://www.kazusa.or.jp/lotus/**), and sequencing efforts are in progress for several others including tomato, maize, *Medicago truncatula*, sorghum (11) and close relatives of *Arabidopsis thaliana*. Researchers also generated expressed sequence tags (ESTs) from many plants including lotus, beet, soybean, cotton, wheat, and sorghum (see **http://www.ncbi.nlm.nih.gov/dbEST/**).

Genome Sequencing

Advances in sequencing technologies provide opportunities in bioinformatics for managing, processing, and analyzing the sequences. Shotgun sequencing is currently the most common method in genome sequencing: pieces of DNA are sheared randomly, cloned, and sequenced in parallel. Software has been developed to piece together the random, overlapping segments that are sequenced separately into a coherent and accurate contiguous sequence (93). Numerous software packages exist for sequence assembly (51), including Phred/Phrap/Consed (**http://www.phrap.org**), Arachne (**http://www.broad.mit.edu/wga/**), and GAP4 (**http://staden.sourceforge.net/overview.html**). TIGR developed a modular, open-source package called AMOS (**http://www.tigr.org/software/AMOS/**), which can be used for comparative genome assembly (102). Current limitations in shotgun sequencing and assembly software remain largely in the assembly of highly repetitive sequences, although the cost of sequencing is another limitation. Recently developed methods continue to reduce the cost of sequencing, including sequencing by using differential hybridization of oligonucleotide probes (48, 62, 101), polymorphism ratio sequencing (16), four-color DNA sequencing by synthesis on a chip (114), and the "454 method" based on microfabricated high-density picoliter reactors (87). Each of these sequencing technologies has significant analytical challenges for bioinformatics in terms of experimental design, data interpretation, and analysis of the data in conjunction with other data (33).

Gene Finding and Genome Annotation

Gene finding refers to prediction of introns and exons in a segment of DNA sequence. Dozens of computer programs for identifying protein-coding genes are available (150). Some of the well-known ones include Genscan (**http://genes.mit.edu/GENSCAN.html**), GeneMarkHMM (**http://opal.biology.gatech.edu/GeneMark/**), GRAIL (**http://compbio.ornl.gov/Grail-1.3/**), Genie (**http://www.fruitfly.org/seq_tools/genie.html**), and Glimmer (**http://www.tigr.org/softlab/glimmer**). Several new gene-finding tools are tailored for applications to plant genomic sequences (112).

Ab initio gene prediction remains a challenging problem, especially for large-sized eukaryotic genomes. For a typical *Arabidopsis thaliana* gene with five exons, at least one exon is expected to have at least one of its borders predicted incorrectly by the ab initio approach (19). Transcript evidence from full-length cDNA or EST sequences or similarity to potential protein homologs can significantly reduce uncertainty of gene identification (154). Such methods are widely used in "structural annotation" of genomes, which refers to the identification of features such as genes and transposons in a genomic sequence using ab initio algorithms and other

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

information. Several software packages have been developed for structural annotation (3, 45, 57, 66). In addition, one can use genome comparison tools such as SynBrowse (**http://www.synbrowser.org/**) and VISTA (**http://genome.lbl.gov/vista/index.shtml**) to enhance the accuracy of gene identification. Current limitations of structural annotation include accurate prediction of transcript start sites and identification of small genes encoding less than 100 amino acids, noncoding genes (such as microRNA precursors), and alternative splicing sites.

An important aspect of genome annotation is the analysis of repetitive DNAs, which are copies of identical or nearly identical sequences present in the genome (78). Repetitive sequences exist in almost any genome, and are abundant in most plant genomes (69). The identification and characterization of repeats is crucial to shed light on the evolution, function and organization of genomes and to enable filtering for many types of homology searches. A small library of plant-specific repeats can be found at **ftp://ftp.tigr.org/pub/data/TIGR_Plant_Repeats/**; this is likely to grow substantially as more genomes are sequenced. One can use RepeatMasker (**http://www.repeatmasker.org/**) to search repetitive sequences in a genome. Working from a library of known repeats, RepeatMasker is built upon BLAST and can screen DNA sequences for interspersed repeats and low complexity regions. Repeats with poorly conserved patterns or short sequences are hard to identify using RepeatMasker due to the limitations of BLAST. To identify novel repeats, various algorithms were developed. Some widely used tools include RepeatFinder (**http://ser-loopp.tc.cornell.edu/cbsu/repeatfinder.htm**) and RECON (**http://www.genetics.wustl.edu/eddy/recon/**). However, due to the high computational complexity of the problem, none of the programs can guarantee finding all possible repeats as all the programs use some approximations in computation, which will miss some repeats with less distinctive patterns. Inevitably, a combination of repeat finding tools is required to obtain a satisfactory overview of repeats found in an organism.

Sequence Comparison

Comparing sequences provides a foundation for many bioinformatics tools and may allow inference of the function, structure, and evolution of genes and genomes. For example, sequence comparison provides a basis for building a consensus gene model like UniGene (18). Also, many computational methods have been developed for homology identification (136). Although sequence comparison is highly useful, it should be noted that it is based on sequence similarity between two strings of text, which may not correspond to homology (relatedness to a common ancestor in evolution), especially when the confidence level of a comparison result is low. Also, homology may not mean conservation in function.

Methods in sequence comparison can be largely grouped into pair-wise, sequence-profile, and profile-profile comparison. For pair-wise sequence comparison, FASTA (**http://fasta.bioch.virginia.edu/**) and BLAST (**http://www.ncbi.nlm.nih.gov/blast/**) are popular. To assess the confidence level for an alignment to represent homologous relationship, a statistical measure (Expectation Value) was integrated into pair-wise sequence alignments (71). Remote homologous relationships are often missed by pair-wise sequence alignment due to its insensitivity. Sequence-profile alignment is more sensitive for detecting remote homologs. A protein sequence profile is generated by multiple sequence alignment of a group of closely related proteins. A multiple sequence alignment builds correspondence among residues across all of the sequences simultaneously, where aligned positions in different sequences probably show functional and/or structural relationship. A sequence profile is calculated using the probability of

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

occurrence for each amino acid at each alignment position. PSI-BLAST (**http://www.ncbi.nlm.nih.gov/BLAST/**) is a popular example of a sequence-profile alignment tool. Some other sequence-profile comparison methods are slower but even more accurate than PSI-BLAST, including HMMER (**http://hmmer.wustl.edu/**), SAM (**http://www.cse.ucsc.edu/research/compbio/sam.html**), and META-MEME (**http://metameme.sdsc.edu/**). A profile-profile alignment is more sensitive than the sequence-profile-based search programs in detecting remote homologs (146). However, due to its high false positive rate, profile-profile comparison is not widely used. Given potential false positive predictions, it is helpful to correlate the sequence comparison results with the relationship observed in functional genomic data, especially the widely available microarray data as discussed in the section Transcriptome Analysis below. For example, when a gene is predicted to have a particular function through sequence comparison, one can gain confidence in the prediction if the gene has strong correlation in gene expression profile with other genes known to have the same function.

Proteins can be generally classified based on sequence, structure, or function. Several sequence-based methods were developed based on sizable protein sequence (typically longer than 100 amino acids), including Pfam (**http://pfam.wustl.edu/**), ProDom (**http://protein.toulouse.inra.fr/prodom/current/html/home.php**), and Clusters of Orthologous Group (COG) (**http://www.ncbi.nlm.nih.gov/COG/new/**). Other methods are based on "fingerprints" of small conserved motifs in sequences, as with PROSITE (**http://au.expasy.org/prosite/**), PRINTS (**http://umber.sbs.man.ac.uk/dbbrowser/PRINTS/**), and BLOCKS (**http://www.psc.edu/general/software/packages/blocks/blocks.html**). The false positive rate of motif assignment is high due to high probability of matching short motifs in unrelated proteins by chance. Other sequence-based protein family databases are built from multiple sources. InterPro (**http://www.ebi.ac.uk/interpro/**) is a database that integrates domain information from multiple protein domain databases. Using protein family information to predict gene function is more reliable than using sequence comparison alone. On the other hand, very closely related proteins may not guarantee a functional relationship (97). One can use structure- or function-based protein families (when available) to complement sequence-based family for additional function information. SCOP (**http://scop.mrc-lmb.cam.ac.uk/scop/**) and CATH (**http://cathwww.biochem.ucl.ac.uk/**) are the two well-known structure-based family resources. ENZYME (**http://us.expasy.org/enzyme/**) is a typical example of a function family.

