In Silico Analysis of Crop Science: Report on the First China-UK Workshop on Chips, Computers and Crops

Ming Chen[1]* and Andrew Harrison[2]

[1] Department of Bioinformatics / State Key Laboratory of Plant Physiology and Biochemistry, College of Life Sciences, Zhejiang University, Hangzhou 310058, China; [2] Departments of Mathematical Sciences & Biological Sciences, University of Essex, Colchester CO4 3SQ, UK.

A workshop on "Chips, Computers and Crops" was held in Hangzhou, China during September 26–27, 2008. The main objective of the workshop was to bring together China and UK scientists from mathematics, bioinformatics and plant molecular biology communities to exchange ideas, enhance awareness of each others' fields, explore synergisms and make recommendations on fruitful future directions in crop science. Here we describe the contributions to the workshop, and examine some conceptual issues that lie at the foundations and future of crop systems biology.

Key words: systems biology, functional genomics, plant, chip, computer, crop

Introduction

Our species is facing several daunting problems, including an increasing shortage of available food, the depletion of global oil reserves, and a mounting scarcity of freshwater. Feeding the growing world population is a major challenge. Plant biology has the potential of providing partial solutions to these daunting problems. To meet future demands on plants, both as food and fuel, we must gain the ability to engineer the design of crop plants predictably. To do this, it is necessary to gain a comprehensive knowledge of how plants function in a relatively short period of time. We must develop our understanding of how plants function to such an extent that we can precisely predict by modeling how a plant will respond to any given genetic manipulation or environmental perturbation. Fortunately, as we step into the 21st century, the tools for achieving this degree of knowledge are available or within grasp.

In the post-genomic era, all the major hurdles to obtaining genomes have been overcome, and increasing numbers of plant genomes are being sequenced and publicly accessible. In order to better understand how genomes function and are regulated, technologies such as microarrays, have allowed investigators to measure simultaneously which genes are turned on or off in a genome in response to an experimental treatment. Techniques that profile changes in gene expression permit the analysis of the expression levels of thousands of genes simultaneously. Meanwhile, proteomic methods reveal the proteins translated from the mRNA molecules as the direct result of gene expression.

In summarizing a systems biology meeting three years ago, an underground swell of research in systems biology was noted (*1*). To continue discussing important issues of crop systems biology, a China-UK workshop on "Chips, Computers and Crops" was held in Hangzhou, China during September 26–27, 2008. This workshop, sponsored by the Biotechnology and Biological Sciences Research Council (BBSRC) of UK, was co-chaired by Dr. Ming Chen (Zhejiang University, China) and Dr. Andrew Harrison (Essex University, UK). Participants included theoreticians, informaticians and plant scientists with a wide range of interests. The attendees from UK included Prof. Sean May, Dr. Zoe Wilson and Dr. Neil Graham (University of Nottingham), Ms. Joanna Rowsell (University of Essex), Dr. Sascha Ott and Dr. John Hammond (University of Warwick), Dr. Chris Needham (University of Leeds), Dr. Hugh Shanahan (Royal Holloway, London), and Dr. Mark Winfield (University of Bristol). The attendees from China included Prof. Lingling Chen (Shandong University of Technology), Prof. Huixia Shou (Zhejiang University), Prof. Xun Yu (Shandong Agricultural University), Prof. Shaojian Zheng (Zhejiang University), Dr. Cheng Xu (Zhejiang University), Dr. Yijun Meng (Zhejiang University), Dr. Yan Lee (Yunnan University), Mr. Junjie Cao (Zhejiang University), Mr.

***Corresponding author.**
E-mail: mchen@zju.edu.cn

Qiang Jiang (Shanghai Jiaotong University), Mr. Dijun Chen (Harbin Medical University) and other colleagues. The discussion covered many topics that are summarized as follows.

