LA-UR -88-450

CONF-880567--2

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

LA-UR--88-450

DE88 006465

TITLE THE HUMAN GENOME -- COMPUTATIONAL CHALLENGES

AUTHOR(S) George I. Bell, T-DO

SUBMITTED TO Proceedings for the Third International
Conference on Supercomputing
Boston, Massachusetts
May 15-20, 1988

Los Alamos

Los Alamos National Laboratory
Los Alamos, New Mexico 87545

THE HUMAN GENOME --
COMPUTATIONAL CHALLENGES

George I. Bell

Theoretical Division, Los Alamos National Laboratory

Los Alamos, NM 87545

Abstract. The deoxyribonucleic acid (DNA) of a human cell contains all the information required for specifying that cell, or indeed the whole person, and constitutes the human genome. Programs are now underway to obtain genetic linkage maps and physical maps of human chromosomes containing the DNA, and large scale efforts will soon begin to provide detailed sequences. The challenges involved in assembling these data into a knowledge base are examined. Computations will play a key role in enabling the scientists to understand the information contained in sequence data. Pattern recognition and string matching algorithms will be of particular importance. Recent results in the use of adaptive networks for pattern detection will be presented.

I. INTRODUCTION AND SCOPE OF THE PROBLEM

DNA is a linear informational polymer. The monomers that are joined together to form the polymer are called nucleotides or bases and are of four different kinds that we may abbreviate as A, C, G or T. Thus the base at any position in the polymer can be specified by two bits. The DNA in a human cell contains about 3 × 10^9 bases from each parent organized into 23 chromosomes including one sex chromosome. It is possible that each chromosome contains a single DNA molecule or more precisely a single DNA double helix consisting of two complementary molecules. We may thus regard the human genome as consisting of 46 bit strings (23 from each parent) containing altogether about 1.2×10^{10} bits.

It is now becoming technically feasible to sequence an entire human genome, thereby obtaining the information required to specify a human being, albeit in highly encrypted form. A variety of workshops and studies, starting with the Santa Fe workshop in March 1986 and including recently reported studies by the Office of Technology Assessment and National Academy of Sciences have examined the motivation, technology requirements and costs of a project to sequence the human genome. A consensus appears to be emerging that the project should be undertaken with initial emphasis on (1) obtaining physical maps of chromosomes, (2) improving and automating the sequencing and mapping technology and (3) informatics - developing the computational tools for assembling, organizing and analyzing the sequence and map information. Before considering physical maps and informatics, let us consider what sorts of information may be found in the DNA.

At the outset we must recognize that no two individuals, save identical twins, will have the identical DNA. On the average, the DNA in unrelated individuals differs at around one base in 500 [1], or overall at about 10^7 bases out of 6×10^9. Thus while one may sequence an arbitrary reference human genome, the study of genetic heterogeneity at specific sites in the genome will probably be a more interesting enterprise to most investigators.

Information in the DNA carries many messages. (1) Some of the DNA codes for proteins, which are also linear polymers that may serve as catalysts (enzymes) signaling or structural materials in a cell. A segment of DNA coding for a protein is called a gene. It is estimated that a human cell has about 10^5 genes, that is it can make 10^5 different kinds of proteins. (2) Additional DNA sequences code for the regulation of gene expression. Sometimes they do this by recognizing and binding specific proteins that facilitate (or block) the reading of the gene in the protein making process. (3) Some of the DNA must be used in determining the structure of the chromosomes and guiding the precisely equal allocation of the chromosomes to two daughter cells in cell division. Little is known about this structural DNA but it does involve repetitive sequences. The DNA in a human cell is about 2m in length yet it fits in the cell nucleus having a diameter of about 5 µm. (4) Finally there is a lot of DNA having no known function and sometimes called "junk DNA." Some is probably parasitic. An example is the "ALU sequence" in humans; about 300 bases in length and present in 3 × 10^5 copies it thus takes up ~ 10^8 bases or more than a percent of the genome. Having no known function, it is probably largely parasitic.

DNA THAT CODES FOR PROTEINS

Much is known about how DNA codes for proteins. The monomers in proteins are called amino acids of which 20 different kinds are used in cells. A three letter code in the DNA specifies an amino acid in the protein with a complex series of intermediates, including ribonucleic acids (RNA) and proteins being involved in the process. The genetic code specifies which amino acid is encoded by every three base combination (codon) and is shown in Table I. The DNA code is first transcribed into a complementary RNA copy in which U replaces T so the table refers to the bases in RNA, A, C, G and U. Note that three of the codons specify "stop" or end of message. A long sequence of codons, uninterrupted by stop codons is a candidate for a proton coding sequence and is called an "open reading frame" (ORF). Translation is initiated at the codon ATG so location of this codon near the start of an ORF further strengthens the case for protein coding. Note that a given base sequence can be read in three different frames. Usually only one of these is used for encoding a protein, though some viruses use overlapping messages in different reading frames.

