LA-UR-91-1372

LA-UR--91-1372

DE91 011418

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

TITLE THE HUMAN GENOME PROJECT

AUTHOR(S) GEORGE I. BELL

MASTER

SUBMITTED TO TALK TO BE GIVEN AT "SUPERCOMPUTING USA / PACIFIC 91"
SANTA CLARA, CA, JUNE 19-21, 1991

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes.

The Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy

Los Alamos National Laboratory
Los Alamos, New Mexico 87545

FORM NO. 836 R4
ST. NO. 2629 5/81

THE HUMAN GENOME PROJECT

George I. Bell
Los Alamos National Laboratory
Los Alamos, NM 87545

1. ABSTRACT

The Human Genome Project will obtain high-resolution genetic and physical maps of each human chromosome and, somewhat later, of the complete nucleotide sequence of the deoxyribonucleic acid (DNA) in a human cell. The talk will begin with an extended introduction to explain the Project to non-biologists and to show that map construction and sequence determination require extensive computation in order to determine the correct order of the mapped entities and to provide estimates of uncertainty. Computational analysis of the sequence data will become an increasingly important part of the project, and some computational challenges are described.

2. Introduction

All of the information that is required to specify a human being is encoded in the DNA of a human cell. That DNA may be regarded as a set of 46 linear molecules, corresponding to the 23 pairs of chromosomes, one from each parent, where each molecule is a chain of nucleotides or bases that are of four different kinds, denoted A, C, G, and T. Thus, as a first approximation, DNA may be considered a long linear message written in a four-letter alphabet — long in that there are about six billion bases in the 46 chromosomes of a human cell. Genes, traditional units of inheritance, are located here and there along the chromosomes. There are of the order of 10^5 genes in the human genome, and a typical gene encodes the information to make a particular protein molecule. These concepts will be explained more fully in the talk and are discussed in Refs. (1)-(3), with relation to the Human Genome Project.

3. Some Computational Problems in Map and Sequence Assembly

An early objective of the Project is to make a high-resolution genetic linkage map of each human chromosome, which is to locate genes and other genetic markers at a uniformly high density along each chromosome. The separation between two markers, on the same chromosome, is estimated from the probability that they are inherited together, i.e., passed from parent to offspring without a chromosome "crossing over" between the markers. Thus, genetic maps are composed from statistical analysis of family studies on the inheritance of linked markers. The data for such analysis are not abundant, and the simultaneous determination of the relative order and spacing of a number of markers, together with uncertainty estimates, is a difficult computational task.

Another kind of map that the Human Genome Project aims to produce is an ordered-clone map of each chromosome. Its importance derives from the fact that the amount of DNA in a chromosome ($\sim 10^8$ bases) is too long to be directly cloned, sequenced, or otherwise manipulated. (In this context, cloning refers to the amplification of a DNA fragment by its incorporation into multiplying cells, namely bacteria or yeast.) Chromosomal DNA can, however, be chopped into overlapping clonable fragments, containing $10^4 - 10^6$ bases, by using restriction enzymes, but there is then the problem of how to order the fragments (clones) into an overall view of the chromosome. This is done by partially characterizing each clone. Two clones that have enough properties in common are then likely to contain a common segment of DNA and thus overlap on the map. The likelihood of overlap, given the data, is estimated from Bayes Theorem together with some model of the chromosome that can be used to estimate the probabilities of the data, given non-overlap or given overlap.

Inasmuch as one is attempting to order $10^3 - 10^4$ clones into a map, it is impossible to exhaustively enumerate all possible orders and assess their merits. Therefore, one first assembles the most confidently overlapping clones into islands ("contigs") and then uses some kind of greedy or branch and bound algorithm to add to and link the islands. Estimates of uncertainty in the resulting maps are themselves quite uncertain. There is, therefore, a tendency to regard all maps as tentative until they have been confirmed by ample redundant data.

Similar problems arise in generating the DNA sequence of a clone, namely, the clone DNA is too long to directly sequence, so that it is cut up into smaller fragments, of length ~ 500 bases, which can be sequenced.

The problem then arises of reassembling the overlapping sequence fragments into an overall consensus sequence of the clone. In this case, the sequence fragments can be pretty well characterized by their sequences, but problems arise from non-random errors in sequencing and the presence of many repetitive sequences in human DNA.

Good estimates of uncertainty in the consensus sequence are not available. This is a serious problem because the determination of such sequences is a major part - and probably the most expensive part - of the Genome Project. In order to minimize cost, one would like to minimize the sequencing redundancy required to produce a consensus sequence with specified limits of uncertainty.

As the maps and sequence data are generated, they must be deposited into databases that are readily accessible to a large variety of queries. At present, such databases are relational, but research on object databases for this use is underway. It appears that the databases will be distributed, with numerous satellite copies of major components. Parallel machines may be attractive hosts for such databases.

