The Plant Cell, Vol. 13, 723, April 2001, www.plantcell.org

IN THIS ISSUE

Everything in Its Place: Conservation of Gene Order among Distantly Related Plant Species

The use of "model" species in biological research is based on the assumption that many of their features are shared among a wide range of related taxa. Thus, it is hoped that many of the genes associated with important traits in crop plants will be identified via homology with their counterparts in Arabidopsis. In addition to a high degree of conservation of individual gene sequences throughout the plant kingdom, comparative genomics has revealed a high degree of conservation in genome structure, or synteny, among closely related taxonomic groups. *Synteny*, from the Greek *syn* (together with) and *taenia* (ribbon), refers to loci contained within the same chromosome. In comparative genomics, it is often used as a synonym for colinearity (which is more properly *conserved synteny*) and refers to some degree of conservation of gene content, order, and orientation between chromosomes of different species or between nonhomologous chromosomes of a single species.

For example, gene content appears to be highly conserved with a remarkable degree of colinearity among the grasses, including the grain crops rice, wheat, maize, barley, sorghum, and millet. This is despite large differences in genome size (e.g., 430 Mbp in rice compared with 16,000 Mbp in wheat), which appear to be attributable principally to differences in the amounts of repetitive DNA (associated mainly with retroelements) in intergenic regions and polyploidy (Bennetzen and Freeling, 1997). Thus, rice appears to be an excellent model for the grasses (Chen et al., 1997; Devos and Gale, 2000).

Macrosynteny can be explored using genetic resources such as restriction fragment length polymorphism maps. Restriction fragment length polymorphism markers derived from one species are hybridized against genomic DNA from one or more related species to create a comparative map. Conserved macrosynteny can be observed over large genetic distances on linkage maps of closely related taxa, such as species or genera within the same family. Typically, macrosynteny can be explored only among relatively closely related species because of the limits of cross-hybridization of markers. However, the distance between two markers on a genetic map can comprise hundreds of genes, and macrosynteny between two species does not necessarily imply the existence of microsynteny or the conservation of local gene repertoire, order, and orientation. For example, Tarchini et al. (2000) found interruptions in microsynteny between rice and maize and cautioned that the use of rice as a model system for other cereals may be complicated by the presence of rapidly evolving gene families and microtranslocations. Conversely, conserved microsynteny is possible between species that lack obvious signs of macrosynteny.

The investigation of microsynteny requires sequencing and annotation of genomic DNA, enabling direct comparison of the sequences using various computational tools. Thus, the completed Arabidopsis genome sequence and growing lists of genomic resources for other plants have been an incredible boon to comparative genomics research.

Arabidopsis exhibits extensive conserved synteny with species from the closely related genera Brassica (O'Neill and Bancroft, 2000) and Capsella (Acarkan et al., 2000) (all three are in the family Brassicaceae). Comparisons of Arabidopsis with more distantly related species also have shown some degree of synteny. For example, significant synteny was reported between soybean (Leguminosae) and Arabidopsis along the entire length of Arabidopsis chromosome 1 and soybean linkage group A2, and blocks of synteny were found among other chromosomes as well (Grant et al., 2000). Soybean and Arabidopsis are estimated to have diverged from a common ancestor 92 million years ago (MYA), compared with 6 to 10 MYA for Arabidopsis and Capsella and 12 to 20 MYA for Arabidopsis/Capsella and Brassica.

Comparisons of Arabidopsis and rice have revealed some degree of synteny spanning the divide between monocotyledonous and dicotyledonous plants (Paterson et al., 1996). However, microcolinearity between Arabidopsis and rice appears to have eroded to the extent that the Arabidopsis genome will be of limited use for map-based gene prediction and isolation in the grasses (Devos et al., 1999). Due to its relatively small genome size, rice has become the model species for the grasses. The sequence of the rice genome was completed recently by two private companies (Syngenta [Basel, Switzerland] and Myriad Genetics [Salt Lake City, UT]) and is on the road to completion for public access (Eckardt, 2000). It will be interesting to see the extent of conserved synteny between rice and other monocot families.

Arabidopsis and tomato (Solanaceae) diverged from a common ancestor an estimated 112 to 156 MYA, which follows closely the divergence of dicotyledonous from monocotyledonous families, estimated at 130 to 200 MYA. Thus, comparisons of Arabidopsis and tomato should offer a snapshot of evolution since the introduction of dicotyledonous plants and provide information

IN THIS ISSUE

relevant to a wide range of families encompassed within the Arabidopsis-tomato clade, which includes legumes, Curcurbitaceae (melons), Rutaceae (citrus), Salicaceae (poplar), Malvaceae (cotton), Rosaceae, Asteraceae, and others (Ku et al., 2000).

Ku et al. (2000) found an extensive network of synteny between a 105-kb segment of tomato chromosome 2 and Arabidopsis chromosomes 2 to 5. They concluded that the dominating factors in the divergence of genome organization between these species have been repeated rounds of large scale genome duplication followed by selective gene loss. These authors predicted that complicated networks of conserved synteny would be common among higher plant families.

In this issue of *The Plant Cell*, Rossberg et al. (pages 979–988) report a remarkable degree of conserved microsynteny between Arabidopsis and tomato based on the sequencing and annotation of a 57-kb region of tomato chromosome 7 and a comparison of this region with the Arabidopsis genome. The tomato sequence chosen included the coding regions for five genes, which were used to search the Arabidopsis genome for similar sequences. Homologous sequences for all five of the genes were found within a 30-kb region corresponding to Arabidopsis chromosome 1. The intergenic regions were greatly expanded in tomato, and three of the five genes appeared in inverted orientation relative to Arabidopsis. The 30-kb region on chromosome 1 was found to be part of a larger segment that shows evidence of duplication on chromosome 3 (Blanc et al., 2000; The Arabidopsis Genome Initiative, 2000). However, only one of the seven genes in the 30-kb region (gene *C*) had a counterpart in the duplicated region on chromosome 3. These data are consistent with the hypothesis that the region of Arabidopsis chromosome 1 that retains synteny with tomato chromosome 7 underwent duplication followed by extensive gene loss within the Arabidopsis genome after divergence from the common ancestor with tomato.

