Integrated studies on plant biology using multiparallel techniques

O liver Fiehn*, Sebastian Kloska‡ and Thomas Altmann‡

Plant biology, especially the fields of molecular genetics and molecular physiology, is currently undergoing a change in paradigm from 'vertical' analysis of the role(s) of one or a few genes to 'horizontal' holistic approaches, studying the function of many or even all of the genes of an organism simultaneously. This change is leading us beyond genomes to transcriptomes, proteomes and metabolomes, and to an understanding of life at an entirely new level. Profiling strategies are putting this change into effect through the generation of large amounts of data, requiring that current bioinformatic approaches adapt and grow in order to make the most of these data.

Addresses
Max-Planck-Institut für Molekulare Pflanzenphysiologie, Am Mühlenberg 1, 14476 Golm, Germany
* e-mail: fiehn@mpimp-golm.mpg.de
‡ e-mail: kloska@mpimp-golm.mpg.de
© e-mail: altmann@mpimp-golm.mpg.de

Current Opinion in Biotechnology 2001, 12:82–86

0958-1669/01/$ — see front matter © 2001 Elsevier Science Ltd. All rights reserved.

Abbreviations
2DE two-dimensional gel electrophoresis
FT-MS Fourier-transform mass spectrometry
LC liquid chromatography/chromatographic
ORF open reading frame

Introduction

Large investments in genome sequencing and gene analysis were originally made with the notion that, although expensive in the short term, they would be highly cost effective in the end. It is already the case that gene discovery using heterologous probes has been largely replaced because of our ability to use sequence information generated through large-scale sequencing programs. The enormous value of partial cDNA sequence (expressed sequence tag, EST) collections was first realised in the paradigm that the complete inventory of genes provides the means for a holistic analytical approach at the molecular level; for the first time an organism can be studied as a genetic system in its entirety. Several multiparallel analytical approaches have been developed that provide the means by which to study gene/gene expression at the transcript (mRNA), protein and metabolite (the substrates and the products of reactions catalysed and controlled by the gene products) levels. The idea is to monitor all (or as many as possible) components of the system at all of these levels. The enormous complexity of the data sets generated through these analyses (each data point may be composed of tens of thousands of values) will require an extension of our current bioinformatic abilities.

The intention of this review is to provide an overview (which, because of space constraints, must remain general) of the implementation and the application of these techniques in the pre-eminent plant model, Arabidopsis. Depending on the resources available, similar approaches are being applied in many other plants.

Transcript analysis

There are several prominent analytical approaches used for transcript profiling: microarray-based approaches (cDNAs or oligonucleotides) [3,4,5••,6••]; sequencing-based approaches, such as serial analysis of gene expression (SAGE) [7] or massively parallel signature sequencing (MPSS) [8••]; and differential-display-based approaches, such as arbitrarily primed (AP) PCR [9] and cDNA-amplified fragment length polymorphism (AFLP) [10]. Of all of these, microarray technology has received the most widespread interest. This technique has been used extensively in yeast and human work, and thousands of expression profiles have been examined (e.g. [11,12••,13,14]). Although the first demonstration of the viability of the microarray technique was achieved with a set of Arabidopsis cDNAs [15], only a limited number of microarray-based expression studies in plants have been published. As the complete genomic sequence of Arabidopsis [16] is now available, a boost in the application of this technology, similar to that previously seen for yeast, is expected to happen soon.

Initial experiments in plants were performed by Desprez et al. [17], who used a set of cDNA fragments representing 864 genes to study expression in light-grown and dark-grown seedlings, and by Ruan et al. [18], who generated expression data for 1443 genes in leaves, roots and floral tissues. Microarray-based differential analysis of gene expression represented by 1701 randomly picked strawberry fruit cDNAs resulted in the identification of sets of genes differentially expressed during fruit ripening [19]. In this way, a novel gene
(SAAT), that plays a crucial role in flavour biogenesis, was identified. Similarly, an array of 150 Arabidopsis cDNAs was used to study and compare wound-induced, dehydrogenase-induced and insect-feeding-induced gene expression in wildtype, ethylene-insensitive and coronatine/jasmonate-insensitive mutants [20•]. Further reports on the use of microarrays of 5524 genes/clones or 2375 genes, respectively, include the identification of nitrate-regulated genes [21] and the observation of coordinated changes in gene-expression patterns in response to pathogen infection or treatment with salicylic acid, methyl jasmonate or ethylene [22]. Recent meeting reports [23,24] and a large number of entries in the Stanford Microarray Database (SMD) [25], however, indicate that a wealth of expression profiling data sets obtained for microarrays of up to 11,000 cDNA fragments will become available soon for Arabidopsis. Furthermore, initial experiments have been done using Arabidopsis oligonucleotide arrays representing 8775 genes [26].

This progress has been rapid, but the use of a number of technical improvements — designed to increase result reproducibility (such as standardising spotting and labelling so as to reduce variability), to produce truly quantitative data, to increase sensitivity and to provide a means by which to compare data obtained from different sets of experiments — is required if this progress is to continue. A study by Lee et al. [27] highlighted the importance of experimental replication in achieving valid information and this, too, is obviously an important factor in experimental design. Finally, common standards and references must be set if the data accumulated by different groups and through the application of different techniques are to be integrated.

