The International Rice Genome Sequencing Project: progress and prospects

T. Sasaki, T. Matsumoto, T. Baba, K. Yamamoto, J. Wu, Y. Katayose, and K. Sakata

The rice genome sequencing project has been pursued as a national project in Japan since 1998. At the same time, a desire to accelerate the sequencing of the entire rice genome led to the formation of the International Rice Genome Sequencing Project (IRGSP), initially comprising five countries. The sequencing strategy is the conventional clone-by-clone shotgun method using P1-derived artificial chromosome/bacterial artificial chromosome (PAC/BAC) libraries from rice variety Nipponbare as a common template resource. As of September 2000, ten countries from this international collaboration had already contributed about 30 Mb of the rice genome sequence. Analysis of the rice genome should facilitate a better understanding of the concept of inheritance in the rice plant and the development of new research endeavors in physiology and biochemistry. Crucial information from nucleotide sequences will be useful for improving breeding technology as one of the ultimate goals of rice genome research.

Rice is undoubtedly a dominant and important staple worldwide. The global production of rice reached 600 million tons in 1999, which is roughly equivalent to the production volume achieved in maize and wheat. However, to meet the expected increase in the population (8 billion by 2020) that depends largely on rice as a direct or indirect source of food, rice production needs to be sustained for at least the next 50 years. One way to address this problem is to ensure the improvement of current rice yields and expand agricultural area for rice production. Conventional breeding, however, has been the sole or principal approach for crop improvement, particularly in the development of many elite rice varieties. The actual situation is far from ideal. The population increase in Asia, Africa, and Latin America is expected to continue beyond modest estimates and breeding of rice varieties that are tolerant of biotic and abiotic stresses on those continents remains a feasible way to meet this challenge.
The boom in rice genome analysis is closely linked to the facts described above. Refinements in genome research technology have progressed well during the past 10 years such that nucleotide sequencing is now 20-fold faster. Simply put in a more concrete concept, sequencing the entire rice genome is a goal that can be realistically accomplished. The development of basic tools such as a fine molecular genetic map with 2,300 markers (Harushima et al 1998), a physical map with YACs (yeast artificial chromosomes) covering 60% of the rice genome (Saji et al 2001), and 10,000 independent ESTs (expressed sequence tags) (Yamamoto and Sasaki 1997) has greatly facilitated the launching of the whole-genome sequencing of rice. Without these tools and information, genome sequences are only meaningless digitized codes that are irrelevant to the study of inheritance. For example, How can genetically unassigned sequences be used to improve rice quality? or How can genes controlling the targeted phenotype be identified? Only by combining genetically reliable data can genome sequence information provide a clue for advanced studies to overcome biotic or abiotic stresses in rice plants.

In 1997, the Japanese government decided to start a new program beginning in fiscal year 1998 to tackle all of the rice genome sequence as an indispensable resource for advanced genomics such as gene discovery or gene expression profiling with an ultimate goal of improving rice quality by using a novel breeding strategy. This announcement stimulated other countries that had a strong interest in genome sequencing, not only of rice but also of other cereal crops in view of the significant synteny shared by these closely related species. Compared with that of rice (430 Mb), the genomic sizes of other important cereal crops are huge: 3,200 Mb for maize and 17,200 Mb for wheat, which make them less attractive as targets for whole-genome sequencing. The utility of the syntenic relationship among cereals has been proven in the discovery of common genes such as the dwarf phenotype in maize (D8) and wheat (Rht1) based on genomic information derived from rice (Peng et al 1999). During the 4th International Symposium of Plant Molecular Biology in 1997 in Singapore, a workshop on rice genome sequencing was held to organize the foundation for an international collaboration, the International Rice Genome Sequencing Project (IRGSP), aimed at accelerating completion of the sequencing of the entire rice genome (http://rjp.dna.affrc.go.jp/Seqcollab.html). In February 1998, the first working group meeting was held in Tsukuba with five countries establishing basic guidelines governing the sequencing collaboration such as sequence accuracy and policies for sequence release. Later in this chapter, sequencing progress at IRGSP will be described.

