1

The use of functional genomics to understand components of plant metabolism and the regulation occurring at molecular, cellular and whole plant levels

Paolo Pesaresi[1,2]

[1]Parco Tecnologico Padano, Via Einstein, Loc. Cascina Codazza, 26900, Lodi, Italy

[2]Dipartimento di Produzione Vegetale, Università Statale di Milano, Via Celoria 2, 20133 Milano, Italy (e-mail: paolo.pesaresi@tecnoparco.org)

1. Introduction

The completion of the genome sequence of the small weed plant *Arabidopsis thaliana* (The Arabidopsis genome initiative 2000), and more recently of rice (Goff et al. 2002; Yu et al. 2002, 2005), has greatly changed the face of plant biology. Knowing the exact sequence and location of all the genes of a given organism is the first step towards understanding how all parts of a biological system work together. Information about the hypothesized function of an unknown gene may be deduced from its sequence homology to other genes of known function. However, genome sequencing projects have revealed the existence of a tremendous amount of biological diversity, with large proportion of genes sharing no homology to genes with known or hypothesized functions. In this respect functional genomics is the key approach to transforming quantity into quality (Borevitz and Ecker 2004; Holtfort et al. 2002). Functional genomics is a general approach toward understanding how the genes of an organism work

P. Ranalli (ed.), Improvement of Crop Plants for Industrial End Uses, 1–26.

together by assigning new functions to unknown genes. For efficient gene function analysis, researchers can choose from a multitude of different methods, most of them derived from genomic research performed on model organisms such as yeast, nematodes, flies and mice, not forgetting the technological spin-offs that were inspired by the human genome project. Arabidopsis populations, mutagenized by random insertion of T-DNA or transposon elements, have been generated with the aim to perform high-throughput reverse genetics studies and comprehensive forward genetics studies of the entire gene compendium (Alonso et al. 2003). Additionally, information about the spatial and temporal expression pattern of a gene can be gained from analysis of qualitative and quantitative changes of messenger RNAs, proteins, and metabolites. These techniques, able to simultaneously analyze large numbers of transcripts, proteins and chemical constituents, have led to the creation of new research fields within functional genomics, named transcriptomics, proteomics, and metabolomics. Each method has its inherent limitations and none of them alone is sufficient to assign a function to a gene of interest. However, the organization of the vast amount of data from the various approaches into central databases allows easy extraction and comparison of meaningful information.

This chapter has the aim to highlight the major approaches that makes up modern plant functional genomics and to describe how they add a new dimension to the comprehension of plant biology with particular emphasis to the model plant, *Arabidopsis thaliana*.

2. Plant genome sequences

The public effort to sequence the genome of the model flowering plant, *Arabidopsis thaliana*, was completed in December 2000 (The Arabidopsis genome initiative 2000), and it was the third complete genome of a higher eukaryote, after *Drosophila melanogaster*, and *Caenorhabditis elegans*. This tiny mustard plant, a common weed of the Brassicaceae family, had been chosen as the first reference plant to be sequenced, because it has several advantages over other species. Its nuclear genome is very small with 115 million base pairs (Mb) of euchromatin out of the estimated 125 Mb total, its generation time is very short, and it is genetically very well characterized. The total number of Arabidopsis genes revealed by the five chromosome sequences was initially estimated at 25,490 and later revised to 30,700 (version 5 annotation), resulting in 11,000-15,000 gene families, with about one quarter of genes believed to be plant specific. The entire *Arabidopsis* genome dataset has been stored and can be retrieved from different user-friendly databases, including TAIR, TIGR, MIPS and NCBI (Table 1.1). In 2002, two groups released the second plant genome sequence, rice. A four time shotgun coverage of *Oryza sativa* ssp. *indica* covered 361 Mb of the estimated 466 Mb (Yu et al. 2002). *Oryza sativa* ssp *japonica* was sequenced to five times shotgun coverage (Goff et al. 2002) and resulted in 372 Mb of non-overlapping sequence from the 12 rice chromosomes and 55,890 genes were identified (version 4 annotation). More recently, improved whole genome shotgun sequences for the genomes of indica and japonica rice have been reported (Yu et al. 2005). Sequences and analysis details of rice genome are available at several databases including TIGR, and Rice Genome research Program (Table 1.1). The completely annotated reference genomes of Arabidopsis and rice certainly serve as a starting point for the large-scale functional analysis of other plant genomes. Indeed, many other plant species have entered the genomics era. In particular, the Joint Genome Institute (JGI) has essentially finished a deep draft genome sequence of the model tree *Populus trichocarpa* (cottonwood). Ten-times shotgun coverage, amounting to 5.5 Gb, is now available for download and BLAST searches at the JGI database (Table 1.1). Moreover, approximately 80,000 ESTs have been sequenced and will certainly be helpful for gene annotation. *Medicago truncatula* has been chosen as the model legume plant for genomics studies. The complete genome of the first legume will certainly speed up the comprehension of the molecular mechanisms responsible for the conversion of molecular nitrogen into usable organic forms (legume/*Rhizobium* symbiosis). Another crop plant that has been used as a model for decades is *Lycopersicon esculentum* (tomato). The genome is about 900 Mb and its sequencing is a priority according to the National Plant Genomics Initiative. At the moment more than 150,000 *Lycopersicon esculentum* ESTs, stored at Cornell University (Table 1.1), are available. Additionally, the sequencing of all twelve chromosomes has been initiated.

3. Genome-wide insertional mutagenesis

One of the most significant findings revealed through analysis of plant genomes is the large number of genes for which no function is known or can be predicted. An essential tool for the functional analysis of these completely sequenced genomes is the ability to create loss-of-function mutations for all the genes (Borevitz and Ecker 2004). A knockout line can provide a crucial second allele when only a single EMS allele is available. Here observation of similar phenotypes in both alleles makes sure that the correct gene was identified. The null genetic background is suitable for transgenic studies that investigate altered expression patterns or test altered proteins. Often redundancy can confound genetic studies by masking phenotypes. This problem can be dealt with by creating double, triple, or greater knockout mutations among multiple gene family members. Although, targeted gene replacement via homologous recombination is extremely facile in yeast, its efficiency in plants does not yet allow for the creation of a set of genome-wide gene disruptions (Gong and Rong 2003; Parinov and Sundaresan 2000). Additionally, gene silencing via the RNA interference (RNAi) method has several drawbacks, including the lack of stable heritability of a phenotype, variable levels of residual gene activity, and the inability to simultaneously silence several unrelated genes (Hannon 2002). Because of these disadvantages, the random insertions of T-DNA or transposon elements has become the strategy of choice to generate loss-of-function Arabidopsis populations. In particular, Alonso and colleagues (2003) have generated about 150,000 transformed plants carrying one or more T-DNA elements. After sequencing analysis, about 88,000 T-DNA integration sites were identified, resulting in the disruption of 21,799 genes, about 74% of the Arabidopsis gene repertoire (T-DNAexpress, Table 1.1). Several other studies have created sequence-index collections of knockout mutations, including the French program Inra/Genoplante (Samson et al. 2002), the German project GABI-Kat (Li et al. 2003), the Japanese group in RIKEN (Kuromori et al. 2004), and the Cold Spring Harbor Laboratory (Table 1.1). Nowadays, there are more than 360,000 Arabidopsis flanking sequences that hit more than 90% of the currently known genes. Nearly all sequences can be searched at the signal T-DNAexpress website and corresponding seeds can be ordered both at the European Arabidopsis stock center and at the Arabidopsis Biological Resource Center (ABRC), located at the Ohio State University (Table 1.1).