Rossberg et al. also investigated Capsella for synteny with the 30-kb region of Arabidopsis chromosome 1 and identified a 27-kb contiguous overlapping sequence (contig) from eight cosmid clones that exhibited almost complete microcolinearity with the Arabidopsis sequence. Capsella homologs were found in this contig for all seven Arabidopsis genes located in the 30-kb region, all seven genes were in the same order and orientation, and the intergenic regions were of similar sizes. Interestingly, gene *C* from Arabidopsis chromosome 1 had greater similarity to Capsella gene *C* than to the homologous gene on Arabidopsis chromosome 3. Hybridization experiments suggested that Capsella also contains another region of homology that may be analogous to the region on Arabidopsis chromosome 3, indicating that duplication of the region that contains gene *C* most likely occurred before the divergence of the Arabidopsis and Capsella genera.

The widespread occurrence of gene duplication and the consequent proliferation of large gene families in plants leads to difficulties in determining orthology between species. Orthologous genes, or orthologs, are genes from different species that are derived from a common ancestor, whereas paralogs are genes within a species that arose from a duplication event. Conserved microsynteny is indicative of orthologs, although Bennetzen (2000) cautioned that microcolinearity can be deceiving in this respect because gene duplications often occur over short distances and can be followed by loss of the original parent gene. Thus, orthology can be defined unambiguously only if a high degree of colinearity is observed in the flanking regions of putative orthologs. These distinctions are important not only for establishing evolutionary histories of plant taxa but also for making assumptions about gene function. A high degree of gene sequence homology between individual genes from different species together with conserved microsynteny suggests true functional orthology, whereas a lack of colinearity (even if associated with a highly conserved individual coding region) is more likely to be indicative of a divergence of gene function and/or a rapidly evolving gene family. For example, Rossberg et al. found that one of the five genes on tomato chromosome 7, gene *D*, which has similarity to WRKY transcription factors, has numerous homologs in other regions of the Arabidopsis genome that were not found to be colinear with the segment identified on Arabidopsis chromosome 1. Thus, analyses of microsynteny also may help to characterize rapidly evolving gene families.

Analyses of many different eukaryotic genomes have revealed that polyploidization and gene duplication are widespread phenomena among eukaryotes. A number of simple diploid genomes, including those of yeast and Arabidopsis, appear to be derived from ancient polyploids. Numerous analyses of the Arabidopsis genome indicate that extensive “genome shuffling” has occurred, characterized by several rounds of large scale duplication followed by gene loss (Blanc et al., 2000; Grant et al., 2000; Ku et al., 2000). The work of Rossberg et al. provides additional evidence for this hypothesis. In fact, it has been suggested that all eukaryotes are derived from ancient polyploids, which had a tendency to evolve to a diploid state through sequence diversification and chromosomal rearrangement (Grant et al., 2000). Polyploidy is particularly widespread among extant plant species relative to other eukaryotes. Comparative genomics may begin to provide answers regarding the evolution and maintenance of the various states of ploidy observed among plants and other eukaryotes.

The work by Rossberg et al. fills an important gap in our knowledge of plant comparative genomics by demonstrating a remarkable degree of conserved microsynteny between distantly related (Arabidopsis and tomato) as well as closely related (Arabidopsis and *C. rubella*) dicotyledonous species. This study, together with that of Ku et al. (2000), suggests that segments of microcolinearity can be exploited to identify orthologous genes in Arabidopsis and distantly related dicots.

Nancy A. Eckardt
News and Reviews Editor

REFERENCES

Acarkan, A., Rossberg, M., Koch, M., and Schmidt, R. (2000). Comparative genome analysis reveals extensive conservation of genome organisation for *Arabidopsis thaliana* and *Capsella rubella*. Plant J. **23,** 55–62.

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

Bennetzen, J.L. (2000). Comparative sequence analysis of plant nuclear genomes: Microcolinearity and its many exceptions. Plant Cell **12,** 1021–1029.

Bennetzen, J.L., and Freeling, M. (1997). The unified grass genome: Synergy in synteny. Genome Res. **7,** 301–306.

Blanc, G., Barakat, A., Guyot, R., Cooke, R., and Delseny, M. (2000). Extensive duplication and reshuffling in the Arabidopsis genome. Plant Cell **12,** 1093–1101.

Chen, M.P., SanMiguel, P., de Oliveira, A.C., Woo, S.-S., Zhang, H., Wing, R.A., and Bennetzen, J.L. (1997). Microcolinearity in *sh2*-homologous regions of the maize, rice and sorghum genomes. Proc. Natl. Acad. Sci. USA **94,** 3431–3435.

Devos, K.M., Beales, J., Nagamura, Y., and Sasaki, T. (1999). *Arabidopsis*-rice: Will colinearity allow gene prediction across the eudicot-monodicot divide? Genome Res. **9,** 825–829.

Devos, K.M., and Gale, M.D. (2000). Genome relationships: the grass model in current research. Plant Cell **12,** 636–646.

Eckardt, N.A. (2000). Meeting Report: Sequencing the rice genome. Plant Cell **12,** 2011–2017.

Grant, D., Cregan, P., and Shoemaker, R.C. (2000). Genome organization in dicots: Genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*. Proc. Natl. Acad. Sci. USA **97,** 4168–4173.

