Full text provided by www.sciencedirect.com

Plant bioinformatics: from genome to phenome

David Edwards[1,2] and Jacqueline Batley[2]

[1]Plant Biotechnology Centre, Primary Industries Research Victoria, Department of Primary Industries, La Trobe University, Bundoora, Victoria 3086, Australia
[2]Victorian Bioinformatics Consortium, Plant Biotechnology Centre, La Trobe University, Bundoora, Victoria 3086, Australia

The vast quantities of diverse biological data generated by recent biotechnological advances have led to the development and evolution of the field of bioinformatics. This relatively new field facilitates both the analysis of genomic and postgenomic data and the integration of information from the related fields of transcriptomics, proteomics, metabolomics and phenomics. Such integration enables the identification of genes and gene products, and can elucidate the functional relationships between genotype and observed phenotype, thereby permitting a system-wide analysis from genome to phenome. With the increasing value and throughput of plant biotechnology, bioinformatics is being called on to integrate the varied data generated by the expanding '-omic' technologies.

Following recent advances in technology and the development of ultra high-throughput research, the field of biotechnology is beginning to suffer from data overload. This has led to the development of a broadening field of science, termed bioinformatics, in which biology and information technology converge. As such, bioinformatics is often considered to be different things by different people. In its most basic form, bioinformatics might be described as 'the structuring of biological information to enable logical interrogation'.

Recent advances in genomic technologies have led to an explosion of data and a huge growth in bioinformatics within both plant biotechnology and the broader biomedical sciences. Applications of bioinformatics have expanded with the so-called '-omic' technologies, and this discipline now sits as an umbrella over biotechnology. The new challenges facing the field of bioinformatics are to provide both complex data integration across the -omic platforms and a direct link between traditional genetics – through the genome, transcriptome, proteome and metabolome – and the observed phenotype of the plant. Researchers now want more than candidate functions for DNA sequences or predicted structures for proteins. Biotechnology demands intelligent searching and filtering of numerous, complex data types to address specific issues, ranging across specialist research fields outside the knowledge of any one individual. Although bioinformatics is expanding its applications alongside the rise of the new -omic and postgenomic technologies, its focus and strengths remain in the analysis of DNA sequences and genomes.

Genomics

Modern bioinformatics came of age with the development of genomic technologies, specifically the ability to produce large amounts of sequence information at an ever-decreasing cost. High-throughput gene discovery by expressed sequence tag (EST) sequencing, initiated in 1991 [1], set the requirement for large and searchable sequence databases. Although EST sequencing is still the standard procedure for gene discovery in many crops, a reduction in the cost of DNA sequencing has led to a move towards whole-genome sequencing.

Plant genomics was revolutionized by the release of the complete *Arabidopsis thaliana* genome sequence by the Arabidopsis Genome Initiative in 2000, four years ahead of schedule [2]. Two years later, the completion of the rice (*Oryza sativa* L ssp. *japonica* Nipponbare) genome sequence by public consortia was announced (see USDA News Release; Table 1). This work was complemented by rice sequencing work undertaken by the agribusinesses Syngenta [3] and Monsanto [4], and a separate research project at the Beijing Genomics Institute that sequenced the rice subspecies *indica* [5]. Owing to similarities at the genomic level between rice and other important crop species [6], completion of the rice genome has had a significant impact on both plant biotechnology and crop bioinformatics.

The availability of complete-genome sequences, as well as the flood of sequence data, is leading to alternative views on how these data can be organized and interrogated. The high level of redundancy in gene discovery programs is being condensed through reference to consensus or complete-genome sequences. If a complete-genome sequence is unavailable for a specific crop, closely related syntenic genomes can be used. The ever-increasing size of DNA sequence databases continues to push bioinformatic capabilities, and there is a growing need to condense redundant data. Database development has been accompanied by progress in tools for data analysis, enabling researchers to annotate sequences more fully and to mine complex interacting data for valuable biological knowledge.

