Part I

Systems Biology: An Overview

Annual Plant Reviews (2009) **35**, 3–40
doi: 10.1111/b.9781405175326.2009.00001.x
www.interscience.wiley.com

Chapter 1

SYSTEMS BIOLOGY: PRINCIPLES AND APPLICATIONS IN PLANT RESEARCH

Gloria M. Coruzzi,[2] Alejandro R. Burga,[1] Manpreet S. Katari[2] and Rodrigo A. Gutiérrez[1,2]

[1] *Departamento de Genética Molecular y Microbiología, Pontificia Universidad Católica de Chile, Santiago, Chile*
[2] *Department of Biology and Center for Genomics and Systems Biology, New York University, New York, NY, USA*

'We cannot solve our problems with the same thinking we used when we created them'
Albert Einstein

Abstract: Plants have played a major role in the geochemical and climatic evolution of our planet. Today, in addition to their fundamental ecological importance plants are essential for humans as the main source of food, provide raw materials for many types of industry and chemicals for medical applications. It is thus daunting to realize how little we understand about plant systems. To date, only approximately 15% of the genes of *Arabidopsis thaliana*, the most explored model system for plant biologists, have been characterized experimentally. Systems biology offers the opportunity to increase our understanding of plants as living organisms, by generating a holistic view of the organism grounded at the molecular level. In this chapter, we discuss the basics of systems biology, the data and tools we need for systems research and how it can be used to produce an integrated view of plant biology. We finish with a discussion of case studies, published examples of plant systems biology research and their impact on our knowledge of plants as integrated systems.

Keywords: network; Arabidopsis; genomics; bioinformatics; modelling; data integration

1.1 Introduction

1.1.1 Systems thinking

What is systems biology? We advocate a definition anchored in the general systems theory: 'The exercise of integrating the existing knowledge about biological components, building a model of the system as a whole and extracting the unifying organizational principles that explain the form and function of living organisms.' Systems thinking is not a new trend, but dates back to the end of the 1800s and the beginning of the 1900s. One of the pioneers of systems thinking was the Russian philosopher Alexandr Bogdanov (1873–1928). His interests and writings ranged from social, to biological and physical sciences. His work anticipated in many important ways Norbert Weiner's '*Cybernetics*' and Ludwig von Bertalanffy's '*General Systems Theory*'. Bogdanov (1980) proposed that all physical, biological and human sciences could be unified by treating them as systems of relationships and by seeking the organizational principles that underlie all systems.

We would like to illustrate systems thinking with the following example. Imagine you are standing in front of 'La Grande Jatte', the painting by the famous pointillist artist George Seurat. If we are close to the painting, we can easily distinguish the small coloured spots, but we lose their pattern, in fact we may not even realize they are part of a composition. Only when we are standing back far enough, we can appreciate the subject and beauty of the painting, as people standing by the lake. Most scientists today are standing very close to their subjects. They know their area of research extremely well, but cannot easily place their knowledge in the global context. The aim of systems biology is to move away from the detail and instead produce a global view of the system under scrutiny. This is achieved not solely by system-wide data collection, integration and analysis. It is also about a change in the research focus from the elements to the interactions and to the discovery of higher levels of organization and the emerging properties at these higher levels of organization.

1.1.2 Complexity and robustness in biological systems

Everyone agrees that biological systems are complex. But what does complexity mean? Is complexity in biological systems just a consequence of the inherent difficulty in understanding them? In this chapter, we would like to use a more precise definition of complexity. We refer to a system as a complex when it exhibits the following characteristics: (1) It is composed of many different elements. (2) The constituent elements interact. (3) The elements and their interactions are dynamic, and are often governed by non-deterministic rules. (4) The elements and/or interactions can be influenced by external

factors. As a consequence of these properties, complex systems exhibit unexpected or emergent behaviour that cannot be understood from studying the parts in isolation. In addition, mathematical models that describe them usually involve nonlinear behaviour. Even the simplest life forms (e.g. mycoplasma) are complex by this definition. Living systems are composed of a large number of elements with diverse chemical properties (nucleic acids, proteins, carbohydrates, lipids, ions and many small molecules or metabolites) that are intricately interconnected. We all know that these elements and connections are dynamic and respond to internal or external factors. For instance, the messenger RNA (mRNA) levels of hundreds of genes that code for hundreds of different proteins can vary over time due to the action of signal transduction cascades activated by developmental or environmental cues.

In addition to complexity, and perhaps because of it, biological systems are robust. This key and ubiquitous feature of biological systems refers to their ability to maintain proper functions despite internal and/or external perturbations and uncertainty (Stelling *et al.*, 2004; Kitano, 2007). Robustness is an example of a property that emerges at the system level and that cannot be understood at the individual part level (Kitano, 2004). The types of perturbations encountered by living systems are varied and include: stochastic noise (e.g. due to low copy number of cellular components), physiological and developmental signals, environmental change and genetic variation. It is important to distinguish robustness from homeostasis. As indicated above, robustness is related to preserving function. In contrast, homeostasis refers to maintaining the state of the system. A system is robust as long as it maintains functionality, even if this is achieved by moving to a new steady state. For instance, during the diauxic shift yeasts drastically changes from anaerobic to aerobic metabolism. This transition involves genome-wide changes in gene expression and the reprogramming of metabolic pathways (DeRisi *et al.*, 1997). In this example, the state of the system is dramatically altered. However, the system is robust as it continues to produce adenosine triphosphate and grow. Robustness allows for changes in the structure and components of the system in response to perturbations, but maintains specific functions (Kitano, 2004).

Robustness can be achieved by multiple mechanisms (for a review see Kitano, 2004). First, negative or positive feedbacks allow for robust dynamic responses in regulatory contexts as diverse as the cell cycle, circadian clock and chemotaxis. Second, robustness can also be achieved by providing multiple means to achieve a specific function. These alternative or fail-safe mechanisms encompass the typically observed phenomena in living systems of redundancy, overlapping function and diversity. Third, the modularity of living systems is an effective mechanism of containing perturbations and damage locally to minimize the effect on the whole system. Fourth, decoupling isolates low-level variation from high-level functionalities. All four different mechanisms are typically found in a living organism.

1.1.3 Robust evolving systems

How can complex biological systems be both robust to mutational perturbations and at the same time accumulate heritable changes that produce new adaptive traits during evolutionary time? Intuitively, robustness appears to act against evolutionary change. The extent of phenotypic change caused by natural selection depends on phenotypically expressed genetic variation that is buffered in a robust system. This is true in the short term, but in the long term silent mutations accumulate that can express under certain conditions (e.g. a specific environmental stress). This phenomenon was originally postulated and studied by Waddington (1959). He showed that once the phenotypic buffering capacity of an organism is exhausted by severe perturbations, altered phenotypes can emerge, that is, the phenotype is no longer 'canalized'. Molecular evidence for Waddington's original observations was presented first in *Drosophila melanogaster* (Rutherford and Lindquist, 1998) where Hsp90 (a heat shock-induced chaperone) was shown to be an evolutionary conserved protein affecting phenotypic variation. Inhibiting Hsp90 by mutation or pharmacological means caused an increase in a wide range of altered phenotypes in Drosophila. Those phenotypes could be fixed through breeding and interestingly became robust and independent of environmental perturbations (temperature) or Hsp90 inhibition. Later studies in *Arabidopsis thaliana* (Queitsch *et al.*, 2002) showed that in plants, reduction of Hsp90 function resulted in an increase in phenotypic variation in the absence of genetic variation. Probably, epigenetic mechanisms are involved in this phenomenon (Sollars *et al.*, 2003); however, the exact molecular mechanisms are still a mystery (Salathia and Queitsch, 2007). Robustness and phenotypic variability are not antagonistic forces, but are intimately related concepts. Isalan *et al*. (2008) recently addressed this relationship in *Escherichia coli*. They systematically explored the effect of adding new edges in the *E. coli* gene network by expressing hundreds of promoter-open reading frame combinations. The open reading frames coded for transcription factors working at different levels of the gene network hierarchy. Remarkably, nearly 95% of the new edges were tolerated (i.e. *E. coli* shows a robust gene network) and some of the new strains grew better than the wild type under diverse selective pressures (Isalan *et al.*, 2008).

1.2 Network biology

How do we approach the complexity of biological systems? In 1736, the famous mathematician Leonhard Euler found a solution to an old problem: 'The seven bridges of Königsberg.' In those days, Königsberg was a city of the Kingdom of Prussia, set on the Pregel River. The city included two large islands that were connected to each other and the mainland by seven bridges. The problem was to decide whether it was possible to walk by a route that crossed each bridge exactly once. Euler did not solve the problem

empirically, but instead made a fundamental abstraction of the problem. First, he eliminated all the features except the landmasses and the bridges. Then, he replaced each landmass with a dot (vertex or node) and each bridge with a line (edge or link). The resulting structure was a graph. Euler realized that the degrees of the nodes were an important property of the graph and he could later demonstrate in terms of degrees that the hypothetic walk was impossible. The degree tells us how many links a particular node has to other nodes. The key to Euler's success was that he could abstract the fundamental elements of the system and their relationships and represent them in a way that favoured their analysis. He knew, for instance, that parameters such as the length of the bridges or whether they were made of stone or wood were irrelevant for his purpose. Similar principles and approaches can apply to complex systems such as living organisms. By abstracting and focusing on the important features for system form and function, a detailed and comprehensive yet tractable view of the system can be obtained. A highly successful and now widespread abstraction to represent complex biological systems uses graphs. The following sections will discuss the basics of how graphs are applied in biological research. (Note: For more details regarding the subjects covered in this section please see Chapter 2 of this volume.)

1.2.1 General principles

When networks, or graphs in a more formal mathematical language, are applied to molecular biology nodes (or vertices) represent the molecules present inside a cell (e.g. proteins, RNAs and/or metabolites) and links (or edges) between nodes represent their biological relationships (e.g. physical interaction, regulatory connections, metabolic reactions). For example, in protein interaction networks, protein complexes can be represented as networks where nodes represent proteins and their edges represent the physical interactions between them. In metabolic networks, nodes can represent metabolites and the edges the metabolic reactions that transform one metabolite into another (Wagner and Fell, 2001). In genetic networks, nodes can represent genes and the link between them a genetic interaction, such as synthetic lethality. But other abstractions are possible. For instance, nodes can also be used to represent biological processes and the edges connecting them indicate the functional relationship between the processes. This network simplification has uncovered basic principles of the structure and organization of different types of molecular networks in many different organisms (Barabasi and Oltvai, 2004) and has been successful for biological research. Examples of the use of networks in plant biology are discussed later in this chapter.

1.2.2 Network properties

Networks have diverse structures and topologies. Prior to the work by Barabasi and Albert (Barabasi and Albert, 1999; Barabasi and Oltvai, 2004),

networks were generally considered as having either regular (e.g. with a square lattice) or random topologies. To construct a random network according to the classical Erdös–Rényi model, two nodes are chosen randomly from a pool of N nodes and a link is established between them (Erdös and Rényi, 1960). This procedure results in a network where the number of edges for each node (degree) follows a Poisson distribution. In random networks, there are a few nodes that are lowly or highly connected and most nodes have roughly the same number of links. In contrast to random networks, one of the common architectural features of naturally occurring networks including molecular networks is to present a scale-free topology (Barabasi and Albert, 1999). Networks with scale-free topology are characterized by a power law degree distribution: $P(k) \sim k^{-\gamma}$, where $P(k)$ is the probability that a selected node has k links and γ is the degree exponent. The γ value for most naturally occurring networks varies between 2 and 3 (Albert *et al.*, 2000). Whereas in random networks most nodes have approximately the same degree or number of connections, in scale-free networks there are many nodes that are poorly connected and few nodes that are very highly connected (Albert *et al.*, 2000). The highest degree nodes (highest number of edges) are typically referred to as 'hubs' and are important for the architecture and function of the network.

One important consequence of the scale-free topology of naturally occurring networks is robustness. Scale-free networks show an extraordinary tolerance to perturbations as compared to random networks of equivalent size. For example, 5% of the nodes in a scale network can fail without affecting the mean path length (Albert *et al.*, 2000). In contrast, an informed targeting of 5% of the most connected nodes results in a doubling of the mean path length (Albert *et al.*, 2000). This phenomenon can be understood based on the low frequency of highly connected nodes. And therefore the probability of randomly targeting a hub in a scale-free network is very low. This observation is correlated with findings in biological networks. In yeast, approximately 20% of proteins with less than 5 connections are essential, in contrast to the 62% when considering proteins with more than 15 connections (Jeong *et al.*, 2001). Removing a hub protein has a high probability of resulting in a lethal phenotype, supporting their fundamental role. Similarly, a malicious hacker trying to affect Internet function may have a higher chance of succeeding by damaging servers that are hubs. It was recently shown that human proteins targeted by the Epstein–Barr virus were enriched for hub proteins (Calderwood *et al.*, 2007). Viruses could have evolved to attack key hub proteins in their host's proteomes.

