MOLECULAR MEDICINE

HUNTING FOR GENES IN COMPUTER DATA BASES

Mark S. Boguski, M.D., Ph.D.

HARDLY a month goes by without a report that some new gene for a human disease has been cloned. Recently, for example, the genes responsible for ataxia–telangiectasia and early-onset Alzheimer's disease were isolated. In the final analysis, such studies always come down to a DNA sequence and whatever biologic function (or functions) we can infer from it. In other words, we may know from genetic-linkage studies that a particular gene sequence (in mutated form) is responsible for a disease, but what clues can a DNA sequence provide about the precise pathophysiology of that disease? Overwhelmingly, the answer now lies in a specialty known as computational biology (sometimes called "bioinformatics"). The main procedure used in this field is comparative sequence analysis by data-base homology searching.

Homology, strictly speaking, means structural similarity due to descent from a common ancestor and is usually used in an anatomical context, as when one states, for instance, that the human arm is homologous to the whale's fin. In the context of molecular biology, two DNA (or protein) sequences are said to be homologous if they are sufficiently similar to suggest that they are derived from a common ancestral gene and thus may be similar in function. The alpha and beta subunits of hemoglobin and myoglobin are all homologous sequences, for example. When one isolates a new gene, the very first step in its analysis is a search for homologues that may shed light on the new sequence by virtue of inferred structural and biochemical similarities among the gene products. I will present some specific cases below, but first I will describe the infrastructure that is necessary to support this activity, which is carried out thousands of times every day by researchers throughout the world.

There are three main components of homology searching: public data bases of DNA and protein sequences, computer algorithms and programs for searching these data bases, and the Internet, which connects individual research laboratories to these centralized data repositories and software tools. GenBank is a comprehensive data base of DNA and protein sequences assembled and distributed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine. GenBank (along with its international collaborators, the Data Library of the European Bioinformatics Institute and the DNA Database of Japan) contains all publicly available information on gene and gene-product sequences and is updated daily. GenBank currently contains information on 492,483 DNA sequences comprising 354 million nucleotide bases and representing more than 11,000 biologic species. New human DNA sequences are added to the data base at a rate of 1500 per day.

A GenBank record consists of not only the DNA sequence itself but also a unique identifier (known as the accession number) and a biologic annotation of the sequence's features, such as exons and introns for genes and coding regions for messenger RNAs. The annotation includes literature citations, and through its Entrez system, NCBI provides direct connections to Medline records, a taxonomy data base, and (when available) the three-dimensional structures of proteins encoded by genes. All this information is available on CD-ROM and over the Internet and World Wide Web. Access is by electronic mail, file-transfer software ("anonymous ftp"), and browsing software such as Netscape Navigator and NCSA (National Center for Supercomputer Applications) Mosaic. One can search for a sequence by key words or author's name. But one of the most powerful methods is to conduct a homology search with the Basic Local Alignment Search Tool (BLAST).

BLAST is a sophisticated computer program capable of rapidly detecting even very distant sequence similarities in a statistically rigorous manner. Typically, one sends a "query sequence" (e.g., a new gene sequence) over the Internet to NCBI, where powerful computers compare this sequence with every sequence in GenBank and report the results, usually within minutes (Fig. 1). The results consist of a list of matching sequences, the statistical significance of the matches, and alignments between the query and matching data-base sequences. It is difficult to overestimate the importance of this technique; it has been used in the analysis of virtually every DNA or protein sequence ever determined. Homology searching is available at a number of sites throughout the world; NCBI alone processes approximately 10,000 search requests daily.