A protein family can be represented in a phylogenetic tree that shows the evolutionary relationships among proteins. Phylogenetic analysis can be used in comparative genomics, gene function prediction, and inference of lateral gene transfer among other things (36). The analysis typically starts from aligning the related proteins using tools like ClustalW (**http://bips.u-strasbg.fr/fr/Documentation/ClustalX/**). Among the popular methods to build phylogenetic trees are minimum distance (also called neighbor joining), maximum parsimony, and maximum likelihood trees (reviewed in 31). Some programs provide options to use any of the three methods, e.g., the two widely used packages PAUP (**http://paup.csit.fsu.edu**), and PHYLIP (**http://evolution.genetics.washington.edu/phylip.html**). Although phylogenetic analysis is a research topic with a long history and many methods have been developed, various heuristics and approximations are used in constructing a phylogenetic tree, as the exact methods are too computationally intense. Hence, different methods sometimes produce significantly different phylogenetic trees. Manual assessment of different results is generally required.

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

TRANSCRIPTOME ANALYSIS

The primary goal of transcriptome analysis is to learn about how changes in transcript abundance control growth and development of an organism and its response to the environment. DNA microarrays proved a powerful technology for observing the transcriptional profile of genes at a genome-wide level (22, 111). Microarray data are also being combined with other information such as regulatory sequence analysis, gene ontology, and pathway information to infer coregulated processes. Whole-genome tiled arrays are used to detect transcription without bias toward known or predicted gene structures and alternative splice variants. Other types of analysis include ChIP-chip [chromatin immunoprecipitation (ChIP) and microarray analysis (chip)] analysis, which combines microarrays with methods for detecting the chromosomal locations at which protein-DNA interactions occur across the genome (23). A related technique uses DNA immunoprecipitation (DIP-chip) to predict DNA-binding sites (80). This review does not cover all available technologies for measuring expression data such as tag-based transcriptional profiling technologies like massively parallel signature sequencing (MPSS) and SAGE (20, 28).

Microarray Analysis

Microarray analysis allows the simultaneous measurement of transcript abundance for thousands of genes (153). Two general types of microarrays are high-density oligonucleotide arrays that contain a large number (thousands or often millions) of relatively short (25–100-mer) probes synthesized directly on the surface of the arrays, or arrays with amplified polymerase chain reaction products or cloned DNA fragments mechanically spotted directly on the array surface. Many different technologies are being developed, which have been recently surveyed by Meyers and colleagues (89). Competition among microarray platforms has led to lower costs and increased numbers of genes per array. Unfortunately, the diversity of array platforms makes it difficult to compare results between microarray formats that use different probe sequences, RNA sample labeling, and data collection methods (142).

Other important issues in microarray analysis are in processing and normalizing data. Some journals require multiple biological replicates (typically at least three) and statistically valid results before publishing microarray results. Replication of the microarray experiment and appropriate statistical design are needed to minimize the false discovery rate. The microarray data must also be deposited into a permanent public repository with open access. A good overview of microarray data analysis can be found in References 37 and 118. The main difficulty of dealing with microarray data is the sheer amount of data resulting from a single experiment. This makes it very difficult to decide which transcripts to focus on for interpreting the results. Even for standardized arrays such as those from Affymetrix, there are still arguments on the optimal statistical treatment for the sets of probes designed for each gene. For example, the *Affycomp* software compares Affymetrix results using two spike-in experiments and a dilution experiment for different methods of normalization under different assessment criteria (27). This information can be used to select the appropriate normalization methods.

Many tools are available that perform a variety of analysis on large microarray data sets. Examples include commercial software such as Gene Traffic, GeneSpring (**http://www.agilent.com/chem/genespring**), Affymetrix's GeneChip Operating Software (GCOS), and public software such as Cluster (41), CaARRAY (**http://caarray.nci.nih.gov/**), and BASE (109). A notable example is Bioconductor (**http://www.bioconductor.org**), which is an open-source and open-development set of routines written for the open-source R statistical analysis package (**http://www.r-project.org**).

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

Observing the patterns of transcriptional activity that occur under different conditions such as genotypes or time courses reveals genes that have highly correlated patterns of expression. However, correlation cannot distinguish between genes that are under common regulatory control and those whose expression patterns just happen to correlate. Recent efforts in microarray analysis have focused on analysis of microarray data across experiments (91). A study by the Toxicogenomics research consortium indicates that "microarray results can be comparable across multiple laboratories, especially when a common platform and set of procedures are used" (7). Meta-analysis can investigate the effect of the same treatment across different studies to arrive at a single estimate of the true effect of the treatment (106, 123).

Tiling Arrays

Typical microarray sample known and predicted genes. Tiling arrays cover the genome at regular intervals to measure transcription without bias toward known or predicted gene structures, discovery of polymorphisms, analysis of alternative splicing, and identification of transcription factor-binding sites (90). Whole-genome arrays (WGAs) cover the entire genome with overlapping probes or probes with regular gaps. The WGA ensures that the experimental results are not dependent on the level of current genome annotation as well as discovering new transcripts and unusual forms of transcription. In plants, similar studies have been performed for the entire *Arabidopsis* genome (127, 143) and parts of the rice genome (70, 79). These studies identified thousands of novel transcription units including genes within the centromeres, substantial antisense gene transcription, and transcription activity in intergenic regions. Tiling array data may also be used to validate predicted intron/exon boundaries (132).

Further work is needed to establish the best practices for determining when transcription has occurred and how to normalize array data across the different chips. Visualization of the output from tiling arrays requires viewing the probe sequences on the array together with the sequence assembly and the probe expression data. The *Arabidopsis* Tiling Array Transcriptome Express Tool (also known as ChipViewer) (**http://signal.salk.edu/cgi-bin/atta**) displays information about what type of transcription occurred along the *Arabidopsis* genome (143). Another tool is the Integrated Genome Browser (IGB) from Affymetrix, a Java program for exploring genomes and combining annotations from multiple data sources. Another option for visualizing such data are collaborations such as those between Gramene (137) and PLEXdb (116), which allow users to overlay probe array information onto a comparative sequence viewer.

The major limitations of WGAs include the requirement of a sequenced genome, the large number of chips required for complete genome coverage, and analysis of recently duplicated (and thus highly homologous) genes.

Regulatory Sequence Analysis

Interpreting the results of microarray experiments involves discovering why genes with similar expression profiles behave in a coordinated fashion. Regulatory sequence analysis approaches this question by extracting motifs that are shared between the upstream sequences of these genes (134). Comparative genomics studies of conserved noncoding sequences (CNSs) may also help to find key motifs (56, 67). There are several methods to search over-represented motifs at the upstream of coregulated genes. Roughly they can be categorized into two classes: oligonucleotide frequency-based (68, 134) and probabilistic sequence-based models (76, 85, 108).

The oligonucleotide frequency-based method calculates the statistical significance of a site based on oligonucleotide frequency tables observed in all noncoding regions of the specific organism's genome. Usually, the

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

length of the oligonucleotide varies from 4 to 9 bases. Hexanucleotide (oligonucleotide length of 6) analysis is most widely used. The significant oligonucleotides can then be grouped as longer consensus motifs. Frequency-based methods tend to be simple, efficient, and exhaustive (all over-represented patterns of chosen length are detected). The main limitation is the difficulty of identifying complex motif patterns. The public Web resource, Regulatory Sequence Analysis Tools (RSAT), performs sequence similarity searches and analyzes the noncoding sequences in the genomes (134).

For the probabilistic-based methods, the motif is represented as a position probability matrix, where the motifs are assumed to be hidden in the noisy background sequences. One of the strengths of probabilistic-based methods is the ability to identify motifs with complex patterns. Many potential motifs can be identified; however, it can be difficult to separate unique motifs from this large pool of potential solutions. Probabilistic-based methods also tend to be computationally intense as they must be run multiple times to get an optimal solution. AlignACE, Aligns Nucleic Acid Conserved Elements (**http://atlas.med.harvard.edu/**), is a popular motif finding tool that was first developed for yeast but has been expanded to other species (107).

COMPUTATIONAL PROTEOMICS

Proteomics is a leading technology for the qualitative and quantitative characterization of proteins and their interactions on a genome scale. The objectives of proteomics include large-scale identification and quantification of all protein types in a cell or tissue, analysis of post-translational modification and association with other proteins, and characterization of protein activities and structures. Application of proteomics in plants is still in its initial phase, mostly in protein identification (24, 96). Other aspects of proteomics (reviewed in 152), such as identification and prediction of protein-protein interactions, protein activity profiling, protein subcellular localization, and protein structure, have not been widely used in plant science. However, recent efforts such as the structural genomic initiative that includes *Arabidopsis* (**http://www.uwstructuralgenomics.org/**) are encouraging.