TABLE I
THE GENETIC CODE

First Position	Second Position*				Third Position
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	GluN	Arg	A
	Leu	Pro	GluN	Arg	G
A	Ileu	Thr	AspN	Ser	U
	Ileu	Thr	AspN	Ser	C
	Ileu	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

* Each entry represents an amino acid. For example "Phe" stands for phenylalanine.

In all multicellular creatures, called eukaryotes and including people, there is much additional complexity in the genes, namely there are long segments of DNA that are never expressed in the translated protein. These segments are called introns or intervening sequences as distinguished from exons or expressed sequences. Introns must be precisely excised from the RNA copy of the DNA, by a series of enzymatic steps, before the RNA copy is translated into the protein. The structure of a typical enkaryote gene is shown in Fig. 1. Some genes contain tens of exons. Although a qualitative description of the base sequence identifying an intron-exon boundary can be given in terms of a "consensus sequence," this is not good enough for a useful aigorithm.

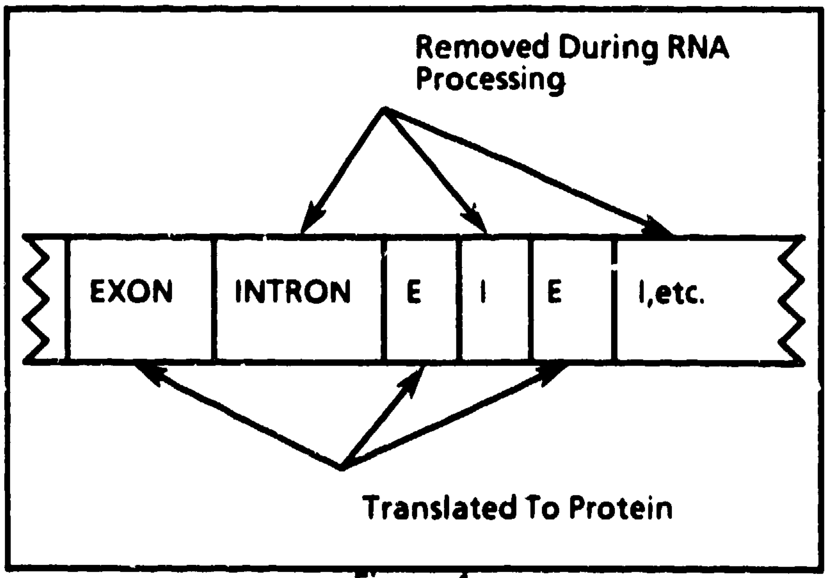

Figure 1
Structure of Typical Eukaryote Gene

II. THE DATABASE -- KNOWLEDGE BASE PROBLEM

The current database of DNA sequences is GenBank [2] assembled by the Los Alamos National Laboratory and the European Molecular Biology Laboratory and distributed by Intelligenetics [3]. It currently contains sequences totaling about 2 × 10^7 bases of which about 2 × 10^6 are from humans. Soon the rate of data entry will exceed 10^6 per month and it will likely be at least ten fold higher within five years. In addition to sequence data, GenBank contains annotation as to source, function, relation to other sequences, etc. Nearly every sequence in GenBank is related to other sequences; they may overlap, they may be alternative forms of the same gene or related genes. These relationships are vital parts of the database and are now entered by people, annotators, associated with the database.

The increasing rate of data entry is forcing GenBank to move toward a system in which most of the data will be submitted in electronically readable form, automatically checked for format and consistency and entered into the database. A relational format with the SYBASE management system is being used. Additionally, annotation of their data entries by the original investigators in electronically readable standard formats will be sought. This will require the provision of friendly if not seductive annotation software to the data originators.

As sequences are obtained in the human genome program, little or nothing will be known about the function of most. A program of automatic annotation should be developed to scan incoming sequences for ORFs, intron-exon boundaries, putative protein coding sequences, repetitive sequences having structural or parasitic properties and any other features. The findings should be entered in the database. Moreover, any significant similarities of the new sequence to other sequences should be noted. In this way one will begin to assemble a "knowledge base" for genetic sequences. Problems of pattern and similarity recognition are noted in the next section.

The sequence database, currently represented by GenBank, is only part of the genetic knowledge base. There are other related databases, including the Protein Information Resource [4] in which the sequences of all known proteins are assembled and also the Human Gene Mapping Library [5] in which the chromosomal locations of many human genes and other markers are given. These locations, which are based on genetic recombination frequencies, constitute a genetic linkage map of the human chromosome [1]. A listing of still other molecular biology databases [6] has been prepared.

In the next few years, another important type of map will be determined, namely physical maps of human chromosomes, and other genomes. In a physical map, the distance between markers is proportional to the number of bases between them and not necessarily to the recombination frequency. Several laboratories are undertaking physical maps of various genomes, including E. Coli, yeast and nematode and of human chromosomes. The method has been to cut the genome with restriction enzymes into a large number of overlapping and potentially clonable fragments. The fragments are then cloned and each clone is then partially characterized, for example by its binding to a number of random probes or by its pattern of lengths on further restriction enzyme digestion. Thus for each clone one determines a number of characteristics, and by assuming that clones overlap if they have a sufficient number of characteristics in common, one attempts to arrange the clones in linear order along the chromosome. This is a non-trivial computational exercise. For a chromosome of length 10^8 bases and a cosmid clone of length 4×10^4, one needs $> 2.5 \times 10^3$ clones to cover the chromosome. It is expected that $\sim 2 \times 10^4$ clones will probably be required for reasonably coverage. Repetitive DNA, nonclonable fragments and ambiguous signatures will complicate the assignment of fragment order. One will need to consider a range of possible orders and to design additional experiments to determine the correct alignment.

Thus the knowledge base for human genetics will soon include genetic linkage maps, with resolutions $\sim 10^6$ bases, physical maps, with resolutions $\sim 10^4$ bases and finally the sequence data. All of these data sets need to be accessible to users throughout the world, simultaneously and in a transparent manner. For example, a scientist studying the gene for cystic fibrosis can identify on the genetic map its approximate chromosomal location, can search the physical map for clones that may include the gene and examine the sequence for possible protein coding and regulatory sequences that may be hallmarks of the disease. The tasks for developers of this knowledge base are to assemble and organize the data from a multitude of sources so that it can be easily accessed and analyzed and so that the results of the analysis can in turn be incorporated into the knowledge base and retrieved by others. If due credit is given to those who conduct the analysis, citations in and growth of the knowledge base may become an accepted form of publication.

III. PATTERN DETECTION AND SEQUENCE SIMILARITY SEARCHING

Molecular biologists would love to be able to deduce the structure and function of a protein from its amino acid sequence or better yet from the DNA sequence of the encoding gene. As explained by R. Jernigan in an accompanying talk this is not possible at present. Nevertheless it is always of interest to compare a new sequence with known ones in order to detect possibly significant similarities which suggest function for the new sequence. In this way it was found that cancer causing genes often resemble naturally occurring cell growth factors or their receptors and many surprising relationships between genes have been found. Inasmuch as exons may be elementary units of evolution [7], a given exon may be shared by many genes. For example, a particular exon may code for a membrane binding portion of a protein and be shared by genes for a variety of proteins having in common only that they bind membranes.

The quantitative characterization of similarity, i.e., the determination of distances between sequences, is complex inasmuch as the significance of, say, a base change is context dependent and non-local. In a protein coding sequence, the effect of a base change on the amino acid sequence can be seen from the genetic code, provided that the reading frame is known. But the same change in an intron or regulatory sequence may be quite different. Insertion or deletion of one or more bases is also common and must be allowed for. Evidently insertion or deletion of a base in a coding sequence will shift the subsequent reading frame and may thus have a large effect on the protein sequence, whiie the same change in an intron may have no effect at all.

Given a set of weights for base changes, insertions, and deletions, an efficient algorithm was devised by Needleman and Wunsch [8], for finding the best alignment between two sequences and the distance between them when optimally aligned. For a recent review see [9]. Because of the length of the comparison strings, eg. 2 × 10^7 for GenBank and 3 × 10^9 for a human genome, considerable effort has gone into optimizing algorithms for current supercomputers, using "hash" tables and/or the vector processing capabilities of CRAY supercomputers. As usual, substantial increases in speed can be gained at the expense of completeness.

Evidently sequence comparison, as all text searching, is intrinsically parallel. Chips designed for text searching are now being considered for similarity comparisons [10]. They may be more useful in searching for precisely defined patterns rather than for general similarity comparisons.

Another general computational problem is that of detecting significant patterns in sequences of unknown function. As noted earlier, a particular problem is that of distinguishing protein coding from non-coding sequences and finding the reading frame. ORFs offer clues and because codons are used with unequal frequencies, protein coding sequences show distinctive non-random patterns [11,12] that can be detected in a long DNA sequence. Considerable success has been found [13] using an adaptive neural network approach that has successfully identified unsuspected protein coding regions and reliably gives the reading frame. The neural network is trained by exposure to a set of known coding and non-coding sequences and adjusts "synaptic weights" as it learns. While these networks may reach impressive levels of discrimination, they do not directly reveal what patterns they have learned to detect.

Perceptron learning algorithms have been used by Stormo [14] in an attempt to locate ribosomal binding sites in an RNA library, but were found to have only limited predictive capabilities. By combining perceptron algorithms with discriminant analysis, DeLisi et al. [15]

were able to predict exon/intron boundaries with around 85% reliability.

In general it has proved difficult to devise algorithms that reliably detect functionally significant patterns in DNA. Neural networks may help. But how do proteins reliably detect these sites while computers fail? One possibility is that deviations of the DNA from its ideal Watson-Crick double helix are recognized by DNA binding proteins [16]. Such deviations may involve local variations of the base sequence but also non-local effects such as the degree of twisting or supercoiling of the helix. Several models for predicting the helical structure of DNA have been proposed [17,18] but have shown only limited predictive capability. More exact calculational approaches are rendered difficult by the large sizes of the molecules (three bases in the double helix have around 600 electrons), by the aqueous environment in which the molecules are found (so that hydrophobic effects are important), and by the importance of electrostatic forces that are difficult to treat accurately [19].

Experimentalists can produce DNA molecules having virtually any desired sequence that can in turn be studied for function, such as the binding of regulatory proteins. There is thus great motivation not only for understanding these results but for developing the capability to predict DNA structure and interaction with proteins.

REFERENCES

[1] White, R. and Lalouel, J.-M., Chromosome Mapping with DNA Markers, Sci. Am 258 (2) 40, 1988.

[2] Burks, C., et al., The GenBank Nucleic Acid Database, Comp Applic. Biosc. 1, 225, 1985; Atencio, et al., The GenBank genetic sequence data bank, Nucleic Acids Res. (in press).

[3] IntelliGenetics, Inc., 700 East El Camino Real, Mountain View, CA 94040 -- BENTON @ BIONET -- 20.ARPA.

[4] George, D. G., Barker, W. C. and Hunt, L. T., The Protein Identification Resource (PIR), Nucl. Acids Res. 14, 11 (1986).

[5] Cohen, I. H., et al., The Human Gene Map, in Genetic Maps, 4th ed., O'Brien, S. J., ed. Cold Spring Harbor (1986).

[6] Lawton, J. R., Martinez, F. and Burks, C., Listing of Molecular Biology Databases, Nucl. Acids Res. (in press -- contact C. Burks, LANL).

[7] Gilbert, W., Nature 271, 501 1978.

[8] Needleman, S. B. and Wunsch, C. D., J. Mol. Biol. 48, 443 (1970).

[9] Goad, W. B., Computational Analysis of Genetic Sequences, Ann. Rev. Biophys. Chem. 15, 79, 1986.

[10] eg., Hollaar, L., "A Testbed for Information Retrieval Research: The Utah Retrieval System Architecture," Proc. SIGIR, 227, 1985; K. I. Yu, et al., "Pipeline for Speed: Fast Data Finder System," Quest, Winter 86/87, pp 5-19.

[11] Fickett, J. W., Nucl. Acids Res. 10, 5303, 1982.

[12] Ulanovsky, L. E. and Trifonov, E. N., Nature 326, 720, 1987.

[13] Barnes, C., Lapedes, A. and Sirotkin, K., manuscript in preparation.

[14] Stormo, G., Schneider, T. D., Gold., L. and Ehrenfeucht, A., Nucl. Acids Res. 10, 2997, 1982.

[15] DeLisi, C., Computers in Molecular Biology: Current Applications and Emerging Trends, Science (in press).

[16] Dickerson, R. E., Sci. Am. 294 (6), 94, 1983.

[17] Calladine, C. R., J. Mol. Biol., 161, 343, 1982.

[18] Tung, C.-S. and Harvey, S. C., Base Sequence, Local Helix Structure, and Macroscopic Curvature of A-DNA and B-DNA, J. Biol. Chem., 261, 3700, 1986.

[19] Soumpasis, M. D., Wiechen, J. and Jovin, T. M., J. Biomol. Struct. Dyn. 4, 535, 1987.