Review Article

The Code of Codes
I: The Science and Technology

Aflab J. Ahmed, Ph.D.
Chicago, Illinois

Abstract

Determination of the complete nucleotide sequence of the human genome is one of the most ambitious scientific projects ever undertaken. The expressed objective of this initiative is to not only decipher the linear order of the basic alphabet of the genome, but also to physically map genetic loci for various human diseases in order to eventually develop rational therapeutic regimes for them. Despite the erstwhile controversy about the relevance of sequencing the entire three billion nucleotides in human DNA, the human genome project is making progress apace. Thus, according to some estimates, it is ahead of the initial estimates of completing the entire sequence in 15 years. Several complementary strategies are being actively pursued to accomplish this task. Among the most promising are approaches utilizing expressed sequence tags (EST) and sequence-tagged sites (STS). Their respective merits and demerits notwithstanding, taken together, these strategies portend to yield a comprehensive map of the human genome and shed light on its dynamics, flux, and evolution.

Keywords: Human genome, linkage analysis, recombinant DNA, yeast artificial chromosomes, contigs, sequence-tagged sites, expressed sequence tags.

Genetics is the discipline that seeks to elucidate the processes by which various organisms pass on their traits of anatomy, physiology, and behavior to their progeny and of how each individual expresses those traits in its life cycle. Genetics, thus, is the central problem of biology. As the basic mechanisms by which transmission and expression of hereditary characteristics are unraveled, new insights are continuously being gained into other problems in biology; the study of immunology, endocrinology, neurobiology; the study of cancers and other diseases; and that of evolution.[1]

Central to genetics is the concept of the gene. Genes, however, act in concert with each other within their chromosomal context. In other words, the concept of the gene presupposes the map of the genes, their locations, and interrelationships and, of course, the sequence of their chemical constituents. Modern genetics began in 1900 with the rediscovery of Mendel's rule first stipulated 35 years earlier.[2] The word "gene" was first coined in 1909. In 1910, it was demonstrated for the first time that a particular gene

From the Department of Neurology
Northwestern University Medical School
Chicago, Illinois

Reprint Requests: Aflab J. Ahmed, Ph.D.
Department of Neurology
Tarry Building 13-715
Northwestern University Medical School
303 East Chicago Ave.
60611-3008

had a locus, i.e., it could be assigned to a specific chromosome. The first genetic map showing relative locations of six genes on one chromosome was published in 1913.[3]

In some 70-plus years since, the concept of gene has dramatically changed and has become more profound, and with it the idea of gene maps and sequences has been refined.[4]

The double helical structure of deoxyribonucleic acid (DNA) is an example of the law of parsimony in correspondence to structural detail to functional requirements. The two strands of DNA intertwine coaxially in a clockwise manner. Each strand is a string of four chemical moieties, called nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T). These nucleotides differ only in the shape of a flat structure, called a (nitrogenous) base, that protrudes from the side. The two strands are at the outside of the structure and are linked across the space between them by the bases that form pairs. The bases are important because their sequence along the DNA strand is the only variable part of its structure. The physical shape of the bases limits them to just two types of pairs as they connect the strands. They pair with perfect fidelity: A to T and C to G. The bases are spatially organized such that exactly 10 base pairs occur in the length of a full turn of the helix.[5,6]

Interestingly, for genetic instructions, no physical or chemical rules determine the sequence of bases; i.e., the order in which they are strung along the helix. Yet, at the same time, paradoxically, the constraints are almost exquisitely stringent: When the sequence on one strand is fixed, the base pairing solely determines the complementary sequence on the opposite strand. In DNA, hence, form and function coalesce. The structure of DNA has the mechanisms built in for passing on of its genetic potential and its expression. Constraints of the base pairing rule predicate that if the strands unwind, each can assemble on itself an exact copy of its complementary previous strand producing two identical helices during cell division. This is gene transmission, in principle. The freedom of the sequence of bases allows encoding with a four-letter alphabet (A, C, G, T) of the entire specification of the substance and regulation of the organism.

