



ELSEVIER

Systems approaches to understanding cell signaling and gene regulation

Commentary

Nicholas J Provart and Peter McCourt

The age of 'omics' is upon us, and scientific papers that reflect this are starting to appear at an ever-increasing rate. The amount of information generated in any 'omics' program is daunting and often overwhelms plant scientists whose main interests relate to cell or developmental biology. For this revolution in data generation to have any impact in plant signaling studies, we must have great confidence in both the quality of the data and our ability to represent it in ways that are meaningful to general plant biologists. Systems biology has begun to address these issues and to provide examples in which the analysis of large data sets has led to biological insights into cell signaling and gene regulation.

Addresses

Department of Botany, University of Toronto, 25 Willcocks Street,
Toronto, Ontario M5S 3B2, Canada
e-mail: mccourt@botany.utoronto.ca

Current Opinion in Plant Biology 2004, 7:605–609

Available online 23rd July 2004

1369-5266/\$ – see front matter

© 2004 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.pbi.2004.07.001

Introduction

The age of genomics, transcriptomics, and proteomics is upon us, and the number of studies that use some aspect of these technologies to study a biological problem is growing exponentially. For the information generated using these technologies to be useful, however, 'gold standard' quality data, good experimental design, multiple data inputs, and robust statistical analysis must be used. This review covers select examples in which these criteria have been applied to address questions relating to cell signaling and gene regulation in plants. For historical reasons, most high-throughput experiments in plants have involved transcript profiling. If analyzed correctly, the data from these experiments can be used to identify interactions between signaling pathways and to identify novel *cis*-acting elements in the promoters of co-regulated genes. Although less mature, proteomics technologies are also beginning to be used to further our understanding of plant signaling. The integration of the data generated by these and other high-throughput technologies with data available in public databases will

be necessary if this information is to be useful to the individual plant biologist.

'Omics' and systems biology: turning water into wine

The past decade has seen a conceptual shift in how we approach biological problems. The advent of high-throughput technologies has led to the rapid accumulation of biological data, ranging from complete genomic sequences and transcript profiles to 'all-by-all' protein–protein interaction maps. Often referred to as 'omics', these massively parallel approaches are usually classified by the biological constituent that they measure [1]. Transcriptomics, for example, attempts to measure the transcript levels of all genes from a given genome. Likewise, proteomics refers to the study of the complete protein complement of a cell, a tissue, an organ and so on. 'Omics' approaches are often applied to a whole organism that is exposed to a specific condition, or that contains a single gene mutation, to assess how a specific treatment or gene contributes to global regulation within that organism.

Although the scale of these experiments is impressive, many biologists get the feeling that this is the best of times and the worst of times. We now have access to quantities of data that would have taken years to obtain in the past, and the sheer amount of data available often overwhelms our ability to understand the biological consequences. A metaphor for these concerns is to think of a plant cell as a large metropolitan city. If we do massive cataloging of the people and their occupations, the buildings, the transit systems, and so on, does this really tell us why Tokyo is Tokyo, and more importantly, why Tokyo is not New York (Figure 1)? How can we use information obtained using 'omics' technologies to build cellular topology maps that are easy to interpret and allow us to navigate to the specific information that we need? The construction of such maps requires a systems approach that is built on concepts and techniques learned from other scientific disciplines such as physics, mathematics, numerical analysis, stochastic processes, and control theory. Systems biology allows the connecting of dots and the building of patterns from the information that is buried in genomes. Thus, although systems biology is dependent on 'omics' technologies for data generation, it really encompasses the design and use of analysis tools, and the development of new ways to represent data that are meaningful and allow insight.

Figure 1



A busy Tokyo intersection. The plant cell can be thought of as a metropolitan city. In daily life, each person (gene) contributes at some level to making the city (cell) thrive and function. High-throughput analyses first identify and names of all of people within the city (sequencing) and their occupations (protein function). We then find out when the people work (expression profiling), where they work (intracellular localization) and with whom they interact (protein interaction). However, this massive cataloging of information does not really tell us why Tokyo is Tokyo and, more importantly, why Tokyo is not New York. That is the challenge of systems biology – to sort through the information and connect the dots.

