POLICY: BIOMEDICINE

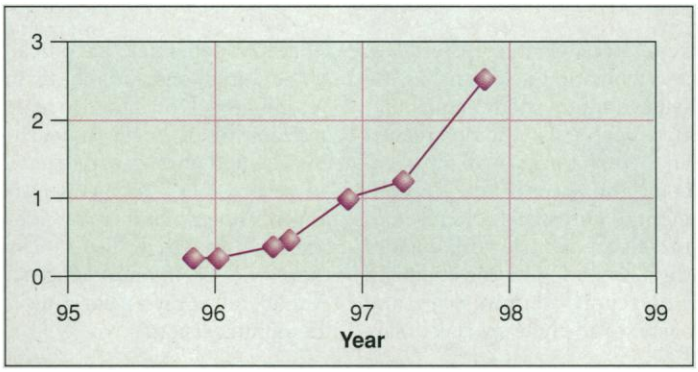
An Independent Perspective on the Human Genome Project

Steven E. Koonin

The U.S. Human Genome Project (HGP) is a joint effort of the Department of Energy and the National Institutes of Health, formally initiated in 1990. Its stated goal is ". . . to characterize all the human genetic material—the genome—by improving existing human genetic maps, constructing physical maps of entire chromosomes, and ultimately determining the complete sequence . . . to discover all of the more than 50,000 human genes and render them accessible for further biological study." The original 5-year plan was updated and modified in 1993 (*1*, *2*).

DOE's Office of Biological and Environmental Sciences recently chartered the JASON group to review the DOE component of the HGP. This group, mainly consisting of physical and information scientists, was asked to consider three areas: technology, quality assurance and quality control, and informatics. This article summarizes the group's findings and recommendations (*3*).

Technology. The present state of the art for determining the sequence of DNA is defined by Sanger sequencing, in which DNA fragments are labeled by fluorescent dyes and separated according to length with polyacrylamide gel electrophoresis (PAGE) (*4*). The base at the end of each fragment can then be visualized and identified by the dye with which it reacts. Although more than 95% of the genome remains to be sequenced, roughly 55 megabases (Mb) have been completed in the past year (see the figure). The world's large-scale sequencing capacity (not all of which is applied to the human genome) is estimated to be roughly 100 Mb per year. It is sobering to contemplate that an average production of 400 Mb will be required each year to complete the human sequence by the target date of 2005.



Percentage of the human genome sequenced to date. Almost 3% of the genome has been sequenced in contiguous stretches longer than 10 kb and is now deposited in publicly accessible databases. Compiled by J. Roach, as described in http://weber.u.washington.edu/~roach/human_genome_progress2.html.

The present technology has only a limited read-length capability (the number of contiguous bases that can be identified from each fragment); the best current practice can read 700 to 800 bases, with perhaps 1000 bases as the ultimate limit. Because the DNA segments of interest are much longer than this [40 kilobases (kb) for a cosmid clone; 100 kb or more for a bacterial artificial chromosome or a gene], the present technology requires that long lengths of DNA be cut into overlapping short segments (~1 kb in length) that can be sequenced directly. The sequences from these shorter pieces must then be assembled into the final sequence. Up to 50% of the effort at some sequence centers goes into this final assembly and finishing of the sequence. The ability to read longer fragments would step up the pace and quality of sequencing.

Apart from the various genome projects, however, there is little pressure to achieve longer read lengths. The 500 to 700 base lengths read by the current technology are well suited to many scientific needs, including pharmaceutical searches, studies of some polymorphisms, and studies of some genetic diseases.

Other drawbacks of the present technology include the time- and labor-intensive nature of gel preparation and running, as well as the comparatively large amounts of sample required, which also increases the cost of reagents and necessitates extra amplification steps.

Thus, the present sequencing technology leaves much to be desired and must be supplanted in the long term if the potential for genomic science is to be fully realized. Promising methods that could be cheaper and faster than PAGE include single-molecule sequencing, mass spectrometric methods, hybridization arrays, and microfluidic capabilities. None of these is sufficiently mature, however, to be a candidate for near-term major scale-up. It is therefore important to support research aimed at improving the present method. Advances in hardware development could, for example, increase the lateral scan resolution of the machine so that more lanes of a gel can be analyzed. The genome community should unify its efforts to enhance the performance of present-day instruments.

Better software will improve the lane tracking, base identification, assembly, and finishing processes. Many of the problems of base identification also occur in the demodulation of signals in communication and magnetic recording systems, and some of the existing literature in these areas should be used by the HGP. The ability to correctly assemble a final sequence without manual editing would markedly speed up the process. It would also be helpful to develop a common set of finishing rules.

Because sequencing technology should (and is likely to) evolve rapidly, the large-scale sequencing centers must be flexible enough to incorporate new technologies. There is a great need to support the development of non-PAGE–based sequencing that goes beyond the current goals of a faster version of PAGE. The funding for such advanced technology is a small fraction of the total HGP but should be increased by approximately 50%.

Quality assurance and quality control. DOE and NIH are recognizing that the HGP must make data accuracy and data quality integral to its execution. A high-quality database can provide useful, densely spaced markers across the genome and enable large-scale statistical studies. A quantitative understanding of data quality across the whole genome sequence is thus almost as important as the sequence itself. Among the top-level steps that should be taken are allocating resources specifically for quality issues and establishing a separate research program for quality assurance and control (perhaps a group at each sequencing center).

The author is professor of Theoretical Physics and vice president and provost at the California Institute of Technology. He led the JASON study reported on in this article. E-mail: koonin@caltech.edu

The stated accuracy goal of the HGP is one error in 10^4 bases, which is set to be less than the polymorphism rate. However, this has been a controversial issue, as genomic data of lower accuracy are still of great utility. For example, pharmaceutical companies searching for genes can use short sequences (400 bases) at an accuracy of one error per 100 bases. The debate on error rates should focus on the level of accuracy needed for each specific scientific objective or use of the genome data. The necessity of finishing sequences without gaps should be subject to the same considerations.