Topics

Crop omics

In the past, plant developmental biologists described how anatomy and morphology are established during ontogenesis, and geneticists identified many regulators. In contrast to the traditional approaches that mostly focus on one or a few genes at a time, highly parallel omics profiling technologies have been developed to probe many genes, transcripts, proteins or metabolites at once. Omics focuses on large scale and holistic data/information to understand life in encapsulated omes such as genome, transcriptome, proteome, metabolome, and interactome. Crop omics is a broad discipline of science and engineering for analyzing the interactions of biological information objects in various omes. Undoubtedly, crop omics technologies will not only make traditional gene-centered approaches obsolete but should instead complement forward plant genetics.

Microarrays are used to detect DNA and RNA on a genome scale. Modern arrays can contain tens of thousands of probes, and a microarray experiment can accomplish genetic tests in parallel. Therefore, arrays have dramatically accelerated many types of investigation. Applications include gene expression profiling, single nucleotide polymorphism (SNP) detection, and detecting post-transcriptional processing. Several talks in the workshop focused on the big picture of microarray analysis, the infrastructure that presently exists, developing annotation pipelines, and quality control assessment.

A presentation entitled "*Arabidopsis* omics and their applications to crops" by **Sean May** started the talks. His group and other Nottingham plant scientists have been at the forefront of international research studying the model plants *Arabidopsis thaliana* and tomato, identifying many of the key genes that regulate their development, coordinating their genome sequencing efforts and, through the Nottingham *Arabidopsis* Stock Centre (NASC), providing underpinning resources to the international scientific community. NASC distributes, collects and preserves seed and DNA resources of *Arabidopsis* and crop species. In addition, it also collects, generates and distributes transcriptomics data (especially Affymetrix) in the form of a primary repository (NASCarrays), which represents the majority of plant transcriptomic experiments in the public domain. NASC now has a public version of Genespring workgroup for community use in analyzing transcriptomic data. It also has a mature integrated genome browser AtEnsEMBL (http://atensembl.arabidopsis.info) (*2*) as part of ukcrop.net, which incorporates both TAIR and MIPS genome annotations and links through it to all of the other databases and resources.

Ming Chen's group developed a comprehensive computational platform for rice microarray annotation and data analysis. The platform provides convenient query for comprehensive annotation compared with other similar databases. Moreover, coupled with existing rice microarray data, it provides online analysis methods from the perspective of bioinformatics. This comprehensive bioinformatics analysis platform is composed of five modules, including data fetching, microarray annotation, sequence analysis, result visualization and data analysis. The integrative functional annotation system *Os*CAS (*Oryza sativa* Chips Annotation System) is a comprehensive web-based system to analyze the results of rice microarray experiments and reveal the potential relationships among genes. This platform is designed to facilitate the further exploration of microarray data within the framework of systems biological research. With a user-friendly web interface, *Os*CAS adopts gene chip probe IDs as inputs to retrieve relevant information under user's designation. In this system, public databases, including GenBank, UniGene, Swiss-Prot, TIGR, KOME, KEGG, Gene Ontology and miRBase are integrated to cover gene information, protein features, metabolic pathways and regulatory factors in rice. Besides the public rice genomic resources, *Os*CAS also includes the reprocessed information from several useful analytical tools such as CSRDB and miRU to guide a deep mining of the primary annotations. *Os*CAS has been successfully applied in annotating large sets of gene chip probe IDs from several rice microarray experiments and efficiently facilitated the further biological experiment design. *Os*CAS runs on a Linux server and is freely available at http://bis.zju.edu.cn/oscas/.

Andrew Harrison reported efforts to mine large repositories of Affymetrix GeneChip data in order to identify systematic biases in the data and to discover novel signals associated with post-transcriptional processing of RNA. Many probes containing contiguous runs of guanine tend to be outliers with respect to the rest of the probeset to which they are assigned. They discovered that rather than being random outliers, these probes with runs of guanine show correlated behavior with each other (*3*). This indicates directly that they are not measuring a biological signal but a bias associated with the biophysics of GeneChips. It is believed that this behavior results from the formation of G-quadruplexes as a result of 4 spatially adjacent 25-mers hybridizing to each other rather than to target.