The human genome project (HGP) was implicit in the discovery of the structure of DNA,[5] quite like the oak in the acorn. Until the elucidation of the DNA structure in 1953, biologists could only do what Mendel had done before them; they could only infer the presence and interactions of genes by their visible signs by analyzing the traits made to manifest in organisms in breeding trials. Mendelian genetics refined this empirical approach to highly sophisticated lengths. It would only be a matter of time, however, that genes would be understood as a specific sequence of bases in DNA. Once the scientists understood that, they could begin to think of analyzing the growth and functioning of organisms from the inside out by identifying functional base sequences and finding out what they determine in living organisms, including human beings. In its totality, the human genome comprises all the genes found in the cells of human beings. As such, it is the master blueprint of life, the code of all codes, in that it contains enciphered in its detail, specific information about all the genes that human beings possess.

Sometime during the mid-1980s the idea materialized that the sequence of the entire human genome, all 3 billion base pairs, should be determined.[7,8] The proposal, which roiled the scientific waters, immediately fueled a controversy that in some form continues even today, albeit not with its erstwhile intensity or urgency. Some biologists worried that such a "big-science" project could divert the scarce resources away from smaller, innovative projects and academic investigators. Equally, there were concerns aired about what seemed to be a real possibility of "dirigisme," a centrally directed program. System analysts wondered how the copious flow of information could be collected, collated, recorded, and retrieved. Seasoned molecular biologists saw little point in sequencing the entire genome when more than 95% of it has no apparent function. Ethicists were concerned about the misuse and abuse of such detailed knowledge of man's genetic endowment. National and commercial rivalries emerged, characterized by polarized rhetoric as to "Who Owns the Human Genome?"[9] One school of thought espoused the notion that the copyright to a DNA sequence belongs to individuals or interests that carry out the sequencing and would be made widely available for a price. This standpoint is robustly disputed by others. Arguments about patentability of the sequence data continue to rage.[10]

For several years, HGP hung fire. Nonetheless, it was apparent that sequencing and mapping were already proceeding stridently in innumerable laboratories in many countries. The issue was not whether the human genome will be sequenced but what strategies should be devised, what priorities assigned, and how the different efforts be optimally coordinated and synchronized.[11,12] Many different organizations, led by the National Institutes of Health (NIH) and the Department of Energy (DOE) in the United States, in conjunction with international consortia such as the Human Genome Organization (HUGO), have in the interim reached consensus on how to advance this most ambitious project ever initiated in the history of biological sciences.

Scientifically, a major issue has been about the approach in sequencing the genome as to whether it would be more meaningful to start with a linkage map and then proceed to the physical map or whether to attempt both concurrently. A linkage map relates genes or cloned DNA sequences to one another along the various chromosomes. This type of mapping has been successfully used to identify genes for a number of diseases.[13,14] It involved analysis of co-inheritance of genetic loci in informative families and the statistical interpretation of whether loci co-segregate more often than would be expected by chance. Once two loci are estab-

lished to be genetically linked, their estimated distances from each other can be calculated from the frequency of recombination. Groups of different loci (defined by suitable DNA polymorphisms) can be studied simultaneously by multipoint linkage analysis. A map is, thus, built in which the distance between the markers is expressed in the unit of recombination, the centimorgan (cM). A map of 1 cM corresponds roughly to 1 megabase (Mb; 10^6 base pairs). However, recombination frequencies per megabase of DNA vary significantly both by gender and chromosomal region. There are higher frequencies of recombination in female meioses such that linkage maps in females have greater apparent distances. Proportionately, telomeric regions of chromosomes undergo more recombination events in males; whereas, centromeric regions of chromosomes undergo more recombination events in males. Consequently, the "rough" 10 cM linkage maps of the human genome already available have uneven and uncertain spacing and need to be substantiated by physical characterization. The optimal resolution of a linkage map is of the order of 1 cM with the caveat that there will often be larger gaps due to the absence of suitable polymorphic markers.