Table 1.1. Websites relevant to plant functional genomics.

Plant genomic databases	
The Arabidopsis Information Resource (TAIR)	http://www.arabidopsis.org/
The Institute for Genomics Research (TIGR)	http://www.tigr.org/plantProjects.shtml
Munich Information center for Protein Sequences (MIPS)	http://mips.gsf.de/projects/plants
National Center for Biotechnology Information (NCBI)	http://www.ncbi.nlm.nih.gov/
Rice Genome research Program (RGP)	http://rgp.dna.affrc.go.jp/
Joint Genome Institute (JGI)	http://www.jgi.doe.gov/
Tomato Expression Database	http://ted.bti.cornell.edu/

Table 1.1. Websites relevant to plant functional genomics.

T-DNAexpress	http://signal.salk.edu/cgi-bin/tdnaexpress
Inra/Genoplante	http://urgv.evry.inra.fr/projects/FLAGdb++/HTML/index.shtml
GABI-Kat	http://www.gabi-kat.de/
RIKEN	http://rarge.gsc.riken.jp/dsmutant/index.pl
Cold Spring Harbor Laboratory	http://genetrap.cshl.org/
European Arabidopsis Stock Center	http://arabidopsis.info/
Arabidopsis Biological resource center (ABRC)	http://www.arabidopsis.org/abrc/
Microarray databases	
NASCArrays	http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl
ArrayExpress	http://www.ebi.ac.uk/arrayexpress/
Gene Expression Omnibus (GEO)	http://www.ncbi.nlm.nih.gov/geo/
GENEVESTIGATOR	https://www.genevestigator.ethz.ch/
Proteomics databases	
Swiss-prot	http://www.expasy.org/sprot/ppap/
Plastid Proteome Database (PPDB)	http://ppdb.tc.cornell.edu/
Arabidopsis Mitochondrial Protein Database (AMPD)	http://www.plantenergy.uwa.edu.au/applications/ampdb/index.html
Aramennon database	http://aramemnon.botanik.uni-koeln.de/
Protein-GFP fusions	http://deepgreen.stanford.edu/
Subcellular location database Arabidopsis (SUBA)	http://www.plantenergy.uwa.edu.au/applications/suba/index.php
Plant Phosphorylation database (PlantP)	http://PlantsP.sdsc.edu.
Metabolomics databases and Computational tools	
Golm Metabolome Database (GMD)	http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html
Kyoto Encyclopedia of Genes and Genomes (KEGG)	http://www.genome.jp/kegg/
BRaunschweig ENzyme DAtabase (BRENDA)	http://www.brenda.uni-koeln.de/
AraCyc	http://www.arabidopsis.org/tools/aracyc
MAPMAN	http://gabi.rzpd.de/projects/MapMan/
MetNet	http://metnet.vrac.iastate.edu/

4. Forward and reverse genetics-essential steps towards understanding gene function

The classical or "forward genetics" approach to gene function analysis aims to identify the sequence change that underlies a specific mutant phenotype (Ostergaard and Yanofsky 2004). In the recent past, the starting point involved mutagenesis of a large number of wild-type seeds by treatment with chemical reagents or irradiation. Such treatments typically introduced single nucleotide changes or small deletions in the genome, resulting in mutant collections to be screened for the phenotype of interest (Feldmann et al. 1994; Greene et al. 2003). Even today, there are still good reasons to screen for mutants in chemically mutagenised populations, including the possibility to have an allelic series, allowing for strong, intermediate and weak alleles of genes that otherwise would produce lethal phenotypes when inactivated. Such conventional forward approach is, however, inefficient when it comes to performing high-throughput functional genomics analyses. Indeed, despite the existence of genetic and physical maps, the map-based cloning still remains a laborious approach (Peters et al. 2003). Screening of populations generated by random insertion mutagenesis, in contrast, has the advantage to allow the isolation of the disrupted gene, causing the observed phenotype, in a much more direct way. Indeed, the precise chromosomal location of individual inserts can be easily obtained by PCR-based methodologies, such as thermal asymmetric interlaced PCR (TAIL; Liu and Whittier 1995), and amplification of insertion mutagenised sites (AIMS; Frey et al. 1998).

Whereas forward genetics starts with the mutant and then leads to the gene, reverse genetics starts with the gene of interest and ends with the corresponding mutant. As mentioned above, databases containing sequence information of Arabidopsis DNA stretches flanking the insertion sites are available and they can be easily searched for the knockout of interest. This approach is particularly useful to verify the role of genes whose functions are already known in other species (Bellafiore et al. 2005). However, loss-of-function lines not always exhibit an evident phenotype. Meinke et al. (2003) estimate that of all the predicted Arabidopsis genes only about 10% of them are expected to result in a detectable loss-of-function phenotype. Indeed, about 65% of Arabidopsis genes appear to be members of families with two or more members, and functional redundancy among closely related genes often obscures their phenotypes. To circumvent this problem, reverse genetics allows mutations in all members of a gene family to be identified. Indeed, simple crosses can be performed to combine mutations in closely related genes, resulting in phenotypes that would otherwise remain hidden. In addition, some phenotypic characteristics may be hard to detect unless the mutated gene is studied in a certain mutant background that more clearly reveals its loss-of-function phenotype (Roeder et al. 2003). Assessing a possible phenotype may also depend on the assay conditions. For instance, mutations in genes that are involved in stress responses may only display a detectable phenotype when subjected to certain environmental challenges. Reverse genetics largely facilitates these kinds of studies by allowing scientists to make qualified guesses on which combination of mutations would give rise to phenotypic changes.

5. Transcriptomics - depicting the expression level of genes

Obtaining mutations in genes of interest is an important and necessary step, however is just one of many powerful tolls that are needed to understand how the function of a gene is carried out. Genome-wide analysis of gene expression can be used as a powerful methodology to provide further information on gene activity. In this sense the development of full genome oligonucleotide-based microarrays, such as the ATH1 by Affymetrix, which represents approximately 23,750 Arabidopsis genes, allow to monitor at the same time the abundance of thousands of mRNA molecules (Redman et al. 2004). The classical experiment involves the simultaneous measurement of the relative concentration of a given transcript in two different samples by competitive two-colour hybridization. Most experiments use reverse transcription for labeling the cDNAs of both samples. cDNA molecules are labeled with different fluorophores (e.g. fluorescein, Cy5, Cy3). After hybridization to the arrayed target molecules the relative measure of gene expression for each gene analyzed is determined with the help of a high-resolution laser scanning device. Scanning is conducted using two different wavelengths, giving a quantitative fluorescence image for the two different probe populations (Lockhart and Winzeler 2000). Expression profiling experiments employing microarrays can be used to address diverse biological problems. For instance, by monitoring different tissues and developmental stages, an atlas can be created that describes the expression pattern of every gene in the genome (Borevitz and Ecker 2004). Knowledge of the timing and expression pattern of genes allows potential network to be created. The biological function of unknown genes in such a network can be inferred under the assumption that genes expressed similarly will be involved in a similar process (Brown and Botstein 1999;