Systems biology data generation: boom or bust

For obvious reasons, plant systems biologists must have complete confidence in the high-throughput data generated by genomic projects. Although the *Arabidopsis* genome was published in 2000 and a draft sequence of two rice cultivars followed in 2002, determining the genetic structure of these genomes is still an ongoing process [2–4]. Gene annotation, for example, which involves determining if a genomic region contains a gene, cannot be predicted with complete confidence strictly from genomic sequence alone because of the complexity of eukaryotic gene structure. A recent approach for the validation of gene annotation used tiled *Arabidopsis* genomic sequences on an Affymetrix GeneChip to identify approximately 6000 new transcription units [5]. Thus, plant researchers must constantly update the gene annotation versions that they use, even for so-called fully sequenced genomes such as that of *Arabidopsis*. Never-

theless, high-throughput data from these and other microarray experiments are filling up databases, such as Gene Expression Omnibus (GEO), NASCArrays, and the Stanford Microarray Database (see Box 1). Depositing transcript profiles in centralized databases not only archives the data but also has the advantage of allowing researchers to compare expression profiles across a variety of experimental conditions, or to search for genes that respond similarly to their gene of interest (e.g. using Expression-Angler [see Box 1]). In addition, several statistical methods have been developed in the past couple of years to aid in the analysis of microarray data, which presents unique problems because the number of replicates is usually low and the variance in expression levels is not constant for all genes. For example, the Significance Analysis of Microarrays (SAM) method developed by Tusher *et al.* [6] uses a modified *t*-test statistic. This modified statistic takes into account the greater variability of expression level detection for genes expressed at low levels, and uses the non-parametric false discovery rate control method to select genes whose expression differs significantly under different conditions. An excellent discussion of this and other statistical methods has been provided by Cui and Churchill [7].

Aside from transcriptomics, another active area of data acquisition is the identification of the structure and function of all the proteins in an organism. Proteomic technologies, which include two-dimensional gel electrophoresis followed by tandem mass spectrometry (MS), are being used to identify all the proteins in a cell at the molecular level. At the same time, yeast two-hybrid interactions and large-scale tandem affinity purification (TAP) tagging are being applied to identify proteins that interact to form complexes. From a genetic perspective, collections of *Arabidopsis* gene-knockout mutants are now publicly available and can be used to determine if a gene or protein of interest has any discernable phenotype [8]. With these data, as with the genomic sequencing data, researchers are at the mercy of the quality of the data or resources. For example, it is worrisome that independent all-by-all protein-interaction maps in yeast based on two-hybrid screening show little overlap [9].

Microarray-based approaches: fish and chips

For historical reasons, transcript profiling has dominated high-throughput genomic studies in plants — this technology has been around longer than others. Transcript-profiling experiments have often followed a simple experimental design in which, for example, plants are subjected to a specific biotic or abiotic stress and then assayed for changes in gene expression. First generation experiments led to the general conclusion that plants often respond to different environments using overlapping batteries of genes, but that these expression outputs most probably require multiple signaling pathways [10–16].

Box 1 Publicly available web resources for *Arabidopsis* systems biology.

<http://www.Arabidopsis.org>

The *Arabidopsis* Information Resource website is the main repository for genomic sequence information, including sequence information from ecotypes other than Columbia, such as Landsberg *erecta* [28]. This website also contains copious information on other aspects of *Arabidopsis* biology.

<http://ssbdjc2.nottingham.ac.uk/narrays/experimentbrowse.pl>

The NASCArrays database contains gene expression data for more than 350 samples. It is possible to perform 'electronic northern' for a given gene of interest on this website, and to see how it responds transcriptionally under given conditions [29].

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gds>

NCBI's Gene Expression Omnibus (GEO) website contains data from more than 135 Affymetrix GeneChip experiments, in addition to information collected from numerous other *Arabidopsis* expression-profiling experiments performed with other platforms [30].

<http://genome-www5.stanford.edu/>

The Stanford Microarray Database contains expression profiles generated by several hundred cDNA microarray experiments performed on *Arabidopsis* [31].

<http://bbc.botany.utoronto.ca:88/>

The University of Toronto Department of Botany's Affymetrix Database is accessible at this site via a tool, called ExpressionAngler, that identifies genes that respond similarly to one's gene of interest across all of the gene profiling experiments in the database (K Toufighi, E Ly, NJ Provart, unpublished).