Physical maps of the genome are based on either cytogenetics or molecular technologies. Maps based on cytogenetics order loci with respect to the visible banding patterns along chromosomes using data from somatic cell hybrids or the more sensitive technique of *in situ* hybridization. Molecular maps characterize large tracts of DNA on the basis of landmarks established by restriction endonuclease cleavage. The resulting DNA fragments are cloned in vectors and different clones are related to one another by virtue of overlapping restriction sites or sequences to give a set of contiguous DNA pieces, or a contig map. The advent of vectors such as yeast artificial chromosomes (YACs) in which DNA fragments as large as 500 kilobases (kb) can be cloned, permits considerably larger contig maps to be constructed in specific chromosomal regions.[15]

Laboratories generating contig maps frequently sequence portions of the DNA in their overlapping clones to make certain, in the first instance, that the clones are, in fact, overlapping. Thus, some portions of the human genome have already been sequenced, albeit they are a mere minuscule fraction of the total genomic DNA, and are distributed at random among the chromosomes. One serious issue in the genome mapping pertains to relating different sequences to one another to construct a definitive overall map. A particular contig map depends on the physical existence of a collection of clones (a library) that may have to be transferred to another laboratory to establish or exclude a possible relationship to another contig map. If clone libraries are not available, or if they degenerate in storage, it becomes difficult to establish continuity in construction of the physical map.

This problem has been resolved by an ingenuous approach that makes use of a common language for landmarks in the human genome.[16] Any sequence of 200 to 500 bp of genomic DNA that can be amplified by the polymerase chain reaction (PCR) constitutes a landmark or sequence-tagged site (STS). Taken together with the data on the most appropriate primers for PCR amplification, these sequences are recorded in an STS data base. An investigator who has constructed a contig map could test the map with other characterized regions of the genome by searching for an established STS with primer data from the STS data bank. If the average spacing of the STSs were 100 kb, a fragment size suitable for YAC cloning, about 30,000 stics (genomic complexity/STS spacing = $3 \times 10^9/10^5 = 30{,}000$) would be necessary. It is estimated that approximately one-half of these already exist.

Ideally, a broad strategy of the HGP should make progress on two fronts. A short-term goal must be to improve the resolution of the linkage map to spacings of about 2 cM. This should provide a framework for physical mapping, the assembly of continuous regions of at least 2 Mb of DNA, each harboring some 20 or more STSs as reference landmarks. The longer-term goals to systematically sequence whole regions, and even chromosomes, ought to be postponed until later. As the technology of automatic sequencing improves to 1,000 kb per day *vis a vis* the current output of 2 kb per day, it would be a more practical prospect to sequence the entire human genome fairly quickly.

Nonetheless, there are alternative views that question the practicality of this approach. A sizeable number of investigators have proposed that the long-term objective of sequencing the entire genome should be postponed in favor or concentrating on transcribed genes, representing the biologically active and physiologically relevant 2 to 3% of the total DNA sequences. This approach selects clones from complementary DNA (cDNA) libraries, which, theoretically, encompass the exonic sequences of all the genes expressed in the tissue from which the library was constructed.[17] The attractive feature of this idea is that it foresees sequencing of an easily manageable stretch of 400 to 500 bp in each clone. This limited sequencing should provide enough working information about the genes to facilitate a search for homologies to known sequences recorded in various sequence data banks. If there should be no homology, the sequence would likely represent part of a novel transcribed gene, and sufficient data would be available from the partial sequence (termed as expressed sequenced tag, or EST) to allow others to recognize the putative gene in their work. Conceptually, this strategy is rather similar to the STS methodology with the exception that the landmarks are exclusively in the transcribed genes only and no attempt is made to link the DNA segments either physically or genetically.