Oliver 2000). Additionally, by looking at the changing expression patterns in response to abiotic or biotic stresses, one can identify the complete set of genes that is involved in a certain biological process. Defining such sets of genes, which are co-regulated under defined conditions, also allows for the identification of "marker genes" that are diagnostic for certain developmental or environmental processes (Holtorf et al. 2002). Alternatively, comparison of the expression profiles of wild-type plants to mutant lacking the activity of a stress- or development-induced transcription factor gene allows the identification of genes modulated by that transcription factor. Since the microarray-based expression profile technology has been established, thousands of arrays have been processed, of which a significant number are publicly available through services and repositories (Table 1.1). The Nottingham Arabidopsis Stock Centre Transcriptomics Service, NASCArrays, is one of them. Currently the database contains 40 experiments made up of about 400 GeneChips from Affymetrix system, but the number is increasing rapidly (Craigon et al. 2004). Other repositories include the ArrayExpress at the European Bioinformatics Institute (Brazma et al. 2003), and the Gene Expression Omnibus (GEO; Edgar et al. 2002) at the National Center for Biotechnology Information. However, the combination of multiple datasets still raises a number of questions concerning their compatibility, in particular when comparing data from different platforms. To overcome this problem, GENEVESTIGATOR (Table 1.1), a database and Web-browser data mining interface, has been created with the peculiarity to contain exclusively data from Affymetrix GeneChip (Zimmermann et al. 2004, 2005). Although data from different experiments may not be pooled for a rigorous expression profiling analysis, one can assume that the large scale combination and analysis of expression data from a single platform like the Affymetrix system allows the identification of biologically meaningful expression patterns of individual genes. Currently, the database covers more than 150 experiments, 28 plant organs and ten growth stages, in addition to responses to 68 environmental factors and to more than 100 genetic modifications. The dataset can be presented in the context of plant development, plant organ, environmental conditions, and mutated genetic backgrounds, both for individual genes or for families of genes, thereby answering questions such as "which other genes are coexpressed with the one of interest?", "in which organ or tissue is expressed the gene of interest?", "in which section of the life cycle the gene of interest is expressed?", or "which environmental stimulus induces the expression of the gene of interest?" The resulting answers can be used to confirm previous hypothesis or generate new hypotheses about gene expression network structures and genetic regulatory networks, resulting in the design of more precise and targeted experiments aimed to gene function discovery.

6. Proteomics – protein compendium and interacting partners

It is mostly proteins that carry out cellular functions, therefore, for a comprehensive understanding of biological functions, the proteome of organelles, cells, and tissues has to be systematically characterised. The promise of proteomics is the precise definition of the function of every protein, and how that function changes in different environmental or developmental conditions, with different modification states of the protein, and with different interacting partners (Roberts 2002). The sequencing of the Arabidopsis and rice genomes has greatly aided the scope for the discovery and exploitation of the plant proteome (Swiss-prot; Table 1.1). Large-scale transcript analysis now allows high-fidelity assessments of the tissue and developmental profiles of probable Arabidopsis proteomes, albeit with the caveat that differences in transcript abundance underlie differences in protein abundance. However, even these experimental approaches largely neglect the cellular compartmentalization of plant cells. The products of thousands of genes in plants are efficiently targeted to particular parts of the cell by elaborate targeting machinery that uses targeting information within the amino acid sequence of proteins (Heazlewood et al. 2005). Identifying protein locations within the plant cell is thus an important step toward a broader understanding of cellular function as a whole, and provides vital assistance in identifying the role of the many proteins currently ascribed to unknown function in plant genome databases. Several routes can be taken to place this cellular compartmentalization perspective on plant genomic data. The use of bioinformatics targeting algorithms to predict where gene products will be located is a simple, low-cost, and rapid way to tackle this issue. An array of such programs, able to predict protein localization into the nucleus, mitochondrion, plastid, peroxisome, and endoplastic reticulum, exists. However, a significant limitation of this approach is the lack of prediction capabilities for compartments, such as the Golgi, vacuole, and plasma membrane. A first prediction aimed to identify nuclear-encoded proteins targeted to chloroplasts was performed by Abdallah et al. (2000). These authors analysed the, at that time, partially sequenced genome of *A. thaliana*, employing the neural network-based program ChloroP (Emanuelsson et al. 1999), and extrapolated a total number of around 2,200 proteins having a chloroplast transit peptide (cTP). Based on the ChloroP algorithm, the TargetP program was, subsequently, developed by Emanuelsson et al. (2000), and it was estimated that more than 3,000 genes of the nuclear genome of *A. thaliana* encode for proteins featuring a cTP. More recently, the accuracy of the four most-widely used cTP predictors, iPSORT (Bannai et al. 2002), TargetP (Emanuelsson et al. 2000), PCLR (Schein et al. 2001) and Predotar (http://urgi.infobiogen.fr/predotar/predotar.html), was re-evaluated on a test set of 2,450 proteins with known subcellular location, and was found to be substantially lower than previously reported (Richly and Leister 2004). A combination of cTP predictors resulted to be superior to any one of the predictors alone and was employed to estimate that around 2,000 different cTP-proteins should exist in *A. thaliana*. Clearly, the large discrepancies among different type of predictors highlights the need of direct experimental approaches to better identify organelle and compartment proteomes. The strategies most commonly used involve cellular fractionation, centrifugation-based purification of an organelle, or cellular compartment and mass spectrometry (MS) to identify peptides (Millar 2004). A series of reports has provided in-depth analyses of chloroplast proteome. Norbert Rolland, Jacques Joyard and colleagues have analyzed a mix of inner and outer envelope membrane proteins of chloroplast from spinach and *A. thaliana* (Ferro et al. 2002; Seigneurin-Berny et al. 1999). Several known, as well as novel, membrane proteins were identified. In their latest, more extensive study with mixed *A. thaliana* chloroplast envelope membranes, more than 100 proteins were identified (Ferro et al. 2003). Almost one third of the identified proteins was reported to have no known function, whereas more than 50% were very likely to be associated with the chloroplast envelope, based on their postulated function or because they were already known envelope proteins. Wolfgang Schröder, Thomas Kieselbach and colleagues also analyzed the lumenal proteome of *A. thaliana* and spinach (Kieselbach et al. 1998; Schubert et al. 2002), resulting in the identification of thirty-six proteins. Similarly, several groups have contributed to the investigation of the mitochondrial proteome (Eubel et al. 2003; Herald et al. 2003; Kruft et al. 2001; Millar et al. 2001; Werhahn and Braun 2002). Recently, a large analysis using non-gel proteomic approaches based on liquid chromatography and tandem MS (LC MS-MS) has provided a set of more than 400 non-redundant proteins from Arabidopsis mitochondria (Heazlewood et al. 2004). The proteome of nuclei, vacuoles, and peroxisomes has also received attention in recent reports (Carter et al. 2004; Fukao et al. 2002; Pendle et al. 2005). A series of studies has also identified proteins among the other intracellular membrane systems, including plasma membranes, Golgi, and endoplasmic reticulum (Alexandersson et al. 2004; Prime et al. 2000; Santoni et al. 1999).

A complementary approach to MS in identifying protein location is the expression and visualization of fluorescence proteins (FPs) attached to proteins of interest. A range of differently coloured fluorescent proteins have been used, including green fluorescent protein (GFP), red fluorescent protein (RFP), yellow fluorescent protein (YFP), and cyan fluorescent protein (CFP), with GFP being the dominant choice. Increasingly referred to as clone-based proteomics, single–protein studies, medium throughput approaches, and even high-throughput GFP screening of protein locations using this technique are currently under way in *Arabidopsis* (Cutler et al. 2000; Koroleva et al. 2005). Many hundreds of proteins have been visualised in this manner to date and form an important dataset for determining subcellular localization. Concerning *Arabidopsis*, it has been calculated that all the different approaches have resulted in the localization of 4,418 proteins, representing approximately 15% of the whole predicted proteome. All these information distributed in a large set of databases (Table 1.1) have been recently collected in the Subcellular location database for Arabidopsis proteins (SUBA; Heazlewood et al. 2005), which provides an integrated understanding of protein localization, encompassing the plastid, mitochondrion, peroxisome, nucleus, plasma membrane, endoplasmic reticulum, vacuole, Golgi, cytoskeleton structures, and cytosol. Of course, the subcellular localization data alone are not sufficient to determine the function of the protein of interest, however they represent an additive value towards the determination of gene function.

