Plant Comparative Genetics after 10 Years

M. D. Gale and K. M. Devos

VIEWPOINT

The past 10 years have seen the discovery of unexpected levels of conservation of gene content and gene orders over millions of years of evolution within grasses, crucifers, legumes, some trees, and *Solanaceae* crops. Within the grasses, which include the three 500-million-ton-plus-per-year crops (wheat, maize, and rice), and the crucifers, which include all the *Brassica* crops, colinearity looks good enough to do most map-based cloning only in the small genome model species, rice and *Arabidopsis*. Elsewhere, knowledge gained in a few major crops is being pooled and applied across the board. The extrapolation of information from the well-studied species to orphan crops, which include many tropical species, is providing a solid base for their improvement. Genome rearrangements are giving new insights into evolution. In fact, comparative genetics is the key that will unlock the secrets of crop plants with genomes larger than that of humans.

Over the past 10 years, plant comparative genetics has shown that the organization of genes within genomes has remained more conserved over longer evolutionary periods than previously imagined. We are left, of course, with many new questions. Two of the most pressing today are "How good is the genetic colinearity between the model plants receiving DNA sequencing attention and their related crop species?" and "What are the useful limits to comparative genetics?" These questions are vital in the planning of the next generation of genome programs, which will include maize and wheat, which both have more DNA than *Homo sapiens*.

In the mid 1980s, when restriction fragment length polymorphism (RFLP) analysis was first applied to plants—tomato and maize in the United States and wheat in the United Kingdom—it became clear that complementary DNA RFLP probes could be cross-mapped to provide anchors that allowed genomes to be compared. Two studies, one that showed that the tomato and potato maps were very similar (*1*) and another that showed that the three diploid genomes that form present-day hexaploid bread wheat had retained almost identical gene orders (*2*), gave the first hints that plant gene linkage arrangements might have remained conserved over long evolutionary periods. Over the past 10 years, close relationships have been demonstrated between the genomes of almost all economic grass crops (Fig. 1), between the *Solanaceae* crops (*3*), between the *Brassica* crops and *Arabidopsis* (*4*), between pines (*5*), between rosaceous fruit tree species (*6*), and between several legumes (*7, 8*).

Several generalities have emerged. Conservation of gene orders, but not intergenic sequences, over millions of years appears to be the rule within plant families. Some genome restructuring occurs, and this restructuring may be more rapid in polyploids than in diploids. Colinearity, however, rides over severalfold differences in genome size and chromosome number.

Is Colinearity Good Enough for Cross-Species Gene Isolation?

If colinearity is perfect, then it should be possible to isolate genes that have been mapped precisely on the genetic map in large genome species by map-based cloning in a smaller genome model species, such as wheat genes in rice or oilseed rape genes in *Arabidopsis*. A map-based cloning approach in rice has been used for the isolation of the wheat *Ph* gene (*9*), which controls chromosome pairing. Similarly, work is under way to isolate *Rpg1*, a stem rust resistance gene in barley, by "walking" in rice (*10*). Although neither walk has yet been concluded successfully, remarkably precise colinearity has been observed over most of the corresponding regions. However, in both cases, breaks in complete correspondence did occur in or near the target regions. These indications that everything may not be perfect at the microlevel are similar to results from human-mouse comparisons, where colinearity is often interrupted by insertions, deletions, and inversions (*11*). One of the few studies to date in which contiguous DNA sequences have been compared showed complete colinearity through three genes in the ~20-kb *Sh2-A1* regions in rice and sorghum (*12*).

Within the *Crucifereae*, colinearity looks to be extremely strong between *Arabidopsis* and oilseed rape, which are said to be only 10 million years apart (*13*), although, as yet, very little genomic sequence is available from the crop genome. One of the first results to emerge from the cross-mapping of *Arabidopsis* genes onto the *Brassica* genomes was that the basic *Arabidopsis* gene set is essentially triplicated in the diploid *Brassica* crops (*14*). The DNA content of the diploid *Brassica* crops, at 480 Mb, is, in fact, about three times that of *Arabidopsis*. Triplicated regions of similar genetic length have been identified that correspond with almost precise colinearity to segments of *Arabidopsis* that carry major flowering time genes (*4, 15*) (Fig. 2). In another fine-mapping study, T. C. Osborn's group at the University of Wisconsin, Madison, has established that the major vernalization-responsive flowering time gene in *Brassica rapa*, *VFR2*, is likely to be a homolog of *FLC*, which is located at the top of *Arabidopsis* chromosome 5 (*16, 17*). Preliminary data from R. Schmidt's lab in Cologne show that there is extensive microcolinearity between a 200-kb region of *Arabidopsis* chromosome 4 and a region of the *Capsella rubella* genome where 17 *Arabidopsis* genes mapped to four *Capsella* cosmid contigs (*18*). Within the contigs, gene orders were completely conserved and distances between genes, where they were established, were highly similar.