Electrophoresis Analysis

Electrophoresis analysis can qualitatively and quantitatively investigate expression of proteins under different conditions (54). Several bioinformatics tools have been developed for two-dimensional (2D) electrophoresis analysis (86). SWISS-2DPAGE can locate the proteins on the 2D PAGE maps from Swiss-Prot (**http://au.expasy.org/ch2d/**). Melanie (**http://au.expasy.org/melanie/**) can analyze, annotate, and query complex 2D gel samples. Flicker (**http://open2dprot.sourceforge.net/Flicker/**) is an open-source stand-alone program for visually comparing 2D gel images. PDQuest (**http://www.proteomeworks.bio-rad.com**) is a popular commercial software package for comparing 2D gel images. Some software platforms handle related data storage and management, including PEDRo (**http://pedro.man.ac.uk/**), a software package for modeling, capturing, and disseminating 2D gel data and other proteomics experimental data. Main limitations of electrophoresis analysis include limited ability to identify proteins and low accuracy in detecting protein abundance.

Protein Identification Through Mass Spectrometry

After protein separation using 2D electrophoresis or liquid chromatography and protein digestion using an enzyme (trypsin, pepsin, glu-C, etc.), proteins are identified by typically using mass spectrometry (MS) (1). In contrast to other protein identification techniques, such as Edman degradation microsequencing, MS provides a high-throughput approach for large-scale protein identification. The data generated from mass spectrometers are often complicated and computational analyses are critical in interpreting the data for protein identification (17, 55). A major limitation in MS protein identification is the lack of open-source software. Most widely used tools are expensive commercial packages. In addition, current statistical models for matches between MS spectra and protein sequences are generally oversimplified. Hence, the confidence assessments for computational protein identification results are often unreliable. There are two types of MS-based protein identification methods: peptide mass fingerprinting (PMF) and tandem mass spectrometry (MS/MS).

Peptide mass fingerprinting. PMF peptide/protein identification compares the masses of peptides derived from the experimental spectral peaks with each of the possible peptides computationally digested from proteins in the sequence database. The proteins in the sequence database with a significant number of peptide matches are considered candidates for the proteins in the experimental sample. MOWSE (99) was an earlier software package for PMF protein identification, and Emowse (**http://emboss.sourceforge.net/**) is the latest implementation of the MOWSE algorithm. Several other computational tools have also been developed for PMF protein identification. MS-Fit in the Protein Prospector (**http://prospector.ucsf.edu/**) uses a variant of MOWSE scoring scheme incorporating new features, including constraints on the minimum number of peptides to be matched for a possible hit, the number of missed cleavages, and the target protein's molecular weight range. Mascot (**http://www.matrixscience.com/**) is an extension of the MOWSE algorithm. It incorporates the same scoring scheme with the addition of a probability-based score. A limitation of PMF protein identification is that it sometimes cannot identify proteins because multiple proteins in the database can fit the PMF spectra. In this case, additional MS/MS experiments are needed to identify the proteins.

Tandem mass spectrometry. MS/MS further breaks each digested peptide into smaller fragments, whose spectra provide effective signatures of individual amino acids in the peptide for protein identification. Many tools have been developed for MS/MS-based peptide/protein identification, the most popular ones being SEQUEST (**http://fields.scripps.edu/sequest/**) and Mascot (**http://www.matrixscience.com/**). Both rely on the comparison between theoretical peptides derived from the database and experimental mass spectrometric tandem spectra. SEQUEST, one of the earliest tools developed for this, produces a list of possible peptide/protein assignments in a protein mixture based on a correlation scoring scheme (145). Mascot, together with its PMF protein identification capacity, uses a similar algorithm as SEQUEST for MS/MS peptide/protein identification. The limitations of these programs are that a significant portion of MS/MS spectra cannot be assigned due to various factors, including sequencing and annotation errors in the search database. In addition, post-translational modifications are currently not handled well using computational approaches.

The de novo sequencing approach based on MS/MS spectra is an active research area (30). Typically the algorithms match the separations of peaks by the mass of one or several amino acids and infer the probable peptide sequences that are consistent with the matched amino acids (25). There are a few popular software packages for peptide de novo sequencing using MS/MS data, including Lutefisk (**http://www.hairyfatguy.com/lutefisk/**) and PEAKS (**http://www.bioinformaticssolutions.com/products/peaks**). One limitation of current de novo methods is that they often cannot provide the exact sequence of a peptide. Instead, several top candidate sequences are suggested.

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

METABOLOMICS AND METABOLIC FLUX

Metabolomics is the analysis of the complete pool of small metabolites in a cell at any given time. Metabolomics may prove to be particularly important in plants due to the proliferation of secondary metabolites. As of 2004, more than 100,000 metabolites have been identified in plants, with estimates that this may be less that 10% of the total (133). In a metabolite profiling experiment, metabolites are extracted from tissues, separated, and analyzed in a high-throughput manner (44). Metabolic fingerprinting looks at a few metabolites to help differentiate samples according to their phenotype or biological relevance (58, 115). Technology has now advanced to semiautomatically quantify >1000 compounds from a single leaf extract (138).

The key challenge in metabolite profiling is the rapid, consistent, and unambiguous identification of metabolites from complex plant samples (110). Identification is routinely performed by time-consuming standard addition experiments using commercially available or purified metabolite preparations. A publicly accessible database that contains the evidence and underlying metabolite identification for gas chromatography-mass spectrometry (GC–MS) profiles from diverse biological sources is needed. Standards for experimental metadata and data quality in metabolomics experiments are still in a very early stage and a large-scale public repository is not yet available. The ArMet (architecture for metabolomics) proposal (61) gives a description of plant metabolomics experiments and their results along with a database schema. MIAMET (Minimum Information About a Metabolomics Experiment) (13) gives reporting requirements with the aim of standardizing experiment descriptions, particularly within publications. The Standard Metabolic Reporting Structures (SMRS) working group (119) has developed standards for describing the biological sample origin, analytical technologies, and methods used in a metabolite profiling experiment.

Metabolite data have been used to construct metabolic correlation networks (121). Such correlations may reflect the net partitioning of carbon and nitrogen resulting from direct enzymatic conversions and indirect cellular regulation by transcriptional or biochemical processes. However, metabolic correlation matrices cannot infer that a change in one metabolite led to a change in another metabolite in a metabolic reaction network (122).

Metabolic flux analysis measures the steady-state flow between metabolites. Fluxes, however, are even more difficult to measure than metabolite levels due to complications in modeling intracellular transport of metabolites and the incomplete knowledge about the topology and location of the pathways in vivo (115). The most basic approach to metabolic flux analysis is stoichiometric analysis that calculates the quantities of reactants and products of a chemical reaction to determine the flux of each metabolite (39). However, this method is numerically difficult to solve for large networks and it has problems if parallel metabolic pathways, metabolic cycles, and reversible reactions are present (140). FluxAnalyzer is a package for MATLAB that integrates pathway and flux analysis for metabolic networks (75).

Flux analysis using ^{13}C carbon labeling data seeks to overcome some of the disadvantages of stoichiometric flux analysis described above (120). More rigorous analysis is needed for full determination of fluxes from all of the experimental data in ^{13}C constrained flux analysis (stoichiometric model with a few flux ratios as constraints) and the stoichiometric and isotopomer balances. Iterative methods have been used to solve the resulting matrix of isotopomer balances, with the nuclear magnetic resonance or gas chromatography measurements used to provide consistency. As more reliable data are collected, one can use ordinary differential equations for dynamic simulations of metabolic networks and combine information about connectivity, concentration balances, flux balances, metabolic control, and pathway optimization. Ultimately, one may integrate all of the information and perform analysis and simulation in a cellular modeling environment like E-Cell (**http://www.e-cell.org/**) or CellDesigner (**http://www.systems-biology.org**).

ONTOLOGIES

The data that are generated and analyzed as described in the previous sections need to be compared with the existing knowledge in the field in order to place the data in a biologically meaningful context and derive hypotheses. To do this efficiently, data and knowledge need to be described in explicit and unambiguous ways that must be comprehensible to both humans and computer programs. An ontology is a set of vocabulary terms whose meanings and relations with other terms are explicitly stated and which are used to annotate data (5, 10, 14, 124). This section introduces the types of ontologies in development and use today and some applications and caveats of using the ontologies in biology.

