A Time to Sequence

Maynard V. Olson

My views on the Human Genome Project first appeared in *Science* in 1987 in the form of a one-word quotation: "Huge" (*1*). Eight years later, in deference to the size and complexity of the current program, I have been allotted more space. I use it here to examine the state of the project and to suggest a path forward.

The basic game plan for an organized Human Genome Project in the United States was established by a National Research Council committee, chaired by Bruce Alberts (*2*). Indeed, I was testifying before this committee in 1987 when my assessment of the project's scale caught a *Science* reporter's attention. The Alberts Committee recommended an early emphasis on genetic linkage mapping and clone-based physical mapping of human DNA. In parallel, the committee recommended research on the technology of DNA sequencing, as well as pilot-scale sequencing of the genomes of model organisms. This approach was viewed as the best way to improve the reliability of DNA sequencing—and to drive down its cost—while simultaneously gathering data of immediate biological value.

To a remarkable degree, the Alberts Committee read the historical and technical trends correctly. It now appears that even its estimates of time scale and cost—15 years at $200 million per year—were about right. In 1987, skeptics could still argue about basic feasibility with some force. Conversely, many of the project's proponents lacked a realistic sense of the diversity of problems that had to be solved before mammalian-scale sequencing would become practical. Even the Alberts Committee's middle-of-the-road recommendations would likely have proven to be overambitious if it were not for several unforeseen developments. Technically, the most important of these has been the emergence of the polymerase chain reaction (PCR) as a primary tool for DNA analysis. Rapid advances in computer technology have also been significant, particularly because they have allowed most of the project's data-analysis and data-management needs to be met by the distributed efforts of small groups of programmers working in close collaboration with experimentalists. Finally, vigorous international participation in the project has materialized, a development that the Alberts Committee strongly encouraged but could not count on.

The policy success of defining and implementing a program of this complexity in the face of rapidly evolving technology—and on a relatively austere budget—provides grounds for satisfaction (*3*). Nonetheless, the project's greatest challenge lies ahead. The preliminary phase of the Human Genome Project emphasized diverse lines of research, many of which could be pursued in conventional molecular biology laboratories. Much of this activity must ultimately be displaced by a more monolithic sequencing program, largely focused on human DNA. Neither the Alberts Committee nor subsequent policy reviews (*4*) provide clear guidance on how or when to carry out this transition. Recently, proponents of an early and aggressive move to very large-scale sequencing of human DNA have emerged from among the leaders of model-organism sequencing initiatives (*5*). In this Policy Forum, I add my support to their proposal. The case in favor of an early transition to human sequencing rests on an assessment of three questions: Are the maps good enough? Is the technology strong enough? and Would it be good policy?

The Maps

Almost certainly, the maps are good enough. This assessment rests on the current state of the maps, the rate at which they are improving, and the advantages of combining the last stages of physical mapping with sequencing. The dominant low-resolution mapping paradigm is sequence-tagged site (STS)–content mapping, applied either to comprehensive yeast artificial chromosome (YAC) libraries (*6*) or to panels of human-rodent hybrid cell lines that contain multiple segments of human DNA [that is, "radiation-hybrid," or RH, cell lines (*7*)]. These forms of mapping define the order of STSs, which are short, unique DNA sequences most commonly detected by PCR assays (*8*). STS ordering is inferred from data on the STS content (that is, presence or absence of particular STSs) in the random segments of the human genome present in a set of clones that has been organized into a "typing resource." The ability of a typing resource to resolve the order of STSs is determined by the average spacing between segment ends, typically 50 to 100 kbp in current resources. Maps with an average spacing between STSs of approximately 100 kbp already exist for perhaps 15% of the genome. Approximately half the genome has been mapped only by whole-genome approaches that thus far have produced average spacings closer to 300 kbp. The balance of the genome is at an intermediate state. There are also regions that have progressed beyond, or even bypassed, the STS-mapping stage, but they constitute only a small fraction of the total.

