REPORT ON THE HUMAN GENOME INITIATIVE for the OFFICE OF HEALTH AND ENVIRONMENTAL RESEARCH

Prepared by the
Subcommittee on Human Genome
of the
Health and Environmental Research Advisory Committee
for the
U.S. Department of Energy
Office of Energy Research
Office of Health and Environmental Research

April 1987

Appendix A. Analysis of Costs.

General. The total cost of sequencing the human genome will certainly fall in the billion dollar range, although it is important to stress that the actual cost will be very sensitive to the state-of-the-art technologies associated with DNA sequencing, and the related requirements for automation of procedures for cloning, mapping, data handling and data analysis. As an example, compare the current and projected future costs for DNA sequencing and their corresponding implications for sequencing the human genome.

Estimated Cost for Determining the DNA Sequence of a Human Genome (Given Unique Fragments)*

SOURCE	COST	GENOME COST
Current commercial laboratories	$1/base	$6 billion
Japanese sequencing machines	$0.17/base	$1 billion
Future cost with automation	$0.01/base	$60 million

*This estimate does not include the cost of isolating and ordering the fragments; it only includes sequencing each DNA strand, or 6 billion bases. Sequencing both strands provides a check on the accuracy of the sequence.

This table illustrates the importance of making substantial initial investments in technology. We emphasize that the above estimates do not include costs for cloning, mapping or data analysis. Thus our proposal for sequencing the human genome would necessarily be staged. The first 5 years would focus on three general objectives: 1) mapping the human genome, 2) development of technology, 3) sequencing of selected chromosomal regions.

Advances in technology are a necessary first step in sequencing the human genome. These advances will make large-scale sequencing and subsequent comparative studies practical and cost-effective. At present the only automated sequencing machines are based on Sanger's method. There are probably distinct advantages to be gained from automating the Maxam-Gilbert method. A detailed comparison of the two approaches should precede a major investment in one of them. Both approaches can also benefit from considerable optimization. A twofold increase in the length of sequence accessible on a single gel lane would cut the cost of sequencing by considerably more than a factor of two. A number of ways to increase this sequencing range, such as pulsed-field techniques, are very promising and need to be tested.

Multiplex sequencing techniques such as those being developed by George Church are still in their infancy. However, their potential attractiveness is so great that a careful evaluation and refinement of such methods is surely warranted before one embarks on large-scale sequencing. Direct physical approaches to sequence determination such as mass spectrometry or scanning tunneling microscopy are speculative, but their potential impact must not be overlooked. Such approaches should be critically tested in the next few years.

Current strategies for using any of the existing sequencing methods are mostly shotgun approaches which sequence random fragments of DNA. These are quite inefficient since they require sequencing the same region many times over. Sequencing of overlapping fragments is needed to determine the order of the fragments; this is called a bottom-up approach. Phased, or top-down approaches, including systematic ordering and mapping, linked library construction, and optimized production of DNA fragments will all result in far less redundancy in the sequencing. These preliminary steps probably represent half of the final cost and require more than half of the skilled labor. Each of these preliminaries to the actual acquisition of sequence data needs full exploration, refinement and optimization. Most of these preliminaries can and should be automated. Very exciting developments, like methods for cloning or purifying large DNA fragments, and schemes for orderly generation of nested sets of DNA pieces are so new that their potential cannot yet be evaluated. However, it is inevitable that some of these methods will have to be incorporated into any effective large scale sequencing effort.

Once the speed, error rate, and cost are appropriate then one can begin the organized and coordinated effort to sequence a reference human genome. The technologies will then be sufficient to sequence other genomes and to examine human polymorphisms. The wide range of technologies that must be developed for this project are outlined below.

Technologies Required for Sequencing the Human Genome

1. Production of DNA fragments containing 100 to 1000 kilobases

    a. Chromosome separation
    b. Sequence-specific chemical and enzymatic scissors (restriction enzymes)
    c. Separation and purification of large fragments
    d. Large-insert cloning

2. Automated DNA handling, mapping and sequencing

    a. DNA preparation
    b. DNA cloning
    c. Physical, restriction fragment, and genetic mapping
    d. Chemical, physical and enzymatic sequencing

3. Data storage and analysis

    a. Immediate data entry with uniform notation
    b. Efficient searching with cross-referencing and access to other data banks
    c. Rapid data distribution
    d. Parallel or concurrent processing
    e. New algorithms for analyzing and interpreting DNA and protein sequences

4. Detection and analysis of DNA, RNA and protein at very low levels

    a. Single molecule analytical methods
    b. Methods for detecting large numbers of DNA fragments simultaneously (multiplexing)

We estimate that the cost of the development of all of these technologies will be about $500 million dollars. The total cost will be near $1 billion and completion of the project will take many years. However, each advance in technology will produce immediate benefits to medicine, agriculture and industry.

Strategy. A substantial effort directed at technology, mapping and pilot-project sequencing can begin immediately. The committee recognizes that implementation of this initiative by DOE has already begun, and it praises the speed and thrust of the effort. $11.5 million has been requested for fiscal year 1988; an amount double this would be more appropriate. Funds spent early in this project will save money later, because each advance in technology will make all the following steps more efficient and less costly. Support of $40 million dollars the first year (fiscal year 1989) and increasing linearly to $200 million dollars by the fifth year (fiscal year 1993) could be used very effectively. We envision three types of grants — to individual investigators, to centers with 3 to 10 senior investigators and to a few large centers that will include mapping, sequencing and interpreting the human genome. In addition to the principal investigators, each project will involve junior scientists and engineers, and students. A total of 2500 professional people might be working on the initiative by 1993. The professional personnel will include molecular biologists, chemists, engineers, physicists, computer scientists and so forth.

Recommended funding levels are:

FISCAL YEAR	$ MILLION	TOTAL
1988	20	20
1989	40	60
1990	80	140
1991	120	260
1992	160	420
1993	200	620
1994	200	820
1995	200	1,020

Reasonable goals to attain by the end of seven years of support at the level requested (by the end of 1995 with $1 billion spent) are:

1) The United States should have the capacity to sequence ten million bases per day.

2) The complete map of each chromosome and an essentially complete sequence of at least one human chromosome should be finished.

Attainment of these goals will prove that the U.S. has the capabilities to continue the process to obtain all the benefits promised. We assume that equivalent progress will have been made in computer algorithms to analyze the sequences, and to characterize medically important genes.