Editor’s Choice Series on Sharing Data and Materials

Sharing the Wealth. The Mechanics of a Data Release from Industry

Steven Rounsley*
Cantata Pharmaceuticals, 300 Technology Square, Cambridge, Massachusetts 02139

In this era of “-omics,” the approach to data collection has widened and become less focused. Researchers will often collect data in a very non-targeted way, generating a large resource that can later be mined to address specific questions. The creation of these resources has created a new challenge in the age-old tug-of-war between self-interest and altruism. Whether in the academic or private sector, the value of genomics-style data resources for the scientific community is a powerful force that has changed the way we think about the scientific endeavor.

Sharing of experimental data is a long-standing tradition in science. However, this “community spiritedness” has traditionally occurred hand in hand with some form of recognition (e.g. publication) and for the purpose of enabling others to reproduce and confirm the results. Without such data sharing, the conclusions cannot be fairly evaluated and, thus, the reputation of the researcher is built on publishing the best data available. However, with the large data resources generated using any of the technologies du jour, the opportunities for publication are limited, and, yet, the community benefits from sharing the data are much greater.

When these resources are created with public funds, such as the hugely successful National Science Foundation (NSF) Plant Genome Program, sharing of data can be made a condition of funding. Such an approach led to the NSF-funded sequencing groups releasing Arabidopsis genomic sequences as they were produced—piece by piece. This, in turn, enabled countless researchers to make breakthroughs using that sequence and led to many more publications than were ever produced about the sequence itself.

When the data resource is generated in the private sector, the obstacles to data sharing are different yet the dilemma is very similar. Having led Arabidopsis genomics projects in both public and private sectors, I have experienced frustrations and rewards of sharing data in both settings. By way of illustration, allow me to describe the process that led to the release in 2000 of the Arabidopsis polymorphism collection by Monsanto’s Cereon Genomics unit.

In 1998, the publicly funded Arabidopsis Genome Initiative (AGI) was reaching full throttle. In July of that year, a little over 28 Mb of the Columbia ecotype genome had been completed. The remainder would be essentially finished over the next 2 years. The availability of large portions of the genome sequence was changing the way people thought about many aspects of Arabidopsis biology, including gene family analysis, molecular evolution, and map-based cloning.

Until then, map-based cloning involved the arduous process of chromosome walking in which large-insert clones such as bacterial artificial chromosomes (BACs) or yeast artificial chromosomes (YACs) were used to first produce a physical map of the region and, subsequently, as a starting point to search for polymorphic markers to allow the segregating mutations to be mapped. If a marker was from a polymorphic region of the genome, then the process could move forward, but this trial and error process was labor intensive and time consuming.

With the availability of large stretches of sequence, the process could be accelerated. Probes could now be designed from precise locations along the BAC clone, and the genic content of the region was known. As a consequence, mapping projects around the world began following the progress of the AGI sequencing effort very closely. Graduate students eagerly awaiting the completion of the next clone were not shy in communicating their anticipation to the AGI sequencing groups. Assuming sequence was available for the region of interest, the limiting factor was finding the polymorphisms in that region of sequence. One popular method was to target any mono-, di-, or trinucleotide repeat as a candidate site for polymorphism, but depending on ecotype pair, this was often unsuccessful.

In 1998, rumors of an Arabidopsis genome project in the private sector started to surface. This caused some initial consternation within the AGI but eventually just led to increased determination to complete the genome in a timely manner. The rumors were in fact true, although the goal and details of the project were slightly different than the AGI had imagined.

Cereon Genomics, which had been established by Monsanto in late 1997, was sequencing the Landsberg erecta ecotype of Arabidopsis using the whole genome shotgun method. This approach to sequencing larger genomes was a widely reported news story in mid-1998 with the establishment of Celera Genomics and its attempt at the human genome. The fact...
that the AGI was generating a high-quality, complete genome sequence for the Columbia ecotype allowed Cereon to tailor its sequencing project for two complementary goals. The first—using the fragments obtained from low-pass sequencing to glimpse at the gene content of Arabidopsis—was transitory and would soon be surpassed by the higher quality, more complete public sequence. The second made use of the differences between the two ecotypes. Could a low-pass genome sequence identify polymorphisms between Columbia and Landsberg *erecta* ecotypes and, therefore, accelerate map-based cloning projects within the company?

Identifying the polymorphisms was not a straightforward task. Due to the reduced quality inherent in low-pass sequencing, many of the differences between the ecotypes were actually sequencing errors rather than polymorphisms. This difference in quality between the two genomes required that the resulting data be considered as predicted polymorphisms rather than actual polymorphisms, but in practice, the data were quite reliable. Upward of 90% of the predicted single nucleotide polymorphisms (SNPs) that were tested were confirmed. In addition to higher than expected reliability, the density of the predictions was also a pleasant surprise. On average, we had a predicted SNP every 3.3 kb and an insertion or deletion every 6.6 kb (Jander et al., 2002). Alignments of high-quality Landsberg *erecta* sequences from GenBank against the Columbia BACs suggested that the real SNP density is actually 10-fold higher, but finding just 10% of the possible polymorphisms provided a density that had the potential to map a mutation to a single gene simply by segregation analysis.