Another important aspect of proteomics concerns the characterization of covalent processing events, such as proteolytic cleavages and/or addition of modifying groups to one or more amino acids, responsible to change the properties of a protein (Huber and Hardin 2004; Mann and Jensen 2003). Far from being mere “decorations”, post-translation modifications of a protein can, indeed, determine its activity state, localization, turnover, and interactions with other proteins. Despite the great importance of these modifications for biological function, their study on a large scale has been hampered by a lack of suitable technologies. Indeed, proteomics has been very successful in identifying proteins in complexes and organelles since only few peptides are needed for protein identification, but they are not enough for complete primary-structure determination. A central consideration in the characterization of the modifications is the need for as large an amount of the protein as possible. Protein modifications are typically not homogenous, and a single gene can give rise to a bewildering number of gene products as a result of different modifications. Up to now most of the protein modifications have been analysed on a one-by-one basis and in many cases by using recombinant expressed proteins, but the real promise of proteomics is to asses systematically the modifications of large number of proteins (Jensen 2000). Recent technological developments has made it increasingly feasible to directly analyse very complex peptide mixtures by LC MS-MS, and a single chromatographic run can result in the identification of hundreds of peptides, increasing substantially the chance of finding modified peptides. Moreover, the complexity of peptide mixtures can be reduced by affinity chromatography. For example, phosphopetides can be captured selectively through their negatively charged phosphogroup on immobilized metal affinity columns (IMAC; Nuhse et al. 2003). Recently, this technique has been used for a comprehensive analysis of phosphorylated membrane proteins in *Arabidopsis*, resulting in the identification of more than 300 phosphorylation sites (Nuhse et al. 2004). This analysis has yielded general principles for predicting other phosphorylation sites in plants and provided indications of specificity determinants for responsible kinases. In addition, more than 50 sites were mapped on receptor-like kinases and revealed an unexpected complexity of regulation. All the data have been collected in a new searchable database for plant phosphorylation sites (PlantP; Table 1.1), resulting in the first database on protein post-translational modifications, similar to the genomics, transcriptomics, and proteomics databases in existence today. Once post-translational modification analyses could be routinely done at proteomics level, the identification of more and more modification sites will dramatically increase, resulting in the development and tuning of algorithms aimed to the prediction of modification sites and to the functional interpretation of post-translational modifications.

The “omics” technologies has made clear that a discrete biological function can only rarely be attributed to an individual molecule. Instead, most biological characteristics arise from complex interactions between the cell constituents, such as proteins, DNA, RNA and small molecules. Therefore a key challenge for biology in the post-genomics era is to understand the structure and the dynamics of the complex intercellular web of interactions that contribute to the structure and function of a living cell. A prolific genome-wide approach to study protein-protein interactions is the yeast two-hybrid assay (Fields and Song 1989). This method constructs an artificial transcription factor, fusing the DNA-binding domain with the first query protein and a transcriptional activation-domain to the second protein or to an expression library. When two proteins bind, a reporter gene is actively transcribed. Thousands of protein interactions have been discovered with this strategy in yeast, *Caenorhabditis elegans*, and *Drosophila melanogaster* and recently this approach has been adopted also for *Arabidopsis* (Hackbush et al. 2005). In particular, the Uhrig group has investigated the interaction network of 3-aa loop-extension (TALE) homeodomain proteins. A combination of cDNA-library screenings and an all-against-all pair wise interaction test revealed the formation of a complex array of homo- and hetero-dimers within the TALE family. In addition, a previously unrecognised plant-specific protein family denominated *Arabidopsis thaliana* ovate family proteins (AtOFPs) was involved in the highly interconnected TALE interaction network. An alternative technique under development for global discovery of protein-protein interactions is “fluorescence resonance energy transfer” (FRET) microscopy (Wouters et al. 2001). Fluorophores within about 60 Å of one another transfer energy, and tagging different fluorophores to a pair of proteins will generate observable resonant phenomena when the pair binds. The broad applicability of this method is not yet clear, but FRET has been successful in specific trials. High-throughput screens are currently in design. This system potentially has the great advantage of reporting both the localization and timing of protein interactions, *in vivo* and in response to experimental conditions or perturbations (Carter 2005). Array technology is another candidate for direct detection of protein-protein interactions. Taking DNA microarrays as a model, the aim is to construct a chip onto which an entire proteome is spotted (Schweitzer et al. 2003; Smith et al. 2005). This protein array would facilitate global screens not only for protein-protein interactions, but also for protein-DNA interactions. Indeed, the latter approach will be extremely useful to reveal the main actors, such as transcription factors and promoter *cis*-regulatory elements, responsible of the complex gene expression regulation mechanisms.

7. Metabolomics – comprehensive non-biased analysis of metabolites

Transcriptomics and proteomics certainly contribute to the description of phenotypes and discover of gene function, however, it is essential that phenotypic effects be described as explicitly as possible. To this aim, metabolites, regarded as the ultimate gene products, represent the direct link between genes and phenotypes. The plant kingdom is able to produce an astonishing wealth of metabolites, ranging from 90,000 to 200,000, both from primary and secondary metabolism (Fiehn 2002). Additionally, metabolites have a much greater variability in the order of atoms and subgroups with respect to the 4-letter code of genes and transcripts and the 20-letter code of proteins, making their characterization extremely demanding in terms of technologies. Accordingly, different analytical approaches have been designed in order to address specific questions. In particular, to directly study the primary effect of a genetic alteration, an analysis can be constraint to the specific substrate and/or the direct product of the corresponding encoded protein. This strategy is called “targeted analysis” and is mainly used for screening purposes. Alternatively, the effects of biotic or abiotic stresses can be monitored on a selected number of predefined metabolites or pathways, by using the “metabolite profiling” strategy. However, quite frequently, genetic or environmental repercussions are not limited to one biological pathway. Indeed, the metabolite levels of unrelated pathways may be altered due to pleiotropic effects. In order to understand these effects, a comprehensive analysis in which all metabolites of a biological system are identified and quantified is needed. Such an approach has been called, “metabolomics”. Metabolomics approaches must aim at avoiding exclusion of any metabolite by using well conceived sample preparation procedures and analytical techniques (Bhalla et al. 2005). To asses the enormous diversity of structurally complex chemical compounds, various approaches have been initiated in the last decade largely due to the tremendous advances in the instrumentation and data handling capabilities (Fukusaki and Kobayashi 2005). In particular, advances in metabolomics analysis owe primarily to improvements in the MS technology that has resulted in formats that are more user-friendly and amenable to biologists. Additionally, combination of mass spectrometry with in-line gas or liquid chromatography (GC-MS and HPLC-MS) has increased the efficiency of separation and identification of molecules. Nuclear Magnetic Resonance (NMR) is another potential very useful technique to be used in metabolomics, since in principle any chemical species that contains protons gives rise to signals. Indeed, NMR is often used for metabolite fingerprinting, where the aim is to look for compositional similarities and explore the overall natural variability (Bligny and Douce 2001; Raamsdonk et al. 2001). Recently, a new technique has been introduced, called Fourier Transform Ion Cyclotron Mass Spectrometry (FT-MS), that allows the study of phenotypic changes associated with metabolism. FT-MS is, indeed, suitable for rapid screening of similarities and dissimilarities in large collections of biological samples, such as plant mutant populations. Separation of the metabolites is achieved solely by ultra-high mass resolution. Identification of the putative metabolite or class of metabolites to which it belongs can then be obtained by determining the elemental composition of the metabolite based upon the accurate mass determination (Brown et al. 2005).

