Resources and Opportunities

Deciphering a Weed. Genomic Sequencing of Arabidopsis

Nancy Federspiel*
Exelixis, Inc., 170 Harbor Way, P.O. Box 511, South San Francisco, California 94083-0511

By the end of 2000, the genomic sequence of the model plant, Arabidopsis, will be completed, annotated, and released. Although the hype over the draft of the human genome sequence may overshadow the significance of this lowly weed, its importance for the future food security of the world’s population may have equal weight for human health with the actual catalog of human genes. In recent months the popular media have enthusiastically reported protests around the world against the use of genetically modified organisms, especially plants. While farmers have been breeding, and thereby modifying, agricultural crops for millennia, protesters feel that the use of recombinant DNA for changing the characteristics of plants is inherently dangerous. Nevertheless, the entire complement of Arabidopsis genes will define the first member of the “other” multicellular biological kingdom and along with the emerging rice genome it will facilitate even more precise manipulations of crop species to achieve targeted goals. In addition, this will be the most accurate and most complete sequence of a higher eukaryotic genome in existence and it will probably retain that status for some years to come. At this historic juncture it is well worth a quick review and acknowledgment of the process and the people whose efforts have made this achievement possible.

It was just over a decade ago that the Arabidopsis Multinational Science Steering Committee, a small group of forward-thinking Arabidopsis researchers, decided that the most valuable goal for the advancement of plant science was the determination of the entire sequence of the approximately 130-Mb Arabidopsis genome. With a target completion date of 2004, new sources of funding and scientists interested in undertaking such a large effort had to be identified without detracting from other aspects of basic plant biology research. Early projects focused on identifying expressed genes through single-pass sequencing of cDNA clones; Tom Newman and colleagues at Michigan State University (Newman et al., 1993), and a group of investigators in France (Höfte et al., 1993) produced a large number of expressed sequence tags (EST) that were collected in the database dbEST (Boguski et al., 1993). Preparations for genomic sequencing in the early 1990s in Europe and the U.S.; in addition, the Japanese developed libraries to support their sequencing efforts in a P1 vector and in a modified BAC vector that could be used in Agrobacterium-mediated plant transformation.

The Europeans were the first to organize a sequencing consortium and obtain funding for a pilot genomic sequencing effort. Led by Mike Bevan at the John Innes Centre in Norwich, 18 European laboratories formed “European Scientists Sequencing Arabidopsis” to sequence approximately 1.9 Mb on chromosome 4. In December 1995, the Kasuza DNA Research Institute in Japan initiated their Arabidopsis sequencing efforts, focusing on chromosome 5. Soon after, in 1996 three U.S. funding agencies, the Department of Energy, Department of Agriculture, and the National Science Foundation as the lead agency requested proposals and funded three pilot sequencing projects in the U.S. Arabidopsis sequencing at The Institute for Genomic Research (TIGR) focused initially on chromosome 2 and was directed by Steve Rounsley and Craig Venter. The second U.S. effort, a consortium of researchers at Cold Spring Harbor Laboratory led by Dick McCombie at the Genome Sequencing Center of Washington University School of Medicine in St. Louis headed by Rick

* E-mail nfedersp@exelixis.com; fax 650–837–8204.
Wilson, and at Applied Biosystems Inc. with Ellson Chen, focused on the top arm of chromosome 4. The third U.S. project, concentrating on chromosome 1, was directed by the SPP Consortium composed of the Stanford DNA Sequencing and Technology Center under Ron Davis and Nancy Federspiel, Joe Ecker’s group at the University of Pennsylvania, and Sakis Theologis’ group at the Plant Gene Expression Center/UC-Berkeley. In the summer of 1996, representatives of all the international sequencing groups met at the first AGI (Arabidopsis Genome Initiative) meeting and established guidelines for sequencing standards and accuracy, data release, and resolution of conflicts. The AGI has played an essential role for the last 5 years in coordinating the international efforts, monitoring progress, allocating regions of the genome to new participants, and redistributing areas to utilize excess sequencing capacity. Subsequent rounds of funding in the U.S. and Europe increased the rate of progress and included an additional European consortium led by the French genome center Genoscope under Marcel Salanoubat and Francis Quetier, focusing on chromosome 3. The combined efforts of all the groups has led to the completion of the Arabidopsis genome 4 years ahead of the original schedule!

Considering that over 40 laboratories have participated in this enormous undertaking, it is interesting to compare the different strategies that were used (or not used) to define the genomic structure of Arabidopsis. These strategies are variations along a gradient from fully random sample sequencing to highly directed localized sequencing. Whole genome shotgun sequencing (in which total genomic DNA is sheared into small pieces, cloned, and sequenced) was proposed as one component of the effort by the SPP Consortium because the resulting random sequences would tag nearly every gene in the genome in the first phase of the project. Directed sequencing, in contrast, produces highly accurate sequence in discrete areas of individual chromosomes, whereas other chromosomal areas remain unknown until late in the process. At the first AGI meeting there was vigorous debate among the representatives over the whole genome shotgun strategy and it was decided by majority vote not to support it due to concerns over how the data could be utilized and incorporated into the final, highly accurate product. Five years later this strategy has become much more accepted in the biological community as an efficient way of surveying a new genome in the initial stages of analysis.