Protein analysis
Protein profiling is the next level of analysis after transcript profiling. In order to achieve efficient separations, two-dimensional gel electrophoresis (2DE) utilises orthogonal physical properties of proteins, such as isoelectric points and apparent molecular masses. 2DE is a long-standing biochemical technique that is now being exploited for plant proteomic research [28••,29•,30] and is especially applicable to the analysis of abundant changes in highly expressed cytosolic proteins [31]. However, 2DE approaches face inherent problems if proteome-wide identification or accurate quantification is wished for. It has been shown that classical 2DE approaches detect only highly expressed proteins even when highly sensitive staining techniques are employed [32••]. Even when spots can be picked up by sensitive stains, quantitative estimations are limited by low dynamic ranges.

Capillary isoelectric focusing is an alternative to 2DE that can be directly coupled to ion cyclotron resonance mass spectrometers (Fourier-transform mass spectrometry, FT-MS) for the analysis of both crude protein mixtures [33••] and complex peptide digests [34]. A newly developed interface [35] has the potential to analyse mixtures of non-digested proteins with less expensive time-of-flight mass spectrometers. To increase sample loading limits, isoelectric focusing in aqueous systems has been coupled to non-porous liquid chromatographic (LC) separations and mass spectrometric peptide identification [36•]. Using reversed phase LC separation and MS/MS/MS fragmentation by ion trap mass spectrometers [37•], peptide sequences can be more efficiently deciphered than in classical tandem MS approaches. This approach might be useful when studying the proteomes of non-sequenced plants. LC/LC/MS techniques allow more sensitive and rapid separations of complex peptide mixtures when cation exchange is coupled to reversed phase LC columns. This strategy has been shown to be superior in speed, robustness and protein coverage, when compared with 2DE analysis [38••]. The major breakthroughs, however, have been made by approaches that attach chemical linkers to specific protein moieties before mass spectrometric determination. For example, an isotope distribution encoded tag has been used to detect and identify low copy number proteins [39•], and isotope coded tags have also been used to accurately quantify protein expression when comparing two different cell states [40••,41•]. By combining these and upcoming strategies, proteomic approaches will be feasible in more biological laboratories, providing more precision, repeatability, higher throughput and greater cost effectiveness than do 2DE approaches.

Metabolite analysis
Unlike transcripts and proteins, metabolites (which can be regarded as the ultimate gene products) have been largely ignored to date. However, the profiling of up to 68 primary metabolites has been successfully demonstrated to be useful for clinical research by differentially comparing healthy human tissues with diseased ones [42••]. In plant research, mass spectrometry has been used for decades to determine metabolic target compounds, but only recently has the idea been pursued of expanding the list of targets in order to profile a limited number of primary metabolites (apricot fruits [43], plant leaves [44] and potato tubers [45•] have now been subjected to metabolic profiling). Moreover, estimates of the total number of metabolites different members of the plant kingdom synthesise, that is estimates of plant metabolome sizes, range from 90,000–200,000. To determine plant metabolomes, therefore, we must cope with the sheer complexity of mass spectra found in plant chromatograms and apply new technologies for efficient de novo compound identification. The development and release of automated mass spectral deconvolution and identification software (AMDIS) tackled the former issue [46]. For the latter, successful de novo identifications of small polar metabolites have already been demonstrated for Arabidopsis leaf extracts, using new mass spectrometric techniques coupled to database queries [47•].

Most recently, high-throughput profiling of metabolic snapshots has been demonstrated for the first time in the context of plant functional genomics [48••]. Profiles of 326 metabolites of two Arabidopsis ecotypes were compared with two single gene mutants. Cluster analysis revealed distinct ‘metabolic phenotypes’ for each of the four genotypes. Integrating the observed metabolic alterations into
hypotheses about changes in biochemical pathways and gene expression levels will be the next task.

**Bioinformatic data management**

A storm of metabolic, protein and expression profiling activity is now inundating biologists with data. As this activity grows in intensity we will be faced with a challenge of previously unknown dimensions. The effects are already visible in the field of expression profiling.

Early transcriptomics experiments used small sample numbers and model organisms with relatively small genomes [49]; future experiments, however, will deal with hundreds of samples and with organisms that have larger, more complex genomes. A preview of what is to come can be seen in the work reported by Hughes *et al.* [12••], in which 300 samples, half of which were done in duplicate, and 63 negative controls were used to characterise undefined ORFs and potential drug targets in yeast (about 3.6 million values). Although the current data output from metabolomic and proteomic work is comparatively small, it will increase in magnitude as our technical abilities advance.

Profiling is reshaping the way we look at publications. The printed page now often provides data overviews only, whereas experimental results are stored on the web [12••,25]. In a more limited way, the analogous case was true for sequence data: it quickly became uncommon to publish sequences in a printed form. Unlike profiling, however, a few central repositories for sequencing data (GenBank and SwissProt, for example) were established in the early days of sequencing, before the influx became torrential. These repositories grew and changed as data accumulated. As no standardised database model exists for profiling data, biologists currently publish results in whatever format is most convenient to them. There is a growing need for a central repository with standardised format requirements, especially as profiling data are so structurally complex. A few groups have proposed data models and an XML-based data exchange format [50] is currently under development, but this format is limited to the smallest common denominator for the platforms currently in use in the field [51].