Similarly to genome, transcriptome and proteome fields, databases storing the flood of data arising from metabolomics analyses are becoming available. The Golm Metabolome Database (Kopka et al. 2005) is the first database that provides public access to custom mass spectral libraries, metabolite profiling experiments, as well as additional information and tools, e.g. with regard to methods, spectral information or compounds (Table 1.1).

Likewise transcriptomics, the primary objective of metabolomics analysis is to associate the relative changes in quantitative metabolite levels with functional assignments. To this aim, different pattern recognition methods such as hierarchical cluster analysis and principal component analysis can be applied to calculate an individual metabolite profile and compare it to other metabolite profiles. Profiles of samples which group into a defined cluster can then be used to define a metabolic phenotype. Once the existence of clusters within the samples is assured, classical statistics such as Student' s *t* test or multiple analysis of variance (MANOVA) can be applied in order to find statistically significant differences of metabolite levels between the clusters (Fiehn 2002). As with other functional genomics approaches, the interpretation of results may become problematic because of the sheer mass of data generated. The fact that biochemical pathways make up highly regulated networks adds to the complexity of the analysis. A way to interpret the data is to intercalate biochemical pathways, whether or not the alterations of metabolite levels or clustering results can be understood by known aspects of enzymatic regulation. With this respect, the Kyoto Encyclopedia of Genes and Genomes (KEGG; Ogata et al. 1999), represents a valuable tool (Table 1.1). In particular, KEGG is a knowledge base for systemic analysis of genes functions in terms of networks of genes and molecules. The major component of KEGG is the pathway database that consists of graphical diagrams of biochemical pathways and some of the known regulatory pathways. There are about 90 reference maps for the metabolic pathways that are manually drawn and continuously updated accordingly to biochemical evidences. In addition to the data collection efforts, KEGG provides various computational tools, such as for reconstructing biochemical pathways from the complete genome sequence and for predicting gene regulatory networks from the gene expression profiles. BRENDA is another database (BRaunschweig ENzyme Database; Schomburg et al. 2004) where a comprehensive collection of enzyme and metabolic information is collected. The database contains data from at least 83,000 different enzymes from 9800 different organisms, classified in approximately 4200 EC numbers. BRENDA includes biochemical and molecular information on classification and nomenclature, reaction and specificity, functional parameters, occurrence, enzyme structure, application, engineering, stability, disease, isolation and preparation, links and literature references (Table 1.1).

A number of different applications of metabolomics analyses can be imagined. For instance, metabolomics is being increasingly used for understanding the cellular phenotypes in response to various types of abiotic or biotic stresses. In one recent study of sulphur deficiency response, general metabolic readjustment was found (Nikiforova et al. 2005). Mutual influences were found between sulphur assimilation, nitrogen imbalance, lipid breakdown, purine metabolism, and enhanced photorespiration. A general reduction of metabolic activity was seen under conditions of depleted sulphur supply. Metabolomics has also been applied to the case of cold stress response (Cook et al. 2004). In particular, a total of 325 metabolites were up-regulated in cold-treated Arabidopsis plants. Also in this case an extensive reconfiguration of several metabolic pathways could be observed. Concerning biotic stresses, Kant et al. (2004) investigated defence responses in tomato plants after infection with spider mites. Although the spider mites had caused little visible damage to the leaves after the first day of infection, they had already induced direct defense responses. For example, proteinase inhibitor activity had doubled. Moreover, at the fourth day after infection, a significant increase in the emission of volatile terpenoids could be observed. Alternatively, metabolomics data can be used to reveal the phenotype of silent mutations. Indeed, the intercellular concentrations of metabolites can reveal the site of action in the metabolic network of a disrupted gene. Moreover, metabolomics might be the ideal tool to investigate the substrate specificity of the several enzyme isoforms and, certainly, has a deep impact in prediction of novel metabolic pathways and in the description of cellular networks *in vivo*.

8. Naturally occurring genetic variation

Genetic variation found in wild strains is probably the most important basic resource for plant biology. In *Arabidopsis* genetic variation has been identified for many traits mainly by direct analysis and comparison of accessions. This evaluation is facilitated by the large collection of more than 300 different accessions collected worldwide, which are publicly available in the stock centers. In particular, genetic variation has been found for resistances to biotic factors such as bacteria, fungi, viruses, insects, and mammals (Koornneef et al. 2004). Variation for disease resistance genes is large and involves pathogens (Holub 2001) and many variants of one of the approximately 200 types of plant disease resistance genes of the so-called NBS-LRR classes (Meyers et al. 2003). Large variation has been also reported for tolerance to abiotic stresses and for developmental, physiological and biochemical traits (Koornneef et al. 2004). The study of genetic variation, certainly, is extremely useful to the identification of gene function. Indeed, despite the fact that mutant approaches have been very powerful for functional analysis, often the definition of gene function is hampered by the genetic background of the analysed accession. In fact, the sort of mutant phenotypes that can be identified depends on the wild-type genotype. For instance, mutant phenotypes of genes for which the wild-type accession carries a natural null allele or a weak allele might not be detected. Examples of loss of function and probably null alleles present in accessions are quite common, as indicated by the 111 Columbia genes found to be partially or completely deleted in the *Ler* accession (Borevitz et al. 2003). Additionally, most variation among accessions is of a quantitative nature due to the effects of allelic variation at several loci, which combined with the environmental effect, determines a continuous phenotypic distribution of the trait in segregating populations (Quantitative Trait Locus, QTL). The analysis or mapping of QTLs, implies the identification of loci, the relative additive effects, the mode of action of each QTL (dominance effects) and, therefore the contribution of genetic interaction between loci, making it relevant to the comprehension of plant networks. To date, fine QTL mapping has been performed on a limited number of plant processes, including flowering time. Indeed, the timing of flowering transition is genetically differentiated among natural populations of *Arabidopsis*, as shown by the large genetic variation observed for this trait among Arabidopsis accessions. Currently, 14 different QTLs, accounting for flowering time differences among Arabidopsis accessions, have been identified. Among them, the strong effect loci *FRIGIDA* (*FRI*) and *Flowering Locus C* (*FLC*) (Johanson et al. 2000; Michaels and Amasino 1999). The isolation of these loci has led to the identification of two novel genes involved in the regulation of flowering, that could not be identified by mutant analyses because the common laboratory early flowering strains carry loss of function alleles at the corresponding loci. The *FRI* locus encodes a protein with no significant homology to any other protein previously identified, whereas *FLC* encodes a MADS-box transcription factor. FRI positively regulates the expression of *FLC*, whereas the FLC protein negatively regulates the expression of other transcription factors involved in the regulation of flowering such as *SOC1*. Further molecular characterization of *FLC* is showing that it is a central integrator of flowering signals from different pathways (Sheldon et al. 2000). Other loci encoding phytochromes and cryptochromes involved in the flowering response to photoperiod have been also identified (Aukerman et al. 1997; Maloof et al. 2001; Mouradov et al. 2002), leading to a progressive clarification of the processes involved in defining the flowering time and highlighting the importance of natural variation investigation in revealing intricate molecular networks.