<http://www-stat.stanford.edu/~tibs/SAM/>

The Significance Analysis of Microarrays (SAM) method for the identification of genes that are significantly differentially expressed in microarray experiments is available at this site. The tool is available as an easy-to-use plug-in for Excel [6].

<http://www.esat.kuleuven.ac.be/~thijs/BioDemo/MotifSampler.html>

The MotifSampler allows the identification of possible novel regulatory motifs in the promoters of co-regulated genes using a probabilistic Gibbs sampling method [18].

http://www.blueprint.org/bind/search/bind_search.html

The Biomolecular Interaction Network Database (BIND) currently contains information on about 85 interactions between various proteins in *Arabidopsis*, as determined by literature parsing, yeast two-hybrid studies, affinity chromatography and so on. Enter NCBI's taxonomic identifier, 3702, to search for interactions in *Arabidopsis* only [32].

<http://www.Arabidopsis.org/tools/aracyc/>

This website provides a bird's eye view of the metabolic pathways in *Arabidopsis*, with the possibility of overlaying expression information.

More recently, researchers have incorporated other criteria into transcript-profiling experiments to generate more meaningful biological inferences. For example, by first classifying 402 transcription factors into hierarchical clusters on the basis of the similarity of their expression profiles over 56 conditions, it was possible to classify these transcription factors into distinct groups that had not been identified previously [17]. This winnowing down of a large number of transcription factors into shorter lists on the basis of statistical criteria not only uncovered patterns but also better directs the choices of which knockout plants should be screened for specific biological phenotypes.

The same dataset was used to identify novel *cis*-acting elements that are required to upregulate genes in response to pathogen attack [17]. Using MotifSampler ([18]; see Box 1), which uses a probabilistic Gibbs sampling method to identify over-represented motifs within groups of sequences, a novel W-box-like element was identified. This information could now be used to screen WRKY-type transcription-factor knockout lines to iden-

tify the transcription factor(s) responsible for the response to pathogen attack. A similar approach was used to assess which *cis*-element could be responsible for coordinating the expression of genes during rice grain filling [19]. The identification of an AACA element, which had previously been shown to be involved in regulating the expression of a rice glutelin gene [20], was shown by computational analysis to be over-represented in the promoters of 103 genes of diverse function that are upregulated over the course of grain filling.

Another important aspect of the generation of transcript-profiling data involves wet-laboratory-based experimental design. This is particularly true in developmental experiments in which the response of specific cell types to certain conditions needs to be monitored. In one fine example, rather than just homogenizing the plant for RNA isolation, cell types from *Arabidopsis* roots were first purified using green fluorescent protein (GFP)-driven cell-specific gene expression, protoplasting and fluorescently activated cell sorting [21]. This method of purification, followed by k-means clustering, allowed the

cataloging of transcription factors present in one or more of eight localized expression domains. For example, eight MYB-type transcription factors, as compared to just one APETALA2 (AP2)-like transcription factor and no WRKY, HD-ZIP, or bHLH transcription factors, are upregulated in the epidermal atrichoblast region only during longitudinal cell expansion. The sets of regionally upregulated genes were also analyzed for overrepresentation of genes that are related to hormone biosynthesis. As expected, the auxin-regulated gene region coincides with the region of the root in which auxin is known to affect vascular development. More importantly, jasmonic acid (JA)-related genes were found in the epidermis of longitudinally expanding cells and in the lateral root cap; and gibberellic acid (GA)-related genes were found in the stele, endodermis, and cortex of the longitudinally expanding section of the root. These observations led to the hypothesis of the existence of JA and GA organizing centers in root development. Similarly, Leonhardt *et al.* [22] used a protoplasting approach to study abscisic acid (ABA) signal transduction at the level of global gene expression in the guard cells of *Arabidopsis*.