It is still not clear whether cross-species gene isolation is a robust technique that can be used in all instances. It is highly likely that the extent of colinearity between grass genomes will depend on the region. The genomics group at DuPont (*19*) has sequenced 330 kb of contiguous rice DNA in the *Adh1-Adh2* region and cross-mapped any genes they found onto maize. The results reflect good colinearity for housekeeping genes but a poor conservation for predicted "genes" with sequences akin to known disease resistance genes. Mapping of disease resistance gene homologs across rice, barley, and foxtail millet has already led to the conclusion that these genes may be evolving faster than most (*20*).

It may still be that the large genome sizes of wheat, barley, and maize are not the obstacle we expect them to be. It is still possible that the genes themselves lie in groups that are in turn separated by long tracts of repeats (*21*). The evidence to date is equivocal but tends to suggest that the amplification of the larger genomes is not random. *A1* and *Sh2* are 21 and 22 kb apart in rice and sorghum, respectively, even though the sorghum genome is nearly twice the size of rice (*12*). In barley, evidence for "gene islands" has been found in three gene regions (*22, 23*). Gene densities in three independent barley bacterial artificial chromosome (BAC) large insert clones were found to be at least 10 times higher than would be predicted from a random gene distribution. We need more data.

The authors are at the John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, Norfolk, UK.

Other Applications

Comparative genetics impacts many other aspects of research and breeding. Fusion of the knowledge arising from decades of independent research in species that are sexually incompatible, but have highly conserved genomes, is a major benefit for both breeders and geneticists. Comparative alignment of genes controlling quantitative and qualitative traits across species shows clearly that quantitative trait loci (QTLs) and major genes are often the same (*24*). Moreover, if the gene underlying the trait in one of the species has already been isolated, a ready candidate for the corresponding QTL is available (*4*). Candidate genes identified by mutant phenotypes in one species can also be examined for effects on related traits in other species (*25*). Another key application is the potential exploitation of a vastly increased range of "allelic" variants, where the availability of transformation technology no longer confines breeders to only the alleles readily available in their own crop.

Taxonomic thinking is also being affected (*26*), and chromosomal rearrangements that describe tribes within the *Poaceae* have been identified. Caution is, however, necessary when one uses rearrangements as a measure of evolutionary divergence. The number of chromosomal rearrangements cannot be assumed to be a measure of evolutionary time. Analyses of rye (*27*), which diverged only about 7 million years ago from wheat, and *Aegilops umbellulata* (*28*), which is even more closely related to wheat, show numbers of chromosomal rearrangements relative to wheat that are similar to those of rice relative to wheat, yet rice is at least 60 million years distant from wheat.

The Monocot-Dicot Divide

The value of *Arabidopsis* and rice as models for dicotyledonous and monocotyledonous plants, respectively, has been considerably strengthened by the comparative approach. But can we extrapolate