The sequencing strategy at RGP

Japonica rice variety Nipponbare was chosen as a template for genome sequencing in the Rice Genome Research Program (RGP) because it has been used as one of the parents of the F2 population used for constructing the linkage map, the DNA resource of the YAC physical map, and as a source of mRNAs used to construct cDNA libraries to obtain ESTs. This variety is also used as a common resource in the IRGSP. In
the RGP, a PAC (P1-derived artificial chromosome) library was used as a vector to establish a genomic library for sequencing. Our library is composed of about 70,000 clones of Sau3A1 fragments of DNA with an average insert size of 112 kb (Baba et al 2000). Another type of Nipponbare library was constructed using MboI as a restriction enzyme that has the same specificity as Sau3A1 but different sensitivity for methylated cytosine in the target sequence, GATC. In the rice genome, it is known that guanine and cytosine are rich in expressed genomic regions that might influence the distribution of fragment sizes and sequence characteristics obtained by both restriction enzymes. Both libraries should be complementary to make a perfectly reconstructed rice genome using DNA fragments.

About half of the clones in the Sau3A1 library are now being used to construct a sequence-ready PAC physical map by polymerase chain reaction (PCR) screening. The PCR primers are designed from the sequences of sequence tagged sites (STS)/EST markers aligned on the linkage map or YAC physical map. The STS markers are derived from 1,350 restriction fragment length polymorphism (RFLP) markers, of which 1,250 originated from rice cDNAs. On the other hand, locating cDNAs on the YAC physical map using a three-dimensional PCR screening method generates the EST markers (Wu et al 1999). Positive YACs obtained by each EST are aligned consistently along the linkage map using the computer software SEGMAP with the help of RFLP markers. In the RGP, 5,200 EST markers have been generated on the rice genome and these are used for ordering PAC clones. Because the positions of EST markers are not defined genetically, the ordering of markers is sometimes ambiguous. A fingerprinting strategy of ordered PACs with HindIII, EcoRV, and BglII is used to overcome this limitation and to estimate the degree of overlapping among the positive PAC clones extracted by EST markers.

Screening with STS/EST markers is invalid for identifying PACs located in genomic regions where no such markers are available. There are two main methods for filling these gaps: one is end-walking and the other is the STC (sequence tagged connector) method. End-walking uses the end-specific sequence derived from the clone flanking the gap as a PCR primer and hence requires the use of PCR against pooled PAC DNAs in several steps. This method is essential for efficient and specific screening of correctly adjacent PACs. STC, on the other hand, requires a data set of partial sequences of both ends of each PAC or BAC clone within a library. After the core PAC is sequenced, its overlapped sequences are computationally searched within STCs and the degree of overlap is estimated by fingerprinting. This method requires a database of STCs for a target library and, so far, a total of 105,000 STCs derived from two types of Nipponbare BAC libraries (HindIII and EcoRI fragments) constructed by Rod Wing of Clemson University and 55,000 STCs from RGP’s MboI BAC library are available. These numbers indicate a ratio of one STC in every 2.7 kb of the rice genome. Considering the average size of BACs (130 kb) in these libraries, 40 points are expected within each BAC and this density seems sufficient for locating adjacent BACs. The overlap of selected BACs must be examined by fingerprinting because a similar sequence, especially a repeat sequence or transposons, might interfere with the specific identification of correctly positioned BACs among picked BACs.
Selected PAC/BAC clones are cultured and their DNAs are purified by a Qiagen Largeconstruct Kit™ to eliminate DNA from the host bacteria and then they are sheared by ultrasonic treatment. The fragmented DNAs are blunt-ended by T4 DNA polymerase and electrophoresed on an agarose gel for size fractionation to obtain about 2-kb and 5-kb fragments. The eluted DNA fragments are ligated into plasmid vector pUC18 and then transformed using E. coli DH5α to construct 2-kb and 5-kb libraries. Usually, about 1,000 clones from both libraries are used as templates for sequencing by the dye-terminator Sanger method from both ends to generate 4,000 sequences with about 500 bp. Using an ABI3700 automated capillary sequencer, it takes one week to collect this amount of data for one PAC/BAC. These data are analyzed using a computer (UNIX) software, Phred/Phrap, to check the sequence quality and then assembled in a correctly oriented manner from one end to the other. If gaps exist, trials to fill them must be performed first by finding a clone spanning the gap and then sequencing it. Another problem is how to raise the sequence quality when the first assemblage fails to give the threshold quality value. These genomic regions often comprise GC-rich, AT-repeat, or inverted repeat sequences. The sequencing technology now available does not necessarily solve all of these problems and, in some cases, unsequenced regions remain. The final finishing step requires consistent size distribution of EcoRI and HindIII fragments with expected fragment sizes from obtained nucleotide sequences.