**Data analysis**

The biggest challenges, though, lie not in data storage, but in data analysis. The goal is to create a holistic view of the interplay between biomolecules for the first time, to understand how proteins and metabolites interact. Bioinformatics is the key to gleaning substantive information from the deluge and making sense of the data. As profiling leads us from genome to transcriptome, proteome and metabolome, bioinformatics will provide us with the information upon which models for the inner workings of the cell can be developed.

From a computational point of view, profiling data can be seen as a set of result vectors of the function $y = f(x)$, where the vector, $y$, represents the data output (mRNA, protein or metabolite levels) and $x$ represents both environmental and internal conditions. Controlled changes in environmental conditions or in gene expression (antisense constructs or ectopic gene expression, for example) can serve as experimental variables. The ultimate goal is the reverse engineering of life, the ability to deduce what the changes in $y$ will be as a function of $x$. Even if we assume that this function follows a linear model (which it probably does not) the data we have represent a heavily underdetermined equation system as we have more unknown factors than result vectors. On the other hand, it is reasonable to assume that large sets of genes and proteins follow synchronised patterns and that the living cell, therefore, contains many redundancies, which might ease the effort.

On the basis of this assumption, attempts at taming profiling data have been made using clustering algorithms that group raw data in an unbiased way [52,53]. Clustering allows genes, proteins or metabolites to be grouped according to their profiles. The basic idea is the elucidation of function by ‘guilt of association’. Clustering can work well when there is already a wealth of knowledge about the pathway in question, but it works less well when this knowledge is sparse.

All clustering methods used so far, however, have significant drawbacks [54••] that make them unsuitable for detecting complex relationships in data networks. The inability to detect non-linear correlations is one such limitation. A new way to cover similarities and correlation of expression profiles based on mutual informational entropy has been proposed by Butte *et al.* [55]. This is a statistical measure for the independency of two signals (e.g. expression level), which makes no assumption about the underlying mathematical function. This approach, however, artificially groups the stepless expression levels into distinct classes.

Until recently, all clustering methods also lacked a means by which to integrate statistics. Each gene or experiment was mapped into n-dimensional expression space without covering the fact that the sharp point within this space is merely a probability cloud. Hughes *et al.* [12••] made an attempt to include statistics in their clustering methods so as to see which clusters were trustworthy.

Profiling bioinformatics is moving towards methods that try to incorporate as much available knowledge as possible. Recently, two groups used statistical methods to correlate expression profiles with potential promoter sequences [56,57]. Marcotte *et al.* [58] tried to elucidate the functions of unknown ORFs on the basis of five different data sources, including expression profiling data. Brown *et al.* [59] used a special neuronal network to identify potential gene functions on the basis of the expression profile and the transcriptional patterns of well-annotated genes.

Protomics and metabolomics are still not producing as much data as transcriptomics, but they will obviously play key roles in understanding phenotypes in the future.
major focus of bioinformatics in the field of metabolomics is currently metabolic pathway simulation and model construction. Schuster et al. [60] are trying to identify every possible metabolic pathway leading to a given substance. The combination of simulated pathway information with pathway databases like KEGG [61] will serve as a model in attempts to combine data from other such domains.

Conclusions
It is clear that experimental work in plant biology will change dramatically in the near future. Technology will soon enable researchers to analyse processes in plants at all levels and to all depths. This will result in us gaining knowledge at a level hinted at before only in science fiction novels. The combination of holistic analytical techniques, bioinformatic tools and the power of genetics will provide the means by which to achieve a true and complete understanding of plant function. This will result in the ability to model plant life and, thus, to rationally design plants with desirable improved characteristics.

Acknowledgement
We thank Megan McKenzie for critically reading the manuscript.

References and recommended reading
Papers of particular interest, published within the annual period of review, have been highlighted as:
• of special interest
** of outstanding interest
Excellent example of high-throughput expression profiling analyse in yeast based on approximately 500 hybridisations. This provides important improvements in data analysis through a combination of cluster analysis with statistics.


34. Tavazoie S, Hughes J, Campbell M, Cho R, Church G: **Statistically significant differences, from 30% to 100-fold ranges, are found for 326 polar and lipophilic metabolites when comparing mutants to wildtype ecotype plants. Pleiotropic effects are found far away from the effect of the single-gene mutations and preliminary interpretation is given.** Proc Natl Acad Sci USA 1999, 96:2907-2912.


49. Tavazio S, Hughes J, Campbell M, Cho R, Church G: **Statistically significant differences, from 30% to 100-fold ranges, are found for 326 polar and lipophilic metabolites when comparing mutants to wildtype ecotype plants. Pleiotropic effects are found far away from the effect of the single-gene mutations and preliminary interpretation is given.** Proc Natl Acad Sci USA 1999, 96:2907-2912.