Joanna Rowsell presented how correlations between all the probe-pair permutations within a GeneChip probeset frequently show "blocks", indicating that subsets of probes measure distinctive isoforms. Such blocks provide signals associated with alternative polyadenylation and alternative splice site selection. They also identified the cases when SNPs cause outliers in probesets. The results indicate that there are no simple heuristics associated with whether SNPs are responsible for outliers.

Using omics to discover biology relevant to crops

An organism's transcriptome is its set of gene transcripts (mRNAs) at a defined spatial and temporal location. Since gene expression is affected markedly by environmental and developmental perturbations, transcriptome divergence among taxa will evolve through adaptive phenotypic selection. However, **Neil Graham** reported that stochastic, evolutionarily neutral processes also drive transcriptome divergence in plants (*4*). Among 14 Brassicaceae (cabbage family) taxa, transcriptome divergence correlates positively with evolutionary distance between taxa and with gene expression diversity within replicate samples. Remarkably, the transcriptomes of functionally homologous tissues sampled from different taxa have diverged more than the transcriptomes of functionally discrete—and highly specialized—tissues from one taxa. These observations are consistent with neutral evolutionary theories.

Monitoring the mineral requirements of crops by assaying soil or foliar concentrations is notoriously unreliable. Agricultural industries routinely apply excessive amounts of fertilisers to maintain crop yields and quality. This is both costly and can lead to unnecessary pollution. An alternative approach for diagnosing the nutritional status of plants is through changes in gene expression. **John Hammond**'s group identified genes whose expression is altered specifically in response to phosphate (Pi) withdrawal in *A. thaliana* and *Brassica oleracea*. More recently, they have used custom microarrays, representing more than 40,000 potato transcripts, to monitor changes in the expression of potato genes under increasing Pi deficiency and the subsequent re-supply of Pi. They then used a support vector machine to define genes that were diagnostic for Pi deficiency. These genes were then used to successfully identify the Pi status of samples taken from field-grown potato plants. Common transcriptional events occur in all three species they have studied, and also in other species, during Pi starvation. By monitoring the expression of these Pi responsive genes, there is potential to detect a physiological P deficiency in crop plants before the lack of Pi affects yield and quality.

Zoe Wilson studied pathways in pollen development of *Arabidopsis*, and discussed the potential mechanism in rice. The *Arabidopsis* male sterility 1 (MS1) mutation results in mature anthers that are devoid of pollen. MS1 may function by modifying the transcription of tapetal-specific genes implicated in pollen wall development, which then regulate pollen wall material secretion and in turn wall development and tapetal programmed cell death (PCD). Alternatively, the MS1 gene may control tapetal development by directly regulating tapetal PCD and breakdown. The *Arabidopsis* MYB26/male sterility 35 (MS35) gene is critical for the development of secondary thickening in the anther endothecium and subsequent dehiscence. MYB26 regulates NST1 and NST2 expression and in turn controls the process of secondary thickening. Therefore, MYB26 appears to function in a regulatory role involved in determining endothecial cell development within the anther and acts upstream of the lignin biosynthesis pathway.

Understanding the nature of photomorphogenesis is a significant, though attainable, challenge of systems biology. In order to get more insight about this progress, **Hugh Shanahan** presented results from the analysis of transcriptomic data taken from the shoot apical meristem and cotyledon of *A. thaliana* at its seedling stage during photomorphogenesis. The data were taken from both tissues when initially exposed to light and at a series of times after continuous exposure to light. No amplification processes have been applied before applying the samples to Affymetrix chips. After constructing a set of genes that are strictly differentially expressed, they found that among 5,620 genes selected, 1/4 of the transcriptome were differentially expressed during photomorphogenesis. So far they have put together a picture of genes up and down regulated and are looking at their classification fits with a picture of particular types of growth.