Ku, H.-M., Vision, T., Liu, J., and Tanksley, S.D. (2000). Comparing sequenced segments of the tomato and *Arabidopsis* genomes: Large-scale duplication followed by selective gene loss creates a network of synteny. Proc. Natl. Acad. Sci. USA **97,** 9121–9126.

O'Neill, C.M., and Bancroft, I. (2000). Comparative physical mapping of segments of the genome of *Brassica oleracea* var. *alboglabra* that are homeologous to sequenced regions of chromosomes 4 and 5 of *Arabidopsis thaliana*. Plant J. **23,** 233–243.

Paterson, A.H., et al. (1996). Toward a unified genetic map of higher plants, transcending the monocot-dicot divergence. Nat. Genet. **14,** 380–382.

Rossberg, M., Theres, K., Acarkan, A., Herrero, R., Schmitt, T., Schumacher, K., Schmitz, G., and Scmidt, R. (2001). Comparative Sequence Analysis Reveals Extensive Microcolinearity in the *Lateral Suppressor* Regions of the Tomato, Arabidopsis, and Capsella Genomes. Plant Cell **13,** 979–988.

Tarchini, R., Biddle, P., Wineland, R., Tingey, S., and Rafalski, A. (2000). The complete sequence of 340 kb of DNA around the rice *Adh1-Adh2* region reveals interrupted colinearity with maize chromosome 4. Plant Cell **12,** 381–391.

The New Biology: Genomics Fosters a "Systems Approach" and Collaborations between Academic, Government, and Industry Scientists

Genomics is changing the face of biology. At first glance, it is largely a change of scale: we move from considering the function of one or a few genes to considering hundreds or thousands of genes at once. The technological advances that brought about this change of scale are leading to the rapid development of still more tools for experimentation and data analysis at the genomic level. We move beyond the genome to consideration of the *transcriptome*, *proteome*, and *metabolome* (all of the transcripts, proteins, and metabolites, respectively, within a cell, tissue, or organism). Looking deeper, we also find that this change of scale is having a dramatic effect on the structure of the scientific community, the manner in which scientific investigation is conducted (and by whom), and how information is disseminated. These topics were the focus of a recent Keystone Symposium sponsored by Monsanto

MEETING REPORT

Co. entitled "Systems Approach to Plant Biology," which took place at Big Sky, Montana, from January 26 to 31, 2001. Symposium organizers Joe Ecker (Salk Institute, La Jolla, CA) and Steve Briggs (Torrey Mesa Research Institute, San Diego, CA) brought together an eclectic group of plant and nonplant scientists from industry and academia at the forefront of genomics to present their research and discuss the future of plant biology. The symposium speaker list was designed both to stimulate further interest in plant systems among non-plant scientists and to inform the plant science community about some of the latest research in genomics from the non-plant world, with the hope of leading plant research in exciting new directions. Plant research already has begun to take the next steps "beyond the genome," as illustrated in numerous presentations given by plant scientists. This report highlights just a few of the presentations given by the 30 speakers and the ~50 posters presented at the symposium.

THE PROTEIN'S THE THING

In the Keynote Address, Roger Brent (The Molecular Sciences Institute, Berkeley, CA) tackled the issue of what is meant by a "systems approach" to biology and what might be learned from it. Many of his ideas are presented in a thought-provoking review of genomic biology published in the journal *Cell* (Brent, 2000). Brent introduced the topic with a summation of the history of biochemistry and molecular biology. In the 1950s, biochemistry was part of the industrial economy that was focused on how things are made and how they are broken apart, which Brent referred to as "Stoffwechsel," or "changing stuff into other stuff." Beginning in the 1990s, the emphasis shifted to decision making: when and where things are made and broken apart and how these events are regulated. Thus, contemporary cell and molecular biology is now part of the "information economy." The primary goal of research at The Molecular Sciences Institute is a "predictive" biology: the ability to make quantitatively accurate predictions about the consequences of environmental perturbations on a biological system. Genomic techniques will pave the way by providing a complete inventory of biological parts and information about how those parts work together to create system outcomes.

Of primary importance on the road to a predictive biology is information about protein–protein interactions, because it is proteins that do the work inside cells, and very often the work of proteins is accomplished via interactions with other proteins. Green fluorescent protein (GFP) and GFP derivatives provide a useful tool for exploring protein interactions. One method that has been used extensively with yeast is expression of GFP fusions with various mutagenized genes; such transformed yeast strains are then screened for deficiencies in protein–protein interactions. Another strategy discussed by Brent was the use of aptamers, which are synthetic nucleic acid ligands, typically composed of RNA, single stranded DNA, or a combination of these with nonnatural nucleotides, that can be generated for specific binding with proteins as well as amino acids, drugs, and other molecules. Theoretically, aptamer affinity agents could be generated against all of the products (polypeptides) of a genome and all of their modification states and used to detect protein functions and protein interactions. The understanding of complex biological systems gained from this type of approach will have important applications in plant as well as animal systems, for example, in the development of new herbicides and pesticides or in identifying proteins associated with quantitative trait loci that control agronomically important traits.

Protein–protein interactions also were the focus of a presentation by Marc Vidal (Dana Farber Cancer Institute, Boston, MA), who spoke about protein interaction mapping in *Caenorhabditis elegans*. His group is involved in building a resource called "ORFeome," which is a collection of all of the open reading frames (ORFs) in the *C. elegans* genome cloned into a specially designed vector (Gateway) that makes it easy to transfer clones into other "destination" vectors through recombination (Walhout et al., 2000; Reboul et al., 2001; Vidal, 2001). The ORF clones are being used to build complex maps of potential protein–protein interactions involved with various developmental processes through a combination of highly automated yeast two-hybrid screens ("interactome mapping") and RNA-interference (RNAi) suppression of gene expression ("phenome mapping").