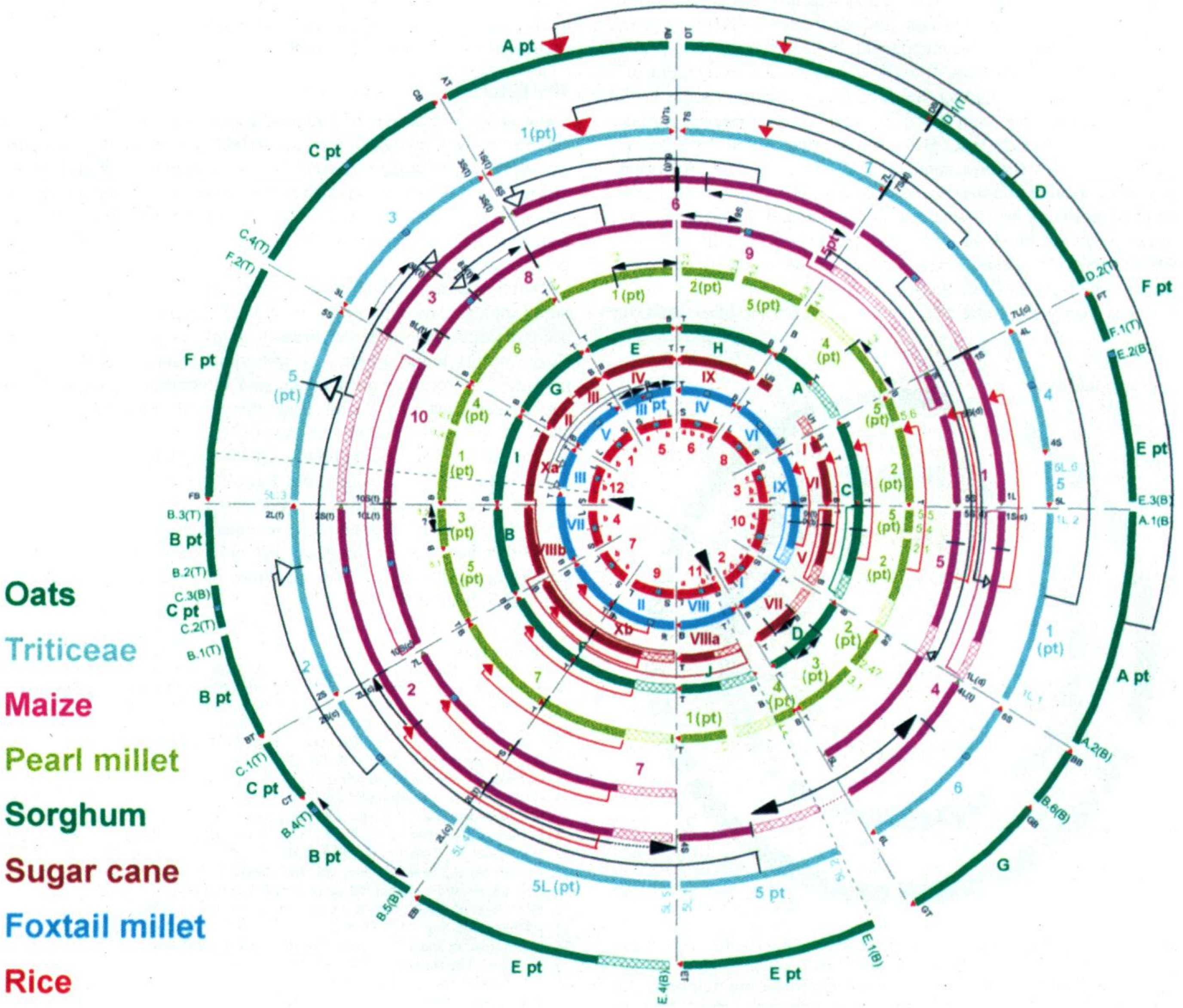

Fig. 1. Twelve grass genomes, one consensus map. Each circle represents the chromosomal complement of a single grass genome. The circles are aligned, in the most parsimonious manner relative to rice, so that radii will pass through different versions of the same genes in the different crops. The data have been drawn from many sources [(*33*); sources listed in (*26*)]. The arrows indicate the inversions and translocations, relative to rice, that are necessary to describe present-day chromosomes. Locations of telomeres (▲) and centromeres (■) are shown where known. Hatched areas indicate chromosome regions for which very little comparative data exist. L, long arm; S, short arm; T, top of chromosome; B, bottom of chromosome; and pt, part.

between the two?

At the level of the gene, we can. For example, having isolated the *Arabidopsis* gibberellin-insensitive dwarfing gene (*29*), an important question in N. P. Harberd's lab at the John Innes Centre (JIC) was whether *GAI* was a homolog of the dwarfing genes that are crucial for the high-yielding semidwarf wheats now grown worldwide. A rice expressed sequence tag (EST) with homology at the amino acid level was found. Mapping of the EST was difficult in rice and wheat but possible in foxtail millet. Excitingly, it mapped in a region that matched the location of the dwarfing genes on the group 4 chromosomes of wheat. Work is now under way to clone the wheat genes and other homoeoalleles, including the maize dominant dwarf, *D8* (*30*).