Types of Bio-Ontologies

A growing number of shared ontologies are being built and used in biology. Examples include ontologies for describing gene and protein function (59), cell types (9), anatomies and developmental stages of organisms (50, 135, 144), microarray experiments (126), and metabolic pathways (84, 151). A list of open-source ontologies used in biology can be found on the Open Biological Ontologies Web site (**http://obo.sourceforge.net/**). Many ontologies on this site are under development and are subject to frequent change. The Gene Ontology (GO) (**www.geneontology.org**) is an example of bio-ontologies that has garnered community acceptance. It is a set of more than 16,000 controlled vocabulary terms for the biological domains of "molecular function," "subcellular compartment," and "biological process." GO is organized as a directed acyclic graph, which is a type of hierarchy tree that allows a term to exist as a specific concept belonging to more than one general term. Other examples of ontologies currently in development are the Sequence Ontology (SO) project (40) and the Plant Ontology (PO) project (**www.plantontology.org**). The SO project aims to explicitly define all the terms needed to describe features on a nucleotide sequence, which can be used for genome sequence annotation for any organism. The PO project aims to develop shared vocabularies to describe anatomical structures for flowering plants to depict gene expression patterns and plant phenotypes.

A few challenges in the development and use of ontologies remain to be addressed, including redundancies in the ontologies, minimal or lack of formal, computer-comprehensive definitions of the terms in the ontologies, and general acceptance by the research and publishing community (10, 14). There is an opportunity for an international repository of ontology standards that could oversee the development and maintenance of the ontologies.

Applications of Ontologies

Ontologies are used mainly to annotate data such as sequences, gene expression clusters, experiments, and strains. Ontologies that have such annotations to data in databases can be used in numerous ways, including connecting different databases, refining searching, providing a framework for interpreting the results of functional genomics experiments, and inferring knowledge (8, 10, 47). For example, one can ask which functions and processes are statistically significantly over-represented in an expression cluster of interest compared to the functions and processes carried out by all of the genes from a gene expression array. Because GO is one of the more well-established ontologies, this section focuses on GO to illustrate applications

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

of ontologies in biology. Ontologies have been used by many model organism databases to annotate genes and gene products (**http://www.geneontology.org/GO.current.annotations.shtml, http://www.geneontology.org/GO.biblio.shtml#annots**). Function annotations of genes using GO have been used mainly in two ways: predicting protein functions, processes, and localization patterns from various data sources (**http://www.geneontology.org/GO.biblio.shtml#predictions**) and providing a biological framework or benchmark set for interpreting results of large-scale probing of samples such as gene expression profiles and protein-protein interactions (**http://www.geneontology.org/GO.biblio.shtml#gene_exp**). In addition, GO annotations have been used to test the robustness of semantic similarity searching methods (83) and to study adaptive evolution (4).

There are several issues in using GO annotations to predict function and to use as a benchmark for large-scale data. One is the misuse or lack of use of evidence codes, which provide the type of evidence that was used to make the annotation (**http://www.geneontology.org/GO.evidence.shtml**). Only about half of the evidence codes refer to direct experimental evidence. Also, several evidence codes are used for indirect evidence, which indicate less certainty in the assertion of the annotation than those made with direct experimental evidence. Other codes are used for computationally derived annotations and have no experimental support and have a higher probability of being incorrect. Researchers and computer programs that use the annotations for inferring knowledge or analyzing functional genomics data should be familiar with these evidence codes in order to minimize misinterpretation of the data. For example, methods to assess relationship between sequence conservation and coexpression of genes and using GO annotations to validate their results should ensure that no annotations using the ISS and IEA evidence codes are used to avoid circular arguments. Similarly, studies that attempt to define biological processes and functions from gene expression data using the GO annotations should ensure that no annotation with inferred from expression pattern (IEP) evidence code is used. The other caveat is that annotations to GO are not equivalently represented throughout GO. When looking for statistical over-representation of GO terms in genes of an expression cluster, there is low statistical power for detecting deviations from expectation for terms that are annotated with a small number of genes (74).

Software for Accessing and Analyzing Ontologies and Annotations

There are a number of software tools for visualizing, editing, and analyzing ontologies and their annotations. The GO Web site maintains a comprehensive list of these tools (**http://www.geneontology.org/GO.tools.shtml**). Some of them are accessible via Web browsers and others have to be installed locally. Tools are also needed to facilitate data integrity checks and more flexible and customizable searching and browsing capabilities to explore these complex networks of concepts. Most of the tools that facilitate analysis of the GO annotations are developed to help interpret gene expression studies. These applications allow researchers to compare a list of genes (for example, from an expression cluster) and identify over-represented GO terms in this list as compared to the whole genome or whole list of genes under study. Most of these software programs use statistical models to provide significance in the over-representation. Recently, Khatri and colleagues reported comparisons of 14 of these tools on their functionalities, advantages, and limitations (74). Finally, most of the bio-ontologies are informal in their semantic representation. Definitions of the terms are provided in natural language, which is fine for human comprehension but does not easily allow computers and software to be developed that can help check for ontology integrity and provide more semantically powerful search functions. More tools are needed that can facilitate the conversion of bio-ontologies to be more formal and computer comprehensive.

DATABASES

Traditionally, biologists relied on textbooks and research articles published in scientific journals as the main source of information. This has changed dramatically in the past decade as the Internet and Web browsers became commonplace. Today, the Internet is the first place researchers go to find information. Databases that are available via the Web also became an indispensable tool for biological research. In this section, we describe types and examples of biological databases, how these databases are built and accessed, how data among databases are exchanged, and current challenges and opportunities in biological database development and maintenance.

Types of Biological Databases

Three types of biological databases have been established and are developed: large-scale public repositories, community-specific databases, and project-specific databases. *Nucleic Acids Research* (**http://nar.oxfordjournals.org/**) publishes a database issue in January of every year. Recently, *Plant Physiology* started publishing articles describing databases (105). Large-scale public repositories are usually developed and maintained by government agencies or international consortia and are places for long-term data storage. Examples include GenBank for sequences (139), UniProt (113) for protein information, Protein Data Bank (32) for protein structure information, and ArrayExpress (100) and Gene Expression Omnibus (GEO) (38) for microarray data. There are a number of community-specific databases, which typically contain information curated with high standards and address the needs of a particular community of researchers. A prominent example of community-specific databases are those that cater to researchers focused on studying model organisms (77, 104, 144) or clade-oriented comparative databases (53, 88, 92, 137). Other examples of community-specific databases include databases focused on specific types of data such as metabolism (151) and protein modification (129). The concept of community-specific databases is subject to change as researchers are widening their scope of research. For example, databases focused on comparing genome sequences recently emerged (e.g., **http://www.phytome.org** and Reference 64). The third category of databases includes smaller-scale, and often short-lived, databases that are developed for project data management during the funding period. Often these databases and Web resources are not maintained beyond the funding period of the project and currently there is no standard way of depositing or archiving these databases after the funding period.

There are some issues in database management. First, there is a general lack of good documentation on the rationale of the design and implementation. More effort is needed to share the experiences via conferences and publications. Also, there are no accepted standards in making databases, schema, software, and standard operating procedures available. In response to this, the National Human Genome Research Institute (NHGRI) has funded a collaborative project called the Generic Model Organism Database (**http://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." Another major issue is that there is a general lack of infrastructure of supporting, managing, and using digital data archived in databases and Web sites in the long term (82). One possibility to alleviate this problem is to create a public archive of biological databases and Web sites to which finished projects

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

could deposit the database, software, and Web sites. There are several projects that are building digital repository systems that can be models for such a repository such as D-Space (**http://dspace.org/**) and the CalTech Collection of Open Digital Archives (CODA; **http://library.caltech.edu/digital/**). Some additional challenges in long-term archiving of data were articulated in a recent National Science Board report (**http://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf**).

Data Representation and Storage

Databases can be developed using a number of different methods including simple file directories, object-oriented database software, and relational database software. Due to the increasing quantity of data that need to be stored and made accessible using the Internet, relational database management software has become popular and has become the de facto standard in biology. Relational databases provide effective means of storing and retrieving large quantities of data via indexes, normalization, referential integrity, triggers, and transactions. Notable relational database software that is freely available and quite popular in bioinformatics is MySQL (**http://www.mysql.com/**) and PostgreSQL (**http://www.postgresql.org/**). In relational databases, data are represented as entities, attributes (properties of the entities), and relationships between the entities. This type of representation is called Entity-Relationship (ER) and database schemas are described using ER diagrams (e.g., TAIR schema at **http://arabidopsis.org/search/schemas.html**). Entities and attributes become tables and columns in the physical implementation of the database, respectively. Data are the values that are stored in the fields of the tables.

Although relational databases are powerful ways of storing large quantities of data, they have limitations. For example, it is not trivial to represent complex relationships between data such as signal transduction pathways. Also, it is difficult to create rich semantic relationships in relational databases to ask the database "what if" types of queries without having extensive software built on top of the database. Another limitation of relational databases is that it is very difficult, if not impossible, to preserve all of the changes that occur to attributes of entities.

