Human Genome Institute: A position paper

We propose that an Institute be established at the University of California, Santa Cruz, whose mission is to map and to sequence the entire human genome. This paper describes some preliminary thoughts regarding the feasibility of the project and how such an Institute might be formed. We hope that we can persuade you that such a proposal is worth serious consideration and that you will be willing to help us further such an examination.

The genetic information that guides development and differentiation resides in the genome, a set of chromosomes composed of DNA molecules. These DNA molecules are sequences of nucleotides that are a blueprint, a recipe, to guide and specify the formation of the organism. Were it possible to "read" the information contained in the nuclectide sequences, we would have access to all the information passed from parent to offspring.

The human genome consists of 24 different DNA molecules (chromosomes) containing a total sequence of approximately 3,000,000,000 nucleotides. The technology now exists to read and decode this sequence. We believe that the time is ripe to begin the systematic process of determining the nucleotide sequence of the entire human genome. This is a very large and formidable undertaking that eventually will be accomplished. We believe a start should be made now.

We see this as a noble and inspiring enterprise. In some respects, like the journeys to the moon, it is simply a "tour de force"; it does not necessarily follow that knowledge of the nucleotide sequence of the human genome will provide deep insights into the physical nature of man. Nevertheless, we are confident that this project will provide an integrating focus for all efforts to use DNA cloning techniques in the study of human genetics. The ordered library of cloned DNA that must be produced to allow the genome to be sequenced will itself be of great value to all human genetics researchers. The project will also provide an impetus for improvements in these techniques, techniques that already have revolutionized the nature of biological research in all areas, from biochemistry to evolution, and promise to have wide-ranging applications in agriculture and medicine.

The scale of this project is indeed awesome. The largest genome so far completely sequenced, the Epstein-Barr virus, is 1.72×10^5 nucleotides in length (Baer et al., 1984) ; the human genome is twenty thousand times larger. However, the technology to carry out this project is already at hand and much of it is routine, and so can be scaled up using bio-engineering procedures. Ultimately we envision that the actual sequence-gathering steps will be automated and computerized.

There are three main components to this project. First, if the 24 different chromosomes (X, Y, and 22 different autosomes) were separated from one another in such a way that each is known, the project would be converted

into 24 smaller projects. Second, the individual chromosomes, each containing on average 150,000,000 nucleotides (150,000kb) must be cut into "clonable" fragments (20-40kb in length), and in a way that allows knowledge of the order of the fragments to be reconstructed. Thirdly, the ordered fragments must be cut into smaller fragments (ca. 0.5kb) to be sequenced. We will next consider the feasibility of each of these elements of the project with regard to existing technologies.

CHROMOSOME ISOLATION. Separation of small chromosomes has already been accomplished (by David Schwartz, Columbia Univ. and by a group at the Livermore Laboratory). Schwartz predicts that the technology for the isolation of large human chromosomes should be in hand in the reasonably near future. In addition, various genes on many of the chromosomes have been cloned and can be used to identify isolated chromosomes. The project could be started with only one isolated chromosome while efforts continue to isolate additional ones. If this does not prove a feasible approach, the project could in any event start at random points within the genome and sequenced segments later assigned to specific chromosomes from already mapped and cloned genes. Nevertheless, this is a significant element of the project and must be considered further.

THE ORDERED LIBRARY. The genome must be broken into relatively small fragments (20-40kb) for cloning. Because the human genome contains substantial amounts of repetitive DNA, construction of chromosomal libraries in a small cosmid vector such as pJB8 (Ish-Horowitcz, 1980) is probably advisable. The existing human genome libraries are in lambda phage vectors and so contain small inserts that are susceptible to recombination events when amplified and maintained in E. coli.

A major element of the overall project will be to determine the relative order of these randomly cloned fragments within the genome. Several methods to do this are already in use:

1) Random Overlap Method. The chromosome is chopped into random fragments and the neighboring fragments identified from overlaps. This method is being used by John Sulston (MRC) to create an ordered library of the genome of the nematode C. elegans (8×10^4kb). We understand that he has tested about 4000 fragments and already has 25% overlaps.