Another interesting feature of naturally occurring networks is their 'small world effect'. In order to understand it, we must first introduce the concept of 'path'. Distance in networks is measured by the number of links that we need to pass when travelling from node A to node B (path from A to B). The path with the smallest number of edges between nodes A and B is the 'shortest path'. If we take all possible pair of nodes in a network and

calculate the average length of all the shortest paths, we obtain a measure called the 'mean path length'. The diameter of the network is the maximal distance (from the shortest path set) between any pair of nodes. Studies regarding social networks performed by Stanley Milgram in 1967 showed that the mean number of individuals required to connect one arbitrary person to another arbitrary person anywhere in the USA is only six; thus the concept of six degrees of separation. Another example of small mean path lengths within a large network is the World Wide Web, with over 800 million nodes, it has a mean path length of only 19 (Albert *et al.*, 1999). The property that every node in these large networks is separated by a few links from all the other nodes in the network is known as the 'small world effect'. The architecture of *Caenorhabditis elegans* nervous system, the power grid of the western United States and the collaboration graph of film actors also show a 'small world effect' (Watts and Strogatz, 1997).

An additional interesting property of networks is clustering. Clustering in networks can be intuitively explained using social networks as an example. In social networks, two of your friends will have a greater chance of knowing one another than two people chosen at random from the entire population. This is due to their common acquaintance with you. This property is quantified with the clustering coefficient. The clustering coefficient for a node k measures how connected are its neighbours. Mathematically, it is calculated as the ratio of the number of edges that exist between the neighbours of k and the total number of edges possible between the neighbours. The clustering coefficient is 1 in a fully connected network. In real-world networks, the clustering coefficient typically has values of 0.1–0.5. These values are much higher than what obtained from random networks.

The combined presence of scale-free topologies and the high degree of clustering in real-world networks is a consequence of their hierarchical structure. Many real-world networks are fundamentally modular, where groups of highly connected nodes can be identified that are poorly connected to nodes outside of the group (for a review see Ravasz *et al.*, 2002). Hartwell *et al.* (1999) defined a biological module as a discrete entity of the cell whose function is separable from those of other modules. Biological network modules have been correlated with biological function in various organisms from yeast to animals to plants (Han *et al.*, 2004; Gunsalus *et al.*, 2005; Gutierrez *et al.*, 2007). An interesting question to be addressed is whether a central control module exists or the control is shared between all the modules in a cell. In addition, an evolutionary implication of modularity is that shifting the connections between modules could potentially change the phenotype of individuals. Therefore, emphasis must be given in the following years to the study of hubs proteins (most highly connected nodes) and their regulation from a systemic perspective. These proteins can occupy key places connecting diverse modules and their study could provide important insights about cellular regulatory mechanisms.

1.3 Experimental approaches for plant systems biology

A systems approach to biology is possible today because of the breakthrough in technologies that allow us to measure and connect the entire complement of defined molecules inside an organism (e.g. transcriptome, proteome). The new influx of high-throughput data is shifting biology from a reductionist to a systems-level view. Practically speaking, systems biology generates and integrates different types of genome-wide data in living organisms to produce models of the system as a whole (Kitano, 2002; Shannon *et al.*, 2003). Ideker, Galitski and Hood (Shannon *et al.*, 2003) described a fundamental framework to carry out research in systems biology: (1) define all the components of the system, (2) systematically perturb and monitor components of the system, (3) reconcile the experimentally observed responses with those predicted by the model and (4) design and perform new perturbation experiments to distinguish between multiple or competing hypotheses. Below, we describe the different data types currently available for plant systems biology.

1.3.1 Enumerating the parts

1.3.1.1 The genome

The first step to understand a living organism from a systems point of view is to enumerate its basic components. The genome of an organism is the complete hereditary information encoded in the DNA (or RNA in some viruses). The first genome of a complex free-living organism to be completed was that of *Haemophilus influenza* (Fleischmann *et al.*, 1995). The tremendous advances in sequencing technologies since then, has catapulted us into the genomic era. Today, the scientific community benefits from the complete genome sequences of many organisms including several plants. In 2000, the complete genome sequence of the plant *A. thaliana* was published (Initiative, 2000). In 2005, the first monocotyledoneous plant was sequenced, *Oryza sativa* (Locke *et al.*, 2005). To date, in addition to Arabidopsis and rice, the complete sequenced genomes of *Chlamydomonas reinhardtii* (Merchant *et al.*, 2007), *Populus trichocarpa* (Tuskan *et al.*, 2006) and *Vitis vinifera* (Jaillon *et al.*, 2007) are available. In addition, at the beginning of 2008, the first genome sequence of a genetically modified plant was reported, for the SunUp *Carica papaya* (Ming *et al.*, 2008). Genomes provide the parts lists as far as genes and their products (proteins and RNAs) are concerned.

Due to lack of financial support or simply due to the large size, the genomes of many species are not and may not be completely sequenced. In addition, much of the DNA in large genomes is thought to correspond to repetitive DNA and transposable elements (Bennet and Ilia, 1995). There are two main approaches for 'sampling' the gene space of genomes without sequencing them completely: (1) sequencing expressed sequence tags (ESTs) (Adams *et al.*, 1991) and (2) sequencing genomic sequences that have been filtered based on

either hypomethylation (Bennetzen *et al.*, 1994; Rabinowicz *et al.*, 1999, 2003a) or High-Cot (Yuan *et al.*, 2003). EST sequencing involves preparing cDNA of the transcribed messages and then randomly sequencing them from either the 3′ or the 5′. The advantage of EST sequencing is that the sequences obtained represent the transcribed regions of the genome. The disadvantages are that ESTs typically do not represent the entire cDNA (only about 500 bp) and they tend to represent only roughly 50% of the total genes in the genome. Large stretches of repetitive DNA and transposable elements tend to be hypermethylated compared to gene coding regions which are unmethylated, for example 95% of the maize exons are thought to be unmethylated (Rabinowicz *et al.*, 2003b). Genomic libraries can be prepared enriched for unmethylated DNA by cloning in bacterial hosts that destroy methylated DNA. Alternatively, the High-Cot method removes repetitive DNA by depending on the rapid association of repetitive DNA. The advantage of sequencing filtered genomic sequences over EST sequencing is that it is possible to obtain the introns and regulatory regions of the gene rather than just the transcribed portion of the gene. Palmer *et al*. (2003) annotated genes from methyl-filtrated sequences and compared them to random genome sequences from maize. They observed a twofold reduction in the number of repetitive reads and fivefold increase in the number of exonic regions in the methyl-filtrated sequences (Palmer *et al.*, 2003). The drawback to this approach is the significant amount of repetitive DNA obtained. Despite the limitations, the different genome sampling methods are still the best alternative to sequencing the entire genome for very large genome sequences. With the advent of highly parallel sequencing technologies, such as the platforms developed by 454 (Margulies *et al.*, 2005) or Solexa Illumina (Bennett, 2004), these limitations may be less relevant in the near future. In addition, in the years to come, the focus may shift from sequencing depth within individual genomes to sampling as many genomes as possible to determine the gene complement of the biosphere.

The first step after obtaining the genome sequence of an organism is the annotation of its genes (identification of their basic elements as exons, introns, regulatory sequences). Several strategies are available for gene discovery: *in silico* gene prediction, information obtained from ESTs, full-length cDNA, tilling and expression arrays, MPSS (massive parallel signature sequencing), SAGE (serial analysis of gene expression) (Alonso and Ecker, 2006). Today, many specialized software are available in the market as well as freeware. Even though *in silico* predictions are an excellent start, they have to be used critically because they are not always accurate. Indeed, gene predictions and genome annotations are typically dynamic improving over time. For example, at least 40% of the original predictions made for the Arabidopsis genome were subsequently found to be wrong and corrected in later releases of the annotation. The next step after gene discovery is gene functional annotation. This functional analysis can be done in several ways: computational predictions, gene expression, mutant analysis and protein–protein interactions, just

to name some of the most common approaches (for an excellent review of this topic please read Alonso and Ecker, 2006). However, high-throughput screens of mutants are fundamental for accessing plant gene functions on a genomic scale (Alonso and Ecker, 2006).

1.3.1.2 Epigenome

Epigenetics is the study of heritable traits that do not involve changes in the DNA sequence. These changes alter the chromatin structure which affect regulation of gene expression. Two well-studied forms of epigenetic regulation in plants are methylation of the DNA molecule at the cytosine base and post-translational modifications of histones (Henderson and Jacobsen, 2007). Heterochromatic regions of plants tend to be highly methylated (Rabinowicz *et al.*, 2003b) which is why sequencing methods such as methyl filtration produce sequences that are enriched in gene coding regions. To understand the importance of DNA or histone modifications for gene regulation, we need to look at DNA or histone modification patterns across the genome and correlate them with gene expression data. Several research groups have carried out such studies in plants. The experimental approach involves isolating nuclear DNA and enriching for the methylated DNA fraction. The methylated DNA is then hybridized to an Arabidopsis whole genome tiling microarray to determine global DNA methylation patterns (Zhang *et al.*, 2006; Zilberman *et al.*, 2007). These studies have uncovered several interesting aspects of the epigenome in Arabidopsis. For example, the heterochromatic regions are heavily methylated, approximately a third of Arabidopsis genes are methylated within the transcribed region, methylation is biased away from the 5′ and 3′ regions in Arabidopsis genes, genes that are methylated in the transcribed regions tend to be expressed at high levels and genes that are methylated in the promoter region tend to be tissue specific (Zhang *et al.*, 2006; Zilberman *et al.*, 2007). Additional studies have been carried out to look at histone modifications (Zhang *et al.*, 2007). Epigenetics adds another level to the regulatory complexity of gene expression. Elucidating the importance and function of epigenetic modifications in the years to come will be crucial to understanding how genomes respond to internal and external perturbations.

1.3.1.3 Transcriptome (including RNAs and small RNAs)

The transcriptome is the set of all the parts of the genome that are expressed as RNA transcripts in one or several populations of cells in a given time and a given environmental condition. Not so many years ago, it was unthinkable to propose measuring the expression of thousands of genes in one experiment. Northern blots were the only choice for molecular biologists and biochemists. Microarray technology has been around since the early 1990s, but the high cost of this technology precluded widespread utilization by the scientific community. Patrick Brown and colleagues made this technology cheaper and easier to do (Schena *et al.*, 1995) and triggered a revolution in transcriptome studies. Interestingly, Arabidopsis was chosen as a case study

to demonstrate microarray technology in 1995. Arabidopsis was utilized because of its small genome and rich EST collection. At the present time, one of the most widely used systems for gene expression studies in Arabidopsis is the ATH1 Genome Array, designed by Affymetrix in collaboration with The Institute for Genome Research. It contains more than 22 000 probes sets representing approximately 24 000 gene sequences on a single array. This technology allows the simultaneous quantitative determination of the expression level of thousands of genes.

Today, the Arabidopsis community benefits from the success of microarray technology. The bottleneck of research is not longer the acquisition of data, but often its analysis. There are several public databases containing microarray data: Genevestigator (https://www.genevestigator.ethz.ch/), NASC-Arrays (http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl), ArrayExpress (http://www.ebi.ac.uk/arrayexpress), The Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) and Stanford Microarray Database (http://genome-www5.stanford.edu) perhaps among the most popular. Hundreds of microarray experiments (available through these resources) have been performed comparing diverse developmental stages, soil conditions and compositions, pathogen infections, oxidative stress and response to diverse chemicals. The information obtained from these experiments has being useful in the detection of new genes or new gene functions involved in particular processes.

It is important to emphasize that the transcriptome does not refer just to messenger RNAs. In addition to mRNAs, ribosomal RNAs (rRNAs) and tRNAs (transfer RNAs), there is great interest in studying the expression of the large and heterogeneous population of small RNAs in plants (Finnegan and Matzke, 2003). These small RNAs can have important roles for regulation of gene expression as well as other roles (Finnegan and Matzke, 2003). Unfortunately, EST libraries and microarray gene chips were not designed to detect small RNAs such as microRNAs (miRNAs). However, with the advancement of sequencing technologies, we are now able to measure miRNAs and other small RNAs (sRNAs), quantify their expression in different cell types and treatments and begin to understand their functional roles in plants (Lu *et al.*, 2005). (Note: For more on the role of sRNA see Chapter 7 of this volume.)