Data Access and Exchange

The most direct, powerful, and flexible way of accessing data in a database is using structured query language (SQL) (**http://databases.about.com/od/sql/**). SQL has a reasonably intuitive and simple syntax that requires no programming knowledge and is suited for biologists to learn without a steep learning curve. However, to use SQL, users need to know the database schema. In addition, some queries that are based on less optimized database structure could result in slow performance and can even sometimes lock the database system. In most databases, access to the data is provided via database access software and graphical user interface (GUI) that allow searching and browsing of the data. In addition to text-based search user interfaces, more sophisticated ways of accessing data such as graphical displays and tree-based browsers are also common.

Although accessing information from a database is fairly easy if one knows which database to go to, it is not as easy to find information if one does not know which database to search. There are several ways to solve this problem such as indexing the content of database-driven pages, developing software that will connect to individual databases directly, or developing a data warehouse of many different data types or database in one site. A relatively new method that is gaining some attention is to use a registry system where different databases that specialize on particular information can declare what data are available in their system and register methods to access their data. Users can send requests to

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

the registry system, which then contact the appropriate databases to retrieve the requested data. Conceptually, this is an elegant way of integrating different databases without depending on the individual databases' schema. However, this relies on the willingness of individual databases to participate in the registry system. This method is called Web services and has been accepted widely by the Internet industry but has not yet been commonly implemented. Projects like BioMOBY (141) and myGRID (125) are implementing this idea for biological databases, but they have not yet been widely used.

Semantics (meaning) and syntax (format) of data need to be made explicit in order to exchange data for analysis and mining. A simple way of formatting data is using a tag and value system (called markup language). An emerging standard for exchanging data and information via the Web is Extensible Markup Language (XML), which allows information providers to define new tag and attribute names at will and to nest document structures to any level of complexity, among other features. The document that defines the meaning of the tags for an XML document is called Document Type Definition (DTD). The use of a common DTD allows different users and applications to exchange data in XML. Although many databases and bioinformatics projects present their data in XML, currently almost every group has their own DTD. Standardization and common use of DTDs for exchanging common data types will be pivotal. There are notable exceptions to this rule including the specification of microarray data, MAGEML (Microarray Gene Expression Markup Language), provided by the Microarray Gene Expression Database Society (MGED) (**http://www.mged.org/**). To a lesser extent, the BIOPAX (**http://www.biopax.org/**) is also becoming a community-accepted standard to describe pathways and reactions. Other than DTDs, biological database communities do not yet have a standard system in software engineering to communicate with each other.

Data Curation

Data curation is defined as any activity devoted to selecting, organizing, assessing quality, describing, and updating data that result in enhanced quality, trustworthiness, interpretability, and longevity of the data. It is a crucial task in today's research environment where data are being generated at an ever-increasing rate and an increasing amount of research is based on re-use of data. In general, some level of curation is done by data generators, but most curation activities are carried out in data repositories. A number of different strategies to curation are used, including computational, manual, in-house, and those that involve external expertise. Assessing data quality involves both determining the criteria for measuring quality and performing the measurements. Data quality criteria for raw data are tied with methods of data acquisition. In many databases, these criteria are not made explicit and the information on the metrics of data-quality assessment is rare.

Curation of data into public repositories should be a parallel and integrated process with publication in peer-reviewed journals. Although much progress has been made in electronic publication and open-access publishing, there is still a gap between connecting the major conclusions in papers and the data that were used to draw the conclusions. In a few cases, data are required to be submitted to public repositories (e.g., sequence data to GenBank, microarray data to ArrayExpress/GEO, and *Arabidopsis* stock data to ABRC). However, there are no such standards established for other data types (e.g., proteomics data, metabolomics data, protein localization, in situ hybridization, phenotype description, protein function information). Standards, specifications, and requirements for publication of data into repositories should be made more accessible to researchers early on in their data-generation and research-activity processes.

One of the most important aspects of today's changing research landscape is

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

the culture of data and expertise sharing. The now famous Bermuda principle (**http://www.gene.ucl.ac.uk/hugo/bermuda.htm**) was extended to large-scale data at a recent meeting (131). In this meeting, the policy for publicly releasing large-scale data prepublication and appropriate conduct and acknowledgment of the uses of these data by the scientific community were discussed. Clearly articulated and community-accepted policies are needed on how data from data repositories should be cited and referenced and how the generators of the data should be acknowledged. Establishing this standard should include journal publishers, database scientists, data generators, funding bodies, and representatives of the user community. Additional challenges and opportunities in database curation were recently articulated (82, 103).

EMERGING AREAS IN BIOINFORMATICS

In addition to some of the challenges and opportunities mentioned in this review, there are many exciting areas of research in bioinformatics that are emerging. In this section, we focus on a few of these areas such as text mining, systems biology, and the semantic web. Some additional emerging areas such as image analysis (117), grid computing (46, 49), directed evolution (29), rational protein design (81), microRNA-related bioinformatics (21), and modeling in epigenomics (43) are not covered due to the limitation of space.

Text Mining

The size of the biological literature is expanding at an increasing rate. The Medline 2004 database had 12.5 million entries and is expanding at a rate of 500,000 new citations each year (26). The goal of text mining is to allow researchers to identify needed information and shift the burden of searching from researchers to the computer. Without automated text mining, much of biomolecular interactions and biological research archived in the literature will remain accessible in principle but underutilized in practice. One key area of text mining is relationship extraction that finds relationships between entities such as genes and proteins. Examples include MedMiner at the National Library of Medicine (128), PreBIND (35), the curated BIND system (2, 6), PathBinderH (155), and iHOP (63). (See Reference 26 for a complete survey of text mining applications.) Results on real-world tasks such as the automatic extraction and assignment of GO annotations are promising, but they are still far from reaching the required performance demanded by real-world applications (15). One key difficulty that needs to be addressed in this field is the complex nature of the names and terminology such as the large range of variants for protein names and GO terms in free text. The current generation of systems is beginning to combine statistical methods with machine learning to capture expert knowledge on how genes and proteins are referred to in scientific papers to create usable systems with high precision and recall for specialized tasks in the near future.

Computational Systems Biology

Classical systems analysis in engineering treats a system as a black box whose inner structure and behavior can be analyzed and modeled by varying internal or external conditions, and studying the effect of the variation on the external observables. The result is an understanding of the inner makeup and working mechanisms of the system (72). Systems biology is the application of this theory to biology. The observables are measurements of what the organism is doing, ranging from phenotypic descriptions to detailed metabolic profiling. A critical issue is how to effectively integrate various types of data, such as sequence, gene expression, protein interactions, and phenotypes to infer biological knowledge. Some areas that require more work include creating coherent validated data sets, developing

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

common formats for pathway data [SBML (65) and BioPAX (**http://www.biopax.org**)], and creating ontologies to define complex interactions, curation, and linkages with text-mining tools. The Systems Biology Workbench project (**http://sbw.kgi.edu/**) aims to develop an open-source software framework for sharing information between different types of pathway models. Other issues are that biological systems are underdefined (not enough measurements are available to characterize the system) and samples are not taken often enough to capture time changes in a system that may occur at vastly different time scales in different networks such as signaling and regulatory networks (98). The long-term goal to create a complete in silico model of a cell is still distant; however the tools that are being developed to integrate information from a wide variety of sources will be valuable in the short term.

Semantic Web

Semantic web is a model to "create a universal mechanism for information exchange by giving meaning, in a machine-interpretable way, to the content of documents and data on the Web" (95). This model will enable the development of searching tools that know what type of information can be obtained from which documents and understand how the information in each document relates to another, which will allow software agents that can use reasoning and logic to make decisions automatically based on the constraints provided in the query (e.g., automatic travel agents, phenotype prediction) (12). Bioinformatics could benefit enormously from successful implementation of this model and should play a leading role in realizing it (95). Current efforts to realize the concepts of the semantic web have been focused on developing standards and specifications of identifying and describing data such as Universal Resource Identifier (URI) and Resource Definition Framework (RDF), respectively (**http://www.w3c.org/2001/sw**). Although implementation of applications using the semantic web is scarce at this point, there are some useful examples being developed such as Haystack (a browser that retrieves data from multiple databases and allows users to annotate and manage the information to reflect their understanding) (**http://www-db.cs.wisc.edu/cidr/cidr2005/papers/P02.pdf**) and BioDash (a drug development user interface that associates diseases, drug progression stages, molecular biology, and pathway knowledge for users) (**http://www.w3.org/2005/04/swls/BioDash/Demo/**).