At the genome level, a key question remains as to whether or to what extent the 240 million years that separate the two main angiosperm groups have eroded conservation of gene order to the point where it is no longer a useful tool with predictive power. An early study (*31*) indicated that as few as 200 rearrangements may distinguish the genomes of sorghum and *Arabidopsis*. New evidence, emerging from comparative mapping at the sequence level, has failed to support this initial claim. Rice DNA sequencing by DuPont of a 330-kb fragment surrounding the *Adh2* locus on chromosome 11 did not reveal colinearity with the *Adh* region on *Arabidopsis* chromosome 1 (*19*). Similarly, J. L. Bennetzen's group found that only two out of eight genes within a contiguous 78-kb sequence on a sorghum BAC were located in adjacent positions on an *Arabidopsis* BAC, and unlinked positions were obtained in *Arabidopsis* for at least three further genes on the BAC (*32*). These latter results tally with data obtained in a JIC–Japanese Rice Genome Program (RGP) collaboration, in which genes from single *Arabidopsis* BACs were shown to map over nearly the entire rice genome (*33*). On the other hand, the Cold Spring Harbor Laboratory group observed conserved colinearity for four out of an estimated six genes when comparing the sequences of a rice BAC and an *Arabidopsis* BAC (*34*). A comparison by A. Kleinhofs' group of the rice sequence obtained around the *Rpg1* gene with *Arabidopsis* chromosome 4 sequence also revealed clusters of similar open reading frames (*35*). So, although there appears to be some evidence of gene conservation between monocot and dicot species, complete colinearity may be limited to few and very small regions. A fuller picture will emerge when data from the rice-sequencing initiative become available to carry out a more precise comparison between the two model species.

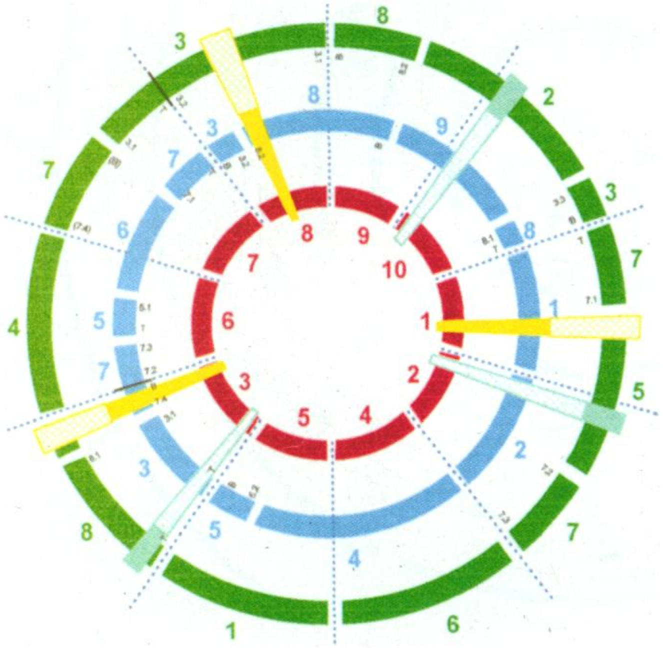

Fig. 2. Ten million years is a short time in crucifer evolution. Although the genomes of *B. rapa* (red circle), *B. oleracea* (blue circle), and *B. nigra* (green circle) have different chromosome numbers, the maps of the three genomes (*4*) can be aligned simply, revealing only a few chromosomal rearrangements that disturb complete colinearity. Moreover, each *Brassica* genome comprises three complete *Arabidopsis* genomes (*14*). The blue regions correspond to a 7.7-Mb *Arabidopsis* chromosome 4 region surrounding *FCA* (*14*). The yellow regions relate to the 2.2-Mb chromosome 5 region surrounding *CONSTANS* (*4*).

Arabidopsis is on course for completion before the original target date of 2003. Rice sequencing has started at RGP in Tsukuba, which is funded for the next 10 years. The new China Rice Genome Program sequencing facility in Shanghai has opened, and it looks as though U.S. funding agencies will match the Japanese effort. Funding is still being sought in Europe, Korea, and Taiwan. Thus, the issue is no longer whether *Arabidopsis* and rice genomic sequences will lead us to all genes in dicots and monocots but rather how best to exploit the sequences as they become available.