1.3.1.4 Proteome

The proteome is the set of all the proteins that are expressed in a given system in a given time and under a defined environmental condition. This term was coined by Mark Wilkins and colleagues in 1995 when studying the smallest known self-replicating organism *Mycoplasma genitalium* (Wasinger and Corthals, 2002). There are many questions one would like to address regarding proteomes: How abundant are the proteins? Where are the proteins located? What are the post-translational modifications in these proteins? Answering these questions is not an easy task. Low-abundance proteins can be extremely difficult to detect and there is no polymerase chain reaction

(PCR)-like method for protein amplification. In addition, the great variability of structures and chemical properties of proteins compared with nucleic acids makes their separation and identification a challenge. The most common approach in proteomics is the use of two-dimensional (2D) gel electrophoresis for separating the proteins (which typically resolves proteins based on charge and molecular weight). The separated proteins are then excised from the gel, fragmented and the fragments analyzed by mass spectrometry. The mass fingerprint of the peptides is then compared with databases for identification (Yates, 2000).

In Arabidopsis, large-scale proteomic efforts have been carried out, for example to determine the proteome of organelles such as the chloroplasts (Friso *et al.*, 2004), mitochondria (Heazlewood *et al.*, 2004) and vacuoles (Mitreva *et al.*, 2004). These studies have shown that many *in silico* predictions of protein sub-cellular localization are wrong and they must be verified experimentally. These studies also suggest that cellular trafficking is more complex than previously thought.

Another problem in proteomic studies has been the coverage of the available techniques. Giavalisco *et al.* (2005) carried out a large-scale study in the hope of achieving a complete coverage of Arabidopsis cells proteome using 2D gel electrophoresis and matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF) mass fingerprinting. Despite sampling different tissues in order to increase the probability of finding novel peptides, they could only find 663 different proteins from 2943 spots. Recently, a proteomic study of *A. thaliana* was carried out (Baerenfaller *et al.*, 2008), giving valuable insights into the proteomic map of diverse organs, developmental stages and undifferentiated cultured cells. The authors identified around 13 000 proteins, which accounts for almost 50% of all Arabidopsis predicted gene models. This study not only allowed the corroboration of gene annotations, but also enriched the genome annotation. For example, the authors provided expression evidence for 57 genes models that were not present in the TAIR7 database. Another interesting result was the identification of 571 organ-specific proteins, which could be useful for the mapping of gene regulatory networks that drive the differentiation programmes in plants. (Note: For more on proteomics in plants see Chapter 8 of this volume.)

1.3.1.5 Metabolome

The metabolome refers to the complete set of small molecule metabolites (such as metabolic intermediates, hormones and other signalling molecules, and secondary metabolites) to be found within a biological system (Oliver *et al.*, 1998). One of the main problems when studying the metabolome is the extreme heterogeneous chemical nature of the compounds considered as constituents of the metabolome. This is especially true in plants. Plants are known to have a tremendous enzymatic capacity, allowing for the production of an estimated ∼200 000 different small molecules (Fiehn, 2002). Finding the

equilibrium between coverage and accuracy of measurement is therefore key in the metabolomics field.

With respect to the technology used to explore the metabolome, classical approaches use gas chromatography–mass spectrometry (GC–MS). This technology detects >300 metabolites in plant tissues (Hirai *et al.*, 2004). One of the first studies of this kind was carried out to determine the mechanism of action of herbicides like acetyl CoA carboxylase and acetolactate synthase inhibitors, using GC–MS (Schauer and Fernie, 2006). They could obtain the metabolic profiles of the treated and control seedlings. These metabolic profiles have also been achieved recently using H-1 NMR (proton nuclear magnetic resonance) (Ott *et al.*, 2003). A new technology used in metabolomics is the Fourier transform-ion cyclotron MS that separates metabolites based on differences in their isotopic masses (Hirai *et al.*, 2004). Crude plant extracts without prior separation of metabolites by chromatography are injected directly in this system. The mass resolution (>100 000) and accuracy (<1 ppm) is superior to other technologies.

Metabolite profiling has evolved from a diagnostic tool used in agriculture to a valuable source of information for gene function prediction in plants. Classically, gene function prediction was made based on studies at the mRNA or protein levels. In the year 2000, Fiehn and co-workers developed a novel tool for plant functional genomics (Fiehn, 2002). Using GC–MS, they automatically quantified 326 distinct compounds from *A. thaliana* leaf extracts, half of those whose chemical structure could be assigned. Four Arabidopsis genotypes were compared (two homozygous ecotypes and a mutant of each ecotype) and distinctive metabolite profiles of each genotype were reported. Data mining tools enabled the assignment of 'metabolic phenotypes'. Metabolite profiling studies have been performed in a diverse array of plant species (Schauer and Fernie, 2006) including: Arabidopsis (Fiehn, 2002), tomato (Schauer *et al.*, 2005), potato (Roessner *et al.*, 2001), rice (Young *et al.*, 2005), strawberry (Aharoni *et al.*, 2002) and eucalyptus (Merchant *et al.*, 2007). Metabolic profiling is also being used in studies of environmental perturbations in order to understand the complex shifts under nutrient limitation and biotic stress. The use of forward and reverse genetics in conjunction with metabolite profiling plays an important role in gene annotation, characterization of biochemical pathways and the identification of new genes for their use in biotechnological applications. (Note: For more details on metabolomics in plants, see Chapter 9 of this volume.)

1.3.1.6 Ionome

The ionome was first described to include all the metals, metalloids and non-metals present in an organism (Lahner *et al.*, 2003), which are involved in a broad range of important biological phenomena including: electrophysiology, signalling, enzymology, osmoregulation and transport (Hesse and Hoefgen, 2006). There is a small and sometimes diffuse boundary between ions and metabolites. Take for example macronutrients (essential elements

used by plants in relatively large amounts for plant growth) like nitrogen, sulfur and phosphorus and metals like manganese, iron, zinc and copper essential for metalloprotein's activity. High-throughput ion-profiling strategies for genomic scale profiling of nutrient and trace element have been utilized for ionome studies (Lahner *et al.*, 2003). Using inductively coupled plasma mass spectrometry (ICP-MS) technology, it was possible to profile shoots from more than 40 000 plant samples (at a rate of 1000/per month) including diverse transferred DNA (T-DNA) insertional mutants and also mutants generated by fast neutron treatment. They have also characterized the ionome of Arabidopsis shoot and seed under standard soil growth conditions. (Note: For more information on ionomics in plants, see Chapter 10 of this volume.)

1.3.2 Systematic characterization of interactions between components

In the previous section, we briefly outlined some of the components that constitute a biological system and for which we have a substantial amount of global experimental data. However, even if we had a complete set of accurate measurements for every molecular component of the cell, we would not be able to reconstruct the system. We cannot think of these components as apart in space and time. The essence of biological systems is to understand not only what the parts are, but also how these parts interact. As discussed in other sections and chapters in this book, it is from the interactions of the parts that the properties typically associated with biological systems emerge. In this section, we shall briefly review the interactions that are better understood at a genome-wide level. It is important to emphasize that Arabidopsis and plants, in general, lag behind other systems in the systematic characterization of the interactions between molecular components. As a community, we should attack this limitation in order to advance plant systems research.

1.3.2.1 Metabolic pathways

Metabolic pathways were classically studied by isotopic radiolabelling of metabolites and biochemical isolation and characterization of the enzymes involved in those pathways. All these efforts resulted in the establishment of linear or cyclic pathways, where metabolites are continuously transformed by enzyme-catalyzed reactions with a defined stoichiometry. One way to represent all those reactions is constructing a network where each node represents a metabolite and a link between two nodes represents the enzymatic reaction that converts one into the other. These metabolic networks show a scale-free topology and small world which probably reflects the need of the system to have a wide and plastic response to environmental perturbations (Wagner and Fell, 2001; Ravasz *et al.*, 2002). Metabolic pathways for plants are stored in public databases such as AraCyc (Mueller *et al.*, 2003) and KEGG (Aoki and Kanehisa, 2005). These databases are built from knowledge in Arabidopsis but also from the enzymes and pathways studied in other

species. Another important source of information is the Arabidopsis Reactome (www.arabidopsisreactome.org), inspired by the human Reactome project. It covers biological pathways ranging from the basic processes of metabolism to high-level processes such as hormonal signalling, and it includes also metabolic pathways for other plant species.

1.3.2.2 Co-expression networks

At the present time, one of the major sources of genome-wide information for Arabidopsis is microarray technology. By comparing gene expression patterns obtained in experiments across a wide range of treatments, networks of co-expression relationships can be built. Basically, the similarity of gene expression between two genes is calculated using a statistic such as the Pearson correlation coefficient. The nodes in the network represent the transcripts and the link between two nodes corresponds to the correlation coefficient between these two genes. Typically, a similarity cut-off is used to decide whether two expression patterns are similar or not and to draw or not the edges between nodes. Alternatively, weighted edges can be used to record the value of the statistic. Gene expression networks are useful to hypothesize gene function. But how many microarray experiments do we need in order to have a robust co-expression network? This issue was recently addressed for Arabidopsis (Aoki *et al.*, 2007). Aoki *et al*. (2007) found that the density of networks derived from microarray experiments from Arabidopsis reached equilibrium when more than 100 arrays were used. This result suggests that 100 experiments are sufficient to build a robust co-expression network. It remains to be determined whether this is insensitive to the experimental factors tested in the microarray experiments utilized. Is the Arabidopsis co-expression network different from the network of another model organism? Bergmann *et al.* (2004) compared the global features of co-expression networks from *Saccharomyces cerevisiae*, *C. elegans*, *E. coli*, *A. thaliana*, *D. melanogaster* and *Homo sapiens*. The co-expression networks built for all these organisms showed a power law degree distribution indicative of scale-free networks. In addition, all these networks exhibited high modularity. The modules were defined as sets of co-expressed genes that share a common function. A number of these core modules were conserved throughout evolution. Examples of core modules identified were: rRNA processing machinery, heat shock response and the proteosome. More recently, it was reported that hub genes from co-expression networks in *A. thaliana* tend to be single-copy genes (Wei *et al.*, 2006). For the positive co-expression network studied, 65% of the hub genes (having 20 or more connections) were found to be single-copy. On the other hand, only 37% of genes having fewer than 20 connections were single-copy genes. Co-expression networks are powerful as they provide a mechanism to relate genes without known function to known genes and help build testable hypothesis.

1.3.2.3 The interactome

The interactome is defined as the complete physical interaction map between the proteins in an organism. Molecular machines (e.g. ribosome, proteasome,

DNA or RNA polymerases), structural protein complexes, signalling cascades are just a few of the many examples of situations in which these interactions are relevant. Thus, defining protein complexes is critical to understanding virtually all aspects of cell function. Interactomes can be generated by various experimental approaches. The most widely used approaches are high-throughput yeast 2-hybrid binding assays and co-affinity purification followed by MS. In Arabidopsis, Van Leene *et al*. (2007) recently reported the development and application of a high-throughput tandem affinity purification (TAP)/MS platform for cell suspension cultures to analyze cell cycle-related protein complexes. Key for their methodology was the fast generation of transgenic cultures overproducing tagged fusion proteins, TAP adapted for plant cells. Using this strategy, they were able to validate 14 interactions that were previously reported and identified 28 new molecular associations. The building of interactomes (protein–protein interaction network) has been useful in the study of many model organisms (Schwikowski *et al.*, 2000; Walhout *et al.*, 2000; Giot *et al.*, 2003) as well as humans (Lehner and Fraser, 2004). Due to the lack of an experimentally determined interactome network for Arabidopsis, several strategies have been tried to predict interactomes. For example, Yu *et al.* proposed transferring annotations between genomes (Yuan *et al.*, 2003). Based on sequence similarity, they could transfer protein–protein and protein–DNA interactions from model organisms (whose interactome was partially available) to Arabidopsis. A similar approach was utilized to predict interactomes for Arabidopsis based on orthologs in *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens* (Geisler-Lee *et al.*, 2007; Gutierrez *et al.*, 2007). Predictions are useful for systems research but far from satisfactory. It is known that experimental approaches for determining protein–protein interactions are not perfect and high-throughput data sets contain a high rate of false positives. Even if they were completely error free, transferring this information to another organism will certainly introduce errors. In addition, it has been reported that at least 14% of Arabidopsis proteins are likely to be found just in the plant lineage (Gutierrez *et al.*, 2004) and homology-based approaches will not predict interactions for these proteins. Small-scale interactome efforts have been reported. For example, the protein interaction map of the MADS Box family of transcription factors (de Folter *et al.*, 2005). It is not necessary to explain this acronym as it is well known in plant biology. Explaining its origin in this context confuses the discussion about protein interactions. A promising new strategy is the use of protein arrays (Popescu *et al.*, 2007). A high confidence Arabidopsis protein interactome defined experimentally is a must in order to advance plant systems biology.