Mark Winfield presented a talk on vernalization and cold acclimation in wheat. With proper experimental design, his group constructed a growth condition with controlled "winter". Then they performed microarray analysis to get a brief overview of changes in transcriptome. As we know, there are many pathways involved in phase change in *Arabidopsis*, which can be either induced by endogenous cues, environmental cues, or both. Gibberellin pathway, autonomous pathway, vernalization pathway and light-dependent pathway together contribute to the integration pathway that leads to floral morphogenesis. Based on this knowledge of *Arabidopsis*, Winfield *et al* used their microarray data to discover vernalization pathway, gibberellin pathway and photoperiod pathway in wheat.

In addition, **Lingling Chen** presented a database of improved gene annotation for bacterial phytopathogens. Features of the database DIGAP include: (1) Some "hypothetical genes" are recognized as non-coding ORFs and the identified non-coding ORFs are very unlikely to encode proteins. (2) The translation initiation sites of all the protein-coding genes are refined. (3) Potential functions of a large number of "hypothetical genes" are predicted. (4) Theoretical gene expression indices CAI and E(g) are calculated to show the gene expression levels. (5) Orthologs of the antibacterial therapeutic drug targets in human and animal species are enumerated in the database, which may provide potential targets for agricultural antibiotic discovery. All of the results provide more accurate information for the research of these phytopathogens and for antibiotic discovery in plant protection.

Towards a systems biology of crops

Using genomic techniques, we can now identify all the genes in a plant, and have successfully sequenced two plant genomes in their entirety—the model organism *Arabidopsis* (*5*) and the crop plant rice (*Oryza sativa*) (*6*, *7*). Moreover, using microarray and proteomic techniques, we now have the ability to resolve which genes are activated or inactivated during development or in response to an environmental change. However, identifying all the genes and proteins (system elements) in an organism is comparable with listing all the parts of an airplane. Although such a list provides a catalog of the individual components, by itself it is not sufficient to understand the complexity underlying the engineered object. What we really need to know before we can intelligently engineer plants is how all these system elements interact.

Systems biology is a new branch of biology that attempts to discover and understand biological properties emerging from the interactions of many system elements (*8*, *9*). The major reason why systems biology is gaining interest today is that progress in molecular biology, particularly in high-throughput genomics and proteomics, is enabling scientists to collect comprehensive datasets on the mechanisms underlying plant growth and plant responses to perturbations. The power of these new tools has led to an explosion of information unparalleled in the history of biology.

Gene regulatory networks govern the functional development and biological processes of cells in all organisms. Genes regulate each other as part of a complex system, of which it is vitally important to gain an understanding. Discovery of the complete gene regulatory networks in plants would allow the development of stress (drought/salt/temperature) resistant crops. Learning large gene regulatory networks with thousands of genes with any certainty from microarray data is extremely challenging. The research by **Chris Needham** aims to build around known networks from the literature on gene regulation, and assesses which other genes are likely to play a regulatory role or in the same regulatory pathways. The gene regulatory networks were modeled with a Bayesian network. The gene expression levels were quantized and a greedy hill climbing search method was used within a network structure learning algorithm. The inclusion of extra genes with the best explanatory power into the model has been demonstrated to be robust. In the analysis, large sets of microarray experiments were used, specifically 2,466 NASC *A. thaliana* microarrays containing gene expression levels of over 20,000 genes in a number of experimental conditions. Initial investigation of this data is very promising. They have learned gene transcription sub-networks regulated by the plant's circadian clock. The network shown was generated from microarray data without using any prior information, and yet the method managed to identify the strong causal relationships between clock components (TOC1, LHY, ELF3, ELF4, CCA1) and to link them to further key regulators of important processes (such as ZAT, myb and GATA transcription factors).