Databases

The rapid growth in DNA sequence information required the development of specific DNA sequence databases. The

Corresponding author: David Edwards (dave.edwards@dpi.vic.gov.au).

www.sciencedirect.com 0167-7799/$ - see front matter © 2004 Elsevier Ltd. All rights reserved. doi:10.1016/j.tibtech.2004.03.002

Table 1. Relevant URLs

Website	URL	Refs
Genomics		
Syngenta	http://www.syngenta.com	[3]
Monsanto	http://www.monsanto.com	[4]
USDA Press Release 18 December 2002	http://www.usda.gov/news/releases/2002/12/0515.htm	
Data and databases		
AceDB	http://www.acedb.org/	
BIOVIZ Genome Viewer	http://www.svgopen.org/2002/abstracts/lewis_et_al__bioviz_genome_viewer.html	[11]
Gene Expression Omnibus	http://www.ncbi.nlm.nih.gov/geo/	
Brassica Gene Ontology Page	http://hornbill.cspp.latrobe.edu.au/cgi-binpub/goindex.pl	
Kyoto Encyclopedia of Genes and Genomes	http://www.genome.ad.jp/kegg/	[48,49]
MicroArray Software Catalogue	http://www.cs.tcd.ie/Nadia.Bolshakova/softwarecatalogue.html	
Directory of MPSS[a] Data Pages	http://mpss.dbi.udel.edu/	
SwissProt	http://www.expasy.org/sprot/	
Bioinformatics		
Wheat: the Big Picture	http://www.wheatbp.net/	
How a Corn Plant Develops	http://maize.agron.iastate.edu/corngrows.html	
ISI Web of Knowledge	http://wok.mimas.ac.uk/	

[a] Abbreviation: MPSS, massively parallel signature sequencing.

largest of these sequence databases emerged in 1986 from the collaboration of GenBank and EMBL, and was joined the following year by the DNA Data Bank of Japan. This meta-sequence database is considered to be the standard repository for public DNA sequences worldwide and contains over 7.4 million plant DNA sequences. Furthermore, the consolidation of public databases led to the application of a common 'feature table' format and common standards for annotation practice. Feature table design provided an extensive vocabulary for describing features and was a precursor to current extensible mark-up language (XML) formats, which provide standards for structuring data across databases.

The GenBank meta-database has remained the repository of choice for DNA sequence data, and it contains vast quantities of organism-specific data. However, in addition to this general sequence data bank, there developed a demand for species-specific sequence databases that could incorporate analytical, visualization and interrogation tools. One of the first such databases, AceDB, was introduced in 1989 (Table 1). AceDB provides a custom database with a graphical user interface and tools for structuring and interrogating genomic data. Although AceDB was initially developed for sequence data from the *Caenorhabditis elegans* genome project, it was rapidly adopted for crop species and remains one of the principal database formats for plant DNA sequences.

Alongside the increase in numbers of sequencing projects, based on numerous crops in different laboratories, has been a concomitant growth in plant databases. AceDB provided an early model for genome databases but the variety of formats has expanded to suit the specific demands of different users. There are now a multitude of database schemata to choose from depending on the requirements of the users, and schemata are frequently modified to suit individual needs.

One of the more significant changes to crop genome databases has been the move towards graphical user interfaces that provide a more user-friendly search environment. Although graphical user interfaces have been developed for AceDB, more recent crop databases have used the Ensembl database schema [7], which has a strong emphasis on graphical user interaction. Ensembl was initially developed as part of the Human Genome Project, and its facility for viewing related data from several different organisms made it an ideal model for the cereal comparative genomic database Gramene [8–10]. A recent advance in crop database interfaces is the use of a standardized client-based scalable vector graphics viewer [11], which enables data views to be manipulated without the need for constant web page updating.

The diversity of database formats and interfaces reflects the needs of various research groups, but such diversity creates a challenge for bioinformatics because it reduces the scope for data integration and interrogation across databases. With the maturation of genomics has come a move towards the adoption of standard data formats and schemata for crop genome information, and it is likely that future databases will be designed with cross-connectivity capabilities as a priority. Complex biological data integration can be also driven by developments in grid computing [12].