One of the most remarkable facts about homology searching is that it is capable of detecting relations among genes that span more than a billion years of molecular evolution. When a new human gene is cloned, a GenBank search may identify the corresponding gene from another species, and chances are good that biochemical information about the nonhuman gene product will provide insight into human disease. For example, a gene responsible for hereditary nonpolyposis colon cancer is homologous to a yeast gene encoding an enzyme that repairs mismatching (mutated) bases in DNA. This observation resulted from a data-base search and greatly advanced our understanding of the causes of cancer. There are many other genes for human diseases that have homologues in yeast (e.g., the genes for cystic fibrosis and neurofibromatosis type 1), with the known function of the yeast gene product providing a crucial insight into the disease. Thus, the real value of homology searching is that it makes synergistic connections between disparate research activities for which no common ground was even suspected.

A recent and dramatic example of this phenomenon is the successful hunt for the gene responsible for atax-

From the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, where reprint requests should be addressed to Dr. Boguski, who may also be reached by electronic mail (boguski@ncbi.nlm.nih.gov).

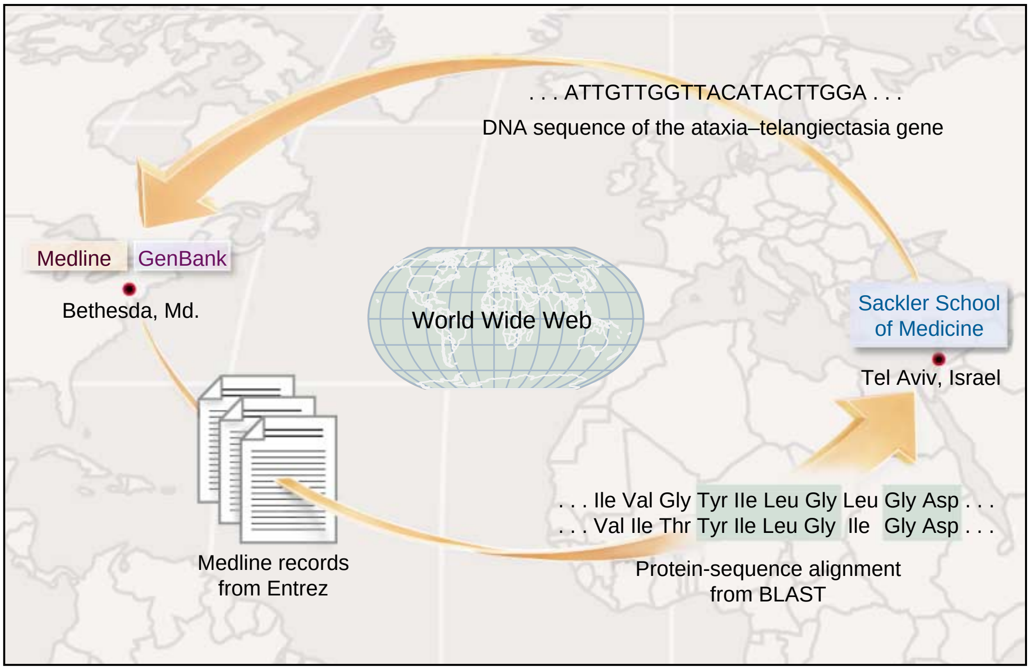
. . . ATTGTTGGTTACATACTTGGA . . .
DNA sequence of the ataxia–telangiectasia gene
Medline
GenBank
Bethesda, Md.
World Wide Web
Sackler School of Medicine
Tel Aviv, Israel
. . . Ile Val Gly Tyr Ile Leu Gly Leu Gly Asp . . .
. . . Val Ile Thr Tyr Ile Leu Gly Ile Gly Asp . . .
Medline records from Entrez
Protein-sequence alignment from BLAST

Figure 1. Data-Base Homology Searching on the Internet.