As the biological systems investigated by experimental biologists are of great complexity, so is their task of finding the best gene or the best transcription factor binding sites to put their experimental efforts into. This decision making can be greatly facilitated by combining the substantial amount of information that is distributed over various types of data (such as conserved genomic sequences, ChIP-chip data and microarray data), making it easily accessible for experimentalists. **Sascha Ott** discussed the detection of potential regulatory regions (or enhancers) from conserved genomic sequences, the analysis of such regions, as well as concepts to integrate a wider range of data in the future.

Yijun Meng presented a genome-wide comparative analysis of rice microRNA-target pairs and characterized their regulatory mechanisms in auxin signaling. Clustering analysis revealed novel auxin-related miRNAs that potentially mediated the signal interactions between auxin and nutrition or stress. MiRNA duplication and expression patterns suggested their evolutionary conservation similarity to protein-coding genes. The characteristics of miRNA promoters were similar with those of the RNA polymerase II-dependent genes. Reciprocal expression patterns indicated many miRNA-target pairs in roots. A feedback model between ARF(s) and miRNA(s) was established based on *cis*-element analysis, target prediction and expression data manipulation.

Perspective

Data manipulation

The amount of biological information in genomics, transcriptomics, proteomics and metabolomics datasets is increasing dramatically. The sheer immensity of the information explosion in plant biology presents new challenges. Researchers should jointly establish ways to organize and store these massive datasets in standard and easily accessible forms. Meanwhile, the increasing size of large biological datasets is leading to the need for ever more sophisticated analysis techniques. Researchers are increasingly encouraged to submit data to a central resource or make data publicly accessible via institutional or personal website. Incentives will be required to ensure this occurs, for example if data submitted to a public resource are used by others, it needs to be highlighted and recognized by the community, by internal and external research evaluations and by funding bodies. Possible suggestions could include a coordinated approach to collect a dataset like microarray data under specific conditions that we could all use and benefit from. However, data error issues need to be always considered; otherwise databases may become just a mass of data.

Development of standards for experimental plant research and for data exchange is a potential benefit from a large network. Appropriate ontologies and standards will be needed to deal efficiently and effectively with the data generated. The greatest challenge in this area is likely to be that new tools or technologies may need to be designed to collect the appropriate information at source rather than setting standards after data capture. One answer to this problem might be to collaborate in large groups to facilitate data collection wholly under standard conditions.

Traditionally trained biologists, who typically have had little experience manipulating the large datasets of the post-genomic era, have to acquire a working knowledge of data storage and access. Training in the use of such massive data archives must become necessary for young biologists. Moreover, currently there are technical and other barriers in applying systems biology approaches to plant research, such as lack of multidisciplinary training, poor communication with other disciplines, and the need to develop common languages. Thus, there is a need for training grants to teach biologists from different fields how to gain competency in data storage, retrieval and analysis.

Interdisciplinary teaching and training

The explosive growth of molecular biology in the last decade has helped to break down many artificial barriers that had been erected between different branches of biology. Now it is necessary that the walls between biology and other scientific disciplines (such as mathematics, computer science and engineering) be breached. Biologists need to become conversant with software developers, data storage archivists, mathematicians, analytical chemists and engineers. If we are to achieve the "*in silico*" plant, biologists need to understand in some depth the models of mathematicians and computer scientists, and need to supply the accurate measurements of many parameters that computational scientists will require to make their models robust. As a consequence, there is a strong need for the development of training programs and the initiation of training grants to foster the intellectual development of both the current and the next generation of scientists in all of the subdisciplines that constitute systems biology. The philosophy of education will also need to change. For example, students need to be trained to thrive in larger and extended research groups, and universities must acknowledge and reward multidisciplinary collaboration and community service by their faculty.