Yeast two-hybrid schemes are based on the work of Fields and Song (1989), who proposed that cDNAs encoding proteins that interact with a known protein could be identified from a cDNA library if the library-encoded proteins all carried transcription activation domains. Many variations on this theme are in use today, but all are based on creating a "trap" for proteins that interact with a known protein of interest (protein X) whereby expression of a selectable marker gene (e.g., conferring the ability to grow on selective media) occurs only if X interacts with another protein carrying the appropriate transcription activation domain. RNAi works by producing a double stranded RNA molecule that targets the corresponding endogenous transcript for degradation by a sequence-specific nuclease complex (Bernstein et al., 2001).

Vidal presented an example of the use of these techniques to build a map of protein interactions associated with the DNA damage response and DNA repair to determine how communication and coordination of events are achieved between repair and cell cycle

MEETING REPORT

checkpoint proteins. ORFs showing high sequence similarity to genes known to be associated with DNA repair in other organisms were cloned into an RNAi vector, and a series of transformation experiments were conducted to "knock out" the corresponding endogenous genes in *C. elegans*. Results from the RNAi experiments were compared with results from yeast two-hybrid experiments to strengthen conclusions about protein–protein interactions, which is particularly useful because yeast two-hybrid experiments are prone to yielding false-positive results. Vidal compared the maps that have been generated thus far to maps of the Missouri River drawn by Lewis and Clark 200 years ago; they are rather crude, but they provide a lot of useful information and will become more accurate and detailed over time.

In addition to discovering and mapping the functions and interactions of proteins in living organisms, there is growing interest in designing new proteins to perform specific tasks. This fascinating topic was the focus of talks presented by Bassil Dahiyat (Xencor Inc., Monrovia, CA), Eileen Tozer (Diversa Corp., San Diego, CA), and Carlos Barbas (Scripps Research Institute, La Jolla, CA). Dahiyat and Tozer spoke about efforts at their respective companies to "evolve" new proteins with useful functions. At Xencor, Dahiyat and his team are involved in "in silico" screening of vast sequence libraries to design "better" proteins (e.g., proteins with improved stability, activity, specificity, potency, or expression). Their procedure uses propriety software to test all possible amino acid sequences for structural similarity to a protein of interest (Dahiyat, 1999). The software generates a manageable list of "best fit" sequences (ranked by an energy score), which are then tested and validated experimentally. Dahiyat described an example in which the structure of the active site of β-lactamase (which destroys β-lactam antibiotics such as penicillin and ampicillin) was screened in silico to direct the construction of a focused library containing 210 alternative sequences that gave the best fit to the original structure. These sequences were cloned into an expression library, the resulting proteins were screened for altered properties, and several mutants were isolated that yielded enhanced antibiotic resistance.

Diversa Corp. is also involved in evolving proteins with enhanced capabilities using their proprietary Gene Reassembly and Gene Site-Saturated Mutagenesis techniques, with an example given of a 30,000-fold improvement in the thermostability of a particular enzyme using the Gene Site–Saturated Mutagenesis technology. Additionally, the company has a discovery program that involves screening for naturally occurring proteins generated from DNA libraries consisting of nucleic acids isolated from uncultured microorganisms obtained from a variety of sites located around the world. Tozer stated that the company is negotiating agreements with many countries that involve benefits sharing, setting up laboratories, capacity building, and sustainable use of genetic resources.

Barbas is engaged in efforts to design transcription factors (in particular DNA binding zinc finger domains) that have the ability to manipulate the expression of endogenous genes. Such proteins could be used to study gene function or as therapeutic agents (i.e., to produce viral resistance in humans or plants). Zinc finger proteins have unique DNA binding characteristics that are conducive to the creation of a "universal" system for the control of gene expression (Segal et al., 1999). A single zinc finger domain, consisting of ~30 amino acid residues folded into a ββα configuration stabilized by chelation of a single zinc ion, binds 3 bp of DNA sequence in a sequence-specific manner. Recognition of longer DNA sequences can be achieved by covalent tandem repeats of multiple zinc finger domains. Theoretically, it should be possible to construct domains for the specific recognition of codons (64 are required for a complete codon recognition alphabet) and to construct a set of polydactyl zinc finger proteins (each having six domains) for the specific recognition of 18-bp sequences, which are sufficient to describe a unique DNA address within all known genomes. Barbas's group prepared a phage display library of more than 1 billion zinc finger proteins, which was "queried" for binding of unique codons against a pool of all 64 codons. Bound complexes were separated from pools using streptavidin-coated beads, binding specificities were measured in an ELISA format, and further refinement of proteins was made using site-directed mutagenesis to produce proteins that bound unique codons. They have completed construction of zinc finger domains that recognize all 32 codons containing either guanine or adenine in the 5′ position (i.e., 5′-GNN-3′ and 5′-ANN-3′). Segal et al. (1999) describe construction of the 16 domains that specifically recognize 5′-GNN-3′ codons.

Barbas's group also has begun to build polydactyl zinc fingers for specific recognition of 18-bp sequences and to design such proteins for the manipulation of endogenous gene expression by fusing them with effector domains (e.g., transcriptional activation or repression domains). For example, polydactyl zinc finger proteins were made that specifically target the mammalian protooncogenes *erbB-2* and *erbB-3*. Although the target sequences in the two genes were identical in 15 of 18 bp, zinc finger proteins were constructed that could selectively regulate (either induce or repress) the expression of either gene (Beerli et al., 2000). In another example involving a plant gene, transgenic Arabidopsis plants were constructed that overexpressed a zinc finger protein targeted against the endogenous *AP3* gene, a floral homeotic gene required for stamen and petal development. The

MEETING REPORT

resulting phenotype was similar to that observed in *AP3* knockout mutants found using more traditional methods, demonstrating the potential of this system for targeted gene regulation in plant systems. The Arabidopsis work was completed in collaboration with Xueni Guan at the Torrey Mesa Research Institute.