Data-base homology searching is carried out more than 10,000 times per day by researchers throughout the world to determine whether a DNA or protein sequence (the "query sequence") resembles any previously described genes or gene products. With the use of electronic mail, World Wide Web browsing software, or other types of client–server computer programs, the query sequence is sent over the Internet to server computers at the National Center for Biotechnology Information in Bethesda, Maryland. With the use of the BLAST program, the query sequence is compared with all the sequences in GenBank, and a report of matching sequences is returned to the sender. The entire process usually takes only a few minutes. In the example shown here, BLAST takes a nucleotide sequence and, using the genetic code, translates it into a predicted protein sequence before performing the data-base search. The report on the search aligns the query sequence with matching sequences, to maximize the number of matching letters (nucleotide bases or amino acid residues). Such alignments also come with P values for the likelihood that these matches could have occurred by chance. In the case of the ataxia–telangiectasia gene product, the protein matched several yeast proteins with P values in the range of 10^{-58} to 10^{-94}. The example shows an alignment between a region of the ataxia–telangiectasia gene product and another protein, with amino acid residues indicated by their three-letter codes. Amino acid residues that are identical in the two proteins are shown in green rectangles.

For a full interpretation of the results of the homology search, it is necessary to examine the literature concerning all biologic aspects of the matched sequences. This step is conveniently performed over the Internet with the Entrez system, which integrates GenBank sequences with Medline records, three-dimensional models of protein structures, and other information resources. Entrez uses Internet-based client–server programs and graphic interfaces to provide "point and click" integrated information retrieval on most popular computer systems. BLAST and Entrez are available free on the World Wide Web at the following address: http://www.ncbi.nlm.nih.gov/.

ia–telangiectasia. This disorder is characterized by progressive cerebellar ataxia in early childhood, immunodeficiency, and a predisposition to lymphoid tumors and other cancers. Although ataxia–telangiectasia is a rare autosomal recessive disorder, about 2.5 million people in the United States are heterozygous carriers of the gene for the disease, and these people have increased radiosensitivity and a dramatically increased risk of several cancers, particularly breast cancer in women. When the gene for ataxia–telangiectasia was finally cloned, after an 18-year effort, a homology search with BLAST revealed in minutes that the gene is biochemically related to a number of yeast enzymes (phosphatidylinositol 3-kinases) critical for cell growth and DNA repair. The search also showed that the gene product is related to other human proteins that are the intracellular targets for the immunosuppressive agents tacrolimus (FK 506) and sirolimus (formerly called rapamycin). Thus, in a single stroke, homology searching revealed biochemical explanations for the phenotypes observed in patients with ataxia–telangiectasia.

The comparison of human genes with those of model organisms such as bacteria, nematodes, and yeast thus provides important insights into the pathophysiology of human disease. For this reason, the Human Genome Project encompasses the sequencing of genomes for several such organisms. The genome of the bacterium *Haemophilus influenzae* has 1.8 million base pairs of DNA encoding 1743 genes, and its sequence was reported in July 1995. The genome of the yeast *Saccharomyces cerevisiae* is composed of 12.5 million base pairs of DNA encoding approximately 6000 genes, and the work on its sequence will be completed by January 1996. The genome of the nematode worm *Caenorhabditis elegans* contains 100 million base pairs of DNA encoding approximately 15,000 genes, and the work on its sequence is scheduled for completion by the end of 1998. The human genome consists of 3 billion base pairs of DNA encoding approximately 50,000 to 100,000 genes and will be completed sometime between 2001 and 2005. Consequently, GenBank will continue to grow exponentially, and gene hunting, with comparative analysis by data-base homology searching, will be a mainstay of biomedical research well into the 21st century. Sequence analysis will find new and expanding applications in molecular diagnosis, vaccine and drug development, and gene therapy.

Recommended Reading

Boguski MS. Bioinformatics. In: Dracopoli NC, Haines JL, Korf BR, et al., eds. Current protocols in human genetics. New York: John Wiley, 1994:11.1.1-11.3.44.

Collins FS. Positional cloning moves from the perditional to traditional. Nat Genet 1995;9:347-50.

Savitsky K, Bar-Shira A, Gilad S, et al. A single ataxia telangiectasia gene with a product similar to PI-3 kinase. Science 1995;268:1749-53.

Tugendreich S, Bassett DE Jr, McKusick VA, Boguski MS, Hieter P. Genes conserved in yeast and humans. Hum Mol Genet 1994;3:1509-17.