Cellular Localization and Spatially Resolved Data

Research in nanotechnology and electron microscopy is allowing researchers to select specific areas of cells and tissues and to image spatiotemporal distributions of signaling receptors, gene expression, and proteins. Laser capture microdissection allows the selection of specific tissue types for detailed analysis (42). This technique has been applied to specific plant tissues in maize and *Arabidopsis* (73, 94). Confocal imaging is being used to model auxin transport and gene expression patterns in *Arabidopsis* (60). Methods in electron microscopy are being applied to image the spatiotemporal distribution of signaling receptors (149). Improved methods in laser scanning microscopes may allow measurements of fast diffusion and dynamic processes in the microsecond-to-millisecond time range in live cells (34). These emerging capabilities will lead to new understanding of cell dynamics.

CONCLUSION

In this review, we attempt to highlight some of the recent advances made in bioinformatics in the basic areas of sequence, gene expression, protein, and metabolite analyses, databases, and ontologies, current limitations in these areas, and some emerging areas. A number of unsolved problems exist in bioinformatics

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

today, including data and database integration, automated knowledge extraction, robust inference of phenotype from genotype, and training and retraining of students and established researchers in bioinformatics. Bioinformatics is an approach that will be an essential part of plant research and we hope that every plant researcher will incorporate more bioinformatics tools and approaches in their research projects.

If the next 50 years of plant biology can be summed into one word, it would be "integration." We will see integration of basic research with applied research in which plant biotechnology will play an essential role in solving urgent problems in our society such as developing renewable energy, reducing world hunger and poverty, and preserving the environment. We will see integration of disparate, specialized areas of plant research into more comparative, connected, holistic views and approaches in plant biology. We will also see more integration of plant research and other biological research, from microbes to human, from a large-scale comparative genomics perspective. Bioinformatics will provide the glue with which all of these types of integration will occur. However, it will be people, not tools, who will enable the gluing. Ways in which biological research will be conducted in 2050 will be much different from the way in which it was done in 2000. Each researcher will spend more time on the computer and the Internet to generate and describe data and experiments, to analyze the data and find other people's data relevant for comparison, to find existing knowledge in the field and to relate it to his or her results into the current body of knowledge, and to publish his or her results to the world.

ACKNOWLEDGMENTS

We are grateful to Blake Meyers, Dan MacLean, Shijun Li, Scott Peck, Mark Lange, Bill Beavis, Todd Vision, Stefanie Hartmann, Gary Stacey, Chris Town, Volker Brendel, and Nevin Young for their critical comments on the manuscript. This work has been supported in part by NSF grants DBI-99,78564, DBI-04,17062, DBI-03,21666 (SYR); ITR-IIS-04,07204 (DX); DBI-02,09809 (JD); USDA grants NRI-2002-35,300-12,619 (JD) and CSREES 2004-25,604-14,708 (DX); NIH grants NHGRI-HG002273, R01-GM65466 (SYR); National Center for Soybean Biotechnology (DX); Pioneer-Hi-Bred (SYR); and Carnegie Canada (SYR).

LITERATURE CITED

1. Aebersold R, Mann M. 2003. Mass spectrometry-based proteomics. *Nature* 422:198–207
2. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, et al. 2005. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res.* 33:D418–24
3. Allen JE, Pertea M, Salzberg SL. 2004. Computational gene prediction using multiple sources of evidence. *Genome Res.* 14:142–48
4. Aris-Brosou S. 2005. Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis. *Mol. Biol. Evol.* 22:200–9
5. Ashburner M, Ball C, Blake J, Botstein D, Butler H, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25:25–29
6. Bader G, Betel D, Hogue C. 2002. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* 31:248–50
7. Bammler T, Beyer RP, Bhattacharya S, Boorman GA, Boyles A, et al. 2005. Standardizing global gene expression analysis between laboratories and across platforms. *Nat. Methods* 2:351–56

8. Bard J. 2003. Ontologies: formalising biological knowledge for bioinformatics. *Bioessays* 25:501–6
9. Bard J, Rhee SY, Ashburner M. 2005. An ontology for cell types. *Genome Biol.* 6:R21
10. Bard JB, Rhee SY. 2004. Ontologies in biology: design, applications and future challenges. *Nat. Rev. Genet.* 5:213–22
11. Bedell JA, Budiman MA, Nunberg A, Citek RW, Robbins D, et al. 2005. Sorghum genome sequencing by methylation filtration. *PLoS Biol.* 3:e13
12. Berners-Lee T, Hendler J, Lassila O. 2001. The Semantic Web. *Sci. Am.* 284:34–43
13. Bino R, Hall R, Fiehn O, Kopka J, Saito K, et al. 2004. Potential of metabolomics as a functional genomics tool. *Trends Plant Sci.* 9:418–25
14. Blake J. 2004. Bio-ontologies-fast and furious. *Nat. Biotechnol.* 22:773–74
15. Blaschke C, Krallinger M, Leon E, Valencia A. 2005. Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinformatics* 6:S16
16. Blazej RG, Paegel BM, Mathies RA. 2003. Polymorphism ratio sequencing: a new approach for single nucleotide polymorphism discovery and genotyping. *Genome Res.* 13:287–93
17. Blueggel M, Chamrad D, Meyer HE. 2004. Bioinformatics in proteomics. *Curr. Pharm. Biotechnol.* 5:79–88
18. Boguski MS, Schuler GD. 1995. ESTablishing a human transcript map. *Nat. Genet.* 10:369–71
19. Brendel V, Zhu W. 2002. Computational modeling of gene structure in Arabidopsis thaliana. *Plant Mol. Biol.* 48:49–58
20. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, et al. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* 18:630–34
21. Brown JR, Sanseau P. 2005. A computational view of microRNAs and their targets. *Drug Discov. Today* 10:595–601
22. Brown P, Botstein D. 1999. Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* 21:33–37
23. Buck MJ, Lieb JD. 2004. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* 83:349–60
24. Canovas FM, Dumas-Gaudot E, Recorbet G, Jorrin J, Mock HP, Rossignol M. 2004. Plant proteome analysis. *Proteomics* 4:285–98
25. Chen T, Kao MY, Tepel M, Rush J, Church GM. 2001. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* 8:325–37
26. Cohen AM, Hersh WR. 2005. A survey of current work in biomedical text mining. *Brief Bioinform.* 6:57–71
27. Cope LM, Irizarry RA, Jaffee HA, Wu Z, Speed TP. 2004. A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics* 20:323–31
28. Coughlan SJ, Agrawal V, Meyers B. 2004. A comparison of global gene expression measurement technologies in Arabidopsis thaliana. *Comp. Funct. Genomics* 5:245–52
29. Dalby PA. 2003. Optimising enzyme function by directed evolution. *Curr. Opin. Struct. Biol.* 13:500–5
30. Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA. 1999. De novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* 6:327–42
31. Densmore LD 3rd. 2001. Phylogenetic inference and parsimony analysis. *Methods Mol. Biol.* 176:23–36

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

32. Deshpande N, Addess KJ, Bluhm WF, Merino-Ott JC, Townsend-Merino W, et al. 2005. The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.* 33:D233–37

33. Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, et al. 2005. Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. *Bioinformatics* 21:1958–63

34. Digman MA, Brown CM, Sengupta P, Wiseman PW, Horwitz AR, Gratton E. 2005. Measuring fast dynamics in solutions and cells with a laser scanning microscope. *Biophys. J.* 89:1317–27

35. Donaldson I, Martin J, de Bruijn B, Wolting C, Lay V, et al. 2003. PreBIND and Textomy—mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics* 4:11

36. Doolittle WF. 1999. Phylogenetic classification and the universal tree. *Science* 284:2124–29

37. Draghici S. 2003. *Data Analysis Tools for DNA Microarrays*. London: Chapman and Hall

38. Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids. Res.* 30:207–10

39. Edwards JS, Palsson BO. 2000. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. USA* 97:5528–33

40. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, et al. 2005. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* 6:R44

41. Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95:14863–68

42. Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, et al. 1996. Laser capture microdissection. *Science* 274:998–1001

43. Fazzari MJ, Greally JM. 2004. Epigenomics: beyond CpG islands. *Nat. Rev. Genet.* 5:446–55

44. Fiehn O. 2002. Metabolomics—the link between genotypes and phenotypes. *Plant Mol. Biol.* 48:155–71

45. Foissac S, Bardou P, Moisan A, Cros MJ, Schiex T. 2003. EUGENE'HOM: a generic similarity-based gene finder using multiple homologous sequences. *Nucleic Acids Res.* 31:3742–45

46. Foster I. 2002. What is the Grid? A three point checklist. In *GRIDToday*, pp. 4. Chicago: Argonne National Lab & University of Chicago