Because efficient screening methods exist for finding new clones that contain a particular STS (*9*), the choice of which clones to sequence at a particular site in the genome can be made immediately before the sequencing is carried out. There is presently healthy competition between cloning systems such as cosmids, P1-based clones, and bacterial artificial chromosomes (BACs), all of which provide plausible ways to clone the DNA that will actually be sequenced (*10*). The recombinant DNA molecules generated by these cloning systems contain 40 to 200 kbp of human DNA. Various "fingerprinting" and "contig-building" strategies allow contigs (that is, collections of overlapping clones that collectively cover the target region) to be built whose lengths are typically a few times the size of the clones from which they are constructed (*11*).

Because the spacings between mapped STSs are already comparable to the sizes of contigs that can be readily seeded around an STS, even current maps would allow much of the genome to be covered with well-mapped clones that are suitable for sequencing. Current mapping projects have enough momentum to reduce average STS spacings to 100 kbp throughout the genome within a year or two. Even with these maps, it is inevitable that there will be many clones sequenced whose precise genomic positions and left-right orientations cannot be determined simply from their STS content. However, it would be sensible to handle these cases by developing additional STSs at the ends of those sequenced clones whose positions and orientations are uncertain, rather than to continue random STS mapping to an unnecessarily high resolution throughout the genome. This strategy would answer the question: How good does the physical map need to be? with the most economical possible answer—just good enough to allow all sequence tracts to be aligned with it.

The resolution of the physical map required to support sequencing exceeds that needed to maintain alignment between the physical map, the genetic linkage map, and the cytogenetic map. Therefore, as the sequence of the human genome emerges, it will be possible to align sequence tracts with the genetic and cytogenetic maps, as well as the physical map, thereby allowing correlations between

The author is in the Department of Molecular Biotechnology, University of Washington, Seattle, WA 98195, USA.

particular sequences and observations on mutant human chromosomes.

The Technology

The question of whether or not sequencing technology is adequate for a near-term, massive increase in the scale of genomic sequencing is more troublesome than the corresponding question about maps. Many participants in the Human Genome Project, including this author, envisioned the project as a vehicle for developing powerful new sequencing tools that would displace the techniques of the 1980s through a combination of fundamental advances and automation.

What has happened instead is arguably a better development for experimental biology. Sequencing methodology has improved incrementally in a way that is leading to convergence, rather than divergence, between the methods employed in "genome centers" and those used in more typical molecular biology laboratories. Following a period of competition between alternative sequencing strategies, a dominant technology has emerged for large-scale genomic sequencing: Clones the size of cosmids or larger are analyzed by random sampling (that is, "shotgun" sequencing), implemented on commercial, four-color fluorescence sequencing instruments (*12*). The optimum size of the starting clones, the level of detail with which these clones should be mapped, and the extent to which random sampling should be supplemented by more "directed" methods remain contentious. However, the important news is that the basic approach works in any of several well-tested variations.

In retrospect, the idea that sequencing technology would be displaced in a few years by fundamentally new approaches was implausible. Major advances in analytical techniques are neither more frequent nor any easier to stimulate programmatically than are other scientific developments. Gel electrophoresis was first used to separate biological macromolecules on 23 January 1954 (*13*). Forty years later, it has changed relatively little while playing a key role in one revolutionary discovery after another in basic biology.

The envisioned transition to high-level automation of DNA sequencing was also unrealistic. The Human Genome Project lacks both the financial and human resources to bring it about. Clearly, contemporary sequencing procedures could be fully automated with a sufficient engineering investment. Less clear is how long it would take, what it would cost, and how competitive the result would be with more labor-intensive methods. The most realistic policy would be to continue to seek efficiency gains through the piecemeal introduction of labor-saving devices. As sequencing is implemented on a larger scale and cost containment becomes a paramount concern, it should become progressively easier to spot bottlenecks that could be overcome by specialized equipment.

An uncomfortable corollary to the emergence of a dominant technology is that it is time to curtail support for competing approaches. Small-scale exploration of genuinely novel approaches remains appropriate. However, it is time to recognize that genomic sequencing is in the coalescence phase of the alternating periods of competition and coalescence by which complex technologies lurch from one generation to the next. During this phase, the dominant technology improves rapidly and declines in cost just because it is dominant.