2) Walking. This is the most commonly used method. One starts with a given fragment and uses it as a hybridization probe to isolate overlapping fragments and these to isolate further overlapping fragments and so on. Thus one can start from one or a number of random points on a chromosome and systematically isolate overlapping fragments and so progress in both directions from these points. Using present technology, a researcher can probably walk 100 kb per month. Thus a human chromosome could be walked by one researcher in 1,500 months, or by 10 researchers in about ten years. However, a new method, called Jumping, should increase the speed of this process by an order of magnitude.

3) Jumping. This method has recently been developed by Hans Lehrach (EMBO Lab., Heidelberg). the chromosome is cut into pieces of up to 300 kb and these are spliced into a lambda EMBL vector genome. The middle segment of the splice is then removed and the lambda with the two ends (ca. 5kb) of the 100kb insert is religated and cloned. These are then used to probe for other overlapping clones. Since a chromosome will produce on average only about 2,000 such clones, sequencing these by either of the above two methods should be relatively straightforward. Each of these clones can then be used to find and sequence the internal fragments initially removed. This procedure should markedly accelerate the process of fragment ordering that will eventually result in a human genome library of cloned and ordered DNA fragments. Lehrach suggests that the speed of jumping should be ~10^6 nucleotides/month/walk, and that one researcher could conduct ~20 simultaneous walks, or ~2×10^7 nucleotides/month/person. Thus it would take ~ 12 person-years to walk the human genome.

CONCLUSION: Using existing technology, it is possible that within a few years, the human genome could be reduced to an ordered set of cloned fragments, fragments that are readily sequencable. This ordered library of the human genome would be in itself a major achievement, and could represent the primary scientific goal of the first ten years. Investigators throughout the world could then, for example, apply to the Institute to have specific fragments of interest targeted as high priority for sequencing. All information obtained would be made available via computer networks to any interested researcher. Specific clones listed in the library would also be distributed.

NUCLEOTIDE SEQUENCING. With existing techniques, a technician might be able to sequence as much as 10kb per week. If methods were improved to about 50kb/wk, 50 technicians could sequence an entire human chromosome in one year. Sequencing techniques are improving rapidly; a promising new method involving the use of fluorescent labels and scanning techniques is now in development (Lloyd M. Smith and Jane Sanders, Caltech Biology Report, 1984). Another approach (Beck and Pohl, 1984) uses direct transfer of DNA fragments onto an immobilizing matrix during electrophoresis, allowing 500-1000 nucleotides to be read from a single gel. Automated film scanners are already being developed by Genentech and Bio-rad. It seems highly likely that the simple rapid application of new techniques developed elsewhere will greatly speed up this process. If one considers the rate of development of DNA sequencing technology over the past decade or so, an exponential curve is obtained (fig. 1). If extrapolated, it suggests that a rate of 50kb per person per week would not be unreasonable to expect by 1988.

The above discussion, albeit superficial, shows that we have at hand the procedures to sequence the human genome. Even with techniques already in existence or near at hand, a reasonably sized institute with 50 technicians could approach completion of the project in 10-20 years. However, technological progress in these areas has been very rapid and is likely to continue. Application of new techniques will ensure success in a much shorter time if a start is made now.

THE INSTITUTE. We envision an institute located on the Santa Cruz campus with its own laboratory building and associated guest residence.

There would be a permanent staff of 50-75 scientists and technicians. 75% effort would go toward the human genome sequencing project, 10% toward development of improved techniques, 10% to facilitate application of findings to basic biological research and to medicine, and 5% to education; there would be continuous "state of the art" cloning and sequencing courses available and internships in the Institute for course graduates.

The Institute would be the repository for information obtained elsewhere and would be accessed by computer links to other nucleic acid databases and to individual research groups. Conferences and workshops would be encouraged; visitors would be housed in the guest residence.

Although the University of California, Santa Cruz would administer the activities of the Institute, it would be governed by a Board of Directors made up of distinguished scientists in the field.

We suggest that the cost of this project would be $25,000,000 for buildings and $5,000,000 per year in operating expenses.

Some of the questions we believe important to address in the first workshop are,

1. What are the major scientific benefits to be gained by such a project?

2. What medical benefits would derive from the project?

3. Is the project feasible?

4. What are the major obstacles to carrying it out?

5. Is a cottage industry approach a superior alternative?

6. Where should we go from here?

This proposal was prepared by Bob Edgar, Harry Noller, and Bob Ludwig.