The finished sequence is then computationally analyzed to extract meaningful information from the sequence. This procedure is called annotation, in which mainly information on the existence of expressed genes is obtained and the expected gene names (identification) or functions are annotated after searching for similarities with nucleotide or amino acid sequence databases. Similarity search tools such as BLASTX and BLASTNW are usually used against databases with nonredundant translated amino acid sequences and with rice ESTs, respectively. In addition, a gene-prediction tool, GENSCAN, is used for predicting the plausible initiation and stop codons along the sequenced nucleotide. Although this tool or its derivative is only well fitted to Arabidopsis or maize genomic sequences, genes in the rice genome are predictable by combining all of the results obtained by these three tools. The predicted results are both available in the form of tables and graphical view.

To show the annotated results described above with the positional information of the sequenced PAC/BAC, we developed a database named INE, INtegrated Rice Genome Explorer, with a graphical view tool that is designed to function with Java applet (http://dna.afrc.go.jp/giot/INE.html) (Sakata et al 2000). This makes it feasible to track the target PAC/BAC basically by genetic markers along the linkage map. In addition, ESTs mapped on YACs are shown on the sequenced PAC/BAC after confirmation of their existence within the YAC sequence. The idea that sequence information should be shown basically by a map-view aims to facilitate gene identification when phenotypes such as mutants or components of QTL are genetically tagged by DNA markers and mapped on a genetic map. The present version of INE will soon be revised for easier use.
Progress of sequencing in the RGP

We are currently equipped with 14 units of ABI3700 capillary type-sequencers used for massive routine sequencing and 10 ABI377 96-well slab-gel-type sequencers used mainly for gap filling and PAC-end sequencing. Robotic machines are employed for plasmid isolation, sequencing reaction, colony picking, or clone transfer from one titer plate to another. For the first assembly of 4,000 shotgun sequence data, each chromatogram file is assembled with computer software, Phred/Phrap or its equivalent, Traceturner. The assembled data are then visualized by another computer software, Consed. If sequence gaps are detected after assembly or if the score of each nucleotide is too low to satisfy a reliable threshold value even after assembly, sequences in these regions must be reanalyzed. In case of gaps, first, a shotgun clone bridging the gap must be located using the Phred/Phrap data, and then this clone should be sequenced after preparing subclones by a shotgun or deletion method. Another choice is to use primer walking by designing PCR primers for a Dye-terminator sequencing reaction using the bridging clone as a template. The former method is cheaper and saves more time than the latter.

The RGP is now focusing on chromosome 1 to complete its sequence. The size of this chromosome is estimated at 52 Mb based on its total genetic distance. PCR screening of the 34,560 PAC clones with about 400 STS/EST markers and a computational search of the flanking sequences of the sequenced PAC/BAC among 120,000 BAC-STCs resulted in 52 contigs covering 27 Mb of chromosome 1 as of August 2000. As mentioned above, the strategy adopted in the RGP to make a sequence-ready physical map largely depends on STS/EST markers; therefore, the genomic region without such markers must be filled by a combination of STC search and fingerprints. The disadvantage of the latter method is the length of time required to obtain sequence information on PAC/BAC to be used for the STC search (about three weeks from the beginning of PAC/BAC culturing to the end of the first shotgun sequencing). However, this is the only method that can be used for filling gaps between genomic regions without any markers and much effort must be given to this approach.

By the end of September 2000, 48 PACs on chromosome 1, 12 PACs on chromosome 6, 2 PACs on chromosome 2, 1 PAC on chromosome 3, and 1 PAC on chromosome 8 had been completely sequenced and annotated. Following these clones, 25 PACs and 2 BACs on chromosome 1, 7 PACs on chromosome 6, and 1 PAC on chromosome 2 had also been completely sequenced and are under annotation. The total sequenced length is 13.0 Mb including the overlapped regions within a contig. These sequences have been registered in the DNA Databank of Japan (DDBJ) and the information is freely available on the Web.

We observed several characteristics of the completed and annotated sequences. The average predicted gene density is one gene in every 5 kb. This ratio suggests the presence of about 80,000 genes within the rice genome, assuming their even distribution. The total number of expressed rice genes was previously estimated at about 20,000–30,000 based on the matching frequency of genomic and expressed genes so far cloned. This large discrepancy in the estimated total gene number could be due to
an overestimation of the gene number and uneven gene distribution within the rice genome. The former seems to be the case because, as in Arabidopsis, a total of 30,000 genes is predicted from the whole-genome sequence. We cannot offer any reasonable explanation regarding the large discrepancy in estimating the total gene number considering that rice is a monocot, whereas Arabidopsis is a dicot. The most plausible reason for this discrepancy may be the uneven gene distribution because the genomic regions sequenced so far have been selected using PCR primers designed by nucleotide sequences from expressed genes. In addition, the restriction enzyme Sau3A1 used for constructing the PAC library is sensitive to methylated cytosine, and intergenic or heterochromatin regions may be resistant to cutting into small fragments with this enzyme. This speculation will be proven by using the BAC library constructed using MboI, which has the same specificity as Sau3A1 but is insensitive to methylation. This MboI BAC library of Nipponbare has been used so far to generate 50,000 STCs, which are used to screen clones to fill gaps between PACs picked by ESTs.