In the real world, accuracy requirements must be balanced against what users need, the cost, and the capability of the sequencing technology to deliver a given level of accuracy. Establishing this balance requires an open dialogue among the sequence producers, sequence users, and the funding agencies, informed by quantitative analyses and experience.

Assays should be developed that can accurately and efficiently measure sequence quality. For example, it would be appropriate to develop, distribute, and use "gold standard" DNA samples that could be used routinely by the whole sequencing community for assessing the quality of the sequence output.

Research into the origin and propagation of errors through the entire sequencing process is fully warranted. We see two useful outputs from such studies: (i) more reliable descriptions of expected error rates in final sequence data, as a companion to database entries; and (ii) "error budgets" to be assigned to different segments of mapping and sequencing processes to aid in developing the most cost-effective strategies for sequencing and other needs.

DOE and NIH should solicit and support detailed Monte Carlo computer simulation of the complete mapping and sequencing processes. The basic computing methods are straightforward: a reference segment of DNA (with all of the peculiarities of human sequence) is generated and subjected to models of all steps in the sequencing process; individual bases are randomly altered according to errors introduced at the various stages; and the final reconstructed segment or simulated database entry is compared with the input segment and errors are noted.

Results from simulations are only as good as the models used for introducing and propagating errors. For this reason, the computer models must be developed in close association with technical experts in all phases of the process being studied, so that they best reflect the real world. This exercise will stimulate new experiments to validate the error-process models and thus will lead to increased experimental understanding of process errors as well.

Improved software is needed to enhance the ability of database centers to check the quality of submitted sequence data before its inclusion in the database. Many of the current algorithms are highly experimental and will be improved substantially over the next 5 years. In addition, an ongoing software quality assurance program should be considered for the large community databases, with advice from commercial and academic experts on software engineering and quality control. It is appropriate for the HGP to insist on a consistent level of documentation, both in the published literature and in user manuals, of the methods and structures used in the database centers that it supports. DOE and NIH should also decide on standards for the inclusion of quality metrics for base identification and DNA assembly along with every database entry submitted.

Informatics. Genome informatics is a child of the information age, a status that brings clear advantages and new hurdles. Managing such a diverse, large-scale, rapidly moving informatics effort is a considerable challenge for both DOE and NIH. The infrastructure supporting the requisite software tools ranges from small research groups (for example, for local special-purpose databases) to large Genome Centers (for process management and robotic control systems) to community database centers (for GenBank and the Genome Database). The resources that each of these groups can put into increasing software sophistication, into ensuring ease of use, and into quality control vary widely. Thus, in informatics areas requiring new research (such as gene finding), a broad-based approach of "letting a thousand flowers bloom" is most appropriate. At the other end of the spectrum, DOE and NIH must impose community-wide standards for software consistency and quality in areas of informatics in which a large user community will be accessing major genome databases.

DOE and NIH should adhere to a bottom-up, customer approach to informatics. Part of this process would be to encourage forums, including close collaborative programs, between the users and providers of informatics tools, with the purposes of determining what tools are needed and of training researchers in the use of new methods.

To ensure that all the database centers are user-oriented and that they are providing services that are genuinely useful to the genome community, each database center should be required to establish its own "users group" (as is done by facilities as diverse as the National Science Foundation's Supercomputer Centers and NASA's Hubble Space Telescope). Further, informatics centers must be critically evaluated as to the actual use of their information and services by the community.

Data formats, software components, and nomenclature should be standardized across the community. If multiple formats exist, it would be worthwhile to invest in systems that can translate among them. Data archiving, data retrieval, and data manipulation should be modularized so that one database is not overextended, and several groups should be involved in the development effort. The community should be supporting several database efforts and promoting standardized interfaces and tools among those efforts.

Final notes. The HGP involves technology development, production sequencing, and sequence utilization. Greater coupling of these three areas can only improve the project. Technology development should be coordinated with the needs and problems of production sequencing, whereas sequence generation and informatics tools must address the needs of data users. Promotion of such coupling is an important role for the funding agencies.

The HGP presents an unprecedented set of organizational challenges for the biology community. Success will require setting objective and quantitative standards for sequencing costs (capital, labor, and operations) and sequencing output (error rate, continuity, and amount). It will also require coordinating the efforts of many laboratories of varying sizes supported by multiple funding sources in the United States and abroad.

A number of diverse scientific fields have successfully adapted to a "big science" mode of operation (nuclear and particle physics, space and planetary science, astronomy, and oceanography are among the prominent examples). Such transitions have not been easy on the scientists involved. However, in essentially all of these cases, the need to construct and allocate scarce facilities has been an important organizing factor. No such centralizing force is apparent in the genomics community, but the HGP is very much in need of the coordination it would produce.

References and Notes

1. F. Collins and D. Galas, *Science* **262**, 43 (1993).
2. The status and challenges of the HGP have been recently reviewed [L. Rowen *et al.*, *ibid.* **278**, 605 (1997)].
3. The MITRE Corporation, JASON Report JSR-97-315 (The MITRE Corporation, McLean, VA, 1997). The participants included S. Block, J. Cornwall, W. Dally, F. Dyson, N. Fortson, G. Joyce, H.J. Kimble, N. Lewis, C. Max, T. Prince, R. Schwitters, P. Weinberger, and W. H. Woodin.
4. For a basic discussion and explanation of the terminogy used, see http://www.ornl.gov/TechResources/Human_Genome/publicat/primer/intro.html