Tools

Primary tools for sequence comparison and assembly have grown in line with an expansion of the datasets that they analyze. Without basic local alignment search tool (BLAST) [13] and related sequence comparison tools, much of the data coming from the many high-throughput sequencing laboratories would be nothing more than strings of letters. BLAST remains the fastest means by which to identify specific sequences in large datasets and enables the rapid annotation of novel sequences. Although BLAST is the standard tool for identifying sequence similarities in large datasets, there are several options for assembling sequence datasets, the choice of which depends on hardware availability, dataset size, data format, structure and the genetic structure of the organism.

Sequence similarity search and assembly tools are the foundation of many software applications for analyzing crop genomic information. The ability to rapidly identify similarities to previously characterized sequences greatly enhances the sequence annotation process and has led to the development of comparative sequence databases, whereas sequence assembly packages both reduce the high level of redundancy in datasets and enable variations in sequence to be identified.

The availability of large sequence datasets permits mining for biological features – for example, single nucleotide polymorphism [14–16] and simple sequence repeat [17] molecular markers – that can be then applied to plant biotechnology research such as genetic trait mapping [18]. The availability of complete-genome sequences enables further mining for novel promoter sequences [19,20] and other regulatory features such as micro-RNAs [21,22]. This tertiary level annotation provides links to both the phenotype and the complex regulatory mechanisms that govern development and response to the environment.

Transcriptomics

The application of microarrays and sequence-based methods to expression profiling has added an extra dimension to current genomic data and has founded several statistics-based disciplines within bioinformatics.

Owing to their extended linear dynamics, sequence-based methods have the potential to determine more accurately quantitative levels of gene expression. Furthermore, they do not require prior sequence information and so have the advantage of being able to identify novel genes or to assess gene expression in uncharacterized plants. With the scaling-up of EST sequencing projects, it is becoming possible to mine these datasets to estimate expression information [23], although this remains more a byproduct of EST sequencing than a true transcriptomic tool.

The predominant methods for sequence-based expression analysis are serial analysis of gene expression (SAGE) [24] and massively parallel signature sequencing (MPSS) [25,26]. Of these, only SAGE has been broadly adopted for plant genomes. Although MPSS provides several major benefits over SAGE, the high costs involved with the process have led to its limited implementation in public plant biotechnology research. The use of MPSS for annotating genomes with expressed genes is likely to lead to the wider adoption of MPSS for crop species, and the availability of public plant MPSS data is increasing (Table 1).

Hybridization-based microarrays have become the transcriptomic tool of choice, probably because they can be used to analyze several samples simultaneously. The rapid implementation of microarrays has been followed by a growth in the bioinformatics of microarray data analysis [27,28]. This field has expanded from the initial examination of twofold differences in expression, to the incorporation of complex statistical models including Lowess normalization, hidden Markov modeling and Bayesian statistics [29–31]. There are also a plethora of software tools for analyzing microarray data (see MicroArray Software Catalogue; Table 1) and, with the continued growth in methods for array data analysis, it is unlikely that a standard, uniform procedure will be developed soon.

There is a direct relationship between genes and their expression, but the process of quantifying microarray-based measurements leads to difficulties in making direct comparisons between experiments. Efforts are being made to standardize microarray experiments and repositories such as Gene Expression Omnibus (Table 1) are being used more widely. With continued developments in the field of microarray data production, significant improvements in data analysis and integration are necessary before these data can be structured for efficient interrogation.

Microarray technology continues to expand: cDNA arrays are being produced for gene expression analysis in many plant species, and complete oligonucleotide-based Unigene arrays are being developed for the major plant species. The development of technologies for 'one-off' custom production of oligonucleotide expression arrays is likely to lead to a spread of microarray technology into niche applications in expression analysis and genotyping. Data produced from these increasing numbers of unique array designs will extend the need for complex data integration and analysis [27].

