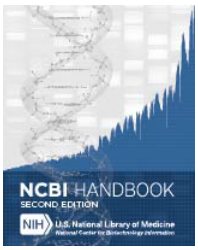
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: Sayers E. NCBI Protein Resources. 2013 Nov 12 [Updated 2013 Nov 21]. In: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013-.
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



NCBI Protein Resources

Eric Sayers, PhD[1]

Created: November 12, 2013; Updated: November 21, 2013.

Introduction

Proteins are the machines of life. They perform almost all of the processes necessary to sustain life, and also form a variety of structural and connective materials that constitute the bulk of our bodies and those of all other organisms. While we can think of a single protein as a discreet polymer of amino acids, the functional form of many proteins in the cell is actually a complex of several individual polymer chains, all working in concert to do a particular task. At NCBI, we therefore provide several resources that represent these various aspects of proteins, ranging from the sequences of individual chains to functional classifications of large protein families.

Protein

The Protein database is the most fundamental NCBI resource for proteins. It contains text records for individual protein sequences derived from a variety of sources, including GenBank, the NCBI Reference Sequence (RefSeq) project and several external databases including UniProtKB/SWISS-Prot and the Protein Data Bank (PDB). It is important to understand that the sequences contained in almost all Protein records (with the exception of PDB) are conceptual translations of an RNA coding sequence, meaning that no one determined the protein sequence experimentally, but rather inferred the sequence from the corresponding RNA. Protein records are available in several formats (including FASTA and XML) and are linked to many other NCBI resources, allowing users to find relevant data such as literature, DNA/RNA sequences, genes, biological pathways and expression and variation data. We also provide pre-computed sets of identical and similar proteins for each sequence as determined by the BLAST algorithm. The BLAST Link (Blink) tool provides a graphical view of these pre-computed sets and provides a variety of filtering tools and links to multiple alignment views.

Structure

The Structure database contains 3D coordinate sets for experimentally-determined structures in PDB. At NCBI, we import these data from PDB and format the data for viewing in Cn3D, the NCBI structure viewer. We also calculate structural similarity between all of these records using the VAST algorithm and allow users to view superpositions of highly similar structures. We also link these structure records to the corresponding sequence records in the Protein database, to literature articles where the structure was reported, and to information about any ligands present in the structure.

Author Affiliation: 1 NCBI; Email: sayers@ncbi.nlm.nih.gov.

Corresponding author.

Conserved Domains (CDD)

The Conserved Domain database (CDD) is a collection of sequence profiles that represent highly conserved domains within protein sequences. Very often these domains have a particular function that is shared between those sequences that contain it. Typically one identifies the presence of a conserved domain in a sequence using the CD-Search tool, and these results provide access to sequence alignments, distance trees, selected literature, and structural views that highlight important elements within the domain. While we curate our own set of domain records, CDD also contains records from external resources such as Pfam and SMART. NCBI provides links to precomputed CD-Search results for all Protein records and also displays such results on each record in the Structure database. CD-Search results are also provided in several other NCBI displays, including protein BLAST results and the graphical sequence viewer.

Protein Clusters

The Protein Clusters database contains sets of proteins annotated on RefSeq genomes from prokaryotes, viruses, fungi, plants, and organelles (mitochondria and chloroplasts). Each protein cluster record consists of a set of protein sequences clustered according to the maximum alignments calculated by BLAST between the individual sequences. We then hypothesize that the proteins within each set are homologous, and on this basis use these clusters to support functional annotation of new RefSeq genomes.