Information on repeated sequences is also available from sequenced PACs. Many types of repeats in the rice genome can be evaluated with the software Miropeats. If the most abundant two-base repeat is TA, then the exact number of repeats cannot be confirmed experimentally. Also, a one-base repeat of G or C and a two-base repeat of GC may be present with short lengths. These types of repeats can seriously interfere with sequencing using any chemistry of the dideoxy nucleotide method.

Transposable elements are annotated and are shown in INE. Several types of elements, such as Gypsy-like, Copia-like, and non-LTR types, are found. The frequency of existence of these transposable elements is one element in every 100 kb among the PACs we have sequenced so far. This frequency is unexpectedly low and for the same reason as to why gene density is higher within these PACs. This situation is very different from that observed in the maize genome, which is rich with transposable elements around the Adh1 gene (San Miguel et al 1998). The difference in the genome size of both species may reflect the contents of such elements between genes.

A similarity search by BLASTX of the NCBI nonredundant protein database frequently matches a hypothetical or putative Arabidopsis gene as the most similar counterpart of the predicted rice gene. Although this reflects the progress of genome sequencing of Arabidopsis and huge amounts of genome sequence data from other plants, no evidence of clear similarity has been shown in the gene order between sequences of rice chromosome 1 and almost all of the five chromosomes of Arabidopsis. A similarity search with BLASTN of rice ESTs from rice variety Nipponbare (the same variety used for genome sequencing) revealed that more than one-third of the predicted genes matched ESTs with a significant score. In many cases, the same predicted gene redundantly hit ESTs.

IRGSP: International Rice Genome Sequencing Project

Rice has a genome size of 430Mb, the second largest species so far targeted for genome sequencing, and its importance as a staple food for about half of the world population is well recognized. This motivated researchers having a strong interest in
The International Rice Genome Sequencing Project: ... rice genome sequencing to organize an international collaborative project to decode the sequence in an accelerated manner encompassing strict sequencing standards common to all participating members (Sasaki and Burr 2000). At the beginning of IRGSP, five countries, Japan, the United States, the United Kingdom, China, and Korea, joined together to discuss and agree upon guidelines on sequence standards, chromosome sharing, and so on. An in-depth description of the rules governing the collaboration is available on the Web at http://rsg.dna.affrc.go.jp/Seqcollab.html. The main points of these guidelines are summarized as follows: (1) japonica rice variety Nipponbare is sequenced as a common DNA resource, (2) the sequencing strategy is clone-by-clone using a PAC/BAC library shared among IRGSP members, (3) accuracy of the sequence must be 99.99%, (4) each member has a mandate to sequence at least 1 Mb per year, and (5) sequence data will be released immediately to the public after the completion of sequencing. As of September 2000, a total of 102 PACs/BACs had been completely sequenced in the IRGSP, covering about 15 Mb of the rice genome. An additional 100 PACs/BACs covering another 15 Mb have also been sequenced although several gap regions remain to be filled.

Thus far, at most 10% of the rice genome has been sequenced. How can we accelerate complete sequencing under limited financial support? Sequencing can be enhanced with funding support from sources other than IRGSP. In April 2000, a private company, Pharmacia/Monsanto, announced the delivery of its rice genomic sequences, BAC clones, and other related data to IRGSP. The data have been generated for a project to produce a rough draft sequence of the rice genome using rice variety Nipponbare. Two key points are expected to complement the IRGSP sequencing strategy: one is to construct a sequence-ready BAC/PAC physical map; the other is to supply shotgun sequence data for each target BAC. An agreement to combine both sets of data from IRGSP and Pharmacia/Monsanto is under negotiation. Although key issues surrounding the merging of data remain to be resolved, a blueprint of the rough draft sequence of the rice genome should enhance early completion of the sequencing project.

References

Notes

Authors’ address: Rice Genome Research Program (RGP), NIAR/STAFF, 1-2, Kannondai 2-chome, Tsukuba, Ibaraki 305-8602, Japan, E-mail: tsasaki@abr.affrc.go.jp.