9. Inferring biological networks

A key aim of postgenomic research is to systematically catalogue all molecules and their interactions within a living cell. There is a clear need to understand how these molecules and the interactions between them determine the function of the enormously complex cell machinery, both in isolation and when surrounded by other cells forming tissues, organs and whole organisms. Although the study of genetic variation, certainly, contributes to the identification of molecular networks, the large datasets generated by high-throughput technologies, the "omics" technologies, represent a great resource to network biology. Bioformatics tools have been developed to collect and organize the datasets provided by transcriptomics, proteomics and metabolomics analyses (see above), however it is necessary to combine them with a portfolio of interpretation tools able to integrate the multiparameter raw data and link them to biological contexts. Examples of such tools are GENMAPP (Dahlquist et al. 2002), Pathway studio (Nikitin et al. 2003), PATHWAY Processor (Grosu et al. 2002) and BIOMINER (Sirava et al. 2002). However their usefulness for plant datasets is restricted, since they were developed for microbial or animal systems. A first plant-specific application, aimed to integrate transcriptomics data to metabolism, is represented by AraCyc (Mueller et al. 2003). The database (Table 1.1) currently contains about 2000 gene annotations in 117 individual pathways. The pathways are summarised figuratively on an overview map, many are available as detailed diagrams, and a tool, the AraCyc Expression Viewer, allows the user to overlay mRNA expression data on the AraCyc pathway diagrams. One of the comprehensive open-source software packages that allows integration of Arabidopsis transcriptomics and metabolomics data is MAPMAN (Thimm et al. 2004). Within MAPMAN (Table 1.1), Arabidopsis genes are grouped in 200 hierarchical categories by a module called TRANSCRIPTSCAVENGER, and hundreds of metabolites are linked to pathways using the METABOLITESCAVENGER module. The IMAGEANNOTATOR module allows the uploading of experimental data, resulting in a quick overview of the pathways together with transcript and/or metabolite contents. An impressive tool in development is represented by MetNet (Wurtele et al. 2003). This bioinformatics package is able to model metabolic and regulatory networks. Currently, only a limited amount of data have been uploaded into MetNet, but the first full production version promises to be an invaluable tool for Arabidopsis researchers. BioPathAt (Lange and Ghassemian 2005) most probably represents the first bioinformatics tool able to integrate transcriptomics, proteomics, and metabolomics data in the context of well-annotated biochemical pathways. In particular, Arabidopsis metabolic pathways have been generated based on the apparent coding capacity of the entirely sequenced Arabidopsis genome and assembled in the BioPathAtMAPS module. A gene list for the enzymes involved in the different pathways has been compiled using literature keyword and sequence-based searches in the TAIR database resulting in the BioPathAtDB module. Roughly 1500 genes/enzymes are present in this module and for all of them the subcellular localization was predicted using the PSORT and TargetP programs. Additionally, information on genes/enzymes expression patterns (organ- and tissue-specificity of transcript abundance), enzyme presence (based on proteomics data), enzyme activity (based on biochemical assays with purified, native proteins or crude protein extracts), and biochemical characteristics of recombinantly expressed isozymes (substrate specificity) are added. A complementary database containing information regarding the organ- and tissue-specific pool sizes of metabolites involved in Arabidopsis biochemical pathways (BioPathAtMETDB) is also available, together with gene-protein, and protein-protein interaction data. Dynamic boxes of different colours, placed on the biochemical pathway maps, are used to visualise patterns of RNA abundances, protein expression and metabolite pools. Additionally, enzyme activators and repressors are connected to biochemical pathways by coloured lines, providing insights into the networks that regulate metabolic pathways. Taken together, the computational tools support the researchers with a multidimensional representation of metabolic networks that certainly help to manage the data complexity and to enhance the knowledge on plant biology.

10. Conclusions

Plant genomics research has entered the phase of high-throughput gene function characterization due to the development of essential genetic tools, including comprehensive sets of sequence-indexed mutant collections, and the employment of the “omics” technologies. Global genome data, such as transcriptome atlases, are providing a holistic picture of gene expression regulation. Proteomics analyses are defining spatial and temporal localization of proteins as well as their specific dependence upon environmental conditions. In addition, the progressive development of methods to probe protein-protein and DNA-protein interactions, as well as posttranslational modifications, with high coverage and reliability will contribute to the modeling of cellular dynamics. Certainly, metabolomics investigations are already bridging the gap between gene products and experimental phenotypes. Throughout recent years it has become increasingly clear, however, that each method has its inherent limitations and none of them alone suffices to unequivocally assign functions to genes. In order to take full benefit from functional genomics, the vast and increasing amount of disparate data types needs to be interconnected and stored in central databases, where information concerning gene sequence, gene expression, protein function, protein interaction, protein localization, phenotype of loss of function line, and metabolic perturbation are linked together. Progress in computational studies of existing data are, therefore, of great importance for data integration as well as for the identification of fundamental properties upon which biological model can be built. The long term goal, as suggested in the ambitious "2010 project" (Chory et al. 2000) is "to understand every molecular interaction in every cell throughout a plant lifecycle. […]. The ultimate expression of our goal is nothing short of a virtual plant which one could observe growing on a computer screen, stopping this process at any point in that development, and with the click of a computer mouse, accessing all the genetic information expressed in any organ or cell under a variety of environmental conditions". Work on other plant species will certainly benefit from Arabidopsis research, both in the use of functional data and in research methodology. Rice, for instance, represents a solid platform for transferring Arabidopsis knowledge and enhancing our understanding of crop species.

References

Abdallah F, Salamini F, Leister D (2000) A prediction of the size and evolutionary origin of the proteome of chloroplasts of *Arabidopsis*. Trends Plant Sci 5: 141-142

Alexandersson E, Saalbach G, Larsson C, kjellbom P (2004) Arabidopsis plasma membrane proteomics identifies components of transport, signal transduction and membrane trafficking. Plant Cell Physiol 45: 1543-1556

Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. Science 301: 653-657

Aukerman MJ, Hirschfeld M, Wester L, Weaver M, Clack T, Amasino RM, Sharrock RA (1997) A deletion in the PHYD gene of the Arabidopsis Wassilewskija ecotype defines a role for phytochrome D in red/far-red light sensing. Plant Cell 9: 1317-1326

Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S (2002) Extensive feature detection of N-terminal protein sorting signals. Bioinformatics 18: 298-305

Bellafiore S, Barneche F, Peltier G, Rochaix JD (2005) State transitions and light adaptation require chloroplast thylakoid protein kinase STN7. Nature 433: 892-895

Bhalla R, Narasimhan K, Swarup S (2005) Metabolomics and its role in understanding cellular responses in plants. Plant Cell Rep 24: 562-571

Bligny R, Douce R (2001) NMR and plant metabolism. Curr Opin Plant Biol 4: 191-196

Borevitz JO, Ecker JR (2004) Plant genomics: The third wave. Annu Rev Genomics Hum Genet 5: 443-477

Borevitz JO, Liang D, Plouffe D, Chang HS, Zhu T et al. (2003) Large-scale identification of single-feature polymorphisms in complex genomes. Genome Res 13: 513-523