BEYOND THE GENOME IN PLANT SYSTEMS

A number of groups have developed large collections of T-DNA or transposon insertion lines in Arabidopsis in efforts to produce mutations for every gene. Many of these lines are publicly available through the Arabidopsis Functional Genomics Consortium (http://afgc.stanford.edu/), the Arabidopsis Biological Resource Center (http://aims.cps.msu.edu/aims/), and the Nottingham Stock Center (http://nasc.nott.ac.uk/). Jose Alonso (Salk Institute, La Jolla, CA) described a new collection of T-DNA insertion mutants of Arabidopsis developed with the laboratory of Joe Ecker (Salk Institute) in collaboration with William Crosby (National Research Council, Saskatoon, Canada). The group is creating a database of DNA sequences flanking the T-DNA inserts; 6000 lines have been sequenced to date. Their results revealed that the T-DNAs are distributed randomly in the Arabidopsis genome. More than 4000 gene knockouts were obtained in their pilot high sequencing study. Unlike in numerous other T-DNA or transposon insertion collections, each line is maintained individually, and the same amount of seed from each line is pooled to create superpools for polymerase chain reaction (PCR)-based screening to find insertions in particular genes of interest. The advantages of maintaining individual lines are equal representation of seed in the screening pools (as opposed to collecting seed from a pool of plants) and ease of verifying knockout mutations (i.e., only a few progeny need to be tested from any individual line). Based on genetic analysis of the mutant lines, the collection has more than 225,000 insertions, and 50% of the lines contain a single insertion. This collection should provide another valuable resource for the scientific community.

A couple of interesting and novel mutant screening techniques were introduced by Nick Carpita (Purdue University, West Lafayette, IN) and Richard Trethewey (Metanomics, Berlin, Germany). Carpita described a high throughput screen for mutants of cell wall biogenesis developed in collaboration with Maureen McCann (John Innes Centre, Norwich, UK). Cell type and function are defined by the cell wall, and although we know the polymers that make up cell walls in a wide variety of plant species, we have little knowledge about wall construction. Developing models of cell wall construction must take into account the variety of cell wall constituents that occurs among different plant groups. Carpita's group elected to focus on Arabidopsis and maize, because they represent the most widely distinct cell walls among angiosperms. They have developed a high throughput screen using Fourier transform infrared (FTIR) microspectroscopy to uncover mutant phenotypes resulting from T-DNA or transposon insertions into genes already known to be wall biogenesis related and to identify novel cell wall mutants (Chen et al., 1998). Cell walls are isolated in a rapid procedure at different times during wall synthesis, and FTIR spectra are collected and stored for later analysis. The group is creating a library of FTIR spectra to classify different cell wall mutant "spectrotypes" and will create a publicly accessible database.

Trethewey is involved in metabolic profiling of Arabidopsis in collaboration with Lothar Willmitzer (Max Planck Institute, Potsdam, Germany). This group created a T-DNA knockout population and uses gas chromatography and mass spectrometry in a high throughput format to create metabolic profiles of mutant lines and different ecotypes of Arabidopsis. They can reliably detect 325 different metabolites, and they have found that statistical analysis of mutant profiles can be predictive and that different ecotypes of Arabidopsis have completely different metabolomes. For example, principal component analysis of the metabolic profiles of two homozygous ecotypes (Col-2 and C24) and a mutant of each ecotype (*dgd1* in the Col-2 background, which is characterized by a 90% reduction in the galactolipid digalactosyldiacylglycerol, and *sdd1-1* in the C24 background, which carries a mutation in a regulatory gene involved in the control of stomatal development) showed that the mutants could be distinguished readily from their respective parental lines. Furthermore, the metabolic phenotypes of the two ecotypes were more divergent than the metabolic profiles of the mutants and their respective parental ecotypes (Fiehn et al., 2000). Thus, metabolic profiling represents another valuable genomic tool. The groups at Metanomics and the Max Planck Institute are working on further development of the technology, such as combining metabolic profiling with radioisotope labeling tracer flux analysis and resolving metabolite levels in different organelles.

Insertional mutation also can be designed specifically to isolate particular types of mutations, which was nicely illustrated in talks given by Detlef Weigel (Salk Institute) and David Ehrhardt (Stanford University, Palo Alto, CA). One of the principal disadvantages of loss-of-function mutational screens is that many plant genes belong to gene families that have overlapping functions, and it may be necessary to knock out multiple genes to produce a phenotype. Gain-of-function mutational screens often can overcome this difficulty. Weigel described an activation tagging approach being used by his

MEETING REPORT

group in collaboration with Joanne Chory (Salk Institute), Chris Lamb (John Innes Centre, Norwich, UK), Marty Yanofsky (University of California, San Diego), and Maria Harrison (Noble Foundation, Ardmore, OK). Arabidopsis was transformed with T-DNA carrying a selectable marker (e.g., Basta) and four copies of the *Cauliflower mosaic virus* 35S promoter. If the insert falls upstream (or sometimes downstream) of a particular gene, its transcription can be activated, causing a gain-of-function phenotype. Several tens of thousands of lines were screened primarily for abnormal morphology. Weigel discussed several mutants that demonstrate the usefulness of this method, including the brassinosteroid biosynthetic mutant *bas1-D* (Neff et al., 1999), the *pap1-D* mutant, which has a bright purple phenotype caused by overexpression of a MYB transcription factor (Borevitz et al., 2000), and the *yucca* mutant, which overexpresses a flavin monooxygenase–like enzyme that encodes a key step in auxin biosynthesis (Zhao et al., 2001). Another mutant, *top1-D*, which was characterized in collaboration with Marty Yanofsky, causes overexpression of a transcription factor that affects fruit development. In each of these cases, the activated gene belonged to a multigene family and knockout mutations either have not been described or were found to exhibit weak or subtle phenotypes (if any), likely due to the existence of other closely related genes with overlapping functions.

