All things protein...

Marilyn D Ritchie, PhD
Professor, Biochemistry and Molecular Biology
Director, Center for Systems Genomics
The Pennsylvania State University
NCBI Protein Tools
NCBI Protein Tools

Protein

The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and function.
SCN5A protein [Homo sapiens]

GenBank: AA144622.1
Fasta  Graphics

LOCUS        AA144622  1983 aa linear PRI 18-MAR-2009
DEFINITION   SCN5A protein [Homo sapiens].
ACCESSION    AA144622
VERSION      AA144622.1 GI: 219521582
DBSOURCE     accession NC144621.1
KEYWORDS     MGC.
SOURCE       Homo sapiens (human)
ORGANISM     Homo sapiens
             Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
             Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
             Catarrhini; Hominidae; Homo.
REFERENCE     1  (residues 1 to 1983)
             Klausner,R.D., Collins,F.S., Wagner,L., Shenmen,C.M., Schuler,G.D.,
             Altschul,S.F., Zeeberg,B., Bueltow,K.K., Schaefer,C.F., Bhat,N.K.,
             Hopkins,R.F., Jordan,M., Moore,T., Max,S.I., Wang,J., Hsieh,F.,
             Diatchenko,L., Marusina,K., Farma,A.A., Rubin,G.M., Hong,L.,
             Stapleton,M., Soares,M.B., Bonaldo,M.F., Casavant,T.T.,
             Scheetz,T.E., Brownstein,M.J., Usdin,T.B., Toshiyuki,S.,
             ...
NCBI Protein Tools

BLAST finds regions of similarity between biological sequences.

New DELTA-BLAST, a more sensitive protein-protein search

BLAST Assembled Genomes

Find Genomic BLAST pages:

Enter organism name or id—completions will be suggested

GO

Basic BLAST

Choose a BLAST program to run.

- nucleotide blast
  - Search a nucleotide database using a nucleotide query
  - Algorithms: blastn, megablast, discontiguous megablast

- protein blast
  - Search protein database using a protein query
  - Algorithms: blastp, psi-blast, phi-blast, delta-blast

- blastx
  - Search protein database using a translated nucleotide query

- tblastn
  - Search translated nucleotide database using a protein query

- tblastx
  - Search translated nucleotide database using a translated nucleotide query
About UniProt

The mission of UniProt is to provide the scientific community with a comprehensive, high quality and freely accessible resource of protein sequence and functional information.

UniProt is comprised of four components, each optimised for different uses:

1) The **UniProt Knowledgebase (UniProtKB)** is the central access point for extensive curated protein information, including function, classification, and cross-reference.

   It consists of two sections:

   - **UniProtKB/Swiss-Prot** which is manually annotated and is reviewed and
   - **UniProtKB/TrEMBL** which is automatically annotated and is not reviewed.
Uniprot

• Mission of UniProt is to provide comprehensive, high quality and freely accessible resource of protein sequence and functional information

• UniProt Knowledgebase (UniProtKB)
  • central access point for extensive curated protein information, including function, classification, and cross-reference
    • UniProtKB/Swiss-Prot which is manually annotated and is reviewed
    • UniProtKB/TrEMBL which is automatically annotated and is not reviewed
Uniprot

• UniProt Reference Clusters (UniRef) databases
  • clustered sets of sequences from the UniProtKB and selected UniProt Archive records to obtain complete coverage of sequence space at several resolutions while hiding redundant sequences

• UniProt Archive (UniParc)
  • comprehensive repository, used to keep track of sequences and their identifiers

• UniProt Metagenomic and Environmental Sequences (UniMES)
  • repository specifically developed for metagenomic and environmental data.
Uniprot

How to use this tool

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences, which can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

1. Enter either a protein or nucleotide sequence or a UniProt identifier (e.g. P00750 or A4_HUMAN or UPI000000001) into the form field.
2. Optionally, change the program parameters with the dropdown menus under the form.
3. Click the Run BLAST button.