Unfortunately, each of the technical disciplines that systems biology attempts to merge has developed separately, and each has evolved its own technical jargon that seems impenetrable and arcane to the outsider. This barrier to communication must be overcome, and proximity is the best way to do that. This might be achieved through cross-disciplinary research opportunities for graduate students and postdoctoral research associates. Funding special centers devoted to systems biology research and staffed by biologists, computer scientists, mathematicians and engineers could be another approach to foster the development of systems biology and to achieve the goal of attaining the "*in silico*" plant.

At this time, a regular training program in plant systems biology is proposed. Students would be based across a distributed network of plant biology laboratories with appropriate, preferably local, theoretical co-supervisors. A training hub should provide computing support, some theoretical training for supervisors as well as students, and networking among groups. Supervisors without current systems biology funding must be included to broaden participation.

International collaboration

Since not every institution will be able to provide the all necessary expertise or indeed enthusiasm for problems in plant biology, collaboration is required. At the project level, individual biologists must identify collaborators, who may be in other institutions. This is not necessarily a barrier to progress, as many already have collaborators at distant locations, but ready access is clearly advantageous. At the community level, a significant increase in the number of theoreticians working on questions in plant science will be required to achieve any large-scale goals in crop systems biology.

Yet there is little present international coordination specific to crop systems biology, nor is there a consensus on the need for such coordination. The 2010 program of the National Science Foundation of USA on *Arabidopsis* functional genomics recommended in its mid-course report a focus on the provision of biological data and informatics resources and on understanding "exemplary networks" to facilitate the transition to systems biology. Locations in continental Europe that are strongly engaged in plant systems biology will be natural partners for a large-scale activity in UK. There are over 350 plant research groups in 42 institutions scattered from Aberdeen to Exeter in UK. Many of these groups are international leaders in their fields. UK can offer scientific expertise in particular: broad excellence in plant science, computational biology and informatics; numerous resources; large datasets such as transcriptomic datasets (NASC); genomic databases; large *in silico* crop breeding databases; wide range of species studied.

In China, crop research takes place mainly in Beijing, Shanghai, Wuhan and Hangzhou areas, including Chinese Academy of Sciences, Peking University, China Agricultural University, Tsinghua University, Zhejiang University and Huazhong Agricultural University. Funding for crop research is improving in China as the National Natural Science Foundation of China (NSFC) and the 973 Program, the two main fundings for basic research, doubled their budget in recent years. Confronted with land degradation, chronic water shortages, and a growing population that already numbers 1.3 billion, China is looking to a transgenic green revolution to secure its food supply. In September 2008, the government rolled out a \$3.5 billion research and development (R&D) initiative on genetically modified (GM) plants. Of the six GM plants including cotton, petunia, tomato, sweet pepper, poplar trees and papaya that China has approved for commercialization, only cotton is grown widely. A new initiative could pave the way for GM versions of the biggest prize of all: rice.

To promote international collaboration, funding agencies are generally keen to fund international collaborations and there exist many highly successful bilateral and trilateral projects. What is needed now are mechanisms to support large, intercontinental projects that make best use of truly exceptional facilities, expertise or resources wherever they are found. For example, there will be growing benefits from the coordination and standardization of experiment, data and model formats, maximizing the inter-lab comparability and thus the value of data and models from different groups. The visibility of larger-scale projects should facilitate interactions with other research communities (such as increase engagement by theoreticians), funding agencies, policy-makers and the public. We suggest that increased core activity or generation of a large dataset requires funding over longer timescales. It is a common experience that the most productive multidisciplinary interactions take several years to develop. Therefore, a five-year duration is a minimum.

Technical challenges and solutions

Crop systems biology shares many challenges with systems biology in other species, particularly multicellular organisms, which are not rehearsed here. However, data from defined plant cell types will be important for many projects. Data on whole organs or seedlings remain relevant in generating preliminary models and also for modeling subsystems that are similar in all cells. Here are some potential solutions to provide cell-type specific data.