1.3.2.4 Regulatory interactions

The transcriptional circuits underlying genetics programmes are characterized by interactions between transcription factors (TFs) and *cis*-regulatory

regions in the promoter regions of target genes. Large-scale efforts to characterize regulatory interactions in Arabidopsis have not been performed. Regulatory interaction predictions have been attempted and at least in a few cases the predictions have been experimentally validated (e.g. Gutierrez *et al.*, 2008). To build regulatory networks, a very useful tool for the plant community is the Arabidopsis Gene Regulatory Information Server (AGRIS; http://arabidopsis.med.ohio-state.edu/). AGRIS contains all Arabidopsis promoter sequences, TFs, and their target genes and functions (Palaniswamy *et al.*, 2006). AGRIS currently houses three linked databases: At*cis*DB (*Arabidopsis thaliana cis*-regulatory database), AtTFDB (*Arabidopsis thaliana* TF database) and AtRegNet (*Arabidopsis thaliana* regulatory network). At*cis*DB currently contains 25 516 promoter sequences of annotated Arabidopsis genes with a description of putative *cis*-regulatory elements (Molina and Grotewold, 2005). AtTFDB contains information on approximately 1770 TFs. These TFs are grouped into 50 families, based on the presence of conserved DNA-binding domains.

1.3.2.5 The phenome: large-scale phenotypic analysis of mutants

Classic genetic approaches aim to understanding gene function by the study of the phenotypes caused by mutations of the gene. Generally speaking, mutations in genes that result in similar phenotypes often imply that the genes function in the same process. This rationale has led to the development of large-scale initiatives to characterize phenotypes caused by the systematic inactivation of genes in the genome of model organisms. Integrating this phenome data with other genome-wide data has led to important advances, for example in the understanding of animal developmental processes (Gunsalus *et al.*, 2005). Currently, the Arabidopsis community benefits from large populations of T-DNA and transposon insertional mutants (Scholl *et al.*, 2000). The location of the T-DNA elements in the genome has been determined using PCR-based amplification of flanking sequences and sequencing (Bevan and Walsh, 2005). To date, there are ∼320 000 sequenced mutant lines in the reference Arabidopsis Columbia genome. The *Arabidopsis* Biological Resource Centre and the Nottingham *Arabidopsis* Stock Centre (NASC) contain seed banks where the global community can request individual knockout lines for functional studies. Many protein-coding genes (around 1600) do not possess insertions in exonic or intronic sequences. According to a simulation (Bevan and Walsh, 2005), if the number of random insertions is doubled, a raise from 94% to 98% is expected in the number coding genes with insertions. RNA interference (RNAi) and targeting induced local lesions in genomes (TILLING) are alternatives to isolate mutants in individual genes. Recently, Shinozaki and co-workers (Kuromori *et al.*, 2006) carried out a pilot study to characterize the Arabidopsis phenome. They constructed 18 000 transposon-insertion lines of Arabidopsis. They selected approximately 4000 transposon-insertional lines which had a Ds transposon inserted in the coding region of genes and visually analyzed their phenotypes. About 140 lines were found

with reproducible and distinguishable phenotypes. They established 8 categories and 43 secondary categories for the description of those mutant lines. The images of the morphological phenotypes observed have been entered into a searchable database (http://rarge.gsc.riken.jp/phenome/) (Kuromori *et al.*, 2006). Although this work is based mainly in observation and description of phenotypes, the information obtained is very important for the functional characterization of new genes. In addition, phenotypic information can be used together with transcriptome and interactome data sets for the construction of powerful functional networks (Gunsalus *et al.*, 2005). A promising approximation has been reported that uses 3D (three dimensional) laser scanners for the collection of morphological information (Kaminuma *et al.*, 2005). Phenome efforts like the ones described will undoubtedly play a key role for understanding gene function in Arabidopsis.

1.3.2.6 Genetic interactions

Some single mutations or gene deletions do not have an obvious altered phenotype, but the phenotype is uncovered when two of those are combined. These epistatic relationships are used to define interactions (edges) between genes (nodes) and build genetic interaction networks. Building genetic interaction networks is important because most heritable traits are affected by interactions between multiple genes. Most of the studies involving genetic interactions have focused on synthetic lethal interactions because they are much easier to score and detect as compared to other subtle changes in phenotype. A synthetic lethal interaction between two genes is defined when the survival of the combined mutations is less than the product of the survival of the two single mutations (Lehner, 2007). If these interactions are carried out in a systematic way, it is possible to construct genetic interaction networks. Nowadays, *S. cerevisiae* is the model organism in which most of the genetic interactions have been mapped (Ooi *et al.*, 2006; Ming *et al.*, 2008). There are two main experimental strategies in yeast: synthetic genetic arrays (SGA) and synthetic lethal analysis of microarrays (SLAM). In the first approach, the starting point is a yeast strain carrying a query mutation that is systematically mated to a library of viable deletion strains and subsequently evaluated phenotypically in parallel. On the other hand, in the second approach, double mutants are constructed (integrative transformation) and assayed as a single pool. Later, microarrays are used to detect the relative growth rates making use of a barcode that each double mutant has in its genome. In addition, genetic interaction screenings (Baugh *et al.*, 2005) and a genetic interaction network have also been generated for *C. elegans*, taking advantage of the bacterial RNAi feeding system (Lehner *et al.*, 2006). It is also possible, although less practical, to do this in mammalian and fly cell cultures (Wheeler *et al.*, 2004). Genetic interaction screenings have also been described in Arabidopsis (Parker *et al.*, 2000), although it is more complicated and laborious to develop a systematic test of all the possible genetic interactions in plants. Genetic interactions can be used to understand

gene function. For instance, the higher the number of shared genetic interactions partners between two genes, the higher the probability that both genes interact physically and share a biological function (Ming *et al.*, 2008). However, a genetic interaction can have diverse explanations and a systemic interpretation of those is lacking. Although the transfer of genetic interaction between species using orthology relationships is a possibility in order to obtain a genetic interaction network in Arabidopsis, preliminary evidence suggests that this step may not be so direct. In contrast to gene essentiality, genetic interactions seem not to be so widely conserved across species (Lehner, 2007). Another strategy consists in the prediction of genetic interactions. Zhong and Sternberg integrated information about anatomical expression patterns, phenotypes, functional annotations, microarray co-expression and protein interactions to predict *C. elegans* genetic interactions (Zhong and Sternberg, 2006). With their model, they could identify 12 genetic interaction partners of the *let-60*/Ras gene and 2 of the *itr-1* gene. It would be a challenge worth pursuing to construct an analogous genetic interaction network for Arabidopsis.

1.4 Strategies for genomic data integration

Here is an interesting question: What do Belgium and systems biology have in common? Well, definitely not their 500 varieties of beer. Interestingly, they both share a motto: '*L'union fait la force*' ('Strength through unity'). Our universe is a noisy place. By noise, we are not just talking about an unintended sound that reaches our ears, but it can also be something unpleasant, unexpected and undesired. The word 'noise' can be traced back to the Latin word *nausea* (feeling of sickness), and there is no doubt that some scientists feel that way when facing it. Noise is a random and generally persistent disturbance that obscures or reduces the clarity of a signal or the result of an experiment. Nowadays, it is common to read in scientific literature that one of the biggest problems when constructing genome-scale models of biological systems is that the underlying data is noisy. This is an inherent property of the high-throughput techniques used for the acquisition of massive data, interactions between proteins, etc. As a consequence of this noise, data sets contain false positives. This occurs, for example due to self-activators in the yeast 2-hybrid technique, and much effort has been invested in solving this problem (Margulies *et al.*, 2005). False-positive interactions can also be 'biological'. This means that the physical interaction between proteins A and B can be true, but maybe this interaction does not make sense in a cellular context, because protein A and B are expressed in different tissues or different cell types or different organelles or even in different periods of times in the same cell during development. That is why special caution must be taken, for example when trying to predict the function of a gene based on a single global data set. When we say 'Strength through unity', we mean that it is possible to increase

the confidence of the edges in a network by integrating different sources of information. The rationale behind this is that it is unlikely that false-positive interactions are reproducible in data sets acquired through different experimental or *in silico* approaches. Data integration is also important as we try to build comprehensive models of a system. The sheer amount of information published daily for any of the model organisms makes it almost impossible to manually store, classify and integrate the data for systems modelling. Bioinformatics tools are key for systems research. Ideally, a high quality integrated model of the organism of choice will speed up the discovery process by allowing experimental scientists to spend time addressing important biological questions in the laboratory instead of navigating through an ocean of genomic information on their computers.

1.4.1 Integration strategies

There are generally four methods that are used as a solution for data integration: (1) Hypertext Navigation, (2) Data Warehouse, (3) Unmediated MultiDB queries and (4) Federated Databases (Karp, 1996). Hypertext Navigation allows the user to query only one database but the results often contain hyperlinks to the equivalent entry in another database. This method is more common for websites that are based on information retrieval. A Data Warehouse retrieves data from multiple resources, translates the formats and puts them in one database. This allows for much faster and more complex queries taking advantage of all the data loaded in the one database; however, translating database formats from one to another is a challenge in itself. The major drawback of this method is that it would be very difficult to keep up with all the new resources becoming available and keeping it all updated. The Unmediated MultiDB Queries allow the databases to remain separate but the query itself extends across all of the databases. The Federated Database is a combination of Data Warehouse and Unmediated MultiDB Queries; it allows the databases to be separate but it contains a federated schema, which dynamically translates to queries in the individual schemas. For a detailed discussion about these methods see Peter Karp's (1996) review.

In addition to simply providing data, many websites also provide applications that analyze the data for the user. Integrating these services has the same obstacles as integrating data, but instead of only querying the data you are also launching an application. Two recent efforts, iSYS (Siepel *et al*., 2001) and BioMOBY (Wilkinson *et al*., 2003), provide infrastructure that allows integration of applications. Both are designed to be 'plug and play' and generic to allow for compatibility with practically any application or data.

1.4.2 Case studies on data integration

A good example of an integration strategy (Hwang *et al*., 2005) is the data integration methodology called 'Pointillist'. This methodology handles multiple

data types from technologies with different noise characteristics. Pointillist integrated 18 data sets containing information about the galactose utilization in yeast. The data included global changes in mRNA and protein abundance, genome-wide protein–DNA interaction data, database information and computational predictions. Even though, the galactose biochemical pathway has been studied for more than 40 years and is one of the best understood in eukaryotic systems, they managed to predict and corroborate experimentally a new relationship between the fructose and the galactose pathways.

Kelley *et al*. (2003) proposed an integration strategy to face the problem of yeast genetic interaction analysis. There are many types of genetics interactions. For example, synthetic lethal interactions in which mutations in two non-essential genes result in a lethal phenotype. Genetic interactions are useful to study pathway organization. Some high-throughput methods for discovering genetic interactions have been developed such as SGA or SLAM. Although these processes can be mostly automated, the bottleneck is the interpretation of the functional significance of each interaction found. In order to find a solution to this problem, Kelley *et al*. assembled a genetic interaction network and a physical interaction network. The former was generated by SGA large-scale screen data and interactions culled from Munich Information Center for Protein Sequences, and the latter was connected by interactions of three types: protein–protein interactions (A and B interact physically), protein–DNA (A binds to regulatory sequence of B) and shared-reaction metabolic relationships (enzymes A and B have a substrate in common). They defined three different interpretations of genetics interactions: (1) between-pathways in which the genetic interaction bridges genes operating in two pathways with redundant or complementary functions, (2) within-pathways in which genetic interactions occur between proteins subunits within a pathway and (3) indirect effects in which the lethal phenotype involves many diverse pathways. The researchers could uncover mechanisms behind many of the observed genetics interactions that were preferentially between-pathway and they could predict new functions for 343 yeast proteins based on the models generated (not validated experimentally).

Another practical example of an integration strategy is the recent work of Hirai *et al*. (2004) in plants in which they integrated transcriptome and metabolomic data. They focused on the study of glucosinolates (GSLs) which are produced by Brassicaceae family. GSLs apparently provide anticarcinogenic, antioxidative and antimicrobial activity. Despite the high biotechnological importance of these compounds, before this research, there was no previous report on the genes regulating methionine-derived aliphatic GSL biosynthesis. The authors used an integrated strategy based on transcriptome co-expression analysis for public data sets and their own data sets together with metabolic profiling. They first generated a data set of condition independent co-expression profiles by calculating the Pearson correlation between all paired combinations of 22 263 Arabidopsis genes. Then they visualized the co-expression network including only pairs of genes whose correlation

coefficient was >0.65. They found that genes involved in aliphatic GSL biosynthesis were clustered in a discrete module together with two uncharacterized genes that coded for TFs: Myb28 and Myb29. Further experiments revealed that Myb28 is a positive regulator for basal-level production of aliphatic GSL and Myb29 presumably plays an accessory function integrating JA signalling and GSL biosynthesis. Interestingly, they could induce the biosynthesis of large amounts of GSLs by overexpressing Myb28 in Arabidopsis-cultured suspension cells that do not normally synthesize them. In addition, they predicted many of the genes involved in the pathway by examining the obtained co-expression network. This clearly shows that it is possible to find genes involved in particular plant processes using integrative systems biology strategies.

