Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 4338–4344, May 1993

Review

The human genome project

Maynard V. Olson

Department of Molecular Biotechnology, University of Washington, Seattle, WA 98195

ABSTRACT **The Human Genome Project in the United States is now well underway. Its programmatic direction was largely set by a National Research Council report issued in 1988. The broad framework supplied by this report has survived almost unchanged despite an upheaval in the technology of genome analysis. This upheaval has primarily affected physical and genetic mapping, the two dominant activities in the present phase of the project. Advances in mapping techniques have allowed good progress toward the specific goals of the project and are also providing strong corollary benefits throughout biomedical research. Actual DNA sequencing of the genomes of the human and model organisms is still at an early stage. There has been little progress in the intrinsic efficiency of DNA-sequence determination. However, refinements in experimental protocols, instrumentation, and project management have made it practical to acquire sequence data on an enlarged scale. It is also increasingly apparent that DNA-sequence data provide a potent means of relating knowledge gained from the study of model organisms to human biology. There is as yet little indication that the infusion of technology from outside biology into the Human Genome Project has been effectively stimulated. Opportunities in this area remain large, posing substantial technical and policy challenges.**

In the United States, the Human Genome Project first took clear form in February 1988, with the release of the National Research Council (NRC) report *Mapping and Sequencing the Human Genome* (1). To a degree remarkable in Federal science policy, this report has had a clear effect on subsequent programmatic activity. With a budget in the current fiscal year of $170 million, jointly administered by the National Institutes of Health (NIH) and the Department of Energy (DOE), a program is underway that conforms closely to the recommendations of the NRC committee. After a 5-year reality check, it is of both scientific and policy interest to examine how the committee's view toward this project has fared in the field.

Background

The human genome is the genetic material in human egg and sperm cells (i.e., germ cells), which contain 3×10^9 base pairs (bp) of DNA. Given the four-letter alphabet of DNA—customarily symbolized with the letters G, A, T, and C—the sequence of 3×10^9 bp corresponds to 750 megabytes of information. If the sequence of the human genome could be determined, it would be possible to store and manipulate it on a desktop computer. However, even the dream of acquiring DNA sequence on this scale is of recent origin. Dramatic progress was made during the 1950s and 1960s in understanding the mechanisms by which genetic information specifies biological structure and function. However, during this era, the information itself remained nearly inaccessible.

A landmark event in DNA analysis came in 1970 with the discovery of site-specific restriction enzymes (refs. 2 and 3; ref. 4, p. 64). These remarkable enzymes have the ability to scan any source of DNA for every occurrence of a particular string of bases—for example, the enzyme *Eco*RI recognizes the string GAATTC. Restriction enzymes cleave both strands of the double helix at their recognition sites. Since the cleavage events are directed by the DNA sequence, they always occur at the same positions in different samples of genomic DNA extracted from any genetically homogeneous source (e.g., different tissues of the same individual human or different individuals sampled from an inbred strain of mice).

Restriction enzymes provide a means to develop precise physical maps of DNA simply by determining the coordinates in base pairs of the sites at which particular enzymes cleave (ref. 4, pp. 66–67; ref. 5). Like topographic maps, physical maps of DNA derive their utility through annotation: mapped landmarks provide reference points relative to which functional DNA sequences such as genes can be localized. Restriction enzymes also facilitate a key step in the cut-and-splice procedures by which recombinant-DNA molecules (i.e., DNA clones) are constructed (ref. 4, pp. 73–74 and pp. 99–124).

The importance of recombinant-DNA technology is often attributed primarily to its synthetic dimension. For example, the ability to design and construct a DNA molecule that programs a bacterium to synthesize a mammalian protein provides a route to large amounts of the pure protein. The ability to alter the structure of the protein through site-directed mutagenesis lends genuine novelty to the resultant biosynthetic opportunities. However, the importance of recombinant-DNA technology in making the Human Genome Project feasible stems from its analytical dimensions. Cloning provides a means to purify individual recombinant-DNA molecules from complex mixtures and then to prepare biochemically useful amounts of the molecules by culturing the microbial strains into which they have been introduced.

A less obvious consequence of the discovery of restriction enzymes was the development of the first practical method of genetic mapping in humans (ref. 4, pp. 519–522; ref. 6). Most human cells contain two copies of each DNA sequence, one of maternal and the other of paternal origin. When a new germ cell is produced, it contains only one copy of the genome, a copy that is a unique mosaic of the two genomes from which it was derived. Genetic mapping involves measuring, through actual inheritance studies in families, the probability that two closely spaced segments of the genome will stay together during germ-cell formation. The mapping requires an ability to distinguish between the two copies of the genome present in the somatic cells from which the germ cells are derived. Subtle differences in the base sequence of different instances of the human genome sometimes alter restriction sites and, hence, restriction-fragment sizes. These alterations are detectable even in complex genomes by a method known as gel-transfer hybridization, which was developed in 1975 (ref. 4, pp. 127–130; ref. 7). In 1987, the first global human genetic map, based on "restriction-fragment-length polymorphisms," was published (8).

As to the actual determination of DNA sequence, reasonably efficient methods first appeared in 1977 (ref. 4, pp. 67–69; refs. 9 and 10). A technique known as

Abbreviations: DOE, Department of Energy; FISH, fluorescence *in situ* hybridization; NIH, National Institutes of Health; NRC, National Research Council; STS, sequence-tagged site; YAC, yeast artificial chromosome.

chain-termination sequencing came to dominate standard practice: it is based on enzymatic DNA synthesis, carried out *in vitro* in the presence of artificial chain-terminating variants of the normal DNA-precursor molecules. By the early 1980s individual sequences exceeding 10^5 bp had been determined (11); however, a more common scale of analysis was 10^3–10^4 bp. Most physical mapping was carried out on a similar scale.

Given the gap between the ability to determine 10^5 bp of DNA sequence in a state-of-the-art laboratory and the 10^9-bp size of the human genome, the NRC committee confronted an enormous problem of scale. Partly because of the obvious need for improved technology and also because of a desire to maximize synergy between genome analysis and studies of biological function, the committee recommended against early emphasis on large-scale sequencing of human DNA. Instead, it advocated comprehensive physical and genetic mapping of the human genome, extensive mapping and sequencing of the smaller genomes of several model organisms, and a systematic effort to develop improved sequencing technology.

Principal Aims of the Human Genome Project

More important than the specific mapping and sequencing objectives of the Human Genome Project are three broader aims that are implicit in these goals:

(*i*) To improve the research infrastructure of human genetics.

(*ii*) To help establish DNA sequence as the primary interface between knowledge of human biology and knowledge of the biology of model organisms.

(*iii*) To launch an open-ended effort to improve the analytical biochemistry of DNA.

For the purposes of this review, progress in the Human Genome Project will be examined relative to these three broad aims.

The Research Infrastructure of Human Genetics

In the context of the Human Genome Project, research infrastructure refers to the biological, informational, and methodological tools with which genetics research is carried out. Intensive genetic analysis of any species is heavily dependent on infrastructure. Particularly important are genetic-linkage maps, physical maps of DNA, and characterized DNA clones. The latter are useful as reagents that can be used to assay for particular short segments of the genome by DNA·DNA hybridization (ref. 12, pp. 188–191) and as the starting points for sequence analysis or functional studies.

Human genetics is uniquely dependent on strong research infrastructure. While model organisms have been extensively bred for the specific purpose of facilitating genetic analysis (13), human genetics is limited to the examination of individuals, families, and populations as they are found in contemporary society. Hence, the NRC committee set ambitious goals for the construction of detailed physical and genetic maps of the human genome, as well as organized collections of cloned human DNA. By design, the goals were too ambitious for the technology of 1988. In retrospect, they were so ambitious that they probably would have overwhelmed the basic methodologies on which the NRC report was based. Fortunately, technical advances since 1988 have exceeded all reasonable expectations.

Much of this progress was made possible by the development of the polymerase chain reaction (PCR). PCR, which is essentially a method of *in vitro* cloning, allows the amplification of specific DNA molecules *in vitro* through cycles of enzymatic DNA synthesis (ref. 4, pp. 79–85). PCR amplification is dependent on a pair of short, synthetic "primers" (i.e., single-stranded DNA molecules whose ends can be extended by DNA polymerase under the direction of template molecules). The test sample provides the template molecules, and the primers direct the amplification to a particular segment of the template DNA, typically a region only a few hundred base pairs in length. Starting with a minute sample of total human DNA, it is possible to amplify any such region 1 billionfold while leaving the rest of the genome at its original concentration.

Widespread application of the PCR depends on an efficient, automated method for the chemical synthesis of the PCR primers. An approach to DNA synthesis based on phosphoramidite chemistry, which became routine in the early 1980s, meets this need (ref. 4, pp. 69–70; refs. 14–16). The first paper on the PCR appeared in 1985 (17) but received little notice; for example, despite its present prominence in genome analysis, it is not mentioned in the NRC report. The explosive growth of PCR applications began with the publication of an important refinement of the PCR protocol in 1989—the use of a thermostable DNA polymerase (18). This refinement allowed the cycles of DNA synthesis, which are analogous to cellular generations, to be driven by simple thermal cycling with no new addition of reagents at each cycle.

By the end of 1989, it was already apparent that the PCR provided a practical means of abstracting large-scale physical maps away from the particular methods used to construct them. This capability required a new choice of landmarks called sequence-tagged sites (STSs; ref. 19). An STS is simply a short, unique sequence of DNA that can be amplified via the PCR. STSs are ideal landmarks during map construction because of the ease with which they can be detected by PCR assays. Equally important is their role in map representation and map use.

Complex physical maps based on restriction sites are of little value as experimental tools unless they are supported by a collection of clones that can be used to detect particular segments of the mapped DNA via DNA·DNA hybridization assays. Comprehensive maps of the human genome would have to be supported by tens of thousands of clones, each of which would have to be maintained as a separate microbial strain. In contrast, STSs can be described in an electronic data base in a form that makes them experimentally accessible in any laboratory. The most critical aspect of an STS description is the DNA sequence of the two primers. Laboratory implementation of an STS simply requires that the two primers be synthesized and the appropriate temperature-cycling regime be carried out.

Most large-scale physical maps are constructed through the process of "contig building." A contig is an organized set of DNA clones that collectively provide redundant cloned coverage of a region that is too long to clone in one piece (ref. 4, pp. 587–588; refs. 20–22). Typically, the clones have random end points, and the contig is described by specifying the amount of overlap between each clone and its nearest neighbors. A procedure referred to as STS-content mapping provides a convenient method of establishing these overlaps (ref. 4, pp. 610–612; ref. 23). In a step that precedes contig building, the STSs are tested to confirm that they occur in a single copy in the genome; then, if two clones share even a single STS, they can be reliably assumed to overlap.

Although the PCR has had a profound effect on physical mapping, other new developments have also improved the prospects for the construction of large-scale physical maps. One such development has been the introduction of the yeast artificial-chromosome (YAC) cloning system, first described in 1987 (ref. 4, pp. 590–592; ref. 24). YACs allow large segments of DNA to be cloned as linear, artificial chromosomes into the yeast host *Saccharomyces cerevisiae*. Even some of the earliest YAC clones were 10 times the size of the largest clones that had been constructed previously. Furthermore, the YAC system appears capable of cloning a higher proportion of the genomic DNA of many organisms

than could be recovered using earlier systems. This point has been most clearly documented during the physical mapping of the genome of the nematode worm *Caenorhabditis elegans* (25).

By 1989, YAC technology had evolved to the point where specific segments of the human genome could be recovered efficiently in YAC clones (26). Soon thereafter, multi-megabase-pair contigs began to appear (23, 27–29), and, in the fall of 1992, complete YAC-based physical maps of human chromosome 21 (30) and the human Y chromosome (31) were published. In these projects, contig construction was largely by STS-content mapping. There is little doubt that the same technology employed on chromosomes Y and 21, as well as on a large segment of the X chromosome (29), has sufficient power to produce highly connected physical maps of the entire human genome.

Another important advance in physical mapping has been the development of fluorescence *in situ* hybridization (FISH) into a routine procedure. This technique employs DNA probes that can detect segments of the human genome by DNA·DNA hybridization on samples of lysed metaphase cells prepared under conditions that preserve the morphology of the condensed human chromosomes. Attachment of fluorescent molecules to the probe DNA allows visualization in the light microscope of the position on a chromosome to which the probe binds. The technique is a refinement of previous *in situ* hybridization methods that depended on radiolabeling of the probes and autoradiographic detection (32). The increases in convenience, reliability, and resolution that have accompanied nonisotopic detection have transformed the role of *in situ* hybridization in physical mapping. The first nonisotopic visualization of single-copy sequences in human chromosomes by *in situ* hybridization was published in 1985 (33). Fluorescence detection of single-copy sequences was introduced in 1987 (34), after which applications expanded rapidly (35–37).

FISH contributes to two aspects of long-range physical mapping. First, it allows individual clones to be mapped at a coarse level long before contig building is complete, thereby providing reagents of immediate use in the analysis of targeted regions. Second, contig maps have discontinuities whenever a site in the genome is missing from the available clone collections. FISH provides a way to order and orient contigs along a chromosome even when occasional discontinuities exist. Early efforts to construct physical maps of human chromosomes, which depended on cosmid clones that are propagated in *Escherichia coli*, yielded relatively small contigs separated by discontinuities (38, 39). YAC-based methods have led to greatly improved continuity, but the need remains for supplementary methods to define the order and orientation of disconnected contigs along chromosomes.

Radiation-hybrid mapping, which involves fragmentation of chromosomes in cultured cells with high doses of x-rays followed by incorporation of the fragments into stable cell lines, provides still another solution to this problem (ref. 4, pp. 608–609; ref. 40). Current protocols for radiation-hybrid mapping are notable for their abandonment of the traditional goal of isolating a single short segment of the human genome in each rodent cell line. Nearly all of the radiation-hybrid lines produced by these protocols contain many unrelated segments of the human genome. Proximity of two STSs, or other markers, is inferred by statistical analysis of the pattern in which they occur in a large collection of cell lines. Closely spaced markers have a higher probability of occurring together in the same cell line than do pairs of markers that are on different chromosomes or are far apart on the same chromosome.

While the PCR, together with such new techniques as YAC cloning, FISH, and radiation-hybrid mapping, has led to a surge of success in physical mapping, PCR-based methods have also transformed genetic mapping. In particular, the PCR has allowed development of a new class of genetic markers that have a particularly high probability of existing in alternate forms in different instances of the human genome.

These markers are based on short, repetitive DNA sequences that are widely distributed in the human genome. A particularly common motif is . . . $(CA)_n$ . . . . At sites where this motif occurs, n, the number of repetitions of the dinucleotide CA, is highly variable from one instance of the human genome to the next (41, 42). Different values of n lead to PCR-amplification products of different lengths when the entire . . . $(CA)_n$ . . . tract is amplified by using primers that flank the repeat; these differences are readily detected by gel electrophoresis. An attractive feature of PCR-detectable genetic markers is that they are simply a special type of STS. As such, they can be readily included as landmarks in physical maps, as well as genetic maps, thereby providing a simple method of interrelating these two types of maps. Many PCR-detectable genetic markers have been integrated into preexisting maps of the human, greatly improving these maps (43). Still more recently, a human genetic map that is completely based on PCR-detectable markers has been constructed (44). Markers of the same type have also transformed genetic mapping in the mouse, whose genome is the same as that of the human (45).

A key test of the effectiveness of the infrastructure-building features of the Human Genome Project is the extent to which its components are being used even before genome-wide physical maps are available. The most critical test involves projects directed at the "positional cloning" of genes associated with heritable diseases. Positional cloning is a strategy that was developed during the 1980s to allow determination of the biochemical basis of the many heritable diseases whose analysis has resisted the more traditional approach of direct biochemical analysis of diseased tissue (46). In general, the biochemical analysis of diseased tissue is rarely effective unless the genetic defect alters a protein whose metabolic role in normal tissue is already understood. Few of the heritable diseases that cause mental retardation, psychosis, congenital malformation, malignant tumors, and other similarly complex effects meet this criterion.

The first step in positional cloning is to localize the "disease" gene by carrying out genetic mapping studies on families with multiple affected members. Studies of the coinheritance of the disease with genetically mapped DNA markers allow determination of the position of the gene in the genome. Actual biochemical identification of the gene still remains a formidable task since the resolution of genetic maps in the human is rarely better than 1 megabase pair (Mbp). Physical mapping and functional studies on the cloned DNA are required to find the gene within the candidate region.

Better physical mapping methods, particularly the combination of YAC cloning and FISH analysis, have improved the prospects for positional cloning. An exemplary baseline case, published just before either of these techniques became widely available, is cystic fibrosis. Final success in the positional cloning of the cystic fibrosis gene required heroic physical mapping efforts that never achieved any semblance of continuous cloned coverage of the candidate region (47). Piecemeal cloning and mapping proved adequate only because the gene was large and in a gene-poor region of the genome.

Subsequent successes with a series of disease genes reveal the influence of improved techniques. Examples in which YACs, FISH, or both figured prominently include the following: fragile-X syndrome (48–50), the most common heritable form of mental retardation; familial adenomatous polyposis (51, 52), a heritable form of colorectal cancer; myotonic dystrophy (53), an adult-onset disease that affects muscle function; Kallmann syndrome (54, 55), a defect in neuronal development; Lowe syndrome (56), a developmental defect affecting the lens, brain, and kidney; and Menkes disease

(57–59), a neurological disease that is lethal in early childhood.

The genes for many other heritable diseases are now under analysis by similar techniques. Particularly impressive is progress on the genetic mapping of diseases such as familial breast and ovarian cancer (60, 61) and early-onset familial Alzheimer disease (62). The genetic analysis of these diseases is complicated by a set of factors that will be encountered increasingly often as positional cloning is applied to complex, adult-onset genetic disorders: suitable families are rare, small, and incomplete (i.e., few grandparents, parents, or siblings of the affected individuals are available); even family members that remain disease-free throughout a normal life span cannot be reliably categorized as unaffected since they may have died from other causes before disease developed; the disease is common enough in the general population that cases with genetic and nongenetic causes occur frequently in the same family. Highly informative genetic markers, such as the PCR-detectable CA-repeat polymorphisms, have helped address these problems since they maximize the likelihood that the segment of the chromosome that bears the disease-causing mutation can be tracked reliably from one generation to the next even when there are many family members that are missing or must be excluded from the study because of their uncertain disease status.

In addition to improved genetic mapping, successful completion of these projects may require further advances in physical mapping and sequencing. Because of the difficulty of the genetic analysis, it is unlikely that disease genes such as the recently described one for early-onset familial Alzheimer disease on chromosome 14 (62) will be localized even to within 1 Mbp by genetic mapping. Thus, its isolation will place great demands on physical mapping resources and techniques for locating genes within cloned DNA. The case of Huntington disease, for which ample family resources are available, is instructive: the gene was genetically mapped to a position near the end of the short arm of chromosome 4 in 1983 (63) but has not yet been identified in cloned DNA. Its position, even now, is known only to within 2.5 Mbp (64).

Still another class of disease genes whose analysis has benefited from new infrastructure are genes whose disruption in somatic cells causes cancer. Particularly in leukemias and lymphomas, a common mechanism by which disease-causing mutations arise is translocation, a process of chromosome breakage and rejoining. The combination of YACs and FISH analysis has simplified the mapping of the chromosomal breakpoints and allowed the isolation of genes whose disruption is the initiating event in several forms of neoplasia (65, 66).

DNA Sequence as an Interface Between Knowledge of Human Biology and Knowledge of the Biology of Model Organisms

Central to the NRC committee's recommendations, which emphasized the importance of sequencing the genomes of model organisms, was the belief that DNA sequence offers a potent means of interrelating diverse aspects of biological knowledge. Events during the past 5 years have strongly reinforced this concept.

Particularly remarkable is the ability of DNA-sequence data to call attention to similarities between biological phenomena that are superficially unrelated. A typical example involves the successful transfer of information from the study of yeast mating to diverse areas of human biology. Yeast cells have two mating types, commonly referred to as **a** and α. In the yeast life cycle, **a** and α cells are the rough counterparts of mammalian germ cells. The yeast counterpart of fertilization involves the fusion of **a** and α cells, a process that is partly mediated by two peptide hormones, **a** factor and α factor. These hormones are named after the cell type that secretes them. They trigger a series of changes in the opposite cell type that prepare the cell for fusion.

The mechanisms through which **a** factor and α factor are synthesized and secreted have been studied in detail by genetic techniques that are particularly well developed in yeast (67). A peculiar feature of **a**-factor secretion is its independence of the pathway through which yeast proteins are normally secreted. A particular gene, *STE6*, encodes a protein that allows **a** factor to leave the cell while bypassing the normal secretory pathway. Sequence analysis of *STE6* revealed unmistakable similarity to the human gene *mdr1* (68, 69). This gene has attracted interest because of its involvement in multiple-drug resistance, a phenomenon in which malignant cells become simultaneously resistant to several of the most commonly used chemotherapeutic agents (70). Once the relatedness of *STE6* and *mdr1* had been established by sequence comparison, it was quickly shown by gene-transfer experiments that the mouse version of the *mdr1* gene will actually substitute in yeast for the function of *STE6*, correcting the inability of yeast cells with mutations in the *STE6* gene to secrete **a** factor (71). The availability of the yeast system opens up a powerful new front for the study of this poorly understood transport mechanism.

Studies of α-factor biosynthesis have proven equally productive in providing insights into human metabolism. While the secretion of α factor follows the normal pathway, its biosynthesis has a feature that is unusual in yeast but relatively common in human cells: it is produced by proteolytic processing of a precursor peptide at a Lys-Arg linkage. Genetic studies revealed that the gene *KEX2* encodes the protease that carries out this processing step. Comparison of the sequence of *KEX2* with all other known DNA sequences revealed strong similarity to a human gene of previously unknown function, c-*fur* (72). Subsequent analysis showed that c-*fur* is a member of a family of human genes that encode proteases that process precursors to many important proteins and peptides including insulin, nerve growth factor, bone morphogenetic protein, and a major component of the AIDS virus (73). There is a long history of direct, biochemical efforts to identify these proteases because of their potential interest as pharmacological targets. These efforts led to the description of a whole series of proteases that are capable of cleaving Lys-Arg linkages under particular *in vitro* conditions but that serve other functions *in vivo*.

These two examples illustrate the strength of the concept, which is fundamental to the Human Genome Project, that DNA sequence provides the key to efficient knowledge transfer between model organisms and human biology. At present, this process requires considerable serendipity, only because available DNA-sequence data on the genomes of both the human and the major model organisms are fragmentary. There has been enough progress on the sequencing of the genomes of *E. coli* (74), *S. cerevisiae* (75), and the nematode worm *C. elegans* (76) to indicate the value of systematic genomic sequencing. However, the real work of determining complete sequences for the genomes of these model organisms still lies ahead.

In humans, the main new source of systematic data has come from the sequencing of cDNAs, cloned DNA copies of the messenger RNA (mRNA) molecules that actually direct protein synthesis (ref. 4, pp. 102–104; ref. 77). This method is a cost-effective way of discovering new human genes because only a small fraction of genomic DNA directly codes for proteins. However, cDNA sequencing is unlikely to replace genomic sequencing as the definitive method of characterizing the complete set of human genes for several reasons: genes contain critical DNA sequences that regulate their expression but are not included in the mRNA; there are common instances both in which one gene produces multiple, substantially different mRNA molecules and in which multiple genes produce nearly identical mRNA molecules, situations that are difficult to sort out

without detailed knowledge of the structures of the corresponding genes and gene families; the cost advantages of cDNA sequencing erode when the goal is accurate, full-length sequences rather than one-pass, partial sequences; no adequate solution has been found to the problem that the mRNA products of different genes are present at widely different concentrations, which vary dramatically in different tissues, different developmental stages, and different metabolic states.

Open-Ended Improvements in the Analytical Biochemistry of DNA

The promise of DNA-sequence comparison as a fundamental tool in biological research emphasizes the need for progressively better methods of DNA analysis, particularly DNA sequencing. A critical feature of this challenge is its open-ended nature. DNA sequencing is a technology, like digital computing, for which there is no obvious point at which further improvements would saturate potential applications. A basic misimpression about the Human Genome Project is that once its narrow goals are met, demands for large-scale DNA sequencing will taper off. DNA sequence data are basically a source of hypotheses, the rigorous testing of which typically requires the acquisition of still more DNA sequence. The determination of a "reference" human sequence will provide a strong incentive to trace genes through evolution with finer grain than the *E. coli*/yeast/worm/fly/mouse/human comparisons on which the NRC committee recommended early emphasis. Finally, the study of individual variation, which plays a central role both in biology and in medicine, poses unbounded demands for DNA-sequence data.

Juxtaposed to this open-ended need for improvements in the efficiency of DNA sequencing is the reality that there has been no obvious increase in the basic efficiency of DNA sequencing during the past decade. The protocols have become more robust, and the skill level required for success has been lowered. Fluorescence-based methods with real-time detection of the products of DNA-sequencing reactions during electrophoresis have eased laboratory management of large projects and decreased the subjectivity of data interpretation (ref. 4, pp. 595–598; ref. 78). Hence, the practicality of large projects is greater now than it was a decade ago. However, it is not apparent that there has been any change in either the efficiency or the accuracy with which an expert DNA sequencer can gather data.

The NRC committee recognized this problem but was overoptimistic about its resolution (ref. 1, p. 2):

> "The technical problems associated with mapping and sequencing the human and other genomes are sufficiently great that a scientifically sound program requires a diversified, sustained effort to improve our ability to analyze complex DNA molecules . . . . Prospects are . . . good that the required advanced DNA technologies would emerge from a focused effort that emphasizes pilot projects and technological development."

NIH and DOE have both made vigorous efforts to steer a significant portion of the project in the recommended directions. However, there is little indication that decisive research momentum has developed in the technology of DNA sequencing. It can be argued that the problem is predominantly cultural rather than technical, relating to the different value systems and research emphases of molecular genetics, on the one hand, and analytical chemistry, applied physics, and engineering, on the other (79).

An illustration of the magnitude of the technical challenge is provided by the gap between the theoretical and actual output of the current generation of DNA-sequencing instruments. Standard commercial instruments now have the capacity to produce ≈30 kilobase pairs (kbp) of raw sequence data per day (78). Allowing for the desirability of determining the sequence of the two redundant strands of DNA independently and for some oversampling of data for each strand, a ratio of raw sequence data to finished data of 5:1 should be achievable. Hence, a single instrument should be capable of producing 6 kbp of finished sequence per day, or ≈2 Mbp per year. In reality, no genome center has yet produced even 1 Mbp of contiguous, finished sequence per year even though such centers typically have many sequencing instruments.

This paradox reflects the present impossibility of integrating all the steps in DNA sequencing into a continuous process that fully utilizes even the capabilities of current sequencing instruments. Although this experience is universal among DNA sequencing laboratories, there is little consensus about which steps in the process are rate limiting, much less what should be done to improve them.

What is clear is that there is a dramatic gap between the advanced biological technologies of molecular genetics and the primitive nonbiological technologies. The latter include the physical manipulation of samples, methods of chemical and physical analysis, process design, quality control, and information handling. These areas are all critical to efforts to scale up bench-top molecular genetics, and most biologists are poorly trained to make the needed innovations. The Human Genome Project has stimulated increased interactions between biologists and scientists and technologists who have the necessary expertise to solve these problems. However, the difficulties of translating these beginnings into major improvements in DNA analysis continue to pose substantial policy challenges.

Conclusions

For an effort that is only in its third year of substantial funding, the Human Genome Project in the United States is making good progress toward its central goals. The policy on which it was based has proven farsighted even in the face of rapid technological change. In the mapping goals of the project, which have dominated the first years, the experimental methods that are leading to success have diverged widely from those extant when the NRC report was issued. Nonetheless, the report's conceptual framework has survived with little alteration.

Examples abound of biological advances that have benefited directly from the early activities of the Human Genome Project. Precise tracking of the cause-and-effect relationships between activities funded through the Human Genome Project in the United States and specific biological advances is neither possible nor desirable. Human genome analysis is a loosely coordinated international endeavor to which funding agencies and scientists in many countries have already made important contributions. Vigorous research activity funded through other Federal programs, private agencies, and industry has also had a major impact. Nonetheless, NIH and DOE programmatic efforts, particularly through their productive investment in YACs, FISH, PCR-detectable DNA polymorphisms, and radiation-hybrid mapping, have clearly achieved good progress toward the mapping goals of the NRC report and also contributed directly to the success of many other research projects in the biomedical sciences.

Like other human endeavors, the Human Genome Project has succeeded best when it has aligned itself with broader trends. Examples include its increasing reliance on PCR, yeast genetics, and fluorescence microscopy. It has succeeded least when it has tried to establish new trends such as the importation of high technology from other areas into biology. This tension is healthy and will undoubtedly remain as the project focuses increased attention on its flagship goal of determining the sequence of the 99% of the human genome about which we still know almost nothing.

Note Added in Proof. The gene that is mutated

in Huntington disease has now been identified (80).

1. National Research Council (1988) *Mapping and Sequencing the Human Genome* (Natl. Acad. Press, Washington, DC).
2. Smith, H. O. & Wilcox, K. W. (1970) *J. Mol. Biol.* **51,** 379–391.
3. Smith, H. O. (1979) *Science* **205,** 455–462.
4. Watson, J. D., Gilman, M., Witkowski, J. & Zoller, M. (1992) *Recombinant DNA* (Freeman, New York), 2nd Ed.
5. Nathans, D. (1979) *Science* **206,** 903–909.
6. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. (1980) *Am. J. Hum. Genet.* **32,** 314–331.
7. Southern, E. M. (1975) *J. Mol. Biol.* **98,** 503–517.
8. Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B. *et al.* (1987) *Cell* **51,** 319–337.
9. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
10. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 560–564.
11. Baer, R., Bankier, A. T., Biggin, M. D., Deininger, P. L., Farrell, P. J., Gibson, T. J., Hatfull, G., Hudson, G. S., Satchwell, S. C., Seguin, C., Tuffnell, P. S. & Barrell, B. G. (1984) *Nature (London)* **310,** 207–211.
12. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. & Watson, J. D. (1989) *Molecular Biology of the Cell* (Garland, New York), 2nd Ed.
13. Fink, G. R. (1988) *Genetics* **118,** 549–550.
14. Beaucage, S. L. & Caruthers, M. H. (1981) *Tetrahedron Lett.* **22,** 1859–1862.
15. Caruthers, M. H. (1985) *Science* **230,** 281–285.
16. Hunkapiller, M., Kent, S., Caruthers, M., Dreyer, W., Firca, J., Giffin, C., Horvath, S., Hunkapiller, T., Tempst, P. & Hood, L. (1984) *Nature (London)* **310,** 105–111.
17. Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A. & Arnheim, N. (1985) *Science* **230,** 1350–1354.
18. Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. & Erlich, H. A. (1988) *Science* **239,** 487–491.
19. Olson, M., Hood, L., Cantor, C. & Botstein, D. (1989) *Science* **245,** 1434–1435.
20. Coulson, A., Sulston, J., Brenner, S. & Karn, J. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 7821–7825.
21. Olson, M. V., Dutchik, J. E., Graham, M. Y., Brodeur, G. M., Helms, C., Frank, M., MacCollin, M., Scheinman, R. & Frank, T. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 7826–7830.
22. Kohara, Y., Akiyama, K. & Isono, K. (1987) *Cell* **50,** 495–508.
23. Green, E. D. & Olson, M. V. (1990) *Science* **250,** 94–98.
24. Burke, D. T., Carle, G. F. & Olson, M. V. (1987) *Science* **236,** 806–812.
25. Coulson, A., Kozono, Y., Lutterbach, B., Shownkeen, R., Sulston, J. & Waterston, R. (1991) *BioEssays* **13,** 413–417.
26. Brownstein, B. H., Silverman, G. A., Little, R. D., Burke, D. T., Korsmeyer, S. J., Schlessinger, D. & Olson, M. V. (1989) *Science* **244,** 1348–1351.
27. Anand, R., Ogilvie, D. J., Butler, R., Riley, J. H., Finniear, R. S., Powell, S. J., Smith, J. C. & Markham, A. F. (1991) *Genomics* **9,** 124–130.
28. Silverman, G. A., Jockel, J. I., Domer, P. H., Mohr, R. M., Taillon-Miller, P. & Korsmeyer, S. J. (1991) *Genomics* **9,** 219–228.
29. Little, R. D., Pilia, G., Johnson, S., D'Urso, M. & Schlessinger, D. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 177–181.
30. Chumakov, I., Rigault, P., Guillou, S., Ougen, P., Billaut, A. *et al.* (1992) *Nature (London)* **359,** 380–387.
31. Foote, S., Vollrath, D., Hilton, A. & Page, D. C. (1992) *Science* **258,** 60–66.
32. Gall, J. G. & Pardue, M. L. (1969) *Proc. Natl. Acad. Sci. USA* **63,** 378–383.
33. Landegent, J. E., Jansen in de Wal, N., van Ommen, G.-J. B., Baas, F., de Vijlder, J. J. M., van Duijn, P. & van der Ploeg, M. (1985) *Nature (London)* **317,** 175–177.
34. Landegent, J. E., Jansen in de Wal, N., Dirks, R. W., Baas, F. & van der Ploeg, M. (1987) *Hum. Genet.* **77,** 366–370.
35. Lawrence, J. B., Villnave, C. A. & Singer, R. H. (1988) *Cell* **52,** 51–61.
36. Trask, B., Pinkel, D. & van den Engh, G. (1989) *Genomics* **5,** 710–717.
37. Lichter, P., Tang, C.-J. C., Call, K., Hermanson, G., Evans, G. A., Housman, D. & Ward, D. (1990) *Science* **247,** 64–69.
38. Stallings, R. L., Torney, D. C., Hildebrand, C. E., Longmire, J. L., Deaven, L. L., Jett, J. H., Doggett, N. A. & Moyzis, R. K. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 6218–6222.
39. Tynan, K., Olsen, A., Trask, B., de Jong, P., Thompson, J., Zimmermann, W., Carrano, A. & Mohrenweiser, H. (1992) *Nucleic Acids Res.* **20,** 1629–1636.
40. Cox, D. R., Burmeister, M., Price, E. R., Kim, S. & Myers, R. M. (1990) *Science* **250,** 245–250.
41. Weber, J. L. & May, P. E. (1989) *Am. J. Hum. Genet.* **44,** 388–396.
42. Litt, M. & Luty, J. A. (1989) *Am. J. Hum. Genet.* **44,** 397–401.
43. NIH/CEPH Collaborative Mapping Group (1992) *Science* **258,** 67–86.
44. Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G. & Lathrop, M. (1992) *Nature (London)* **359,** 794–801.
45. Dietrich, W., Katz, H., Lincoln, S. E., Shin, H. S., Friedman, J., Dracopoli, N. C. & Lander, E. S. (1992) *Genetics* **131,** 423–447.
46. Collins, F. S. (1992) *Nature Genet.* **1,** 3–6.
47. Rommens, J. M., Iannuzzi, M. C., Kerem, B. S., Drumm, M. L., Melmer, G., Dean, M., Rozmahel, R., Cole, J. L., Kennedy, D., Hidaka, N., Zsiga, M., Buchwald, M., Riordan, J. R., Tsui, L. C. & Collins, F. S. (1989) *Science* **245,** 1059–1065.
48. Verkerk, A. J. M. H., Pieretti, M., Sutcliffe, J. S., Fu, Y. H., Kuhl, D. P. A. *et al.* (1991) *Cell* **65,** 905–914.
49. Kremer, E. J., Pritchard, M., Lynch, M., Yu, S., Holman, K., Baker, E., Warren, S. T., Schlessinger, D., Sutherland, G. R. & Richards, R. I. (1991) *Science* **252,** 1711–1714.
50. Oberle, I., Rousseau, F., Heitz, D., Kretz, C., Devys, D., Hanauer, A., Boue, J., Bertheas, M. F. & Mandel, J. L. (1991) *Science* **252,** 1097–1102.
51. Kinzler, K. W., Nilbert, M. C., Su, L. K., Vogelstein, B., Bryan, T. M. *et al.* (1991) *Science* **253,** 661–665.
52. Groden, J., Thliveris, A., Samowitz, W., Carlson, M., Gelbert, L. *et al.* (1991) *Cell* **66,** 589–600.
53. Fu, Y. H., Pizzuti, A., Fenwick, R. G., Jr., King, J., Rajnarayan, S., Dunne, P. W., Dubel, J., Nasser, G. A., Ashizawa, T., de Jong, P., Wieringa, B., Korneluk, R., Perryman, M. B., Epstein, H. F. & Caskey, C. T. (1992) *Science* **255,** 1256–1258.
54. Legouis, R., Hardelin, J. P., Levilliers, J., Claverie, J. M., Compain, S., Wunderle, V., Millasseau, P., Le Paslier, D., Cohen, D., Caterina, D., Bougueleret, L., Delemarre-Van de Waal, H., Lutfalla, G., Weissenbach, J. & Petit, C. (1991) *Cell* **67,** 423–435.
55. Franco, B., Guioli, S., Pragliola, A., Incerti, B., Bardoni, B., Tonlorenzi, R., Carrozzo, R., Maestrini, E., Pieretti, M., Taillon-Miller, P., Brown, C. J., Willard, H. F., Lawrence, C., Persico, M. G., Camerino, G. & Ballabio, A. (1991) *Nature (London)* **353,** 529–536.
56. Attree, O., Olvios, I. M., Okabe, I., Bailey, L., Nelson, D. L., Lewis, R. A., McInnes, R. R. & Nussbaum, R. L. (1992) *Nature (London)* **358,** 239–242.
57. Vulpe, C., Levinson, B., Whitney, S., Packman, S. & Gitschier, J. (1993) *Nature Genet.* **3,** 7–13.
58. Chelly, J., Tumer, Z., Tonnesen, T., Petterson, A., Ishikawa-Brush, Y., Tommerup, N., Horn, N. & Monaco, A. P. (1993) *Nature Genet.* **3,** 14–19.
59. Mercer, J. F. B., Livingston, J., Hall, B., Paynter, J. A., Begy, C., Chandrasekharappa, S., Lockhart, P., Grimes, A., Bhave, M., Siemieniak, D. & Glover, T. W. (1993) *Nature Genet.* **3,** 20–25.
60. Hall, J. M., Lee, M. K., Newman, B., Morrow, J. E., Anderson, L. A., Huey, B. & King, M.-C. (1990) *Science* **250,** 1684–1689.
61. Hall, J. M., Friedman, L., Guenther, C., Lee, M. K., Weber, J. L., Black, D. M. & King, M.-C. (1992) *Am. J. Hum. Genet.* **50,** 1235–1242.
62. Schellenberg, G. D., Bird, T. D., Wijsman, E. M., Orr, H. T., Anderson, L., Nemens, E., White, J. A., Bonnycastle, L., Weber, J. L., Elisa Alonso, M., Potter, H., Heston, L. L. & Martin, G. M. (1992) *Science* **258,** 668–671.
63. Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., Watkins, P. C., Ottina, K., Wallace, M. R., Sakaguchi, A. Y., Young, A. B., Shoulson, I., Bonilla, E. & Martin, J. B. (1983) *Nature (London)* **306,** 234–238.
64. Davies, K. (1992) *Nature (London)* **357,** page before 95.
65. Djabali, M., Selleri, L., Parry, P., Bower, M., Young, B. D. & Evans,

G. A. (1992) *Nature Genet.* **2,** 113–118.
66. Ziemin-van der Poel, S., McCabe, N. R., Gill, H. J., Espinosa, R., III, Patel, Y., Harden, A., Rubinelli, P., Smith, S. D., LeBeau, M. M., Rowley, J. D. & Diaz, M. O. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 10735–10739.
67. Botstein, D. & Fink, G. R. (1988) *Science* **240,** 1439–1443.
68. Kuchler, K., Sterne, R. E. & Thorner, J. (1989) *EMBO J.* **8,** 3973–3984.
69. McGrath, J. P. & Varshavsky, A. (1989) *Nature (London)* **340,** 400–404.
70. Gottesman, M. M. & Pastan, I. (1988) *J. Biol. Chem.* **263,** 12163–12166.
71. Raymond, M., Gros, P., Whiteway, M. & Thomas, D. Y. (1992) *Science* **256,** 232–234.
72. Fuller, R. S., Brake, A. J. & Thorner, J. (1989) *Science* **246,** 482–486.
73. Barr, P. J. (1991) *Cell* **66,** 1–3.
74. Daniels, D. L., Plunkett, G., III, Burland, V. & Blattner, F. R. (1992) *Science* **257,** 771–778.
75. Oliver, S. G., van der Aart, Q. J. M., Agostoni-Carbone, M. L., Aigle, M., Alberghina, L. *et al.* (1992) *Nature (London)* **357,** 38–46.
76. Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., Dear, S., Coulson, A., Craxton, M., Durbin, R., Berks, M., Metzstein, M., Hawkins, T., Ainscough, R. & Waterston, R. (1992) *Nature (London)* **356,** 37–41.
77. Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., Kerlavage, A. R., McCombie, W. R. & Venter, J. C. (1991) *Science* **252,** 1651–1656.
78. Hunkapiller, T., Kaiser, R. J., Koop, B. F. & Hood, L. (1991) *Science* **254,** 59–67.
79. Olson, M. V. (1991) *Anal. Chem.* **63,** 416A–420A.
80. The Huntington's Disease Collaborative Research Group (1993) *Cell* **72,** 971–983.