Use of current cell cultures for baseline studies

Cell culture systems can produce diverse data that are readily amenable to modeling, including timecourse data after chemical interventions. These have been used to study plant cell cycle and senescence processes. However, plant cell suspension cultures do not obviously represent a particular cell type in the intact plant, in terms of metabolism, receptor expression or differentiation. Development of stable, differentiated cell cultures would greatly facilitate plant systems biology, but there are alternatives as following.

Cell purification or extraction of components from defined cell types

Certain plant cells can be physically separated at high purity, including pollen and stomatal guard cells. Other techniques exist to extract RNA, but potentially also protein and metabolites, from single cell types of intact plants. These techniques include laser-capture microdissection, fluorescently tagging and purifying protoplasts of a single cell type, sampling single cell contents using micropipettes, or the extraction of mRNA from polyribosomes that carry a cell-type-specific protein tag. Limited amounts of samples are recovered, permitting transcriptomic assays that include amplification, as well as assays for specific molecular species. Other profiling methods (proteomic and metabolomic) may be limited by the sample amounts, so continued technical development to increase the sensitivity of these methods is desirable.

Integration of biological data and databases

Data integration is one of the most challenging problems facing bioinformatics today. Researchers have to interpret many types of information from a variety of sources: lab instruments, public databases, gene expression profiles, raw sequence traces, single nucleotide polymorphisms, chemical screening data, proteomic data, putative metabolic pathway models, and many others. There are a number of techniques, approaches and products available to help scientists tackle this increasingly complex issue (*10*). These include: application-level integration (using middleware to integrate disparate data sources is usually referred to as a federated database approach); data-level integration without semantic cleaning (integrating data at the data layer through indices, database links, and memory-mapped data structures); and data-level integration with semantic cleaning (data warehouse). However, the need for standards efforts is imperative. There is currently an effort underway to standardize domain-specific ontologies and vocabularies to support interoperability of data and software components. Another approach is to standardize domain-specific analytical data models to help integrate public data with proprietary data across all life science domains in an enterprise. Moreover, XML based data storage and transfer are widely used.

Achieving the *in silico* crop plant

Successful modeling of crop plants is the ultimate goal of crop systems biology. Clearly, useful modeling will depend on having a large amount of high-quality quantitative information about all aspects of biological processes. Many new types of data have to be systematically determined. The ideal program will mimic the performance of a cell or tissue or organism. The modeling of intact higher plants will be especially challenging because of the differential responsiveness of various cell types to a given perturbation. The collection of the comprehensive data needed for modeling might initially be most successful using single-cell microorganisms or higher plant cells grown in defined liquid cultures.

To model plant response accurately, a multitude of software programs of the sorts widely used by engineers (to name a few, parameter optimization, flux balance analysis, systems analysis, and computer model simulations) need to be adapted for the study of plants. The integration of modeling with experimental work will derive many new insights, with greater complexity, and hopefully greater impact on the global problems we face. A key benefit of modeling is that we can rapidly do many "virtual experiments" and select those that are most interesting for the real world. For example, we could assess the tradeoffs "*in silico*" of different means for increasing the water use efficiency of a given crop given a set of possible genotypes and a range of possible environmental conditions. The generation of phenotype phase diagrams would enable plant biologists to predict with accuracy which factors are limiting to plant growth under different environmental conditions.

Potential risks and alternatives

With the more recent advent of genome scale data, we have gone from studying one gene–one protein at a time to the whole genome. In the new era of systems biology, our challenge will be to incorporate information on all genes and proteins in a cell into a composite model of interacting components. Despite the power of its promise, systems biology is still in its infancy.

Coordinating research in a dispersed network of laboratories will have significant benefits in addition to the added value in direct research outputs. The network will broaden the coordination and standardization of experiment, data and model formats through the community more effectively than a single center, facilitating development of public repositories and maximizing the value of data and models from many groups. The visibility of a large-scale project with a clear goal should facilitate interactions with other research communities, funding agencies, policymakers and the public.