Participants in the AGI then utilized the more directed approaches with BAC or other large-insert clones that were mapped to specific chromosomes as substrates for all the sequencing efforts. European Scientists Sequencing Arabidopsis, composed of a large number of laboratories with varying sequencing capacities, chose early on to invest in a large mapping effort to establish an overlapping set of clones distributed along their chosen chromosomal locations. Once the list of clones was defined, individual clones were allocated to the different participating laboratories for sequencing, and each laboratory could operate independently of the others on its own set of clones. Using this approach, the first large contiguous stretch of Arabidopsis genomic sequence was completed and it was revealed that the predicted gene density in Arabidopsis is quite high, averaging 1 gene/5 kb (Bevan et al., 1998). An alternative approach was used by many of the other sequencing groups, which required less initial effort on mapping clones; this strategy instead required a database of the end sequences of all the available BAC clones. To initiate the process, “seed” BACs were mapped to widely distributed sites along a chromosome and then completely sequenced. The sequence of each seed BAC was then compared with the end-sequence database to identify the BAC clones that overlapped each end by a minimum amount. These BACs were in turn sequenced completely to extend the region of contiguous sequence and the process was repeated until the “contigs” merged to form the entire sequence of a chromosomal arm (centromeric repeats are difficult to sequence and require specialized strategies). This approach led to the completion of the sequence of chromosome 2 (Lin et al., 1999) and the mapping strategy resulted in the completion of the sequence of chromosome 4 (Mayer et al., 1999).

Although most sequencing groups relied heavily on manual labor and, later in the project, some commercial automation for sample preparation, the SPP Consortium used the Arabidopsis sequencing project as the testbed for new robotic instrumentation developed at Stanford to reduce the cost and increase the throughput for sequencing (Marziali et al., 1997). The AGI agreed to follow the so-called “Bermuda standards” for sequence accuracy established by the Human Genome Project so that the resulting final product should have less than one error per 10,000 bases. However, adhering to the Bermuda standards for immediate data release to GenBank proved to be more problematic for some AGI groups due to their sources of funding and other internal issues. As a result, public release of sequence data ranged from depositing rough draft sequence in GenBank within 24 h of assembly to posting of sequence data on individual web sites, to releasing to GenBank only after finishing to high quality and annotation. This was a continuing source of discussion and negotiation among AGI members throughout the project, especially in the latter stages when re-allocation of clones required data exchange to confirm overlaps and fill gaps in the sequence. Publication of the sequence of the remaining three chromosomes, as well as a discussion of the structure and content of the entire genome, is scheduled for the end of 2000 and will ensure that all AGI sequence data is in the public
domain (The Arabidopsis Genome Initiative, in press). The accumulation of approximately 130 million pieces of Arabidopsis sequence in the database certainly does not in and of itself provide the key to understanding how this model plant works or how to modify economically important crop species for improvement in nutrition, response to stress, and other traits required for long-term agricultural sustainability. It is analogous to now having a book published in a foreign language, divided into five chapters (for the five chromosomes), but having no punctuation to separate the words and sentences and no dictionary for translation. Annotation of the sequence, or defining the interesting features such as genes, repetitive sequences, regulatory regions, etc., will initiate the process of producing a Rosetta stone for the plant genome. Members of the AGI have all been responsible for annotating the clones that were sequenced in the individual groups, and diverse methods have been used to provide these gene labels along the genomic sequence. In general, sequence matches with the EST database, with the non-redundant protein sequence database, and results of gene prediction programs are the bases of annotation.

Numerous annotation issues must be addressed in any large-scale genomic sequencing effort such as this. ESTs generally represent only a fraction of the genes in an organism, and the individual sequence tags usually define only a portion of each gene. Comparisons to characterized genes from other organisms provide another method of identifying conserved genes, but intron/exon junctions, start/stop sequences, and regulatory regions may or may not be similarly conserved. Gene prediction programs often produce conflicting output that must be interpreted by a human annotator whose final choice may not be biologically accurate. In addition, characterization of a "hypothetical protein" that has no database matches may be correct on the day the annotation was completed and submitted to GenBank, but may be incorrect the following day after a new set of sequences from any source is deposited into the database. To help basic researchers during the course of the genome project, several of the AGI members created databases to automatically predict coding sequences for GenBank records lacking annotation, to identify protein motifs and protein similarities, and to provide regular updates of new entries (DATAMAP, Palm et al., 2000; http://www.kazusa.or.jp/kaos/; http://www.mips.biochem.mpg.de/proj/thal/).

When the complete Arabidopsis sequence has been compiled from the individual BAC sequences, TIGR is funded to provide a unified set of annotation that will provide consistent nomenclature and current database matches (http://www.tigr.org/tdb/ath1/htmls/ath1.html). However, this again will be a snapshot in time of the existing data and knowledge, and ongoing efforts to curate the sequence data in a stable, accessible database will be required for full development of the long-term value of the Arabidopsis sequencing project. The Arabidopsis Information Resource is one such possible resource, and their efforts are described in another article in this issue. Despite the best efforts of programmers attempting to improve gene prediction software and annotators, a large percentage of putative genes will remain essentially guesses without corroborating experimental evidence. Full-length cDNA sequencing, characterization of tagged mutant lines, and expression profiling with arrays are among the approaches that will greatly aid in deciphering the meaning of the Arabidopsis genome sequence. The knowledge gained from this major endeavor will ultimately define the basic building blocks of this model plant and provide the tools for manipulating desired traits in other more complex plants. It is to be hoped that the precise definition of what is changed in a genetically modified plant will allow a more exact understanding of the outcome of this manipulation. In this way, political, environmental, and biological issues can be addressed so that the scientific community and the general public alike can be informed of the merits of the new agricultural genomics technologies.

Received September 7, 2000; accepted September 18, 2000.

LITERATURE CITED


Newman T, Bruijn, FJ de, Green P, Keegstra K, Kende H, McIntosh L, Ohlrogge J, Raikkel N, Somerville, S, Tho-