Brazma A, Parkinson H, Sarkans U, Shojatalab M, Milo J et al. (2003) ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 31: 68-71

Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. Nat Genet 21: 33-37

Brown SC, Kruppa G, Dasseux JL (2005) Metabolomics applications of FT-ICR mass spectrometry. Mass Spectrom Rev 24: 223-231

Carter GW (2005) Inferring network interactions within a cell. Brief Bioinform 6: 380-389

Carter C, Pan S, Zouhar J, Avila EL, Girke T, Raikhel NV (2004) The vegetative vacuole proteome of *Arabidopsis thaliana* reveals predicted and unexpected proteins. Plant Cell 16: 3285-303

Chory J, Ecker JR, Briggs S, Caboche M, Coruzzi GM et al. (2000) National Science Foundation-Sponsored Workshop Report: "The 2010 Project" functional genomics and the virtual plant. A blueprint for understanding how plants are built and how to improve them. Plant Physiol 123: 423-426

Cook D, Fowler S, Fiehn O, Thomashow MF (2004) A prominent role for the CBF cold response pathway in configuring the low-temperature metabolome of *Arabidopsis*. Proc Natl Acad Sci U S A 101: 15243-15248

Craigon DJ, James N, Okyere J, Higgins J, Jotham J, May S (2004) NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. Nucleic Acids Res 32: 575-577

Cutler SR, Ehrhardt DW, Griffitts JS, Somerville CR (2000) Random GFP::cDNA fusions enable visualization of subcellular structures in cells of *Arabidopsis* at a high frequency. Proc Natl Acad Sci U S A 97: 3718-3723

Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat Genet 31: 19-20

Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30: 207-210

Emanuelsson O, Nielsen H, von Heijne G (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci 8: 978-984

Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300: 1005-1016

Eubel H, Jansch L, Braun HP (2003) New insights into the respiratory chain of plant mitochondria. Supercomplexes and a unique composition of complex II. Plant Physiol 133: 274-286

Feldmann KA, Malmberg RL, Dean C (1994) Mutagenesis in *Arabidopsis*. In *Arabidopsis* (Meyerowitz E and Somerville CR eds) Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, pp 137-172

Ferro M, Salvi D, Riviere-Rolland H, Vermat T, Seigneurin-Berny D, Grunwald D, Garin J, Joyard J, Rolland N (2002) Integral membrane proteins of the chloroplast envelope: identification and subcellular localization of new transporters. Proc Natl Acad Sci U S A 99: 11487-11492

Ferro M, Salvi D, Brugiere S, Miras S, Kowalski S, Louwagie M, Garin J, Joyard J, Rolland N (2003) Proteomics of the chloroplast envelope membranes from *Arabidopsis thaliana*. Mol Cell Proteomics 2: 325-345

Fields S, Song O (1989) A novel genetic system to detect protein-protein interactions. Nature 340: 245-246

Fiehn O (2002) Metabolomics – the link between genotypes and phenotypes. Plant Mol Biol 89: 235-249

Frey M, Stettner C, Gierl A (1998) A general method for gene isolation in tagging approaches: amplification of insertion mutagenised sites (AIMS). Plant J 13: 717-721

Fukao Y, Hayashi M, Nishimura M (2002) Proteomic analysis of leaf peroxisomal proteins in greening cotyledons of *Arabidopsis thaliana*. Plant Cell Physiol 43: 689-696

Fukusaki E, Kobayashi A (2005) Plant metabolomics: potential for practical operation. J Biosci Bioeng 100: 347-354

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

Gong M, Rong YS (2003) Targeting multi-cellular organisms. Curr Opin Genet Dev 13:215-220

Greene EA, Codomo CA, Taylor NE, Henikoff JG, Till BJ et al. (2003) Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in *Arabidopsis*. Genetics 164: 731-40

Grosu P, Townsend JP, Hartl DL, Cavalieri D (2002) Pathway Processor: a tool for integrating whole-genome expression results into metabolic networks. Genome Res 12: 1121-1126

Hackbusch J, Richter K, Muller J, Salamini F, Uhrig J (2005) A central role of *Arabidopsis thaliana* ovate family proteins in networking and subcellular localization of 3-aa loop extension homeodomain proteins. Proc Natl Acad Sci U S A 102: 4908-4912

Hannon GJ (2002) RNA interference. Nature 418: 244-251

Heazlewood JL, Tonti-Filippini JS, Gout AM, Day DA, Whelan J, Millar AH (2004) Experimental analysis of the Arabidopsis mitochondrial proteome highlights signaling and regulatory components, provides assessment of targeting prediction programs, and indicates plant-specific mitochondrial proteins. Plant Cell 16: 241-256

Heazlewood JL, Tonti-Filippini J, Verboom RE, Millar AH (2005) Combining experimental and predicted datasets for determination of the subcellular location of proteins in *Arabidopsis*. Plant Physiol 139: 598-609

Herald VL, Heazlewood JL, Day DA, Millar AH (2003) Proteomic identification of divalent metal cation binding proteins in plant mitochondria. FEBS Lett 537: 96-100

Holtfort H, Guitton MC, Reski R (2002) Plant functional genomics. Naturwissenschaften 89: 235-249

Holub EB (2001) The arms race is ancient history in *Arabidopsis*, the wildflower. Nat Rev Genet 2: 516-527

Huber SC, Hardin SC (2004) Numerous posttranslational modifications provide opportunities for the intricate regulation of metabolic enzymes at multiple levels. Curr Opin Plant Biol 7: 318-322

Jensen ON (2000) Modification-specific proteomics: strategies for systematic studies of post-translationally modified proteins. Proteomics: A trend guide. Blackstock W, Mann M (eds) Elsevier Science London

Johanson U, West J, Lister C, Michaels S, Amasino RM, Dean C (2000) Molecular analysis of FRIGIDA, a major determinant of natural variation in Arabidopsis flowering time. Science 290: 344-347

Kant MR, Ament K, Sabelis MW, Haring MA (2004) Differential timing of spider mite-induced direct and indirect defenses in tomato plants. Plant Physiol. 135: 483-495

Kieselbach T, Hagman-Andersson B, Schröder WP (1998) The thylakoid lumen of chloroplasts. Isolation and characterization. J Biol Chem 273: 6710-6716

Koornneef M, Alonso-Blanco C, Vreugdenhil D (2004) Naturally occurring genetic variation in *Arabidopsis thaliana*. Annu Rev Plant Biol 55: 141-172

Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B et al. (2005) GMD@CSB.DB: the Golm Metabolome Database. Bioinformatics 21: 1635-1638

Koroleva OA, Tomlinson ML, Leader D, Shaw P, Doonan JH (2005) High-throughput protein localization in *Arabidopsis* using Agrobacterium-mediated transient expression of GFP-ORF fusions. Plant J 41: 162-174

Kruft V, Eubel H, Jansch L, Werhahn W, Braun HP (2001) Proteomic approach to identify novel mitochondrial proteins in *Arabidopsis*. Plant Physiol 127: 1694-1710

Kuromori T, Hirayama T, Kiyosue Y, Takabe H, Mizukado S et al. (2004) A collection of 11800 single-copy Ds transposon insertion lines in *Arabidopsis*. Plant J 37: 897-905

Lange BM, Ghassemian M (2005) Comprehensive post-genomic data analysis approaches integrating biochemical pathway maps. Phytochemistry 66: 413-451

Li Y, Rosso MG, Strizhov N, Viehoever P, Weisshaar B (2003) GABI-Kat SimpleSearch: a flanking sequence tag (FST) database for the identification of T-DNA insertion mutants in *Arabidopsis thaliana*. Bioinformatics 19: 1441-1442

