An Integrated Database of Biological Sequences, Scientific Literature, and 3-Dimensional Structure

Stephen H. Bryant
bryant@ncbi.nlm.nih.gov

National Center for Biotechnology Information (NCBI),
National Library of Medicine, National Institutes of Health,
Bethesda, MD 20894, USA.

Abstract

NCBI maintains a database combining biological sequences and associated Medline citations. This information may be accessed over the Internet via an easy-to-use browser, Entrez. With Entrez one may rapidly discover what is known about a biological molecule, exploring "links" between sequences and citations, the homologous "neighbors" of a sequence, and "neighbor" citations, which discuss the structure and function of related molecules. A recent addition to Entrez is information describing 3-dimensional structure, allowing one to examine directly the structure of a biological molecule or its homologs.

The 3-dimensional structure database used by Entrez is called MMDB, for Molecular Modeling Data Base. It consists of messages in the ASN.1 language, which are translated automatically to in- memory data structures using C routines available in the NCBI toolkit. The MMDB data specification gives a complete description of the chemical structure of a macromolecular assembly, with unambiguous linkage to atomic coordinates. This data organization is intended to facilitate the computations required for homology modeling, which is based on alignment of the residue sequence, or, more generally, on comparison of the chemical graphs of different molecules, including their non-polymer components.

NCBI supports an active research program in comparative molecular analysis techniques which make use of 3-dimensional structure information. Computational methods for sequence-structure "threading" provide a means to generate model structures for proteins which are very distantly related to a protein of known structure. Methods for rapid structure-structure comparison provide a means to detect architectural similarities and/or distant evolutionary relationships among proteins of known structure. These methods currently make use of the NCBI "PKB" research database. In the near future, however, they will operate as client software accessing MMDB, and in this way provide examples of how applications developed elsewhere may make use of this resource.