Proteomics

The term 'proteomics' was coined in the mid 1990s [32,33] on the back of the success of 'genomics' and has since come to incorporate many aspects of protein biochemistry. The bioinformatics of proteomics predates the term in the form of databases of predicted protein sequences, which were mostly an outcome of the growth of genomic and high-throughput sequencing. Proteomics currently encompasses databases of protein sequences, databases of predicted protein structures and, more recently, databases of protein expression analysis, and the field is expanding with emerging technologies.

The principal protein sequence database remains SwissProt (Table 1), which was established in 1986 as a repository for predicted protein sequences and now contains multi-level protein data [34,35]. In the bid to link the genome and the proteome with associated phenotypes, there has been a push towards the prediction of protein structures in relation to their sequence. This drive has come mainly from the pharmaceutical industry, although structure prediction has applications in plant biotechnology research.

The development of more accurate algorithms for predicting protein structure is moving protein structure elucidation out of the laboratory and into the hands of bioinformaticists [36]. As more protein structures are identified, the relationship between structure and function becomes easier to predict. Databases of protein structures and comparative structure tools also facilitate the identification of common structures and predicted functions [37,38]. Another challenge facing the plant biotechnology and bioinformatic research community is the translation of complete-genome DNA sequence data into protein structures and predicted functions: such a step will provide the vital link between the genetics of an organism and its expressed phenotype. A comparison of the numbers of current plant protein sequences with predicted structures suggests that there is much scope for research in this area.

Knowledge of the structure and function of every protein would revolutionize the field of proteomics. A further challenge is the high-throughput determination of protein expression patterns. Protein expression is predominantly determined by two-dimensional gel electrophoresis and protein spot characterization through molecular mass determination. Two-dimensional gel electrophoresis technology is progressing towards the detection of smaller and smaller quantities of protein, and the application of fluorescent dye labeling enables the accurate determination of quantitative differences between two samples [39]. This, along with highly accurate methods for molecular mass determination and databases of predicted protein fragments, permits the rapid identification of not only the complete predicted protein sequence and the related DNA gene sequence, but also any post-translational modifications such as phosphorylation [40–43].

Recent advances in protein detection arrays and high-throughput antigen studies are being applied in the biomedical sector [44–46]. These advances are likely to have an impact on plant biotechnology, particularly in the fields of pharmaceutical and/or nutriceutical development, although applications of these technologies in plant biotechnology are limited at present.

Proteomics has significant prospects for advancing our understanding in plant biotechnology owing to its direct relationship with gene and transcript data. Proteomes also have a strong influence on the measured phenotype of the plant, either directly through protein content or function or indirectly through the relationship of a protein with the metabolome. The potential for bioinformatics to structure and integrate -omic data, therefore, relies on an ability to model both the proteome and its interactions.

Metabolomics

Like proteomics, metabolomics was derived from the field of biochemistry and involves the analysis (usually high throughput or broad scale) of small-molecule metabolites and polymers such as starch. The foundations of metabolomics are descriptions of biological pathways and current metabolomic databases, such as Kyoto Encyclopedia of Genes and Genomes [47,48] (Table 1), are frequently based on well-characterized biochemical pathways.

On a more applied level, the bioinformatics of metabolomics involves the identification and characterization of a broad range of metabolites through reference to quantitative biochemical analysis. Although this field is relatively new, there have been significant recent advances [49] and there is scope for many direct applications in plant biotechnology [50,51].

Metabolomics might be considered to be the key to integrated systems biology because it is frequently a direct gauge of desired phenotype [52], measuring quantitative and qualitative traits such as starches in cereal grains or oils in oilseeds. Moreover, metabolomes can be correlated with genetics through proteomes, transcriptomes and genomes and therefore bypass the more traditional quantitative trait locus approach applied to molecular crop breeding. One of the challenges for bioinformatics will be the structuring and integration of these diverse types of data for the emerging field of systems biology [49,53,54].