IV. Problems in Sequence Analysis

As the Genome Project generates new sequence information, we would like to be able to read the message thereby provided. How can this be done?

At present, the most powerful way to obtain information about a new sequence is by comparison with a database of all known sequences, such as GenBank. In such comparisons, one considers all possible alignments of the new query sequence (of length q bases) with the database (of length n), seeking regions of similarity and imposing penalties for mismatches, insertions, and deletions. There are standard dynamic programming algorithms for performing this search in a time $t = cqn$, where c is a constant depending on the computer systems, code, etc. Unfortunately, they are too slow for routine use, even on the fastest supercomputer. For example, GenBank has $n \simeq 5 \times 10^7$, and a particular code for a single processor CRAY XMP had $c \simeq 10^{-6}$ sec. Thus, to compute a query sequence of $q = 10^4$ with GenBank would take $\sim 5 \times 10^7$ sec $\simeq$ 140 hours. Faster results are obtain using massively parallel computers since evidently the database can be partitioned among all of the processors. For example, on the 64000 processor CM2, c was around 10^{-8} sec. However still greater speed is needed since the database is growing rapidly and queries are numerous. Two approaches are providing the increase in speed.

In the first (lexical) approach, one initially compares the sequences, not allowing for insertions or deletions, searching for regions that have a large number of words (short subsequences) in common. This comparison can be done very rapidly, for example, by constructing an index of words in the database that is then compared to words in the query. Once regions having a high density of common words have been identified, the dynamic programming algorithm can be used, if desired, to improve the local alignment. Codes, implementing such approaches, permit ready comparisons of query sequences with GenBank on scientific workstations or any more powerful computer[4].

The second approach is to construct special purpose VLSI chips that will carry out the dynamic programming algorithms. Experience with these to date is more limited than with lexical methods, but several are being (or have been) developed and are expected to be quite powerful and useful.

Even when a new sequence does not show any significant similarity to known sequences, we would like to be able to read its message and at the very least to identify sites of possible functional significance. Molecular biologists know a lot about the non-random use of words in genetic sequences and the association of patterns of word use with functional sites. For example, it is known that there is a non-random use of three-letter words and pairs thereof in DNA sequences that code for proteins (genes). Moreover, there are preferred sequences near the beginnings of genes that promote gene expression, etc. Unfortunately for human DNA, such features are generally soft. And when human DNA sequences are scanned using algorithms (including neural nets or rule based expert systems) to detect such features, the results may show good sensitivity (most known sites are correctly identified) but poor selectivity (i.e., the rate of false positive is unacceptably high). The problem is not due to any lack of computer power but more to a lack of critical data, including crystallographic data on the binding of enzymes to DNA together with data on the effects of replacing individual bases by alternatives within a functional site. In addition, it is quite possible that we do not really understand just what the proteins (often enzymes) that bind to DNA and RNA are really recognizing - that, for example, in addition to the local sequence, some structural motifs that may be influenced by torsional stress on the DNA double helix are important. If so, there are serious computational problems associated with trying to predict the detailed structural state of a DNA helix subject to torsional stress. In particular, under an unwinding stress, certain local sequences will be prone to adopt a left handed helical conformation as contrasted with the usual right handed helix. Other local palindromic sequences may extrude cruciform structures or the double helix may just come apart. The estimation of probabilities for these various competing conformations is a complex combinatorial problem[5].

Finally, I refer to the problems of predicting the structure of a protein from its amino acid sequence – a subject to be discussed by the next two speakers. This is a classical problem of great importance for the Human Genome Project. The sequence of bases in the DNA corresponding to a gene determines the sequence of amino acids in the protein do ed therefrom. Therefore, as one generates new DNA sequences and, hopefully, learns to identify those coding for a protein, there will be a need to better predict the likely structure and function of the proteins. In addition, as if the protein folding problem is not already hard enough, we would like to be able to predict the interaction of a folded protein with an appropriate (or arbitrary) DNA sequence.

V. References

(1) *Mapping and Sequencing the Human Genome*, Commission on Life Sciences, National Research Council. National Academic Press, Washington, D.C. (1988)

(2) *Mapping Our Genes, The Genome Projects: How Big, How Fast?*, Office of Technology, Congress of the United States (1988)

(3) *Understanding Our Genetic Inheritance*, The U. S. Human Genome Project: The First Five Years, FY 1991-1995, National Technological Information Service. U. S. Department of Commerce, Springfield, VA 22161 (1990)

(4) W. R. Pearson, *Methods in Enzymology* 183:63 (1990)

(5) C. Benham in *Mathematical Methods for DNA Sequences*, M. S. Waterman, ed. CRC Press, Boca Raton, FL, 1989