Ehrhardt described a novel technique for creating "localization tags" using GFP::cDNA fusions. A large collection of transgenic lines of Arabidopsis was created by transformation with a cDNA library constructed of cDNAs fused to GFP, and fluorescence microscopy was used to screen for alterations in the normal pattern of GFP distribution to the cytoplasm and nucleoplasm of Arabidopsis cells (Cutler et al., 2000). Of 5700 transgenic seedlings screened, 120 (~2%) exhibited heritable non-wild-type GFP distribution patterns. These were sorted into a number of phenotypic classes based on the appearance of GFP localization, such as endoplasmic reticulum membrane, vacuolar membrane, chromosomes, "Q-balls," and "tiny bubbles." The largest class (43 lines) belonged to a group called "torus," which showed GFP fluorescence localized to a number of motile organelles with a variety of forms that changed shape and often formed little linear extensions. Their appearance was similar to that of peroxisome-localized markers, the cDNA markers in this class were almost always found to contain "out-of-frame" fusions, and a number of the torus cDNA sequences were found to encode a short string of amino acids at the C terminus that is similar to known peroxisome-localized proteins. These data suggest that peroxisomes are targeted with relatively high frequency by chance, and Ehrhardt warned that other researchers should be careful to ensure that their own GFP fusions are not inadvertently targeted to the peroxisome. This approach provides a visualization tool for subcellular structure and dynamics that may give clues to gene functions, identify novel cellular compartments, aid in the isolation of organelles, and identify useful proteins to target for genetic engineering. This interesting collection of subcellular localization markers includes one that showed a dynamic subcellular response to wounding, a possible marker for plasmodesmata, and a possible cytokinetic intermediate. Information on this project and on obtaining clones and transgenic lines can be found at http://deepgreen.stanford.edu.

THE RICE GENOME AND BEYOND

On January 26, Syngenta and Myriad Genetics announced the completion of sequencing of the rice genome. Syngenta, formed in November 2000, is the merger product of the Novartis and Zeneca agribusinesses. Whole genome shotgun sequencing of the rice genome was completed by Myriad Genetics, and sequence assembly and genome mapping was performed at the Torrey Mesa Research Institute (TMRI; a subsidiary of Syngenta). A number of TMRI scientists were at the meeting to report on the company's projects in rice genomics. Steve Goff (TMRI, San Diego, CA) reported on the completed rice genome sequence. Whole genome shotgun sequencing does not have as high a degree of accuracy as the "BAC-by-BAC" approach currently being used in a public rice genome project (led by the Rice Genome Project [RGP] based in Tsukuba, Japan), but it has the advantage of producing results more rapidly and at lower cost. The public rice genome sequencing project was described recently by Eckardt (2000), and current information on the project's status can be found at http://rgp.dna.affrc.go.jp/Seqcollab.html. Goff reported that TMRI was interested in sequencing rice because of the small genome and the large degree of conserved synteny with other cereals and because information coming from other genome projects (e.g., human and Arabidopsis) suggests that only a fraction of the expressed genes typically are discovered through cDNA sequencing. The sequence was assembled from 7.1 million sequencing runs. Repeats and organellar sequences were removed, and the assembly of linked contiguous minimally overlapping fragments was aided by the use of ~120,000 bacterial artificial chromosome (BAC) end sequences obtained from the Clemson University Genomic Institute (funded by Syngenta, then Novartis) and 332,000 SAGE (*s*erial *a*nalysis of *g*ene *e*xpression) tags generated at TMRI.

MEETING REPORT

Goff reported that the assembled sequence is estimated to cover 99% of the genome, based on a comparison of Syngenta's sequence with the sequences of six BACs completed by the public RGP. Preliminary analysis of the genome suggests that there are ~50,000 rice genes, although Goff cautioned that this was a rough estimate and that better "rice-trained" gene prediction programs are needed. Of the 27,421 predicted proteins in the Arabidopsis genome, ~80% showed homology with rice, and 3% exhibited a very high degree of homology. Approximately one-third of the predicted proteins in Arabidopsis (8800) may be plant specific, because they show homology within the rice genome but not within any other non-plant species genome in public databases. There are no clear regions of conserved macrosynteny between Arabidopsis and rice, and analysis of 17 Arabidopsis genes flanking the *FRIGIDA* locus shows no evidence of microsynteny with rice. However, there are large regions of conserved synteny between rice, maize, barley, and wheat, and rice should prove to be an excellent model for these and other grain crops.

Meanwhile, the RGP has renewed its commitment to complete the public sequence of the rice genome in the shortest possible time. Junshi Yazaki (Institute of the Society of Techno-Innovation of Agriculture, Forestry, and Fisheries, Tsukuba, Japan) reported on an RGP rice functional genomics project. This group has constructed cDNA microarrays using ~9000 cDNA clones obtained from different organs (e.g., developing seed, seedling, panicle, root, etc.). Currently, the RGP is focusing on two major areas, stress responses and tissue- or organ-specific gene expression, with the goal of establishing a transcriptional database for rice. Further information on this project can be found at http://microarray.rice.dna.affrc.go.jp.

