



THE NATIONAL INSTITUTES OF HEALTH WWW.NIH.GOV

National Center for Biotechnology Information

You may not have heard of the National Center for Biotechnology Information (NCBI), but chances are good that you've used one of our resources, such as the PubMed database of biomedical literature or the BLAST DNA and protein sequence similarity search tool. And you're not alone: Each day, over 2 million people access NCBI databases and download a total of more than 3 terabytes (trillion bytes) of data.

Mission

Congress established NCBI in 1988 as a division of the U.S. National Library of Medicine and a national resource for molecular biology information. Since then, NCBI has become a leading source for public biomedical databases, software tools for analyzing molecular and genomic data, and research in computational biology. Technological advances and reduced costs for sequencing have led to a staggering volume of molecular data that necessitates computerized databases and analysis tools. NCBI's mission is to find new approaches to deal with the increasing volume and complexity of data, and to provide researchers with improved access to analysis and computing tools in order to better understand genes and their role in health and disease.

Organization

NCBI's programs and activities fall into three general categories: databases and software, research, and outreach/education. The Information Engineering Branch, the largest of NCBI's three branches, is responsible for the design, development, and maintenance of NCBI databases and software tools. The Computational Biology Branch conducts basic and applied research in computational, mathematical, and theoretical problems in molecular biology and genetics. The branch has two core groups: one working on research intended to improve NCBI databases and tools, and the other focusing on independent research. The Information Resources Branch provides support and training for NCBI's databases and services and manages its computing infrastructure.

Databases and Software

NCBI creates and maintains over 40 databases for the medical and scientific communities as well as the general public; these include literature, molecular, and genomic databases. NCBI's core literature database is PubMed, which provides abstracts and citations for millions of articles from thousands of biomedical journals. PubMed records include links to full-text versions of the articles (when available) from NCBI's PubMed Central (PMC) electronic archive and from journal websites, as well as links to information from other NCBI sites. For example, key scientific terms in PubMed abstracts are linked to explanatory information in NCBI's Bookshelf, a growing collection of biomedical books that can be searched electronically. Links also are provided to related information in the genomic and molecular databases. The sequencing of the human genome, completed in 2003, marked the beginning of a new era in the evolution of biological science and laid a new foundation for research in many disciplines, including the genetic causes of disease. NCBI provides integrated, linked resources for genomic information intended to aid researchers in this effort. Some of NCBI's core genomic resources are GenBank, an annotated collection of all publicly available DNA sequences; RefSeq, a curated collection of DNA, RNA, and protein sequences; dbSNP, a database of single nucleotide polymorphisms (areas of the genome that have been found to vary among humans); and the Influenza Virus Resource, which provides flu sequence data. NCBI also offers databases on protein sequences, protein structure, chromosomal aberrations in cancer, genes and gene expression, and taxonomy. One of NCBI's newest databases is PubChem, which aims to offer comprehensive information on the biological activities of small molecules, including the results of high-throughput screening to assess the effects of compounds on target proteins. The majority of NCBI databases are linked through its Entrez search engine,

which provides integrated access to literature, sequence, mapping, taxonomy, and structural data. The system facilitates the process of research and discovery by linking records and terms to related information across NCBI databases. Another key tool is BLAST, which, in a matter of seconds, identifies similar gene and protein sequences to the sequence being queried.

Basic and Applied Research

NCBI has a multi-disciplinary research group composed of molecular biologists, biochemists, computer scientists, mathematicians, research physicians, and structural biologists concentrating on basic and applied research in computational molecular biology. Together they are studying fundamental biomedical problems, including comparative genomics, proteomics, molecular evolution, and disease. These investigators not only make important contributions to basic science but also serve as a wellspring of new methods for applied research activities, including enhancements to NCBI's publicly available databases and software tools. To read more about NCBI's research, visit www.ncbi.nlm.nih.gov/CBBresearch.

Outreach and Education

NCBI provides a variety of educational materials and programs to aid researchers, librarians, educators and others in the use of its databases and software tools. As sequence tools and data have grown more complex, the need for training has increased, and NCBI has expanded its educational efforts accordingly. Training classes are held frequently around the country and on the NIH campus in Bethesda, Maryland.

New Directions

Genome-wide association (GWA) studies: In December 2006, NCBI launched a GWA database called dbGaP. The database will provide a central location for archiving and distributing phenotype and genotype data from a variety of studies ranging from heart disease, women's health, and diabetes, to environmental factors in disease. Connecting phenotype and genotype data offers the potential for increased understanding of basic biological processes affecting human health, improved disease prediction and patient care, and ultimately the realization of personalized medicine. The initial release of dbGaP contains data from an age-related macular degeneration study and from a Parkinson's disease study. Several additional studies are expected to be added in the coming year, including the Framingham SHARE Study, which will associate genotype data with phenotype data collected in the landmark Framingham Heart Study.

The Discovery Initiative: NCBI's databases are highly integrated. For example, a PubChem record may have links to records for chemically similar compounds, to a protein structure that was crystallized along with the chemical, to relevant journal articles, etc. These records, in turn, are linked to other related records. While this extensive network of links provides users with vast opportunities for exploration and for making the kinds of connections that underlie the discovery process, many users do not go beyond retrieving the basic results from a search query. The Discovery Initiative aims to improve the presentation of results so that users are more readily drawn to these related data that could lead to serendipitous discoveries.

To learn more about NCBI, visit our website at www.ncbi.nlm.nih.gov or email us at: info@ncbi.nlm.nih.gov.

David J. Lipman, M.D.
Director,
National Center for Biotechnology Information