The Future

Some areas of research will require more input if we are going to exploit this new information to best effect. For example, researchers are beginning to realize that chromosomal duplications are complicating factors in many map-based applications. Increasing numbers of duplications of varying sizes and ages are being found in most genomes. Indeed, it now seems unlikely that the pure diploid plant genome exists.

Another area in which input is desperately needed is comparative bioinformatics, the only means by which plant geneticists can hope to become conversant with the breadth of plant genome work today. Cross-species bioinformatics is still in its infancy (*36*); however, progress has recently been made, and internationally agreed-upon coordinated ways of working have emerged from a meeting of U.S. and U.K. database curators this year (*37*).

The difficulties and unknowns notwithstanding, comparative genetics is the key to extending our knowledge of plant genomes and plant genes. Even though the major cereal genomes contain more DNA than the human genome, it is already possible to formulate a maize or wheat genome program. We now know we do not have to start from scratch. It is just a matter of how much we can borrow through comparative genetics.

References

1. M. W. Bonierbale *et al.*, *Genetics* **120**, 1095 (1988).
2. S. Chao *et al.*, in *Proceedings of the 7th International Wheat Genetics Symposium*, T. E. Miller and R. M. D. Koebner, Eds. (Institute of Plant Science Research, Cambridge Laboratory, Cambridge, UK, 1988), pp. 493–498.
3. S. D. Tanksley *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 6419 (1988).
4. U. Lagercrantz *et al.*, *Plant J.* **9**, 13 (1996).
5. D. Neale, personal communication.
6. G. King, personal communication.
7. N. F. Weeden *et al.*, *J. Hered.* **83**, 123 (1992).
8. D. Menancio-Hautea *et al.*, *Theor. Appl. Genet.* **86**, 797 (1993).
9. T. Foote *et al.*, *Genetics* **147**, 801 (1997).
10. A. Kilian *et al.*, *Plant Mol. Biol.* **35**, 187 (1997).
11. E. A. Carver and L. Stubbs, *Genome Res.* **7**, 1123 (1997).
12. M. Chen *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 3431 (1997).
13. J. Muller, *Bot. Rev.* **47**, 1 (1981).
14. U. Lagercrantz and D. J. Lydiate, *Genetics* **144**, 1903 (1996).
15. A. C. Cavell *et al.*, *Genome* **41**, 62 (1998).
16. T. C. Osborn *et al.*, *Genetics* **146**, 1123 (1997).
17. T. C. Osborn, personal communication.
18. R. Schmidt, personal communication.
19. R. Tarchini, A. Rafalski, S. Tingey, personal communication.
20. D. Leister *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 370 (1998).
21. A. Barakat *et al.*, *ibid.* **94**, 6857 (1997).
22. R. Panstruga *et al.*, *Nucleic Acids Res.* **26**, 1056 (1998).
23. P. Schulze-Lefert, personal communication.
24. M. G. Pereira and M. Lee, *Theor. Appl. Genet.* **90**, 380 (1995).

25 P. C. Bailey *et al.*, *ibid.*, in press.
26. M. D. Gale and K. M. Devos, *Proc. Natl. Acad. Sci.* **95**, 1971 (1998).
27. K. M. Devos *et al.*, *Theor. Appl. Genet.* **85**, 673 (1993).
28. H. Zhang *et al.*, *ibid.* **96**, 69 (1998).
29. J. Peng *et al.*, *Genes Dev.* **11**, 3194 (1997).
30. N. P. Harberd, personal communication.
31. A. H. Paterson *et al.*, *Nature Genet.* **14**, 380 (1996).
32. J. L. Bennetzen, personal communication.
33. K. M. Devos, unpublished data.
34. R. A. Martienssen, L. Parnell, W. R. McCombie, personal communication.
35. A. Kleinhofs, personal communication.
36. S. W. Cartinhour, *Plant Mol. Biol.* **35**, 241 (1997).
37. M. D. Gale *et al.*, recommendations from the BBSRC-USDA Bilateral Plant Bioinformatics Planning and Coordination Meeting, Llangollen, UK, 22 to 24 March 1998.

Databases in Genomic Research

William M. Gelbart

VIEWPOINT

Genome-related databases have already become an invaluable part of the scientific landscape. The role played by these databases will only increase as the volume and complexity of relevant biology data rapidly expand. We are far enough into the genome project and into the development of these databases to assess their attributes and to reexamine some of the conceptual organizations and approaches they are taking. It is clear that there are needs for both highly detailed and simplified database views, the latter being especially needed to make expert domain data more accessible to nonspecialists.