1.4.3 Software tools for systems biology

The large-scale genomic data available to biologists today require the use of a variety of software tools to process and analyze the data. A systems approach to understanding biology involves an iterative process of data integration, building a model, designing experiments to support the model and generating new hypothesis based on new data (Gutierrez *et al.*, 2005). There are several tools that are specific for each stage, however, few of them span across several stages and even fewer allow for the iterative analysis. In Chapter 5, several systems biology tools are discussed in detail: Sungear (Poultney *et al.*, 2007), Genevestigator (Zimmermann *et al.*, 2004), MapMan (Thimm *et al.*, 2004) and Cytoscape (Shannon *et al.*, 2003). In this section, we will highlight some additional tools that help biologists perform systems-level analysis. Please note that in no way are we providing a comprehensive list of all software tools available. In addition, this is an active and dynamic field with new tools constantly being generated.

Several environments have been developed in the past years that permit data integration and modelling (Endy and Brent, 2001). Such software allows detailed mathematical representation of cellular processes (e.g. Gepasi (Mendes, 1997) and Virtual Cell (Loew and Schaff, 2001)), as well as qualitative representations of cellular components and their interactions (e.g. Cytoscape (Shannon *et al.*, 2003) and Osprey (Breitkreutz *et al.*, 2003)). The Arabidopsis eFP browser (Winter *et al.*, 2007) provides a graphical representation of gene expression plotted on drawings of the different tissue and developmental stages based on data from AtGenExpress (Schmid *et al.*, 2005). Generally, quantitative models are directed towards specific cellular processes of interest and are built by representing existing literature as a system of mathematical equations. Quantitative models are powerful because they describe a system in detail (Endy and Brent, 2001). Unfortunately, they require a very detailed understanding of the system under scrutiny. This information is not readily obtained for every component in a biological process, and less so in every model organism. In fact, there are still many gaps in our qualitative

understanding of biological systems. For example, most of the genes in Arabidopsis have not yet been experimentally characterized. The crucial first stage in building quantitative models is constructing a map of the interaction network to be analyzed (Schwender *et al.*, 2003).

Because genes encoding proteins that participate in the same pathway or are part of the same protein complex are often co-regulated in their expression patterns, clustering techniques (e.g. Eisen *et al.*, 1998) are commonly applied to genome-wide expression measurements to hypothesize the role of uncharacterized genes. In addition, in experiments where the biologist is comparing a treated versus a control sample, statistical methods such as Rank Products (Breitling *et al.*, 2004) can be used to determine the genes that are differentially expressed between the experimental conditions. To learn the biological significance of such lists of genes (e.g. co-regulated genes or differentially expressed genes), a common next step is to evaluate the frequency of occurrence of functional attributes using structured functional annotations such as the gene ontology (GO) (Ashburner *et al.*, 2000). Several software packages to automate this type of analysis now exist (e.g. Onto-Express (Khatri *et al.*, 2002), GoMiner (Zeeberg *et al.*, 2003), GOSurfer (Zhong *et al.*, 2004), FatiGO (Al-Shahrour *et al.*, 2004)). While advanced data analysis tools for exploiting genomic data are rapidly emerging, one of the drawbacks of current software tools is that they are highly specialized to each platform. This translates in the need to travel through several different web services or to use various stand-alone applications to be able to analyze the large data sets characteristic of genomic research, a cumbersome and inefficient process not amenable to iterative *in silico* exploration and experimentation.

Due to the increased use of networks to represent interaction data, viewers and graph drawing software such as Pajek (Batagelj and Mrvar, 1998), daVinci (Wilkinson and Links, 2002), Graphlet (Wilkinson and Links, 2002) and Graphviz (Wilkinson and Links, 2002), developed for general purpose graph drawing and visualization, are now used to organize and display biological data as graphs. Other tools such as MintViewer (Zanzoni *et al.*, 2002), GeneInfoViz (Zhou and Cui, 2004), PIMRider (Hybrigenics, 2004), GenMAPP (Dahlquist *et al.*, 2002), MAPPFinder (Doniger *et al.*, 2003), TopNet (Yu *et al.*, 2004), MAPMAN (Thimm *et al.*, 2004) and PaVESy (Ludemann *et al.*, 2004), each developed in the context of specific applications, allow certain data types to be displayed and provide some data analysis tools. Other advanced software environments such as Cytoscape (Shannon *et al.*, 2003), Osprey (Breitkreutz *et al.*, 2003), VisANT (Hu *et al.*, 2004), PathwayAssist (Wilkinson and Links, 2002) and PathBlazer (Wilkinson and Links, 2002) include various methods to layout the network graphs, support connection to external databases and include more sophisticated data analysis tools (Shannon *et al.*, 2003).

In many cases, biologists need to use many if not all of the tools mentioned and data management becomes an issue. Several institutions, for example MetNet (Wurtele *et al.*, 2007) and The Bio-Array Resource for Arabidopsis

Functional Genomics who created Arabidopsis eFP browser (Winter *et al.*, 2007), have created entire suites of wonderful applications to analyze genomic data but there is absolutely no interaction between them. VirtualPlant is an attempt to provide a platform for systems biology where biologists can maintain their genomic data and execute different data visualization and analysis tools. VirtualPlant's GeneCart supports the iterative nature of systems biology analysis and allows users to save results from one analysis and feed it into another. Applications available on VirtualPlant range from simple list functions, microarray data analysis to gene network visualization and analysis.

1.5 Systems biology in plant research

The development of systems biology approaches and tools for Arabidopsis has enabled the generation of testable biological hypotheses derived from the integration and analysis of genomic data within a network context. In this section, we briefly discuss proof of principle studies where a systems approach has uncovered testable hypotheses for regulatory networks in Arabidopsis.

1.5.1 Qualitative network models and genome-wide expression data define carbon/nitrogen molecular machines in Arabidopsis

In plants, there is ample evidence that C-signals interact with nitrogen (N) status (Gutierrez *et al.*, 2007) to control genes in metabolic pathways, including N-assimilation and amino acid synthesis (e.g. Gutierrez *et al.*, 2007; Palenchar *et al.*, 2004) as well as broad kinds of developmental mechanisms such as germination (Alboresi *et al.*, 2005), shoot growth (Walch-Liu *et al.*, 2000, 2005; Rahayu *et al.*, 2005; Krouk *et al.*, 2006), root growth (Forde, 2002) or flowering (Raper *et al.*, 1988; Rideout *et al.*, 1992). While these studies confirm the existence of a complex CN-responsive gene regulatory network in plants, the possible mechanisms for CN sensing and signalling remain unknown. While transcriptome studies conducted in the pre-systems biology era, confirmed the existence of genome-wide CN-responses (Palenchar *et al.*, 2004; Price *et al.*, 2004), those studies were not able to model responses or reveal underlying mechanisms. For example, at least four general mechanisms for CN sensing can be proposed: (1) N-response independent of C, (2) C-response independent of N, (3) C- and N-interaction or (4) a unified CN-response (Gutierrez *et al.*, 2007). To support or reject these models for C/N sensing, plants subjected to a systematic matrix of C and N treatments were analyzed using a systems approach (Gutierrez *et al.*, 2007). Hierarchical cluster analysis ($\geq$0.5 correlation) of transcriptome data generated in this study uncovered many gene clusters whose average expression patterns showed statistically significant CN interactions (analysis of variance, $p < 0.01$), suggesting that

C and N interaction (Model 3) is a prominent mode of regulation by N in Arabidopsis roots. To gain a global but detailed view of how transcriptional responses to carbon and N nutrient treatments affects plants as a system, an Arabidopsis multi-network was created and used to query gene clusters whose members respond to C/N according to Models 1–3, for highly connected sub-networks defined using a graph clustering algorithm developed by (Ferro *et al.*, 2003). This analysis revealed a set of 'molecular machines' comprised of highly connected genes involved in metabolic, cellular or signalling pathways, whose expression is regulated by C and N metabolites. One such CN-responsive sub-network is involved in responses to auxin, as it contains 13 genes in the auxin response pathway (including the auxin receptor), 5 auxin efflux carriers and 2 auxin transport proteins. Validation using time-course studies suggest that the phytohormone auxin acts as a regulator of plant growth in response to C and/or N availability. In addition, the CN-responsive gene network was shown to contain a significant proportion of regulatory proteins including: 299 TFs and 27 genes that are known targets of miRNAs. This latter result implicated a new role for miRNAs in post-transcriptional regulation of gene expression by CN metabolite signals in plants.

1.5.2 A systems approach identifies an organic nitrogen-responsive gene network regulated by the master clock gene CCA1

A second systems biology study, addressed the mechanisms by which distinct forms of N are sensed as signals for N status (Gutierrez *et al.*, 2008). There is growing evidence that the N 'input' (nitrate) and 'output' (Glu/Gln) of the N-assimilatory pathway serve not only as N-metabolites, but also as N-signals that are sensed and transduced, to control genes regulating plant metabolism and development (for NO_3^- see Forde, 2002; Sollars *et al.*, 2003; Scheible *et al.*, 2004; for Glu/Gln see Oliveira and Coruzzi, 1999; Rawat *et al.*, 1999). Studies with nitrate reductase nulls, unable to reduce nitrate, have shown that nitrate can serve as a 'metabolic N-signal', to regulate genes in N-uptake/assimilatory pathway and to control root development in response to nitrate availability in the soil (Forde, 2002). There is also ample, though less direct evidence, that the assimilated forms of organic-N such as glutamate (Glu) or glutamine (Gln) serve as signals for levels of organic-N or C:N status, to repress further N-assimilation, or lateral root growth, when organic-N stores are replete (Oliveira and Coruzzi, 1999; Rawat *et al.*, 1999). To identify the regulatory networks regulated by these N-signals, transcriptome analysis identified a set of 834 genes regulated in response to inorganic-N and/or organic-N treatments. To identify the global processes regulated by these N-signals, the Arabidopsis multi-network (Gutierrez *et al.*, 2007) was queried for network connectivity between these N-regulated genes. Of the 834 N-regulated genes, 368 were connected by a variety of edges including

metabolic and regulatory connections. To identify potential 'master' regulators of this N-responsive gene sub-network, N-regulated TFs were ranked based on their number of predicted regulatory connections. At the top of the list, with 47 connections to targets in the N-regulated gene network, is a Myb family TF which is the central clock control gene *CCA1* (Daniel *et al.*, 2004). Exploration of the network 'neighbourhood' surrounding the *CCA1* hub predicts that *CCA1* is an important regulator of an N-assimilation gene network. Specifically, the model predicts that overexpression of *CCA1* would induce the expression of the *GLN1.3* gene and repress the expression of *ASN1* and *GDH1*, a result which was validated using 35S::*CCA1* lines (*CCA1-ox*) (Sollars *et al.*, 2003). Chromatin-IP assays using CCA1 antibodies confirmed binding of CCA1 to the promoter regions of *GLN1.3*, *GDH1* and *bZIP1* and ASN1. These results indicate that the circadian clock regulates N-assimilation (Farre *et al.*, 2005) in response to N availability by transcriptional control by *CCA1* which in turn targets genes central to the N-assimilatory pathway in plants. Specifically, in this model Glu-repression of CCA1 leads to downregulation of GLN1.3, and upregulation of ASN1. This regulation leads to the conversion of metabolically reactive Gln to inert Asn, used for N-storage, when levels of Glu/Gln are abundant.

CCA1 is a key component involved in a negative feedback loop at the centre of the circadian clock (Locke *et al.*, 2005; McClung, 2006). The finding that *CCA1* mRNA levels are regulated by organic N-sources, suggest that N-signals act as an input to the circadian clock. To test this new hypothesis derived from the systems approach, Arabidopsis seedlings were subject to pulses of inorganic-N or organic-N at intervals spanning a circadian cycle and determined the effects on the phase of the oscillation in *CCA1::LUC* expression. Each N-treatment resulted in subtle (2 h) but stable phase shifts in CCA1::LUC expression, indicating that N-status serves as an input to the circadian clock. These alterations in phase response curves are consistent with N-signals mediating a weak (type 1) clock resetting observed for other metabolic signals (Bunning and Moser, 1973; Bollig *et al.*, 1978; Kondo, 1983). The emerging view of the circadian clock as a key integrator of metabolic and physiologic processes (Farre *et al.*, 2005; McClung, 2006) is that it receives input not only from environmental stimuli but also from metabolic pathways, many of which are subject to circadian regulation. Thus, the clock is known to regulate genes in N-assimilation (Pilgrim *et al.*, 1993; Farre *et al.*, 2005), this systems study suggests a new hypothesis that N in turn feeds back to the circadian clock, at least in part through the N-regulation of *CCA1* expression.