47. Fraser AG, Marcotte EM. 2004. A probabilistic view of gene function. *Nat. Genet.* 36:559–64

48. Frazer KA, Chen X, Hinds DA, Pant PV, Patil N, Cox DR. 2003. Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. *Genome Res.* 13:341–46

49. Gannon D, Alameda J, Chipara O, Christie M, Duke V, et al. 2005. Building grid portal applications from a Web service component architecture. *Proc. IEEE* 93:551–63

50. Garcia-Hernandez M, Berardini TZ, Chen G, Crist D, Doyle A, et al. 2002. TAIR: a resource for integrated Arabidopsis data. *Funct. Integr. Genomics* 2:239–53

51. Gibbs RA, Weinstock GM. 2003. Evolving methods for the assembly of large genomes. *Cold Spring Harb. Symp. Quant. Biol.* 68:189–94

52. Goff SA, Ricke D, Lan TH, Presting G, Wang R, et al. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* 296:92–100

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

53. Gonzales MD, Archuleta E, Farmer A, Gajendran K, Grant D, et al. 2005. The Legume Information System (LIS): an integrated information resource for comparative legume biology. *Nucleic Acids Res.* 33:D660–65

54. Gorg A, Obermaier C, Boguth G, Harder A, Scheibe B, et al. 2000. The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* 21:1037–53

55. Gras R, Muller M. 2001. Computational aspects of protein identification by mass spectrometry. *Curr. Opin. Mol. Ther.* 3:526–32

56. Guo H, Moose SP. 2003. Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. *Plant Cell* 15:1143–58

57. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, et al. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31:5654–66

58. Harrigan GG, Goodacre R, eds. 2003. *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*. Boston: Plenum

59. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, et al. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32:D258–61

60. Heisler MG, Ohno C, Das P, Sieber P, Reddy GV, et al. 2005. Patterns of auxin transport and gene expression during primordium development revealed by live imaging of the Arabidopsis inflorescence meristem. *Curr. Biol.* 15:1899–911

61. Jenkins H, Hardy N, Beckmann D, Draper J, Smith AR, et al. 2004. A proposed framework for the description of plant metabolomics experiments and their results. *Nat. Biotechnol.* 22:1601–6

62. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, et al. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science* 307:1072–79

63. Hoffmann R, Valencia A. 2004. A gene network for navigating the literature. *Nat. Genet.* 36:664

64. Horan K, Lauricha J, Bailey-Serres J, Raikhel N, Girke T. 2005. Genome cluster database. A sequence family analysis platform for Arabidopsis and rice. *Plant Physiol.* 138:47–54

65. Hucka M, Finney A, Bornstein BJ, Keating SM, Shapiro BE, et al. 2004. Evolving a lingua franca and associated software infrastructure for computational systems biology: the Systems Biology Markup Language (SBML) Project. *Syst. Biol.* 1:41–53

66. Hudek AK, Cheung J, Boright AP, Scherer SW. 2003. Genescript: DNA sequence annotation pipeline. *Bioinformatics* 19:1177–78

67. Inada DC, Bashir A, Lee C, Thomas BC, Ko C, et al. 2003. Conserved noncoding sequences in the grasses. *Genome Res.* 13:2030–41

68. Jensen LJ, Knudsen S. 2000. Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics* 16:326–33

69. Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* 431:569–73

70. Jiao Y, Jia P, Wang X, Su N, Yu S, et al. 2005. A tiling microarray expression analysis of rice chromosome 4 suggests a chromosome-level regulation of transcription. *Plant Cell* 17:1641–57

71. Karlin S, Altschul SF. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* 87:2264–68

72. Kell DB, Brown M, Davey HM, Dunn WB, Spasic I, Oliver SG. 2005. Metabolic footprinting and systems biology: the medium is the message. *Nat. Rev. Microbiol.* 3:557–65

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

73. Kerk NM, Ceserani T, Tausta SL, Sussex IM, Nelson TM. 2003. Laser capture microdissection of cells from plant tissues. *Plant Physiol.* 132:27–35

74. Khatri P, Draghici S. 2005. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 21:3587–95

75. Klamt S, Stelling J, Ginkel M, Gilles ED. 2003. FluxAnalyzer: exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps. *Bioinformatics* 19:261–69

76. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262:208–14

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

78. Lewin B. 2003. *Genes VIII*. Upper Saddle River, NJ: Prentice Hall

79. Li L, Wang X, Xia M, Stolc V, Su N, et al. 2005. Tiling microarray analysis of rice chromosome 10 to identify the transcriptome and relate its expression to chromosomal architecture. *Genome Biol.* 6:R52

80. Liu X, Noll DM, Lieb JD, Clarke ND. 2005. DIP-chip: rapid and accurate determination of DNA-binding specificity. *Genome Res.* 15:421–27

81. Looger LL, Dwyer MA, Smith JJ, Hellinga HW. 2003. Computational design of receptor and sensor proteins with novel functions. *Nature* 423:185–90

82. Lord P, Macdonald A. 2003. *e-Science Curation Report–Data Curation for e-Science in the UK: An Audit to Establish Requirements for Future Curation and Provision*. Twickenham, UK: Digital Archiving Consultancy Ltd.

83. Lord PW, Stevens RD, Brass A, Goble CA. 2003. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 19:1275–83

84. Mao X, Cai T, Olyarchuk JG, Wei L. 2005. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics* 21:3787–93

85. Marchal K, Thijs G, De Keersmaecker S, Monsieurs P, De Moor B, Vanderleyden J. 2003. Genome-specific higher-order background models to improve motif detection. *Trends Microbiol.* 11:61–66

86. Marengo E, Robotti E, Antonucci F, Cecconi D, Campostrini N, Righetti PG. 2005. Numerical approaches for quantitative analysis of two-dimensional maps: a review of commercial software and home-made systems. *Proteomics* 5:654–66

87. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–80

88. Matthews DE, Carollo VL, Lazo GR, Anderson OD. 2003. GrainGenes, the genome database for small-grain crops. *Nucleic Acids Res.* 31:183–86

89. Meyers BC, Galbraith DW, Nelson T, Agrawal V. 2004. Methods for transcriptional profiling in plants. Be fruitful and replicate. *Plant Physiol.* 135:637–52

90. Mockler TC, Ecker JR. 2005. Applications of DNA tiling arrays for whole-genome analysis. *Genomics* 85:1–15

91. Moreau Y, Aerts S, Moor B, Strooper B, Dabrowski M. 2003. Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet.* 19:570–77

92. Mueller LA, Solow TH, Taylor N, Skwarecki B, Buels R, et al. 2005. The SOL Genomics Network. A comparative resource for solanaceae biology and beyond. *Plant Physiol.* 138:1310–17

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

93. Myers EW. 1995. Toward simplifying and accurately formulating fragment assembly. *J. Comput. Biol.* 2:275–90

94. Nakazono M, Qiu F, Borsuk LA, Schnable PS. 2003. Laser-capture microdissection, a tool for the global analysis of gene expression in specific plant cell types: identification of genes expressed differentially in epidermal cells or vascular tissues of maize. *Plant Cell.* 15:583–96

95. Neumann E. 2005. A life science Semantic Web: Are we there yet? *Sci. STKE* 283:pe22

96. Newton RP, Brenton AG, Smith CJ, Dudley E. 2004. Plant proteome analysis by mass spectrometry: principles, problems, pitfalls and recent developments. *Phytochemistry* 65:1449–85

97. Noel JP, Austin MB, Bomati EK. 2005. Structure-function relationships in plant phenylpropanoid biosynthesis. *Curr. Opin. Plant Biol.* 8:249–53

98. Papin JA, Reed JL, Palsson BO. 2004. Hierarchical thinking in network biology: the unbiased modularization of biochemical networks. *Trends Biochem. Sci.* 29:641–47

99. Pappin DJ, Hojrup P, Bleasby AJ. 1993. Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* 3:327–32

100. Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, et al. 2005. ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* 33:D553–55

101. Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, et al. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294:1719–23

102. Pop M, Phillippy A, Delcher AL, Salzberg SL. 2004. Comparative genome assembly. *Brief Bioinform.* 5:237–48

103. Rhee SY. 2004. Carpe diem. Retooling the publish or perish model into the share and survive model. *Plant Physiol.* 134:543–47

104. Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, et al. 2003. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res.* 31:224–28

105. Rhee SY, Crosby B. 2005. Biological databases for plant research. *Plant Physiol.* 138:1–3

106. Rhodes D, Yu J, Shanker K, Deshpande N, Varambally R, et al. 2004. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl. Acad. Sci. USA* 101:9309–14