BLAST

Target database: UniProtKB
E-Threshold: 10
Matrix: Auto
Filtering: None
Gapped: yes
Hits: 250

Run Blast in a separate window.
**Align**

**Protein sequences (FASTA) or UniProt identifiers**

1. Enter either protein sequences in FASTA format or UniProt identifiers into the form field, for example:
   - TPA_HUMAN
   - TPA_PIG

2. Click the **Run Align** button.
Aligning multiple sequences

- Highlights areas of similarity which may be associated with specific features that have been more highly conserved than other regions
  - These regions in turn can help classify sequences or to inform experiment design
- Important step for phylogenetic analysis, which aims to model the substitutions that have occurred over evolution and derive the evolutionary relationships between sequences
- Clustal Omega improves on ClustalW in a number of ways - alignment accuracy and improved scaling to many sequences are the main results
PDB

• Single worldwide repository of information about the 3D structures of large biological molecules, including proteins and nucleic acids
• Understanding the shape of a molecule helps to understand how it works
• Can be used to help deduce a structure's role in human health and disease, and in drug development
• Structures in the archive range from tiny proteins and bits of DNA to complex molecular machines like the ribosome
• PDB archive is available at no cost to users
• PDB archive is updated each week at the target time of Wednesday 00:00 UTC (Coordinated Universal Time)
• Most recent release is timestamped and linked on every page in the top right header.
Molecule of the Month

Apoptosomes

"To be, or not to be"—that question is continually being asked by each of your cells. Your cells are preprogrammed to die on command. This is essential during development of large organisms like ourselves, where cells work together, growing and dying to shape our complicated bodies. It is also essential throughout our adult lives, to remove damaged or infected or cancerous cells. The machinery for cell death is always silently present in cells, but can be instantly mobilized if the choice is made to die.

Full Article

Protein Structure Initiative Featured System
Peptidyl-carrier Proteins

Bacteria are creative chemists, constantly discovering new ways to build unusual molecules for exploiting natural resources and fighting
### Query Parameters:

**Taxonomy**
- Tree Search for Bacteria (eubacteria)

### Query Refinements: Select an item or pie chart

<table>
<thead>
<tr>
<th>Organism</th>
<th>Taxonomy</th>
<th>Experimental Method</th>
<th>X-ray Resolution</th>
<th>Release Date</th>
<th>Polymer Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Escherichia coli (4888)</td>
<td>Bacteria only (37229)</td>
<td>X-ray (36100)</td>
<td>less than 1.5 Å (2857)</td>
<td>before 2000 (3123)</td>
<td>Protein (36115)</td>
</tr>
<tr>
<td>Escherichia coli K-12 (2076)</td>
<td>Bacteria/Eukaryota (770)</td>
<td>Solution NMR (1907)</td>
<td>1.5 - 2.0 Å (12776)</td>
<td>2000 - 2005 (6451)</td>
<td>Mixed (2102)</td>
</tr>
<tr>
<td>Bacillus subtilis (1037)</td>
<td>Bacteria/Other (182)</td>
<td>Electron Microscopy (270)</td>
<td>2.0 - 2.5 Å (12172)</td>
<td>2005 - 2010 (12673)</td>
<td>RNA (120)</td>
</tr>
<tr>
<td>Thermus thermophilus HB8 (995)</td>
<td>Bacteria/Viruses (87)</td>
<td>Hybrid (28)</td>
<td>2.5 - 3.0 Å (5737)</td>
<td>2010 - today (16093)</td>
<td></td>
</tr>
<tr>
<td>Mycobacterium tuberculosis (925)</td>
<td>Bacteria/Archaea (32)</td>
<td>Solid-State NMR (19)</td>
<td>3.0 and more Å (2569)</td>
<td>this year (2752)</td>
<td></td>
</tr>
<tr>
<td>Staphylococcus aureus (798)</td>
<td>Bacteria/Eukaryota/Viruses (14)</td>
<td>Neutron Diffraction (9)</td>
<td>more choices...</td>
<td>this month (84)</td>
<td></td>
</tr>
<tr>
<td>Pseudomonas aeruginosa (677)</td>
<td>Bacteria/Eukaryota/Other (12)</td>
<td>Electron Crystallography (6)</td>
<td>more choices...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other (27135)</td>
<td>Other (14)</td>
<td>Fiber Diffraction (1)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Enzyme Classification
- 3: Hydrodrolases (7032)
- 2: Transferases (5586)
- 1: Oxidoreductases (4683)
- 4: Lyases (2077)
- 5: Isomerases (1313)
- 6: Ligases (980)

### SCOP Classification
- Alpha and beta proteins (a/b) (6190)
- Alpha and beta proteins (α+β) (4077)
- All beta proteins (3250)
- All alpha proteins (2485)
- Multi-domain proteins (alpha and ... (569)
- Small proteins (379)
- Membrane and cell surface proteins (377)
- Other (338)

### Protein Symmetry
- Asymmetric (18125)
- Cyclic (14374)
- Dihedral (4600)
- Tetrahedral (225)
- Octahedral (78)
- Helical (48)
- Icosahedral (12)
- More choices...