1.5.3 Cell-specific nitrogen responses mediate developmental plasticity

In this application of systems biology to Arabidopsis development, cell-specific transcriptome analysis was analyzed using a systems approach to ask how N-nutrient signals coordinate expression of gene networks in specific

cell types of roots (Gifford *et al.*, 2008). G*reen fluorescent protein* (GFP)-marked transgenic lines spanning the cell types of the root were transiently treated (2 h) with nitrate (5 mM) or control KCl treatments, and GFP expressing cells were sorted and isolated using fluorescence-activated cell sorting (Gifford *et al.*, 2008). Five thousand three hundred ninety-six N-regulated transcripts were identified in which 87% were significantly N-regulated in at least one, but not *all* cell-types profiled. Only 771 transcripts responded to N-treatment across *all* cell-types examined. Thus, cell sorting greatly increased the sensitivity to detect transcriptional regulation in specific cell types by N treatment that were largely hidden in studies of whole roots from nitrate-treated plants (Sollars *et al.*, 2003; Scheible *et al.*, 2004; Gutierrez *et al.*, 2007). This cell-specific N-response transcriptome data also revealed new mechanisms by which N regulates developmental processes in roots. One example is the cell-specific regulation of a transcriptional circuit derived from a systems analysis of the data and validated to mediate lateral root outgrowth in response to N. Importantly, the N-regulated response of the miR167 targets (Wu *et al.*, 2006) was only detectable with cell-specific resolution. Validation studies showed that miR167 is specifically expressed and N-regulated in the pericycle (PmiR:GUS line and mature RNA) and that N-regulation of its target ARF8 is abrogated in a mutant in which the miR167 binding site of ARF8 is mutated (Wu *et al.*, 2006). These findings fill a gap in our knowledge of how multi-cellular organisms cope with N-responses at the cellular level.

1.5.4 Quantitative models of defined molecular processes

One of the best examples of systems biology in action in plants was demonstrated by James Locke and colleagues (reviewed in Ueda, 2006) where they studied the circadian clock in Arabidopsis. The first molecular component of the Arabidopsis central oscillator identified was the feedback loop between the genes *LHY* (*late elongated hypocotyl*), *CCA1* (*circadian clock associated*) and *TOC1* (*timing of cab expression*) (for a historical perspective on the clock see (McClung, 2006). *LHY* and *CCA1* are partially redundant genes that are activated by light and their products repress *TOC1.* Locke and colleagues analyzed experimental data from various circadian studies (Matsushika *et al.*, 2000; Mizoguchi *et al.*, 2002; Kim *et al.*, 2003) including *cca1/lhy* double mutant (Locke *et al.* 2005), and determined that the feedback loop was not sufficient to explain many of the circadian rhythms observed. For example, the feedback loop does not explain why there is a 12-h delay in *LHY/CCA1* after activation of *TOC1*. The feedback model also does not explain why TOC1 levels drop much earlier (dusk) than *LHY/CCA1* activation. Locke and colleagues proposed a new model with two loops and two new factors, X and Y, to account for the experimental observations. In the new model, factor Y is light induced and activates *TOC1* and factor X activates *LHY/CCA1* and itself is activated by TOC1. They next looked at experimental data to identify candidate genes to carry out the X or Y functions. The data analyzed suggested that *GI*

(*GIGANTEA*) could play the role of Y. X remains to be identified. Circadian study of *gi/lhy/cca1* showed a decreased level of expression of TOC1 (Locke *et al.*, 2006). Similar studies allowed Locke and colleagues to create another model with three loops to incorporate other genes *Pseudo Response Regulatory* (*PRR7*) and *PRR9* that have also circadian rhythmic expression patterns (Farre *et al.*, 2005). The new loop is a feedback loop between *PRR7/PRR9* and *LHY/CCA1.* The work by Locke and colleagues contains an iterative nature of analyzing data, building a model, validating a model and then developing new hypotheses, which is a trademark of systems biology.

1.5.5 Modelling the behaviour and tissue and organs

One of the beautiful aspects of development is the creation of shapes and patterns. How do single cells develop into multi-cellular organisms? This is a key question in plant biology that can benefit from systems approaches. Several examples of tissue, organ or plant modelling have been reported. For example, the growth dynamics underlying the development of *Antirrhinum* petals were modelled (Prusinkiewicz and Rolland-Lagan, 2006). They determined three types of parameters: the rate of increase in size (growth rate), the degree to which growth occurs preferentially in any direction (anisotropy) and the orientation angle of the main direction of growth (direction). The model allowed them to conclude that the key parameter determining the petal asymmetry of the *Antirrhinum* is the direction of growth rather than the regional differences in growth rate. In a different study, Mundermann *et al*. (2005) systematically took thousands of measurements, such as length, width and other parameters that describe shapes, in frequent time intervals and created a model representing plant development. In their model the plant is broken into four basic parts: apices, internodes, leaves and flowers, and each flower is further broken into pedicel, carpel, sepals, petals and stamens. The measurements are taken for each of the individual parts and the simulation of the model assembles all the components into a 3D growing plant. The model is written in L-system-based modelling language L +C (Karwowski and Prusinkiewicz, 2003). The validation of the model was the generation of a simulated image that looked very similar to a wild-type Arabidopsis. A similar model is also available for rice (Watanabe *et al.*, 2005). These plant models can serve as references for future studies in plant development and morphology.

There are several different computational techniques that have been developed for processing of data, such as images and experimental measurements, and construction of simulation models (reviewed in Prusinkiewicz and Rolland-Lagan, 2006). One such model is reaction-diffuse where the pattern forming substance is assumed to diffuse across cells (Turing, 1952; Gierer and Meinhardt, 1972). Meyerowitz and colleagues used this model to study the shape and size of the shoot apical meristem (SAM) and expression domains of the *WUSCHEL* (*WUS*) gene (Jonsson *et al.*, 2005). *WUS* is known to be

involved in SAM development and is known to induce the ligand *CLAVATA3* (*CLV3*) which binds to the *CLAVATA1* (*CLV1*) receptor kinase. *CLV1* then activates a signalling cascade that represses *WUS* (Sharma *et al.*, 2003). In this study, the authors used in vivo confocal microscopy and image-processing algorithms to obtain quantitative measurements of gene expression. The measurements were subsequently used as input for the reaction-diffuse model, which allowed them to predict the outcome of laser cell ablation experiments (Reinhardt *et al.*, 2003) and removal of *CLV3* signal (Fletcher *et al.*, 1999). The model correctly simulated creation of two new *WUS* domains after the ablation of the central cells from the SAM and the expansion of *WUS* expression with the removal of *CLV3*. (Note: For more on systems biology applications to plant development and evolution, see Chapters 11 and 12 of this volume.)

1.6 Conclusion

Hodgkin and Huxley (1952) described a model that explained in mathematical terms the ionic mechanisms underlying the initiation and propagation of action potentials in the squid giant axon. They could even predict with some success how a neuron works. Both researchers integrated their chemical, biological, physical and mathematical knowledge in order to understand this phenomenon, and they can be fairly called pioneers of systems biology. One of the main problems today when integrating diverse disciplines like biology, physics or computer science is actually the integration of the people who work in those fields. These scientists have, for example diverse backgrounds and cultures for publishing, communicating their results and doing research. In addition, their 'languages' can differ substantially, making communication a difficult task. This is another challenge systems biology must overcome in the coming years, and new educational strategies are definitely needed in universities. Wilson (2000) wrote: 'The love of complexity without reductionism makes art; the love of complexity with reductionism makes science.' This is clearly a challenge for a new generation of scientists and students who must coherently and elegantly integrate the vast amount of biological information available in order not only to make and validate predictions, but to understand the underlying principles of living matter. Perhaps art and science are different worlds, but they have something in common: the creative process. You may call it inspiration or scientific induction, but those will only favour the prepared minds. Are you prepared for systems biology?

Acknowledgements

We regret that many interesting articles could not be cited due to space limitations. Research in RAG's laboratory in this area is supported by grants from FONDECYT (1060457), ICGEB (CRPCHI0501) and ICM-MIDEPLAN (MN-PFG P06–009-F). Research in GC's laboratory in this area

is supported by grants from NIH (GM032877), NSF (IOB 0519985), DOE (DEFG02–92ER20071). Collaborative work by RAG's and GC's labs on systems biology is supported by NSF (DBI-0445666).

References

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., *et al.* (1991) Complementary, D.N.A sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651–1656.

Aharoni, A., Ric de Vos, C.H., Verhoeven, H.A., Maliepaard, C.A., Kruppa, G., Bino, R., *et al.* (2002) Nontargeted metabolome analysis by use of Fourier Transform Ion Cyclotron Mass Spectrometry. *Omics* **6**, 217–234.

Albert, R., Jeong, H. and Barabasi, A.L. (1999) Diameter of the World-Wide Web. *Nature* **401**, 130–131.

Albert, R., Jeong, H. and Barabasi A.L. (2000) Error and attack tolerance of complex networks. *Nature* **406**, 378–382.

Alboresi, A., Gestin, C., Leydecker, M.T., Bedu, M., Meyer, C. and Truong, H.N. (2005) Nitrate, a signal relieving seed dormancy in Arabidopsis. *Plant Cell Environ* **28**, 500–512.

Alonso, J.M. and Ecker, J.R. (2006) Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in Arabidopsis. *Nat Rev Genet* **7**, 524–536.

Al-Shahrour, F., Diaz-Uriarte, R. and Dopazo, J. (2004) FatiGO: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics* **20**, 578–580.

Aoki, K., Ogata, Y. and Shibata, D. (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant Cell Physiol* **48**, 381–390.

Aoki, K.F. and Kanehisa, M. (2005) Using the KEGG database resource. *Curr Protoc Bioinformatics* Chapter 1, Unit 1 12.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29.

Baerenfaller, K., Grossmann, J., Grobei, M.A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., *et al.* (2008) Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. *Science* **320**, 938–941.

Barabasi, A.L. and Albert, R. (1999) Emergence of scaling in random networks. *Science* **286**, 509–512.

Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101–113.

Batagelj, V. and Mrvar, A. (1998) Pajek – program for large network analysis. *Connections* **21**, 47–57.

Baugh, L.R., Wen, J.C., Hill, A.A., Slonim, D.K., Brown, E.L. and Hunter, C.P. (2005) Synthetic lethal analysis of Caenorhabditis elegans posterior embryonic patterning genes identifies conserved genetic interactions. *Genome Biol* **6**, R45.

Bennet, M.D. and Leitch, I.J. (1995) Nuclear DNA amounts in angiosperms. *Ann Bot* **76**, 113–176.

Bennett, S. (2004) Solexa Ltd. *Pharmacogenomics* **5**, 433–438.

Bennetzen, J.L., Schrick, K., Springer, P.S., Brown, W.E. and SanMiguel, P. (1994) Active maize genes are unmodified and flanked by diverse classes of modified, highly repetitive DNA. *Genome* **37**, 565–576.

Bergmann, S., Ihmels, J. and Barkai, N. (2004) Similarities and differences in genome-wide expression data of six organisms. *Plos Biol* **2**, 85–93.

Bevan, M. and Walsh, S. (2005) The Arabidopsis genome: a foundation for plant research. *Genome Res* **15**, 1632–1642.

Bogdanov, A. (1980) *Essays in Tektology: The General Science of Organization* (Seaside, CA: Intersystems Publications).

Bollig, I., Mayer, K., Mayer, W.-E. and Engelmann, W. (1978) Effects of cAMP, theophylline, imidazole, and 4-(3,4-dimethoxybenzyl)-2-imidazolidone on the leaf movement rhythm of Trifolium repens – a test of the cAMP-hypothesis of circadian rhythms. *Planta* **141**, 225–230.

Breitkreutz, B.-J., Stark, C. and Tyers, M. (2003) Osprey: a network visualization system. *Genome Biol* **4**, r22.1–r22.4.

Breitling, R., Armengaud, P., Amtmann, A. and Herzyk, P. (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett* **573**, 83–92.

Bunning, E. and Moser, I. (1973) Light-induced phase shifts of circadian leaf movements of phaseolus: comparison with the effects of potassium and of ethyl alcohol. *Proc Natl Acad Sci USA* **70**, 3387–3389.

Calderwood, M.A., Venkatesan, K., Xing, L., Chase, M.R., Vazquez, A., Holthaus, A.M., *et al.* (2007) Epstein-Barr virus and virus human protein interaction maps. *Proc Natl Acad Sci USA* **104**, 7606–7611.

Dahlquist, K., Salomonis, N., Vranizan, K., Lawlor, S. and Conklin, B. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet* **31**, 19–20.