Other -omics technologies

Genomics has spawned a plethora of related -omics terms that frequently relate to established fields of research. Of these terms, 'phenomics', the high-throughput analysis of phenotypes, has probably the biggest application in plant biotechnology. The great plasticity of plant genomes in producing various phenotypes from little genetic variation has provided both challenges and opportunities for crop improvement. The detailed and systematic analysis of phenotype requires both a data repository and a means of structured interrogation. The field of phenomics developed from the phenotypic characterization of mutant plants, the descriptions of which have been published in volumes that frequently use structured ontological terms (e.g. see Refs [55–57]). The storage of these data in searchable databases, together with the application of phenomics to high-throughput analysis [58], plant development (e.g. see Wheat: the Big Picture and How a Corn Plant Develops; Table 1) and natural variation [59], creates the final link in the chain from the genetics of crop development to crop production.

Another area of data integration has, from its inception, required data structuring and query systems. The current literature databases in the field of 'bibliomics' (bibliographic reference data management) were founded on printed reference lists. These have since been integrated into web-searchable forms, such as the ISI Web of Knowledge (Table 1). Although there has been some integration of bibliographic resources in species-specific databases, such integration has generally required considerable manual input. Thus, there is scope in bioinformatics for the automated integration of bibliographic references with biological datasets [60].

Data integration: from genome to phenome

Bioinformatics arose from the need to structure and to interrogate the ever-increasing quantity and forms of biological data being generated through the developing -omic technologies. As these technologies continue to grow, so does the need for such integration and interrogation across the various types of data and scientific disciplines. Precise data integration requires the formal annotation of data with relational terms, and this is an essential driver behind applications of bioinformatics in the development of systems biology.

Although manual annotation of DNA sequences has been considered to be the 'gold standard' because it is less likely to accumulate and build on previous errors, automated annotation by sequence comparison tools is making a resurgence owing to its lower cost and the reduction in bias or variation that is inherent in manual annotation [61]. Whereas primary annotation of sequences is usually performed using sequence comparison tools (BLAST searches of DNA and protein databases), secondary annotation (e.g. genetic or physical map position, gene expression data and predicted protein structure) provides data integration and greater insight into potential gene function.

A current limitation of complex annotation and integration is the lack of agreed formats across databases. This is being addressed by the use of 'gene ontology' terms for protein and gene sequences [62,63], minimum information about a microarray experiment (MIAME) standards for microarray experiments [64], and plant ontologies for broader plant-based database information [65]. Data integration is also being assisted through the use of agreed XML standards. At present, only the *Arabidopsis* and rice plant genomes have primary gene ontology annotation. However, sequence comparison tools permit the application of this primary annotation to related species. For example, *Brassica* EST sequences can be identified through their similarity to *Arabidopsis* sequences with specific gene ontology annotation (see *Brassica* Gene Ontology Page; Table 1).

The formal annotation of diverse datasets is complemented by the parallel analysis of related data in the emerging field of systems biology. The integration and structured interrogation of metabolome and transcriptome datasets are already yielding results [66], providing the basis for the integration of genome and phenome data. Linking gene expression, protein sequence and protein structure data with genetic and physical map data will integrate genetics, genomics, transcriptomics and proteomics. The further incorporation of metabolomic data and data from phenotype studies will close the loop and create the foundation for advanced knowledge bases – in other words, meta-integrated databases that facilitate queries across whole-systems biology.

References

1 Adams, M.D. *et al.* (1991) Complementary-DNA sequencing – expressed sequence tags and human genome project. *Science* 252, 1651–1656

2 The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815.

3 Goff, S.A. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). *Science* 296, 92–100

4 Barry, G.F. (2001) The use of the Monsanto draft rice genome sequence in research. *Plant Physiol.* 125, 1164–1165

5 Yu, J. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp *indica*). *Science* 296, 79–92.

6 Moore, G. *et al.* (1995) Grasses, line up and form a circle. *Curr. Biol.* 5, 737–739

7 Hubbard, T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.* 30, 38–41

8 Ware, D.H. *et al.* (2002) Gramene, a tool for grass genomics. *Plant Physiol.* 130, 1606–1613

9 Ware, D.H. *et al.* (2002) Gramene: a resource for comparative grass genomics. *Nucleic Acids Res.* 30, 103–105

10 Jaiswal, P. *et al.* (2002) Gramene: development and integration of trait and gene ontologies for rice. *Comp. Funct. Genomics* 3, 132–136

