Perspective

A Common Language for Physical Mapping of the Human Genome

MAYNARD OLSON, LEROY HOOD, CHARLES CANTOR, DAVID BOTSTEIN

IN A REPORT ISSUED IN JANUARY 1988, THE NATIONAL Research Council (NRC) Committee on the Mapping and Sequencing of the Human Genome, on which the present authors served, recommended a staged mapping and sequencing project with early emphases on physical mapping of human DNA, mapping and sequencing of the genomes of model organisms, and the development of sequencing technology (*1*). As the Committee's recommendations on physical mapping are beginning to be implemented on a substantial scale, it is timely to review these recommendations in the light of recent technical advances. In particular, the polymerase chain reaction (PCR) (*2*), a method that has only come into widespread use during the past 2 years, seems to us to offer a path toward a physical map that largely circumvents two problems that were prominent in the NRC Committee's discussions. One of these was the difficulty of merging mapping data gathered by diverse methods in different laboratories into a consensus physical map. The second was the logistics and expense of managing the huge collections of cloned segments on which the mapping data would depend almost absolutely.

By allowing short DNA sequences to be detected easily with high specificity and sensitivity, PCR makes practical the use of DNA sequence itself to define the basic landmarks on the physical map. We advocate the use of short tracts of single-copy DNA sequence (that is, sequences that occur only once in the genome) that can be easily recovered at any time by PCR as the landmarks that define position on the physical map. Construction of a physical map would then be seen as the determination of the order and spacing of DNA segments, each of which is identified uniquely by such a sequence. This will solve the problem of merging data from many sources, eliminate the need for large clone archives, and define a physical map that can evolve smoothly and naturally toward the ultimate goal of a complete DNA sequence of the human genome.

Physical mapping: A hybrid technology. The physical map of the human genome envisioned by the NRC report as the precursor of sequencing was a hybrid of a "restriction map" and a "contig map." Following the paradigm introduced by Nathans in the early 1970s for the case of SV40, restriction maps show the order and distances between cleavage sites of site-specific restriction endonucleases (*3*). This type of mapping has been extended to much larger genomes, such as that of *Escherichia coli*, by exploiting the ability to separate very large restriction fragments with pulsed-field gel electrophoresis (*4*). Contig maps represent the structure of contiguous regions of the genome by specifying the overlap relationships among a set of clones (*5*). Contig maps are dependent on the continuing existence of a particular underlying clone collection; the generation and most uses of these maps depend on detailed analysis of individual clones.

Hybrid maps draw on the complementary strengths of restriction maps and contig maps. Pure restriction maps are difficult to construct, primarily because the sites for the most suitable enzymes are distributed nonrandomly and are sometimes blocked by the action of methylation systems that covalently modify DNA in vivo. Furthermore, restriction maps fail to address the need of most map users for ready access to the cloned DNA. Pure contig maps are also difficult to construct because these maps lose continuity at any point where clones are unavailable or overlap relationships are unclear. Indeed, extrapolation from past experience suggests that a contig map of a human chromosome of average size would be likely to contain between 200 and 1000 gaps. In a hybrid map, restriction maps based on the direct analysis of uncloned DNA—as well as data from other low-resolution mapping sources such as linkage mapping, cytogenetics, and somatic cell genetics—are used to orient and align a series of contigs. In favorable cases, the resultant maps have good long-range continuity and are supported by clone collections that cover a high fraction of the mapped region.

Sequence-tagged sites (STSs) will enhance the hybrid mapping strategy. The present proposal is not an alternative to the strategy described for mapping the human genome: the STS proposal redefines the end product, and is not itself a new mapping method. The idea would be to "translate" all types of mapping landmarks into the common language of STSs. Virtually any useful mapping method uses cloned DNA segments as landmarks, regardless of whether they are members of contigs, segments that contain an unusual restriction site, probes that detect genetically mapped DNA polymorphisms, or sequences that hybridize in situ to particular cytogenetic bands. In practice, the translation of any of these examples to produce an STS would simply require sequencing a short tract of DNA from the clone that defines the landmark.

In most instances, 200 to 500 bp of sequence define an STS that is operationally unique in the human genome (that is, can be specifically detected via PCR in the presence of all other genomic sequences). A PCR assay for an STS could be implemented simply by synthesizing two short (~20 nucleotides) oligodeoxynucleotides, chosen to be complementary to opposite strands and opposite ends of the sequence tract. A DNA sample would be tested for the presence of the sequence by testing its capacity to serve as a template for the in vitro synthesis of the tract in the presence of these two oligodeoxynucleotide "primers." The procedure involves many automated cycles of DNA synthesis in a standard laboratory thermocycler; consequently, when the assay is positive, such large amounts of product are made that it can be detected without radioactive labeling.

The overwhelming advantage of STSs over mapping landmarks defined in other ways is that the means of testing for the presence of a particular STS can be completely described as information in a database. No access to the biological materials that led to the definition or mapping of an STS is required by a scientist wishing to assay a DNA sample for its presence. An entry in the STS database would not only include raw sequence data on which a PCR-based STS assay could be based, but also would include detailed instructions for implementing a well-tested PCR assay. From such information alone, the assay could be implemented by any laboratory within 24 hours.