Daniel, X., Sugano, S. and Tobin, E.M. (2004) CK2 phosphorylation of CCA1 is necessary for its circadian oscillator function in Arabidopsis. *PNAS* **101**, 3292–3297.

de Folter, S., Immink, R.G., Kieffer, M., Parenicova, L., Henz, S.R., Weigel, D., *et al.* (2005) Comprehensive interaction map of the Arabidopsis MADS Box transcription factors. *Plant Cell* **17**, 1424–1433.

DeRisi, J.L., Iyer, V.R. and Brown, P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686.

Doniger, S., Salomonis, N., Dahlquist, K., Vranizan, K., Lawlor, S. and Conklin, B. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol* **4**, R7.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* **95**, 14863.

Endy, D. and Brent, R. (2001) Modelling cellular behaviour. *Nature* **409**, 391–395.

Erdös, P. and Rényi, A. (1960) On the evolution of random graphs. *Publ Math Inst Hung Acad Sci* **5**, 17–61.

Farre, E.M., Harmer, S.L., Harmon, F.G., Yanovsky, M.J. and Kay, S.A. (2005) Overlapping and distinct roles of PRR7 and PRR9 in the Arabidopsis circadian clock. *Curr Biol* **15**, 47–54.

Ferro, A., Pigola, G., Pulvirenti, A. and Shasha, D. (2003) Fast clustering and minimum weight matching algorithms for very large mobile backbone wireless networks. *Int J Found Comput Sci* **14**, 223–236.

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

Finnegan, E.J. and Matzke, S.A. (2003) The small RNA world. *J Cell Sci* **116**, 4689–4693.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**, 496–512.

Fletcher, J.C., Brand, U., Running, M.P., Simon, R. and Meyerowitz, E.M. (1999) Signaling of cell fate decisions by CLAVATA3 in Arabidopsis shoot meristems. *Science* **283**, 1911–1914.

Forde, B.G. (2002) Local and long-range signaling pathways regulating plant responses to nitrate. *Annu Rev Plant Biol* **53**, 203–224.

Friso, G., Giacomelli, L., Ytterberg, A.J., Peltier, J.B., Rudella, A., Sun, Q., *et al.* (2004) In-depth analysis of the thylakoid membrane proteome of Arabidopsis thaliana chloroplasts: new proteins, new functions, and a plastid proteome database. *Plant Cell* **16**, 478–499.

Geisler-Lee, J., O'Toole, N., Ammar, R., Provart, N.J., Millar, A.H. and Geisler, M. (2007) A predicted interactome for Arabidopsis. *Plant Physiol* **145**, 317–329.

Giavalisco, P., Nordhoff, E., Kreitler, T., Kloppel, K.D., Lehrach, H., Klose, J., *et al.* (2005) Proteome analysis of Arabidopsis thaliana by two-dimensional gel electrophoresis and matrix-assisted laser desorption/ionisation-time of flight mass spectrometry. *Proteomics* **5**, 1902–1913.

Gierer, A. and Meinhardt, H. (1972) A theory of biological pattern formation. *Kybernetik* **12**, 30–39.

Gifford, M.L., Dean, A., Gutierrez, R.A., Coruzzi, G.M. and Birnbaum, K.D. (2008) Cell-specific nitrogen responses mediate developmental plasticity. *Proc Natl Acad Sci USA* **105**, 803–808.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., *et al.* (2003) A protein interaction map of drosophila melanogaster. *Science* **302** (5651), 1727–1736.

Gunsalus, K.C., Ge, H., Schetter, A.J., Goldberg, D.S., Han, J.D., Hao, T., *et al.* (2005) Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis. *Nature* **436**, 861–865.

Gutierrez, R.A., Green, P.J., Keegstra, K. and Ohlrogge, J.B. (2004) Phylogenetic profiling of the Arabidopsis thaliana proteome: what proteins distinguish plants from other organisms? *Genome Biol* **5**, R53.

Gutierrez, R.A., Lejay, L.V., Dean, A., Chiaromonte, F., Shasha, D.E. and Coruzzi, G.M. (2007) Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in Arabidopsis. *Genome Biol* **8**, R7.

Gutierrez, R.A., Shasha, D.E. and Coruzzi, G.M. (2005) Systems biology for the virtual plant. *Plant Physiol* **138**, 550–554.

Gutierrez, R.A., Stokes, T.L., Thum, K.E., Xu, X., Obertello, M., Katari, M.S., *et al.* (2008) Systems approach identifies an organic nitrogen-responsive gene network that is regulated by the master clock control gene *CCA1. Proc Natl Acad Sci USA* **105** (12), 4939–4944.

Han, J.D., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88–93.

Hartwell, L.H., Hopfield, J.J., Leibler, S. and Murray, A.W. (1999) From molecular to modular cell biology. *Nature* **402**, C47–C52.

Heazlewood, J.L., Tonti-Filippini, J.S., Gout, A.M., Day, D.A., Whelan, J. and Millar, A.H. (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.

Henderson, I.R. and Jacobsen, S.E. (2007) Epigenetic inheritance in plants. *Nature* **447**, 418–424.

Hesse, H. and Hoefgen, R. (2006) On the way to understand biological complexity in plants: S-nutrition as a case study for systems biology. *Cell Mol Biol Lett* **11**, 37–56.

Hirai, M.Y., Yano, M., Goodenowe, D.B., Kanaya, S., Kimura, T., Awazuhara, M., *et al.* (2004) Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. *Proc Natl Acad Sci USA* **101**, 10205–10210.

Hodgkin, A.L. and Huxley, A.F. (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *J Physiol* **117**, 500–544.

Hu, Z., Mellor, J., Wu, J. and DeLisi, C. (2004) VisANT: an online visualization and analysis tool for biological interaction data. *BMC Bioinformatics* **5**, 17.

Hwang, D., Smith, J.J., Leslie, D.M., Weston, A.D., Rust, A.G., Ramsey, S., *et al.* (2005) A data integration methodology for systems biology: experimental verification. *Proc Natl Acad Sci USA* **102**, 17302–17307.

Hybrigenics (2004) http://pim.hybrigenics./.

Initiative. T.A.G. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796–815.

Isalan, M., Lemerle, C., Michalodimitrakis, K., Horn, C., Beltrao, P., Raineri, E., *et al.* (2008) Evolvability and hierarchy in rewired bacterial gene networks. *Nature* **452**, 840–845.

Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467.

Jeong, H., Mason, S.P., Barabasi, A.L. and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. *Nature* **411**, 41–42.

Jonsson, H., Heisler, M., Reddy, G.V., Agrawal, V., Gor, V., Shapiro, B.E., *et al.* (2005) Modeling the organization of the WUSCHEL expression domain in the shoot apical meristem. *Bioinformatics* **21** (Suppl 1), i232–i240.

Kaminuma, E., Heida, N., Tsumoto, Y., Nakazawa, M., Goto, N., Konagaya, A., *et al.* (2005) Three-dimensional definition of leaf morphological traits of Arabidopsis in silico phenotypic analysis. *J Bioinform Comput Biol* **3**, 401–414.

Karp, P.D. (1996) A strategy for database interoperation. *J Comput Biol* **2** (4), 573–583.

Karwowski, R. and Prusinkiewicz, P. (2003) Design and implementation of the L+C modeling language. *Electron Notes Theor Comput Sci* **86** (2), 1–19.

Kelley, B.P., Sharan, R., Karp, R.M., Sittler, T., Root, D.E., Stockwell, B.R., *et al.* (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *PNAS* **100**, 11394–11399.

Khatri, P., Draghici, S., Ostermeier, G.C. and Krawetz, S.A. (2002) Profiling gene expression using onto-express. *Genomics* **79**, 266–270.

Kim, J.Y., Song, H.R., Taylor, B.L. and Carré, I.A. (2003) Light-regulated translation mediates gated induction of the Arabidopsis clock protein LHY. *EMBO J* **22**, 935–944.

Kitano, H. (2002) Computational systems biology. *Nature* **420**, 206–210.

Kitano, H. (2004) Biological robustness. *Nat Rev Genet* **5**, 826–837.

Kitano, H. (2007) Towards a theory of biological robustness. *Mol Syst Biol* **3**, 137.
Kondo, T. (1983) Phase shifts of potassium uptake rhythm in Lemna gibba G3 due to light, dark or temperature pulses. *Plant Cell Physiol* **24**, 659–665.
Krouk, G., Tillard, P. and Gojon, A. (2006) Regulation of the high-affinity NO_3^- uptake system by a NRT1.1-mediated 'NO_3^--demand' signalling in *Arabidopsis*. *Plant Physiol* **142**, 1075–1086.
Kuromori, T., Wada, T., Kamiya, A., Yuguchi, M., Yokouchi, T., Imura, Y., *et al.* (2006) A trial of phenome analysis using 4000 Ds-insertional mutants in gene-coding regions of Arabidopsis. *Plant J* **47**, 640–651.
Lahner, B., Gong, J., Mahmoudian, M., Smith, E.L., Abid, K.B., Rogers, E.E., *et al.* (2003) Genomic scale profiling of nutrient and trace elements in Arabidopsis thaliana. *Nat Biotechnol* **21**, 1215–1221.
Lehner, B. (2007) Modelling genotype-phenotype relationships and human disease with genetic interaction networks. *J Exp Biol* **210**, 1559–1566.
Lehner, B., Crombie, C., Tischler, J., Fortunato, A. and Fraser, A.G. (2006) Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. *Nat Genet* **38**, 896–903.
Lehner, B. and Fraser, A.G. (2004) A first-draft human protein-interaction map. *Genome Biol* **5**, R63.
Locke, J.C., Kozma-Bognar, L., Gould, P.D., Feher, B., Kevei, E., Nagy, F., *et al.* (2006) Experimental validation of a predicted feedback loop in the multi-oscillator clock of Arabidopsis thaliana. *Mol Syst Biol* **2**, 59.
Locke, J.C., Southern, M.M., Kozma-Bognar, L., Hibberd, V., Brown, P.E., Turner, M.S., *et al.* (2005) Extension of a genetic network model by iterative experimentation and mathematical analysis. *Mol Syst Biol* **1**, 0013.
Loew, L.M. and Schaff, J.C. (2001) The virtual cell: a software environment for computational cell biology. *Trends Biotechnol* **19**, 401–406.
Lu, C., Tej, S.S., Luo, S., Haudenschild, C.D., Meyers, B.C. and Green, P.J. (2005) Elucidation of the small RNA component of the transcriptome. *Science* **309**, 1567–1569.
Ludemann, A., Weicht, D., Selbig, J. and Kopka, J. (2004) PaVESy: pathway visualization and editing system. *Bioinformatics* **20** (16), 2841–2844.
Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380.
Matsushika, A., Makino, S., Kojima, M. and Mizuno, T. (2000) Circadian waves of expression of the APRR1/TOC1 family of pseudo-response regulators in Arabidopsis thaliana: insight into the plant circadian clock. *Plant Cell Physiol* **41** (9), 1002–1012.
McClung, C.R. (2006) Plant circadian rhythms. *Plant Cell* **18**, 792–803.
Mendes, P. (1997) Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem Sci* **22**, 361–363.
Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., *et al.* (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* **318**, 245–250.
Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J.H., *et al.* (2008) The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature* **452**, 991–996.
Mitreva, M., McCarter, J.P., Martin, J., Dante, M., Wylie, T., Chiapelli, B., *et al.* (2004) Comparative genomics of gene expression in the parasitic and free-living nematodes Strongyloides stercoralis and Caenorhabditis elegans. *Genome Res* **14**, 209–220.
Mizoguchi, T., Wheatley, K., Hanzawa, Y., Wright, L., Mizoguchi, M., Song, H.R., *et al.*