Implications of the adoption of systems approaches in crop plant research include two points. The first is the selection of model species for systems research. The model species *Arabidopsis* has been realistically used for plant systems biology. We may need more model species for crop systems biology within a reasonable timeframe. Monocot plant rice is naturally intensively researched for years. While domesticated grass crops such as wheat are vital to human existence, understanding how genes in these crops function will be important to future crop improvement. The wild grass species *Brachypodium distachyon* (Brachypodium) has been proposed as a new model plant system for grass crops (*11*, *12*). Brachypodium is closely related to a large number of exceptionally important crops, including wheat, and has the necessary biological attributes desired in a model plant system. Achieving the availability of this new model plant system will accelerate the discovery of agriculturally important genes that can be used to improve the productivity of wheat and other grass crops.

The second is the conflict of research priorities between translational and fundamental research in plant labs. The "models to crops" agenda will have significant effects on the plant biology community at the same time as the systems biology effort develops. Smaller academic labs may be unable to sustain activity at both interfaces. For larger groups and the community more generally, there is a risk of diluting research effort in both areas.

Conclusion

This report highlights the recent workshop on "Chips, Computers and Crops", as well as visions of new and emerging research in the area of crop systems biology. As we accumulate more and more high-quality data, mathematical modeling of biological processes will become ever more successful; that is, the responses of plants to genetic manipulations and environmental perturbations will become increasingly predictable. This will enable us to engineer plants quickly and with foresight as to the cost benefit of the proposed "reprogramming" of the plant genome.

Under certain collaboration frameworks, the provision of exponentially accumulating biological data of crops could be coordinated internationally. Multiple initiatives will be necessary to develop crop systems biology on the scale required to tackle the challenge of a whole-crop systems model. UK will be in a strong position to contribute to some elements. A distributed research activity, including a fraction of the 200+ *Arabidopsis* research groups in UK together with their interdisciplinary collaborators, could establish a leading position in plant systems biology if it were suitably coordinated. China intends to push for GM crops in the next decade, and initiate budgets of billions of dollars. Scientists have strong willingness to collaborate with copartners in the entire world. We possess all the resources and talents necessary; what we now require is an international resolve to marshal the talents and facilities necessary to reach the goal of "*in silico*" crop plant.

Acknowledgements

We are grateful to BBSRC for funding this meeting, to many participants at the workshop, and to individuals who have contributed their efforts to this report. Related studies (to MC) are partially supported by grants from NSFC (No. 30771326), the 863 Program (No. 2008AA10Z125), and the Ministry of Education, China.

References

1. Chen, M. 2005. Systems biology brings life sciences closer. *Genomics Proteomics Bioinformatics* 3: 194-196.
2. James, N., *et al.* 2007. AtEnsEMBL. *Methods Mol. Biol.* 406: 213-227.
3. Upton, G.J., *et al.* 2008. G-spots cause incorrect expression measurement in Affymetrix microarrays. *BMC Genomics* 9: 613.
4. Broadley, M.R., *et al.* 2008. Evidence of neutral transcriptome evolution in plants. *New Phytol.* 180: 587-593.
5. *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796-815.
6. Yu, J., *et al.* 2001. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296: 79-92.
7. Goff, S.A., *et al.* 2001. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296: 92-100.
8. Ideker, T., *et al.* 2001. A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.* 2: 343-372.
9. Kitano, H. 2002. Systems biology: a brief overview. *Science* 295: 1662-1664.
10. Giffiths, K. and Resnick, R. 2000. Approaches to integrating biological data. Tutorial in the Eighth International Conference on Intelligent Systems for Molecular Biology, La Jolla, USA.
11. Draper, J., *et al.* 2001. *Brachypodium distachyon.* A new model system for functional genomics in grasses. *Plant Physiol.* 127: 1539-1555.
12. Opanowicz, M., *et al.* 2008. *Brachypodium distachyon*: making hay with a wild grass. *Trends Plant Sci.* 13: 172-177.