Although the Landsberg *erecta* sequencing project had been completed, its value continued to grow as each AGI BAC was released. Therefore, a process was established to continually update the collection of polymorphisms with each AGI release. The value of this resource was illustrated internally by the successful and rapid map-based cloning of several genes of interest to the company. As a consequence, the scientists involved in the project began to realize the impact this data set would have on the whole research community if only it could be made available. It was decided that we would make a proposal to management. Our eagerness to do this was likely connected to the fact that the majority of the scientists involved had recently made the move from academic positions to Cereon Genomics. The next challenge was to make the proposal attractive to management. Our eagerness to do this was likely connected to the fact that the majority of the scientists involved had recently made the move from academic positions to Cereon Genomics. The next challenge was to make the proposal attractive to management. The gist of the request was that we had spent millions of dollars building a resource, had only just started to gain any value from it, but we’d like to give it away. On its face, it seemed to be an idea that would be difficult to justify to senior management.

However, beyond altruism, there were a number of reasons that made this a good idea. First, in our business, cloning a gene was just one step in a long process toward a product. Preventing others from cloning genes didn’t help us get to those products any faster. In fact, a whole community could clone genes faster than we could, and each of those genes would be available after publication. Second, good will with plant biologists is a valuable commodity and was an important goal for Cereon and Monsanto, as was enhancing acceptance of agricultural biotechnology by furthering plant biology research in the academic sector. These reasons and the general feeling that it was the right thing to do were readily accepted by Cereon’s and Monsanto’s management, and approval was given to explore various mechanisms for the release—with the caveat that any plan to release the data must not allow it to be used by Monsanto’s industry competitors.

And therein lies the central issue that inhibits many academic-industry relationships. Not the fear that the academic collaborator will benefit from the data, but a fear that the industry competition will indirectly benefit from it. Giving a company’s competitors an advantage is generally not a good strategy for increasing shareholder value.

Different mechanisms for the data release were considered over a period of several months. Each proposed mechanism was evaluated for its ability to achieve the goals of the data release: enable the plant biology community in a major way, increase community goodwill, and protect Monsanto’s interests by preventing access to industry competitors.

Initially, a limited collaboration-style model was considered. This would provide the marker set to existing Monsanto collaborators under a confidentiality agreement. This plan was quickly rejected because it benefited only a small number of researchers and had a high overhead cost, involving individual confidentiality agreements. A broader access to the data set would be needed to achieve the goals of the data release.

The second mechanism was making the data set available to any academic researchers using a standard material transfer agreement (MTA). This had the advantage of using a standard approach to protecting Monsanto’s intellectual property. However, the reach-through rights common with such agreements were very unpopular with the community, and, in this case, were unjustified. Even using a more relaxed MTA would not remove the large administrative overhead needed for such a program.

The plan that was ultimately successful took a very different approach. Maximizing community access was achieved by teaming up with the NSF-funded Arabidopsis community database, The Arabidopsis Information Resource (TAIR). This solution was ideal for many reasons. First, the nature of the data set lent itself to online rather than physical distribution—the
traditional approaches designed for sharing clones and seeds were not needed here. Second, the TAIR collaboration would simultaneously advertise the availability of the data to the whole community and place it alongside the other community resources—a perfect location. Third, the generosity of the TAIR group presented an extremely economical way of making the data available—the establishment and maintenance of an electronic distribution mechanism for just this data would be costly.

Unfortunately, making the decision to collaborate with TAIR for the data release was just the beginning, and much work remained. Online distribution via TAIR required a different approach to the legal aspects of the data release. First, the terms of the traditional MTA were reworked to produce a more relaxed agreement between the researcher and Monsanto. This agreement prohibited redistribution of the entire data set but otherwise placed few restrictions on the academic researcher. The entity that was being protected by the agreement was the data set as a whole; therefore, individual polymorphisms with utility could be freely used and up to 20 could be published—to support a map-based cloning paper, for example.

Traditional MTA’s require signatures by both parties before they were binding. Coupling online distribution with the requirement for signatures would be very clumsy. Therefore, we wanted to implement a “click” agreement in which a user would agree to the terms by clicking the “I agree” button. Although many examples of such agreements existed on the Web, surprisingly there was very little case law to guide lawyers in crafting the agreement. This added to the time needed to set up the collaboration.

Last but not least, an agreement between Monsanto and TAIR was needed. This was simply to formalize the agreement between the two organizations but was made more complicated by the fact that TAIR was not a single organization but an NSF-funded project at two different institutions. Thus, before the collaboration could get started, the TAIR-Monsanto agreement needed approval from the Carnegie Institution of Washington, the National Center for Genome Resources, and the NSF.

On May 3, 2000, the Cereon Arabidopsis Polymorphism Collection was made available to the academic community through the TAIR Web site. Several updates were made available in the following months as the AGI project completed the Columbia genome; now, the final collection contains 56,670 polymorphisms. Since the first release, almost 1,000 researchers have registered for access to the database. In February 2002, a similar agreement was put in place making the original data from low-pass sequencing of Landsberg erecta available thru TAIR.

Despite the long path from initial idea to final implementation, every step was driven by the understanding, shared by all involved, that this was a worthwhile endeavor. The long and involved process was as much due to the fact that many aspects of this had never been done before as caused by the corporate realities of intellectual property protection. With the success of this and other data releases, I hope that researchers in industry will feel more emboldened to take suggestions for new industry-academic collaborations to their management. Given the appropriate project and a willingness to challenge conventional wisdom on proprietary data, it is possible to find a solution that benefits the entire research community—public and private. It is certainly a legacy that Cereon Genomics can be proud of.

Received March 24, 2003; returned for revision July 16, 2003; accepted July 16, 2003.

LITERATURE CITED