Policy Implications

Even if the maps and technology are judged adequate, there remains a question as to whether or not it is a good idea to divert resources from other activities to large-scale human sequencing. Program areas that would be adversely affected, together with brief arguments supporting their importance, are summarized below.

Technology. Further technological development would reduce the cost of human sequencing and allow the sequencing of other genomes. Overinvestment in sequencing capacity on the basis of current technology would suppress innovation and create large facilities that would rapidly become obsolete.

Informatics. Data collection is outstripping current capabilities to annotate, store, retrieve, and analyze maps and sequences. Better computational tools will be necessary before biologists will be able to make effective use of the data.

Disease. An important motivation for the Human Genome Project is to make it easier to analyze human genetic diseases. Activities such as intensive mapping of expressed-sequence tags and light sampling of genomic sequence provide the cheapest and fastest route to this goal.

Gene function. Advances in molecular biology are most effectively driven by functional studies. The Human Genome Project should partition its resources between gene discovery and studies of the functions of the genes that are being discovered.

Genetic variation. Much of the biological interest in the human genome lies in generic variation and its relation to phenotype.

Model organisms. Basic cellular mechanisms can be studied more effectively in model organisms than in the human. The lessons learned in these systems are often readily transferable to the human because of the evolutionary conservation of critical genes. The list of model organisms under analysis could be expanded at modest cost since most model organisms have relatively small genomes.

Human resources. The development of genome centers and other laboratories with expertise in state-of-the-art methods is as important a goal as data collection. These laboratories are essential training resources and ensure widespread access to genome analysis tools. Continuity in the support of current programs should not be endangered by rapid shifts in programmatic emphasis.

These arguments underscore the need to maintain some balance amongst the Human Genome Project's diverse goals. They also make clear that genome analysis will face expanding, rather than contracting, opportunities once the human genome has been sequenced. Nonetheless, at the present juncture, the more compelling scientific and policy arguments favor a tightly focused Human Genome Project.

Genetic first principles favor early acquisition of a complete genomic sequence. The digital information that underlies biochemistry, cell biology, and development can be represented by a simple string of G's, A's, T's, and C's. This string is the root data structure of an organism's biology. Genetic and cytogenetic maps, as well as vast amounts of biochemical data, can be overlaid on the genome sequence in a natural way.

The financial costs of delay would exceed plausible savings from gains in efficiency. The Human Genome Project presently has a budget of approximately $200 million per year in the United States alone. The current cost of converting good STS maps to genomic sequence appears to be in the range of $0.20 to $0.40 per base pair. Costs will undoubtedly decline as economies of scale are realized. Hence, the total cost of producing a high-quality human sequence is likely to be less than $1 billion. Given present budgetary levels, the wait-and-see costs of an overly cautious policy would mount to $1 billion in just a few years. In all likelihood, the hidden costs of delayed availability of the data would be still larger because the sequence of the human genome would have broad effects on the efficiency of biomedical research.

Goal-oriented science projects are bad policy unless they have a well-defined objective. A vaguely defined Human Genome Program would be a bad compromise between targeted and investigator-initiated research. The more discipline that the project displays in setting priorities, the less it will threaten the curiosity-driven, small-laboratory science that is the best route to sustained scientific innovation. By shortening the path from observation to hypothesis to experimental test, the sequence of the human genome will empower small laboratories to attack problems in human biology that are presently beyond the reach of even the largest research teams.

International participation will be favored by an unequivocal commitment to very large-scale sequencing of human DNA. Different countries have diverse methods of organizing and supporting science. Efforts to negotiate common programs will collide with this diversity unless the goal and time schedule for a project are both clear. If the Human Genome Project in the United States moves decisively toward genomic sequencing, many other countries may be expected to join the effort, each mobilizing the needed support in its own way. The European yeast sequencing effort, spearheaded by the European Economic Community, achieved precisely this result after its pioneering commitment to obtain a complete sequence of the *Saccharomyces* genome. Increased international participation will allow sharing of the high financial cost of the Human Genome Project, while also securing a legacy of joint human participation in this important step in our genetic self-characterization.

