

Correct protein function depends on the three-dimensional (3D), or folded, structure the proteins assume in biological environments; thus, understanding protein structure will be essential in determining gene function. DNA sequences will be translated into amino acid sequences, and researchers will try to make inferences about functions either by comparing protein sequences with each other or by comparing their specific 3-D structures (Fig. 15).

Because the 3-D structure patterns (motifs) that protein molecules assume are much more evolutionarily conserved than amino acid sequences, this type of homology search could prove more fruitful. Particular motifs may serve similar functions in several different proteins, information that would be valuable in genome analyses.

Currently, however, only a few protein motifs can be recognized at the sequence level. Continued development of analytic capabilities to facilitate grouping protein sequences into motif families will make homology searches more successful.

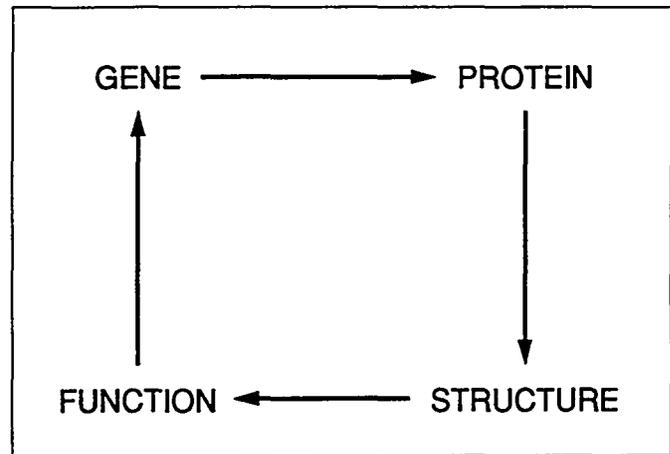


Fig. 15. Understanding Gene Function. Understanding how genes function will require analyses of the 3-D structures of the proteins for which the genes code.

Mapping Databases

The Genome Database (GDB), located at Johns Hopkins University (Baltimore, Maryland), provides location, ordering, and distance information for human genetic markers, probes, and contigs linked to known human genetic disease. GDB is presently working on incorporating physical mapping data. Also at Hopkins is the Online *Mendelian Inheritance in Man* (OMIM) database, a catalog of inherited human traits and diseases.

The Human and Mouse Probes and Libraries Database (located at the American Type Culture Collection in Rockville, Maryland) and the GBASE mouse database (located at Jackson Laboratory, Bar Harbor, Maine) include data on RFLPs, chromosomal assignments, and probes from the laboratory mouse.

Sequence Databases

Nucleic Acids (DNA and RNA)

GenBank®, the European Molecular Biology Laboratory (EMBL) sequence database, and the DNA Database of Japan (DDBJ) house over 70 Mb of sequence from more than 2500 different organisms. Compiled from both direct submissions and journal scans, GenBank is supported at IntelliGenetics (Mountain View, California) and Los Alamos National Laboratory (LANL) through a contract from the NIH National Institute of General Medical Sciences. Although responsibility for GenBank will move to the National Center for Biotechnology Information (NCBI) of the National Library of Medicine in September 1992, LANL will continue to handle direct data submissions from authors. International collaborations with EMBL and DDBJ will also continue. NCBI is also developing GenInfo, a data archive that will eventually offer integrated access to other databases.