



A Science Primer

National Center for Biotechnology Information

About NCBI	NCBI at a Glance	A Science Primer	Databases and Tools
Human Genome Resources	Model Organisms Guide	Outreach and Education	News

About NCBI
Site Map

Science Primer:

Just the Facts: A Basic Introduction to the Science Underlying NCBI Resources

Genome Mapping

Molecular Modeling

SNPs

ESTs

Microarray
Technology

Molecular Genetics

Pharmacogenomics

Phylogenetics

BIOINFORMATICS

Over the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an explosive growth in the biological information generated by the scientific community. This deluge of genomic information has, in turn, led to an absolute requirement for computerized databases to store, organize, and index the data and for specialized tools to view and analyze the data.

The completion of a "working draft" of the human genome--an important milestone in the Human Genome Project--was announced in June 2000 at a press conference at the White House and was published in the February 15, 2001 issue of the journal *Nature*.

What Is a Biological Database?

A **biological database** is a large, organized body of persistent data, usually associated with computerized software designed to update, query, and retrieve components of the data stored within the system. A simple database might be a single file containing

many records, each of which includes the same set of information. For example, a record associated with a nucleotide sequence database typically contains information such as contact name, the input sequence with a description of the type of molecule, the scientific name of the source organism from which it was isolated, and often, literature citations associated with the sequence.

For researchers to benefit from the data stored in a database, two additional requirements must be met:

- easy access to the information
- a method for extracting only that information needed to answer a specific biological question

The data in [GenBank](#) are made available in a variety of ways, each tailored to a particular use, such as data submission or sequence searching.

At NCBI, many of our databases are linked through a unique search and retrieval system, called **Entrez**. [Entrez](#) (pronounced ahn' tray) allows a user to not only access and retrieve specific information from a single database but to access integrated information from many NCBI databases. For example, the Entrez Protein database is cross-linked to the Entrez Taxonomy database. This allows a researcher to find

taxonomic information (**taxonomy** is a division of the natural sciences that deals with the classification of animals and plants) for the species from which a protein sequence was derived.

What Is Bioinformatics?

Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned. At the beginning of the "genomic revolution", a bioinformatics concern was the creation and maintenance of a database to store biological information, such as nucleotide and amino acid sequences. Development of this type of database involved not only design

Biology in the 21st century is being transformed from a purely lab-based science to an information science as well.

issues but the development of complex interfaces whereby researchers could both access existing data as well as submit new or revised data.

Ultimately, however, all of this information must be combined to form a comprehensive picture of normal cellular activities so that researchers may study how these activities are altered in different disease states. Therefore, the field of bioinformatics has evolved such that the most pressing task now involves the analysis and interpretation of various types of data, including nucleotide and amino acid sequences, protein domains, and protein structures. The actual process of analyzing and interpreting data is referred to as **computational biology**. Important sub-disciplines within bioinformatics and computational biology include:

- the development and implementation of tools that enable efficient access to, and use and management of, various types of information
- the development of new algorithms (mathematical formulas) and statistics with which to assess relationships among members of large data sets, such as methods to locate a gene within a sequence, predict protein structure and/or function, and cluster protein sequences into families of related sequences

Why Is Bioinformatics So Important?

The rationale for applying computational approaches to facilitate the understanding of various biological processes includes:

- a more global perspective in experimental design
- the ability to capitalize on the emerging technology of **database-mining** - the process by which testable hypotheses are generated regarding the

Although a human disease may not be found in exactly the same form in animals, there may be sufficient data for an animal model that allow researchers to make inferences about the process in humans.

function or structure of a gene or protein of interest by identifying similar sequences in better characterized organisms

Evolutionary Biology

New insight into the molecular basis of a disease may come from investigating the function of homologs of a disease gene in model organisms. In this case, **homology** refers to two genes sharing a common evolutionary history. Scientists also use the term homology, or homologous, to simply mean similar, regardless of the evolutionary relationship.

NCBI's [COGs database](#) has been designed to simplify evolutionary studies of complete genomes and to improve functional assignment of individual proteins.