107. Roberts C, Nelson B, Marton M, Stoughton R, Meyer M, et al. 2000. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* 287:873–80

108. Roth FP, Hughes JD, Estep PW, Church GM. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* 16:939–45

109. Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg Å, Peterson C. 2002. BioArray Software Environment: a platform for comprehensive management and analysis of microarray data. *Genome Biol.* 3:software0003.1–.6

110. Schauer N, Steinhauser D, Strelkov S, Schomburg D, Allison G, et al. 2005. GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Lett.* 579:1332–37

111. Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270:467–70

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

112. Schlueter SD, Dong Q, Brendel V. 2003. GeneSeqer@PlantGDB: gene structure prediction in plant genomes. *Nucleic Acids Res.* 31:3597–600

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

114. Seo TS, Bai X, Kim DH, Meng Q, Shi S, et al. 2005. Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides. *Proc. Natl. Acad. Sci. USA* 102:5926–31

115. Shanks JV. 2005. Phytochemical engineering: combining chemical reaction engineering with plant science. *AIChE J.* 51:2–7

116. Shen L, Gong J, Caldo RA, Nettleton D, Cook D, et al. 2005. BarleyBase—an expression profiling database for plant genomics. *Nuceic Acids Res.* 33:D614–18

117. Sinha U, Bui A, Taira R, Dionisio J, Morioka C, et al. 2002. A review of medical imaging informatics. *Ann. NY Acad. Sci.* 980:168–97

118. Slonim DK. 2002. From patterns to pathways: gene expression data analysis comes of age. *Nat. Genet.* 32:502–8

119. SMRS Working Group. 2005. Summary recommendations for standardization and reporting of metabolic analyses. *Nat. Biotechnol.* 23:833–38

120. Sriram G, Fulton DB, Iyer VV, Peterson JM, Zhou R, et al. 2004. Quantification of compartmented metabolic fluxes in developing soybean embryos by employing biosynthetically directed fractional ^{13}C labeling, two-dimensional [^{13}C, ^{1}H] nuclear magnetic resonance, and comprehensive isotopomer balancing. *Plant Physiol.* 136:3043–57

121. Steuer R, Kurths J, Fiehn O, Weckwerth W. 2003. Interpreting correlations in metabolomic networks. *Biochem. Soc. Trans.* 31:1476–78

122. Steuer R, Kurths J, Fiehn O, Weckwerth W. 2003. Observing and interpreting correlations in metabolomic networks. *Bioinformatics* 19:1019–26

123. Stevens J, Doerge R. 2005. Combining Affymetrix microarray results. *BMC Bioinformatics* 6:57

124. Stevens R, Goble CA, Bechhofer S. 2000. Ontology-based knowledge representation for bioinformatics. *Brief Bioinform.* 1:398–414

125. Stevens RD, Robinson AJ, Goble CA. 2003. myGrid: personalised bioinformatics on the information grid. *Bioinformatics* 19(Suppl.)1:i302–4

126. Stoeckert CJ Jr, Causton HC, Ball CA. 2002. Microarray databases: standards and ontologies. *Nat. Genet.* 32(Suppl.):469–73

127. Stolc V, Samanta MP, Tongprasit W, Sethi H, Liang S, et al. 2005. Identification of transcribed sequences in Arabidopsis thaliana by using high-resolution genome tiling arrays. *Proc. Natl. Acad. Sci. USA* 102:4453–58

128. Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, Weinstein JN. 1999. MedMiner: an internet text-mining tool for biomedical information, with application to gene expression profiling. *BioTechniques* 27:1210–17

129. Tchieu JH, Fana F, Fink JL, Harper J, Nair TM, et al. 2003. The PlantsP and PlantsT Functional Genomics Databases. *Nucleic Acids Res.* 31:342–44

130. The_Arabidopsis_Genome_Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408:796–815

131. The_Wellcome_Trust. 2003. *Sharing Data from Large-Scale Biological Research Projects: A System of Tripartite Responsibility*. Fort Lauderdale, FL: Wellcome Trust

132. Toyoda T, Shinozaki K. 2005. Tiling array-driven elucidation of transcriptional structures based on maximum-likelihood and Markov models. *Plant J.* 43:611–21

133. Trethewey R. 2004. Metabolite profiling as an aid to metabolic engineering in plants. *Curr. Opin. Plant Biol.* 7:196–201

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

134. van Helden J. 2003. Regulatory sequence analysis tools. *Nucleic Acids Res.* 31:3593–96

135. Vincent PL, Coe EH, Polacco ML. 2003. Zea mays ontology—a database of international terms. *Trends Plant Sci.* 8:517–20

136. Wan X, Xu D. 2005. Computational methods for remote homolog identification. *Curr. Protein Peptide Sci.* 6:527–46

137. Ware DH, Jaiswal P, Ni J, Yap IV, Pan X, et al. 2002. Gramene, a tool for grass genomics. *Plant Physiol.* 130:1606–13

138. Weckwerth W, Loureiro M, Wenzel K, Fiehn O. 2004. Differential metabolic networks unravel the effects of silent plant phenotypes. *Proc. Natl. Acad. Sci. USA* 101:7809–14

139. Wheeler DL, Smith-White B, Chetvernin V, Resenchuk S, Dombrowski SM, et al. 2005. Plant genome resources at the national center for biotechnology information. *Plant Physiol.* 138:1280–88

140. Wiechert W, Mollney M, Petersen S, de Graaf AA. 2001. A universal framework for 13C metabolic flux analysis. *Metab. Eng.* 3:265–83

141. Wilkinson M, Schoof H, Ernst R, Haase D. 2005. BioMOBY successfully integrates distributed heterogeneous bioinformatics Web Services. The PlaNet exemplar case. *Plant Physiol.* 138:5–17

142. Woo Y, Affourtit J, Daigle S, Viale A, Johnson K, et al. 2004. A comparison of cDNA, oligonucleotide, and affymetrix GeneChip gene expression microarray platforms. *J. Biomol. Tech*. 15:276–84

143. Yamada K, Lim J, Dale JM, Chen H, Shinn P, et al. 2003. Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science* 302:842–46

144. Yamazaki Y, Jaiswal P. 2005. Biological ontologies in rice databases. An introduction to the activities in Gramene and Oryzabase. *Plant Cell Physiol.* 46:63–68

145. Yates JR 3rd, Eng JK, McCormack AL, Schieltz D. 1995. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.* 67:1426–36

146. Yona G, Levitt M. 2002. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J. Mol. Biol.* 315:1257–75

147. Yu J, Hu S, Wang J, Wong GK, Li S, et al. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* 296:79–92

148. Yuan Q, Ouyang S, Wang A, Zhu W, Maiti R, et al. 2005. The institute for genomic research Osa1 rice genome annotation database. *Plant Physiol.* 138:18–26

149. Zhang J, Leiderman K, Pfeiffer JR, Wilson BS, Oliver JM, Steinberg SL. 2006. Characterizing the topography of membrane receptors and signaling molecules from spatial patterns obtained using nanometer-scale electron-dense probes and electron microscopy. *Micron* 37:14–34

150. Zhang MQ. 2002. Computational prediction of eukaryotic protein-coding genes. *Nat. Rev. Genet.* 3:698–709

151. Zhang P, Foerster H, Tissier CP, Mueller L, Paley S, et al. 2005. MetaCyc and AraCyc. Metabolic pathway databases for plant research. *Plant Physiol.* 138:27–37

152. Zhu H, Bilgin M, Snyder M. 2003. Proteomics. *Annu. Rev. Biochem.* 72:783–812

153. Zhu T, Wang X. 2000. Large-scale profiling of the Arabidopsis transcriptome. *Plant Physiol.* 124:1472–76

154. Zhu W, Schlueter SD, Brendel V. 2003. Refined annotation of the Arabidopsis genome by complete expressed sequence tag mapping. *Plant Physiol.* 132:469–84

155. Ding J, Viswanathan K, Berleant D, Hughes L, Wurtele ES, et al. 2005. Using the biological taxonomy to access biological literature with PathBinderH. *Bioinformatics* 21:2560–62

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

DISCLOSURE STATEMENT

J.D. is a PI of the PLEXdb database that focuses on using Affymetrix GeneChips for cross-species comparison.

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

Annual Review of Plant Biology

Volume 57, 2006

Contents

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.

INDEXES

ERRATA

An online log of corrections to *Annual Review of Plant Biology* chapters (if any, 1977 to the present) may be found at http://plant.annualreviews.org/

Annu. Rev. Plant Biol. 2006.57:335-360. Downloaded from arjournals.annualreviews.org
by Stanford University Robert Crown Law Lib. on 05/10/07. For personal use only.