GENOME-WIDE GENE EXPRESSION ANALYSIS

Oligonucleotide and cDNA microarrays are used widely for large scale gene expression analysis. Stacey Harmer (Scripps Research Institute, La Jolla, CA) presented a good example of the value of genome-wide expression analysis from experiments conducted in collaboration with the Torrey Mesa Research Institute. Affymetrix "GeneChip" Arabidopsis oligonucleotide arrays comprising 8200 genes were used to investigate circadian clock–controlled gene expression (Harmer et al., 2000). As expected, many types of photosynthetic genes were found to be clock regulated. Clock-regulated gene clusters also were identified for other processes, which, although not entirely surprising, have not been documented so extensively. For example, a group of 12 genes associated with nitrate and sulfate assimilation exhibited a peak in expression levels at dawn. Harmer noted that nitrate and sulfate assimilation require a lot of reducing power, so it may be advantageous to plants for genes associated with these processes to be clock regulated for activity during the day. Other clock-controlled genes involved in metabolism include a cluster of genes associated with starch mobilization (e.g., a putative starch kinase, fructose bisphosphate aldolases, and sucrose phosphate synthase) that peak during the subjective night, a time when plants must rely on stored sugars. Harmer's group also identified a novel nine-nucleotide clock-controlled promoter element, called the "evening element," which was present in the upstream region of 31 clock-controlled genes whose expression peaked at the end of the day. Mutation of this conserved sequence was found to disrupt circadian rhythmicity, and fusion of a promoter element that contained this sequence with the luciferase reporter gene was sufficient to induce rhythmicity in luciferase expression. An important observation about Harmer's research was that the expression of ~6% of the genes on the array changed significantly within the 4-hr sampling intervals of the experiments conducted. Researchers who conduct large time course experiments, for which sampling for a "single time point" could take up to a few hours, might wish to design experiments accordingly.

Some methods of analyzing gene expression on a genome-wide scale do not rely on previous knowledge of gene sequence. Marc Zabeau (University of Ghent, Belgium) discussed the use of one such method, cDNA-AFLP (for amplified fragment length polymorphisms), for transcript profiling of plant cell cycle genes. The plant cell cycle includes processes that are not found in mammals or yeast, such as endoreduplication and synthesis of a rigid cell wall. Approximately 100 cell cycle regulation genes have been identified in plants thus far, and Zabeau's group hopes to characterize many more using genomic technologies. The cDNA-AFLP technique involves construction of a cDNA library that is subjected to restriction enzyme digestion, adapter ligation, and selective PCR amplification. Zabeau's group has modified the standard procedure by preparing cDNA biotinylated at the 3′ end, followed by digestion with one restriction enzyme and capture of the 3′ ends with streptavidin-coated beads, and then digestion with a second restriction enzyme followed by adapter ligation. This reduces the complexity of the sample by producing just one diagnostic fragment per transcript. They chose to use the tobacco cell line BY2 to investigate cell cycle control because, unlike Arabidopsis cell lines, the cell cycle in this system can be synchronized. Tobacco BY2 cells were synchronized by treatment with aphidicoline, and samples were collected for cDNA-AFLP analysis at 12 time points after release from aphidicoline blocking. Approximately

20,000 AFLP tags were generated, and 1487 were identified as candidates for cell cycle–modulated gene tags, including many corresponding to rare abundance transcripts. Clusters of genes were identified that showed cell cycle–modulated expression. For example, the S-phase gene cluster included DNA and RNA polymerases, the G2/M transition period was marked by the expression of cyclins and cyclin-dependent kinases, and the M-phase cluster included pectinesterases and ubiquitin. More than 1100 cell cycle–modulated tags were sequenced, and 780 were found to correspond to novel or previously uncharacterized genes. Work is in progress to isolate full-length cDNAs corresponding to these genes, and the group is progressing toward a comprehensive catalog of cell cycle genes.

GENE EXPRESSION ANALYSIS: THE NEXT STEP

Some groups are recognizing a need to develop new techniques for high throughput monitoring of gene expression on a cellular or subcellular level during different stages of development. Hinanit Koltai (North Carolina State University, Raleigh) presented a method for high throughput, in situ PCR-based localization of gene transcripts (Koltai and Bird, 2000). The method involves tissue sectioning followed by the transfer of tissue subsections to a liquid phase in a 96-well format for PCR amplification of specific transcripts. The 96-well format permits robotic processing of large numbers of samples, and labeled PCR products can be recovered from the 96-well supernatant and sequenced to verify the specificity of the histological staining pattern. Koltai presented results from experiments conducted to demonstrate the utility of this method using tissues from Arabidopsis, *Medicago truncatula*, and tomato. For example, transcripts of the *PHANTASTICA* and *KNOTTED* (*Tkn2*) genes in tomato tissue colocalized to the same region of shoot and root meristems. Additional features of *Tkn2* expression were noted that were not detected previously, such as expression in mature leaves. The method also was used to monitor the expression of GFP reporter constructs and to confirm the ablation of target gene expression in transgenic "antisense" plants.

June Medford (Colorado State University, Fort Collins) described a method for in vivo imaging of plants based on optical coherence microscopy (OCM). An OCM signal is produced by differential photon reflectance caused by variations in refractive indices within plant tissue. Compared with a light micrograph image, the OCM signal shows very little signal from the cell wall. An OCM image typically shows a three-dimensional view of the cytoplasm, nucleus, and organelles without the surrounding cell walls. Visualization can be obtained through 1 mm of tissue (far beyond the range of confocal microscopy), and many features of a live plant growing in soil (e.g., developing ovules, stomata, flower parts, pollen) can be viewed in a completely noninvasive manner. Medford's group has designed a platform for growing Arabidopsis in special trays and robotics for moving plants into position for imaging, which will enable the measurement of at least 500 plants per day. They are working on further modifications of the system to allow for in vivo monitoring of gene expression by designing a reporter molecule that will produce a strong OCM signal. Polyhydroxybutyrate (PHB) was identified as a good candidate for such a reporter, and Medford showed images of plants expressing PHB directed to the chloroplast under the control of the *Cauliflower mosaic virus* 35S promoter. PHB was found to give a good OCM signal, and gene expression images from transgenic plants were consistent with chloroplast localization. The group plans to explore the use of other promoters and other potential reporter molecules.