Genomic databases are public windows on the high-throughput genome projects. In a sense, the success or failure of genome projects depends on the availability and utility to the scientific community of the data that are produced. Further, the very thrust of high-throughput science is the creation of large, well-organized, and rigorous sets of data. With this greatly increased biological data set that needs to be traversed, a variety of centralized databases are required to present these data in digestible chunks. Given the nature of biology and of database technology, it is probably impossible to determine in advance the database needs of the biological research community, but periodic retrospective analysis is certainly warranted. In this way, success stories can be identified, systematic problems can be assessed, and important gaps in the range of database coverage can be addressed. Having lived a dual existence as both a provider and a consumer of database information, I would like to offer my perspectives on where the genomic/genetic databases presently are and some of the issues that need to be addressed in the near term.

The Current Database Landscape

It is not my intention to exhaustively review the array of important genome-related databases that abound on the Internet. Rather, I would like to make some general classifications and comments. Genome-related databases can be broken into two major groups: generalized and specialized (or expert domain) databases. Generalized databases include the GenBank/EMBL/DDBJ archives of nucleic acids sequences and the PIR and SwissProt polypeptide sequence databases. Such databases capture and present information on particular classes of molecules, without any phylogenetic or functional exclusions. In contrast, the specialized databases do have more limited purviews, such as those organized around a specific model organism or around a type of biological function, such as protein family databases.

Interestingly, none of these generalized or specialized databases solely contain genome project data, but rather they are a mosaic of data from genome projects intermixed with those from the broader scientific community. This is in fact a recognition that the genome projects do not have exclusive license to produce any particular type of data—they are just much larger scale and frequently more accurate or self-consistent sources of particular types of information. In contrast, although the contributions of the community might lack as much data consistency and breadth of coverage, these possible deficiencies are offset by the greater expertise behind the individual contributions, which often are the culmination of years of focused research. The scientific community is best served by seamless integration of the high-throughput genome project data with the focused contributions of high-expertise groups.

Nothing makes a stronger case for such integration than a consideration of our current ability to decipher the information embedded in genomic DNA. The elucidation of the full genomic DNA sequence of humans, for example, has been referred to as the Rosetta Stone of human biology, which implies that it will allow us to elucidate all of the information encapsulated in this DNA sequence. However, it might be more appropriate to liken the human genomic sequence to the Phaestos Disk: an as yet undeciphered set of glyphs from a Minoan palace on the island of Crete. With regard to understanding how to make sense of the A's, T's, G's, and C's of genomic sequence, by and large we are functional illiterates.

Consider all of the structural information required to build a polypeptide chain and all of the regulatory information required to deploy that polypeptide in the correct sets of cells at the proper developmental times and in the requisite quantities. If every set of such information were analogous to one sentence in the instruction manual that we call the genome, a reasonable current assessment is that we have a partial but still quite incomplete knowledge of how to identify and read certain nouns (the structures of the nascent polypeptides and protein-coding exons of mRNAs). Our ability to identify the verbs and adjectives and other components of these genomic sentences (for example, the regulatory elements that drive expression patterns or structural elements within chromosomes) is vanishingly low. Further, we do not understand the grammar at all—how to read a sentence, how to weave the different sentences together to form sensible paragraphs describing how to build multicomponent proteins and other complexes, how to elaborate physiological or developmental pathways, and so on. Finally, we have little knowledge of how to identify and intepret structural information in the genome, such as boundary domains and other punctuation that separate different polypeptide-coding sentences from one another.

Were we to be able to read the genomic instruction manual in the same way we can read a book written in a language we understand, we might not need a huge support system of scientific databases. However, we are nowhere close to being at this point with regard to the genome. For now, the genomic sequence of an organism is written in a language we barely comprehend. However, through the work of the scientific community, we can attach biological meaning to limited regions of the sequence. Until we vastly improve our ability to actually read genomic DNA, we should work toward the goal of attaching all available experimental information as annotations to the framework, or reference, genomic DNA. This should be an important focus for model organism databases, in which substantial genetic information can serve as genomic annotation. Ordinarily, the task of framework sequence annotation should fall to one of the organism-specific expert domain databases. These groups have the specific

The author is in the Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA.