The Future of DNA Sequencing: Methods and Applications

Charles R. Cantor, Natalia Broude, Takeshi Sano, Marek Przetakiewicz and Cassandra L. Smith
Center for Advanced Biotechnology and Departments of Biomedical Engineering and Pharmacology, Boston University, Boston, MA 02215

THE HISTORY OF LARGE SCALE DNA SEQUENCING

The Human Genome Project entails the determination of the DNA sequence of a set of samples that will represent a single human genome. The identity of the particular base, A, T, C or G, at each position along 24 pieces of DNA encompassing one example of each of the human (H. sapiens) chromosomes, will be revealed. A total of about 3 billion bases will be identified. Along the way, the DNA sequence of several model genomes will be determined including a worm, C. elegans, a fly, D. melanogaster, a yeast, S. cerevisiae, and a bacterium, E. coli. All of these model genomes together amount to less than 20% of one human genome. In addition, depending on the efficiency and cost of the methods finally employed for most of the sequencing, we will also see revealed some or all of the DNA sequence of the common laboratory mouse, M. musculus. It is clear that, with the technology currently in hand, it will be possible to obtain the set of fragments needed for this DNA sequencing effort (this process is called physical mapping) and then read the sequence of each of these fragments for a total cost approximating the original estimate of $3 billion.

The Human Genome Project is the first example of a large scale organized DNA sequencing effort. Previous DNA sequencing projects largely focused on individual genes and rarely produced more than 100,000 bases of continuous DNA sequence. The major purpose of the human genome project is to find the estimated 80,000 to 100,000 different genes encoded for by this DNA and make them conveniently available for further study. The cost of finding genes one at a time by conventional human molecular genetic methods is prohibitive. It is estimated that the cost of finding just one human disease gene, that for cystic fibrosis, was far more than $100 million, and there are at least 5000 genes known to be involved in human inherited diseases.

The human genome project aims at finding all the genes in a fairly systematic manner, but this does not mean that we will know the identity of these genes as soon as they are found. Current technology allows us to look at unknown DNA sequence and estimate, fairly accurately, which segments are actually translated into proteins. However, very little hints are provided about the function of these proteins or their cellular locations and levels of expression. Currently about 200 million base pairs of DNA sequence are accessible in public databases like GenBank. Ordinarily, what is done is to compare any new
sequence against all known sequences and hope that a statistically significant match is found with a gene whose function is already understood. This happens in about 40% of all cases, and the newly-found genes are assumed to have a related or similar function to its pre-existing counterpart. The remaining newly-found genes begin, at least, as enigmas.

A new approach to gene finding has almost made the original concept of the genome project obsolete. Instead of looking directly at genomic DNA, the sequence information is obtained from DNA copies of expressed cellular messenger RNAs (mRNAs). These copies are called complementary DNAs (cDNAs), and a collection of the cDNAs from particular sample is called a cDNA library. The advantage of examining cDNAs is that one can focus only on genes that are actually expressed in the particular sample of interest. The disadvantage of cDNAs is that, unlike the genome, in which most genes are present at the same concentration (one copy per chromosome), mRNAs differ in abundance over at least 4 orders of magnitude. Thus a finite sized sample of cDNAs will contain large numbers of duplicates for commonly expressed genes and a relatively poor representation of most rarely expressed genes. There are ways to even out the distribution of cDNAs [1], but even if this is not done, it turns out that cDNA sequencing is a far more efficient strategy for discovering genes than direct examination of genomic DNA [2]. Thus it is likely that most human genes will be found as cDNAs over the next few years. This still entails a considerable amount of DNA sequencing; one can estimate that at least a million cDNAs will be sequenced to find most of the human genes.

Even with most genes identified, the full genomic human DNA sequence remains an attractive and highly desirable target. Information about cellular locations and levels of gene expression lie, usually, not in the genes themselves but in flanking DNA sequences. Sometimes this information can lie very far away from the actual coding region of the gene. (One case is known where sequences important in regulating expression of the protein hemoglobin lie 20,000 base pairs away from the gene itself.) In addition, the actual encoding of genes in DNA sequence is usually discontinuous. Blocks of expressed sequences called exons are spaced by often much larger blocks of sequences (introns) that are removed during the process of maturation of an mRNA. However, some of the signals for proper processing and removal of intron sequences lie within the introns themselves. Errors in these signals lead to improper processing, and this is a fairly common mechanism for inherited human disease. As a result, for full comprehension, the genomic DNA sequence must be available for study.

**Why Will We Need Large Scale DNA Sequencing After the Genome Project?**

As far as studies of human biology are concerned, the Human Genome Project is a beginning, not a conclusion. Human functions that are absolutely essential are likely to be encoded by genes that are virtually identical in all members of the species. Presumably these key functions are so demanding that any alteration in the genes involved makes an individual unviable. Thus we