INDUSTRY AND ACADEMIA: BUILDING BRIDGES

There is a growing commercialization of the biological sciences. Interestingly, the research focus in industry has shifted perceptibly from product development to technology development and science. Steve Briggs spoke about the need to build bridges between academic and industry scientists involved in genomic research, because each of these worlds has something the other needs. The goal of geneticists in both industry and academic institutions is to match genes with traits. There is a tremendous amount of diversity in biological systems, such that no company can cover all of "biology." However, companies can cover all of "genomics" (i.e., genomics technologies), whereas the scale and cost of genomics often is prohibitive for academic researchers. Briggs said the answer, therefore, is for genomics companies to interact and collaborate with a lot of biologists. Briggs and Xun Wang (TMRI) spoke about genomics research at TMRI, which is designed to support collaborations between industry and academic scientists. The research approach at TMRI is to combine geneticists with analytical chemists and computer scientists to solve genomics research problems. There are two "discovery groups" (Plant Health and Consumer Health) that interact with different "technology groups" in structural genomics, RNA dynamics, bioinformatics, and protein and metabolite dynamics. The institute actively solicits proposals from academic scientists for collaborative research. This arrangement allows the biologists (both academic collaborators and company researchers) to focus on their specialty, biology, instead of spending their time on bioinformatics

MEETING REPORT

and high throughput technology development. Research presented by Stacey Harmer (see above) is an example of a collaboration between TMRI and academic scientists. Briggs maintained that, so far, the institute has not turned down a single proposal from academic researchers, and it welcomes collaboration with small laboratories and young scientists. TMRI is committed to sharing scientific recognition and commercial benefits with academic collaborators and to rapid patent filing and publication of results. More information can be found at http://www.nadii.com/. TMRI may provide a model for how industry–academic collaborations in genomic research can best proceed in the future.

Nancy A. Eckardt
News and Reviews Editor

REFERENCES

Beerli, R.R., Dreier, B., and Barbas III, C.F. (2000). Positive and negative regulation of endogenous genes by designed transcription factors. Proc. Natl. Acad. Sci. USA **97,** 1495–1500.

Bernstein, E., Caudy, A.A., Hammond, G., and Hannon, G.J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature **409,** 363–366.

Borevitz, J.O., Xia, Y., Blount, J., Dixon, R.A., and Lamb, C. (2000). Activation tagging identifies a conserved MYB regulator of phenylpropanoid biosynthesis. Plant Cell **12,** 2383–2394.

Brent, R. (2000). Genomic biology. Cell **100,** 169–183.

Chen, L.M., Carpita, N.C., Reiter, W.D., Wilson, R.H., Jeffries, C., and McCann, M.C. (1998). A rapid method to screen for cell-wall mutants using discriminant analysis of Fourier transform infrared spectra. Plant J. **16,** 385–392.

Cutler, S.R., Ehrhardt, D.W., Griffitts, J.W., and Somerville, C.R. (2000). Random GFP::cDNA fusions enable visualization of subcellular structures in cells of *Arabidopsis* at high frequency. Proc. Natl. Acad. Sci. USA **97,** 3718–3723.

Dahiyat, B.I. (1999). In silico design for protein stabilization. Curr. Opin. Biotechnol. **10,** 387–390.

Eckardt, N.A. (2000). Meeting Report: Sequencing the rice genome. Plant Cell **12,** 2011–2017.

Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R.N., and Willmitzer, L. (2000). Metabolite profiling for plant functional genomics. Natl. Biotechnol. **18,** 1157–1161.

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

Harmer, S.L., Hogenesch, J.B., Straume, M., Chang, H.-S., Han, B., Zhu, T., Wang, X., Kreps, J.A., and Kay, S.A. (2000). Orchestrated transcription of key pathways in Arabidopsis by the circadian clock. Science **290,** 2110–2113.

Koltai, H., and Bird, D.M. (2000). High throughput cellular localization of specific plant mRNAs by liquid-phase in situ reverse transcription-polymerase chain reaction of tissue sections. Plant Physiol. **123,** 1203–1212.

Neff, M.M., Nguyen, S.M., Malancharuvil, E.J., Fujioka, S., Noguchi, T., Seto, H., Tsubuki, M., Honda, T., Takatsuto, S., Yoshida, S., and Chory, J. (1999). BAS1: A gene regulating brassinosteroid levels and light responsiveness in Arabidopsis. Proc. Natl. Acad. Sci. USA **96,** 15316–15323.

Reboul, J., et al. (2001). Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in *C. elegans*. Nat. Genet. **27,** 332–336.

Segal, D.J., Dreier, B., Beerli, R.R., and Barbas III, C.F. (1999). Toward controlling gene expression at will: Selection and design of zinc finger domains recognizing each of the 5′-GNN-3′ DNA target sequences. Proc. Natl. Acad. Sci. USA **96,** 2758–2763.

Vidal, M. (2001). A biological atlas of functional maps. Cell **104,** 333–339.

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

Zhao, Y., Christensen, S.K., Fankhauser, C., Cashman, J.R., Cohen, J.D., Weigel, D., and Chory, J. (2001). A role for flavin monooxygenase-like enzymes in auxin biosynthesis. Science **291,** 306–309.