Equally exciting is the potential for uncovering evolutionary relationships and patterns between different forms of life. With the aid of nucleotide and protein sequences, it should be possible to find the ancestral ties between different organisms. Thus far, experience has taught us that closely related organisms have similar sequences and that more distantly related organisms have more dissimilar sequences. Proteins that show a significant sequence conservation, indicating a clear evolutionary relationship, are said to be from the same **protein family**. By studying **protein folds** (distinct protein building blocks) and families, scientists are able to reconstruct the evolutionary relationship between two species and to estimate the time of divergence between two organisms since they last shared a common ancestor.

Phylogenetics is the field of biology that deals with identifying and understanding the relationships between the different kinds of life on earth.

Protein Modeling

The process of evolution has resulted in the production of DNA

sequences that encode proteins with specific functions. In the absence of a protein structure that has been determined by X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy, researchers can try to predict the three-dimensional structure using **protein or molecular modeling**. This method uses experimentally determined protein structures (**templates**) to predict the structure of another protein that has a similar amino acid sequence (**target**).

Although molecular modeling may not be as accurate at determining a protein's structure as experimental methods, it is still extremely helpful in proposing and testing various biological hypotheses. Molecular modeling also provides a starting point for researchers wishing to confirm a structure through X-ray crystallography and NMR spectroscopy. Because the different genome projects are producing more sequences and because novel protein folds and families are being determined, protein modeling will become an increasingly important tool for scientists working to understand normal and disease-related processes in living organisms.

The Four Steps of Protein Modeling

- Identify the proteins with known three-dimensional structures that are related to the target sequence
- Align the related three-dimensional structures with the target sequence and determine those structures that will be used as templates
- Construct a model for the target sequence based on its alignment with the template structure(s)
- Evaluate the model against a variety of criteria to determine if it is satisfactory

Genome Mapping

Genomic maps serve as a scaffold for orienting sequence information. A few years ago, a researcher wanting to localize a gene, or nucleotide sequence, was forced to manually map the genomic region of interest, a time-consuming and often painstaking process. Today, thanks to new technologies and the influx of sequence data, a number of high-quality, genome-wide maps are available to the scientific community for use in their research.

Computerized maps make gene hunting faster, cheaper, and more practical for almost any scientist. In a nutshell, scientists would first use a genetic map to assign a gene to a relatively small area of a chromosome. They would then use a physical map to examine the region of interest close up, to determine a gene's precise location. In light of these advances, a researcher's burden has shifted from mapping a genome or genomic region of interest to navigating a vast number of Web sites and databases.

Map Viewer: A Tool for Visualizing Whole Genomes or Single Chromosomes

NCBI's [Map Viewer](#) is a tool that allows a user to view an organism's complete genome, integrated maps for each chromosome (when available), and/or sequence data for a genomic region of interest. When using Map Viewer, a researcher has the option of selecting either a "Whole-Genome View" or a "Chromosome or Map View". The Genome View displays a schematic for all of an organism's chromosomes, whereas the Map View shows one or more detailed maps for a single chromosome. If more than one map exists for a chromosome, Map Viewer allows a display of these maps simultaneously.

Using [Map Viewer](#), researchers can find answers to questions such as:

- Where does a particular gene exist within an organism's genome?

- Which genes are located on a particular chromosome and in what order?
- What is the corresponding sequence data for a gene that exists in a particular chromosomal region?
- What is the distance between two genes?

The rapidly emerging field of bioinformatics promises to lead to advances in understanding basic biological processes and, in turn, advances in the diagnosis, treatment, and prevention of many genetic diseases. Bioinformatics has transformed the discipline of biology from a purely lab-based science to an information science as well. Increasingly, biological studies begin with a scientist conducting vast numbers of database and Web site searches to formulate specific hypotheses or to design large-scale experiments. The implications behind this change, for both science and medicine, are staggering.

[Back to top](#)

Revised: March 29, 2004.

NCBI

NLM

NIH

[Privacy Statement](#)

[Disclaimer](#)

[Accessibility](#)