11 Lewis, C.T. *et al.* (2003) The *Brassica/Arabidopsis* comparative genome browser: a novel approach to genome browsing. *J. Plant Biotechnology* 5, 197–200

12 Rungsarityotin, W. *et al.* (2002) Grid computing and bioinformatics development. A case study on the *Oryza sativa* (rice) genome. *Pure Appl. Chem.* 74, 891–897

13 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410

14 Barker, G. *et al.* (2003) Redundancy based detection of sequence polymorphisms in express sequence tag data using AutoSNP. *Bioinformatics* 19, 421–422

15 Batley, J. *et al.* (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiol.* 132, 84–91

16 Somers, D.J. *et al.* (2003) Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. *Genome* 46, 431–437

17 Robinson, A.J. *et al.* (2004) Simple sequence repeat marker loci discovery using SSR primer. *Bioinformatics* 10.1093/bioinformatics/bth104 (www.bioinformatics.oupjournals.org)

18 Morgante, M. and Salamini, F. (2003) From Plant genomics to breeding practice. *Curr. Opin. Biotechnol.* 14, 214–219

19 Qui, P. (2003) Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. *Biochem. Biophys. Res. Commun.* 309, 495–501

20 Ettwiller, L.M. *et al.* (2003) Discovering novel *cis*-regulatory motifs using functional networks. *Genome Res.* 13, 883–895

21 Rhoades, M.W. *et al.* (2002) Prediction of plant microRNA targets. *Cell* 110, 513–520

22 Nelson, P. *et al.* (2003) The microRNA world: small is mighty. *Trends Biochem. Sci.* 28, 534–540

23 Rafalski, J.A. *et al.* (1998) New experimental and computational approaches to the analysis of gene expression. *Acta Biochim. Pol.* 45, 929–934

24 Velculescu, V.E. *et al.* (1995) Serial analysis of gene expression. *Science* 270, 484–487

25 Brenner, S. *et al.* (2000) Gene expression analysis by massively parallel signal sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* 18, 630–634

26 Brenner, S. *et al.* (2000) *In vitro* cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proc. Natl. Acad. Sci. U. S. A.* 97, 1665–1670

27 Moreau, Y. *et al.* (2003) Comparison and meta analysis of microarray data: from the bench to the computer desk. *Trends Genet.* 19, 570–577

28 Goodman, N. (2002) Biological data becomes computer literate: new advances in bioinformatics. *Curr. Opin. Biotechnol.* 13, 68–71

29 Leung, Y.F. and Cavalieri, D. (2003) Fundamentals of cDNA data analysis. *Trends Genet.* 19, 649–659

30 Pan, W. (2003) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* 14, 546–554

31 Detours, V. *et al.* (2003) Integration and cross-validation of high-throughput gene expression data: comparing heterogeneous data sets. *FEBS Lett.* 546, 98–102

32 Wilkins, M. *et al.* (1996) From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Nat. Biotechnol.* 14, 61–65

33 Wilkins, M., *et al.* eds (1997) *Proteome Research: New Frontiers in Functional Genomics*, Springer-Verlag

34 Boeckmann, B. *et al.* (2003) The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31, 365–370

35 Gasteiger, E. *et al.* (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* 31, 3784–3788

36 Schwede, T. *et al.* (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.* 31, 3381–3385

37 Maggio, E.T. and Ramnarayan, K. (2001) Recent developments in computational proteomics. *Trends Biotechnol.* 19, 266–272

38 Norin, M. and Sundström, M. (2002) Structural proteomics: developments in structure-to-function predictions. *Trends Biotechnol.* 20, 79–84