Dynamic resource allocation works. The Human Genome Project in the United States has achieved consistent scientific success by allocating nearly all its resources through peer-reviewed grants that extend for 3 or 4 years. Competition within this system is intense, and many grants are not renewed even when they have met or exceeded their goals. This paradox is unavoidable in an applied science project with sequentially dependent objectives. While there are inefficiencies associated with this system, they pale beside those that result when permanent institutions are created that tie science to the past rather than the future.

There is a less abstract argument for moving ahead with human sequencing: That is what the money is for. The Human Genome Project was not sold to the U.S. Congress as a generalized vehicle for increasing support for molecular genetics, medical genetics, bioinformatics, or instrumentation development. It was sold on the grounds that sequencing the human genome would be immensely useful, was becoming technically feasible, and would not happen by itself. The foundations of this argument are worth revisiting. Substantial public resources have been invested in studies of the molecular genetics of model organisms. This investment was largely motivated by the perceived relevance of the research to human health. The Human Genome Project was designed both to make the human system easier to study directly and to increase the "bandwidth" for knowledge transfer between model-organism and human biology. The power of genome analysis to facilitate these goals is already well demonstrated (*14*). Completion of the sequence of the human genome and the sequences of the genomes of key model organisms will mobilize the full benefits of this new approach to biology.

While huge, the central task of the Human Genome Project is bounded by one of the most remarkable facts in all of science: The development of a human being is guided by just 750 megabytes of digital information. In vivo, this information is stored as DNA molecules in an egg or sperm cell. In a biologist's personal computer, it could be stored on a single CD-ROM. The Human Genome Project should get on with producing this disk, on time and under budget.

REFERENCES

1. R. Lewin, *Science* **235**, 747 (1987).
2. National Research Council, *Mapping and Sequencing the Human Genome* (National Academy Press, Washington, DC, 1988).
3. D. E. Koshland Jr., *Science* **266**, 199 (1994).
4. F. Collins and D. Galas, *ibid.* **262**, 43 (1993).
5. E. Marshall, *ibid.* **267**, 783 (1995); *ibid.* **268**, 1270 (1995).
6. E. D. Green and M. V. Olson, *ibid.* **250**, 94 (1990); E. D. Green and P. Green, *PCR Methods Appl.* **1**, 77 (1991); S. Foote, D. Vollrath, A. Hilton, D. C. Page, *Science* **258**, 60 (1992); I. Chumakov *et al.*, *Nature* **359**, 380 (1992).
7. D. R. Cox, M. Burmeister, E. R. Price, S. Kim, R. M. Myers, *Science* **250**, 245 (1990).
8. M. Olson, L. Hood, C. Cantor, D. Botstein, *ibid.* **245**, 1434 (1989).
9. E. D. Green and M. V. Olson, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 1213 (1990).
10. J. C. Pierce, B. Saver, N. Sternberg, *ibid.* **89**, 2056 (1992); H. Shizuya *et al.*, *ibid.*, p. 87794; P. A. Ioannou *et al.*, *Nat. Genet.* **6**, 84 (1994).
11. A. Coulson, J. Sulston, S. Brenner, J. Karn, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 7821 (1986); M. V. Olson *et al.*, *ibid.*, p.7826; Y. Kohara, K. Akiyama, K. Isono, *Cell* **50**, 495 (1987).
12. T. Hunkapiller, R. J. Kaiser, B. F. Koop, L. Hood, *Science* **254**, 59 (1991); R. Wilson *et al.*, *Nature* **368**, 32 (1994); M. Johnston *et al.*, *Science* **265**, 2077 (1994); R. D. Fleischmann *et al.*, *ibid.* **269**, 496 (1995).
13. O. Smithies, *Genetics* **139**, 1 (1995).
14. M. V. Olson, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 4338 (1993).

For the opportunity to participate in a discussion of the issues raised in this Policy Forum, go to the following URL (http://sci.aaas.org/aaas/policy).