UNRAVELING THE GENOME OF RICE

Stephen A. Goff from Syngenta's Torrey Mesa Research Institute in San Diego in California, USA, describes the work which has led to the elucidation of the rice genome

Introduction

Approximately half of the world's population relies on rice for a majority of their caloric intake. Over the next several decades, rice consumption and demand is expected to increase due to human population growth. Since little new acreage is available to increase rice cultivation, higher yields will be needed to meet this anticipated demand, and at the same time rice improvement and nutritional enhancement will need to accelerate. A better understanding of the genes controlling rice development, grain yield, and grain quality will facilitate necessary improvements in this staple crop.

In addition to rice being an important cereal crop for human consumption, it has become clear that rice is a model for other important crops like corn, wheat, and barley. The rice genome is considerably smaller than the genomes of the other major cereal crops. The size of the rice genome is estimated at 420-450 million base pairs. For reference, the human genome is approximately 3 billion base pairs. Maize, barley, and wheat also have significantly larger genomes of 3, 5, and 16 billion base pairs respectively. Although the genome sizes of these important crops vary significantly, it is believed that each of these major crops will have essentially the same complement of genes located in similar positions relative to one another. Many rearrangements have taken place since these crops diverged evolutionarily fifty-to-seventy million years ago. However, comparisons of the physical and genetic maps of cereal genomes have led to the conclusion that a significant amount of co-linearity of gene order exists among the various cereal genomes studied (Ahn *et al*., 1993; Bennetzen and Freeling, 1993; Chen and SanMiguel, 1998; Chen *et al*., 1997; Ahn and Tanksley, 1993; Moore *et al*., 1995; Moore, 1995; Moore *et al*., 1997; Van Deynze *et al*., 1995a; Van Deynze *et al*., 1995b) Genes found in all organisms sequenced to date are related to previously discovered genes or novel. In the case of rice about one-third of the genes encode proteins similar to ones found only in the recently sequenced model plant Arabidopsis. Many rice genes encode proteins found in both animal and bacterial cells, serving both enzymatic and regulatory functions.

The smaller genome size of rice results in a higher gene density relative to the other cereals, and makes rice an attractive target for gene discovery efforts and complete genome sequencing. In a collaborative effort with Myriad Genetics, Syngenta's Torrey Mesa Research Institute has recently completed the draft sequencing of the rice genome. This project is the crop equivalent of the much-publicized human genome project recently reported in the journals *Science* (Venter *et al*., 2001) and *Nature* (International Human Genome Sequencing Consortium, 2001). A publicly funded project to sequence the rice genome to a high degree of accuracy is also underway, and is expected to be complete within a few years. The public rice genome project was reviewed in *Pesticide Outlook* in 1999 (Sasaki, 1999).

Generating the rice physical map

The genetic material (DNA) from rice was isolated and fragmented into hundreds of thousands of small (1000–2000 base pairs) pieces and amplified in bacteria. Several million DNA sequencing reactions were generated using robotic liquid handlers and processed on state-of-the-art automated DNA sequencing instruments. The resulting sequences were used to assemble the rice genome. This approach is called a "shotgun" sequencing strategy, and has recently proven efficient for large genome organisms such as the fruit fly (Adams *et al*., 2000; Myers *et al*., 2000) and the human genome (Venter *et al*., 2001; International Human Genome Sequencing Consortium, 2001). The sequence information covers the entire rice genome approximately six-fold, a redundancy slightly higher than that used to generate the human genome draft sequence (Venter *et al*., 2001). DNA fingerprinting techniques similar to those used in forensic medicine or paternity testing were used to align the rice draft sequence information to the known genetic map. The resulting rice genome map is estimated to cover greater than 99% of the rice genome when compared to reference sequence available in public databases. Rice is the second plant to be sequenced to this degree of completion and the first crop plant to be sequenced. In December 2000, the completed sequence of the model plant *Arabidopsis thaliana* was reported (The *Arabidopsis* Genome Initiative, 2000). *Arabidopsis* is a dicot in the mustard family, and is distant relative of the monocot rice.

Discovering the rice genes

A variety of data supports the estimate of 50,000 genes in the complete rice genome. This figure is consistent with previous estimates, although the final precise number of genes in rice will require many years of additional confirmation work. Analysis of the *Arabidopsis* genome resulted in approximately 25,000 predicted genes (The *Arabidopsis* Genome Initiative, 2000). *Arabidopsis* shares approximately 9000 genes with rice that have not yet been found in other species and are likely to be involved in plant-specific

DOI: 10.1039/b106290p

functions. A few thousand predicted genes from *Arabidopsis* are not found in rice.