Liu YG, Whittier RF (1995) Thermal asymmetric interlaced PCR: automatable amplification and sequencing of insert end fragments from P1 and YAC clones for chromosome walking. Genomics 25: 674-681

Lockhart DJ, Winzeler EA (2000) Genomics, gene expression and DNA arrays. Nature 405: 827-836

Maloof JN, Borevitz JO, Dabi T, Lutes J, Nehring RB et al. (2001) Natural variation in light sensitivity of *Arabidopsis*. Nat Genet 29: 441-446

Mann M, Jensen ON (2003) Proteomic analysis of post-translational modifications. Nat Biotechnol 21: 255-261

Meinke DW, Meinke LK, Showalter TC, Schissel AM, Mueller LA, Tzafrir I (2003) A sequence-based map of Arabidopsis genes with mutant phenotypes. Plant Physiol 131: 409-418

Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW (2003) Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. Plant Cell 15: 809-834

Michaels S, Amasino RM (1999) FLOWERING LOCUS C encodes a novel MADS domain protein that acts as a repressor of flowering. Plant Cell 11: 949-956

Millar AH (2004) Location, location, location: surveying the intracellular real estate with proteomics in plants. Funct Plant Biol 31: 563-571

Millar AH, Sweetlove LJ, Giege P, Leaver CJ (2001) Analysis of the Arabidopsis mitochondrial proteome. Plant Physiol 127: 1711-1727

Mouradov A, Cremer F, Coupland G (2002) Control of flowering time: interacting pathways as a basis for diversity. Plant Cell 14 :111-130

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

Nikiforova VJ, Kopka J, Tolstikov V, Fiehn O, Hopkins L, Hawkesford MJ, Hesse H, Hoefgen R (2005) Systems rebalancing of metabolism in response to sulfur deprivation, as revealed by metabolome analysis of Arabidopsis plants. Plant Physiol 138: 304-318

Nikitin A, Egorov S, Daraselia N, Mazo I (2003) Pathway studio - the analysis and navigation of molecular networks Bioinformatics Applications Note 19: 1-3

Nuhse TS, Stensballe A, Jensen ON, Peck SC (2003) Large-scale analysis of in vivo phosphorylated membrane proteins by immobilized metal ion affinity chromatography and mass spectrometry. Mol Cell Proteomics 2: 1234-1243

Nuhse TS, Stensballe A, Jensen ON, Peck SC (2004) Phosphoproteomics of the Arabidopsis plasma membrane and a new phosphorylation site database. Plant Cell 16: 2394-2405

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27: 29-34

Oliver S (2000) Guilt-by-association goes global. Nature 403: 601-603

Ostergaard L, Yanofsky MF (2004) Establishing gene function by mutagenesis in *Arabidopsis thaliana*. Plant J 39: 682-696

Parinov S, Sundaresan V (2000) Functional genomics in Arabidopsis: large-scale insertional mutagenesis complements the genome sequencing project. Curr Opin Biotechnol 11: 157-161

Pendle AF, Clark GP, Boon R, Lewandowska D, Lam YW et al. (2005) Proteomic analysis of the Arabidopsis nucleolus suggests novel nucleolar functions. Mol Biol Cell 16: 260-269

Peters JL, Cnudde F, Gerats T (2003) Forward genetics and map-based cloning approaches. Trends Plant Sci 8: 484-491

Prime TA, Sherrier DJ, Mahon P, Packman LC, Dupree P (2000) A proteomic analysis of organelles from *Arabidopsis thaliana*. Electrophoresis 21: 3488-3499

Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A et al. (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. Nat Biotechnol 19: 45-50

Redman JC, Haas BJ, Tanimoto G, Town CD (2004) Development and evaluation of an Arabidopsis whole genome Affymetrix probe array. Plant J 38: 545-561

Richly E, Leister D (2004) An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of *Arabidopsis* and rice. Gene 329: 11-16

Roberts, JK (2002) Proteomics and a future generation of plant molecular biologists. Plant Mol Biol 48: 143-154

Roeder AH, Ferrandiz C, Yanofsky MF (2003) The role of the REPLUMLESS homeodomain protein in patterning the *Arabidopsis* fruit. Curr Biol 13: 1630-1635

Samson F, Brunaud V, Balzergue S, Dubreucq B, Lepiniec L et al. (2002) FLAGdb/FST: a database of mapped flanking insertion sites (FSTs) of *Arabidopsis thaliana* T-DNA transformants. Nucleic Acids Res 30: 94-97

Santoni V, Doumas P, Rouquie D, Mansion M, Rabilloud T, Rossignol M (1999) Large scale characterization of plant plasma membrane proteins. Biochimie 81: 655-661

Schein AI, Kissinger JC, Ungar LH (2001) Chloroplast transit peptide prediction: a peek inside the black box. Nucleic Acids Res 29: E82

Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D (2004) BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res 32: 431-433

Schubert M, Petersson UA, Haas BJ, Funk C, Schroder WP, Kieselbach T (2002) Proteome map of the chloroplast lumen of *Arabidopsis thaliana*. J Biol Chem 277: 8354-8365

Schweitzer B, Predki P, Snyder M (2003) Microarrays to characterize protein interactions on a whole-proteome scale. Proteomics 3: 2190-2199

Seigneurin-Berny, D., Rolland, N., Garin, J. and Joyard, J (1999) Technical Advance: Differential extraction of hydrophobic proteins from chloroplast envelope membranes: a subcellular-specific proteomic approach to identify rare intrinsic membrane proteins. Plant J 19: 217-228

Sheldon CC, Rouse DT, Finnegan EJ, Peacock WJ, Dennis ES (2000) The molecular basis of vernalization: the central role of FLOWERING LOCUS C (FLC). Proc Natl Acad Sci U S A 97: 3753-3758

Sirava M, Schafer T, Eiglsperger M, Kaufmann M, Kohlbacher O, Bornberg-Bauer E, Lenhof HP (2002) BioMiner--modeling, analyzing, and visualizing biochemical pathways and networks. Bioinformatics 18: 219-230

Smith MG, Jona G, Ptacek J, Devgan G, Zhu H, Zhu X, Snyder M (2005) Global analysis of protein function using protein microarrays. Mech Ageing Dev 126:171-175

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

Thimm O, Blasing O, Gibon Y, Nagel A, Meyer S, Kruger P, Selbig J, Muller LA, Rhee SY, Stitt M (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J 37: 914-939

Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W (2004) GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. Plant Physiol 136: 2621-2632

Zimmermann P, Hennig L, Gruissem W (2005) Gene-expression analysis and network discovery using Genevestigator. Trends Plant Sci 10: 407-409

Werhahn W, Braun HP (2002) Biochemical dissection of the mitochondrial proteome from *Arabidopsis thaliana* by three-dimensional gel electrophoresis. Electrophoresis 23: 640-646

Wouters FS, Verveer PJ, Bastiaens PIH (2001) Imaging biochemistry inside cells. Trends Cell Biol 11: 203-211

Wurtele ES, Li J, Diao L, Zhang H, Foster CM et al. (2003) MetNet: Software to build and modelt he biogenetic lattice of *Arabidopsis.* Comparative and Functional Genomics 4: 239-245

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

Yu J, Wang J, Lin W, Li S, Li H et al. (2005) The genome of *Oryza sativa*: a history of duplication. PLoS Biol 3: 266-281