39 Heazlewood, J.L. and Millar, A.H. (2003) Integrated plant proteomics – putting the green genomes to work. *Funct. Plant Biol.* 30, 471–482

40 Watson, B.S. *et al.* (2003) Mapping the proteome of barrel medic (*Medicargo truncatula*). *Plant Physiol.* 131, 1104–1123

41 Koller, A. *et al.* (2002) Proteomic survey of metabolic pathways in rice. *Proc. Natl. Acad. Sci. U. S. A.* 99, 11969–11974

42 Gallardo, K. *et al.* (2003) Proteomics of *Medicago truncatula* seed development establishes the time frame of diverse metabolic processes related to reserve accumulation. *Plant Physiol.* 133, 664–682

43 Bae, M.S. *et al.* (2003) Analysis of the *Arabidopsis* nuclear proteome and its response to cold stress. *Plant J.* 36, 652–663

44 Templin, M.F. *et al.* (2002) Protein microarray technology. *Trends Biotechnol.* 20, 160–166

45 Zhu, H. and Snyder, M. (2002) 'Omic' approaches for unravelling signalling networks. *Curr. Opin. Cell Biol.* 14, 173–179

46 Lee, Y.S. and Mrksich, M. (2002) Protein chips: from concept to practice. *Trends Biotechnol.* 20, S14–S18

47 Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto encylopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30

48 Kanehisa, M. *et al.* (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.* 30, 42–46

49 Fernie, A.R. (2003) Metabolome characterisation in plant systems analysis. *Funct. Plant Biol.* 30, 111–120

50 Roessner, U. *et al.* (2002) Metabolic profiling and biochemical phenotyping of plant systems. *Plant Cell Rep.* 21, 189–196

51 Fiehn, O. *et al.* (2000) Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* 18, 1157–1161

52 Fiehn, O. (2002) Metabolomics – the link between genotypes and phenotypes. *Plant Mol. Biol.* 48, 155–171

53 Sumner, L.W. *et al.* (2003) Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* 62, 817–836

54 Weckworth, W. (2003) Metabolomics in systems biology. *Annu. Rev. Plant Physiol.* 54, 669–689

55 Neuffer, M.G., *et al.* eds (1996) *Mutants of Maize*, Cold Spring Harbor Laboratory Press

56 Bowman, J. ed. (1994) *Arabidopsis: An Atlas of Morphology and Development*, Springer-Verlag

57 Kiesselbach, T. ed. (1999) *The Structure and Reproduction of Corn*, Cold Spring Harbor Laboratory Press

58 Parvin, B. *et al.* (2002) BioSig: an imaging bioinformatic system for studying phenomics. *Comput.* 35, 65–71

59 Nevo, E. (2001) Evolution of genome–phenome diversity under environmental stress. *Proc. Natl. Acad. Sci. U. S. A.* 98, 6233–6240

60 Raychaudhuri, S. *et al.* (2003) The computational analysis of scientific literature to define and recognise gene expression clusters. *Nucleic Acids Res.* 31, 4553–4560

61 Kasukawa, T. *et al.* (2003) Development and evaluation of an automated annotation pipeline and cDNA annotation system. *Genome Res.* 13, 1542–1551

62 Camon, E. *et al.* (2003) The Gene Ontology (GOA) project: implementation of GO in Swiss-Prot, TrEMBL and Interpro. *Genome Res.* 13, 662–672

63 Consortium, T.G.O. (2001) Creating the gene ontology resource: design and implementation. *Genome Res.* 11, 1425–1433

64 Brazma, A. *et al.* (2002) Minimum information about a microarray experiment (MIAME) – towards standards for microarray data. *Nat. Genet.* 29, 365–371

65 Bruskiewich, R. *et al.* (2002) The Plant Ontology™ Consortium and plant ontologies. *Comp. Funct. Genomics* 3, 137–142

66 Urbanczyk-Wochniak, E. *et al.* (2003) Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep.* 4, 989–993