Studying rice gene expression

The predicted genes of rice were used to construct a rice gene "microarray" which is being used for gene expression analysis. This microarray consists of several hundred thousand miniature spots capable of signaling how strongly each gene is expressed in a given sample. These rice gene microarrays have been used to study gene expression in different stages of rice development and in various rice tissues such as the filling grain. This technology allows researchers to determine which genes of the plant are turned on in response to pathogen infection, drought conditions, low or high temperature, insect predation, and a number of other environmental conditions that all plants must face. Identifying the genes expressed under these conditions and stages of development will allow researchers to determine which genes are likely to enhance the plant's ability to remain healthy and provide the highest yield. Accelerated traditional breeding or genetic enhancements via biotechnology become possible once these gene functions are clearly understood. Relatively small changes in gene expression can result in dramatic phenotypic consequences. For example, the teosinte branched mutant of maize converts the plant toward the multiple stalk morphology found in teosinte. This morphological change is due to an expression difference of a regulatory protein rather than an altered protein (Wang *et al*., 1999).

Studying rice proteins and metabolites

Recent advancements in the analysis of proteins and metabolites have been added to the suite of technologies used to study rice and other cereal crops. Coupling rapid analytical techniques such as liquid chromatography (LC) and mass spectrometry (MS) with software capable of identifying proteins from the genome sequence information is allowing researchers to study plant responses to biotic and abiotic factors such as pathogen infection and chemical treatments.

These approaches are also allowing researchers to study the tissue distribution and subcellular localization of specific proteins and small molecules. These advancements in the field of "proteomics" and "metabolite profiling" are being applied to further understand plant development, responses to the environment, and responses to application of crop protection compounds. LC and MS technologies will further bring together crop breeding and molecular analysis to enhance the nutrition, increase the yields, and tailor the responses of crops to specific chemical applications.

Summary and future outlook

With the completion of the model plant *Arabidopsis* genome and now the completion of the rice genome, comparative analysis of plants is poised to move ahead rapidly. Molecular approaches have been developed to further explore crop improvement. The developed technologies will enable higher crop yields and more stress-tolerant crops as the human genome sequence will be used to improve human health, disease detection, and disease prevention. In the next few years, specific genes and metabolic or regulatory pathways will be elucidated. Avenues toward crop improvement will be opened, and traditional plant breeding will adapt molecular analysis technology to enhance crop varieties. Genetic diversity in wild plants will become accessible, and both transgenic as well as accelerated traditional breeding will be used to maintain a pace of crop improvement able to provide for a growing population.

References

Adams, M. D.; *et al.* (2000). The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195.

Ahn, S.; Anderson, J.A.; Sorrells, M. E.; Tanksley, S. D. (1993). Homoeologous relationships of rice, wheat and maize chromosomes, *Mol. Gen. Genet*. **241**, 483–490.

Ahn, S.; Tanksley, S. D. (1993). Comparitive linkage maps of the rice and maize genomes. *Proceedings of the National Academy of Sciences*., **90**, 7980–7984.

Bennetzen, J. L.; Freeling, M. (1993). Grasses as a single genetic system: genome composition, colinearity and compatibility. *Trends Genet*. **9**, 259–261.

Chen, M.; SanMiguel, P.; Bennetzen, J. L. (1998). Sequence organization and conservation in *sh2/a1*-homologous regions of sorghum and rice. *Genetics* **148**, 435–443.

Chen, M.; SanMiguel, P.; de Oliveira, A. C.; Woo, S. S.; Zhang, H.; Wing, R. A.; Bennetzen, J. L. (1997). Microcolinearity in *sh2* homologous regions of the maize, rice, and sorghum genomes. *Proeedings of the National Academy of Sciences,* **94**, 3431–3435

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Moore, G.; Foote, T.; Helentjaris, T.; Devos, K.; Kurata, N.; Gale, M. (1995). Was there a single ancestral chromosome? *Trends in Genetics* **11**, 81–82.

Moore, G. (1995). Cereal genome evolution: pastoral pursuits with 'Lego' genomes. *Current Opinion in Genetics & Development* **5**, 717–24.

Moore, G.; Aragon-Alcaide, L.; Roberts, M.; Reader, S.; Miller, T.; Foote, T. (1997). Are rice chromosomes components of a holocentric chromosome ancestor? *Plant Molecular Biology*, **35**, 17-23.

Myers, E. W.; *et al.*, (2000). A whole genome assembly of *Drosophila*. *Science* **287**, 2196–2204.

Sasaki, T. (1999) The rice genome project. *Pesticide Outlook*, **10**, 114–116.

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

Van Deynze, A. E.; Nelson, J. C.; O'Donoughue, L. S.; Ahn, S. N.; Siripoonwiwat, W.; Harrington, S. E.; Yglesias, E. S.; Braga, D. P.; McCouch, S. R.; Sorrells, M. E. (1995a). Comparative mapping in grasses. Oat relationships. *Molecular & GeneralGenetics,* **249**, 349–356.

Van Deynze, A. E.; Nelson, J. C.; Yglesias, E. S.; Harrington, S. E.; Braga, D. P.; McCouch, S. R.; Sorrells, M. E. (1995b). Comparative mapping in grasses. Wheat relationships. *Molecular & General Genetics* **248**, 744–754.

Venter, J. C.; *et al.* (2001). The sequence of the human genome. *Science* **291**, 1304–1351.

Wang, R. L.; Stec, A.; Hey, J.; Lukens, L.; Doebley, J. (1999). The limits of selection during maize domestication. *Nature*, **398**(6724), 236–239.