Proteomics-based approaches

Although proteomics has lagged behind transcriptomics, the functions of all proteins and how they form complexes during growth and development are now beginning to be systematically explored in plants [23,24]. For example, the yeast two-hybrid system, which detects protein–protein interaction, was used to screen more than 5 million protein pairs from rice to identify potential protein–protein interactions [24]. Although yeast two-hybrid analysis is notorious for generating both false-positive and false-negative results, dynamic biological patterns began to emerge when this interaction dataset was combined with gene expression data from rice plants that had been subjected to various growth environments and with other information available on rice. For example, a PROTEIN PHOSPHATASE2A (PP2A) regulatory subunit, which is induced upon cold treatment, interacts with an inositol phosphatase-like protein (IPP) and a stress-regulated 14-3-3 protein. The IPP also interacted with a drought-repressible zinc-finger protein, and the 14-3-3 protein interacts with an ATP synthase that localizes to the *oa3.1* drought-tolerance-related quantitative trait locus (QTL) of rice. Although these integration points intuitively make sense on a biological level, systems analysis is now allowing the application of this type of analysis to unrelated datasets. For example, Bayesian statistical methods, which allow the combination of dissimilar types of data, were used in yeast to build a protein–protein interaction network from independent sources such as cell co-localization data, RNA co-expression data and co-essentiality data [25]. This statistical analysis gave a protein–protein interaction map that is more accurate than those created using standard two-hybrid-interaction

datasets. Thus, as more high-throughput datasets from different independent approaches become available, the reliability of new overall maps will increase.

Conclusions

High-throughput data acquisition and systems approaches are generating many intriguing insights into plant signal transduction and gene regulation. In coming back to our analogy of the cell as a city, it is obvious that many of the conclusions being reached at present are actually hypotheses that must be tested using traditional genetic or cell-biological methods. Nonetheless, the availability of large-scale knockout collections will accelerate the wet-laboratory work necessary to provide an understanding of the biological roles of the various players in signal transduction and gene regulation. In addition, the availability of publicly available gene expression datasets should increase our confidence in the hypotheses that are being generated. With this said, however, the goal is still to understand Tokyo — and we are a long way from this. Novel approaches, such as the automated image analysis of real-time laser confocal microscopic observations of GFP fusions (e.g. in shoot apical meristem development), will provide a deeper understanding of the cell–cell and long-distance signaling mechanisms in plant growth and development. Such observations can be used to generate differential equations, which can be numerically solved to yield computer models of plant development and signaling [26,27]. Thus, in spite of the apparent profusion of data, orders of magnitude more data will be needed to reach the goal of complete understanding of signal transduction and gene regulation in plants.

References

1. Zhu H, Snyder M: 'Omic' approaches for unraveling signaling networks. *Curr Opin Cell Biol* 2002, **14**:173-179.
2. *Arabidopsis* Genome Initiative: Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 2000, **408**:796-815.
3. Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H *et al.*: A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 2002, **296**:92-100.
4. Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X *et al.*: A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 2002, **296**:79-92.
5. Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M *et al.*: Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* 2003, **302**:842-846.
6. Tusher VG, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 2001, **98**:5116-5121.
7. Cui X, Churchill GA: Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol* 2003, **4**:210.
8. Weisshaar B: Reverse genetic stocks. Multinational *Arabidopsis* steering committee annual report 2003. pp. 18-21.
9. Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M: Bridging structural biology and genomics:

- assessing protein interaction data with known complexes.** *Trends Genet* 2002, **18**:529-536.
10. Cheong YH, Chang HS, Gupta R, Wang X, Zhu T, Luan S: **Transcriptional profiling reveals novel interactions between wounding, pathogen, abiotic stress, and hormonal responses in *Arabidopsis*.** *Plant Physiol* 2002, **129**:661-677.
 11. Fowler S, Thomashow MF: ***Arabidopsis* transcriptome profiling indicates that multiple regulatory pathways are activated during cold acclimation in addition to the CBF cold response pathway.** *Plant Cell* 2002, **14**:1675-1690.
 12. Kreps JA, Wu Y, Chang HS, Zhu T, Wang X, Harper JF: **Transcriptome changes for *Arabidopsis* in response to salt, osmotic, and cold stress.** *Plant Physiol* 2002, **130**:2129-2141.
 13. Seki M, Narusaka M, Ishida J, Nanjo T, Fujita M, Oono Y, Kamiya A, Nakajima M, Enju A, Sakurai T *et al.*: **Monitoring the expression profiles of 7000 *Arabidopsis* genes under drought, cold and high-salinity stresses using a full-length cDNA microarray.** *Plant J* 2002, **31**:279-292.
 14. Glazebrook J, Chen W, Estes B, Chang HS, Nawrath C, Metraux JP, Zhu T, Katagiri F: **Topology of the network integrating salicylate and jasmonate signal transduction derived from global expression phenotyping.** *Plant J* 2003, **34**:217-228.
 15. Lorenzo O, Piqueras R, Sanchez-Serrano JJ, Solano R: **ETHYLENE RESPONSE FACTOR1 integrates signals from ethylene and jasmonate pathways in plant defense.** *Plant Cell* 2003, **15**:165-178.
 16. Rabbani MA, Maruyama K, Abe H, Khan MA, Katsura K, Ito Y, Yoshiwara K, Seki M, Shinozaki K, Yamaguchi-Shinozaki K: **Monitoring expression profiles of rice genes under cold, drought, and high-salinity stresses and abscisic acid application using cDNA microarray and RNA gel-blot analyses.** *Plant Physiol* 2003, **133**:1755-1767.
 17. Chen W, Provart NJ, Glazebrook J, Katagiri F, Chang HS, Eulgem T, Mauch F, Luan S, Zou G, Whitham SA *et al.*: **Expression profile matrix of *Arabidopsis* transcription factor genes suggests their putative functions in response to environmental stresses.** *Plant Cell* 2002, **14**:559-574.
 18. Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P, Moreau Y: **A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling.** *Bioinformatics* 2001, **17**:1113-1122.
 19. Zhu T, Budworth P, Chen W, Provart N, Chang H-S, Guimil S, Su W, Estes B, Zou G, Wang X: **Transcriptional control of nutrient partitioning during rice grain filling.** *Plant Biotech J* 2003, **1**:59-70.
 20. Wu C, Washida H, Onodera Y, Harada K, Takaiwa F: **Quantitative nature of the Prolamin-box, ACGT and AACCA motifs in a rice glutelin gene promoter: minimal cis-element requirements for endosperm-specific gene expression.** *Plant J* 2000, **23**:415-421.
 21. Birnbaum K, Shasha DE, Wang JY, Jung JW, Lambert GM, Galbraith DW, Benfey PN: **A gene expression map of the *Arabidopsis* root.** *Science* 2003, **302**:1956-1960.
 22. Leonhardt N, Kwak JM, Robert N, Waner D, Leonhardt G, Schroeder JI: **Microarray expression analyses of *Arabidopsis* guard cells and isolation of a recessive abscisic acid hypersensitive protein phosphatase 2C mutant.** *Plant Cell* 2004, **16**:596-615.
 23. Kersten B, Feilner T, Kramer A, Wehrmeyer S, Possling A, Witt I, Zanol MI, Stracke R, Lueking A, Kreutzberger J *et al.*: **Generation of *Arabidopsis* protein chips for antibody and serum screening.** *Plant Mol Biol* 2003, **52**:999-1010.
 24. Cooper B, Clarke JD, Budworth P, Kreps J, Hutchison D, Park S, Guimil S, Dunn M, Luginbuhl P, Ellero C *et al.*: **A network of rice genes associated with stress response and seed development.** *Proc Natl Acad Sci USA* 2003, **100**:4945-4950.
 25. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: **A Bayesian networks approach for predicting protein-protein interactions from genomic data.** *Science* 2003, **302**:449-453.
 26. Shapiro BE, Levchenko A, Meyerowitz EM, Wold BJ, Mjolsness ED: **Cellerator: extending a computer algebra system to include biochemical arrows for signal transduction simulations.** *Bioinformatics* 2003, **19**:677-678.
 27. Peak D, West JD, Messinger SM, Mott KA: **Evidence for complex, collective dynamics and emergent, distributed computation in plants.** *Proc Natl Acad Sci USA* 2004, **101**:918-922.
 28. Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M *et al.*: **The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community.** *Nucleic Acids Res* 2003, **31**:224-228.
 29. Craigon DJ, James N, Okyere J, Higgins J, Jotham J, May S: **NASCArrays: a repository for microarray data generated by NASC's transcriptomics service.** *Nucleic Acids Res* 2004, **32** (Database issue):D575-D577.
 30. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, Rapp BA: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2001, **29**:11-16.
 31. Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA *et al.*: **The Stanford Microarray Database.** *Nucleic Acids Res* 2001, **29**:152-155.
 32. Bader GD, Betel D, Hogue CW: **BIND: the Biomolecular Interaction Network Database.** *Nucleic Acids Res* 2003, **31**:248-250.