### Protein Stoichiometry
- Homomer (18766)
- Monomer (15804)
- Heteromer (2888)
- More choices...

### Membrane Proteins
- Alpha-Helical (1032)
- Beta-Barrel (331)
- Monotopic Membrane Proteins (84)
PDB
Pfam

Pfam 27.0 (March 2013, 14831 families)

The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). More...

Quick Links

- **Sequence Search**: Analyze your protein sequence for Pfam matches
- **View a Pfam Family**: View Pfam family annotation and alignments
- **View a Clan**: See groups of related families
- **View a Sequence**: Look at the domain organisation of a protein sequence
- **View a Structure**: Find the domains on a PDB structure
- **Keyword Search**: Query Pfam by keywords

**Jump to**

Enter any accession or ID to jump to the page for a Pfam family or clan, UniProt sequence, PDB structure, etc.

Or view the help pages for more information

Recent Pfam blog posts

- [Moving to xfam.org](http://pfam.xfam.org/) (posted 1 May 2014)

http://pfam.xfam.org/
DIP

Database of Interacting Proteins

The DIP™ database catalogs experimentally determined interactions between proteins. It combines information from a variety of sources to create a single, consistent set of protein–protein interactions. The data stored within the DIP database were curated, both, manually by expert curators and also automatically using computational approaches that utilize the knowledge about the protein–protein interaction networks extracted from the most reliable, core subset of the DIP data. Please, check the reference page to find articles describing the DIP database in greater detail.

This page serves also as an access point to other projects related to DIP, such as The Database of Ligand-Receptor Partners (DLRP) and JDIP.

DIP Pages

- **NEWS**: Announcements about the most recent additions and changes to the database.
- **REGISTRATION/ACCOUNT**: Registration and account maintenance. Registration is required to gain access to most of the DIP features. Registration is free to the members of the academic community. Trial accounts for the commercial users are also available. Please, consult Terms of Use for further details.
- **STATISTICS**: Detailed information about the current state of the database as well as some statistics on server usage.
- **SATELLITES**: DIP-related projects, such as DLRP and JDIP.
- **SERVICES**: DIP-derived services.
- **ARTICLES**: DIP in press. Both, papers published on DIP as well as a list of publications referring to DIP.
- **SEARCH**: Database search. This is the starting point of the database exploration. Once the initial protein is found through keyword or sequence searches the interaction network can be explored by interactively following the interaction links.

http://dip.doe-mbi.ucla.edu/dip/Main.cgi
DM²

Domain Mapping of Disease Mutations

Search:

Examples:
Protein (Accessions: NP_009228, Gis: 237681121)
Gene (GeneIDs: 3757, Gene Symbols: BRCA1)
Domain (Names: SapB, CD Accessions: cd00007)
Mutation (dbSNP: rs104886112, OMIM: 225280, SwissProt: VAR_016877)
Description (binding protein, hinge region, etc.)
Disease Name (Breast Cancer, Parkinson Disease, etc.)

The Worldwide Protein Data Bank (wwPDB) consists of organizations that act as deposition, data processing and distribution centers for PDB data. Members are: RCSB PDB (USA), PDBe (Europe) and PDBj (Japan), and BMRB (USA). The wwPDB’s mission is to maintain a single PDB archive of macromolecular structural data that is freely and publicly available to the global community.

18-September-2014

Inclusion of Large Structures in the Main PDB Archive

The wwPDB recently combined entries that represent large structures (such as ribosomes) across multiple PDB files (SPLIT entries) into single files. These combined structures have been issued new PDB IDs and are represented in the archive in both PDBx/mmCIF and PDBML formats.

more

FULL NEWS

Questions? info@wwpdb.org

Questions???