(2002) LHY and CCA1 are partially redundant genes required to maintain circadian rhythms in Arabidopsis. *Dev Cell* **2**, 629–641.
Molina, C. and Grotewold, E. (2005) Genome wide analysis of Arabidopsis core promoters. *BMC Genomics* **6**, 25.
Mueller, L.A., Zhang, P. and Rhee, S.Y. (2003) AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiol* **132**, 453–460.
Mundermann, L., Erasmus, Y., Lane, B., Coen, E. and Prusinkiewicz, P. (2005) Quantitative modeling of Arabidopsis development. *Plant Physiol* **139**, 960–968.
Oliveira, I.C. and Coruzzi, G.M. (1999) Carbon and amino acids reciprocally modulate the expression of glutamine synthetase in Arabidopsis. *Plant Physiol* **121**, 301–310.
Oliver, S.G., Winson, M.K., Kell, D.B. and Baganz, F. (1998) Systematic functional analysis of the yeast genome. *Trends Biotechnol* **16**, 373–378.
Ooi, S.L., Pan, X., Peyser, B.D., Ye, P., Meluh, P.B., Yuan, D.S., *et al.* (2006) Global synthetic-lethality analysis and yeast functional profiling. *Trends Genet* **22**, 56–63.
Ott, K.H., Aranibar, N., Singh, B. and Stockton, G.W. (2003) Metabonomics classifies pathways affected by bioactive compounds. Artificial neural network classification of NMR spectra of plant extracts. *Phytochemistry* **62**, 971–985.
Palaniswamy, S.K., James, S., Sun, H., Lamb, R.S., Davuluri, R.V. and Grotewold, E. (2006) AGRIS and AtRegNet. A platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant Physiol* **140**, 818–829.
Palenchar, P., Kouranov, A., Lejay, L. and Coruzzi, G. (2004) Genome-wide patterns of carbon and nitrogen regulation of gene expression validate the combined carbon and nitrogen (CN)-signaling hypothesis in plants. *Genome Biol* **5**, R91.
Palmer, L.E., Rabinowicz, P.D., O'Shaughnessy, A.L., Balija, V.S., Nascimento, L.U., Dike, S., *et al.* (2003) Maize genome sequencing by methylation filtration. *Science* **302**, 2115–2117.
Parker, J.S., Cavell, A.C., Dolan, L., Roberts, K. and Grierson, C.S. (2000) Genetic interactions during root hair morphogenesis in Arabidopsis. *Plant Cell* **12**, 1961–1974.
Pilgrim, M.L., Caspar, T., Quail, P.H. and McClung, C.R. (1993) Circadian and light-regulated expression of nitrate reductase in *Arabidopsis*. *Plant Mol Biol* **23**, 349–364.
Popescu, S.C., Popescu, G.V., Bachan, S., Zhang, Z., Seay, M., Gerstein, M., *et al.* (2007) Differential binding of calmodulin-related proteins to their targets revealed through high-density Arabidopsis protein microarrays. *Proc Natl Acad Sci USA* **104**, 4730–4735.
Poultney, C.S., Gutierrez, R.A., Katari, M.S., Gifford, M.L., Paley, W.B., Coruzzi, G.M., *et al.* (2007) Sungear: interactive visualization and functional analysis of genomic datasets. *Bioinformatics* **23**, 259–261.
Price, J., Laxmi, A., St Martin, S.K. and Jang, J.C. (2004) Global transcription profiling reveals multiple sugar signal transduction mechanisms in Arabidopsis. *Plant Cell* **16**, 2128–2150.
Prusinkiewicz, P. and Rolland-Lagan, A.G. (2006) Modeling plant morphogenesis. *Curr Opin Plant Biol* **9**, 83–88.
Queitsch, C., Sangster, T.A. and Lindquist, S. (2002) Hsp90 as a capacitor of phenotypic variation. *Nature* **417**, 618–624.
Rabinowicz, P.D., McCombie, W.R. and Martienssen, R.A. (2003a) Gene enrichment in plant genomic shotgun libraries. *Curr Opin Plant Biol* **6**, 150–156.
Rabinowicz, P.D., Palmer, L.E., May, B.P., Hemann, M.T., Lowe, S.W., McCombie, W.R., *et al.* (2003b) Genes and transposons are differentially methylated in plants, but not in mammals. *Genome Res* **13**, 2658–2664.
Rabinowicz, P.D., Schutz, K., Dedhia, N., Yordan, C., Parnell, L.D., Stein, L., *et al.*

(1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat Genet* **23**, 305–308.

Rahayu, Y.S., Walch-Liu, P., Neumann, G., Romheld, V., von Wiren, N. and Bangerth, F. (2005) Root-derived cytokinins as long-distance signals for NO_3^--induced stimulation of leaf growth. *J Exp Bot* **56**, 1143–1152.

Raper, C.D., Jr., Thomas, J.F., Tolley-Henry, L. and Rideout, J.W. (1988) Assessment of an apparent relationship between availability of soluble carbohydrates and reduced nitrogen during floral initiation in tobacco. *Bot Gaz* **149**, 289–294.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. and Barabasi, A.L. (2002) Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555.

Rawat, S.R., Silim, S.N., Kronzucker, H.J., Siddiqi, M.Y. and Glass, A.D. (1999) *AtAMT1* gene expression and NH4+ uptake in roots of *Arabidopsis thaliana*: evidence for regulation by root glutamine levels. *Plant J* **19**, 143–152.

Reinhardt, D., Frenz, M., Mandel, T. and Kuhlemeier, C. (2003) Microsurgical and laser ablation analysis of interactions between the zones and layers of the tomato shoot apical meristem. *Development* **130**, 4073–4083.

Rideout, J.W., Raper, C.D., Jr. and Miner, G.S. (1992) Changes in ratio of soluble sugars and free amino nitrogen in the apical meristem during floral transition of tobacco. *Int J Plant Sci* **153**, 78–88.

Roessner, U., Willmitzer, L. and Fernie, A.R. (2001) High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. *Plant Physiol* **127**, 749–764.

Rutherford, S.L. and Lindquist, S. (1998) Hsp90 as a capacitor for morphological evolution. *Nature* **396**, 336–342.

Salathia, N. and Queitsch, C. (2007) Molecular mechanisms of canalization: Hsp90 and beyond. *J Biosci* **32**, 457–463.

Schauer, N. and Fernie, A.R. (2006) Plant metabolomics: towards biological function and mechanism. *Trends Plant Sci* **11**, 508–516.

Schauer, N., Steinhauser, D., Strelkov, S., Schomburg, D. and Allison, G. (2005) GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Lett* **579**, 1332.

Scheible, W.R., Morcuende, R., Czechowski, T., Fritz, C., Osuna, D., Palacios-Rojas, N., *et al.* (2004) Genome-wide reprogramming of primary and secondary metabolism, protein synthesis, cellular growth processes, and the regulatory infrastructure of Arabidopsis in response to nitrogen. *Plant Physiol* **136**, 2483–2499.

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., *et al.* (2005) A gene expression map of Arabidopsis thaliana development. *Nat Genet* **37**, 501–506.

Scholl, R.L., May, S.T. and Ware, D.H. (2000) Seed and molecular resources for Arabidopsis. *Plant Physiol* **124**, 1477–1480.

Schwender, J., Ohlrogge, J.B. and Shachar-Hill, Y. (2003) A flux model of glycolysis and the oxidative pentosephosphate pathway in developing brassica napus embryos. *J Biol Chem* **278**, 29442–29453.

Schwikowski, B., Uetz, P. and Fields, S. (2000) A network of protein-protein interactions in yeast. *Nat Biotechnol* **18**, 1257–1261.

Siepel, A., Farmer, A., Tolopko, A., Zhuang, M., Mendes, P., Beavis W, *et al.* (2001) ISYS:

a decentralized, component-based approach to the integration of heterogeneous bioinformatics resources. *Bioinformatics* **17**, 83–94.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–2504.

Sharma, V.K., Carles, C. and Fletcher, J.C. (2003) Maintenance of stem cell populations in plants. *Proc Natl Acad Sci USA* **100** (Suppl 1), 11823–11829.

Sollars, V., Lu, X., Xiao, L., Wang, X., Garfinkel, M.D. and Ruden, D.M. (2003) Evidence for an epigenetic mechanism by which Hsp90 acts as a capacitor for morphological evolution. *Nat Genet* **33**, 70–74.

Stelling, J., Sauer, U., Szallasi, Z., Doyle, F.J., III and Doyle, J. (2004) Robustness of cellular functions. *Cell* **118**, 675–685.

Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., *et al.* (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.

Turing, A.M. (1952) The chemical basis of morphogenesis. *Philos Trans R Soc Lond B Biol Sci* **237**, 37–72.

Ueda, H.R. (2006) Systems biology flowering in the plant clock field. *Mol Syst Biol* **2**, 60.

Van Leene, J., Stals, H., Eeckhout, D., Persiau, G., Van Slijke, E., Van Isterdael, G., *et al.* (2007) A tandem affinity purification-based technology platform to study the cell cycle interactome in Arabidopsis thaliana. *Mol Cell Proteomics* **6**, 1226–1238.

Waddington, C.H. (1959) Canalization of development and genetic assimilation of acquired characters. *Nature* **183**, 1654–1655.

Wagner, A. and Fell, D.A. (2001) The small world inside large metabolic networks. *Proc Biol Sci* **268**, 1803–1810.

Walch-Liu, P., Filleur, S., Gan, Y. and Forde, B.G. (2005) Signaling mechanisms integrating root and shoot responses to changes in the nitrogen supply. *Photosynth Res* **83**, 239–250.

Walch-Liu, P., Neumann, G., Bangerth, F. and Engels, C. (2000) Rapid effects of nitrogen form on leaf morphogenesis in tobacco. *J Exp Bot* **51**, 227–237.

Walhout, A.J., Sordella, R., Lu, X., Hartley, J.L., Temple, G.F., Brasch, M.A., *et al.* (2000) Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science* **287**, 116–122.

Wasinger, V.C. and Corthals, G.L. (2002) Proteomic tools for biomedicine. *J Chromatogr B Analyt Technol Biomed Life Sci* **771**, 33–48.

Watanabe, T., Hanan, J.S., Room, P.M., Hasegawa, T., Nakagawa, H. and Takahashi, W. (2005) Rice morphogenesis and plant architecture: measurement, specification and the reconstruction of structural development by 3D architectural modelling. *Ann Bot* (*Lond*) **95**, 1131–1143.

Watts, D. and Strogatz, S. (1997) Collective dynamic of 'small-world' networks. *Nature* **393**, 440–442.

Wei, H., Persson, S., Mehta, T., Srinivasasainagendra, V., Chen, L., Page, G.P., *et al.* (2006) Transcriptional coordination of the metabolic network in Arabidopsis. *Plant Physiol* **142**, 762–774.

Wheeler, D.B., Bailey, S.N., Guertin, D.A., Carpenter, A.E., Higgins, C.O. and Sabatini, D.M. (2004) RNAi living-cell microarrays for loss-of-function screens in Drosophila melanogaster cells. *Nat Methods* **1**, 127–132.

Wilkinson, M.D., Gessler, D., Farmer, A. and Stein, L. (2003) The BioMOBY project

explores open-source, simple, extensible protocols for enabling biological database interoperability. *Proc Virt Conf Genom Bioinf* **3**, 16–26 (ISSN 1547-383X).

Wilkinson, M.D. and Links, M. (2002) BioMOBY: an open source biological web services proposal. *Brief Bioinform* **3**, 331–341.

Wilson, E. (2000) *Consilience: The Unity of Knowledge* (New York, NY: Aldred A. Knopf).

Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G.V. and Provart, N.J. (2007) An 'electronic fluorescent pictograph' browser for exploring and analyzing large-scale biological data sets. *PLoS ONE* **2**, e718.

Wu, M.F., Tian, Q. and Reed, J.W. (2006) Arabidopsis microRNA167 controls patterns of *ARF6* and *ARF8* expression, and regulates both female and male reproduction. *Development* **133**, 4211–4218.

Wurtele, E., Li, L., Berleant, D., Cook, D., Dickerson, J.A., Ding, J., *et al.* (2007) MetNet: systems biology software for arabidopsis. In *Concepts in Plant Metabolomics*, B.J. Nikolau and E.S. Wurtele, eds (Heidelberg: Springer), 145–158.

Yates, J.R., III (2000) Mass spectrometry. From genomics to proteomics. *Trends Genet* **16**, 5–8.

Young, N.D., Cannon, S.B., Sato, S., Kim, D., Cook, D.R., Town, C.D., *et al.* (2005) Sequencing the genespaces of Medicago truncatula and Lotus japonicus. *Plant Physiol* **137**, 1174–1181.

Yu, H., Zhu, X., Greenbaum, D., Karro, J. and Gerstein, M. (2004) TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucl Acids Res* **32**, 328–337.

Yuan, Y., SanMiguel, P.J. and Bennetzen, J.L. (2003) High-Cot sequence analysis of the maize genome. *Plant J* **34**, 249–255.

Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M. and Cesareni, G. (2002) MINT: a molecular INTeraction database. *FEBS Lett* **513**, 135–140.

Zeeberg, B., Feng, W., Wang, G., Wang, M., Fojo, A., Sunshine, M., *et al.* (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol* **4**, R28.

Zhang, X., Clarenz, O., Cokus, S., Bernatavichute, Y.V., Pellegrini, M., Goodrich, J., *et al.* (2007) Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. *PLoS Biol* **5**, e129.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W., Chen, H., *et al.* (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell* **126**, 1189–1201.

Zhong, S., Storch, K.F., Lipan, O., Kao, M.C., Weitz, C.J. and Wong, W.H. (2004) GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space. *Appl Bioinformatics* **3**, 261–264.

Zhong, W. and Sternberg, P.W. (2006) Genome-wide prediction of C. elegans genetic interactions. *Science* **311**, 1481–1484.

Zhou, M. and Cui, Y. (2004) GeneInfoViz: constructing and visualizing gene relation networks. *In Silico Biol* **4**, 0026.

Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T. and Henikoff, S. (2007) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet* **39**, 61–69.

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