STSs facilitate assimilation of mapping data from diverse sources into an

M. Olson is a professor of genetics, Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110. L. Hood is director, NSF Science and Technology Center for Biotechnology, Division of Biology, California Institute of Technology, Pasadena, CA 91125. C. Cantor is the director, Human Genome Center, Lawrence Berkeley Laboratory, Berkeley, CA 94720. D. Botstein is vice president–science, Genentech, Inc., South San Francisco, CA 94080.

evolving physical map. Although its independence of stored biological materials is the central virtue of an STS map, there is a corollary virtue that could prove equally powerful. By providing a common language for physical mapping projects, the use of STSs would allow incorporation of any type of physical mapping data into the evolving map. It would be straightforward, for example, for someone constructing contigs in a given region to scan the contigs for the presence of STSs assigned to the region by another method. Similarly, it would be straightforward to compare two contigs constructed in different laboratories. The importance of having a common language that would facilitate comparisons cannot be overemphasized. The central managerial problem for the human genome project will be to introduce sufficient project accountability and quality control to make sure that resources are used efficiently and that the final product is valid. Without a common language, it is not obvious how this challenge can be met.

Adoption of an STS standard would impart to physical mapping some of the most attractive features of genetic mapping. A crude STS map of the whole genome could be constructed rather quickly. Many of the distance estimates would be poorly defined and some of the site orderings would be uncertain. Nonetheless, this map could evolve smoothly toward a more refined product with inputs from many laboratories and methodologies. Significantly, the dichotomy between "big" and "small" laboratories would disappear. Some large mapping programs will undoubtedly be required to produce a global, high-resolution map, but small laboratories could both draw data from the evolving map and contribute to it as they pursue the detailed analysis of local regions. Finally, STS maps of local regions—and ultimately the whole genome—would converge smoothly toward "exactness," as DNA sequence data accumulate. The only limitation would then be the degree of DNA sequence polymorphism within the human population.

Implementation of the STS proposal. We recommend that several steps be taken now to establish the STS map as the centerpiece of the human physical mapping effort. An STS standard should be adopted that specifies the information required to define an STS and the way in which new STS assays should be tested. Planning for a central STS database and a review process for data entries should also begin. International discussions should be initiated to maximize the likelihood that use of the new common language would cross national boundaries.

To guide resource allocation, we need to set a 5-year goal for the resolution of the STS map. The main practical requirement is that the resolution should be high enough to make regeneration of cloned coverage of any region straightforward. A PCR-based STS assay could be used to screen libraries directly; alternatively, PCR-amplified STS could be labeled by standard methods and used for colony screening. Indeed, testing of the usefulness of labeled, amplified product as a single-copy hybridization probe should be part of the STS standard. If this criterion is met, STSs would actually be much better "reagents" than clones themselves for use during the screening of new libraries. The ability to regenerate a cloned region is a critical concept both because it alleviates the need for large, permanent clone archives and because it protects against "clone obsolescence." Cloning technology is certain to continue to evolve rapidly, and molecular biologists 10 years from now are not going to want to base their work on clones that were prepared in the 1980s.

With respect to the resolution that will be required for an STS map with good practical coverage, a key question is the capacity of cloning vectors. Present cosmid cloning systems have maximum capacities of approximately 40 kb (*6*). Yeast artificial chromosome (YAC) vectors can be used to clone segments of several hundred kilobase pairs (*7*). A plausible 5-year goal for the resolution of an STS map therefore might be an average spacing of 100 kb, requiring the mapping of 30,000 STSs. At this resolution, one-step recovery would be possible for most regions of the genome by cloning in the YAC system, but not in cosmids. If one-step recovery in cosmids still appears to be an essential goal in 5 years, it will be more attractive to take the STS map to higher resolution than to retreat to the concept of permanent clone archives that have to be cataloged, stored, and shipped in order to be useful. Such archives, if based on cosmid technology, would have to contain several hundred thousand clones; not only are the sheer numbers intimidating, but the stability of these clones over time and during regrowth would be constantly open to question.

Existing useful clones can be converted to STSs. Major momentum could be imparted to the human genome project if the scientific community begins at once to convert existing sets of mapped DNA probes to STSs. It is likely that 2000 to 3000 useful probes could be identified for which approximate map positions are already known. These probes could be converted to STSs on a contract basis. Once accomplished, the database would comprise a direct precursor to a low-resolution physical map. Many of the inter-site spacings would have to be estimated from linkage or cytogenetic data, but the very process of refining these estimates would lay the basis for ultimate integration of the physical, genetic, and cytogenetic maps of the human.

In conclusion, the technical means have become available to root the physical map of the human genome firmly in the DNA sequence itself. Sequence information is the natural language of physical mapping. Lest we replay the failed effort to build the Tower of Babel, it would be wise to move decisively toward its adoption.

REFERENCES

1. National Research Council, *Mapping and Sequencing the Human Genome* (National Academy Press, Washington, DC, 1988).
2. R. K. Saiki *et al.*, *Science* **230**, 1350 (1985); R. K. Saiki *et al.*, *ibid.* **239**, 487 (1988).
3. D. Nathans, *ibid.* **206**, 903 (1979).
4. C. L. Smith *et al.*, *ibid.* **236**, 1448 (1987).
5. A. Coulson *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 7821 (1986); M. V. Olson *et al.*, *ibid.*, p. 7826; Y. Kohara, K. Akiyama, K. Isono, *Cell* **50**, 495 (1987).
6. B. Hohn and J. Collins, *Gene* **11**, 291 (1980); G. A. Evans, K. Lewis, B. E. Rothenberg, *ibid.* **79**, 9 (1989).
7. D. T. Burke, G. G. Carle, M. V. Olson, *Science* **236**, 806 (1987); B. H. Brownstein *et al.*, *ibid.* **244**, 1348 (1989).