One drawback of this approach is that the cDNA libraries do not necessarily contain representative clones from all transcribed genes in a particular tissue. Some genes may well be transcribed during successive, temporally regulated windows of stage-specific development such that the

messenger ribonucleic acid (mRNA) is not captured in this library. Likewise, many genes exist in multiple copies that potentially could lead to redundant sequencing of identical clones. In addition, families of similar genes might be difficult to distinguish with partial sequencing with the consequence that some members of the family that harbor sequence divergence might be missed completely. Further, cDNA sequencing yields no information at all about the crucial promoter and other regulatory sequences in the nontranscribed, yet indispensable, upstream regions of the gene(s) of interest. Nevertheless, initial experiments, using a human brain cDNA library, have demonstrated a very high proportion of clones for which partial sequencing indicates that they are transcribed from genes without homologies to any of the already known genes deposited in different data banks. These newly discovered genes, clearly, are of significant relevance to the biology of brain function, and would not have been recognized as such by random total sequencing of large stretches of genomic DNA.

Needless to say, the HGP is proceeding with complementary strategies.[18] Obviously, a detailed linkage map is critical to provide a framework within which, for instance, disease genes could be placed expediently and with a higher degree of confidence than is feasible today. Sequencing of cDNAs and EST landmarks have the capacity to describe transcribed genes of significance in the biology of different tissues. The long-term goal, however, remains to sequence the entire human genome, filling in the information for nontranscribed regions and ultimately to seek to answer the question as to whether the bulk of the human genome is, indeed, devoid of any significant function.[19]

Editor's note: The Code of Codes part II will appear in the April issue of JIMA.

References

1. Judson HF: The eighth day of creation. New York: Simon and Schuster, 1978.
2. Mendel G: Experiments in plant hybridization. translated in: C.A. Davern, ed. Genetics: A scientific American reader. San Francisco: WH Freeman and Co., 1981.
3. Peters JA, ed.: Classic papers in genetics. Englewood Cliffs, NJ: Prentice-Hall, 1964.
4. Watson JD: The human genome project: past, present and future. Science 1990:244-8.
5. a. Watson JD, Crick FHC: Molecular structure of nucleic acids. A structure of deoxyribose nucleic acid. Nature 1953;171:737.
   b. Watson JD, Crick FHC: Genetical implications of the structure of deoxyribonucleic acid. Nature 1953;171:964.
6. Watson JD: The double helix: A personal account of the discovery of the structure of DNA. New York: W. W. Norton and Co., 1980.
7. McCusick VA: Mapping and sequencing the human genome. New England J Med 1989;320:910.
8. Morton NE: Parameters of the human genome. Proc Natl Acad Sci 1991;88:7474.
9. a. Maddox J: Ownership and the human genome. Nature 1994;371:363.
   b. Dickson D: Consortium plans public map of genome. Nature 1994;371:551.
   c. Maddox J: Genes and patent law. Nature 1994;371:271.
11. Siniscalo M: On the strategies and priorities for sequencing the human genome: A personal view. Trends in Genetics 1987;3:182.
12. Canton CR: Orchestrating the human genome project. Science 1990;248:49.
13. Dulbecco RA: A turning point in cancer research: Sequencing the human genome. Science 1986;231:1055.
14. Friedmann T: The human genome project -- some implications of extensive reverse genetic medicine. Amer J Human Genetics 1990;46:407.
15. Green ED, Olson MV: Systematic screening of yeast artificial-chromosome libraries by use of polymerase chain reaction. Proc Natl Acad Scie 1990;87:1213.
16. Olson M, Hood L, Cantor C, Botstein D: A common language for physical mapping of the human genome. Science 1989;245:1434.
17. Adams MD, Kelley JM, Gocayne JD, et al.: Complementary DNA sequencing: Expressed sequenced tags and human genome project. Science 1991;252:1651.
18. Cook-Deegan R: The gene wars: Science, politics and the human genome. New York: W. W. Norton and Co., 1994.
19. Willis C: Exons, introns and talking genes. New York: Basic Books, 1991.